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

Running Head (40 characters): Wildlife trade data from the Internet

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

Word count (Abstract through Literature Cited): 6008

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 efforts. 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 finding relevant websites and collecting data. Here, we present an accessible guide for internet-based wildlife trade surveillance, which uses a systematic method to automate data collection from relevant websites. The guide is easily adaptable to the multitude of trade-based contexts including different focal taxa or derived parts, and locations of interest. Furthermore, as wildlife trade on the Internet becomes more widespread, the ability to collect large amounts of data on traded wildlife will become possible and desirable. Using a case study where we monitor 53 websites, we demonstrate the capabilities and limitations of this kind of large-scale surveillance system. We collected over half a million unique listings in a year and estimate that it would take over two years for one person to clean every listing. We propose that the development of machine learning methods for automation of data collection and processing become a priority and be tested for a variety of different contexts of wildlife trade-related web data.

Introduction

The wildlife trade is an influential driver of species endangerment, source of invasive species, spread of diseases, and criminal activity ('t Sas-Rolfes et al. 2019). Trade occurs across a variety of physical settings, including 'brick and mortar' stores, wet markets, pet stores – and increasingly on the Internet (Siriwat & Nijman 2020). Reliable data on the quantity and composition of the wildlife trade (legal and illegal) is needed to inform decisions about conservation, biosecurity and law enforcement efforts, and develop human behavior change campaigns; yet this data is often not collected and/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). Accordingly, recent efforts to describe and quantify the wildlife trade have turned to the Internet (e.g., Alfino & Roberts 2018). The Internet itself (i.e., the World Wide Web or simply the Web) is categorized by three distinct "layers": the surface web, the deep web, and the dark web (Figure 1; Bergman 2001). The surface web includes any website that can be accessed without logging in or invitation and is indexed by search engines. The deep web includes websites that require either logging in or an invitation to view (e.g., 'private' social media groups, private messaging apps) and may or may not be indexed by search engines. The dark web contains purposefully hidden content, requires special software to access, and is not indexed by search engines (Chen 2011; CRS 2017). Most Internet-based wildlife trade (legal and illegal) is currently occurring on the surface web and deep web (i.e., e-commerce sites, forums, social media: Sung & Fong 2018; Hinsely et al. 2016; IFAW 2018), with minimal evidence to suggest that a negligible amount of wildlife trade is occurring on the dark web (Roberts & Hernandez-Castro 2017).

To date, there have been a variety of creative uses of data collected from the surface and deep web to inform conservation, biosecurity/invasion science, and law enforcement efforts for the illegal trade (references herein). 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 serve as one line of support for 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 in the trade, serving to disentangle correlates of introduction and establishment (Stringham & Lockwood 2018). Also, 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). Social media websites have been used to track the intensity of legal and illegal trade (Jensen et al. 2019). Online access to news outlets (i.e., Google News: https://news.google.com/) has allowed for systematic investigations into wildlife seizures reported in the news (Indraswari et al. 2020).

As researchers increasingly turn to the Internet as a source of information on the wildlife trade, and as the trade of wildlife increases over the Internet, having a unified method for using the Internet to obtain data on the wildlife trade would be helpful. Such a methodology or guide does not currently exist. While previous studies have independently determined how to find relevant websites and collect data, we argue that describing a systematic approach is useful for two main reasons: (i) repeatability and transparency of methods, and (ii) as a primer for future research. Outlining repeatable steps will facilitate repeatable methods for using the Internet as a data source (finding websites, data collection, etc.). Further, there exist many contexts of the wildlife trade that have yet to be explored. A systematic guide can be applied to new contexts of the trade, including new locations and different focal taxa, or derived parts and commodities.

Here, we present an accessible guide to using the Internet (i.e., surface web) to gather data on the wildlife trade. We developed the methodology through our collective knowledge of working with web data and the wildlife trade. We combine the principles of systematic reviews (Koricheva et al. 2013), computer science (Mitchell et al. 2018), data science (Han et al. 2011), and wildlife sciences. 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. We do not intend this paper to be adopted as a strict protocol, as the Internet is highly transient and there needs to be the flexibility to adapt to changing contexts and technology. Furthermore, we posit that, as wildlife trade becomes more pervasive throughout the Internet, future studies and e-surveillance programs will want to increase the number of websites that are monitored to obtain better spatially and temporally resolved data. To explore the implications of this transition, we follow this guide with a case study that explores the Internet trade of vertebrate pets across three countries (Australia, United States, and United Kingdom), where we monitor and collect data from 53 websites. Our case study highlights the current limitations of scaling up studies from few websites to many and we discuss useful future research directions.

Methods

Guide to using the Internet to monitor and quantity the wildlife trade

Our guide is specified 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. Knowledge of computer programming is not required to follow this guide. We focus on e-commerce and marketplace-like websites but other frameworks are available for news outlets and social media (Toivonen et al. 2019; Sonricker Hansen 2012).

1. Defining the scope and purpose of the project

As a first step, defining a specific research question(s) or aim(s) is necessary, since the scope and purpose of the project will influence every subsequent step of the methodology. 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 monitoring for months to years). Other specifics may include the type of website (Appendix S1). On a practical note, the research questions may be influenced by the available data. Thus, there may 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); gathering an inventory of non-native reptiles and amphibians sold as pets in the United States (Stringham and Lockwood 2018); and exploring the social network structure of sellers of horticultural orchids (Hinsely et al. 2016).

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

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 (Koricheva et al. 2013). The two differences are: first, instead of searching the scientific literature, the Internet is searched (via search engines), and, second, not all candidate results will be used for data collection (Section 3). Often, social media groups or accounts will be highly relevant but fail to show up in search engine results. We recommend performing similar searches within the social media website itself (di Minin et al. 2019). It is important to note that the Internet is highly transient: companies/traders go out of business and new ones arise. Websites found in searches can differ in composition and function if surveyed at a later point in time. If the goal is long-term monitoring, then searches may need to be conducted at regularly-timed intervals to revise the list of current candidate websites. Outside of the Internet, there are likely other ways of finding relevant websites such as interviewing a specific community of practice (e.g., reptile keepers and traders). To the best of our knowledge, this method has not been explored, but merits future investigation.

2.1 Defining search phrases

Search phrases are composed of combinations of relevant keywords. We recommend developing a suite of keywords for each target taxa (e.g., species name, trade 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" (Appendix S2). Example search phrases may be: "snakes for sale Australia", "marine fish forum USA", or "orchid store UK". These search phrases should be in the language(s) spoken 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 used among scientists) or names of breeds/morphs/mutations (e.g., Lyons and Natusch 2013), which were not considered in the initial formulation of search phrases.

2.2 Using search engines to perform searches

Search engines 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. For some social media websites with 'private' groups (e.g., Facebook, MeWe), the search can occur within the website itself for relevant groups or by adding the name of the website as a keyword in the search. Once a keyword phrase is searched, the search engine will likely return millions of URLs per phrase. We recommend choosing a cutoff point which balances the quality of search results with search effort (e.g., 20 or 50 results per search). For more information on using search engines see Appendix S3.

2.3 Classifying search engine results

After obtaining URLs from search results, each URL 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 key inclusion criterion can be whether target taxa are being 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 literally ship live animals to your 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).

3. Selecting target websites to monitor

After obtaining the list of candidate websites, the next step is to select which candidate 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 should be collected. One metadata attribute of websites is web traffic statistics, which includes information such as the number of page views per month (see Appendix S4 for more information). Other sources of metadata can be gathered by the researchers themselves, including the type of website and which target taxa are being traded (more than one target taxa can be traded on any one website). In addition, if the website is a classifieds, forum, or social media group, the researchers can conduct a back-of-the-envelope calculation of the average number of posts or listings per day as a metric of popularity. All relevant metadata attributes should be considered when deciding which candidate websites to choose for data collection. Ultimately, expert opinion (i.e., the researchers) 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).

4. Collecting and storing 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 3; Singrodia et al. 2019). Web scrapers organize the contents of a website into a structured tabular format (For more information on web scrapers see Appendix S5). Since each website differs in its underlying structure, custom web scrapers will need to be built 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 (Toivonen et al. 2019). Choosing manual or automatic 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 (Appendix

S5), 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. 2019)

5. Data Cleaning

Data cleaning involves curating each listing for attributes that could not be automatically extracted, but are required for the analysis, such as: species name, quantity, price, or location. Depending on the project, only certain attributes need to be cleaned from the data. For example, if creating an inventory of species, only the species name needs to be resolved. 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 (see our Case Study below). Therefore, data cleaning should be efficiently targeted only for necessary attributes. 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 text box where the user can write anything. Our experience with websites involving the wildlife trade is with the latter, which takes more time to clean. If collecting data manually, simultaneously cleaning data during collection is possible and likely desirable. For information about possible automated data cleaning methods, see Discussion.

5.1 Resolving species names

Resolving the species name of a listing or post is one of the most important parts of data cleaning, and will vary depending on the website. Some pet stores and specialist classifieds websites explicitly state the scientific name while other sites may mention common names or trade names, which complicates species identifications. 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 rank as possible. In some cases, pictures accompanying listings may aid in identification. However, in other cases, online traders may not provide enough information in the listing to identify to species, which is an unfortunate limitation of web data.

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 (Gallagher et al. 2020). In addition, it will enable the researcher to easily look up upstream taxonomy (i.e., Family and Order; R package *taxize*: Chamberlain & Szocs 2013) for analysis. We provide code to gather upstream taxonomy when provided with a taxonomic ID (Appendix S6).

Case study: trade of "exotic" vertebrate pets across three countries

We present a case study that follows the above recommendations. We sought to quantify and compare the trade of live vertebrate "exotic" (i.e., non-domesticated) pet animals occurring online in three majority English speaking countries: Australia, the United States, and the United Kingdom. Detailed methods for our case study can be found in Appendix S7.

Results

From our case study, we retrieved 5,250 search results (URLs) and, using our inclusion criteria, were left with 304 candidate websites of which we selected 53 websites to collect data from. We chose 37 stores, 13 classifieds, 2 forums, and 1 adoption website (Appendix S8). Each website traded/sold one or more of our target taxa (vertebrates) as pets in one of our target locations: US, UK, or Australia.

From all target websites, as of May 19, 2020, we have collected 559,625 unique listings (i.e., non-duplicated listings) with an estimated rate of around 714,000 unique listings per year (Appendix S9). On average, pet stores contained fewer total number of unique listings and a lower rate of new unique listings per week compared to classifieds and forums (Figure 4). The median number of unique listings per year for a store was c. 1,100 and for classifieds/forums was c. 11,000. Yet, the median value was not indicative of every store or classifieds/forums as there was variation within and overlap between their distributions. Still, classifieds/forums had a higher rate of new listings not observed by any store (rate of over > c. 10,000 total listings per year).

Around half the stores (n = 16) and one classifieds website included the scientific name of the species being sold (Figure 5). The remaining 36 websites did not indicate the scientific name (21 stores, 12 classifieds, 2 forums, 1 adoption website). In total, a disproportionate number of listings did not contain the scientific name (95.0%), where an estimated c. 679,000 listings per year will not contain the scientific name. We estimated that it would take around 2.2 years for a single person full-time to manually clean the data (Appendix S10).

Discussion

As more of the global human population shifts to using the Internet, and as ethical and disease concerns of physical markets arise (e.g. COVID-19; Mallapaty 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). There are both advantages and disadvantages to using the Internet as a source of data on the wildlife trade. The main advantage is the ease of data gathering compared to physical market or store surveys, 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 physical store/market surveys. However, there are several caveats and disadvantages to relying on the

Internet data of the wildlife trade. 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. For instance, there may be some contexts where all trade occurs 'on the ground' and not over the Internet. In these contexts, the Internet will provide no useful information to researchers. 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). This is one vital limitation of web data; wildlife traded online represents only the potential of trade. Supplementing data collected online with physical store and/or surveys is a more holistic approach that may be more impactful when considering applied outcomes (e.g., Rowley et al. 2016).

For this guide, we focused on the surface web: websites open to the public without a login or invitation. From our case study findings and other aforementioned research, wildlife trade occurring on the surface web is extremely abundant. However, not all Internet-related wildlife trade occurs on the surface web. Wildlife trade has also been documented in abundance on the deep web, such as in private social media groups (e.g., Facebook: Siriwat & Nijman 2018) and private text messaging apps (e.g., WhatsApp: Setiawan et al. 2019). Our methods of finding relevant websites and automated data collection apply to the deep web with some caveats. In particular, there are certain parts of the deep web that won't appear in search engine results (i.e., private text messaging apps). Furthermore, an added difficulty to monitoring the deep web is that access will require some degree of infiltration to obtain a login or approval to join. In addition, collecting data from deep web sites may require additional ethical considerations, especially if collecting personally identifiable information (Sula 2016). The dark web remains elusive; there is a lack of evidence that wildlife is abundantly traded on the dark web (Roberts & Hernandez-Castro 2017), however, neither has the possibility been conclusively discounted, as not all dark web URLs have been (or can be) searched. Our method of finding relevant websites is not applicable for the dark web because dark web websites are not indexed by search engines. However, if dark web websites are identified by other means, e.g., expert consultation, automated data collection procedures would be similar to those presented in this guide (Cunliffe et al. 2019).

We identified an enormous amount of data available on the live exotic pet trade occurring on the Internet. From 53 websites, we collected data at an estimated rate of over half a million unique listings per year, which is certainly in the realm of "big data" for wildlife research (Dobson et al. 2020). Importantly, we identified a key bottleneck from data collection to analysis – data cleaning, i.e., converting unstructured 'messy' raw data collected from websites to useful data for analysis. Most of the raw data collected from websites were not ready for analysis (only 5% of listings contained the scientific name), and therefore a person will need to manually 'clean' the data to extract information relevant to the analysis. We estimate, for our case, this would take one person working full time around 2 years to clean every listing we collected in one year. For the vast majority of researchers, this amount of time and/or human resources will not be available to them. One option if 'too much' data is collected would be to look at a random subset of the data and evaluate if the subset is representative of the data as a whole. One method to evaluate this is using species accumulation curves (Ugland et al. 2003); if this curve saturates, then potentially the random subset is representative of all the species in the entire dataset (e.g., Nelufule et al. 2020). Conversely, cleaning data may be manageable if the project aim is restricted to either a smaller set of species (or single species), a small number of target websites are chosen, or there is a short timeframe for data collection. Our results on the distribution in the rate of new listings for stores and classifieds (c. 1,000 vs c. 10,000 unique listings per year) can help researchers estimate the amount of resources needed to complete a project.

Automated data cleaning of wildlife trade web data is not yet available, 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 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). Therefore, there will always be the need for human data cleaning. 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 (e.g., Humair et al. 2015).

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 for the illegal trade. 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.

Acknowledgements

We thank Talia Wittmann for graphic design of the figures and Stephanie Moncayo for calculating the rate of data cleaning. 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').

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.

Dobson ADM et al. 2020. Making Messy Data Work for Conservation. One Earth **2**:455–465. Elsevier.

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)

Han J, Pei J, Kamber M. 2011. Data Mining: Concepts and Techniques. Elsevier.

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

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.

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.

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.

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.

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.

Mallapaty S. 2020. China set to clamp down permanently on wildlife trade in wake of coronavirus. Nature. Available on https://www.nature.com/articles/d41586-020-00499-2 (accessed May 2020)

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.

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

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.

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.

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.

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.

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.

Ugland KI, Gray JS, Ellingsen KE. 2003. The species–accumulation curve and estimation of species richness. Journal of Animal Ecology **72**:888–897.

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.

Figures

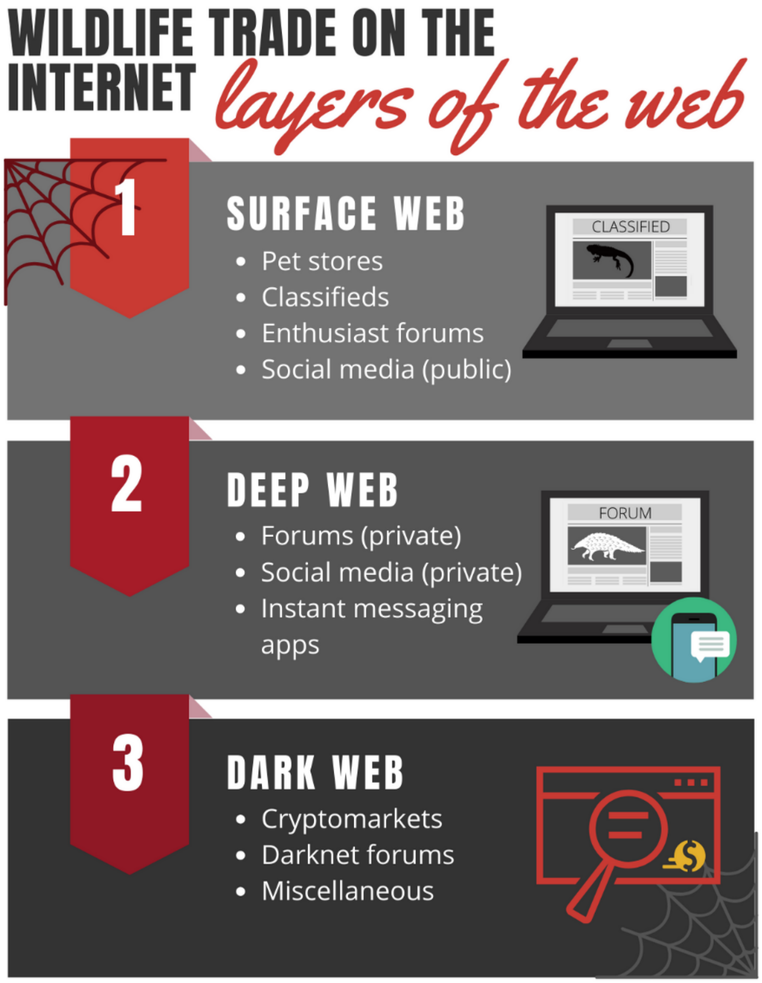


Figure 1.

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 find and monitor. The section of our guide related to 'finding candidate websites' is exclusive to the surface web, which includes websites that can be found through search engines. However, data collection techniques outlined in our guide can be applied to the surface, deep, and dark web.

Figure 2.

Flowchart of our guide to using the Internet to monitor and quantify the wildlife trade.

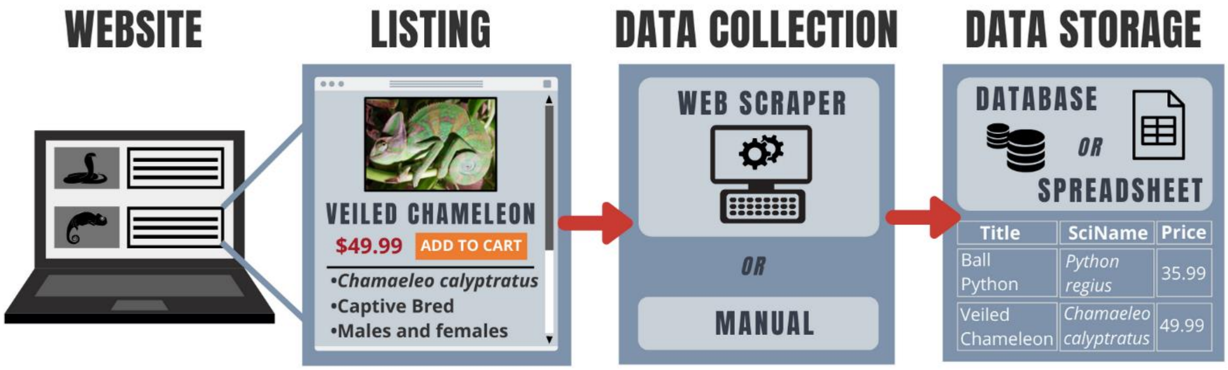


Figure 3.

Data collection and data storage procedure for websites trading wildlife. Websites have underlying HTML code that the web scrapers parse in order to extract relevant information, which can then be stored in a database or spreadsheet. This process can be repeated for different websites using different 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. Most pet store websites aren't updated daily and therefore collecting data weekly or fortnightly can be appropriate. For popular classifieds and social media groups, 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. Conversely, most forums keep an archive of all posts and don't remove old posts, therefore it's less essential that daily data collection occurs. If data collection is to occur frequently, we recommend using automated data collection because manual data collection is more time consuming. However, there is an obvious trade-off between the resources invested in creating web scrapers and the quantity of data that will be collected. Chameleon photo credits: Chris Kade.

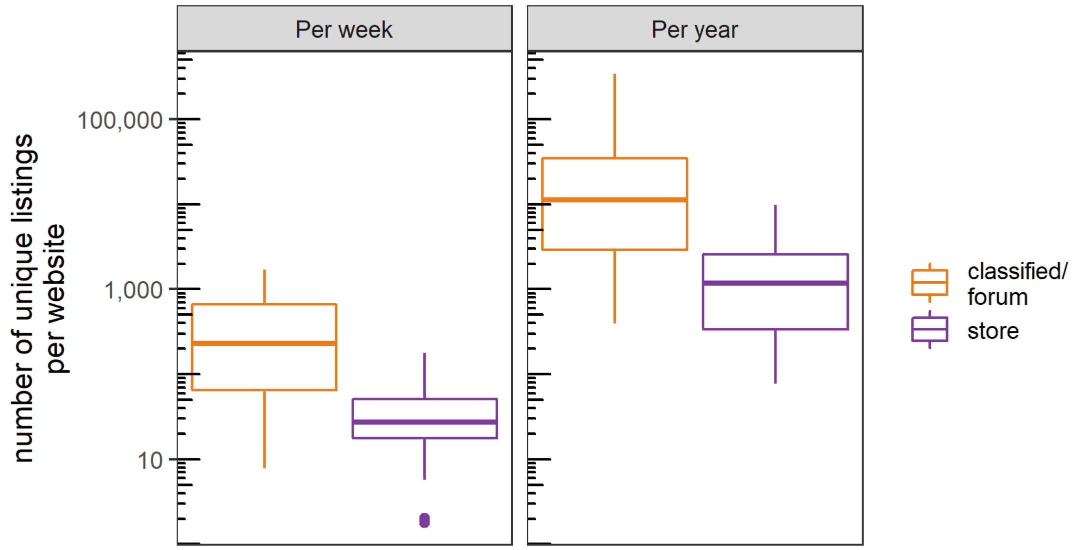


Figure 4.

The number of new unique listings per week (observed) and per year (estimated: Appendix S9) by type of website (n = 15 for classifieds/forums and n = 37 for stores). Note the y axis is on a $\log_{10}$ scale. One adoption website was excluded.

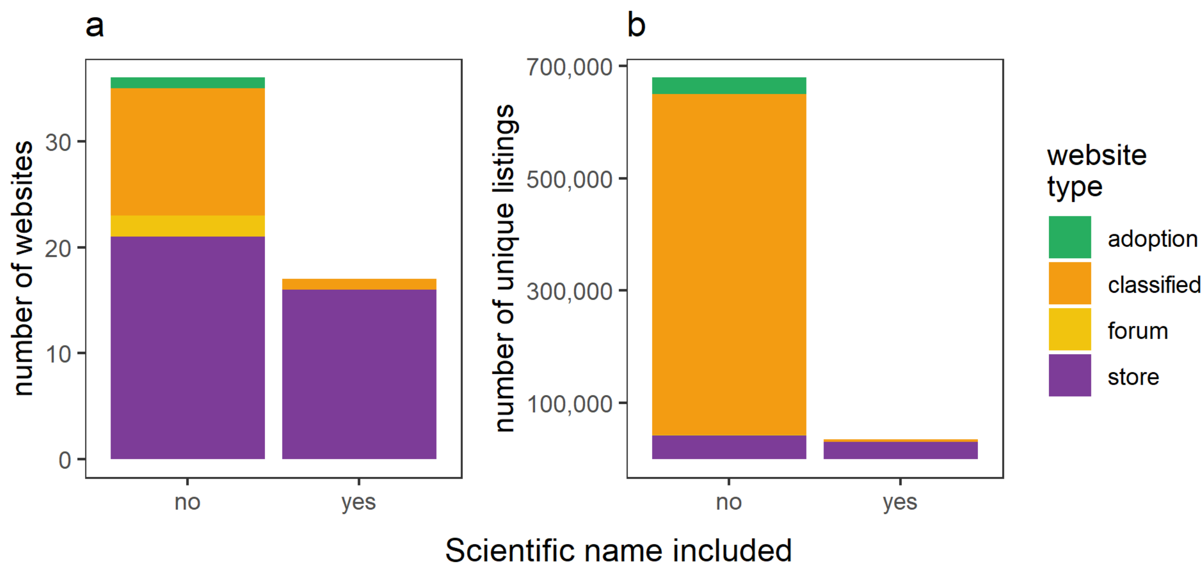


Figure 5.

The number of (a) websites that specify the scientific name of species being traded (n = 37 for stores, n = 13 for classifieds, n = 2 for forums, and n = 1 for adoption websites) for target websites in our case study. (b) The annual number of unique listings with and without a scientific named (estimated: Appendix S9) by website type.

Supporting Information

Table of Contents

Appendix S1: Types of websites relevant to wildlife trade research

We categorize seven types of websites that can be relevant to the wildlife trade:

1. **Pet stores** are often physical storefronts that have websites where they list what species they are selling. Sometimes pet stores will specify whether they can ship, and to where. Other times the online pet stores do not have a physical storefront and is exclusively an online store that ships directly to consumers. Pet stores reliably give either scientific names and/or common names of the species they are selling and their price.
2. **Classifieds** are websites where individual users 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). 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.
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 forums do not remove "sold" listings and usually an entire archive of all sales are kept on the website. Either the common, scientific, or trade name is provided. The location and price are usually provided.

4. **Social media** websites including Facebook and MeWe (among others) have ‘groups’ with a particular purpose where people can join. Some groups focus on trading particular taxa or products. The posts are similar in structure to forums. Some groups are open to the public and others require an invitation (e.g., ‘private’ groups). In addition, individual stores or breeders may maintain social media accounts where they post updates about what species are in stock. Social media websites are among the most popularly-used websites. Facebook has recently implemented a policy that bans the selling of animals (live and derived-parts) on its platform (https://www.facebook.com/policies/commerce/prohibited_content/animals), however these efforts have been ineffective as trade continues to occur (author’s direct observations).
5. **Lost and Found** websites allow users to report a lost or found pet. They are structurally similar to classifieds websites. They provide useful information if exploring invasive species risks. They are usually only available for highly visible species such as turtles and birds (Vall-llosera & Cassey 2017). The species name (scientific, common, or trade name) as well as the location and date is usually provided.
6. **Adoption** websites post pet animals that are available for adoption. This is considered the secondary market for pets. They are structurally similarly to classifieds websites.
7. **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

Pet stores

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.

Classifieds

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.

Forums

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

Social Media

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.

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.

Lost and Found

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.

Adoption

None known to authors.

News

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.

Appendix S2: Table of search phrases

Table of search keywords used to generate search phrases for our case study. 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 stopped each search after 50 search results (i.e., 5 pages of 10 URLs per page) before moving on to the next search.

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

Currently, the most popular search engine is Google (https://www.google.com). Certain search engines offer APIs (application programming interfaces), which can automate the search process by iterating over each search phrase using computer programming (e.g., Bing: Thelwall and Sud 2012). Because search engines can use the user's location to provide personalized result (such as through being logged in to a Google account), extra steps must be taken to ensure that the search engine provides location-relevant results (https://policies.google.com/technologies/location-data). This is especially true if the location you want to find websites for is not the location from where you are accessing the Internet. One way to control the location is to use advanced search features (i.e., Google: https://www.google.com/advanced_search), which allows the user to specify which country to restrict a search to. In addition, using a VPN (virtual private network) may alleviate this issue. Also, some 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.

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). We provide code to access the Alexa API (Appendix S9), which is available through a paid subscription. 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 made from 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, hiring data scientists or contractors to code web scrapers is also an option. 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.

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 is possible. If the data collection is recurrent, then establishing a system to schedule web scrapers to run at regular intervals is important. This 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 (Appendix S1), 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, we decided 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.

Ethical considerations of web scrapers

Ethics approval may be required to collect information from certain websites where personally identifiable material is collected, including social media sites (Zimmer 2010). Care should be taken to ensure de-identified information is used in the final publication. In addition, caution should be taken when storing and sharing data that can be personally identifiable (Harriman & Patel 2014). Web scraping currently encompasses a legal gray area (Zamora 2019), and thus researchers are encouraged to acquire ethics approval prior to using them. Further, 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). This 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.

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

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

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.

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

Appendix S6: Code for gathering upstream taxonomic information using the *taxize* package

To be made public once manuscript is published.

Appendix S7: Methods of Case study - trade of “exotic” vertebrate pets across three countries

1. Define the scope and purpose of the project

We sought to quantify and compare the trade of live vertebrate “exotic” (i.e., non-domesticated) pet animals occurring online in three majority English speaking countries: Australia, the United States, and the United Kingdom. We wanted to include a variety of different website types: pet stores, enthusiast forums, classifieds, and adoption websites.

2. Finding candidate websites where wildlife is traded

We defined a series of search phrases centered around the vertebrate taxa of interest (freshwater aquarium fish, marine aquarium fishes, reptiles, amphibians, and birds), the type of websites (store, classified, forum, etc.), and location. We provide the keywords and search phrases we used in Appendix S2. In total, through all combinations of keywords, we created 105 search phrases. We used the Google search engine to explore our search phrases and stored the top 50 results per search (i.e., 5 pages of results with 10 URLs per page).

We classified each search result as relevant or irrelevant depending on the following inclusion criteria: (1) the target taxa is being traded on this website; and (2) website users are trading in one of the target locations. Since all online transactions are potentially representative of animals being traded, we considered all websites where one can acquire an animal directly (i.e., direct shipping) or indirectly (i.e., facilitating in-person exchange).

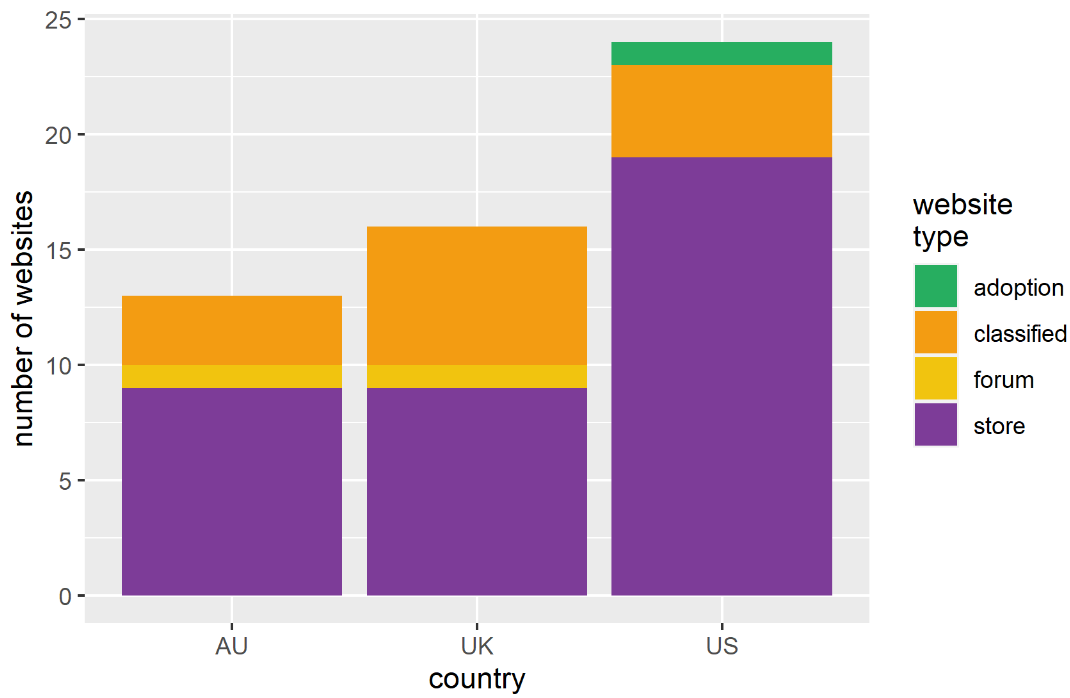
3. Selecting target websites to monitor

We gathered available relevant metadata on each candidate website. For each of our candidate websites, we retrieved Alexa web ranking and the number of page visits per month (if available;

Appendix S3). For each classifieds and forum website, we calculated the approximate rate of new listings (i.e., how many listings posted in the last month). In addition, we calculated the number of times a website showed up in all searches and considered this to be an approximate metric of popularity. We used the above metadata to subjectively choose which websites to collect data from (i.e., our target websites). Further, we wanted a representative number of each type of website (forum, classifieds, store) for each taxa and location. Therefore, we chose at least 3 target websites (if available) for each combination of website type, taxa, and location. We chose to keep the names of websites anonymous as this is considered good ethical practice so as not to interfere with or compromise trading behavior.

4. Collecting and storing data from websites

For each target website, we coded our own web scrapers to collect data in the programming language Python using the libraries bs4, requests, and selenium. We varied how often to collect data depending on the type of website. For pet stores, we collected data once a week, for popular classifieds, once a day, and for less popular classifieds/forums, once every two to three days. The last web scraper was completed in November 2019 and we intend to continue to collect data for at least 2 years. We stored all of the collected data on a local MySQL database. We detected and marked duplicate listings after every data collection event. For stores, we decided that if two or more listings share the exact title and exact text description, they are duplicates. For classifieds websites, we decided that if two or more listings share the same title and the same username, they are duplicates.



Appendix S8: Figure of the number of websites for each country by type of website

Number of websites for each country by type of website chosen in our case study.

Appendix S9: Calculating annual number of listings per website

For each website we coded individual web scrapers to collect data. Each web scraper started at a different date. Therefore, to estimate the annual number of unique listings for each website, we used the following equation:

$$N_i = n_i \times \frac{52}{w_i}$$

where N_i is the estimated number of unique listings per year for website i, n_i is the current number of unique listings collected and w_i is the number of weeks data has been collected for website i.

Appendix S10: Calculating estimated time to clean data

We estimated the total time to clean listings by assuming: (i) a rate of cleaning of 161 listings per hour (estimated by research assistant), (ii) the number of work days in a year is 252 days, and the (iii) work day has 7.5 hours. We assumed that we would need to clean every listing that does not provide a scientific name, which is an estimated 679,000 listings. This results in around 2.2 years of cleaning for one person dedicated solely to data cleaning. See following formula:

$$679{,}000\ listings \times \frac{1\ hour}{161\ listings} \times \frac{1\ working\ day}{7.5\ hours} \times \frac{1\ year}{252\ working\ days} = 2.2\ years$$

Appendix S11: Code for accessing Amazon Alexa Web Information Services API

To be released upon publication