Conservation Biology

Essay

A guide to using the internet to monitor and quantify the wildlife trade

Oliver C. Stringham [1,2*] Adam Toomes [1] Aurelie M. Kanishka,[1,3] Lewis Mitchell,[2] Sarah Heinrich [1] Joshua V. Ross,[2] and Phillip Cassey[1]

[1]Invasion Science & Wildlife Ecology Lab, University of Adelaide, Adelaide, SA, 5005, Australia
[2]School of Mathematical Sciences, University of Adelaide, Adelaide, SA, 5005, Australia
[3]Fenner School of Environment and Society, The Australian National University, Canberra, ACT, 2601, Australia

Abstract: The unrivaled growth in e-commerce of animals and plants presents an unprecedented opportunity to monitor wildlife trade to inform conservation, biosecurity, and law enforcement. Using the internet to quantify the scale of the wildlife trade (volume and frequency) is a relatively recent and rapidly developing approach that lacks an accessible framework for locating relevant websites and collecting data. We produced an accessible guide for internet-based wildlife trade surveillance. We detailed a repeatable method involving a systematic internet search, with search engines, to locate relevant websites and content. For data collection, we highlight web-scraping technology as an efficient way to collect data in an automated fashion at regularly timed intervals. Our guide is applicable to the multitude of trade-based contexts because researchers can tailor search keywords for specific taxa or derived products and locations of interest. We provide information for working with the diversity of websites used in wildlife trade. For example, to locate relevant content on social media (e.g., posts or groups), each social media platform should be examined individually via the site's internal search engine. A key advantage of using the internet to study wildlife trade is the relative ease of access to an increasing amount of trade-related data. However, not all wildlife trade occurs online and it may occur on unobservable sections of the internet.

Keywords: big data, dark web, deep web, e-commerce, pet trade, social media, surface web, web scraping

Resumen: Una Guía para Usar el Internet para Monitorear y Cuantificar el Mercado de Fauna

El crecimiento incomparable del comercio en línea de animales y plantas representa una oportunidad sin precedentes para monitorear el mercado de fauna y así orientar a la conservación, la bioseguridad y la aplicación de la ley. El uso del internet para cuantificar la escala del mercado de fauna (volumen y frecuencia) es una estrategia relativamente reciente y de rápido desarrollo que carece de un marco de trabajo accesible para la localización de sitios web relevantes y para la recolección de datos. Realizamos una guía accesible para la vigilancia del mercado de fauna en internet. Detallamos un método repetible que involucra una búsqueda sistemática por internet, por medio de buscadores, para localizar sitios web y contenidos relevantes. Para la recolección de datos, resaltamos la tecnología de *web scraping* como una manera eficiente de obtener datos de manera automatizada a intervalos regulares de tiempo. Nuestra guía puede aplicarse a la multitud de contextos basados en el mercado porque los investigadores pueden adaptar las palabras de búsqueda a taxones específicos o productos derivados y a localidades de interés. Proporcionamos información para poder trabajar con la diversidad de sitios web que se usan para el mercado de fauna. Por ejemplo, para localizar contenido relevante en las redes sociales (p. ej.: publicaciones o grupos), cada plataforma social debería ser examinada individualmente por medio del buscador interno del sitio. Una ventaja importante de usar el internet para estudiar el mercado de fauna es el acceso relativamente sencillo a una creciente cantidad de datos relacionados con el mercado. Sin embargo, no todo el mercado de fauna ocurre en línea y puede que suceda en secciones inobservables del internet.

Palabras Clave: comercio en línea, macrodatos, mercado de mascotas, redes sociales, web oscura, web profunda, web superficial, web scraping

*email:oliverstringham@gmail.com
***Article Impact Statement**: The internet is a vast source of wildlife trade data; our generalizable framework allows researchers to explore new contexts of the trade.*
Paper submitted May 27, 2020; revised manuscript accepted November 27, 2020.

DOI: 10.1111/cobi.13675

Background

The wildlife trade is an influential driver of species endangerment, spread of invasive species and diseases, and provisioning of criminal activity ('t Sas-Rolfes et al. 2019). Wildlife trade occurs across a variety of physical and virtual settings, including brick-and-mortar stores, wet markets, and digital platforms on the internet (e.g., Alfino & Roberts 2018). Reliable data on the quantity and composition of the wildlife trade (legal and illegal) are vital for informing decisions about conservation, biosecurity, and law enforcement and developing campaigns to change human behavior. Yet these data are rarely collected or are difficult to obtain (Regueira & Bernard 2012; Eskew et al. 2020). In recent years, the internet has played an increasingly important role in facilitating trade in wildlife (Siriwat & Nijman 2020).

Researchers have used data from the internet in various ways to inform wildlife trade research and assist in practical management, including law enforcement. These studies have generally been small in scale (i.e., monitoring one or few websites), but have nonetheless revealed the utility of the internet to describe different aspects of the wildlife trade. In the context of conservation, classified and advertisement websites have been used to estimate intensity of trade and support increases in the legal protection of high-risk species (Rowley et al. 2016). For biological invasions, online pet stores have been used to inventory non-native species (Stringham & Lockwood 2018). Lost-and-found websites have been used to estimate propagule pressure, a major determinant of non-native establishment probability (Cassey et al. 2018), for commonly held exotic pets (e.g., turtles [Kikillus et al. 2012]). In terms of assisting law enforcement, listings from online classified and advertisement websites have been used to quantify the illegal trade (Ye et al. 2020), and social media websites have been used to track the intensity of legal and illegal trade (Jensen et al. 2019).

As the volume and frequency of wildlife trade increases over the internet, having a unified method for using the internet to obtain data on the wildlife trade becomes more critical for researchers. However, such a method, or guide, does not currently exist. By outlining a guide with repeatable steps, we hope to facilitate reproducible methods for using the internet as a data source (including finding websites, data collection, and curation). Further, a guide can serve as a primer for investigating unexplored contexts of the trade, including new locations and different focal taxa, or emerging trade in derived organism parts and commodities.

We produced an accessible guide to using the internet to gather data on the wildlife trade. We developed the method based on our collective knowledge of working with web data and the wildlife trade combined with the methods used in prior published studies. Our goal was for this guide to be used by scientists, nongovernmental organizations (NGOs), government agencies, and other parties who wish to utilize the internet as a source of data on wildlife trade.

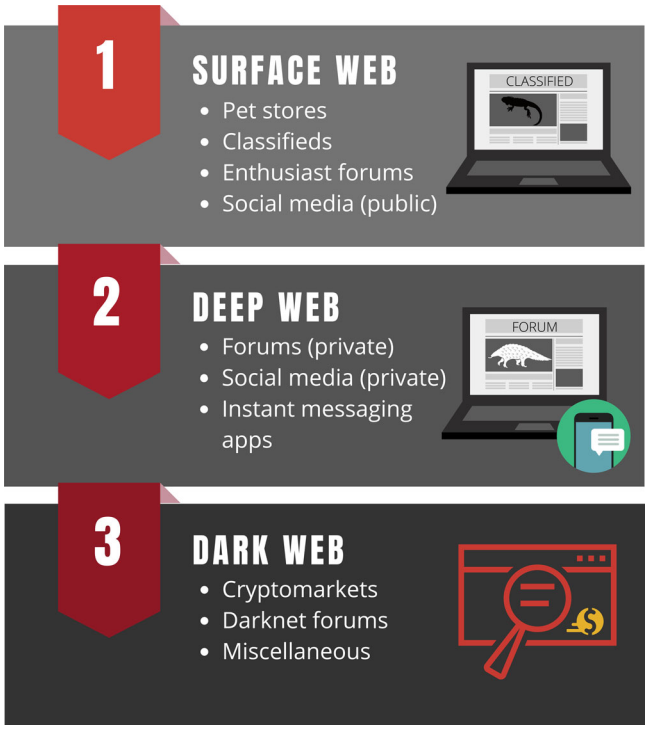

Figure 1. The 3 layers of where wildlife trade occurs on the internet and their components.

Structure of the Internet

The internet (i.e., the World Wide Web or simply the web) is categorized into 3 distinct layers: surface web, deep web, and dark web (Fig. 1) (Bergman 2001). Each layer differs in 1 of 2 factors: whether it is accessible without logging in or invitation (i.e., is publicly viewable) and whether it is indexed by a search engine (i.e., will appear as a result in a search engine). The surface web includes any website that is publicly viewable and is indexed by search engines (e.g., e-commerce websites). The deep web includes websites or online content that require either logging in or an invitation to view (e.g., social media, private messaging apps). Some deep web sites may be indexed by a search engine (e.g., public Facebook or Twitter posts), whereas others may not (e.g., WhatsApp). The dark web contains purposefully hidden content that requires specialized software to access, requires either logging in, or an invitation to view, and is not indexed by any search engine (Chen 2011; CRS 2017). The degree to which researchers can find relevant wildlife trade content on the internet is

influenced by how findable a website or content is (e.g., can a search engine find it). Further, the ethical considerations of collecting data depend in part on how accessible the website is (e.g., is deceit or limited disclosure required to gain access to the content).

Data Available on the Internet

Data availability on wildlife trade varies by website and even within a website (Toivonen et al. 2019) (Appendix S1). On a basic level, online advertisements (i.e., listings or posts) are provided in the form of text, pictures, and videos. Foremost, the name of the species, taxa, or derived product traded is usually stated. Characteristics of the traded taxa or product can include quantity (number, size, volume), age, sex, size, color, morph, and provenance (domestic bred, wild caught, or harvested). The physical location of the advertisement (i.e., city) and metadata on the advertisement itself, such as the number of page views and username of the trader, may be provided. Further, the current purpose for which the wildlife is being used (pet, medicinal, food, etc.) and the rationale for trading the wildlife (e.g., profit, lifestyle change) can sometimes be ascertained from advertisements with open text fields. These attributes may aid in understanding motives associated with wildlife trade participation or consumption (i.e., conservation culturonomics) (Ladle et al. 2016).

Guide to Using the Internet to Monitor and Quantify the Wildlife Trade

Our guide has 6 steps (Fig. 2): 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 results. We generalized the guide for websites found in any layer of the internet (including social media) and detailed how to adapt this guide to different languages and countries. Figure 3 shows 2 hypothetical case studies that accompany and contextualize each step of the guide. For more generalized frameworks on working with social media and online news data, refer to Toivonen et al. (2019) and Sonricker Hansen et al. (2012), respectively.

Defining the Scope and Purpose of the Project (Step 1)

At a minimum, it is essential to decide which species, taxa, or derived products are of interest, the location or locations of interest, and the time frame for data collection (i.e., a single snapshot versus ongoing monitoring for months to years). Considering what type of website (Appendix S1) or layer of the internet may be appropriate. The research questions that can be answered are influenced by the data available on the internet. Thus, there will likely need to be some exploration of the websites and the kind of data they provide (steps 2–3). Examples of project aims include quantifying the trade in parrots in different regions of China (Ye et al. 2020), investigating the sale of pangolin-leather boots in the United States (Heinrich et al. 2019), and exploring the social network structure of sellers of horticultural orchids (Hinsley et al. 2016).

Figure 2. Flowchart of guide to using the internet to monitor and quantify the wildlife trade.

Finding Candidate Websites Where Specific Taxa and Wildlife Products Are Traded (Step 2)

Search engines can be used to find candidate websites (e.g., e-commerce sites, forums). Finding relevant social media content requires special considerations, which we detail below. Outside of the search engines, other approaches to finding candidate websites and choosing target websites include interviewing a specific community of practice (e.g., reptile keepers and traders) or collaborating with other researchers actively engaged in online wildlife-trade monitoring (e.g., governmental

Figure 3. Two hypothetical case studies (columns) following the first 5 steps of our guide to for internet-based wildlife trade surveillance: identify taxa or products of interest; find candidate websites or social media content; selecting websites or content to monitor; collecting data; and cleaning data. The first study (left column) is of trade in non-native ornamental plants in online plant shops or nurseries in Australia (i.e., open web) (sensu Lenda et al. 2014). Keywords are generated that apply to the species of interest, including scientific and trade names and qualifiers, such as for sale *or* store*, are included to create search phrases for the search engines (details in Appendix S2). Search engines (Google and DuckDuckGo) provide a list of candidate websites from which a subset is chosen based on inclusion criteria (store sells ≥1 species of interest and store offers to ship plants or seeds interstate) (green checkmark). Web scrapers are used to collect data on a biweekly basis for 1 year (green check marks on calendar). Data cleaning is less intensive (1 hourglass) than in second case study because data from individual stores tend to be more organized than social media sites (Appendix S1). Because this study explores many species, linking each traded taxa to the Global Biodiversity Information Facility (GBIF) facilitates efficient data cleaning and analysis. For the trade of exotic leather boots made from pangolin skins occurring on social media in the United States (right column) (Heinrich et al. 2019), preliminary investigation revealed several*

agencies or NGOs). It is important to note that the internet is transient: traders go out of business and new ones emerge. Thus, websites found at one time can differ in composition and function if surveyed later. If the goal is long-term monitoring, we suggest revising the list of current relevant websites at regularly timed intervals.

For the surface web, finding candidate websites involves three steps: defining keyword phrases to search, using a search engine to perform searches, and classifying the relevance of each search result. This part of the method is akin to the process of finding relevant scientific articles in a systematic review or meta-analysis (i.e., PRISMA method) (Koricheva et al. 2013). However, instead of searching the scientific literature, the internet is searched (via search engines), and not all candidate results will be used for data collection.

Search phrases are a combination of relevant keywords. We recommend developing a suite of keywords for each target taxa (e.g., species name, common name, product name), type of websites (Appendix S1), and location of interest. Other useful keywords include adding the terms *for sale* or *buy*. Example search phrases may be: "*snakes for sale Australia*," "*marine fish forum USA*," or "*orchid store UK*" (detailed example in Appendix S2). These search phrases should be in the language(s) written in the location of interest. There may be a need to refine keywords after exploratory investigation of search engine results. In particular, there may be trade names (i.e., names for species or taxa used in the wildlife trade community, but not commonly used among scientists), local or regional names or names of breeds, morphs, and mutations (e.g., Lyons & Natusch 2013) that are not captured in the initial formulation of search phrases.

Search engines (e.g., Google) use proprietary algorithms to return a list of URLs (i.e., website addresses) when a search phrase is input. Search engine algorithms consider the relevance of the keywords, the popularity of the website (i.e., number of page views), and, increasingly, the location of where the search occurs (Langville & Meyer 2011). The results from a search engine are expected to change over time due to: changes to the search engine algorithm, changes to website popularity metrics, emergence of new websites, or a change in the location of where the search is performed. Once a keyword phrase is searched, the search engine will likely return millions of URLs per phrase. We recommend choosing a cutoff point that balances the quality of search results with search effort (Appendix S3). Because search engines can use the user's location to provide personalized results (e.g., Google: https://policies.google.com/technologies/location-data), extra steps must be taken to ensure that the search engine provides location- and language-relevant results. One way to control the location is to use advanced search features (e.g., https://www.google.com/advanced_search), which allows the researcher to specify which country and languages to restrict a search to. In addition, using a virtual private network may alleviate location issues. For more information on search engines, see Appendix S3.

Websites on the deep web indexed on search engines will be findable with the same approach outlined for the surface web (e.g., private forums). Currently, aside from expert consultation or interviewing communities of practice, there are no generalizable or automated methods for locating content on the deep web that is not indexed by search engines (e.g. WhatsApp, WeChat, other private messaging apps), or on the dark web. While some algorithms exist for querying deep websites (e.g., Liakos et al. 2016), the actual implementation of these algorithms as web crawlers must be tailored for each individual instance and require unique login details. This severely limits any large-scale monitoring efforts.

After obtaining URLs from search results, each will need to be categorized as relevant or irrelevant. Relevance is subjective, and we recommend defining inclusion and exclusion criteria depending on the scope and purpose of the study. One obvious inclusion criterion is whether the target taxon is traded on the website. Another criterion can be the type of transaction that occurs on the website. Specifically, on the internet, there are varying levels of directness of trade. For instance, some e-commerce companies will ship live animals or products to a customer's doorstep (e.g., pet stores) (Holmberg et al. 2015). On the other hand, there are websites that only facilitate the transaction of selling wildlife online and leave it up to the individuals in the transaction to conduct the exchange (e.g., classifieds: Sung & Fong 2018).

Social media websites vary in structure and format (Appendix S1) (Toivonen et al. 2019). For our purposes, we categorized content found on social media websites into two categories: consolidated and unconsolidated. The differences between each category influence how

hashtags used in their sale. These hashtags are supplied to the internal search engines of the social media sites (Facebook and Instagram). All posts returned from the search engine become the data (i.e., unconsolidated social media content). Data collection occurs every other day because social media content tends to be updated frequently. Data cleaning takes longer than for the other study because more listings are collected and listings are structured as open-text boxes, which must be read and parsed by humans to verify what is being advertised. Natural language processing and associated tools (i.e., fuzzy string matching) can be used to narrow down the number of listings needed to be cleaned (e.g., using text classification models to identify and remove irrelevant posts). The GBIF logo is used with permission (GBIF Secretariat, Copenhagen, https://www.gbif.org).

researchers find relevant social media content related to wildlife trade. Consolidated social media content includes groups dedicated to a particular purpose (e.g., ornamental orchid traders) in which users share content that is only viewable by other group members (e.g., Facebook groups). Social media groups function similarly to forum websites. Unconsolidated social media content consists of users posting to the social media platform at large or to a group of followers. Twitter, for example, is mostly public, where all Tweets (i.e., posts) are viewable by all users. Some social media websites, such as Facebook, may have both consolidated and unconsolidated content.

Social media websites have their own internal search engine, which searches through content of the specific social media site. Thus, for consolidated social media content, we recommend adapting our approach outlined for the surface web (i.e., using search phrases) for internal search engines to find relevant social media groups (e.g., Siriwat & Nijmans 2020). These groups can then be classified by their relevance and considered for monitoring. For unconsolidated social media content, we recommend simply using the internal search engine to search for relevant posts. The posts returned by the search engine become the data itself (e.g., Xu et al. 2019), where the classification and selection steps of this guide are skipped. Many social media users utilize hashtags (#), which are user-generated tags relating to the post's content (e.g., #ivory). Thus, for social media sites, determining what hashtags are used for a specific context of wildlife trade may yield more relevant search results than keyword phrases for both consolidated and unconsolidated social media content (e.g., Morgan & Chng 2018).

Application programming interfaces (APIs) may be available for some social media websites. These may allow for bulk searches (i.e., more than one search at once) and streamlined data collection. Filters may be available in advanced search options of internal search engines or in APIs to restrict search results to certain countries and languages. Finally, social media companies have allowed users to adjust their privacy settings so that only their "followers" or a preselected group of users can view their posts. Content with privacy restrictions may be hidden from internal search engines or API results.

Selecting Target Sites to Monitor (Step 3)

After obtaining the list of candidate websites, the next step is to select which websites to collect data from (i.e., target websites). This step of the framework is the most subjective, and therefore some level of justification and transparency should be provided when choosing target websites. To make informed decisions on selecting target websites, metadata on candidate websites can be collected. For surface websites, one metadata attribute is web traffic statistics, which includes information such as the number of page views per month (details in Appendix S4). In addition, for any website, researchers can calculate the average number of posts or listings per day and use this as a proxy for popularity. Ultimately, researcher discretion is needed to choose target websites because measures of website metadata are not available for all candidate websites and project relevance is not always straightforward to quantify. The number of target websites chosen varies based on the project aim and the resources available to collect and clean data. Again, expert opinion and communities of practice can provide opinions on what websites are most relevant.

Collecting Data from Websites (Step 4)

Data collection is either manual or automated. Manual data collection involves visiting the website and recording the taxa or product is being traded and the desired associated attributes (e.g., price, location). Automated data collection involves constructing web scrapers to visit the website and extract desired relevant information (Fig. 4) (Singrodia et al. 2019). Web scrapers organize the contents of a website into a structured tabular format (for more information on web scrapers and data storage, see Appendix S5). Because each website differs in its underlying structure, custom web scrapers need to be coded for each website individually. A few highly visited websites may have APIs that allow for easy collection of data; this is more likely to be the case for social media websites (e.g., Twitter) (Toivonen et al. 2019). The choice of manual or automated data collection depends on how long and how often data are being collected because it takes technical expertise and time to build web scrapers, which may not be necessary if the number of target websites is small and the data-collection window is short (e.g., Heinrich et al. 2020). These methods of data collection apply to websites and content on the surface web, deep web (including social media), and dark web as long as researchers have access to the website or content (e.g. Cunliffe et al. 2019).

Ethics approval is required to collect information from the internet, especially when personally identifiable material is collected, including, but not limited to, social media sites (Zimmer 2010). Care should be taken to ensure de-identified information is used for analyses and subsequent publication (Harriman & Patel 2014; Sula 2016). Furthermore, collecting data from any deep or dark web site requires ethics approval (Tai et al. 2012) because deceit or limited disclosure of research aims may be required to obtain a login or approval to join the site. Also, automated data collection processes (i.e., web scraping) are a legal gray area (Zamora 2019). Thus, we encourage researchers to acquire ethics approval prior to using them. For specific recommendations of ethical practice, refer to Appendix S6.

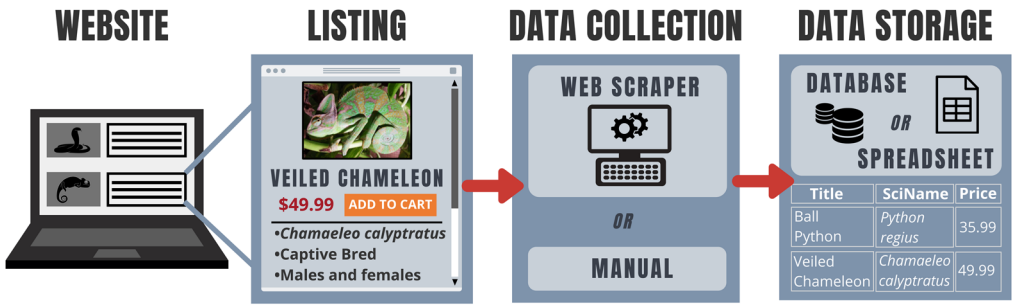

Figure 4. Data collection and storage procedure for websites trading wildlife. Website underlying HTML code is parsed by web scrapers to extract relevant information. The process is repeated for different websites with custom web scraper code (details in Appendix S5). Frequency of data collection depends on the nature of the website, including how often the website is updated. If data collection occurs frequently, automated data collection should be used because manual collection is time consuming. However, there is a trade-off between the resources invested in creating web scrapers and the quantity of data collected. Chameleon photo by Chris Kade.

Data Cleaning (Step 5)

Data cleaning involves curating each listing (i.e., post or advertisement) for attributes that could not be automatically extracted, but are required for the analysis, such as species name, quantity, price, or location. Data cleaning is often a tedious and time-consuming task (Freitas & Curry 2016) and could be the most time-consuming part of the entire project. The amount of cleaning required depends on the structure of the website and varies by individual website (Appendix S1). For instance, some websites may have a separate field for species names, whereas others may just have a free-form open text box where the user can type anything. Our experience with websites involving the wildlife trade is with the latter, which takes substantially more time to clean. If collecting data manually, simultaneously cleaning data during collection is possible and likely desirable.

Resolving the species name in a listing or post is one of the most important aspects of data cleaning. Some pet stores and specialist classifieds websites explicitly state the scientific name, whereas other sites may mention common names, trade names, or simply supply a photo. For all practical purposes, identifications down to the rank of species are needed for effective action on conservation, biosecurity, and crime (Rhyne et al. 2012). Therefore, we recommend identifying the taxa to the most specific taxonomic level possible. If pictures are provided, taxonomic experts can aid in species identification. Yet, the quality of pictures may be too poor to properly identify species. In some instances, online traders may simply not provide enough information in the listing for species-level identification.

If monitoring many species, we recommend relating the species or taxon name to a taxonomic database (e.g., GBIF 2020). Doing so facilitates conformation to taxonomic names by avoiding synonyms and misspellings (Gallagher et al. 2020). In addition, it enables the researcher to easily acquire upstream taxonomy (e.g., family and order). We recommend the R package taxize, which automates the gathering of upstream taxonomy if supplied a scientific name or database identifier for the taxa of interest (Chamberlain & Szocs 2013).

Advantages and Caveats of Web Data

The ease of gathering data from the internet is the main advantage compared with surveying physical markets or stores, especially if one uses automated data collection techniques (i.e., web scrapers). Furthermore, using the internet could potentially allow for a more complete picture of the trade both spatially and temporally than would normally be possible for researchers or organizations who have limited resources for traditional surveys. However, the internet is not a panacea for monitoring the wildlife trade, and relying on the internet for data on the wildlife trade has several disadvantages. First, not all trade occurs or is observable online (e.g., bushmeat trade) (McNamara et al. 2019). The degree to which trade occurs online depends on the type of trade (i.e., pet, derived products, food, etc.), the taxa, the country or culture in question (i.e., internet use varies by country) (Pew Research Center 2016), and possibly the target or consumer group. To the best of our knowledge, there are no estimates of the ratio of physical versus online trade for any context. Another downside is that it is difficult, if not impossible, to verify the validity of online listings of wildlife (i.e., fake or scam versus genuine advertisements). Supplementing data collected online with physical surveys is a more holistic approach that may be more useful when considering applied outcomes (e.g., Rowley et al. 2016).

Considerations for the Deep and Dark Web

Wildlife trade on the surface web and indexed deep web (e.g., social media) is extremely abundant (IFAW

2018; Sung & Fong 2018; Xu et al. 2020). The unindexed deep web, such as private text messaging apps (e.g., WhatsApp, Facebook Messenger), has remained relatively unexplored until recently (e.g., Setiawan et al. 2019; Sanchez-Mercado et al. 2020); thus, the extent of trade is unknown. Given the ease of access of private messaging apps and the anonymity they provide, we hypothesize that trade is also abundant on the unindexed deep web. The dark web remains elusive. While there is evidence that wildlife is not traded on common dark web marketplaces, this does not discount the potential for trade to be occurring elsewhere on the dark web (Harrison et al. 2016; Roberts & Hernandez-Castro 2017). Further, future policies enacted in response to concerns of wildlife trade may shift the balance of where wildlife trade occurs on the internet (Roe et al. 2020). Specifically, new regulations or improved enforcement of illegal trade can unintentionally drive trade away from the open and indexed deep web to the unindexed deep web and dark web (Nijman 2020; Appendix S7), ultimately making it more difficult for researchers to locate wildlife trade online.

Websites and content on the deep and dark web present several challenges for researchers. First, finding websites that trade wildlife on the unindexed deep and dark web is difficult because they are not accessible by search engines. This is an unfortunate reality for researchers, but reflects an intentional design to keep this information private. Further, obtaining access to deep and dark websites often requires researchers to use deceit for successful infiltration. Using deceit requires ethics approval and infiltration requires skills and training that conservation researchers may not have (e.g., remaining anonymous). Thus, interdisciplinary collaborations with criminologists, sociologists, computer scientists, and agencies that specialize in infiltrating and tracking cybercrime (e.g., law enforcement) are beneficial.

Automated Data Cleaning

Automated data cleaning of wildlife trade web data has not been attempted. However, there is potential from computer science subfields, such as machine learning, to help with cleaning messy data (Norouzzadeh et al. 2018; Lamda et al. 2019). Tools relevant to wildlife trade websites are image classification and text classification (e.g., deep learning and natural language processing) (Di Minin et al. 2018; Silge & Robinson 2020), which can potentially use images or text to identify certain attributes of a given listing, such as the species being traded. However, there is a paucity of applications of these tools and fields to web data of the wildlife trade specifically (Xu et al. 2019). Underlying all these machine-learning tools are training sets, which are a representative sample of listings that have been manually classified by a person for the machine-learning algorithm to use (Lamda et al. 2019). The larger the training set, the more likely the machine-learning model will perform well (Norouzzadeh et al. 2018). There will always be the need for manual data cleaning and labeling. One major barrier to successful implementation of automated data cleaning tools for wildlife trade data is the number of species involved in the trade, where research contexts can encompass hundreds to thousands of species and wildlife parts or derivatives (e.g., Humair et al. 2015).

Conclusions

As more of the global human population shifts to using the internet and as ethical and disease concerns of physical markets arise (Roe et al. 2020), the online trade of wildlife is poised to increase. Thus, the internet is, and will continue to be, an invaluable source of data (Lavorgna 2014). Despite the limitations of data collected from the internet, there are vast opportunities to inform conservation, biosecurity, and law enforcement objectives. Current strategies of researchers using small-scale monitoring (i.e., 1 or few websites) should continue to provide insight into specific taxon and product contexts (Sung & Fong 2018). With the development of machine learning tools to clean messy web data, there will be the possibility of creating large-scale (i.e., for many websites) automated systems to detect illegal trade to help inform law enforcement and conservation efforts. Likewise, early risk-screening and rapid-response systems may be possible for invasive species (e.g., Suiter & Sferrazza 2007; Marshall Meyers et al. 2020), especially for exotic pets and ornamental plants whose online trade is commonplace (Lenda et al. 2014; Lockwood et al. 2019). Regardless of the ultimate application, our guide can serve as a primer and starting point to establishing research agendas related to wildlife trade occurring on the internet.

Acknowledgments

We thank T. Wittmann for graphic design of the figures, S. Moncayo for data curation in a previous version of this article, and J. Maher for help with our hypothetical case studies. This work was supported by funding from the Centre for Invasive Species Solutions (PO1-I-002: Understanding and Intervening in Illegal Trade in Non-Native Species).

Supporting Information

Additional information is available online in the Supporting Information section at the end of the online article.

The authors are solely responsible for the content and functionality of these materials. Queries (other than absence of the material) should be directed to the corresponding author.

Literature Cited

Alfino S, Roberts DL. 2018. Code word usage in the online ivory trade across four European Union member states. Oryx **54:**494–498.

Bergman MK. 2001. White paper: the deep web: surfacing hidden value. Journal of Electronic Publishing **7**. https://doi.org/10.3998/3336451.0007.104.

Cassey P, Delean S, Lockwood JL, Sadowski JS, Blackburn TM. 2018. Dissecting the null model for biological invasions: a meta-analysis of the propagule pressure effect. PLOS Biology (e2005987) https://doi.org/10.1371/journal.pbio.2005987.

Chamberlain S, Szocs E. 2013. Taxize: taxonomic search and retrieval in R. F1000Research **2:**191.

Chen H. 2011. Dark web: exploring and data mining the dark side of the web. Springer-Verlag, New York.

CRS (Congressional Research Service). 2017. Dark web. CRS Reports, Washington, D.C. Available from https://crsreports.congress.gov/product/pdf/R/R44101 (accessed May 2020).

Cunliffe J, Décary-Hêtu D, Pollak TA. 2019. Nonmedical prescription psychiatric drug use and the darknet: a cryptomarket analysis. International Journal of Drug Policy **73:**263–272.

Di Minin E, Fink C, Tenkanen H, Hiippala T. 2018. Machine learning for tracking illegal wildlife trade on social media. Nature Ecology & Evolution **2:**406–407.

Eskew EA, White AM, Ross N, Smith KM, Smith KF, Rodríguez JP, Zambrana-Torrelio C, Karesh WB, Daszak P. 2020. United States wildlife and wildlife product imports from 2000–2014. Scientific Data **7:**1–8.

Freitas A, Curry E. 2016. Big data curation. Pages 87–118 in Cavanillas JM, Curry E, Wahlster W, editors. New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe. Springer International Publishing, Cham, Switzerland.

Gallagher RV, et al. 2020. Open science principles for accelerating trait-based science across the tree of life. Nature Ecology & Evolution **4:**294–303.

GBIF (Global Biodiversity Information Facility). 2020. What is GBIF? GBIF, Copenhagen. Available from https://www.gbif.org/what-is-gbif (accessed May 2020).

Harriman S, Patel J. 2014. The ethics and editorial challenges of internet-based research. BMC Medicine **12:**124.

Harrison JR, Roberts DL, Hernandez-Castro J. 2016. Assessing the extent and nature of wildlife trade on the dark web. Conservation Biology **30:**900–904.

Heinrich S, Ross JV, Cassey P. 2019. Of cowboys, fish, and pangolins: US trade in exotic leather. Conservation Science and Practice **1:** e75.

Heinrich S, Toomes A, Gomez L. 2020. Valuable stones: the trade in porcupine bezoars. Global Ecology and Conservation **24:**e01204.

Hinsley A, Lee TE, Harrison JR, Roberts DL. 2016. Estimating the extent and structure of trade in horticultural orchids via social media. Conservation Biology **30:**1038–1047.

Holmberg RJ, Tlusty MF, Futoma E, Kaufman L, Morris JA, Rhyne AL. 2015. The 800-pound grouper in the room: asymptotic body size and invasiveness of marine aquarium fishes. Marine Policy **53:**7–12.

Humair F, Humair L, Kuhn F, Kueffer C. 2015. E-commerce trade in invasive plants. Conservation Biology **29:**1658–1665.

IFAW (International Fund for Animal Welfare). 2018. Disrupt: wildlife cybercrime. IFAW, Yarmouth, Massachusetts. Available from https://d1jyxxz9imt9yb.cloudfront.net/resource/218/attachment/original/IFAW_-_Disrupt_Wildlife_Cybercrime_-_FINAL_English_-_new_logo.pdf (accessed May 2020).

Jensen TJ, Auliya M, Burgess ND, Aust PW, Pertoldi C, Strand J. 2019. Exploring the international trade in African snakes not listed on CITES: highlighting the role of the internet and social media. Biodiversity and Conservation **28:**1–19.

Kikillus KH, Hare KM, Hartley S. 2012. Online trading tools as a method of estimating propagule pressure via the pet-release pathway. Biological Invasions **14:**2657–2664.

Koricheva J, Gurevitch J, Mengersen K. 2013. Handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton, New Jersey.

Ladle RJ, Correia RA, Do Y, Joo G-J, Malhado AC, Proulx R, Roberge J-M, Jepson P. 2016. Conservation culturomics. Frontiers in Ecology and the Environment **14:**269–275.

Lamba A, Cassey P, Segaran RR, Koh LP. 2019. Deep learning for environmental conservation. Current Biology **29:**R977–R982.

Langville AN, Meyer CD. 2011. Google's PageRank and beyond: the science of search engine rankings. Princeton University Press, Princeton, New Jersey.

Lavorgna A. 2014. Wildlife trafficking in the internet age. Crime Science **3:**5.

Lenda M, Skórka P, Knops JMH, Moroń D, Sutherland WJ, Kuszewska K, Woyciechowski M. 2014. Effect of the internet commerce on dispersal modes of invasive alien species. PLOS ONE (e99786) https://doi.org/10.1371/journal.pone.0099786.

Liakos P, Ntoulas A, Labrinidis A, Delis A. 2016. Focused crawling for the hidden web. World Wide Web **19:**605–631.

Lockwood JL, et al. 2019. When pets become pests: the role of the exotic pet trade in producing invasive vertebrate animals. Frontiers in Ecology and the Environment **17:**323–330.

Lyons JA, Natusch DJD. 2013. Effects of consumer preferences for rarity on the harvest of wild populations within a species. Ecological Economics **93:**278–283.

Marshall Meyers N, Reaser JK, Hoff MH. 2020. Instituting a national early detection and rapid response program: needs for building federal risk screening capacity. Biological Invasions **22:**53–65.

McNamara J, Fa JE, Ntiamoa-Baidu Y. 2019. Understanding drivers of urban bushmeat demand in a Ghanaian market. Biological Conservation **239:**108291.

Morgan J, Chng S. 2018. Rising internet-based trade in the critically endangered ploughshare tortoise *Astrochelys yniphora* in Indonesia highlights need for improved enforcement of CITES. Oryx **52:**744–750.

Nijman V. 2020. Illegal trade in Indonesia's national rare animal has moved online. Oryx **54:**12–13.

Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, Clune J. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences of the United States of America **115:**E5716–E5725.

Pew Research Center. 2016. Smartphone ownership and internet usage continues to climb in emerging economies. Pew Research Center, Washington, D.C. Available from https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/ (accessed May 2020).

Regueira RFS, Bernard E. 2012. Wildlife sinks: quantifying the impact of illegal bird trade in street markets in Brazil. Biological Conservation **149:**16–22.

Rhyne AL, Tlusty MF, Schofield PJ, Kaufman L, Morris JA, Bruckner AW. 2012. Revealing the appetite of the marine aquarium fish trade: the volume and biodiversity of fish imported into the United States. PLOS ONE (e35808) https://doi.org/10.1371/journal.pone.0035808.

Roberts DL, Hernandez-Castro J. 2017. Bycatch and illegal wildlife trade on the dark web. Oryx **51:**393–394.

Roe D, Dickman A, Kock R, Milner-Gulland EJ, Rihoy E, 't Sas-Rolfes M. 2020. Beyond banning wildlife trade: COVID-19, conservation and development. World Development **136:**105121.

Rowley JJL, Shepherd CR, Stuart BL, Nguyen TQ, Hoang HD, Cutajar TP, Wogan GOU, Phimmachak S. 2016. Estimating the global trade in Southeast Asian newts. Biological Conservation **199:**96–100.

Sánchez-Mercado A, Cardozo-Urdaneta A, Moran L, Ovalle L, Arvelo MÁ, Morales-Campos J, Coyle B, Braun MJ, Rodríguez-Clark KM. 2020. Social network analysis reveals specialized trade in an endangered songbird. Animal Conservation **23:**132–144.

Setiawan A, Iqbal M, Halim A, Saputra RF, Setiawan D, Yustian I. 2020. First description of an immature Sumatran striped rabbit (*Nesolagus netscheri*), with special reference to the wildlife trade in South Sumatra. Mammalia **84:**250–252.

Silge J, Robinson D. 2017. Text mining with R: A Tidy approach. 1st edition. O'Reilly Media, Beijing; Boston.

Singrodia V, Mitra A, Paul S. 2019. A review on web scrapping and its applications. pp. 1-6, 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India. https://doi.org/10.1109/ICCCI.2019.8821809.

Siriwat P, Nijman V. 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:**454–461.

Sonricker Hansen AL, Li A, Joly D, Mekaru S, Brownstein JS. 2012. Digital surveillance: a novel approach to monitoring the illegal wildlife trade. PLOS ONE (e51156) https://doi.org/10.1371/journal.pone.0051156.

Stringham OC, Lockwood JL. 2018. Pet problems: biological and economic factors that influence the release of alien reptiles and amphibians by pet owners. Journal of Applied Ecology **55:**2632–2640.

Suiter K, Sferrazza S. 2007. Monitoring the sale and trafficking of invasive vertebrate species using automated internet search and surveillance tools. Pages 90–93 in Witmer WG, Pitt WC, Fagerstone KA, editors. Managing vertebrate invasive species: proceedings of an international symposium. USDA/APHIS Wildlife Services, National Wildlife Research Center, Fort Collins, Colorado.

Sula CA. 2016. Research ethics in an age of big data. Bulletin of the Association for Information Science and Technology **42:**17–21.

Sung Y-H, Fong JJ. 2018. Assessing consumer trends and illegal activity by monitoring the online wildlife trade. Biological Conservation **227:**219–225.

't Sas-Rolfes M, Challender DWS, Hinsley A, Veríssimo D, Milner-Gulland EJ. 2019. Illegal wildlife trade: scale, processes, and governance. Annual Review of Environment and Resources **44:**201–228.

Tai MC-T. 2012. Deception and informed consent in social, behavioral, and educational research (SBER). Tzu Chi Medical Journal **24:**218–222.

Toivonen T, Heikinheimo V, Fink C, Hausmann A, Hiippala T, Järv O, Tenkanen H, Di Minin E. 2019. Social media data for conservation science: a methodological overview. Biological Conservation **233:**298–315.

Xu Q, Cai M, Mackey TK. 2020. The illegal wildlife digital market: an analysis of Chinese wildlife marketing and sale on Facebook. Environmental Conservation **47:**206–212.

Xu Q, Li J, Cai M, Mackey TK. 2019. Use of machine learning to detect wildlife product promotion and sales on Twitter. Frontiers in Big Data **2:**28.

Ye Y-C, Yu W-H, Newman C, Buesching CD, Xu Y, Xiao X, Macdonald DW, Zhou Z-M. 2020. Effects of regional economics on the online sale of protected parrots and turtles in China. Conservation Science and Practice **2:**e161.

Zamora A. 2019. Making room for big data: web scraping and an affirmative right to access publicly available information online. Journal of Business, Entrepreneurship and the Law **12:**203–228.

Zimmer M. 2010. "But the data is already public": on the ethics of research in Facebook. Ethics and Information Technology **12:**313–325.