

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

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

Oliver C. Stringham^{1,2}, Adam Toomes¹, Aurelie M. Kanishka^{1,3}, Lewis Mitchell², Sarah Heinrich¹, Joshua V. Ross², Phillip Cassey¹

- 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

16 **Abstract**

17 The unrivalled growth in e-commerce of animals and plants presents an unprecedented opportunity to
18 monitor wildlife trade to inform conservation, biosecurity, and law enforcement efforts. Using the
19 Internet to quantify the scale of the wildlife trade (volume, frequency) is a relatively recent and rapidly
20 developing approach, which currently lacks an accessible framework for finding relevant websites and
21 collecting data. Here, we present an accessible guide for internet-based wildlife trade surveillance,
22 which uses a systematic method to automate data collection from relevant websites. The guide is easily
23 adaptable to the multitude of trade-based contexts including different focal taxa or derived parts, and
24 locations of interest. Furthermore, as wildlife trade on the Internet becomes more widespread, the
25 ability to collect large amounts of data on traded wildlife will become possible and desirable. Using a
26 case study where we monitor 53 websites, we demonstrate the capabilities and limitations of this kind
27 of large-scale surveillance system. We collected over half a million unique listings in a year and estimate
28 that it would take over two years for one person to clean every listing. We propose that the
29 development of machine learning methods for automation of data collection and processing become a
30 priority and be tested for a variety of different contexts of wildlife trade-related web data.

31

32 **Introduction**

33 The wildlife trade is an influential driver of species endangerment, source of invasive species, spread of
34 diseases, and criminal activity ('t Sas-Rolfes et al. 2019). Trade occurs across a variety of physical
35 settings, including 'brick and mortar' stores, wet markets, pet stores – and increasingly on the Internet
36 (Siriwat & Nijman 2020). Reliable data on the quantity and composition of the wildlife trade (legal and
37 illegal) is needed to inform decisions about conservation, biosecurity and law enforcement efforts, and
38 develop human behavior change campaigns; yet this data is often not collected and/or is difficult to
39 obtain (Regueira & Bernard 2012; Eskew et al. 2020).

40

41 In recent years, the Internet has played an increasingly important role in facilitating the wildlife trade
42 (Siriwat & Nijman 2020). Accordingly, recent efforts to describe and quantify the wildlife trade have
43 turned to the Internet (e.g., Alfino & Roberts 2018). The Internet itself (i.e., the World Wide Web or
44 simply the Web) is categorized by three distinct "layers": the surface web, the deep web, and the dark
45 web (Figure 1; Bergman 2001). The surface web includes any website that can be accessed without
46 logging in or invitation and is indexed by search engines. The deep web includes websites that require
47 either logging in or an invitation to view (e.g., 'private' social media groups, private messaging apps) and
48 may or may not be indexed by search engines. The dark web contains purposefully hidden content,
49 requires special software to access, and is not indexed by search engines (Chen 2011; CRS 2017). Most
50 Internet-based wildlife trade (legal and illegal) is currently occurring on the surface web and deep web
51 (i.e., e-commerce sites, forums, social media: Sung & Fong 2018; Hinsely et al. 2016; IFAW 2018), with
52 minimal evidence to suggest that a negligible amount of wildlife trade is occurring on the dark web
53 (Roberts & Hernandez-Castro 2017).

54

55 To date, there have been a variety of creative uses of data collected from the surface and deep web to
56 inform conservation, biosecurity/invasion science, and law enforcement efforts for the illegal trade
57 (references herein). These studies have generally been small in scale (i.e., monitoring one or few
58 websites), but have nonetheless revealed the utility of the Internet to describe different aspects of the
59 wildlife trade. In the context of conservation, classified websites have been used to estimate intensity of
60 trade and serve as one line of support for increases in the legal protection of high-risk species (Rowley et
61 al. 2016). For biological invasions, online pet stores have been used to inventory non-native species in
62 the trade, serving to disentangle correlates of introduction and establishment (Stringham & Lockwood
63 2018). Also, lost and found websites have been used to estimate propagule pressure, a major
64 determinant of non-native establishment probability (Cassey et al. 2018), for commonly held exotic pets
65 (e.g., turtles: Kikillus et al. 2012). In terms of assisting law enforcement, listings from online classifieds
66 have been used to quantify the illegal trade (Ye et al. 2020). Social media websites have been used to
67 track the intensity of legal and illegal trade (Jensen et al. 2019). Online access to news outlets (i.e.,
68 Google News: <https://news.google.com/>) has allowed for systematic investigations into wildlife seizures
69 reported in the news (Indraswari et al. 2020).

70

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

79 guide can be applied to new contexts of the trade, including new locations and different focal taxa, or
80 derived parts and commodities.

81

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

96 **Methods**

97 ***Guide to using the Internet to monitor and quantify the wildlife trade***

98 Our guide is specified in six steps (Figure 2): (1) defining the scope and purpose of the project; (2) finding
99 candidate websites; (3) selecting target websites to monitor; (4) collecting and storing data from
100 websites; (5) cleaning data; and (6) analysis. Here, we detail the process of each step leading up to

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

104

105 *1. Defining the scope and purpose of the project*

106 As a first step, defining a specific research question(s) or aim(s) is necessary, since the scope and
107 purpose of the project will influence every subsequent step of the methodology. At a minimum, it is
108 essential to decide which species, taxa, and/or derived products are of interest, the location(s) of
109 interest, and the timeframe for data collection (i.e., one-time snapshot, versus monitoring for months to
110 years). Other specifics may include the type of website (Appendix S1). On a practical note, the research
111 questions may be influenced by the available data. Thus, there may need to be some exploration of the
112 websites and the kind of data they provide (Steps 2 – 3). Examples of project aims include: quantifying
113 the trade in parrots in different regions of China (Ye et al. 2020); gathering an inventory of non-native
114 reptiles and amphibians sold as pets in the United States (Stringham and Lockwood 2018); and exploring
115 the social network structure of sellers of horticultural orchids (Hinsely et al. 2016).

116

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

118 Finding candidate websites involves three steps: (1) defining keyword phrases to search; (2) using a
119 search engine to perform searches; and (3) classifying the relevance of each search result. This part of
120 the methodology is akin to the process of finding relevant scientific papers in a systematic review or
121 meta-analysis (Koricheva et al. 2013). The two differences are: first, instead of searching the scientific
122 literature, the Internet is searched (via search engines), and, second, not all candidate results will be
123 used for data collection (Section 3). Often, social media groups or accounts will be highly relevant but
124 fail to show up in search engine results. We recommend performing similar searches within the social

125 media website itself (di Minin et al. 2019). It is important to note that the Internet is highly transient:
126 companies/traders go out of business and new ones arise. Websites found in searches can differ in
127 composition and function if surveyed at a later point in time. If the goal is long-term monitoring, then
128 searches may need to be conducted at regularly-timed intervals to revise the list of current candidate
129 websites. Outside of the Internet, there are likely other ways of finding relevant websites such as
130 interviewing a specific community of practice (e.g., reptile keepers and traders). To the best of our
131 knowledge, this method has not been explored, but merits future investigation.

132

133 2.1 Defining search phrases

134 Search phrases are composed of combinations of relevant keywords. We recommend developing a suite
135 of keywords for each target taxa (e.g., species name, trade name, common name, product name), type
136 of websites (Appendix S1), and location of interest. Other useful keywords include adding the terms “for
137 sale” or “buy” (Appendix S2). Example search phrases may be: “snakes for sale Australia”, “marine fish
138 forum USA”, or “orchid store UK”. These search phrases should be in the language(s) spoken in the
139 location of interest. There may be a need to refine keywords after exploratory investigation of search
140 engine results. In particular, there may be trade names (i.e., names for species or taxa used in the
141 wildlife trade community but not used among scientists) or names of breeds/morphs/mutations (e.g.,
142 Lyons and Natusch 2013), which were not considered in the initial formulation of search phrases.

143

144 2.2 Using search engines to perform searches

145 Search engines use proprietary algorithms to return a list of URLs (i.e., website addresses) when a search
146 phrase is input. Search engine algorithms consider the relevance of the keywords, the popularity of the
147 website (i.e., the number of page views), and, increasingly, the location of where the search occurs
148 (Langville and Meyer 2011). The results from a search engine are expected to change at any point in

149 time for a number of reasons including: changes to the search engine algorithm, changes to website
150 popularity metrics, the emergence of new websites, and a change in the location of where the search is
151 performed. For some social media websites with ‘private’ groups (e.g., Facebook, MeWe), the search
152 can occur within the website itself for relevant groups or by adding the name of the website as a
153 keyword in the search. Once a keyword phrase is searched, the search engine will likely return millions
154 of URLs per phrase. We recommend choosing a cutoff point which balances the quality of search results
155 with search effort (e.g., 20 or 50 results per search). For more information on using search engines see
156 Appendix S3.

157

158 2.3 Classifying search engine results

159 After obtaining URLs from search results, each URL will need to be categorized as relevant or irrelevant.
160 Relevance is subjective and we recommend defining inclusion/exclusion criteria depending on the scope
161 and purpose of the study. One key inclusion criterion can be whether target taxa are being traded on
162 the website. Another criterion can be the type of transaction that occurs on the website. Specifically, on
163 the Internet, there are varying levels of “directness” of trade. For instance, some e-commerce
164 companies will literally ship live animals to your doorstep (e.g., pet stores: Holmberg et al. 2015) and
165 there are less direct websites that facilitate the transaction of selling wildlife online, but leave it up to
166 the individuals in the transaction to conduct the exchange (e.g., classifieds: Sung & Fong 2019).

167

168 *3. Selecting target websites to monitor*

169 After obtaining the list of candidate websites, the next step is to select which candidate websites to
170 collect data from (i.e., target websites). This step of the framework is the most subjective and therefore
171 some level of justification and transparency should be provided when choosing target websites. To
172 make informed decisions on selecting target websites, metadata on candidate websites should be

173 collected. One metadata attribute of websites is web traffic statistics, which includes information such
174 as the number of page views per month (see Appendix S4 for more information). Other sources of
175 metadata can be gathered by the researchers themselves, including the type of website and which
176 target taxa are being traded (more than one target taxa can be traded on any one website). In addition,
177 if the website is a classifieds, forum, or social media group, the researchers can conduct a back-of-the-
178 envelope calculation of the average number of posts or listings per day as a metric of popularity. All
179 relevant metadata attributes should be considered when deciding which candidate websites to choose
180 for data collection. Ultimately, expert opinion (i.e., the researchers) is needed to choose target websites,
181 because measures of website metadata are not available for all candidate websites and project
182 relevance is not always straightforward to quantify. The number of target websites chosen will vary
183 based on the project aim(s) and the resources available to collect and clean data (Sections 4 and 5).

184

185 *4. Collecting and storing data from websites*

186 Data collection can occur in one of two main ways: manual or automated. Manual data collection
187 involves visiting the website and recording what taxa/product is being traded along with desired
188 associated attributes (e.g., price, location). Automated data collection involves constructing “web
189 scrapers” to visit the website and extract desired relevant information (Figure 3; Singrodia et al. 2019).
190 Web scrapers organize the contents of a website into a structured tabular format (For more information
191 on web scrapers see Appendix S5). Since each website differs in its underlying structure, custom web
192 scrapers will need to be built for each website individually. A few highly visited websites may have APIs
193 that allow for easy collection of data; this is more likely to be the case for social media websites
194 (Toivonen et al. 2019). Choosing manual or automatic data collection will depend on how long and how
195 often data is being collected, as it takes technical expertise and time to build web scrapers (Appendix

196 S5), which may not be necessary if the number of target websites is small and the data collection
197 window is short (e.g., Heinrich et al. 2019)

198

199

200 *5. Data Cleaning*

201 Data cleaning involves curating each listing for attributes that could not be automatically extracted, but
202 are required for the analysis, such as: species name, quantity, price, or location. Depending on the
203 project, only certain attributes need to be cleaned from the data. For example, if creating an inventory
204 of species, only the species name needs to be resolved. Data cleaning is often a tedious and time-
205 consuming task (Freitas & Curry 2016) and could possibly be the most time-consuming part of the entire
206 project (see our Case Study below). Therefore, data cleaning should be efficiently targeted only for
207 necessary attributes. The amount of cleaning required will depend on the structure of the website and
208 will vary by individual website (Appendix S1). For instance, a website may have a separate field for
209 species names, while another may just have one free form text box where the user can write anything.
210 Our experience with websites involving the wildlife trade is with the latter, which takes more time to
211 clean. If collecting data manually, simultaneously cleaning data during collection is possible and likely
212 desirable. For information about possible automated data cleaning methods, see Discussion.

213

214 5.1 Resolving species names

215 Resolving the species name of a listing or post is one of the most important parts of data cleaning, and
216 will vary depending on the website. Some pet stores and specialist classifieds websites explicitly state
217 the scientific name while other sites may mention common names or trade names, which complicates
218 species identifications. For all practical purposes, identifications down to the rank of species are needed
219 for effective action on conservation, biosecurity, and crime (Rhyne et al. 2012). Therefore, we

220 recommend identifying the taxa to the most specific taxonomic rank as possible. In some cases, pictures
221 accompanying listings may aid in identification. However, in other cases, online traders may not provide
222 enough information in the listing to identify to species, which is an unfortunate limitation of web data.

223

224 If monitoring many species, we recommend relating the species/taxa name to a taxonomic database
225 (e.g., GBIF 2020). Doing so will facilitate conformation to taxonomic names by avoiding synonyms
226 (Gallagher et al. 2020). In addition, it will enable the researcher to easily look up upstream taxonomy
227 (i.e., Family and Order; R package *taxize*: Chamberlain & Szocs 2013) for analysis. We provide code to
228 gather upstream taxonomy when provided with a taxonomic ID (Appendix S6).

229

230 ***Case study: trade of “exotic” vertebrate pets across three countries***

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

235

236 **Results**

237 From our case study, we retrieved 5,250 search results (URLs) and, using our inclusion criteria, were left
238 with 304 candidate websites of which we selected 53 websites to collect data from. We chose 37 stores,
239 13 classifieds, 2 forums, and 1 adoption website (Appendix S8). Each website traded/sold one or more of
240 our target taxa (vertebrates) as pets in one of our target locations: US, UK, or Australia.

241

242 From all target websites, as of May 19, 2020, we have collected 559,625 unique listings (i.e., non-
243 duplicated listings) with an estimated rate of around 714,000 unique listings per year (Appendix S9). On

244 average, pet stores contained fewer total number of unique listings and a lower rate of new unique
245 listings per week compared to classifieds and forums (Figure 4). The median number of unique listings
246 per year for a store was c. 1,100 and for classifieds/forums was c. 11,000. Yet, the median value was not
247 indicative of every store or classifieds/forums as there was variation within and overlap between their
248 distributions. Still, classifieds/forums had a higher rate of new listings not observed by any store (rate of
249 over > c. 10,000 total listings per year).

250

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

257 **Discussion**

258 As more of the global human population shifts to using the Internet, and as ethical and disease concerns
259 of physical markets arise (e.g. COVID-19; Mallapaty 2020), the online trade of wildlife is poised to
260 increase. Thus, the Internet is and will continue to be an invaluable source of data (Lavorgna 2014).

261 There are both advantages and disadvantages to using the Internet as a source of data on the wildlife
262 trade. The main advantage is the ease of data gathering compared to physical market or store surveys,
263 especially if using automated data collection techniques (i.e., web scrapers). Furthermore, using the
264 Internet could potentially allow for a more complete picture of the trade both spatially and temporally
265 than would normally be possible for researchers or organizations who have limited resources for
266 physical store/market surveys. However, there are several caveats and disadvantages to relying on the

267 Internet data of the wildlife trade. First, not all trade occurs nor is observable online (e.g., bushmeat
268 trade; McNamara et al 2019). The degree to which trade occurs online will depend on: the type of trade
269 (i.e., pet, derived products, food, etc.), the taxa, the country or culture in question (i.e., Internet use
270 varies by country; Pew Research Center 2016), and possibly by target/consumer group. For instance,
271 there may be some contexts where all trade occurs ‘on the ground’ and not over the Internet. In these
272 contexts, the Internet will provide no useful information to researchers. To the best of our knowledge,
273 there are no estimates of the ratio of physical versus online trade for any context. Another downside is
274 that it is difficult, if not impossible, to verify the validity of online listings of wildlife (i.e., fake or scam
275 versus genuine advertisements). This is one vital limitation of web data; wildlife traded online
276 represents only the potential of trade. Supplementing data collected online with physical store and/or
277 surveys is a more holistic approach that may be more impactful when considering applied outcomes
278 (e.g., Rowley et al. 2016).

279

280 For this guide, we focused on the surface web: websites open to the public without a login or invitation.
281 From our case study findings and other aforementioned research, wildlife trade occurring on the surface
282 web is extremely abundant. However, not all Internet-related wildlife trade occurs on the surface web.
283 Wildlife trade has also been documented in abundance on the deep web, such as in private social media
284 groups (e.g., Facebook: Siriwat & Nijman 2018) and private text messaging apps (e.g., WhatsApp:
285 Setiawan et al. 2019). Our methods of finding relevant websites and automated data collection apply to
286 the deep web with some caveats. In particular, there are certain parts of the deep web that won’t
287 appear in search engine results (i.e., private text messaging apps). Furthermore, an added difficulty to
288 monitoring the deep web is that access will require some degree of infiltration to obtain a login or
289 approval to join. In addition, collecting data from deep web sites may require additional ethical
290 considerations, especially if collecting personally identifiable information (Sula 2016). The dark web

291 remains elusive; there is a lack of evidence that wildlife is abundantly traded on the dark web (Roberts &
292 Hernandez-Castro 2017), however, neither has the possibility been conclusively discounted, as not all
293 dark web URLs have been (or can be) searched. Our method of finding relevant websites is not
294 applicable for the dark web because dark web websites are not indexed by search engines. However, if
295 dark web websites are identified by other means, e.g., expert consultation, automated data collection
296 procedures would be similar to those presented in this guide (Cunliffe et al. 2019).

297

298 We identified an enormous amount of data available on the live exotic pet trade occurring on the
299 Internet. From 53 websites, we collected data at an estimated rate of over half a million unique listings
300 per year, which is certainly in the realm of “big data” for wildlife research (Dobson et al. 2020).
301 Importantly, we identified a key bottleneck from data collection to analysis – data cleaning, i.e.,
302 converting unstructured ‘messy’ raw data collected from websites to useful data for analysis. Most of
303 the raw data collected from websites were not ready for analysis (only 5% of listings contained the
304 scientific name), and therefore a person will need to manually ‘clean’ the data to extract information
305 relevant to the analysis. We estimate, for our case, this would take one person working full time around
306 2 years to clean every listing we collected in one year. For the vast majority of researchers, this amount
307 of time and/or human resources will not be available to them. One option if ‘too much’ data is collected
308 would be to look at a random subset of the data and evaluate if the subset is representative of the data
309 as a whole. One method to evaluate this is using species accumulation curves (Ugland et al. 2003); if this
310 curve saturates, then potentially the random subset is representative of all the species in the entire
311 dataset (e.g., Nelufule et al. 2020). Conversely, cleaning data may be manageable if the project aim is
312 restricted to either a smaller set of species (or single species), a small number of target websites are
313 chosen, or there is a short timeframe for data collection. Our results on the distribution in the rate of

314 new listings for stores and classifieds (c. 1,000 vs c. 10,000 unique listings per year) can help researchers
315 estimate the amount of resources needed to complete a project.

316

317 Automated data cleaning of wildlife trade web data is not yet available, however, there is potential from
318 computer science subfields, such as machine learning, to help with cleaning messy data (Lamda et al.
319 2019; Norouzzadeh et al. 2020). Tools relevant to wildlife trade websites are image classification and
320 text classification (e.g., Deep learning and Natural Language Processing: Di Minin et al. 2018; Silge &
321 Robinson 2020), which can potentially use images or text to identify certain attributes of a given listing,
322 such as the species being traded. However, there is a paucity of applications of these tools/fields to web
323 data of the wildlife trade specifically (Xu et al. 2019). Importantly, underlying all of these machine
324 learning tools are training sets, which are a representative sample of listings that have been manually
325 classified by a human for the machine learning algorithm to use (Lamda et al. 2019). The larger the
326 training set, the more likely the machine learning model will perform better (Norouzzadeh et al. 2020).
327 Therefore, there will always be the need for human data cleaning. One major barrier to successful
328 implementation of automated data cleaning tools for wildlife trade data is the number of species
329 involved in the trade, where research contexts can encompass hundreds to thousands of species (e.g.,
330 Humair et al. 2015).

331

332 Despite the limitations of data collected from the Internet, there are vast opportunities to inform
333 conservation, biosecurity, and law enforcement objectives. Current strategies of researchers using
334 small-scale monitoring (i.e., one or few websites) should continue to provide insight into specific
335 taxa/products contexts (Sung & Fong 2018). With the development of machine learning tools to clean
336 messy web data, there will be the possibility of creating large-scale (i.e., for many websites) automated
337 systems to detect illegal trade to help inform law enforcement and conservation efforts for the illegal

338 trade. Likewise, early risk-screening and rapid response systems may be possible for invasive species
339 (e.g., Marshall Meyers 2020; Suiter & Sferrazza 2007), especially for 'exotic' animals and ornamental
340 plants whose online trade is commonplace (Lockwood et al. 2019; Lenda et al. 2014). Regardless of the
341 ultimate application, our guide can serve as a primer and starting point to establishing research agendas
342 related to wildlife trade occurring on the Internet.

343

344 **Acknowledgements**

345 We thank Talia Wittmann for graphic design of the figures and Stephanie Moncayo for calculating the
346 rate of data cleaning. This work was supported by funding from the Centre for Invasive Species Solutions
347 (PO1-I-002: 'Understanding and intervening in illegal trade in non-native species').

348

349

350 **References**

- 351 Alfino S, Roberts DL. 2018. Code word usage in the online ivory trade across four European Union
352 member states. *Oryx*:1–5. Cambridge University Press.
- 353 Bergman MK. 2001. White Paper: The Deep Web: Surfacing Hidden Value. *Journal of Electronic*
354 *Publishing* **7**. Available from <http://hdl.handle.net/2027/spo.3336451.0007.104>.
- 355 Cassey P, Delean S, Lockwood JL, Sadowski JS, Blackburn TM. 2018. Dissecting the null model for
356 biological invasions: A meta-analysis of the propagule pressure effect. *PLOS Biology*
357 **16**:e2005987. Public Library of Science.
- 358 Chamberlain S, Szocs E. 2013. “taxize - taxonomic search and retrieval in R.” *F1000Research*.
359 <http://f1000research.com/articles/2-191/v2>.
- 360 Chen H. 2011. *Dark Web: Exploring and Data Mining the Dark Side of the Web*. Springer Science &
361 Business Media.
- 362 Congressional Research Service, March 10, 2017 , “Dark Web”. Available on
363 <https://crsreports.congress.gov/product/pdf/R/R44101> (accessed May 2020)
- 364 Cunliffe J, Décary-Hêtu D, Pollak TA. 2019. Nonmedical prescription psychiatric drug use and the
365 darknet: A cryptomarket analysis. *International Journal of Drug Policy* **73**:263–272.
- 366 Di Minin E, Fink C, Tenkanen H, Hiippala T. 2018. Machine learning for tracking illegal wildlife trade on
367 social media. *Nature Ecology & Evolution* **2**:406–407. Nature Publishing Group.
- 368 Dobson ADM et al. 2020. Making Messy Data Work for Conservation. *One Earth* **2**:455–465. Elsevier.
- 369 Eskew EA, White AM, Ross N, Smith KM, Smith KF, Rodríguez JP, Zambrana-Torrel C, Karesh WB,
370 Daszak P. 2020. United States wildlife and wildlife product imports from 2000–2014. *Scientific*
371 *Data* **7**:1–8.
- 372 Freitas A, Curry E. 2016. Big Data Curation. Pages 87–118 in J. M. Cavanillas, E. Curry, and W. Wahlster,
373 editors. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big*

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

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

378 GBIF: The Global Biodiversity Information Facility (2020) “What is GBIF?”. Available from
379 <https://www.gbif.org/what-is-gbif> (accessed May 2020)

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

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

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

385 Holmberg RJ, Tlusty MF, Futoma E, Kaufman L, Morris JA, Rhyne AL. 2015. The 800-Pound Grouper in the
386 Room: Asymptotic Body Size and Invasiveness of Marine Aquarium Fishes. *Marine Policy* **53**:7–
387 12.

388 Humair F, Humair L, Kuhn F, Kueffer C. 2015. E-commerce trade in invasive plants. *Conservation Biology*
389 **29**:1658–1665.

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

393 International Fund for Animal Welfare. 2018. “Disrupt: Wildlife Cybercrime”. Available on
394 [https://d1jyxxz9imt9yb.cloudfront.net/resource/218/attachment/original/IFAW -
395 Disrupt Wildlife Cybercrime - FINAL English - new logo.pdf](https://d1jyxxz9imt9yb.cloudfront.net/resource/218/attachment/original/IFAW_-_Disrupt_Wildlife_Cybercrime_-_FINAL_English_-_new_logo.pdf) (Accessed May 2020)

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

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

401 Koricheva J, Gurevitch J, Mengersen K. 2013. *Handbook of Meta-analysis in Ecology and Evolution*.
402 Princeton University Press.

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

405 Langville AN, Meyer CD. 2011. *Google’s PageRank and Beyond: The Science of Search Engine Rankings*.
406 Princeton University Press.

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

408 Lenda M, Skórka P, Knops JMH, Moroń D, Sutherland WJ, Kuszewska K, Woyciechowski M. 2014. Effect
409 of the Internet Commerce on Dispersal Modes of Invasive Alien Species. *PLOS ONE* **9**:e99786.
410 Public Library of Science.

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

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

415 Mallapaty S. 2020. China set to clamp down permanently on wildlife trade in wake of coronavirus.
416 *Nature*. Available on <https://www.nature.com/articles/d41586-020-00499-2> (accessed May
417 2020)

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

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

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

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

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

431 Pew Research Center, February, 2016, “Smartphone Ownership and Internet Usage Continues to Climb
432 in Emerging Economies”. Available on
433 [https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-](https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/)
434 [continues-to-climb-in-emerging-economies/](https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/) (Accessed May 2020)

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

437 Rhyne AL, Tlusty MF, Schofield PJ, Kaufman L, Morris JA, Bruckner AW. 2012. Revealing the Appetite of
438 the Marine Aquarium Fish Trade: The Volume and Biodiversity of Fish Imported into the United
439 States. *PLoS ONE* **7**. Available from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3357433/>
440 (accessed April 2, 2020).

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

443 Rowley JLL, Shepherd CR, Stuart BL, Nguyen TQ, Hoang HD, Cutajar TP, Wogan GOU, Phimmachak S.
444 2016. Estimating the global trade in Southeast Asian newts. *Biological Conservation* **199**:96–100.

445 Setiawan A, Iqbal M, Halim A, Saputra RF, Setiawan D, Yustian I. 2019. First description of an immature
446 Sumatran striped rabbit (*Nesolagus netscheri*), with special reference to the wildlife trade in
447 South Sumatra. *Mammalia* **1**. De Gruyter. Available from
448 [https://www.degruyter.com/view/journals/mamm/ahead-of-print/article-10.1515-mammalia-](https://www.degruyter.com/view/journals/mamm/ahead-of-print/article-10.1515-mammalia-2018-0217/article-10.1515-mammalia-2018-0217.xml)
449 [2018-0217/article-10.1515-mammalia-2018-0217.xml](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).

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

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

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

456 Siritwat P, Nijman V. 2020. Wildlife trade shifts from brick-and-mortar markets to virtual marketplaces: A
457 case study of birds of prey trade in Thailand. *Journal of Asia-Pacific Biodiversity*. Available from
458 <http://www.sciencedirect.com/science/article/pii/S2287884X2030042X> (accessed April 24,
459 2020).

460 Sonricker Hansen AL, Li A, Joly D, Mekar S, Brownstein JS. 2012. Digital Surveillance: A Novel Approach
461 to Monitoring the Illegal Wildlife Trade. *PLoS ONE* **7**. Available from
462 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3517447/> (accessed February 14, 2020).

463 Stringham OC, Lockwood JL. 2018. Pet problems: Biological and economic factors that influence the
464 release of alien reptiles and amphibians by pet owners. *Journal of Applied Ecology* **55**:2632–
465 2640.

466 Suiter K, Sferrazza S. 2007. MONITORING THE SALE AND TRAFFICKING OF INVASIVE VERTEBRATE SPECIES
467 USING AUTOMATED INTERNET SEARCH AND SURVEILLANCE TOOLS:5.

468 Sula CA. 2016. Research Ethics in an Age of Big Data. *Bulletin of the Association for Information Science
469 and Technology* **42**:17–21.

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

472 't Sas-Rolfes M, Challender DWS, Hinsley A, Veríssimo D, Milner-Gulland EJ. 2019. Illegal Wildlife Trade:
473 Scale, Processes, and Governance. *Annual Review of Environment and Resources* **44**:201–228.

474 Toivonen T, Heikinheimo V, Fink C, Hausmann A, Hiippala T, Järv O, Tenkanen H, Di Minin E. 2019. Social
475 media data for conservation science: A methodological overview. *Biological Conservation*
476 **233**:298–315.

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

479 Xu Q, Li J, Cai M, Mackey TK. 2019. Use of Machine Learning to Detect Wildlife Product Promotion and
480 Sales on Twitter. *Frontiers in Big Data* **2**. Available from
481 <https://www.frontiersin.org/articles/10.3389/fdata.2019.00028/full> (accessed February 14,
482 2020).

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

486
487



489

490 *Figure 1.*

491 Where wildlife trade occurs on the Internet. Within the Internet, there are three “layers” of where
492 websites can exist: the surface web, the deep web, and the dark web. As wildlife trade moves to
493 websites on the deep and dark web, it becomes increasingly obfuscated (denoted by darkening gray
494 background), making it more difficult for researchers to find and monitor. The section of our guide
495 related to ‘finding candidate websites’ is exclusive to the surface web, which includes websites that can
496 be found through search engines. However, data collection techniques outlined in our guide can be
497 applied to the surface, deep, and dark web.

498

499

500

501

502

503

504

USING THE INTERNET TO MONITOR WILDLIFE TRADE

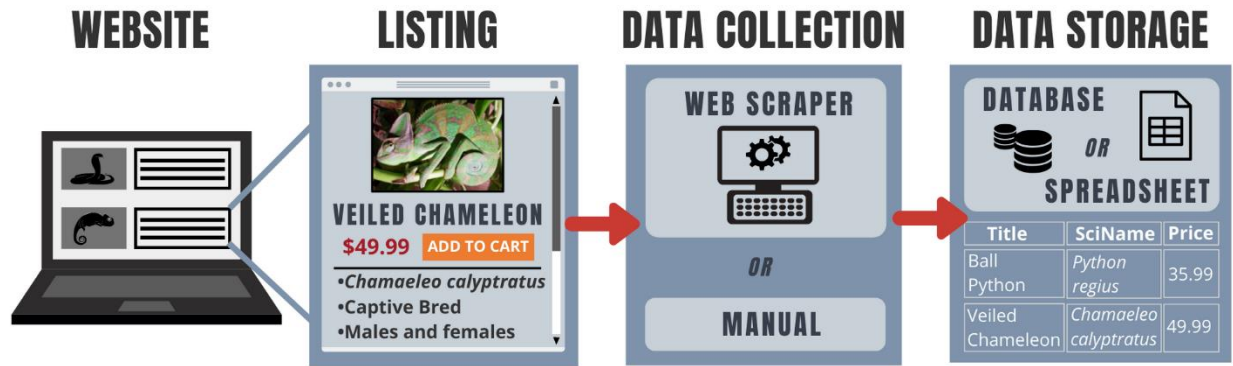


505

506 *Figure 2.*

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

508

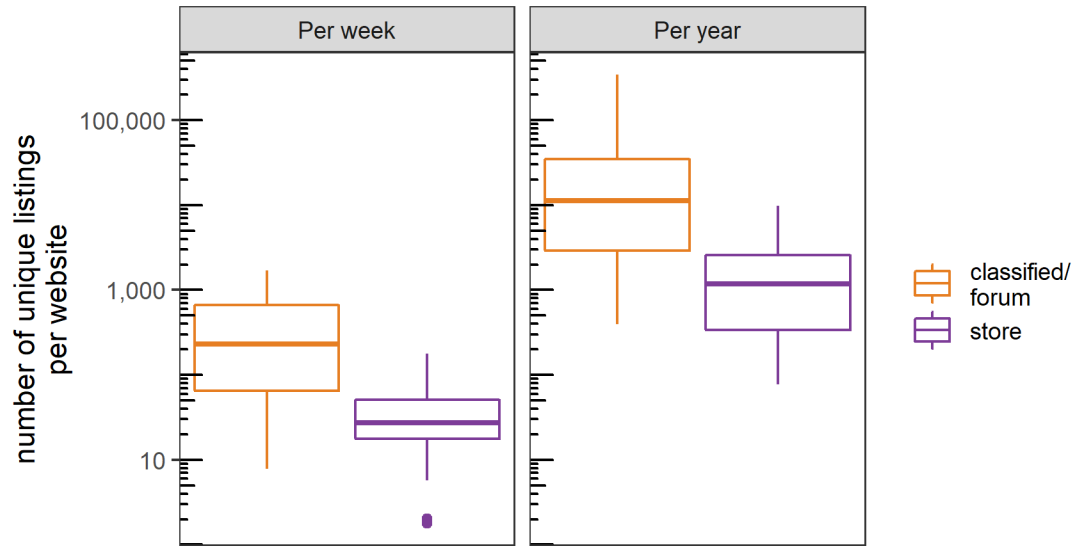


509

510 *Figure 3.*

511 Data collection and data storage procedure for websites trading wildlife. Websites have underlying
 512 HTML code that the web scrapers parse in order to extract relevant information, which can then be
 513 stored in a database or spreadsheet. This process can be repeated for different websites using different
 514 custom web scraper code (see Appendix S5 for more information). The frequency with which to collect
 515 data will depend on the nature of the website, including how often the website is updated. Most pet
 516 store websites aren't updated daily and therefore collecting data weekly or fortnightly can be
 517 appropriate. For popular classifieds and social media groups, data collection will likely be daily or every
 518 two to three days. Some classified websites make listings 'inaccessible' once the seller finds a buyer.
 519 Therefore, for these types of websites, it's important to collect data more frequently in order to capture
 520 listings before they are removed. Conversely, most forums keep an archive of all posts and don't remove
 521 old posts, therefore it's less essential that daily data collection occurs. If data collection is to occur
 522 frequently, we recommend using automated data collection because manual data collection is more
 523 time consuming. However, there is an obvious trade-off between the resources invested in creating web
 524 scrapers and the quantity of data that will be collected. Chameleon photo credits: Chris Kade.

525



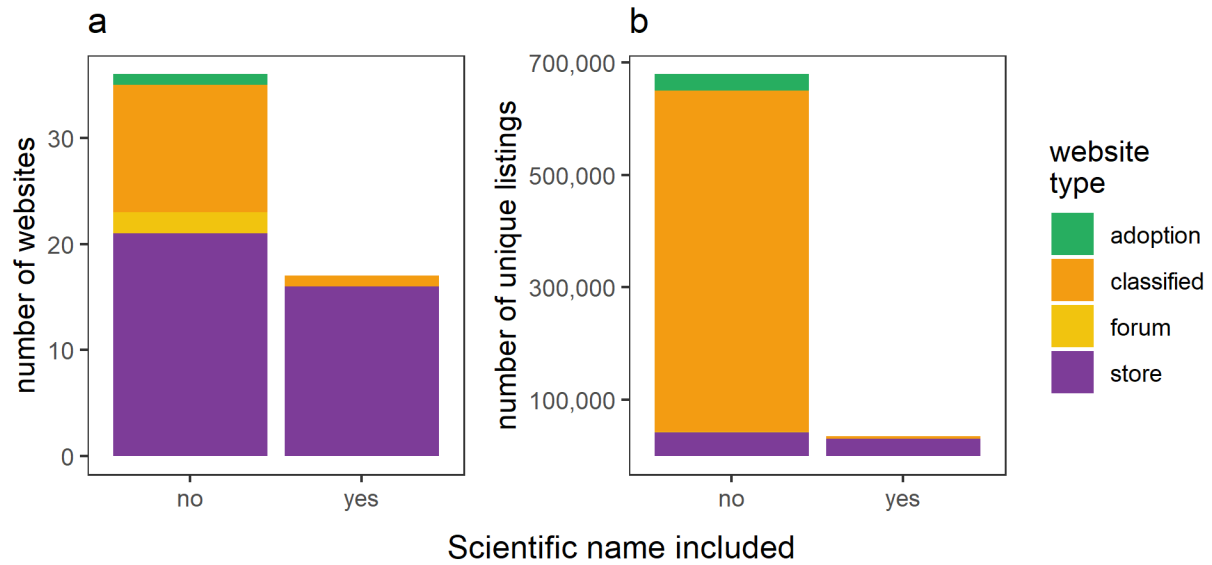
526

527 *Figure 4.*

528 The number of new unique listings per week (observed) and per year (estimated: Appendix S9) by type
 529 of website (n = 15 for classifieds/forums and n = 37 for stores). Note the y axis is on a log₁₀ scale. One
 530 adoption website was excluded.

531

532



533

534 *Figure 5.*

535 The number of (a) websites that specify the scientific name of species being traded (n = 37 for stores, n =

536 13 for classifieds, n = 2 for forums, and n = 1 for adoption websites) for target websites in our case

537 study. (b) The annual number of unique listings with and without a scientific named (estimated:

538 Appendix S9) by website type.

539

540 **Supporting Information**

541 *Table of Contents*

542 Appendix S1: Types of websites relevant to wildlife trade research

543 Appendix S2: Table of search phrases

544 Appendix S3: Further information on search engines

545 Appendix S4: Further information on web traffic statistics

546 Appendix S5: Further information on web scrapers, data storage, and marking duplicates

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

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

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

550 Appendix S9: Calculating annual number of listings per website

551 Appendix S10: Calculating estimated time to clean data

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

553

554

555 **Appendix S1: Types of websites relevant to wildlife trade research**

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

- 557 1. **Pet stores** are often physical storefronts that have websites where they list what species they
558 are selling. Sometimes pet stores will specify whether they can ship, and to where. Other times
559 the online pet stores do not have a physical storefront and is exclusively an online store that
560 ships directly to consumers. Pet stores reliably give either scientific names and/or common
561 names of the species they are selling and their price.
- 562 2. **Classifieds** are websites where individual users can post their animal/wildlife/products they
563 wish to trade. They usually appear on screen in reverse chronological order where the most
564 recent listings appear first. Some classified websites are exclusive to particular taxa (e.g., only
565 reptiles), while others have separate categories for multiple taxa (e.g., a bird section and a
566 reptile section). Classified listings often contain some form of the taxa or product name:
567 scientific, common, trade names. However, this will vary by website, by taxa, and by individual
568 traders. Most classified websites remove listings once they are “sold”. Price is usually provided
569 by the user and therefore a distribution of prices for a given species or products can be derived.
570 The location of the sale is usually given as well.
- 571 3. **Forums** are specialist websites where enthusiasts discuss various aspects of the taxa of interest.
572 Many forums have a dedicated marketplace subforum where trading occurs. The marketplace
573 subforums are structurally similar to classified websites. One key difference is that users can
574 comment below the initial post asking clarifying questions. From these questions it may be
575 possible to determine if the transaction/sale took place. Another difference is that forums do
576 not remove “sold” listings and usually an entire archive of all sales are kept on the website.
577 Either the common, scientific, or trade name is provided. The location and price are usually
578 provided.

- 579 4. **Social media** websites including Facebook and MeWe (among others) have ‘groups’ with a
580 particular purpose where people can join. Some groups focus on trading particular taxa or
581 products. The posts are similar in structure to forums. Some groups are open to the public and
582 others require an invitation (e.g., ‘private’ groups). In addition, individual stores or breeders may
583 maintain social media accounts where they post updates about what species are in stock. Social
584 media websites are among the most popularly-used websites. Facebook has recently
585 implemented a policy that bans the selling of animals (live and derived-parts) on its platform
586 (https://www.facebook.com/policies/commerce/prohibited_content/animals), however these
587 efforts have been ineffective as trade continues to occur (author’s direct observations).
- 588 5. **Lost and Found** websites allow users to report a lost or found pet. They are structurally similar
589 to classifieds websites. They provide useful information if exploring invasive species risks. They
590 are usually only available for highly visible species such as turtles and birds (Vall-Ilosera & Cassey
591 2017). The species name (scientific, common, or trade name) as well as the location and date is
592 usually provided.
- 593 6. **Adoption** websites post pet animals that are available for adoption. This is considered the
594 secondary market for pets. They are structurally similar to classifieds websites.
- 595 7. **News** websites contain news from either print or electronic news companies. For the wildlife
596 trade, many seizures of illegal wildlife are often reported in the news and may be used as a
597 source of data.

598

599 **Select references**

600 Pet stores

601 Holmberg RJ, Tlusty MF, Futoma E, Kaufman L, Morris JA, Rhyne AL. 2015. The 800-Pound Grouper in the
602 Room: Asymptotic Body Size and Invasiveness of Marine Aquarium Fishes. *Marine Policy* 53:7–
603 12.

604

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

608

609 Classifieds

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

613

614 Forums

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

617

618 Social Media

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

622

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

625

626 Lost and Found

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

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

631

632 Adoption

633 None known to authors.

634

635 News

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

639

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

641 **Appendix S2: Table of search phrases**

642 Table of search keywords used to generate search phrases for our case study. The “taxa” column refers
643 to the taxa of interest; the “location” refers to our target locations, and “website type” refers to the
644 website types of interest. We stopped each search after 50 search results (i.e., 5 pages of 10 URLs per
645 page) before moving on to the next search.

646
647 We obtained the search phrases by performing all combinations of “taxa”, “location”, and “website
648 type”, using the follow search phrase templates:

- 649 1. Buy {taxa} {location}
- 650 2. {taxa} for sale OR purchase {location}
- 651 3. {taxa} {website type} {location}

652

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

653

654

655

656 **Appendix S3: Further information on search engines**

657 Currently, the most popular search engine is Google (<https://www.google.com>). Certain search engines
658 offer APIs (application programming interfaces), which can automate the search process by iterating
659 over each search phrase using computer programming (e.g., Bing: Thelwall and Sud 2012). Because
660 search engines can use the user's location to provide personalized result (such as through being logged
661 in to a Google account), extra steps must be taken to ensure that the search engine provides location-
662 relevant results (<https://policies.google.com/technologies/location-data>). This is especially true if the
663 location you want to find websites for is not the location from where you are accessing the Internet.
664 One way to control the location is to use advanced search features (i.e., Google:
665 https://www.google.com/advanced_search), which allows the user to specify which country to restrict a
666 search to. In addition, using a VPN (virtual private network) may alleviate this issue. Also, some websites
667 can opt out of appearing on search engines (Carl Drott et al. 2002), so if a website is known to be
668 important, but does not appear in the search engine results it may still be worth considering it as a
669 candidate website.

670 **References**

- 671 Carl Drott M. 2002. Indexing aids at corporate websites: the use of robots.txt and META tags.
672 Information Processing & Management 38:209–219.
673
674 Thelwall M, Sud P. 2012. Webometric research with the Bing Search API 2.0. Journal of Informetrics
675 6:44–52.
676

677 **Appendix S4: Further information on web traffic statistics**

678 Many websites have web traffic statistics (i.e., metadata) that have been recorded by third party
679 companies. For a given website, these traffic statistics can include: the number of page views per
680 month, the rank/popularity, the country where the website is most popular, and more. One provider of
681 website metadata is Amazon Alexa Web Information Services (<https://www.alexa.com/siteinfo>), which
682 also has an API (https://aws.amazon.com/marketplace/pp/B07Q71HJ3H?ref=srh_res_product_title).
683 We provide code to access the Alexa API (Appendix S9), which is available through a paid subscription.
684 There are a couple of caveats to using web traffic statistics. First is that traffic statistics are calculated for
685 the entire website (i.e., website domain). If the website's only purpose is to trade the target taxa, then
686 this will not be an issue (i.e., online pet store). However, for many websites, there are other reasons
687 people visit the website than to trade the target taxa. For example, the web traffic statistics for eBay, a
688 popular American e-commerce marketplace, would pertain to all trade on eBay and would therefore be
689 unrepresentative of the specific trade. This makes it difficult to compare traffic statistics between
690 websites. In addition, it's important to note that web traffic statistics are not available for all websites.
691 Given these caveats, we recommend using web traffic data as only one line of evidence in choosing a
692 target website.

693

694

695 **Appendix S5: Detailed information on web scrapers and data storage**

696 Background on web scrapers

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

711

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

718

719 Data storage

720 Data collected by web scrapers must be stored in a way that is retrievable for cleaning and subsequent
721 analysis. Data storage can be achieved by using spreadsheets or databases (i.e., Database Management
722 Systems such as MySQL). The choice is dependent on the researcher's familiarity with either, and the
723 frequency or total number of data collection events to be stored. Regardless of the data storage
724 technique, since the fields or columns will likely differ between websites (Appendix S1), the researcher
725 will need to organize and collate data for each website separately.

726

727 Duplicated listings

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

734

735 Ethical considerations of web scrapers

736 Ethics approval may be required to collect information from certain websites where personally
737 identifiable material is collected, including social media sites (Zimmer 2010). Care should be taken to
738 ensure de-identified information is used in the final publication. In addition, caution should be taken
739 when storing and sharing data that can be personally identifiable (Harriman & Patel 2014). Web scraping
740 currently encompasses a legal gray area (Zamora 2019), and thus researchers are encouraged to acquire
741 ethics approval prior to using them. Further, web scrapers can cause 'harm' to the targeted website
742 because they take up bandwidth on the website's server (Zamora 2019). Care should be taken not to

743 overwhelm the targeted website with the web scraper by spacing out visits to the website (i.e., a few
744 seconds between navigating pages). This is especially important in web scraper development. Some
745 websites may have an auto block feature, where they will block an IP address if too many visits occur in
746 a short amount of time.

747

748 **References**

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

750

751 Harrison J. 2020. RSelenium: R Bindings for 'Selenium WebDriver'. R package version 1.7.7.

752 <https://CRAN.R-project.org/package=RSelenium>

753

754 Harriman S, Patel J. 2014. The ethics and editorial challenges of internet-based research. BMC Medicine
755 12:124.

756

757 Holmberg RJ, Tlusty MF, Futoma E, Kaufman L, Morris JA, Rhyne AL. 2015. The 800-Pound Grouper in the
758 Room: Asymptotic Body Size and Invasiveness of Marine Aquarium Fishes. Marine Policy 53:7–
759 12.

760

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

763

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

766

767 Wickham H. 2019. rvest: Easily Harvest (Scrape) Web Pages. R package version
768 0.3.5. <https://CRAN.R-project.org/package=rvest>

769

770 Zamora A. 2019. Making Room for Big Data: Web Scraping and an Affirmative Right to Access Publicly
771 Available Information Online. Journal of Business, Entrepreneurship and the Law 12:203–228.

772

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

775

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

777 To be made public once manuscript is published.

778

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

780 1. Define the scope and purpose of the project

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

785

786 2. Finding candidate websites where wildlife is traded

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

793

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

799

800 3. Selecting target websites to monitor

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

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

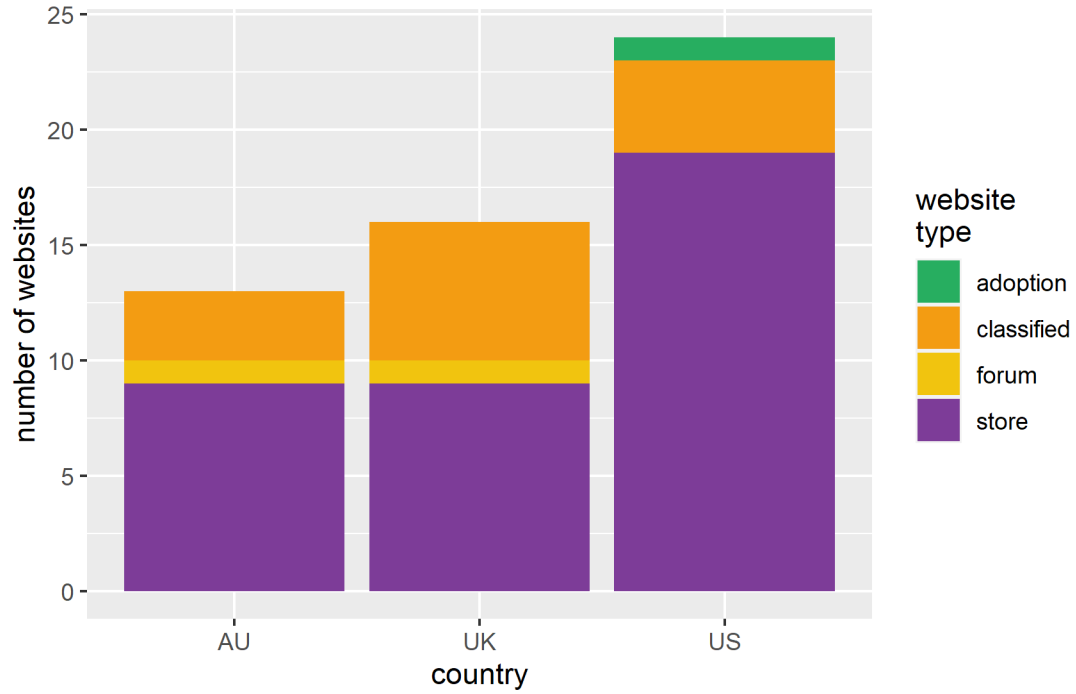
811

812 4. Collecting and storing data from websites

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

822

823



824

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

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

827

828 **Appendix S9: Calculating annual number of listings per website**

829

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

833
$$N_i = n_i \times \frac{52}{w_i}$$

834

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

837

838

839

840 **Appendix S10: Calculating estimated time to clean data**

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

846

847
$$679,000 \text{ listings} \times \frac{1 \text{ hour}}{161 \text{ listings}} \times \frac{1 \text{ working day}}{7.5 \text{ hours}} \times \frac{1 \text{ year}}{252 \text{ working days}} = 2.2 \text{ years}$$

848

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

850 To be released upon publication

851

852

853

854

855