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], Phillip Cassey[1]

1. School Biological Sciences, University of Adelaide, SA 5005, Australia
2. School of Mathematical Sciences, University of Adelaide, SA 5005, Australia
3. Fenner School of Environment and Society, The Australian National University, Canberra, ACT 2601, Australia

Corresponding author: Oliver Stringham, oliverstringham@gmail.com

Keywords: big data, e-commerce; Internet; pet trade; social media; web-scraping

Abstract

The unrivalled 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, frequency) is a relatively recent and rapidly developing approach, which currently lacks an accessible framework for locating relevant websites and collecting data. Here, we present an accessible guide for Internet-based wildlife trade surveillance, which uses a repeatable and systematic method to automate data collection from relevant websites. Our guide is adaptable to the multitude of trade-based contexts including different focal taxa or derived parts, and locations of interest. We provide information for working with the diversity of websites that trade wildlife, including social media platforms. Finally, we discuss the advantages and limitations of web data, including the challenges presented by trade occurring on clandestine sections of the Internet (e.g., deep and dark web).

Background

Introduction

The wildlife trade is an influential driver of species endangerment; as well as spreading invasive species and diseases, and provisioning 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) is vital for informing decisions about conservation, biosecurity, and law enforcement; and developing human behavior change campaigns. Yet this data is rarely collected, or is difficult to obtain (Regueira & Bernard 2012; Eskew et al. 2020). In recent years, the Internet has played an increasingly important role in facilitating the wildlife trade (Siriwat & Nijman 2020).

To date, 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 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 classifieds 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 methodology, 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 parts and commodities.

Here, we present an accessible guide to using the Internet to gather data on the wildlife trade. We developed the methodology through our collective knowledge of working with web data and the wildlife trade, combined with the methods used in prior published studies. Our goal is for this guide to be used by scientists, NGOs, government agencies, and other interested parties, who wish to utilize the Internet as a source of data on the wildlife trade.

Structure of the Internet

The Internet (i.e., the World Wide Web or simply the Web) is categorized into three distinct ‘layers’: the surface web, the deep web, and the dark web (Figure 1; Bergman 2001). Each layer differs in one of two factors: whether it’s 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), while 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 will be influenced by how 'findable' a website/content is (e.g., can a search engine find it). Further, the ethical considerations of collecting data will in part depend on how 'accessible' the website is (e.g., is deceit or limited disclosure required to gain access to the content; Section 4.1).

What data is available?

Data availability on wildlife trade varies by website and even within a website (Appendix S1; Toivonen et al. 2019). On a basic level, online advertisements (i.e., listings or posts) are provided in the form of text, pictures and/or 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 or wild caught/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.) along with 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 around 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

We specify our guide in six steps (Figure 2): (1) defining the scope and purpose of the project; (2) finding candidate websites; (3) selecting target websites to monitor; (4) collecting and storing data from websites; (5) cleaning data; and (6) analysis. Here, we detail the process of each step leading up to analysis. We generalize this guide for websites found in any layer of the Internet (including social media) and discuss how to adapt this guide to different languages and countries. Further, in Figure 3, we provide two hypothetical case studies to accompany and contextualize each step of the guide. For more generalized frameworks on working with social media and online news outlets, refer to Toivonen et al. (2019) and Sonricker Hansen et al. (2012), respectively.

1. Defining the scope and purpose of the project

At a minimum, it is essential to decide which species, taxa, and/or derived products are of interest, the location(s) of interest, and the timeframe for data collection (i.e., one-time snapshot, versus ongoing monitoring for months to years). Further, considering what type of website (Appendix S1) or layer of the Internet may be appropriate. On a practical note, the research questions that can be answered will be 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 US (Heinrich et al. 2019); exploring the social network structure of sellers of horticultural orchids (Hinsley et al. 2016).

2. Finding candidate websites where specific taxa and wildlife products are traded

Here, we detail a method to find candidate websites (e.g., e-commerce sites, forums) by using search engines. Finding relevant social media content requires special considerations, which we detail in Section 2.4. Outside of the search engines, other approaches to finding candidate websites and choosing target websites (Section 3) 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 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 point in 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.

2.1 Surface Web

For the surface web, finding candidate websites involves three steps: (1) defining keyword phrases to search; (2) using a search engine to perform searches; and (3) classifying the relevance of each search result. This part of the methodology is akin to the process of finding relevant scientific papers in a systematic review or meta-analysis (i.e., PRISMA methodology: 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 (Section 3).

2.1.1 Defining search phrases

Search phrases are composed of 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" (See Appendix S2 for a detailed example). 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/regional names, or names of breeds, morphs, and mutations (e.g., Lyons and Natusch 2013), which are not captured in the initial formulation of search phrases.

2.1.2 Using search engines to perform searches

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., the number of page views), and, increasingly, the location of where the search occurs (Langville and Meyer 2011). The results from a search engine are expected to change at any point in time for a number of reasons including: changes to the search engine algorithm, changes to website popularity metrics, the emergence of new websites, and 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 VPN (virtual private network) may alleviate location issues. For more information on search engines, see Appendix S3.

2.2 Deep web and dark web (non-social media)

Websites on the deep web indexed on search engines will be findable using the same approach outlined for the surface web (e.g., private forums). Currently, there are no generalizable or automated methods for locating deep web content or websites not indexed by search engines (e.g. WhatsApp, WeChat, other private messaging apps) nor dark web content, outside of expert consultation or interviewing communities of practice. 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.

2.3 Classifying search engine results

After obtaining URLs from search results, each will need to be categorized as relevant or irrelevant. Relevance is subjective and we recommend defining inclusion/exclusion criteria depending on the scope and purpose of the study. One obvious inclusion criterion is whether the target taxa 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) and there are less direct websites that facilitate the transaction of selling wildlife online, but leave it up to the individuals in the transaction to conduct the exchange (e.g., classifieds: Sung & Fong 2019).

2.4 Social media

2.4.1 Types of social media content

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

2.4.2 Finding social media 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; Section 2.1) for internal search engines to find relevant social media 'groups' (e.g., Siriwat & Nijmans 2020). These 'groups' can then be classified by their relevance (Section 2.3) and considered for monitoring (Section 3). 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 (Section 2.3; Section 3). Importantly, many social media users utilize hashtags (denoted by the number sign: #), 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).

APIs (Application Programming Interfaces) may be available for some social media websites, which may allow for 'bulk' searches (i.e., more than one search at once) and streamlined data collection (Section 4). 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 pre-selected group of users can view their posts. Importantly, content with privacy restrictions may be hidden from the internal search engine or API results.

3. Selecting target sites to monitor

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 (see Appendix S4 for more information). 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 will vary based on the project aim(s) and the resources available to collect and clean data (Sections 4 and 5). Again, expert opinion and communities of practice can provide opinions on what websites are most relevant.

4. Collecting data from websites

Data collection can occur in one of two main ways: manual or automated. Manual data collection involves visiting the website and recording what taxa/product is being traded along with desired associated attributes (e.g., price, location). Automated data collection involves constructing “web scrapers” to visit the website and extract desired relevant information (Figure 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). Since 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). Choosing manual or automated data collection will depend on how long and how often data is being collected, as 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 deep web (including social media) and dark web, as long as researchers have access to the website/content (e.g. Cunliffe et al. 2019).

4.1 Ethical considerations of collecting data from the Internet

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, ethics approval for collecting data from any deep or dark web sites will include obtaining a login or approval to join (Tai et al. 2012); since deceit or limited disclosure of research aims may be required. Also, automated data collection processes (i.e., web scraping) currently encompasses a legal gray area (Zamora 2019), and thus we encourage researchers to acquire ethics approval prior to using them. For specific recommendations of ethical practice, refer to Appendix S6.

5. Data Cleaning

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 possibly be the most time-consuming part of the entire project. The amount of cleaning required will depend on the structure of the website and will vary by individual website (Appendix S1). For instance, a website may have a separate field for species names, while another may just have one 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.

5.1 Resolving species names

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 while 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 as possible. If pictures are provided, taxonomic experts can aid in species identification. Yet, pictures may be too poor in quality to properly identify species. In some instances, online traders may simply not provide enough information in the listing to identify to species level.

If monitoring many species, we recommend relating the species/taxa name to a taxonomic database (e.g., GBIF 2020). Doing so will facilitate conformation to taxonomic names by avoiding synonyms and misspellings (Gallagher et al. 2020). In addition, it will enable 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 to surveying physical markets or stores, especially if using 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 nor is observable online (e.g., bushmeat trade; McNamara et al 2019). The degree to which trade occurs online will depend 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 by target/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 impactful when considering applied outcomes (e.g., Rowley et al. 2016).

Considerations for the Deep and Dark Web

Currently, wildlife trade on the surface web and indexed deep web (e.g., social media) is extremely abundant (Sung & Fong 2018; Xu et al. 2020; IFAW 2018). The unindexed deep web, such as private text messaging apps (e.g., WhatsApp; Facebook Messenger), has remained relatively unexplored until recently (e.g., Sanchez-Mercado et al. 2020; Setiawan et al. 2019), 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 (Roberts & Hernandez-Castro 2017). Further, future policies enacted in response to conservation, and the criminological and welfare 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 (Section 4.1) 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) will be 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 (Lamda et al. 2019; Norouzzadeh et al. 2020). 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/fields to web data of the wildlife trade specifically (Xu et al. 2019). Importantly, underlying all of these machine-learning tools are training sets, which are a representative sample of listings that have been manually classified by a human(s) 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 better (Norouzzadeh et al. 2020). Importantly, there will always be the need for human data cleaning and labelling. 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/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., one or few websites) should continue to provide insight into specific taxa/products 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., Marshall Meyers 2020; Suiter & Sferrazza 2007), especially for 'exotic' animals and ornamental plants whose online trade is commonplace (Lockwood et al. 2019; Lenda et al. 2014). 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.

References

Alfino S, Roberts DL. 2018. Code word usage in the online ivory trade across four European Union member states. Oryx:1–5. Cambridge University Press.

Bergman MK. 2001. White Paper: The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing **7**. Available from http://hdl.handle.net/2027/spo.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 **16**:e2005987. Public Library of Science.

Chamberlain S, Szocs E. 2013. “taxize - taxonomic search and retrieval in R.” F1000Research. http://f1000research.com/articles/2-191/v2.

Chen H. 2011. Dark Web: Exploring and Data Mining the Dark Side of the Web. Springer Science & Business Media.

Congressional Research Service, March 10, 2017 , “Dark Web”. Available on 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. Nature Publishing Group.

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 J. M. Cavanillas, E. Curry, and W. Wahlster, editors. New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big

Data in Europe. Springer International Publishing, Cham. Available from https://doi.org/10.1007/978-3-319-21569-3_6 (accessed May 13, 2020).

Gallagher RV et al. 2020. Open Science principles for accelerating trait-based science across the Tree of Life. Nature Ecology & Evolution **4**:294–303. Nature Publishing Group.

GBIF: The Global Biodiversity Information Facility (2020) "What is GBIF?". 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: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.

International Fund for Animal Welfare. 2018. "Disrupt: Wildlife Cybercrime". Available on 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.

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.

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 **9**:e99786. Public Library of Science.

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. Cambridge University Press.

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 **115**:E5716–E5725. National Academy of Sciences.

Pew Research Center, February, 2016, "Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies". Available on 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 **7**. Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3357433/ (accessed April 2, 2020).

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

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. 2019. First description of an immature Sumatran striped rabbit (Nesolagus netscheri), with special reference to the wildlife trade in South Sumatra. Mammalia **1**. De Gruyter. Available from https://www.degruyter.com/view/journals/mamm/ahead-of-print/article-10.1515-mammalia-2018-0217/article-10.1515-mammalia-2018-0217.xml (accessed April 3, 2020).

Silge J, Robinson D. (n.d.). Text Mining with R. Available from https://www.tidytextmining.com/ (accessed May 14, 2020).

Singrodia V, Mitra A, Paul S. 2019. A Review on Web Scrapping and its Applications. Pages 1–6 2019 International Conference on Computer Communication and Informatics (ICCCI).

Siriwat P, Nijman V. 2018. Illegal pet trade on social media as an emerging impediment to the conservation of Asian otters species. Journal of Asia-Pacific Biodiversity **11**:469–475.

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. Available from http://www.sciencedirect.com/science/article/pii/S2287884X2030042X (accessed April 24, 2020).

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 **7**. Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3517447/ (accessed February 14, 2020).

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:5.

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:1–7. Cambridge University Press.

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**. Available from https://www.frontiersin.org/articles/10.3389/fdata.2019.00028/full (accessed February 14, 2020).

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 **n/a**: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.

Figures

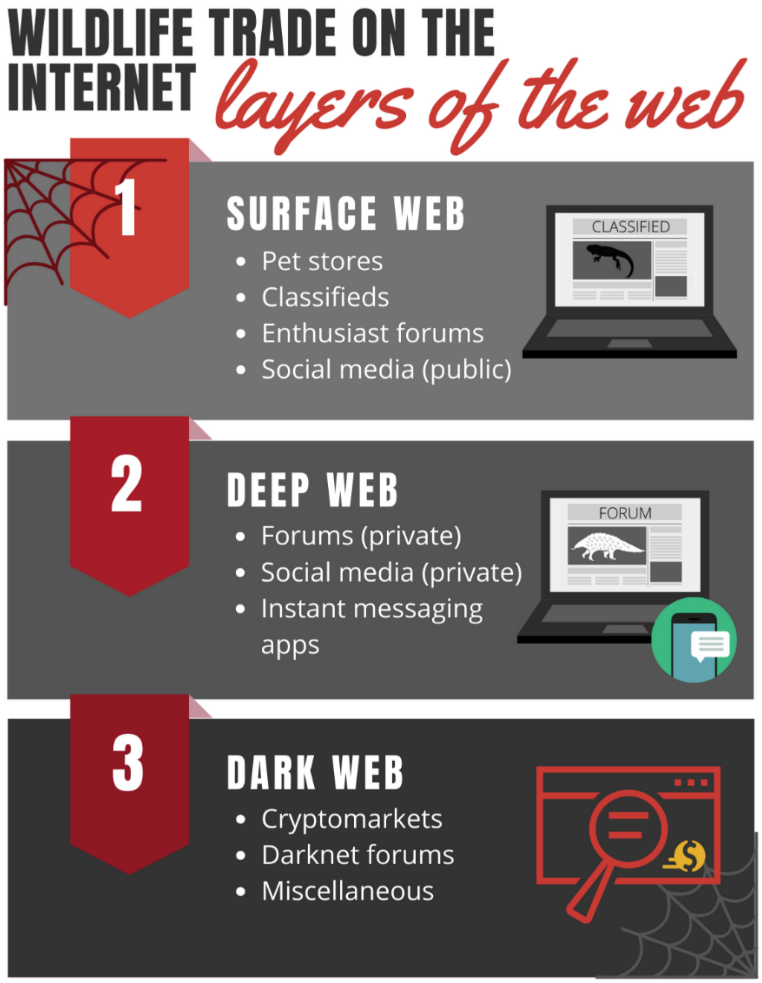

Figure 1.

Where wildlife trade occurs on the Internet. Within the Internet, there are three 'layers' of where websites can exist: the surface web, the deep web, and the dark web. As wildlife trade moves to websites on the deep and dark web, it becomes increasingly obfuscated (denoted by darkening gray background), making it more difficult for researchers to detect and monitor (Appendix S7).

Figure 2.

Flowchart of our guide to using the Internet to monitor and quantify the wildlife trade. Each number corresponds to a step (and subheading title) described in text. Adjusting our guide to a specific taxa/product, language, or location will occur in step 2 (Section 2), where search phrases will be tailored to a specific context (and language) and search engines will be restricted to a particular country. Search and selection of website/content will vary if exploring social media and other deep web content (Section 2.4). Collecting data is a similar process for all websites regardless of where on the web the website exists (Section 4; Figure 1). Cleaning data involves processing the collected data so that is can be analyzed (e.g., verifying the species traded). Machine learning and natural language processing tools have the potential to help speed up the data cleaning process.

Figure 3.

Two hypothetical case studies using the Internet to study wildlife trade. The first (left column) explored the trade of alien ornamental plants found in online plant shops/nurseries in Australia (i.e., open web; sensu Lenda et al. 2014). For this study, keywords were generated based on the alien plant species of interest (along with their scientific and trade names). In addition, qualifiers such as "for sale" or "store" were added to the keywords to create search phrases for the search engines (See appendix S2 for details). Next, the search engines (Google and DuckDuckGo) provided a list of candidate websites, from which a subset were chosen based on inclusion criteria (Section 3) for data collection (designated by green checkmark). For this hypothetical, the inclusion criteria were: (1) if the store sold one or more of the species of interest; and (2) if the store offered to ship plants or seeds interstate. Next, web scrapers were constructed for each website and data collected on a bi-weekly basis for one year (indicated by green check marks on calendar). Finally, data cleaning was undertaken. Since data collected from individual stores/shops tend to be more organized than other types of websites (Appendix S1) data cleaning was less intensive compared to the next case study (denoted by one hourglass icon). Further, since this study explored many species, linking each traded taxa to a taxonomic database (e.g., Global Biodiversity Information Facility: GBIF) facilitated data analysis at the species level (Section 5.1). The second hypothetical case study (right column) explored the trade of exotic leather boots made from pangolin skins occurring on social media in the United States (sensu Heinrich et al. 2019). Preliminary investigation revealed several hashtags were used in sale of pangolin-leather boots. These hashtags were supplied to the internal search engines of the social media sites (Facebook and Instagram). For this study, all posts returned from the search engine become the data itself (i.e., unconsolidated social media content; Section 2.4.2). Data collection occurred every other day since social media content tends to be updated frequently. Finally, data cleaning took longer for this study because: (i) more listings were collected compared to stores, and (ii) the listings were 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). GBIF logo credits: https://www.gbif.org.

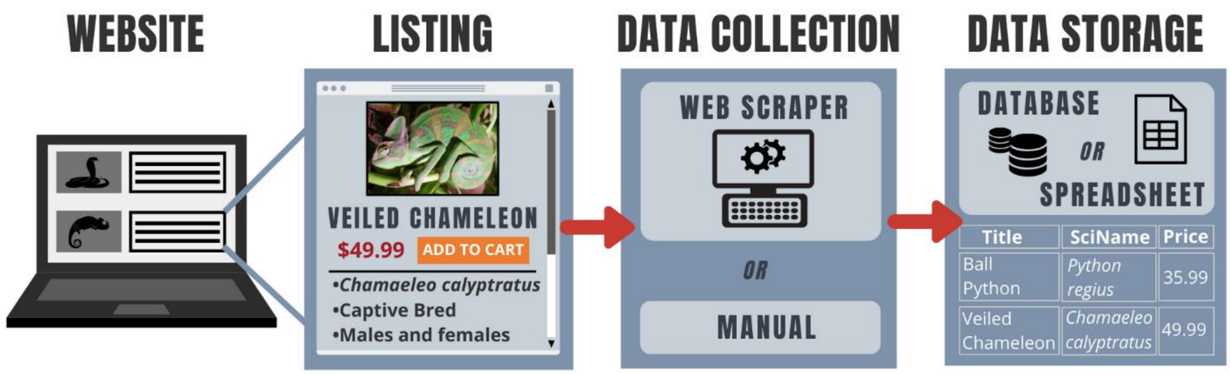

Figure 4.

Data collection and data storage procedure for websites trading wildlife. Websites have underlying HTML code that web scrapers can parse to extract relevant information, which can then be stored in a database or spreadsheet. This process can be repeated for different websites using custom web scraper code (see Appendix S5 for more information). The frequency with which to collect data will depend on the nature of the website, including how often the website is updated. If data collection is to occur frequently, we recommend using automated data collection because manual data collection is laborious and time consuming. However, there is a trade-off between the resources invested in creating web scrapers and the quantity of data that will be collected. Chameleon photo credits: Chris Kade.

Supporting Information

Table of Contents

Appendix S1: Types of websites relevant to wildlife trade research

We categorize eight types of websites/platforms, which can be relevant to the wildlife trade, and we provide select examples (with references) that use each type of website as a data source.

1. **Shops/stores** are often physical storefronts that have websites where they list the species or wildlife products they are selling (e.g., pet stores, ornamental plant nurseries, medicinal shops). Sometimes stores will specify whether they can ship, and to where. Other times the online stores do not have a physical storefront and are exclusively an online store that ships directly to consumers. Pet stores and ornamental plant shops reliably provide either scientific names and/or common names of the species they are selling and their price. In our experience, most store websites are not updated daily and therefore collecting data weekly or fortnightly can be appropriate.

   <u>Select references</u>

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.

Nelufule T, Robertson MP, Wilson JRU, Faulkner KT, Sole C, Kumschick S. 2020. The threats posed by the pet trade in alien terrestrial invertebrates in South Africa. Journal for Nature Conservation:125831.

2. **Classifieds,** including **e-commerce** websites, are websites where individuals or companies can post their animal/wildlife/products they wish to trade. They usually appear on screen in reverse chronological order where the most recent listings appear first. Some classified websites are exclusive to particular taxa (e.g., only reptiles), while others have separate categories for multiple taxa (e.g., a bird section and a reptile section), and other sites may not have a distinct category for any taxa. Classified listings often contain some form of the taxa or product name:

scientific, common, trade names. However, this will vary by website, by taxa, and by individual traders. Most classified websites remove listings once they are "sold". Price is usually provided by the user and therefore a distribution of prices for a given species or products can be derived. The location of the sale is usually given as well. For popular classifieds, data collection will likely be daily or every two to three days. Some classified websites make listings 'inaccessible' once the seller finds a buyer. Therefore, for these types of websites, it's important to collect data more frequently in order to capture listings before they are removed. Further, most dark web marketplaces (i.e., crypto-markets) function nearly identically to surface web classifieds/e-commerce websites. In the US, a popular open-web classifieds website is craigslist (https://www.craigslist.org) and in Australia there is Gumtree (https://www.gumtree.com.au/).

Select references

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

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 n/a:e161.

3. **Forums** are specialist websites where enthusiasts discuss various aspects of the taxa of interest. Many forums have a dedicated marketplace subforum where trading occurs. The marketplace subforums are structurally similar to classified websites. One key difference is that users can comment below the initial post asking clarifying questions. From these questions it may be possible to determine if the transaction/sale took place. Another difference is that users of forums do not typically remove "sold" listings. Either the common, scientific, or trade name is provided. The location and price are usually provided. Most forums keep an archive of all posts

and don’t remove old posts, therefore regular data collection efforts are less essential compared to classifieds.

Select reference

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

4. **Lost and Found** websites allow users to report a lost or found pet. They are structurally similar to classifieds websites. They may provide useful information if exploring invasive species risks. They are usually only available for highly visible species such as domesticated mammals (cats, dogs, rabbits), turtles, and birds. The species name (scientific, common, or trade name) as well as the location and date is usually provided.

Select references

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.

Vall-llosera M, Cassey P. 2017. Leaky doors: Private captivity as a prominent source of bird introductions in Australia. PLOS ONE 12:e0172851.

5. **Adoption** websites post pet animals that are available for adoption. This is considered the secondary market for pets. They are structurally similar to classifieds websites.

6. **News** websites contain news from either print or electronic news companies. For the wildlife trade, many seizures of illegal wildlife are often reported in the news and may be used as a source of data.

Select references

Indraswari K, Friedman RS, Noske R, Shepherd CR, Biggs D, Susilawati C, Wilson C. 2020. It's in the news: Characterising Indonesia's wild bird trade network from media-reported seizure incidents. Biological Conservation 243:108431.

TRAFFIC International (2020) Wildlife Trade Portal. Available at www.wildlifetradeportal.org.

7. **Social media** websites vary drastically in their structure and content relating to the wildlife trade. Broadly, content on social media websites can be separated into: (i) 'groups' with a particular purpose where people can join and (ii) users that post to the social media platform at large, or to a group of followers (Main text, Section 2.4). Some 'groups' focus on trading particular taxa or products. The posts are similar in structure to forums, except with usually less organization. Some 'groups' are open to the public (e.g., users with a login to the social media platform can view) and others require an invitation or approval to join (e.g., 'private' groups). In addition, individual stores, breeders, or traders may maintain social media accounts where they advertise wildlife. Data collection frequency will be similar to that of classifieds. Social media websites are among the most popularly used websites. Examples of social media websites investigated for wildlife trade in the primary literature include Facebook, Twitter, Instagram, and YouTube.

<u>Select references</u>

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.

Kitson, H., & Nekaris, K. A. I. (2017). Instagram-fuelled illegal slow loris trade uncovered in Marmaris, Turkey. Oryx, 51(3), 394.

Measey J, Basson A, Rebelo AD, Nunes AL, Vimercati G, Louw M, Mohanty NP. 2019. Why Have a Pet Amphibian? Insights From YouTube. Frontiers in Ecology and Evolution 7. Frontiers. Available from https://www.frontiersin.org/articles/10.3389/fevo.2019.00052/full

Van TP, Luu VQ, Tien TV, Leprince B, Khanh LTT, Luiselli L. 2019. Longitudinal monitoring of turtle trade through Facebook in Vietnam. The Herpetological Journal 29:48–56.

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. Available from https://www.frontiersin.org/articles/10.3389/fdata.2019.00028/full (accessed February 14, 2020).

8. **Private messaging apps** including WhatsApp and Facebook messenger (among others) are dedicated apps/platform for instant messaging between two or more people. Search engines do not index private messaging apps, so it is not possible to find individual chats using the Internet. The type of information provided about traded taxa in private messaging is likely similar to classifieds and forums. Once access is granted to a private messaging group, exporting a 'log' of the entire chat is a commonly available feature, thus negating the need for web scrapers.

Select references

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.

Stoner, S & Shepherd, C. 2020. Using intelligence to tackle the criminal elements of the illegal trade in Indian Star Tortoises Geochelone elegans in Asia. Global Ecology and Conservation. 23. e01097. 10.1016/j.gecco.2020.e01097.

Appendix S2: Table of search phrases from an example study

We provide a table of keywords used to generate search phrases for an example study quantifying the exotic pet trade in three countries (United States, United Kingdom, and Australia). The “taxa” column refers to the taxa of interest; the “location” refers to our target locations, and “website type” refers to the website types of interest. We obtained the search phrases by performing all combinations of “taxa”, “location”, and “website type”, using the follow search phrase templates:

1. Buy {taxa} {location}
2. {taxa} for sale OR purchase {location}
3. {taxa} {website type} {location}

Taxa	**Location**	**Website Type**
freshwater aquarium fish	United States	Forum
marine aquarium fish	United Kingdom	Store
pet birds	Australia	Breeder
exotic pet reptiles		Adoption
exotic pet amphibians		Classifieds

Appendix S3: Further information on search engines

Application Programming Interfaces (APIs)

Certain search engines offer APIs, which can automate the search process by iterating over each search phrase using computer programming (e.g., Bing: Thelwall and Sud 2012). Currently, Microsoft's search engine, Bing, offers an API (https://azure.microsoft.com/en-au/services/cognitive-services/bing-web-search-api/) while Google does not.

Some websites do not appear in search engines

Some surface websites can opt out of appearing on search engines (Carl Drott et al. 2002), so if a website is known to be important, but does not appear in the search engine results it may still be worth considering it as a candidate website. Checking a website's "robots.txt" file will reveal if they have opted out of appearing on search engines (https://technicalseo.com/tools/robots-txt/).

Choosing a cutoff point

Search engines return millions of URLs per search. Thus, choosing a cutoff point to stop recording resultant URLs is important to optimize search effort. While there can be various methods of choosing a cutoff point, it is important that the chosen method is transparent and repeatable. One semi-quantitative method to decide a cutoff point can be to explore the cumulative proportion of relevant results as a function of cutoff point. The point at which the curve flattens, or begins to flatten, can be considered an optimal cutoff point.

References

Carl Drott M. 2002. Indexing aids at corporate websites: the use of robots.txt and META tags. Information Processing & Management 38:209–219.

Thelwall M, Sud P. 2012. Webometric research with the Bing Search API 2.0. Journal of Informetrics 6:44–52.

Appendix S4: Further information on web traffic statistics

Many websites have web traffic statistics (i.e., metadata) that have been recorded by third party companies. For a given website, these traffic statistics can include: the number of page views per month, the rank/popularity, the country where the website is most popular, and more. One provider of website metadata is Amazon Alexa Web Information Services (https://www.alexa.com/siteinfo), which also has an API (https://aws.amazon.com/marketplace/pp/B07Q71HJ3H?ref_=srh_res_product_title). There are a couple of caveats to using web traffic statistics. First is that traffic statistics are calculated for the entire website (i.e., website domain). If the website's only purpose is to trade the target taxa, then this will not be an issue (i.e., online pet store). However, for many websites, there are other reasons people visit the website than to trade the target taxa. For example, the web traffic statistics for eBay, a popular American e-commerce marketplace, would pertain to all trade on eBay and would therefore be unrepresentative of the specific trade. This makes it difficult to compare traffic statistics between websites. In addition, it's important to note that web traffic statistics are not available for all websites. Given these caveats, we recommend using web traffic data as only one line of evidence in choosing a target website.

Appendix S5: Detailed information on web scrapers and data storage

Background on web scrapers

Web scrapers are computer code that convert unstructured web data into a structured data format (i.e., tabular data format; Singrodia et al. 2019). Coding web scrapers involves technical expertise (Mitchell 2018). Outside of learning to code their own web scrapers, researchers may hire data scientists or contractors to code web scrapers. There are several open-source programming languages that can be used to code web scrapers. Some examples include the language Python with libraries bs4 (https://www.crummy.com/software/BeautifulSoup/), requests (Chandra & Varanasi 2015), and selenium (https://selenium-python.readthedocs.io/). Web scraping is possible in other programming languages including R with the packages RSelenium (Harrison 2020) or rvest (Wickham 2019). In addition, there are “no code” web scrapers, which is “point and click” software that facilitates building of web scrapers without knowledge of programming (de S Sirisuriya 2015). Since web scrapers rely on the underlying HTML of a website, if a website changes its HTML structure (i.e., an update in the website layout), the web scraper may ‘break’ and will need to be updated. There must be a separate custom web scraper coded for each target website (Mitchell 2018; Holmberg et al 2015). In addition to tabular text data, web scrapers can also be programmed to download images.

Web scrapers can cause ‘harm’ to the targeted website because they take up bandwidth on the website’s server (Zamora 2019). Care should be taken not to overwhelm the targeted website with the web scraper by spacing out visits to the website (i.e., a few seconds between navigating pages). Some websites specify the amount of time to ‘wait’ in between visits in their “robots.txt” file (called crawl-delay). Spacing out visits is especially important in web scraper development. Some websites may have an auto block feature, where they will block an IP address if too many visits occur in a short amount of time.

Running web scrapers takes computing resources, however, most modern computers can handle running several web scrapers simultaneously without issues. Alternatively, setting up web scrapers to run on a cloud server or a separate dedicated computer may be desirable. If the data collection is recurrent, then establishing a system to schedule web scrapers to run at regular intervals is possible through built-in software available on all popular computer operating systems (Windows: Task Scheduler, Mac/Linux: cron).

Data storage

Data collected by web scrapers must be stored in a way that is retrievable for cleaning and subsequent analysis. Data storage can be achieved by using spreadsheets or databases (i.e., Database Management Systems such as MySQL). The choice is dependent on the researcher's familiarity with either, and the frequency or total number of data collection events to be stored. Regardless of the data storage technique, since the fields or columns will likely differ between websites, the researcher will need to organize and collate data for each website separately.

Duplicated listings

Determining and marking duplicated listings is an important post data-collection step. Detecting duplicates can be achieved by selecting a column(s) to search for duplicates. If more than one row contains the exact value for the selected column(s) then it can be labelled as a duplicate. For instance, for a pet store, a researcher may decide that if two or more listings share the exact title and exact text description, they are duplicates. Other rules/assumptions can be made depending on the specific website. Labelling unique listings with a unique identifier can help to integrate the raw data with the data cleaning.

References

Chandra, RV, Varanasi, BS. 2015. Python requests essentials. Packt Publishing Ltd.

Harrison J. 2020. RSelenium: R Bindings for 'Selenium WebDriver'. R package version 1.7.7. https://CRAN.R-project.org/package=RSelenium

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.

Mitchell R. 2018. Web Scraping with Python: Collecting More Data from the Modern Web. O'Reilly Media, Inc.

De S Sirisuriya SCM. 2015. A Comparative Study on Web Scraping. Available from http://ir.kdu.ac.lk/handle/345/1051 (accessed May 13, 2020).

Wickham H. 2019. rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.5. https://CRAN.R-project.org/package=rvest

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.

Appendix S6: Recommendations for ethical practice of wildlife e-commerce surveillance

Minimizing harm to research participants (i.e., Internet users) is a key ethical consideration when conducting online surveillance (Buchanan 2010). Additionally, it may also be important to protect the identity of researchers, as wildlife trade activity may at times involve criminal behavior (e.g., Décary-Hétu & Aldridge 2015). Thus, we recommend researchers follow the following ethical practices:

- **Compliance with legislation and acquiring ethics approval.** Approval from relevant ethics committees should always be obtained prior to conducting research and projects should be planned with consideration of relevant legislation (e.g., Australian Government 2020).
- **De-identification of identifiable/re-identifiable data.** Some data collected by researchers can be used to identify an individual person (i.e., identifiable data). This data will require de-identification in order to maintain anonymity of the participants and thus minimize any potential harm to them. Types of data that require de-identification include (but are not limited to): names of participants, user names of participants, age of participants, and locations of online posts (e.g., street address or even postcode). The ease with which data can be de-identified will be largely dependent on the structure of a given website. For example, some sites may have a specific field for usernames, which will be straightforward to de-identify after data collection. In other instances, identifiable information may be present in an unstructured manner in free-form text boxes. This presents a much greater logistical challenge, especially if researchers are collecting large quantities of data that cannot feasibly be manually processed. Finally, in some instances, it is worth considering de-identification of the name of the website(s), group(s), or platform(s) where researchers are collecting data from (e.g., Hinsley et al. 2016). Doing so will preserve researcher anonymity and may prevent future behavioral changes of participants (e.g., switching to trading on a more secure website: Appendix S7).

- **Secure storage of research data.** Data should be stored in a secure manner with transparency around which researchers, institutions or third parties will be granted access to data with different levels of de-identification (Samuel and Buchanan 2020). It is important to consider optimal ways to store 'big data' while maintaining security (e.g., cloud-based storage; Buchanan and Ess 2016). Consideration of data storage should also extend beyond the anticipated lifespan of any given research project.
- **Consideration of limited disclosure of research.** Whether or not researchers can ethically justify the decision to withhold or partially withhold disclosure of their identity or research aims and methodology is highly context dependent. Given the number and potential anonymity of website users whose wildlife trade activity may be monitored, it would be logistically unfeasible to contact everyone in a manner that preserves their identity. However, the same cannot necessarily be said for contacting website administrators or subsite moderators. In fact, for deep web platforms, contact may be necessary in order to gain access to a particular subsection of a website (such as a subforum). It should be noted that such contact may induce changes in wildlife trade behavior that would reduce the value of surveillance research (e.g., individuals who would otherwise participate in illegal trade may choose to do so in an undetectable manner if they are aware that research is taking place). Therefore, the importance of outcomes likely to result from research must be weighed against the importance of gaining consent from research participants.

Australian Government, 2020. National Statement On Ethical Conduct In Human Research. Available at: http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf [Accessed 09 November 2020].

Buchanan, E., 2010, "Internet Research Ethics: Past, Present, Future," in The Blackwell Handbook of Internet Studies, C. Ess and M. Consalvo, (eds.), Oxford: Oxford University Press.

Buchanan, E. and C. Ess. 2016. “Ethics in Digital Research,” in The Handbook of Social Practices and Digital Everyday Worlds, M. Nolden., G. Rebane, M. Schreiter (eds.), Springer.

Décary-Hétu, D, Aldridge, J. 2015. Sifting through the net: Monitoring of online offenders by researchers. European Review of Organised Crime 2, 122-141.

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.

Samuel, G, Buchanan, E, 2020. Guest Editorial: Ethical Issues in Social Media Research. SAGE Publications Sage CA: Los Angeles, CA.

Appendix S7: Consequences of regulations

As more concerns around the wildlife trade emerge (e.g., ethical, disease risk, etc.), governments may impose stricter laws around the online trade of wildlife and/or companies may impose stricter self-regulations of their own websites (Roe et al. 2020). Previous research on trade bans suggest that stricter regulations can have unintended consequences, such as increasing or redirecting trade instead of eliminating it (Challender et al. 2015). Thus, we hypothesize that as regulations around wildlife trade become stricter for the open and indexed deep web, traders will likely move to the either the unindexed deep web (e.g., private messaging apps) or potentially the dark web to avoid detection. One recent incident highlights this possibility. Facebook recently implemented a ban on the sale of animals (live and derived parts) on its website (https://www.facebook.com/policies/commerce/prohibited_content/animals). The efficacy of the ban in reducing trade on Facebook has not been evaluated; however, trade did not stop on Facebook because of the ban (Nijman 2020). Instead, users adjusted how they advertised wildlife, presumably to avoid detection. Users started using code names or acronyms for species and stopped including asking prices (Nijman 2020; author's personal observations). Further, users directed all questions about the advertisement to a private chat (Facebook messenger: Nijman 2020; author's personal observations). In this case, the result of stricter regulations served to decrease the amount of available information for researchers. We note that while stricter regulations may decrease information available to researchers, regulations may have the intended consequence of reducing trade overall.

References

Challender DWS, Harrop SR, MacMillan DC. 2015. Towards informed and multi-faceted wildlife trade interventions. Global Ecology and Conservation 3:129–148.

Nijman V. 2020. Illegal trade in Indonesia's National Rare Animal has moved online. Oryx 54:12–13.

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.