A Flexible and Scalable Approach for Collecting Wildlife Advertisements on the Web

Juliana Barbosa
New York University

Sunandan Chakraborty
Indiana University

Juliana Freire
New York University

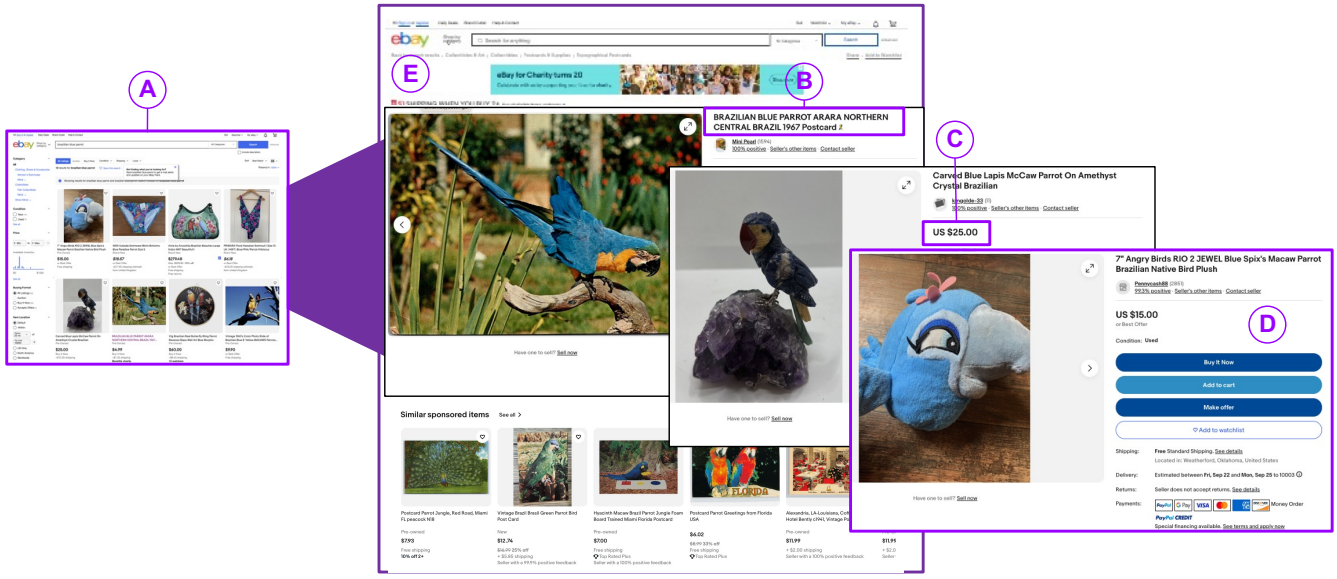


Figure 1: We describe a pipeline that automatically collects a dataset of wildlife advertisements (ads) from e-commerce websites. Issuing a query (A) "Brazilian blue parrot", we can find and follow links to several product pages (E). For each product (D), we extract product attributes such as (B) product title and (C) price. A challenge in constructing this dataset is how to distinguish ads for wildlife products from other types of products, such as postcards and toys.

ABSTRACT

Wildlife traffickers are increasingly carrying out their activities in cyberspace. As they advertise and sell wildlife products in online marketplaces, they leave digital traces of their activity. This creates a new opportunity: by analyzing these traces, we can obtain insights into how trafficking networks work as well as how they can be disrupted. However, collecting such information is difficult. Online marketplaces sell a very large number of products and identifying ads that actually involve wildlife is a complex task that is hard to automate. Furthermore, given that the volume of data is staggering, we need scalable mechanisms to acquire, filter, and store the ads, as well as to make them available for analysis. In this paper, we present a new approach to collect wildlife trafficking data at scale. We propose a data collection pipeline that combines scoped crawlers for data discovery and acquisition with foundational models and machine learning classifiers to identify relevant ads. We describe a dataset we created using this pipeline which is, to the best of our knowledge, the largest of its kind: it contains almost a million ads obtained from 41 marketplaces, covering 235 species and 20 languages.

KEYWORDS

Wildlife, data-mining, web-crawl, dataset

ACM Reference Format:
Juliana Barbosa, Sunandan Chakraborty, and Juliana Freire. 2018. A Flexible and Scalable Approach for Collecting Wildlife Advertisements on the Web. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Wildlife trafficking is one of the most common illicit activities, with human, social, and economic consequences [23]. It not only exacts a considerable toll on society through increased criminality [6]

and environmental devastation [25], but may also expose us to unforeseen health and bio-safety hazards [5, 20].

Globalization has had a significant impact on the wildlife trade, leading to a notable expansion from traditional markets to online platforms. Several factors contribute to this transformation, such as increased connectivity, more efficient supply chains, and the rise of e-commerce platforms [29]. While this creates challenges, it also opens new opportunities. As criminals use technology, they leave traces of their activity on the web. These traces have been used in studies that have significantly improved our understanding of online wildlife trafficking. But these studies are often limited to specific species [12, 14, 24, 26, 33], regions [26, 31, 34] or specific sites. To better understand the traffic of wildlife, it is crucial to have large-scale data resources [9, 13] that have a broader coverage of species, regions, and platforms.

Collecting Wildlife Ads at Scale: Opportunities and Challenges. Online marketplaces are a rich source for obtaining traces of wildlife trafficking [15]. Ads obtained from these sites make it possible for researchers to answer important questions regarding the general dynamics of the online trade, e.g., its volume, species targeted, their origin parts sold, and asking prices[31]. However, this opportunity comes with several challenges.

Finding and collecting data to support uncovering illicit activities is difficult [20]. Online marketplaces sell many different types of products, hence a keyword-based search in such sites retrieves products belonging to a variety of categories. Figure 1 we can see 3 possible results that can be found on a search for "Brazilian blue parrot": a postcard, a parrot carved on cristal, and a stuffed parrot – no real animals are returned. Identifying ads that actually include animals and animal parts for sale is important to streamline the data collection, i.e., to reduce the retrieval of irrelevant pages, and for the downstream data analysis. Machine learning classifiers can be trained to perform this identification. However, the process of training a new model is complex and costly due to the scarcity of labeled data. Moreover, traffickers deliberately hide their actions, making the acquisition of suitable training data, characterized by its authenticity and specificity, even harder to obtain.

For these data to be useful for analysis, we must first extract structured information about the products (the attributes in an ad) from the unstructured product pages. Since the structure and format of different sites vary widely with respect to layout, content structure, and presentation, it is challenging to extract data consistently [11, 22, 35]. Often, extraction scripts (scrapers) must be hand-crafted for specific web sites. Embedded metadata from HTML markup is another source of structured data for products. But just like the HTML descriptions, it can come in different formats and not all sites publish the information. This problem is compounded due to the fact the sites are dynamic and often change how their pages (and metadata) are structured. Consequently, scaling up the data collection process to cover a large number of sites can be time-consuming, both to create and continuously update the scrapers.

Our Contribution. We take a first step towards enabling the large-scale collection of online wildlife ads. We have designed a flexible pipeline to collect product pages that are published on different sites and extract information that is useful for analysis and exploration of wildlife trafficking, such as pricing data, images, and sellers. As we discuss in Section 2, this pipeline is scalable and flexible – it can be customized for different types of collections (e.g., the set of species of interest, the platforms to be crawled); can integrate a wide range of models to identify relevant products and perform extraction; and it can store data in easily accessible cloud storage platforms such as s3. To demonstrate the effectiveness of the pipeline, we describe a preliminary dataset we derived using web pages collected over 34 days which contains almost a million ads from 41 marketplaces, covering 235 species and 20 different languages.

Related Work. Our work is related to the field of data collection, focused on the domain of illicit wildlife trade. Keskin et al. [20] offers a comprehensive perspective on the illicit wildlife trade, shedding light on the critical challenges encountered in this domain. One of the most significant challenges highlighted is the scarcity of data, exacerbated by the fact that the available data tends to be biased toward specific regions and particular species. For instance, Cardoso et al. [8] trained a neural network to identify pangolins. Kulkarni and Di Minin [21] highlight the increasing efforts towards wildlife conservation using machine learning, but confirm the challenges of finding good quality labeled data.

Given that e-commerce has become a new point for wildlife trade, some studies analyzing web-crawled data [37] are emerging. Many of the existing studies have used manual searches to gather evidence of the online wildlife trade, particularly products from endangered species [18, 32]. This labor-intensive approach though accurate has limitations, in terms of being slow and hard to scale. Automated computational methods have emerged as a promising solution, as they can continuously monitor a wide range of species across the digital landscape without heavy reliance on human resources. However, existing studies focusing on automated detection have been used to identify very specific endangered species like cacti [27], elephant ivory [16], and orchids [17].

Our research aims to address the data gap by enabling the creation of diverse and extensive datasets that include a wide range of animal species and web-market locations. The dataset we describe is, to the best of our knowledge, the first to cover a large variety of endangered species being advertised in multiple countries.

2 DATA COLLECTION PIPELINE

Figure 2 provides an overview of the key components of the data-collection pipeline. We describe them in detail below.

Seeds and Site Selection. To collect pages using a web crawler, we need to provide as input a set of *seed* URLs that serve as entry points for the crawl. The crawler follows each seed, downloads the corresponding web pages, and recursively follows links extracted from these pages.

For obtaining wildlife-related ads, we generate the seeds dynamically, based on the underlying URLs that e-commerce websites use for their search "forms", combined with a list of names of endangered animals as search queries. Since different websites have distinct search query patterns, thus we need to compile a list of patterns for the selected sites. For example, if we want to search for ads on `ebay.com` using the query `KEYWORD`, the pattern of URL for the search form is: `/sch/i.html_from=R40&_nkw=KEYWORD&_sacat`. Then, we can substitute the query `KEYWORD` with the names in the list of endangered animals.

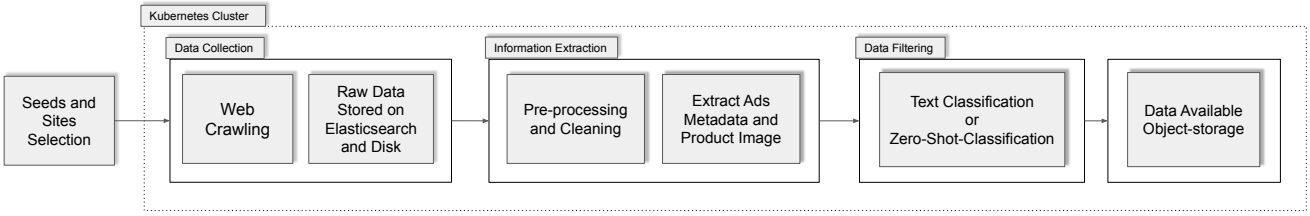


Figure 2: Data Collection Pipeline

Table 1: Data overview: attributes and their descriptions, and the number of records that contain the attributes

Attributes	Description	# of records
url	The ad URL	954,684
title	Product Advisement title	946,732
text	The page text	954,684
product	Name of the product	954,684
description	Description of the product	805,449
domain	Website where the product is posted	954,684
image	URL of the image	787,185
retrieved	time when the page was downloaded	954,684
category	The category listed for that product	25,038
production date	Production date of the product	5,786
price	Price of the product	682,652
currency	Currency of the price	679,717
seller	Seller name	8,910
seller_type	the category the seller is listed	27,483
location	Location of product	25,150
zero_shot_label	zero shot classifier results	954,684
zero_shot_prob	zero shot label probability	954,684
id	UUID used as filename for images	954,684

To construct our dataset, in collaboration with domain experts, we created a list of animals provided by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) that includes endangered species (*CITES Appendix I*) – species that are threatened with extinction and are therefore provided with the highest level of protection.

We generated patterns for a curated list of 49 distinct e-commerce websites, which include 20 eBay platforms operating in various countries. The initial list of websites was selected from the results emerging from web searches using keywords related to the species and other wildlife products. For example, we obtain the top sites resulting from search queries (e.g.,"tiger taxidermy"), and sites that appear in multiple results were compiled to form the initial set.

We use the species name and their respective English names as keywords to generate the seeds. Each of these animals can have one or more English names, leading to a total of 1017 keyword queries. The final list contains 49,833 seeds that include one unique URL for each keyword in each domain (see Figure 1(A)). In Section 3 we describe in detail the dataset collected from these seeds. Note, however, that our approach is general and other lists of animals and websites can be used, depending on the goals of the data collection.

Data Collection. We use the open-source ACHE crawler to perform a *scoped crawl* [1, 7] starting with the 49,833 seeds described above. ACHE downloads all pages from the seed URLs and extracts the links from the pages. Instead of following all links, ACHE scopes its search and only follows links in the domains associated with the seed URLs.

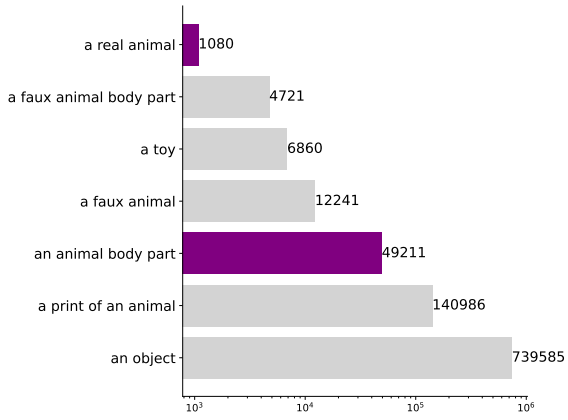
Information Extraction. The crawler retrieves a set of pages that includes ads for individual products. As illustrated in Figure 1(E), these pages have a lot of information that is not pertinent to the actual product, including eBay ads and sponsored items. Therefore, we need to identify and extract the information associated with the product so that we can produce, for each product, a record containing its relevant attributes, e.g., *Price, Seller, Product type, Description, and Product image*. But doing so is challenging due to the diversity of content and structure used by different sites. We implemented a set of strategies that we combined to address this challenge.

Extract page content We use BeautifulSoup [30] to parse the HTML content on the page and extract the page title and text content.

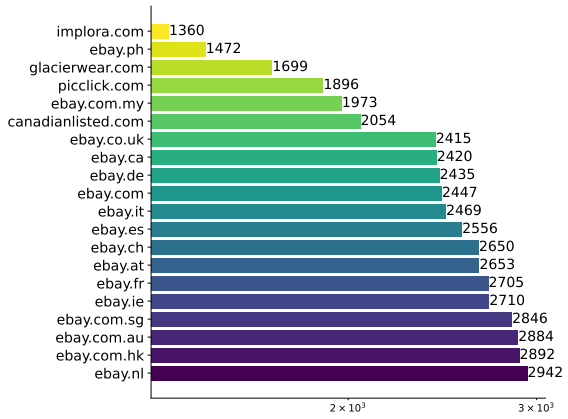
Extract product attributes We scrape the HTML page to extract specific attributes such as price and seller. With MLscraper [4], we can automatically perform this extraction by just providing the tool with a few examples of the desired output based on the information available on a sample HTML page. *Extract product metadata* Some pages have embedded metadata that contains information about the product. To obtain this information, we use Extruct, [2], a library designed for extracting embedded metadata from HTML markup.

Data Filtering. Table 1) shows a summary of the data derived by the extraction step. While there are almost 1 million records, many of the products are not real animals (or animal products). As illustrated in Figure 1, there are postcards, plush toys, and decorative items that are returned for queries that search for wildlife. To filter out *irrelevant* products, we use *text-classification* and *zero-shot-classification* using large-language models (LLMs). We can choose any textual attribute as input to the classifiers, in our dataset we used the name of the product. The pipeline is flexible with respect to the model it uses to perform the task, e.g., we can choose any model available on HuggingFace [3] for zero-shot or text classification.

For Zero-Shot Classification, we use a fine-tuned model of *XLM-RoBERTa* [10]. The model [19] is fine-tuned on a combination of data in 15 languages, but can also be effective in other languages since RoBERTa was trained in 100 different languages. We provide the model one hypothesis: *'This product advertisement is about .'*, and the candidate labels (*a real animal, a toy, a print of an animal, an object, a faux animal, an animal body part, a faux animal body part*). The output consists of the labels and their respective probabilities, which are added as attributes in our dataset.



(a) Distribution of zero-shot-classification labels from hypotheses: *"This product advertisement is about:"*



(b) Top 20 domains where real animal and an animal body part classes are founded

Figure 3: Distribution of zero-shot classes and domains

Implementation Details. The pipeline runs on a Kubernetes cluster and is deployed with Docker images. The information extraction and model classification are executed as Kubernetes Jobs, sending data to S3-like object storage (MinIO) as parquet files. We can access the data, both on Elasticsearch and MinIO, from a JupyterHub also available on the cluster, and using DuckDB, we can execute SQL-style queries directly from the object store. This makes it possible to easily access and explore the data.

3 THE WILDLIFE AD DATA COLLECTION

Data Throughput. To give an idea of the crawler throughput, we deployed a crawler on 2023-08-08. As of 2023-09-05 (after 34 days), it had retrieved over 11 million pages. The average time it takes to fetch a single page is 695.50 milliseconds (ms). This represents the typical response time for page retrieval. Due to politeness constraints, we avoid overloading the web servers where the pages reside, this way we limit the number of requests to the same server. Running the actual pipeline takes longer: the current implementation processes approximately 145,000 pages per day.

Data Overview. We tested our pipeline using 954,684 pages and derived the dataset described in Table 1. Each record corresponds to a web page and includes the URL, domain, as well as the time when the page was downloaded. As expected, not all pages contain all attributes. For instance, the price is available for 682,652 pages. Some attributes are rarer – seller information was extracted for only 8,910 pages and location for 25,150.

Figure 3a shows the distribution of the zero-shot classes in the dataset. The classifier identified 1,080 products as being "real animals" and 49,211 as being "an animal body part". While the zero-shot classifier is not fool-proof, the low percentage of (potentially) relevant products underscores the importance of having an automated data collection and processing pipeline.

The ability to scour a very large number of pages and sites opens the opportunity to obtain data at a scale not previously possible. For example, our collection includes sites in different countries. Figure 3b shows the top 20 domains in which the classes, "a real animal" and "an animal body part", are found. We can notice that from these 20 domains, 8 are from domains of non-English speaking countries, such as the Netherlands and Hong Kong.

4 DISCUSSION

The pipeline we designed and implemented as a first step towards enabling large-scale data collection for wildlife products sold on the Web. Using this pipeline, it is possible to collect datasets that cover a wide range of data sources, in different countries, and that include a large set of species. We have designed the system to be flexible, i.e., it can be configured for specific tasks such as collections that focus on a specific species or sites; and by making it open source, we hope that the community will be able to collectively create and share datasets that can provide insights into wildlife trafficking.[1]

An important goal of our design was to make the pipeline extensible. In this paper, we describe our current implementation and specific choices we have made for the different components. However, it is possible to replace these components. For example, alternative classifiers could be applied for data filtering, and different scraper mechanisms can be used for extraction.

There are a number of directions we intend to pursue in future work to improve our pipeline. The use of zero-shot models for data filtering is promising, but there is room for improvement. We are exploring the use of fine-tuned models such as DistilBert [28] as well as multi-modal models (which use both text and images) [38] to improve the classification accuracy, in particular, to distinguish animal products from toys, prints, clothes, etc.

While rule-based systems such as MLScraper greatly simplify the task of extracting structured information from unstructured web pages, they are brittle and can fail in practice. We would like to investigate the integration of deep-learning-based techniques (e.g., [35, 36]) to create extractors that are robust and able to perform extraction from diverse websites, without the need for site-specific training.

Acknowledgments. This work was funded by the NSF Disrupting Operations of Illicit Supply Networks (D-ISN) program.

[1]The code will be available on GitHub. We will also publish the dataset.

REFERENCES

[1] 2023. ACHE Crawler. https://github.com/VIDA-NYU/ache/.
[2] 2023. Extruct. https://github.com/scrapinghub/extruct.
[3] 2023. HuggingFace. https://huggingface.co/.
[4] 2023. Mlscraper. https://github.com/lorey/mlscraper.
[5] A Alonso Aguirre, Meredith L Gore, Matt Kammer-Kerwick, Kevin M Curtin, Andries Heyns, Wolfgang Preiser, and Louise I Shelley. 2021. Opportunities for transdisciplinary science to mitigate biosecurity risks from the intersectionality of illegal wildlife trade with emerging zoonotic pathogens. *Frontiers in Ecology and Evolution* 9 (2021), 604929.
[6] Michelle Anagnostou and Brent Doberstein. 2022. Illegal wildlife trade and other organised crime: A scoping review. *Ambio* 51, 7 (2022), 1615–1631.
[7] Luciano Barbosa and Juliana Freire. 2007. An adaptive crawler for locating hidden-web entry points. In *Proceedings of the 16th international conference on World Wide Web*. 441–450.
[8] Ana Sofia Cardoso, Sofiya Bryukhova, Francesco Renna, Luís Reino, Chi Xu, Zixiang Xiao, Ricardo Correia, Enrico Di Minin, Joana Ribeiro, and Ana Sofia Vaz. 2023. Detecting wildlife trafficking in images from online platforms: A test case using deep learning with pangolin images. *Biological Conservation* 279 (2023), 109905.
[9] Sandra Charity and Juliana Machado Ferreira. 2020. Wildlife trafficking in Brazil. *TRAFFIC International, Cambridge, United Kingdom* 140 (2020).
[10] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).
[11] Xin Luna Dong, Hannaneh Hajishirzi, Colin Lockard, and Prashant Shiralkar. 2020. Multi-modal information extraction from text, semi-structured, and tabular data on the web. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3543–3544.
[12] Lalita Gomez and Chris R Shepherd. 2019. Bearly on the radar–an analysis of seizures of bears in Indonesia. *European Journal of Wildlife Research* 65, 6 (2019), 89.
[13] Timothy C Haas and Sam M Ferreira. 2015. Federated databases and actionable intelligence: using social network analysis to disrupt transnational wildlife trafficking criminal networks. *Security Informatics* 4, 1 (2015), 1–14.
[14] Lauren Harrington, David Macdonald, and Neil D'Cruze. 2019. Popularity of pet otters on YouTube: evidence of an emerging trade threat. *Nature Conservation* 36 (2019).
[15] Jo Hastie and Tania McCrea-Steele. 2014. *Wanted-dead or alive: exposing online wildlife trade*. International Fund for Animal Welfare.
[16] Julio Hernandez-Castro and David L Roberts. 2015. Automatic detection of potentially illegal online sales of elephant ivory via data mining. *PeerJ Computer Science* 1 (2015), e10.
[17] Amy Hinsley, Tamsin E Lee, Joseph R Harrison, and David L Roberts. 2016. Estimating the extent and structure of trade in horticultural orchids via social media. *Conservation Biology* 30, 5 (2016), 1038–1047.
[18] IFAW. 2008. Killing with keystrokes: an investigation of the illegal wildlife trade on the world wide web. (2008).
[19] joeddav. 2023. *joeddav/xlm-roberta-large-xnli*. https://huggingface.co/joeddav/xlm-roberta-large-xnli
[20] Burcu B Keskin, Emily C Griffin, Jonathan O Prell, Bistra Dilkina, Aaron Ferber, John MacDonald, Rowan Hilend, Stanley Griffis, and Meredith L Gore. 2022. Quantitative investigation of wildlife trafficking supply chains: A review. *Omega* (2022), 102780.
[21] Ritwik Kulkarni and Enrico Di Minin. 2023. Towards automatic detection of wildlife trade using machine vision models. *Biological Conservation* 279 (2023), 109924.
[22] Bill Yuchen Lin, Ying Sheng, Nguyen Vo, and Sandeep Tata. 2020. Freedom: A transferable neural architecture for structured information extraction on web documents. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1092–1102.
[23] Nicholas Magliocca, Aurora Torres, Jared Margulies, Kendra McSweeney, Inés Arroyo-Quiroz, Neil Carter, Kevin Curtin, Tara Easter, Meredith Gore, Annette Hübschle, et al. 2021. Comparative analysis of illicit supply network structure and operations: Cocaine, wildlife, and sand. *Journal of illicit economies and development* 3, 1 (2021), 50–73.
[24] Rowan O Martin, Cristiana Senni, and Neil C D'Cruze. 2018. Trade in wild-sourced African grey parrots: Insights via social media. *Global Ecology and Conservation* 15 (2018), e00429.
[25] Sean L Maxwell, Richard A Fuller, Thomas M Brooks, and James EM Watson. 2016. Biodiversity: The ravages of guns, nets and bulldozers. *Nature* 536, 7615 (2016), 143–145.
[26] David L Roberts, Katya Mun, and EJ Milner-Gulland. 2022. A systematic survey of online trade: trade in Saiga antelope horn on Russian-language websites. *Oryx* 56, 3 (2022), 352–359.
[27] Maurizio Sajeva, Claudio Augugliaro, Matthew J Smith, and Elisabetta Oddo. 2013. Regulating internet trade in CITES species. *Conservation Biology* 27, 2 (2013), 429.
[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv 2019. *arXiv preprint arXiv:1910.01108* (2019).
[29] Penthai Siriwat and Vincent Nijman. 2020. Wildlife trade shifts from brick-and-mortar markets to virtual marketplaces: A case study of birds of prey trade in Thailand. *Journal of Asia-Pacific Biodiversity* 13, 3 (2020), 454–461.
[30] Beautiful Soup. 2023. *Beautiful Soup Documentation*. https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[31] Sarah Stoner. 2014. Tigers: exploring the threat from illegal online trade. *TRAFFIC Bulletin* 26, 1 (2014), 26–30.
[32] TRAFFIC. 2019. Wildlife Crime linked to the Internet: new TRAFFIC report highlights experiences from China. (2019).
[33] Sofia Venturini and David L Roberts. 2020. Disguising elephant ivory as other materials in the Online Trade. *Tropical Conservation Science* 13 (2020), 1940082920974604.
[34] English Vietnamese. 2016. A rapid assessment of e-commerce wildlife trade in Viet Nam. *TRAFFIC Bulletin* 28, 2 (2016), 53.
[35] Qifan Wang, Yi Fang, Anirudh Ravula, Fuli Feng, Xiaojun Quan, and Dongfang Liu. 2022. Webformer: The web-page transformer for structure information extraction. In *Proceedings of the ACM Web Conference 2022*. 3124–3133.
[36] Qifan Wang, Li Yang, Jingang Wang, Jitin Krishnan, Bo Dai, Sinong Wang, Zenglin Xu, Madian Khabsa, and Hao Ma. 2022. SMARTAVE: Structured Multimodal Transformer for Product Attribute Value Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 263–276.
[37] Qing Xu, Mingxiang Cai, and Tim K Mackey. 2020. The illegal wildlife digital market: an analysis of Chinese wildlife marketing and sale on Facebook. *Environmental conservation* 47, 3 (2020), 206–212.
[38] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549* (2023).