Building and Visualizing a Database from Wildlife Trafficking Reports with NLP

Authors

Devin Coughlin

Maylee Gagnon

Victoria Grasso

Faculty Advisors

Renata Konrad

Kyumin Lee

Graduate Advisor

Guanyi Mou

Sponsors

Meredith Gore

Patricia Raxter

A Major Qualifying Project
Submitted to the Faculty of the
WORCESTER POLYTECHNIC INSTITUTE
In partial fulfillment of the requirements for the
Degree of Bachelor of Science in Computer Science

Abstract

Experts combating wildlife trafficking manually sift through articles about seizures and arrests, which is time consuming and makes identifying trends difficult. In this project, we applied natural language processing techniques to extract data from reports published by the Eco Activists for Governance and Law Enforcement. We expanded on spaCy's pre-trained pipeline and added custom named entities which enabled us to identify approximately 40% of wildlife trafficking events. We created visualizations to display trends in the extracted data, which are accessible on our website, *Wildlife Trafficking in Africa*. This project is an initial solution to automatically extract, collect, and display, wildlife trafficking data for experts to easily analyze.

Acknowledgments

We would like to thank our sponsors Meredith Gore, PhD and Patricia Raxter, PhD for their help and guidance during this project. We also want to thank Guanyi Mou, graduate student at WPI, for his support. Finally, we wish to thank our faculty advisors, Renata Konrad, PhD and Kyumin Lee, PhD for their continual advice and encouragement. Our project would not have been possible without all of their feedback and assistance.

This material is based upon work supported by the National Science Foundation under grant NSF-2039951. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

List of Figures

1 Introduction

The International Consortium on Combating Wildlife Crime, a leading expert in the field, defines wildlife trafficking as “the taking, trading (supplying, selling, or trafficking), importing, exporting, and processing, possessing, obtaining and consumption of wild fauna and flora, including timber and other forest products, in contravention of national or international law” (*Wildlife Crime | CITES*, n.d.). The United Nations Office of Drugs and Crime (UNODC) recognized nearly 6,000 species being illegally traded as of 2019 (UNODC, 2020). Pangolins, tigers, elephants, and their associated products, such as scales, skins, and tusks, are examples of commonly trafficked commodities. As of 2018, illegal wildlife trade impacts at least 149 countries and territories (UNODC, 2020). The World Wildlife Fund valued wildlife trade to be at least $19 billion per year, making it the fourth largest illegal global trade (WWF, 2012). Wildlife trafficking is a global issue that endangers the targeted flora and fauna, ecosystems, and human welfare (*Wildlife, Forest & Fisheries Crime Module 1 Key Issues*, 2019).

Many organizations work to combat illegal wildlife trafficking and to conserve wildlife including the World Wildlife Fund, International Union for Conservation of Nature, and the Eco Activists for Governance and Law Enforcement, also known as EAGLE. EAGLE aims to impede wildlife trafficking and related corruption through civic activism and collaborating with governments. They do so through investigations, arrests, prosecutions, and publicity (EAGLE Network, 2022). As a result of EAGLE’s efforts, more than 2,000 wildlife traffickers have been jailed. The organization covers nine countries, Cameroon, Congo, Gabon, Togo, Senegal, Benin, Côte D’Ivoire, Burkina Faso, and Uganda. EAGLE releases monthly briefings containing details of their recent work including seizures and wildlife trafficker arrests. This data could be leveraged to identify trends and insights about wildlife trafficking.

1.1 Project Goal

The motivation behind this project arose from discussion with domain experts Patricia Raxter, PhD and Meredith Gore PhD, who emphasized the issue of wildlife trafficking especially within Africa. Patricia Raxter, PhD is a Subject Matter Expert on transnational wildlife crime. Meredith Gore, PhD is an associate professor at the University of Maryland whose expertise is in the human dimensions of wildlife management, and environment and resource policy.

There is a need for an easy process to read through wildlife trafficking reports and media stories to find trends across illegal wildlife trade seizures. Currently this is a manual process, as investigators often have to sift through data manually. Experts interviewed acknowledged the current process is tedious and articles often are not shared between all the experts. Since the information in these briefings is disjointed it can be difficult for experts to parse through and identify insights. The reports, as seen in Figure 1, contain descriptions of wildlife trafficking seizures within different countries across a certain period.

Figure 1 Excerpt of Monthly Briefing

Summary

7 traffickers arrested with several human bones in Congo. They were arrested during 3 separate operations during a major crackdown on the network. Their illegal activity first came to light in May 2021.

2 traffickers arrested with 2 elephant tusks, weighing just under 40 kg in Congo. One of the traffickers is a pastor and the second is a military man. The military man was extremely violent and tried to escape during the arrest. He attempted to corrupt officers when behind bars.

3 traffickers arrested in Congo with 18 African Grey Parrots. 2 of the traffickers are from the Democratic Republic of Congo. They were arrested during 3 follow up operations from last month's operation when 2 were arrested with 28 African grey parrots.

2 traffickers arrested with 2 elephant tusks in Gabon. They arrived the scene of transaction with each carrying a backpack with concealed ivory. The tusks were cut into 6 pieces.

Note. Eagle Network Briefing December 2021 Page 1, 2021, Eagle (https://www.eagle-enforcement.org/data/files/eagle-briefing-december-2021-public.pdf).

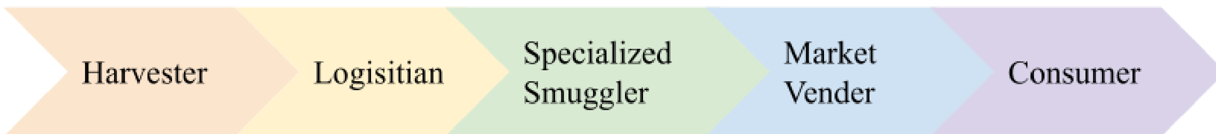
Typically, several seizures happen within a few countries over the course of a single month, this means those combating wildlife trafficking have an enormous amount of data to find, track and analyze. In response, our goal was to design and develop a pipeline to extract data from the EAGLE reports using machine learning techniques and effectively display the data to help experts in the field examine instances of wildlife trafficking.

2 Background

2.1 Roles within Wildlife Trafficking System

Wildlife trafficking market supply chains involve many people assuming various roles. In their 2016 article, Phelps et al., developed a typology framework of the key actors involved in wildlife trafficking. The three main roles were harvesters, intermediaries, and consumers. Harvesters gather the plants and animals from the wild. There are a variety of reasons people choose to harvest such as for individual use, recreation, and commercial gain. Intermediaries transport and sell goods and consist of seven sub-roles: logistician, specialized smuggler, government colluder, third party, processor, launderer, and vendor. Logistician's responsibilities include financing and planning the transportation of goods. This may be done directly or at a distance. Specialized smugglers excel at moving illegal goods across borders without being intercepted. Smugglers may involve government colluders through bribery, and other third parties such as for transportation. Most of these actors intentionally get involved, however sometimes the third parties unknowingly support trade. Processors convert the wildlife into other products such as by skinning animals. Launderers and vendors sell the products to consumers and other intermediaries. They may complete sales in person through markets or digitally through online platforms. Intermediaries are typically the ones embroiled in wildlife trafficking seizures. The last category of actors that Phelps et al. (2016) identified were consumers, which use the final goods. An example of a supply chain an illicit good may pass through can be seen below in Figure 2.

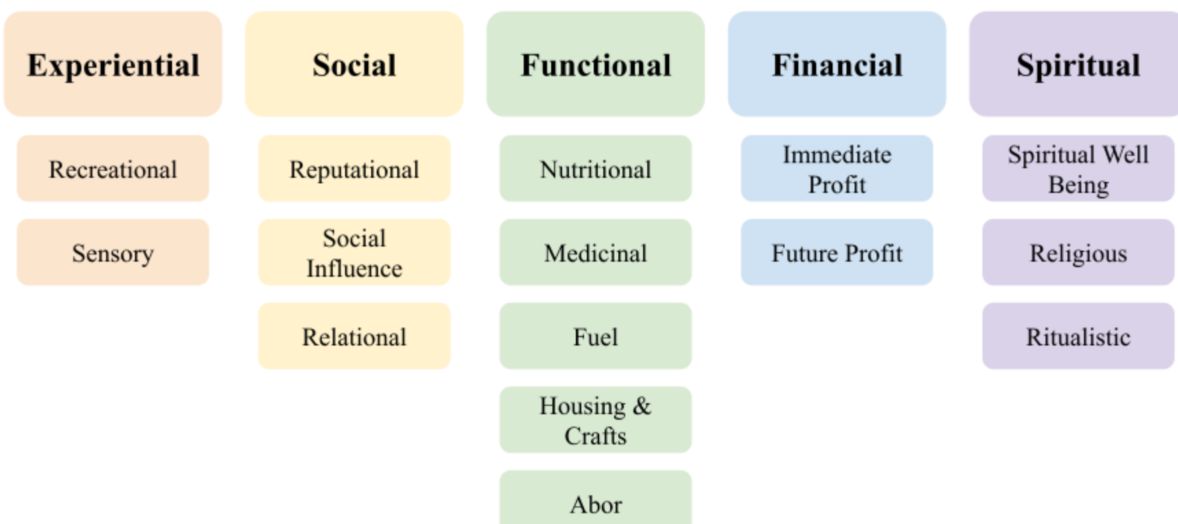
Figure 2 Example of Supply Chain



2.2 Motivations for Committing Wildlife Trafficking

The demand for illegal wildlife stems from a variety of consumer intentions. Thomas-Walters et al. (2021) identified a typology of wildlife consumer motivations. The core five motivational categories, shown in Figure 3, are experiential, social, functional, financial, and spiritual (Thomas-Walters et al., 2021).

Figure 3 Framework for the Motivation Behind the Use of Wildlife Products



Note. Adapted From "Motivations for the use and consumption of wildlife products," by Thomas-Walters, L., Hinsley, A., Bergin, D., Burgess, G., Doughty, H., Eppel, S., MacFarlane, D., Meijer, W., Lee, T. M., Phelps, J., Smith, R. J., Wan, A. K. Y., & Veríssimo, D., 2021, *Conservation Biology*,, *35*(2), p. 483–491 (https://doi.org/10.1111/cobi.13578)

These categories can be demonstrated in different scenarios. For example, pangolins are highly trafficked animals across Asia and Africa, despite an international trade ban on all eight species since 2017. Pangolin usage falls under the functional category as the meat is eaten as food and the scales are used in traditional medicine (Segniagbeto et al., 2021). An example of a social motivation is a consumer collecting elephant ivory. Elephant ivory is highly trafficked, and many believe it demonstrates status (Segniagbeto et al., 2020). Some traffickers are motivated by financial aspirations as wildlife trafficking is a substantial illegal business. It is difficult to quantify a monetary value, however one estimate by the UN in 2016 valued wildlife trafficking to be between $7 billion and $23 billion a year (*TRAFFIC | Wildlife Crime*, n.d.). Although these motivations for wildlife trafficking may draw people to participate, there are some serious consequences to the illegal activity.

2.3 Consequences of Wildlife Trafficking

Wildlife trafficking has widespread implications for species, the spread of zoonotic disease and convergence with other types of illicit crime. Wildlife trafficking can diminish species populations which is especially threatening to endangered species. For instance, all four Asian pangolin species are on the International Union for Conservation of Nature (IUCN) Red List of Threatened Species as “Critically Endangered” or “Endangered” and have declined approximately 90% since the 1960’s (Ingram et al, 2017). One study estimated that at least 850,602 pangolins have been taken from Africa, based only on seizure data since 2009 (USAID WA BiCC, 2021). These seizures likely account for as few as 10% of all illicit shipments suggesting a much larger true number (USAID WA BiCC, 2021). Consumers’ desire for the best specimens, such as the largest or most colorful, prevents the fittest from reproducing, potentially causing reduced fitness in the following generations (Rosen and Smith, 2010). Selective poaching of the most attractive Mexican parrot species based on characteristics such as body size, coloration, and ability to imitate human speech, contributes to their current highly threatened status (Tella & Hiraldo, 2014). Consumer demands may also create sex-ratio imbalances within populations leading to reduced reproduction rates (Schloenhardt, 2020). This can result in even lower population size and further harm to the species populations. However, wildlife trafficking threatens the welfare of not only the target animals but also by standing animals in some cases. For instance, nets used to catch the endangered totoaba macdonaldi fish frequently also catch the vaquita porpoise (Schloenhardt, 2020). Wildlife trafficking also threatens the stability of ecosystems due to the introduction of invasive species. One model of alien reptiles successfully settling in Australia predicted that 5 out of 28 species would likely become established if released (García-Díaz et al., 2017).

Another consequence of wildlife trafficking is the possible spread of zoonotic diseases. Multiple diseases have originated from animals including Nipah virus in Malaysia, Hendra virus in Australia, and Ebola virus in Africa (Medicine, 2015). A 2012 report analyzed samples of nonhuman primate and rodent species, such as chimpanzee and cane rats, from several international airports. The nonhuman primate samples contained retroviruses (simian foamy virus) and/or herpesviruses (cytomegalovirus and lymphocryptovirus) which indicate that illegal bushmeat may lead to spread of pathogens (Smith et al, 2012).

Wildlife trafficking coincides with other forms of criminal activity including but not limited to, arms trafficking, drug trafficking, cash smuggling, counterfeit goods, and criminal violence (Anagnostou & Doberstein, 2021). Anagnostou and Doberstein (2021) identified diversification, enabling crimes, parallel trafficking, shared smuggling routes, and threat finance, as the five most frequent forms of convergence. In their literature review they found the most commonly reported form of convergence was criminal networks diversifying their profits by trafficking both rhino horns and drugs (Anagnostou & Doberstein, 2021). The financial potential of dealing in natural resources attracts criminal organizations and motivates them to traffic items such as timber and endangered species (van Uhm & Nijman, 2020). In their study, Van Uhm and Nijman (2020) found the Congo Basin to be a hotspot for smugglers which traffic not only weapons and ammunition but also illegal timber and animal products (van Uhm & Nijman, 2020).

2.4 Wildlife Trafficking in Africa

Although wildlife trafficking is a global issue, Africa is a region of particular concern as it contains major exporting hubs (Environmental Investigation Agency, 2020). The EAGLE organization works with nine countries in Africa. As seen in Figure 4 the countries are Cameroon, Congo, Gabon, Togo, Senegal, Benin, Côte D’Ivoire, Burkina Faso, and Uganda.

Figure 4 Map of the Eagle Network

Note. Created with Tableau. (2021). Tableau Desktop (Professional Edition) [Computer software]. https://www.tableau.com/

Africa is home to many types of flora and fauna. As of 2015 2,471 amphibian, bird, and mammal species are known to live in West Africa (Mallon et al., 2015). In one study a team visited the largest open-air market in West Africa, Dantokpa, which is in Cotonou, Benin. There they observed many animal goods being sold including baboon skulls, african grey parrots, chameleons, leopard skins, pangolins, and tortoises (Balinga & Stroud, 2020). Species such as the grey parrot and pangolin fall under the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) Appendix I list of threatened animals. The Environment Investigation Agency identified West and Central Africa as a prominent source of elephant ivory and pangolin scales to Asia (Environmental Investigation Agency, 2020). Wildlife trafficking criminals look for environments that are well connected to consumer markets and present minimal law enforcement risks. The Economic Community of West African States (ECOWAS) agreement implemented free movement of people and goods in West Africa contributing to more porous borders (ENACT & INTERPOL, 2018). Aside from wildlife trafficking, there are other forms of organized crime present in West Africa including drug trafficking, human trafficking, counterfeit goods, and stolen motor vehicles (ENACT & INTERPOL, 2018). There is evidence of well-established routes used for moving these various illegal goods (ENACT & INTERPOL, 2018). This shows the feasibility of wildlife trafficking occurring in the region. Wildlife trafficking is not a new phenomenon in Africa. The region has had well established commercial domestic markets for ivory and other wildlife products for centuries (Environmental Investigation Agency, 2020).

3 Related Work

3.1 Machine Learning Techniques Identifying Illicit Sales

There are some studies that monitor illegal wildlife trade online using machine learning techniques. One earlier research article from 2012 utilizes natural language processing and machine learning algorithms to obtain wildlife trade reports from the Internet (Hansen et al., 2012). These wildlife reports were then compiled into a database and displayed on an interactive visualization where users can sort through the reports. This interactive visualization organized reports by location and provided three category tags to improve filtering based on what was discussed in the report. The categories were (i) Breaking, for articles with live animal or wildlife product seizures with specific date, location, and species information; (ii) Warning, for articles with information about historically known illegal wildlife trade activities; and (iii) Context, for alerts on illegal wildlife trade but it does not include specific incident information on product seizures or trade routes used. The interactive visualization was displayed on a publicly available website; however, at the time of writing, it is no longer available to access. Our intent was to replicate the visualization created by Hansen et al. (2012), by organizing information by location and providing law enforcement and scholars the details needed to fight illegal wildlife trade. Patel et al., (2015) utilized Hansen et al.'s (2012) data visualization to analyze countries that support the illegal wildlife trade. Patel et al. (2015) analyzed how interventions in different countries would impact illegal wildlife trade. This study showed the importance of creating a centralized database of instances of wildlife trafficking for law enforcement to monitor.

Miller et al., (2019) used machine learning along with web scraping and data visualization to create a process called the Dynamic Data Discovery Engine (DDDE). This process had six stages to best detect online transactions and sales activities involving any CITES-listed plant or animal. DDDE went through three case studies that analyzed a different animal or plant. The study conducted by Miller et al. (2019) differs from what we wanted to achieve because we wanted our process to analyze all animals or plants in the EAGLE reports instead of selecting certain animals. Miller et al. (2019) displayed data visualizations based on their data from the orchids, pangolins, and ivory case studies. They used word clusters to organize the data found on the websites collected during the case studies. As we extracted data from the EAGLE reports, we did not do any web scraping and only focused on using machine learning and data visualization to create a well categorized database along with a detailed visualization.

Another relevant study utilized machine learning, specifically natural language processing, to analyze and classify Twitter messages with wildlife trafficking keywords (Xu et al., 2019). The authors collected tweets containing keywords and code words related to elephant ivory and pangolins. Out of 138,357 tweets found, only 53 tweets were suspected to promote the sale of elephant ivory. In contrast, not a single pangolin-related tweet was found during their study. Although we used named entity recognition, which is a form of natural language processing, we extracted data from EAGLE reports and not the internet. This article's study acts as one approach to detecting wildlife trafficking.

A conference paper examined using machine learning techniques to identify wildlife trafficking on social media but considered visual content as well as verbal (Nalluri et al., 2021). The study used neural networks and deep learning to find code words related to wildlife trade on Twitter by using images of Rhino species to train the model. Although we did not train any machine learning models, it is interesting to note how the authors were able to work with images to extract information about wildlife trade.

Web scraping, natural language processing, deep learning, and visualization techniques can be used to collect information on other forms of illicit trade. For example, Mackey et al., (2020) used these techniques to collect, analyze, identify, and enable reporting of suspected fake, counterfeit, and unapproved COVID-19-related health care products from Twitter and Instagram. They found over 6 million tweets from March to April associated with suspect marketing and sale of COVID-19 health products. Although the authors used a web scraper to obtain the tweets and Instagram posts, they used an "unsupervised" NLP approach to group and summarize all the content of filtered social media data. The researchers also used a deep learning classifier to conduct supervised classification of all filtered social media posts. The project resulted in a prototype data dashboard that could be used by public health officials, drug regulatory authorities, and law enforcement agencies to visualize characteristics of illegal selling posts (including location, a list of top-related hashtags, images of suspect products, and other metadata).

The above body of literature demonstrates what has already been attempted regarding the use of machine learning techniques to analyze and visualize data. Although some of the discussed studies obtain their data from the internet and not reports, we used some of their approaches for categorizing and visualizing data. The existing studies discussed above apply different types of machine learning techniques to trafficking topics than what we used. These studies prove how unique our project is because we tied a specific machine learning technique, named entity recognition, to a larger wildlife trafficking dataset, EAGLE reports that span multiple countries and species of animals.

3.2 Wildlife Trafficking Domain

We analyzed other studies that examined general aspects of wildlife trafficking. A 2010 study looked at seizure reports on wildlife trafficking to analyze the scope and scale of the trade (Rosen & Smith, 2010). The authors sorted 12 years of seizure records compiled by TRAFFIC into a categorical database. The study used similar categories to ones we used, including date of seizure, country of seizure, country of origin, country of destination, method of transport, accompanying documentation, label on shipment, species name, form (live, product), number of individuals or items, weight, CITES appendix, intended use, condition of live specimens, date of offense or arrest, and date of trial or sentencing. Although the authors organized the data in a database, the study did not describe using any machine learning techniques to organize the data.

Stringham et al., (2021) developed a guide that helped scientists, nongovernmental organizations, government agencies, and other parties use the internet to monitor wildlife trade. Their guide contained 6 steps: define the scope and purpose of the project, find candidate websites, select target websites to monitor, collect and store data from websites, clean data, and analyze data. Although we use the EAGLE reports to find wildlife trafficking data, this study illustrates alternative ways to find data using the internet.

The UNODC assembled the World Wildlife Seizure (WISE) database, and in 2016 released their first World Wildlife Crime Report to gain a better perspective of global wildlife trafficking. The database sources data from the annual, biennial, and special reports from CITES members, the United States Fish and Wildlife Service Law Enforcement Management Information System, and the European Commission Enforcement Working Group (UNODC, 2016). The majority of the data is from 2005 through 2014 and consists of 132,144 seizures from 120 countries (UNODC, 2016). In 2020, the UNODC continued their assessment and examined trends in global wildlife trafficking. Our project shared a similar goal of aggregating wildlife trafficking data in a database to be better utilized by experts. Collecting data from sources that are released frequently would allow for a database to be updated more often with recent events.

3.3 Project Contributions Based on Previous Studies

The main contributions of our work to address the problem of identifying and visualizing illegal wildlife trafficking included using machine learning techniques to parse data and effectively display it for experts to analyze. Our approach was different from previous studies that used web scraping and natural language processing because we worked with named entity recognition to connect trafficking events listed in EAGLE reports. Within the EAGLE reports, we identified many different species of animals trafficked in the countries of West Africa. Previous studies only looked at a limited number of species being trafficked and that restricts the scope of the problem. Through the EAGLE reports we developed a better sense of how widespread trafficking is occurring across the Africa region. We also used data visualization to show the data from the EAGLE reports so investigators can easily identify where these wildlife trafficking incidents are being reported and what is being trafficked. The approach used by Hansen et al., (2012) is similar to the model designed, which is to use a map and document the locations of the different trafficking events that occur. Hansen et al (2012) used manual data entry to create the visualization. In contrast our approach used machine learning techniques and a database to collect and display our data. We anticipate that by developing a different approach to examine wildlife trafficking, the ability of the anti-trafficking community to address this widespread crime will be strengthened.

4 Methodology

4.1 Overview

This project was divided into three phases as shown in Figure 5. The first phase was to create a custom natural language processing (NLP) pipeline using spaCy (*SpaCy*, 2021) to process EAGLE reports and apply named entity recognition (NER) and relation extraction to identify information valuable to wildlife trafficking analysts. The second phase was creating a database to organize and store the extracted information. The third phase was to effectively visualize the contents of the database (Tableau, 2021).

Figure 5 Project Flow

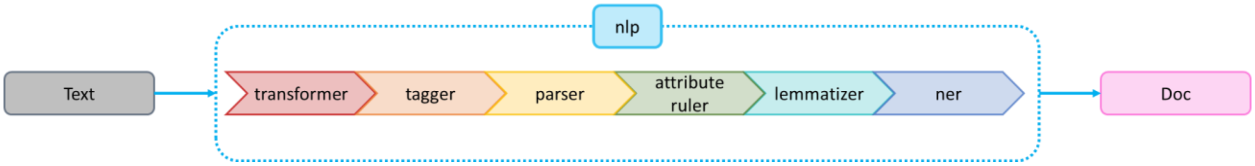
4.2 Data

Monthly EAGLE reports beginning January 2017 through December 2021 were downloaded as pdf files. This range of reports was used as data because they were all written following the same format consisting of a front cover page followed by a one-page summary. This format enabled us to simplify the process of creating our pipeline and extracting data. Reports before January 2017 were not used because they did not follow the same format. Every EAGLE report contains both text and images, but this project focused on the text and ignored the images. More specifically, we focused on just the summary page of each article because it already condensed most of the relevant information in the report and reduced the complexity of the problem.

4.3 NLP Pipeline

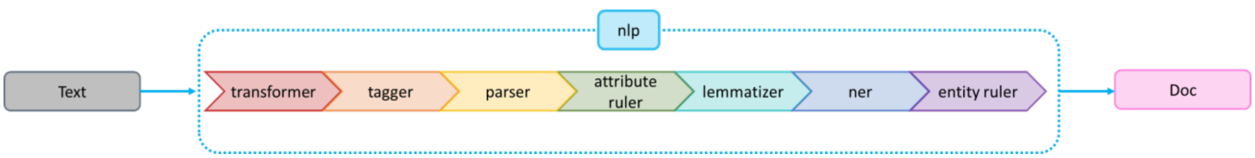
The first step towards creating our NLP pipeline was to look for existing frameworks and models to adapt. One study which compared NER approaches using Python's spaCy, Apache OpenNLP, and TensorFlow, found spaCy to have the best overall performance (Shelar, 2020). SpaCy also had many pre-trained tools and could be easily customized to fit our project. We used the pre-trained English transformer pipeline (roberta-base) "en_core_web_trf" which consisted of a transformer, tagger, parser, attribute ruler, lemmatizer, and NER components. The entire base pipeline can be seen in Figure 6.

Figure 6 SpaCy Base English Transformer Pipeline



The output of the pipeline is the tokenized text, called a doc, from which we can access useful information including named entity types, part of speech tags, dependencies, and much more. In its base form, this pipeline's NER component identifies different entities including dates, geo-political entities, quantities, organizations, and names, but it would not identify domain relevant entities, such as animals. To recognize custom entities, we added spaCy's entity ruler to the end of our pipeline as shown in Figure 7.

Figure 7 Customized Pipeline



The entity ruler component allowed us to use a rules-based approach to create custom named entities to identify in the text: ANIMAL and PRODUCT. Because these entities are finite sets, for each entity type we created a list of those entities. Starting from a collection of reference lists of animals by common English name from multiple websites, we combined the lists, removed duplicate mentions of the same animal, and reduced the list down to 286 animals ("List of Animal Names," 2022; "All Animals A-Z List," n.d.; *Animals*, 2021). This list can be referenced in Appendix A. The decision to exclude an animal from our list involved whether the name was too specific. For example, we include the term "sea turtle" but do not include more specific types of sea turtle such as "leatherback". Most of the excluded animals were species of insects, fish, and birds, because they most often had a common umbrella term. Some exceptions to this rule were made for animals known to be commonly trafficked and or have rarer and more targeted subspecies. For example, the African versus the Asian elephant. A handful of rare animals which were missing from the list were also added by hand because they are known to be trafficking targets. We created a list of nine common animal products as well. This list can be referenced in Appendix B. After creating both lists, we added custom extensions, a spaCy label we used to associate an item in our list with a matching token in the doc text.

Entities extracted from the EAGLE reports could be in singular or plural form (for example tusk and tusks). When adding instances to the database we wanted labels to be consistent. We utilized the inflect Python package to reformat plural instances into singular form (Dyson & Coombs, 2022). This library generates singular noun inflections based on the Oxford English Dictionary and Fowler's English Language Usage (Dyson & Coombs, 2022).

4.4 Database

To design and build the database we first needed to know what information to store. So to learn what is of interest to domain specialists, we informally interviewed Patricia Raxter, PhD, who is a senior analyst that studies illegal wildlife trafficking. She reads through EAGLE reports and many other organization's articles on a daily basis as part of her job. We observed her read through different reports and asked her which information she and her colleagues consider to be the most valuable. She highlighted the animal species and products being trafficked (teeth, tusks, fur, etc.), the trafficking routes, and the methods of transportation as some of the most important things she looks for. We used her input to help us devise an Entity Relationship Diagram (ERD) for the database schema, which is shown in Figure 12 in the results section. To build the database in our codebase we used the python library SQLite3 (SQLite, 2022). Once the database was built, we wrote a function to automatically add the output of the pipeline to the database. We manually reviewed the database as the final step to reduce some error. This mainly included deleting empty rows or correcting other obvious errors such as ivory being associated with pangolins.

4.5 Visualization / Website

Once the database was populated, we wanted to display it for those working with the data. Data can be presented in different forms, such as line graphs, bar graphs, and maps. To decide which style would be best for investigators analyzing the data we explored various visualization software tools including HighCharts (*HighCharts*, 2022), Matplotlib (*Matplotlib*, 2021), and Tableau (Tableau, 2021). We decided to use Tableau due to its established reputation, comprehensive features, and visually appealing graphics. First, we researched the different styles of visualization in further detail to see what was feasible on Tableau. Then using Tableau Desktop, we formatted our visualizations into a story and published our work to Tableau Public. Uploading our visualizations to Tableau Public allowed us to embed the graphics into a website hosted on GitHub. Which graphics we used had to find a balance between something that experts who analyze wildlife trafficking could use as a professional tool, and something that would be visually engaging and comprehensible for non-experts.

5 Results / Discussion

5.1 Named Entity Recognition

We added two custom named entities, ANIMAL and PRODUCT, to the Named Entity Recognition (NER) component of spaCy's pipeline, which already had the capability to recognize multiple types of named entities. A full list of which can be seen in Appendix C with descriptions. In Figure 8 you can see a visualization of spaCy's NER model without our custom named entities on a section of the EAGLE monthly report for April 2021. There is no recognition of types of animals or animal products. In Figure 9 however, is the same section of text visualized including our custom named entities. Animal products are highlighted in green while animal species are highlighted in gray. The addition of our custom named entities allows us to recognize any animal species on our compiled list of animals, as well as any animal product on our list of products.

Figure 8 Base SpaCy Named Entity Recognition

Figure 9 Custom SpaCy Named Entity Recognition

5.1.1 Accuracy Analysis

To analyze the accuracy of the pipeline we compared its results to what a human reader would extract from the same report. Fifteen reports in total were used for testing, three randomly selected from each year between 2017 through 2021. One person individually read five reports, one from each year, and manually recorded the information in the same format the pipeline would add an entry to the database. Then the report was run through the pipeline, and the entries created by the pipeline were compared to the reader's entries. The fields we compared for each entry were arrest count, country, product name, animal species, quantity, and weight. We used four categories to represent each possible outcome of an entry added by the pipeline to the database: fully correct, partially correct, undetected, and unrelated.

A fully correct entry is defined as every part the reader identified was correctly entered by the pipeline to the database. For example, in the June 2021 article, we manually identified one trafficker who was arrested with 2 leopard skins in Senegal. The pipeline correctly added 1 into the arrest field, Senegal into the country field, skin into the product field, leopard into the animal field, and 2 into the quantity field. This perfect entry proves the pipeline can make the connection between related values and add it into the database. A partially correct entry is defined as at least one of the six fields entered correctly into the database by the pipeline. A partially correct entry means one or more fields have incorrect data in the entry, whether that is the wrong country associated with a product, a wrong quantity value entered, or even a different number of traffickers arrested. One example of a partially correct entry is in the March 2019 article. We manually identified the event as 2 traffickers were arrested for having 3 elephant tusks in Gabon. Although the pipeline correctly identified Gabon as the country, tusk as the product, elephant as the animal, and 3 as the quantity, it incorrectly entered the arrest field as 1.

The pipeline sometimes does not identify the correct correlations between entities, and this causes the wrong data to be added to an entry. The undetected classification means an entry we extracted manually was not detected by the pipeline and therefore not in the database. For example, this sentence, “One of the biggest ape traffickers in the African continent was arrested in Guinea,” was in the February 2017 briefing and counted as an event manually but not acknowledged by the pipeline. We believe this is due to the way this sentence is formed. Most of the summary is structured to have the word trafficker as one of the first four words in a sentence when a new event is being described. Therefore, we coded the pipeline to acknowledge a new event when this happens. When a new event is being described and does not follow the pattern, we told the pipeline to identify, the event is not recorded. Finally, the unrelated category is when the pipeline extracts information from the article that was not identified as a new event by a human reader. An example can be found when the pipeline added a row to the database from the sentence in the April 2018 article, “He was trafficking wildlife products to Mali and other countries.” That sentence was in the middle of a paragraph and added information to an already documented event. The country Mali describes what would have been the destination of the trafficked products. However useful data, this country is not describing a new event, so it was deemed incorrectly added by the pipeline.

The distribution of the results for each report is graphed in Figure 10, and the ratio of the pipeline outcomes total results for the pipeline are shown in Figure 11. Of its entries to the database, 11.7% were fully correct, and 28.1% were partially correct. This implies a correctness of almost 40%. An entry identified by the reader but undetected in the database happened only 29.7% of the time. 30.5% of the pipeline’s outcomes were classified as unrelated. This reflects one of the most challenging issues with coding the pipeline, which is properly identifying correlation between entities in events. This will be talked about in more detail in the limitations section.

Figure 10 Pipeline Outcomes Over Report

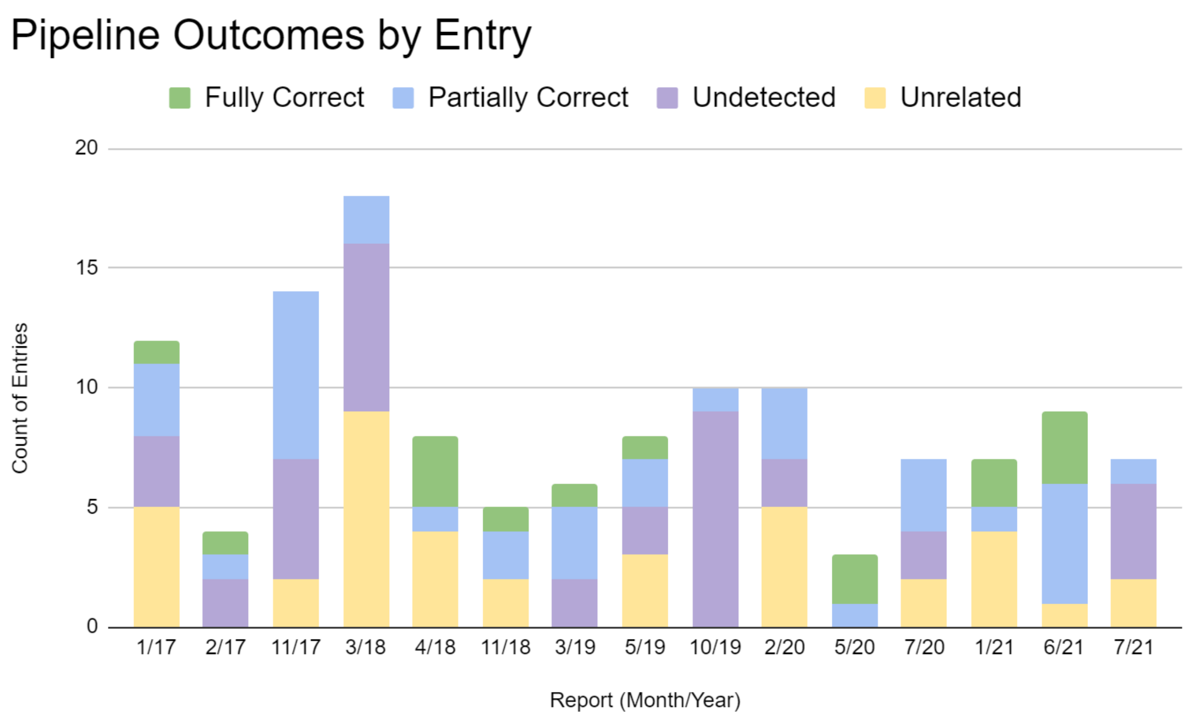
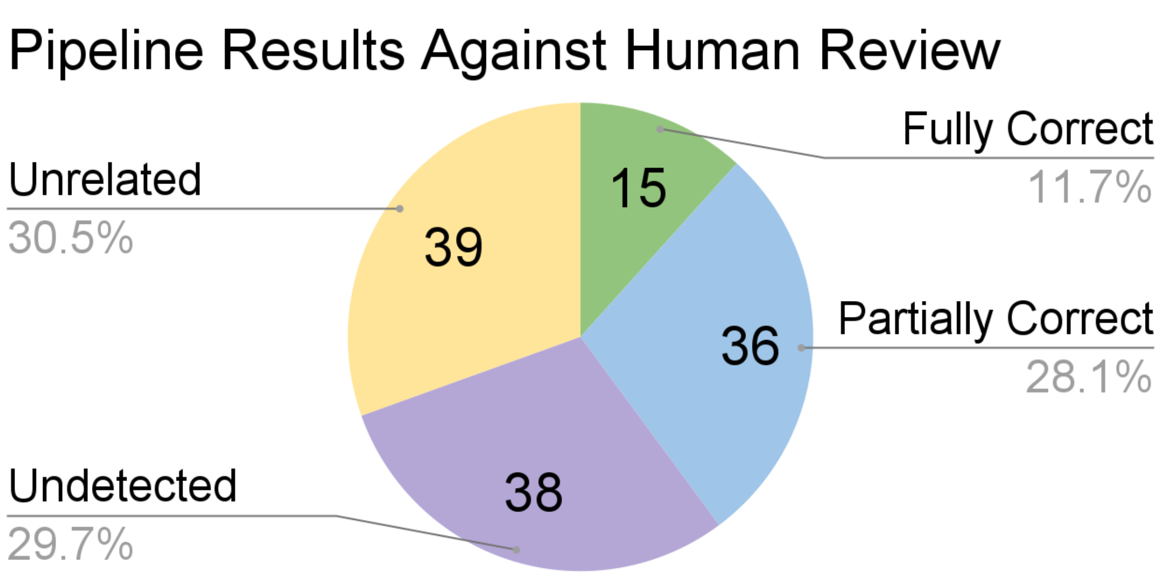


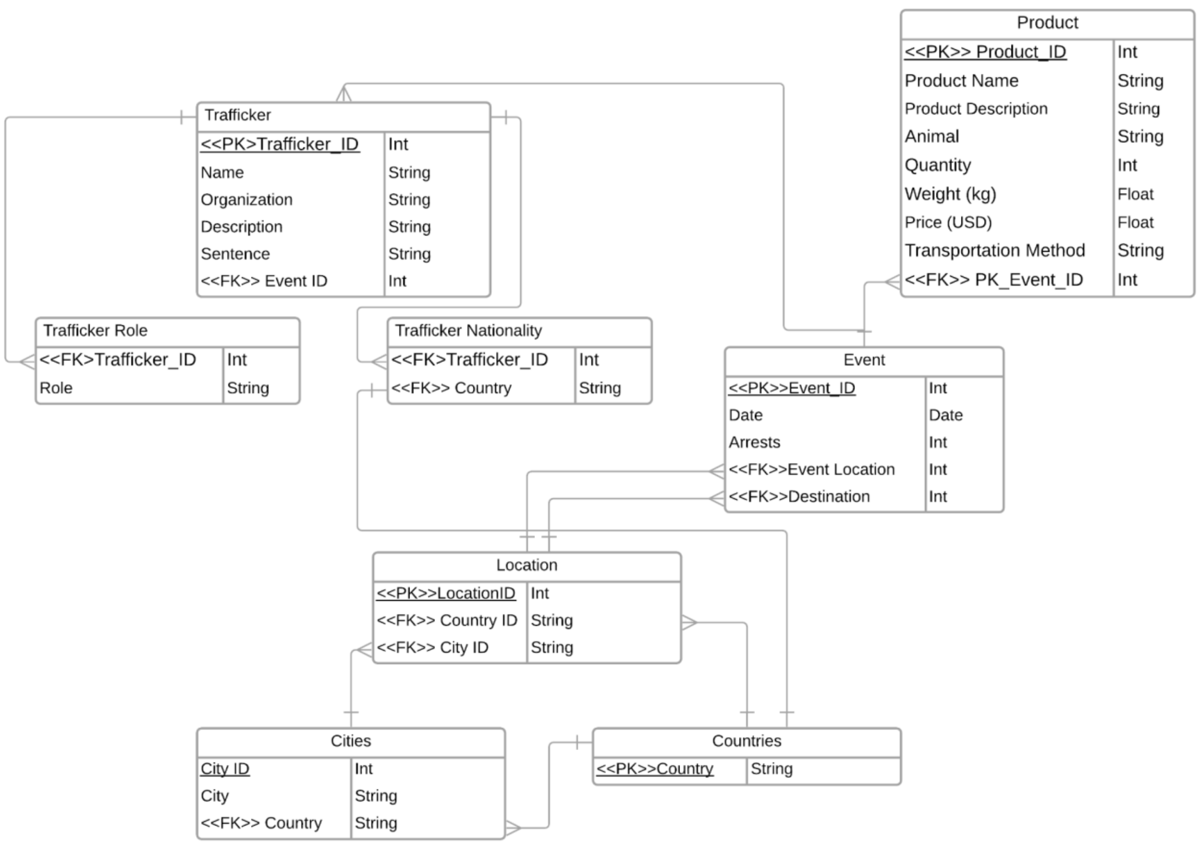
Figure 11 Pipeline Correctness Breakdown



5.2 Database

The database used to store the pipeline results is stored to a local comma separated file (.csv), which can be downloaded and opened in excel. This is a desirable format for many domain experts because it is easy to use and view. Along with the raw database containing all the pipeline results, another version was created which has been manually cleaned for use in our visualizations. Both versions can be downloaded at the website from the About tab.

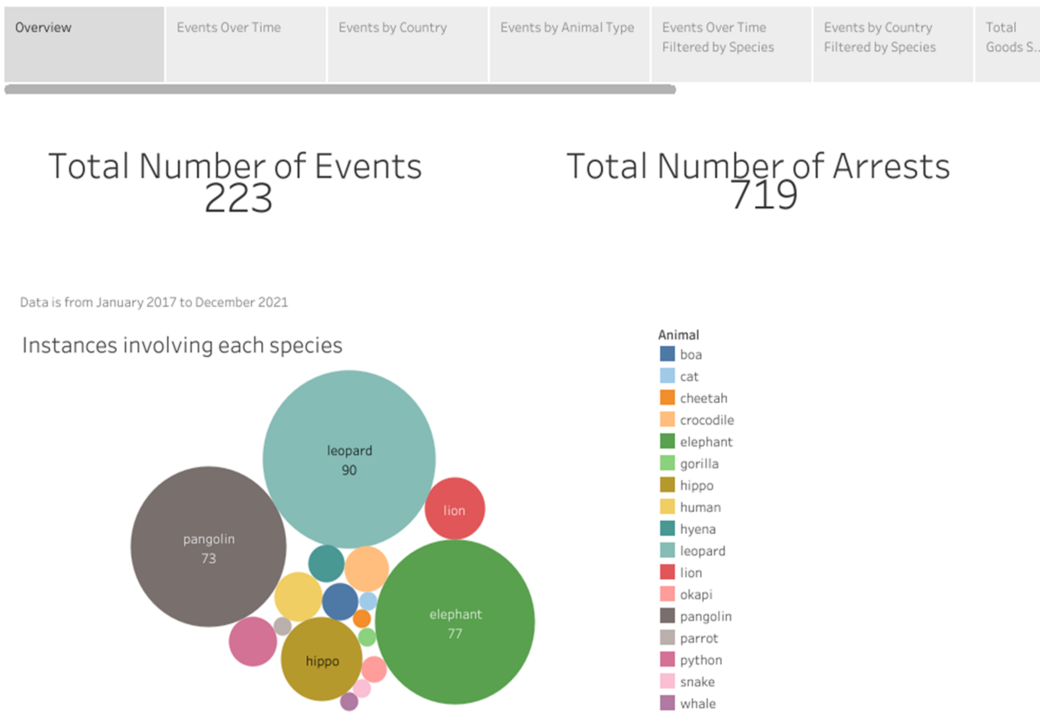
Figure 12 Database Entity Relationship Diagram



5.3 Visualizations / Website

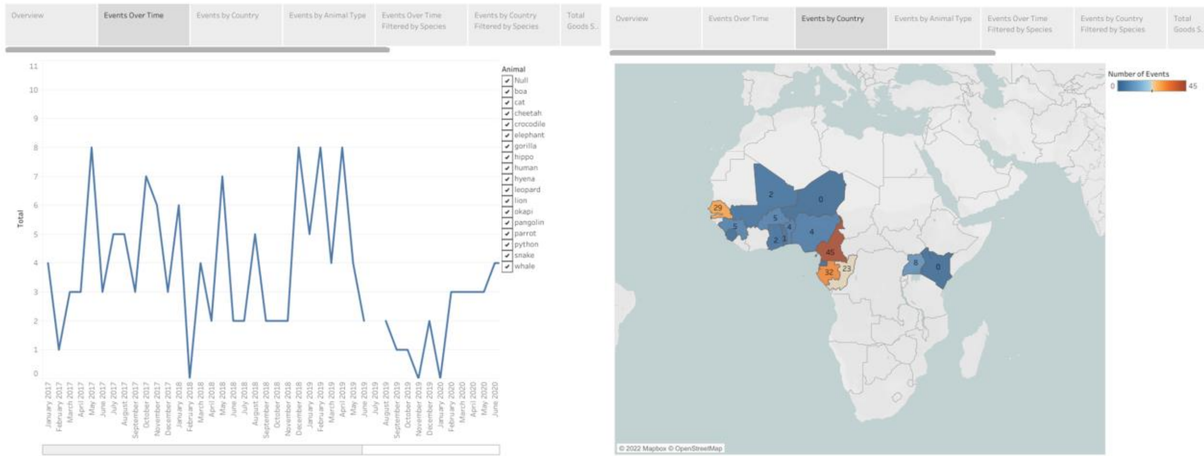
The .csv database file was uploaded into Tableau where we created 10 visualizations displaying the data. The story format, which is a sequence of visualizations, is used to incorporate all of the visualizations on the website. The first story slide, as seen in Figure 13, is a general overview of the data collected. It displays the total number of events and total number of arrests along with a bubble map highlighting the most mentioned species. When this project was completed, the data collected for the visualizations was from January 2017 to December 2021 (July 2019 and August 2020 were unavailable) and displayed a total of 223 events and 719 total arrests. The bubble chart below these totals showed leopards, elephants, and pangolins were the most trafficked species. This opening story gives viewers a general idea of the data collected and prefaces the other stories.

Figure 13 Overview



The next five stories show the distribution of data grouped by events. The graph on the left of Figure 14, shows the number of events over time. It displays a line graph that sums the number of events for each monthly report. This visualization shows which months had higher trafficking events. The next graph in Figure 14, shows the event count by country. It is displayed on a choropleth map with a number label for the event total. One is able to hover over each country to display the country name and count. This graph shows what countries are hotspots for wildlife trafficking. However, there are some caveats to using this type of map, which will be discussed below.

Figure 14 Events Over Time and Events by Country

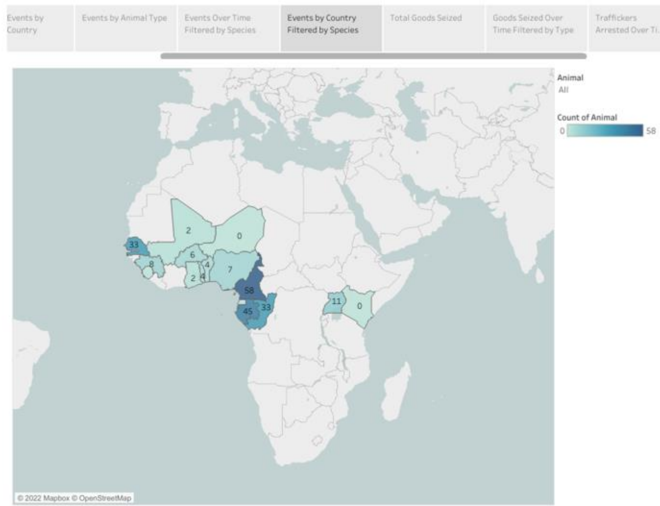


The next couple of stories, as seen in Figure 15, use bar graphs to display the events grouped by animal species. The first shows the animal species on the y-axis and time on the x-axis and the bars represent the count of events. This could be beneficial in analyzing which animals are trafficked more often than others. It can be challenging when trying to display information about various animals and animal products because the standard amounts can vary greatly. For example, it is fairly common for there to be several pounds of pangolin scales but only a few elephant tusks collected in one seizure. To address this, we included filters on our graphs so a user can select a particular type to examine. The other visualization about events displays years on the y-axis and months on the x-axis and groups the data by animal species totals. It is different from the previous bar graph because it is meant to filter the data to only show one animal species at a time. Grouping by one species allows viewers to analyze the data more clearly as the visualization is not overloaded with lots of data.

Figure 15 Events by Animal Type and Events Over Time Filtered by Species

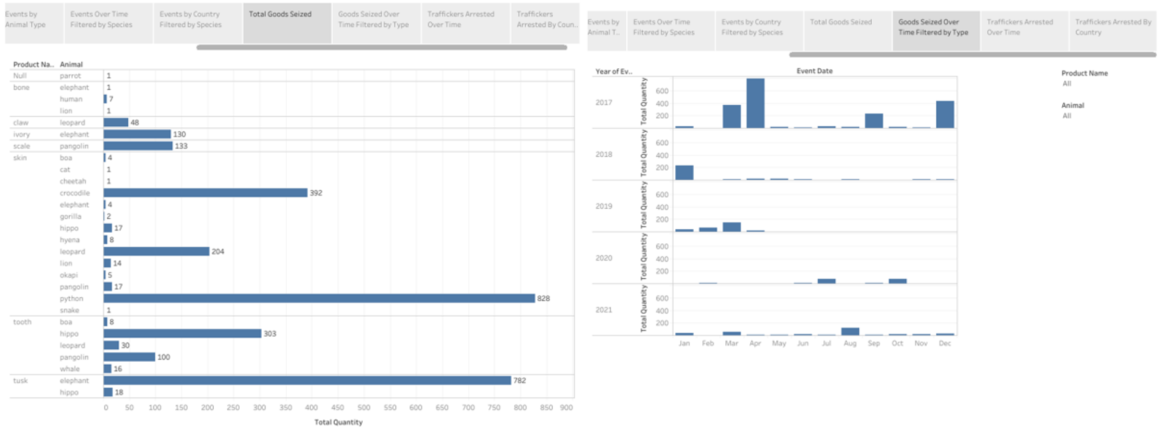
The final visualization using event data, Figure 16, is another choropleth map that groups the count of events by country for a specific animal species.

Figure 16 Events by Country Filter by Species



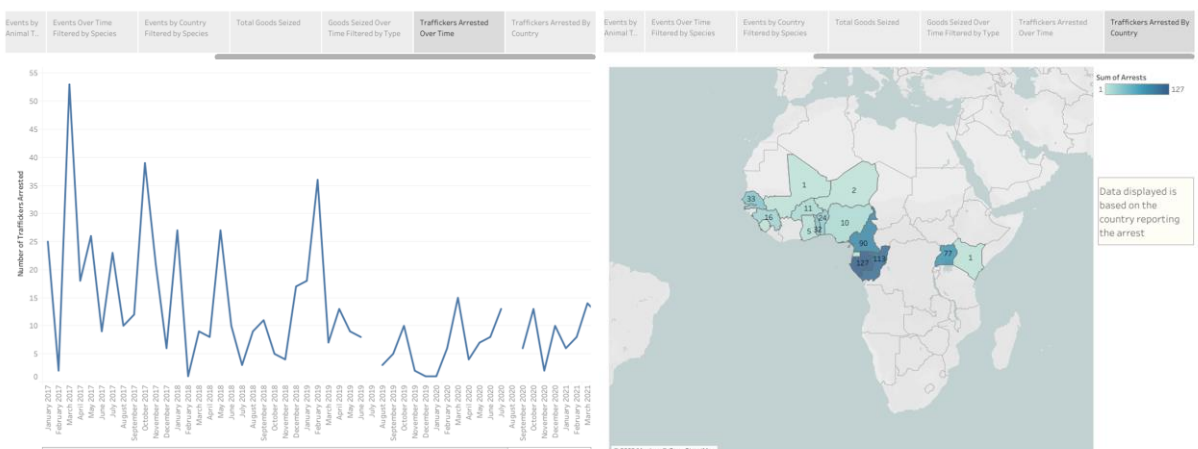
The next two stories, see Figure 17, on the website show product specific data. The visualizations are both bar graphs. The first of two visualizations show total goods seized. It is product data based on the quantity field in the table, which means it is showing the value of how many products have been collected. We grouped by product type first and animal species second so that analysts can find trends for each product. Seeing the distribution of animal types for a given product could be useful, such as comparing tusks sourced from elephants and rhinos. The way the data is presented provides an overview for all the different product and animal types. The other visualization filters the product data, so it shows the totals of product quantities over time for a specific product or animal species. This allows closer analysis on one product/animal species.

Figure 17 Total Goods Seized and Goods Seized Over Time Filter by Type



The final two stories, Figure 18, on the website display visualizations about arrest data. The first is a line graph that shows the total values of traffickers arrested over time while the other visualization shows the arrest data by country. The line graph is useful in seeing what months had higher arrests. The choropleth map shows the number of traffickers arrested in the different countries and displays which countries have reported more arrests than others.

Figure 18 Traffickers Arrested Over Time and Traffickers Arrested by Country

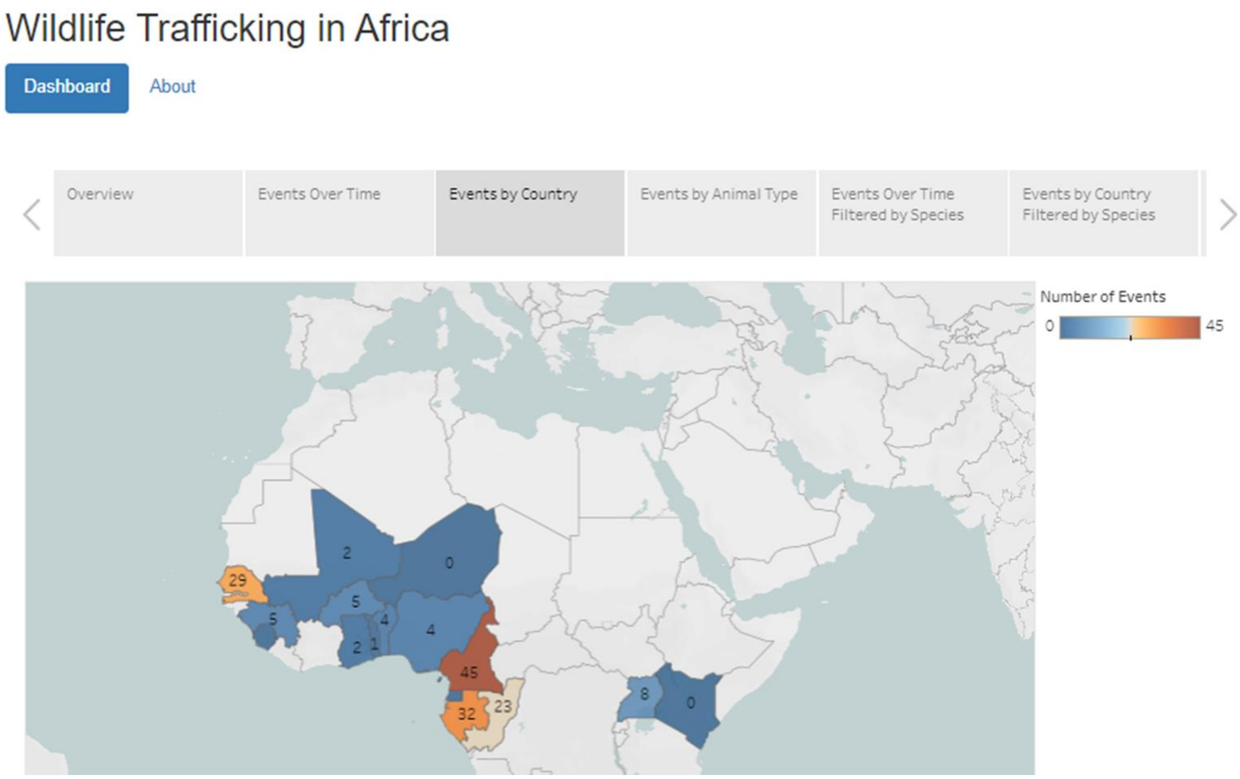


We implemented different choropleth maps to display information regarding the location specific data. Choropleth maps color predefined areas in proportion to the statistical variable being displayed. They are used to easily visualize the variation across regions, however there are several drawbacks to these types of maps. The data is displayed only at the country level since that was the information, we were able to extract from the documents. Countries represent a large geographic region, so the details within a given country are lost. For instance, the trafficking may be predominantly occurring along the coast or in a particular city and this would not be identifiable in the map. Additionally, dividing the map by countries creates the illusion of boundaries at the country borders when this may not reflect reality. Countries with larger geographic areas may visually dominate the map potentially distorting the importance of the data. Likewise, countries that occupy a smaller area may be more difficult to view on the map. This issue could be improved by trying to collect data on a more local level which could be useful for experts.

From our discussions with Meredith Gore, PhD and Patricia Raxter, PhD, it is useful to have the ability to look at the data in a very filtered and specific format along with the overview of all the data. They explained that people in law enforcement look for detailed and filtered information (like data for a specific species or in a certain region) so they can analyze where they should be putting their attention. Whereas policy makers like to see the overview of all data to find stand out places to put money and time. Patricia Raxter, PhD also stated that when looking at wildlife trafficking data, she compares data to find outlying trends. To satisfy all the different ways experts look at the data, we created many forms of visualizations.

Once the visualizations were completed, they were uploaded to Tableau Public and inserted on the Dashboard page of the website, see Figure 19. The website can be viewed at https://wildlifemqp.github.io/Visualizations/. The website also has a Dashboard Definitions section to explain commonly used terms in the dashboard. Below that is a disclaimer to provide viewers context about our data and usage of maps. The About page on the website gives background to wildlife trafficking and our project. The project goal is described as well as where our data was extracted from. There is also a section on the database layout and links to download the most recent CSV files. The final section on the about page links a few resources about wildlife trafficking for users to learn more.

Figure 19 Dashboard of Website



6 Limitations

One of the biggest challenges of this project was identifying when two or more references in the text refer to the same thing. This is known as the coreference problem. An example of the coreference problem, as seen in Figure 20, is distinguishing a person's name in one sentence as the same entity as the pronoun used in the next sentence.

Figure 20 Example of Coreference Problem

Gompei is a goat. He has four legs and two horns.

Finding ways to handle coreference was integral to our project because identifying when an entity has already been referenced is necessary to prevent mistaking it for a new event. Unfortunately, the coreference problem currently does not have a complete solution, so we had to develop our own solutions. We also had trouble pulling weight and quantity fields from EAGLE reports because the weight and quantity was not always near the associated animal in the text. This meant that there were no features, such as a connecting edge on the dependency tree, relating the weight or quantity to the animal. Identifying the relationship between two or more entities in text is called relation extraction. Relation extraction is also a very hard task. These limitations were the main cause of error in our pipeline.

To visualize what the pipeline extracted, we used Tableau which had its own limitations. Tableau limited us to displaying the data in the formats they provided. We were also only able to create these visualizations with cleaned data because data that is not in the correct format or entered incorrectly would be excluded from the visualizations.

We were also limited by what data we were able to collect. EAGLE reports include a summary preceding the more comprehensive report. Since we only processed the summary section of each report, not all the information available was run through the pipeline. Therefore, many of the details are missed.

Limitations outside of our control include being restricted by what was reported in the EAGLE reports. For instance, names of traffickers were typically omitted. Additionally, not all events include the same details. For instance, the weight of a product may not be specified. Also, our data underestimates the true magnitude of wildlife trafficking occurring, since reported seizures represent only a portion of the actual wildlife trafficking occurring. Our data is also limited to countries actively involved in combating wildlife trafficking with the EAGLE organization and skews towards countries more involved in combating wildlife trafficking, as they are the ones more likely to execute and report seizures. These limitations were out of our control, so we did our best to create a project that could extract what was given and still accomplish our goal.

7 Future Work

The pipeline could be enhanced by extracting additional information from the text. Details such as information about the means of transportation and the backgrounds of the traffickers could be useful to examine. Recording more details about the product could also be useful, such as the price and any obfuscation upon seizure. Subsequently this data could be used to generate additional visualizations. Identifying more specifics about the location like city, airport, port, etc. would also be interesting.

Additionally, the pipeline accuracy needs improvement. As previously mentioned, the two biggest challenges were the co-reference and relation extraction problems. One potential way to develop the pipeline is to explore Prodigy which allows users to manually annotate examples within the dataset to train an AI model (*Prodigy*, 2022). It can be used to select named entity recognition training data and natural language processing relations. This approach would address relation extraction by training a model to recognize related entities directly from the text, rather than creating rules for a rules-based approach. This could also be used to link words together such as an animal name and a product type so the model can classify an animal-product entity. Ideally the framework would reach an acceptable accuracy threshold and eliminate the need for manual intervention.

This project could be expanded to use additional data sources. Drawing upon other articles and reports would increase the size of the dataset and offer a better representation of the true scale of wildlife trafficking. Additionally, pulling data from news articles would allow the data to be kept up to date more regularly than monthly reports.

8 Conclusion

Wildlife trafficking domain experts lack a single centralized network for finding and sharing recent data and news. This makes it extremely difficult and time consuming for many researchers, analysts, and law enforcement, to stay informed about current wildlife trafficking across the globe. Furthermore, the absence of one exhaustive wildlife trafficking database interferes with experts' ability to make connections between events and gain big-picture understanding. We believe our natural language processing pipeline with its domain specific named entity recognition, expandable database, and interactive website with various illustrative visualizations, serve as a framework to solve these issues. With continued improvements, our work could become the comprehensive tool for wildlife trafficking data.

References

All Animals A-Z List. (n.d.). *AZ Animals*. Retrieved February 23, 2022, from https://a-z-animals.com/animals/

Anagnostou, M., & Doberstein, B. (2021). Illegal wildlife trade and other organised crime: A scoping review. *Ambio*. https://doi.org/10.1007/s13280-021-01675-y

Animals: 1000+ Animal Names. (2021, April 10). https://visualdictionary.org/animals/

Balinga, M., & Stroud, A. (2020). *Ministry of Environment and Sustainable Development. Understanding threats to West African biodiversity and linkages to wildlife trafficking: Benin field assessment report*. https://pdf.usaid.gov/pdf_docs/PA00XGP5.pdf

Dyson, P., & Coombs, J. (2022). *Inflect*. https://pypi.org/project/inflect/

EAGLE. (2022). *EAGLE Network - Eco Activists for Governance and Law Enforcement*. https://www.eagle-enforcement.org/

ENACT, & INTERPOL. (2018). *INTERPOL Overview of serious and organized crime in West Africa 2018*. ENACT Africa. https://enactafrica.org/research/analytical-reports/interpol-overview-of-serious-and-organised-crime-in-west-africa-2018

Environmental Investigation Agency. (2020). *Out of Africa*. https://eia-international.org/wp-content/uploads/Out-of-Africa-FINAL.pdf

García-Díaz, P., Ross, J. V., Woolnough, A. P., & Cassey, P. (2017). The illegal wildlife trade is a likely source of alien species. *Conservation Letters*, *10*(6), 690–698. https://doi.org/10.1111/conl.12301

Hansen, A. L. S., Li, A., Joly, D., Mekaru, S., & Brownstein, J. S. (2012). Digital surveillance: A novel approach to monitoring the illegal wildlife trade. *PLOS ONE*, *7*(12), e51156. https://doi.org/10.1371/journal.pone.0051156

HighCharts. (2022). https://www.highcharts.com/

Ingram, D. J., Coad, L., Abernethy, K. A., Maisels, F., Stokes, E. J., Bobo, K. S., Breuer, T., Gandiwa, E., Ghiurghi, A., Greengrass, E., Holmern, T., Kamgaing, T. O. W., Obiang, A.-M. N., Poulsen, J. R., Schleicher, J., Nielsen, M. R., Solly, H., Vath, C. L., Waltert, M., … Scharlemann, J. P. W. (2018). Assessing Africa-Wide Pangolin exploitation by scaling local data. *Conservation Letters*, *11*(2), e12389. https://doi.org/10.1111/conl.12389

List of animal names. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=List_of_animal_names&oldid=1072909128

Mackey, T. K., Li, J., Purushothaman, V., Nali, M., Shah, N., Bardier, C., Cai, M., & Liang, B. (2020). Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram. *JMIR Public Health and Surveillance*, *6*(3), e20794. https://doi.org/10.2196/20794

Mallon, D. P., Hoffmann, M., & McGowan, P. J. K. (2015). *An IUCN situation analysis of terrestrial and freshwater fauna in West and Central Africa*. International Union for Conservation of Nature. https://doi.org/10.2305/IUCN.CH.2015.SSC-OP.54.en

Matplotlib (3.5.0). (2021). [Computer software]. https://matplotlib.org/

Medicine, I. of. (2015). *Emerging viral diseases: The One Health Connection: Workshop summary* (E. R. Choffnes & A. Mack, Eds.). The National Academies Press. https://www.nap.edu/catalog/18975/emerging-viral-diseases-the-one-health-connection-workshop-summary

Miller, C., Pay, J., & Smtih, J. (2019). *Detecting online environmental crime markets*. Global Initiative. https://globalinitiative.net/wp-content/uploads/2019/01/TGIATOC-DetectingOnlineMarkets-Web-1.pdf

Nalluri, S., Kumar, S. J. R., Soni, M., Moin, S., & Nikhil, K. (2021). A survey on identification of illegal wildlife trade. In C. Kiran Mai, B. V. Kiranmayee, M. N. Favorskaya, S. Chandra Satapathy, & K. S. Raju (Eds.), *Proceedings of International Conference on Advances in Computer Engineering and Communication Systems* (pp. 127–135). Springer. https://doi.org/10.1007/978-981-15-9293-5_11

Patel, N. G., Rorres, C., Joly, D. O., Brownstein, J. S., Boston, R., Levy, M. Z., & Smith, G. (2015). Quantitative methods of identifying the key nodes in the illegal wildlife trade network. *Proceedings of the National Academy of Sciences*, *112*(26), 7948–7953. https://doi.org/10.1073/pnas.1500862112

Phelps, J., Biggs, D., & Webb, E. L. (2016). Tools and terms for understanding illegal wildlife trade. *Frontiers in Ecology and the Environment*, *14*(9), 479–489. https://doi.org/10.1002/fee.1325

Prodigy. (2022). Prodigy. https://prodi.gy

Rosen, G. E., & Smith, K. F. (2010). Summarizing the evidence on the international trade in illegal wildlife. *EcoHealth*, *7*(1), 24–32. https://doi.org/10.1007/s10393-010-0317-y

Schloenhardt, A. (2020). Wildlife trafficking: causes, characteristics, and consequences. In *Wildlife trafficking: the illicit trade in wildlife, animal parts, and derivatives*. Carl Grossmann. 9783941159440

Segniagbeto, G. H., Agbodji, K. T., Leuteritz, T. E. J., Dendi, D., Fa, J. E., & Luiselli, L. (2020). Insights into the illegal ivory trade and status of elephants in Togo, West Africa. *African Journal of Ecology*, *n/a*(n/a). https://doi.org/10.1111/aje.12748

Segniagbeto, G. H., Assou, D., Agbessi, E. K. G., Atsri, H. K., D’Cruze, N., Auliya, M., Fa, J. E., & Luiselli, L. (2021). Insights into the status and distribution of pangolins in Togo (West Africa). *African Journal of Ecology*, *59*(2), 342–349. https://doi.org/10.1111/aje.12809

Shelar, H., Kaur, G., Heda, N., & Agrawal, P. (2020). Named Entity Recognition Approaches and Their Comparison for Custom NER Model. *Science & Technology Libraries*, *39*(3), 324–337. https://doi.org/10.1080/0194262X.2020.1759479

Smith, K. M., Anthony, S. J., Switzer, W. M., Epstein, J. H., Seimon, T., Jia, H., Sanchez, M. D., Huynh, T. T., Galland, G. G., Shapiro, S. E., Sleeman, J. M., McAloose, D., Stuchin, M., Amato, G., Kolokotronis, S.-O., Lipkin, W. I., Karesh, W. B., Daszak, P., & Marano, N. (2012). Zoonotic viruses associated with illegally imported wildlife products. *PLoS ONE*, *7*(1), e29505. https://doi.org/10.1371/journal.pone.0029505

spaCy (3.2). (2021). [Computer software]. https://spacy.io/

SQLite. (2022). *SQLite3* (3.37.2) [Computer software]. https://www.sqlite.org/index.html

Stringham, O. C., Toomes, A., Kanishka, A. M., Mitchell, L., Heinrich, S., Ross, J. V., & Cassey, P. (2021). A guide to using the internet to monitor and quantify the wildlife trade. *Conservation Biology*, *35*(4), 1130–1139. https://doi.org/10.1111/cobi.13675

Tableau. (2021). *Tableau Desktop* (Professional Edition) [Computer software]. https://www.tableau.com/

Tella, J. L., & Hiraldo, F. (2014). Illegal and legal parrot trade shows a long-term, cross-cultural preference for the most attractive species increasing their risk of extinction. *PLoS ONE*, *9*(9), e107546. https://doi.org/10.1371/journal.pone.0107546

Thomas-Walters, L., Hinsley, A., Bergin, D., Burgess, G., Doughty, H., Eppel, S., MacFarlane, D., Meijer, W., Lee, T. M., Phelps, J., Smith, R. J., Wan, A. K. Y., & Veríssimo, D. (2021). Motivations for the use and consumption of wildlife products. *Conservation Biology*, *35*(2), 483–491. https://doi.org/10.1111/cobi.13578

UNODC. (2016). *World wildlife crime report: Trafficking in protected species*. https://www.unodc.org/documents/data-and-analysis/wildlife/World_Wildlife_Crime_Report_2016_final.pdf

UNODC. (2020). Introduction. In *World wildlife crime report*. https://www.unodc.org/documents/data-and-analysis/wildlife/2020/WWLC20_Chapter_1_Introduction.pdf

USAID WA BiCC. (2021). *Overview of the current status of pangolins and pangolin conservation in West Africa*.

van Uhm, D. P., & Nijman, R. C. C. (2020). The convergence of environmental crime with other serious crimes: Subtypes within the environmental crime continuum. *European Journal of Criminology*, 147737082090458. https://doi.org/10.1177/1477370820904585

Wildlife crime | CITES. (n.d.). Retrieved October 6, 2021, from https://cites.org/eng/prog/iccwc/crime.php

Wildlife, Forest & Fisheries Crime Module 1 Key Issues: Implications of Wildlife Trafficking. (2019). //www.unodc.org

WWF. (2012, December 12). *Illegal wildlife trade threatens national security, says WWF report*. https://wwf.panda.org/wwf_news/?207054/Illegal-wildlife-trade-threatens-national-security-says-WWF-report

Xu, Q., Li, J., Cai, M., & Mackey, T. K. (2019). Use of machine learning to detect wildlife product promotion and sales on Twitter. *Frontiers in Big Data*, *2*, 28. https://doi.org/10.3389/fdata.2019.00028

Appendix A: List of Animals

aardvark
abalone
african buffalo
african elephant
african wild dog
albatross
alligator
alpaca
american black bear
anaconda
andean mountain cat
angelfish
ant
anteater
antelope
ape
arctic wolf
armadillo
asian black bear
asian elephant
bactrian camel
badger
bald eagle
bandicoot
barracuda
bass
bat
bear
beaver
bee
beetle
bengal tiger
betta
bighorn sheep
bilby
bison
black panther
blackbird
blue sheep
blue tang
blue whale
bluefin tuna
boar
bobcat
bonobo
bowhead whale
brown bear
butterfly
caiman
camel
cape buffalo
capybara
caracal
cassowary
caribou
carp
cat
caterpillar
catfish
centipede
chameleon
cheetah
chimpanzee
chinchilla
chipmunk
cicada
civet
clam
clouded leopard
clownfish
coati
cobra
cockatoo
coral
cormorant
cougar
cow
coyote
crab
crane
cricket
crocodile
crow
dalmatian
deer
devil ray
dog
dolphin
donkey
dove
dragonfly
duck
dugong
eagle
earthworm
eel
elephant
elk
emu
falcon
ferret
fennec fox
firefly
fish
flamingo
flying squirrel
fox
frog
gaur
gazelle
gecko
gibbon
giant panda
giraffe
goat
goldfish
goose
gopher
gorilla
grasshopper
gray whale
green anole
guinea pig
hamster
hare
hawk
hedgehog
hen
hippo
hippopotamus
hoatzin
hornbill
horse
human
hummingbird
hyena
iguana
indri
jackal
jackrabbit
jaguar
javan rhino
jellyfish
kangaroo
killer whale
king vulture
kingfisher
kitten
koala
komodo dragon
koi
lemur
leopard
lion
lizard
llama
lobster
lynx
macaw
magpie
mallard
mammoth
manatee
mandrill

Appendix A (cont.): List of Animals

margray
marmoset
manta ray
mantis
mantis shrimp
marlin
marten
meerkat
millipede
mink
mongoose
monkey
moon bear
moose
mosquito
moth
mouse
narwhal
newt
numbat
ocelot
octopus
okapi
opossum
orca
orangutan
ostrich
otter
owl
ox
oyster
panda
pangolin
panther
parrot
peacock
peccary
pelican
penguin
pheasant
pig
pigeon
piranha
polar bear
possum
prairie dog
puma
puppy
puffin
python
quail
quokka
rabbit
raccoon
rat
rattlesnake
raven
red dear
red fox
reindeer
rhinoceros
Rhinoceros
hornbill
robin
rooster
salamander
salmon
sand cat
scarab
sea lion
sea turtle
sea urchin
seagull
seahorse
seal
serow
serval
shark
sheep
shrimp
skunk
sloth
snail
snake
snow leopard
sparrow
spider
squid
squirrel
starfish
stingray
stork
swallow
swan
swordfish
taipan
tamarin
tapir
tasmanian devil
tiger
toad
tortoise
toucan
tree frog
tuna
turkey
turtle
vulture
wallaby
walrus
wasp
water buffalo
whale
wolf
wolverine
wombat
woodpecker
wren
xerus
yak
zebra

Appendix B: List of Animal Products

bone
claw
fur
ivory
meat
scale
skin
teeth
tusk

Appendix C: SpaCy Default Entities

PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, initutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Objects, vehicles, foods, etc. (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, sogns, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage, including "%"
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"First", "second", etc
CARDINAL	Numerals that do not fall under another type