Full Title: Internet scientific name frequency as an indicator of cultural salience of biodiversity

Author Names: Ricardo A. Correia[1,2,*], Paul Jepson[2], Ana C. M. Malhado[1], Richard J. Ladle[1,2]

Author affiliations:

[1]Institute of Biological and Health Sciences, Federal University of Alagoas, Av. Lourival Melo Mota, s/n, Tabuleiro do Martins, 57072-90, Maceió, AL, Brazil

[2]School of Geography and the Environment, University of Oxford, Oxford OX1 3QY, United Kingdom

* Corresponding author:

E-mail: rahc85@gmail.com (RAC)

Keywords:

Biodiversity, birds, culturomics, Internet salience, public perception, scientific names, vernacular names

Highlights:

- Global scale monitoring of biodiversity perceptions is challenging
- Culturomics is a potential solution, but has inherent technical challenges
- Salience of vernacular species names is influenced by linguistic variability
- Web frequency of vernacular and scientific species names is strongly correlated
- Scientific species names should be used to assess digital salience of biodiversity

Abstract:

Public interest in nature is an important driver of the success of conservation actions, such that increasing public awareness of biodiversity has become a major conservation goal (i.e. Aichi Target 1). Macro-scale monitoring of public interest towards nature has thus far been difficult, but the enormous quantity of information generated by the internet allows for new approaches using culturomic techniques. For example, other things being equal, we would expect that the vernacular (common) names of charismatic species with high levels of public interest (e.g. tiger, elephant) to appear on more web-pages than less 'cultural' species. Nevertheless, deriving metrics from such data is challenging because vernacular names often have multiple meanings (e.g. teal, jaguar) that could significantly bias culturomic metrics of cultural visibility. Scientific binomial names of species potentially avoid this problem because Latin is a 'dead' language and the scientific name typically applies only to the biological organism. Here, we investigate whether standard scientific names: i) are a robust proxy of web salience of vernacular species names, and; ii) have the same statistical relationship with vernacular species names across different cultural and language groups. Automated internet searches were carried out for scientific and vernacular names from a global bird species list and six national bird species lists (Australia, Brazil, Indonesia, Spain, Tanzania and USA). For national searches the results were restricted to country web domains. We found strong and consistent correlations between vernacular and scientific species names at both global and country level, independent of language and cultural differences. The universality of this relationship suggests that the web salience of scientific species names is a robust, cross-cultural indicator of species 'culturalness'. Potential applications of this indicator include: i) the development of new indicators to assess public perceptions of biodiversity; ii) systematic identification of species with high cultural visibility; iii) empirical identification of the

biogeographic, ecological, morphological and cultural characteristics of species that influence cultural visibility, globally and in different cultural settings, and; iv) near real-time monitoring of changes in species 'culturalness'. The capture and processing of internet data is technically non-trivial, but can be replicated at low cost and has enormous potential for the creation of new macro-scale metrics of human-nature interactions.

1. Introduction

Social factors are well known to play a key-role in the success of conservation actions (Bennett et al., 2017a; Bennett et al., 2017b; Ehrlich, 2002; Mascia et al., 2003). Public awareness, perceptions, attitudes and engagement with nature and biodiversity can all have a significant influence on the final outcome of conservation efforts (Fischer and Young, 2007; Novacek, 2008). Quantifying and mapping variations in the public perception of biodiversity therefore has the potential to positively contribute to diverse conservation interventions (Jepson and Barua, 2015; Nghiem et al., 2016; Roll et al., 2016; Veríssimo et al., 2014). Indeed, the importance of monitoring public awareness of biodiversity is increasingly being recognized at all scales of conservation action, from local community projects to the development of international policy. For example, the first target of the Aichi Biodiversity Targets (agreed by the Convention on Biological Diversity in 2010) states that "By 2020, at the latest, people are aware of the values of biodiversity and the steps they can take to conserve and use it sustainably". Parties to the CBD are now expected to endeavour in efforts to increase awareness of biodiversity and its values. But how can the progress towards this target be assessed, particularly at the global scale?

The internet, with its enormous and increasing geographical and demographic reach, provides novel opportunities to develop large-scale quantitative metrics of public interest in and visibility of biodiversity (Ladle et al., 2016). Such an approach requires the adoption of 'big data' methods (Hampton et al., 2013), inferring human interest and sentiment towards the environment from the digital representation of words and images. The formal study of human culture through the analysis of changes in word frequencies in large bodies of texts (*corpora*) is known as *culturomics* (Michel et al., 2011). In a culturomic context, the frequency at which web-sites mention the names of species

(hereafter referred to as *internet salience*) can be used as a metric of cultural visibility/interest (Correia et al., 2016; Żmihorski et al., 2013) (hereafter referred to as the property of '*culturalness'*).

The validity of such a metric rests on the assumption that web content broadly reflects the interests, concerns and everyday lives of the human population that generates it. At the country level, there is strong evidence that this is the case. For example, Correia et al. (2016) demonstrated that internet salience of common names (in Portuguese) of species belonging to four highly visible groups of Brazilian birds (toucans, woodpeckers, hummingbirds, parrots) was most strongly associated with metrics of familiarity such as the size of the human population within the species' geographic range. Likewise, Schuetz et al. (2015) observed that internet searches for the common names of 68 resident bird species in the USA were positively associated with estimates of their population densities. Such studies strongly suggest that internet salience of a species can be broadly considered as an indicator of its cultural visibility. Familiarity is only one of several factors that contribute to cultural visibility, which is a product of the interaction between a species' phenotypic and biogeographic traits, and the attitudes, values and culture of the publics with which it interacts (Correia et al., 2016; Ducarme et al., 2013; Jepson and Barua, 2015; Lorimer, 2007).

Despite the promising results of these initial studies, expanding the use of metrics of species 'culturalness' based on the internet salience across cultural and languages barriers poses a significant technical challenge. This is because vernacular names often have loose and/or multiple meanings; in such cases, searches for species names will return results for other cultural entities as well as the biological species. For example, in English the word 'teal' is the vernacular name for a genre of small duck (*Anas crecca*), but since the 1920s it has also been used to refer to a popular shade of bluish/green used in clothes

and paints. Another example is the word ‘jaguar’, a theronym for the South American felid (*Panthera onca*) and, since the 1940s, a luxury car brand. Clearly, searching for ‘teal’ or ‘jaguar’ in available digital corpora would generate considerable ‘noise’ relating to, respectively, the popular colour and the aspirational car brand (Ladle et al., 2016). Furthermore, comparisons of relative internet salience in countries with different languages would inevitably produce significant and unavoidable biases.

Clearly, linguistic variability represents a significant challenge for generating universal metrics of species ‘culturalness’ based on the internet salience of vernacular names. However, this challenge could be largely circumvented if a strong relationship exists between the frequency of occurrence of vernacular and scientific species names in the internet corpus. This is because scientific nomenclature is universal, uses a ‘dead’ language (Latin) and a scientific name refers exclusively (with very few exceptions) to the biological organism. Here, we test the hypothesis that the internet salience of scientific species names and vernacular species names is highly correlated and independent of which country hosts the internet sites or the language used to generate the web content. The degree to which this hypothesis holds true and the characteristics of existing outliers provide an assessment of the robustness of deploying the internet salience of scientific names as a cross-cultural indicator of species ‘culturalness’.

2. Material and methods

The inherent properties of ‘big data’, such as volume and velocity, generate exciting new opportunities for the study of social phenomena (Kitchin, 2013; Ruppert, 2013). However, such properties also bring about new challenges for researchers. These include the development of advanced analytical skills and a critical understanding of social-technological system through which data is produced (Kitchin, 2014). The three main

139 sequential challenges (see Ladle et al., 2016) arising from the application of culturomic
140 approaches to environmental and nature conservation issues are: i) identifying the most
141 appropriate digital corpora to answer the research question; ii) obtaining data from the
142 selected digital corpus or corpora, and; iii) analysing the data. Here, we tackle the
143 challenge of dealing with language variability during data retrieval and analysis.
144 Specifically, we address the problem of "onyms" that create noise and bias in culturomic
145 samples (Tzanis, 2015). While many types of "onyms" occur in digital corpora referring
146 to species, most common examples include synonyms, homonyms and theronyms (see
147 Table 1).

148 We use data extracted from the World Wide Web, which is largely comprised of
149 web-sites and blogs. We chose this digital corpus due to its wide geographical reach and
150 because it is increasingly being used by scientists to investigate relationships between
151 environment, society and culture (Ladle et al., 2016). Furthermore, the varied nature of
152 the contributions that compose this corpus may reduce potential biases associated with
153 the predominance of scientific language in other digital corpora (such as Google Scholar
154 or Web of Science) and the use of colloquial language that predominates on social media
155 and microblogging corpora derived from platforms such as Facebook or Twitter (Giustini
156 and Wright, 2009).

157 There are over one billion registered web-sites worldwide (Internet Live Stats,
158 2016) which most internet users access through search engines such as Google Search or
159 Microsoft Edge. We extracted data from the World Wide Web using Google's Search
160 Engine, which currently claims over 70% of the global search engine market share
161 (NetMarketShare, 2016). Much of Google Search Engine's success is attributable to its
162 personalization algorithms that filter and rank the most relevant search results to the users
163 based search history and location (Dou et al., 2007; Hannak et al., 2013; Kliman-Silver

et al., 2015). Whilst such algorithms benefit the general user, they pose a considerable problem for researchers seeking replicability. To address this problem, we took advantage of Google's Custom Search Engine API (Application Programming Interface) which allows users to carry out repeated searches under the same specifications, increasing replicability and standardizing data retrieval.

Searches were carried out globally, using a list of worldwide bird species obtained from the International Union for the Conservation of Nature Red List of Threatened Species (IUCN, 2015), and at the country level using national bird species lists obtained from Avibase (Lepage, 2015). Two types of searches were carried out using either the vernacular or the scientific names of the species as search strings. All searches were carried out by using quoted search strings (e.g. "European Robin", "*Erithacus rubecula*"), restricting results to exact matches of the search string. For the global search, vernacular species names were searched in English as it is the most represented language on the internet (Ronen et al., 2014; Web Technology Surveys, 2016). Country level searches were carried out for six countries – Australia, Brazil, Indonesia, Spain, Tanzania and USA – and results were restricted to country web domains only. The most represented language in each country's web domains (English for Australia, Tanzania and USA, and Portuguese, Spanish and Indonesian for Brazil, Spain and Indonesia, respectively) was used for vernacular species name searches at the country level. All searches were carried out during March 2016 and the number of web-pages returned by the search was log-transformed and used as a metric of internet salience (Correia et al., 2016; Żmihorski et al., 2013).

The relationship between the internet salience of vernacular and scientific species names was assessed using Pearsons's product moment correlation coefficient. We also analysed the influence of two country-level characteristics that could influence the

189 relationship between vernacular and scientific species names: species richness (number
190 of species occurring in a given country) and internet penetration (the proportion of the
191 population that has access to the internet). Generalized linear models (GLMs) with
192 Gaussian distribution and identity-link function were used to explore how Pearson's
193 correlation scores were associated with bird species diversity and internet penetration at
194 the country level.

195 Finally, we carried out outlier detection to identify and characterize deviations
196 from the observed relationship at both the global and country level. Outlier identification
197 was carried out using function *uni.plot* in package *mvoutlier* (Filsmozer and
198 Gschwandtner, 2015) available for R Software, and considered two dimensions: i) the log
199 sum of web-sites mentioning the vernacular and scientific name, and; ii) the difference
200 between log sum of web-sites mentioning the vernacular name and the log sum of web-
201 sites mentioning the scientific name.

202 Outliers were classified in one of five qualitative categories:

203 **Category 1**: species with more than one commonly used vernacular name (*vernacular*
204 *polyonymous species*).

205 **Category 2**: species for which vernacular names have more than one meaning
206 (*vernacular homonymous species*).

207 **Category 3**: species that have undergone recent taxonomic revision or that have been
208 recently discovered (*taxonomically revised species*).

209 **Category 4**: species with a vernacular name score above the 95^{th} percentile that do not
210 show clear linguistic or taxonomic biases (*super-salient species*).

211 **Category 5**: species that do not fit any of the above criteria (*other outliers*).

212 Chi-square tests were used to assess significant differences in the occurrence of
213 different outliers at the global and country level. All analyses were carried out using R

214 Software, and figures were produced in the same software using the *ggplot2* graphics
215 package (Wickham, 2009).
216
217 **3. Results and Discussion**
218 *3.1. Relationship between the internet salience of vernacular and scientific species names*
219 Global searches using vernacular species names returned an average of approximately 11
220 000 webpages mentioning each bird species (mean = 10 873.1, std. error = 4 372.7; Table
221 2). Searches using scientific names returned a much lower number of webpages,
222 averaging around 1 600 webpages (mean = 1 623.9, std. error = 48.5). Nevertheless, we
223 found a strong and significant correlation between vernacular species names and
224 scientific species names at the global level (Pearson's r=0.775, p-value<0.001; Fig. 1).
225 This indicates that the proportion of webpages mentioning a species scientific name out
226 of the total universe of websites mentioning the species remains somewhat constant
227 across species.
228 At the country level, our searches highlighted large differences between species
229 representation across the countries sampled (Table 2). The USA were by far the country
230 where bird species were most represented on the internet, both in terms of vernacular
231 names (mean = 11 333.4, std. error = 2 008.7) and scientific names (mean = 2 379.0, std.
232 error = 102.8). At the opposite end of the spectrum we find Tanzania, where most species
233 were represented in very few webpages through vernacular names (mean = 7.0, std. error
234 = 0.3) and scientific names (mean = 4.7, std. error = 0.1). The most likely explanation for
235 the low representation of bird species on Tanzanian webpages observed in our sample is
236 likely to be the fact that less than 5% of the population have internet access (see below).
237 The fact that Tanzania is a multilingual country (there are more than 100 languages

spoken in the country, with the most commonly spoken being Bantu Swahili and English) could also have contributed for this result.

Despite the large differences in the web representation of bird species between countries, a strong correlation between vernacular and scientific species names was also observed at the country level (Fig. 2). The strength of the relationship varies for the six countries analysed; Australia, Brazil, Spain and the USA show comparable and relatively high correlation scores (Pearson's $r \geq 0.600$, p-value < 0.001), whereas Indonesia (Pearson's $r = 0.469$, p-value < 0.001) and Tanzania (Pearson's $r \geq 0.497$, p-value < 0.001) show somewhat lower correlations. These results strongly corroborate recent research suggesting that species scientific names can be used as a proxy the representation of vernacular names on the internet (Jaric et al., 2016). The observed relationships at the country level also suggest that this relationship stands across different cultural and linguistic settings, and strongly supports the hypothesis that species representation in digital corpora can be broadly viewed as proxy of species' 'culturalness' (Correia et al., 2016).

3.2. Factors affecting the relationship between vernacular and scientific species name salience

We found no significant relationship (GLM; $\beta < 0.001$, std. error < 0.001, $p = 0.899$) between the number of species in either of the six study countries and the strength of the relationship between vernacular and scientific species name internet salience at the country level (Fig. 3a). However, there was a significant relationship (GLM; $\beta = 0.002$, std. error < 0.001, $p = 0.015$) between internet penetration and the strength of the vernacular name-scientific name relationship (Fig. 3b). This finding indicates that in countries with lower internet penetration there is more variability in the ratio of scientific

to vernacular name frequency in the country's digital corpora. Such a pattern probably reflects a simple sampling effect; the larger and more representative the internet population is in a given country, the more likely it is that scientific names and vernacular will appear together in the same web content on a random basis. For example, it is well known that digital corpora are more complete for the major language groups (Funk and Rusowsky, 2014; Ronen et al., 2014). Hence, despite the big data approach used in this work, large datasets can still be subject to sampling biases (Boyd and Crawford, 2012; Kitchin, 2014), and these are likely to be more prevalent in countries with relatively low internet penetration. Another plausible explanation for this observation could be that lower internet penetration may be associated with particular sectors of society being over- or under-represented on the internet (Graham et al., 2015). The extent of such biases is likely to decrease with time as internet access becomes more ubiquitous, but researchers should nevertheless be aware of potential implications and aim to account for them in culturomic research (Ladle et al., 2016).

3.3. Outlier analysis

Our outlier analysis and classification revealed four main categories of outlier clusters associated with different socio-linguistic characteristics of either the vernacular or scientific species names (Table 3). These clusters are clearly visible in the plots representing relationship between vernacular and scientific name salience (Figure 2) and are comprised of: i) vernacular polyonymous species; ii) vernacular homonymous species; iii) taxonomically revised species, and; iv) 'super-salient' species. Other outliers that do not fit any of the above criteria also exist in our data-sets (Category 5), and their outlier status is likely to be associated with cultural rather than linguistic dynamics.

Vernacular polyonymous species (Category 1) are species that are known by several or many common names (e.g. *Bittern*, *Great Bittern*, *Boomer*, *Butterbump*). For these species, unless all vernacular names are used in the automated search the vernacular name salience will be lower than expected according to vernacular-scientific name relationship (country or global). Such species are mostly clustered parallel to the scientific name axis in the plots (Fig. 4), and were particularly common in Australian (55.2% of outliers) and Indonesian (54.5% of outliers) data-sets (Table 3). Vernacular polyonymous species are caused by the introduction of common name variants into popular culture. This sometimes happens naturally, such as the evolution of regional dialects (e.g. in the UK the Lapwing *Vanellus vanellus* is still commonly referred to by the older term, 'peewit'). Moreover, new names have been introduced into popular discourse through the scientific standardization of vernacular names. These 'scientized' neologisms are frequently longer and more formal than those more frequently used by the public (e.g. 'European Blackbird' in place of 'Blackbird').

Vernacular homonymous species (Category 2) have higher than expected vernacular species name salience caused by multiple meanings of the common name in popular culture. For example, 'Teal' (see introduction) falls into this category. Such outliers most commonly occur in the upper diagonal of the plot (Fig. 4), and were particularly common in the Brazilian dataset (42.5% of outliers, Table 3) – possibly influenced by the relatively recent colonization of this geographic region by Portuguese speaking Europeans. 'Onyms' are relatively common in biological datasets (Tzanis, 2015) and can have different origins or causes (Table 1), but there is no simple *a priori* method to identify these.

Taxonomically revised species (Category 3) are outliers characterized by lower than expected scientific name salience. Such species are usually located parallel to the

312 vernacular name axis (Fig. 4), and their occurrence is associated with either new scientific
313 names (newly discovered species or taxonomic splits), or minor alterations to the spelling
314 of Latin names due to factors such as gender agreement (David and Gosselin, 2011).
315 Outliers caused by taxonomic revisions were particularly common in the Tanzanian
316 (50.0% of outliers) and US (46.9% of outliers) data-sets (Table 3). The rise of molecular
317 taxonomy is undoubtedly accelerating such revisions (Isaac et al., 2004) with a knock on
318 influence of representation in digital corpora.

319 Finally, species where the relationship between vernacular and scientific species
320 internet salience is strongly skewed towards vernacular names, but where there is no
321 evidence for bias deriving from multiple meanings, are classified as 'super-salient'
322 species (Category 4). These are species that are deeply culturally embedded and have
323 exceptional cultural visibility, and are located at the top right edge of the plot (Fig. 4).
324 Super-salient species were frequent in the global (21.2% of outliers) and Indonesian
325 (19.5% of outliers) data-sets (Table 2). Examples we identified include the Barn Owl *Tyto*
326 *alba*, a globally widespread, semi-nocturnal flying owl that lives in close proximity to
327 humans and has long played an important role in rat control, and the Yellow-crested
328 Cockatoo *Cacatua sulphurea*, a common pet species that is found in the wild across
329 Indonesia and East-Timor. Once again, a detailed analysis of the data would be required
330 to clearly distinguish this type of outlier from vernacular homonymous species.
331 Particularly, it may be necessary to sample species representations on the internet in more
332 than one language to obtain an accurate assessment of 'super-salient' species in
333 multilingual countries, such as Tanzania, as species salience may be particularly
334 associated with certain language groups.

335 The four outlier types identified above were by far the most prevalent (Table 3).
336 Their analysis may be used to deepen our understanding of the interactions between

biological species entities and different cultural contexts, and pave the way for novel metrics that can be used to assess public perceptions of biodiversity. For example, such analysis could be used to: i) identify iconic and well-known species for use in conservation outreach and marketing. In the broadest context, scientific name internet saliency can be thought of as a macro-scale metric of cultural interest, capturing key components of cultural ecosystem services such as aesthetic enjoyment; ii) generate simple metrics of country-level taxonomic activity within taxa (taxonomically revised species), and; iii) track the cultural dynamics of particular species, for example after rediscovery, extinction or a notable conservation intervention (Ladle et al., 2016).

Conclusions

We show a strong and consistent relationship between vernacular and scientific names of bird species at a global and national scale, largely independent of cultural context and web-site language. This relationship is more robust for countries with high levels of internet content generation due to sampling size effects. Given the strength and universality of this relationship, we tentatively conclude that internet salience of scientific names can be used as a cross-cultural indicator of species 'culturalness'. Assuming the validity of this indicator, there are many potential benefits for conservation practice and research, including: i) the development of new macro-scale indicators to assess public perceptions of biodiversity; ii) systematic identification of species with high cultural visibility; iii) empirical identification of the biogeographic, ecological, morphological and cultural characteristics of species that influence cultural visibility, globally and in different cultural settings, and; iv) near real-time monitoring of changes in species 'culturalness'.

Furthermore, important information can be gained from examining the relationship between scientific and vernacular species name internet salience. Especially important in this respect is the identification of 'super-salient' species, with levels of cultural visibility far higher than predicted based on their scientific name saliency. A detailed analysis of these species can provide a deeper understanding of the factors driving cultural perceptions of species and the fluid processes through which they become embedded in human culture. Ultimately, the further development of culturomic indices applied to conservation can greatly contribute towards the study of human-nature interactions and the monitoring of global conservation objectives such as the Aichi Biodiversity Targets.

Acknowledgements

This work was funded by the Brazilian National Council for Scientific and Technological Development CNPq: RAC (grants #158841/2015-8 & #400325/2014-4), PJ (#400325/2014-4), ACCM (#310349/2015-0), RJL (#310953/2014-6). We are also thankful to Berta Martín-López and one anonymous referee, whose comments helped to improve the manuscript.

References

Bennett, N.J., Roth, R., Klain, S.C., Chan, K., Christie, P., Clark, D.A., Cullman, G., Curran, D., Durbini, T.J., Epstein, G., Greenberg, A., Nelson, M.P., Sandlos, J., Stedman, R., Teel, T.L., Thomas, R., Veríssimo, D., Wyborn, C., 2017a. Conservation social science: Understanding and integrating human dimensions to improve conservation. Biol Conserv 205, 93-108.

Bennett, N.J., Roth, R., Klain, S.C., Chan, K.M.A., Clark, D.A., Cullman, G., Epstein, G., Nelson, M.P., Stedman, R., Teel, T.L., Thomas, R.E.W., Wyborn, C., Curran, D., Greenberg, A., Sandlos, J., Veríssimo, D., 2017b. Mainstreaming the social sciences in conservation. Conserv Biol 31, 56-66.

Boyd, D., Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society 15, 662–679.

Correia, R.A., Jepson, P.R., Malhado, A.C.M., Ladle, R.J., 2016. Familiarity breeds content: assessing bird species popularity with culturomics. Peerj 4, e1728.

David, N., Gosselin, M., 2011. Gender agreement of avian species-group names under Article 31.2. 2 of the ICZN Code. Bulletin of the British Ornithologists' Club 131, 102-115.

Dou, Z., Song, R., Wen, J.-R., 2007. A large-scale evaluation and analysis of personalized search strategies, Proceedings of the Sixteenth International World Wide Web Conference, Banff, Alberta, Canada, pp. 581-590.

Ducarme, F., Luque, G.M., Courchamp, F., 2013. What are "charismatic species" for conservation biologists. BioSciences Master Reviews 10, 1-8.

Ehrlich, P.R., 2002. Human natures, nature conservation, and environmental ethics. Bioscience 52, 31-43.

403 Filsmozer, P., Gschwandtner, M., 2015. mvoutlier: Multivariate outlier detection based
404 on robust methods. R package version 2.0.6.
405 Fischer, A., Young, J.C., 2007. Understanding mental constructs of biodiversity:
406 Implications for biodiversity management and conservation. Biol Conserv 136, 271-
407 282.
408 Funk, S.M., Rusowsky, D., 2014. The importance of cultural knowledge and scale for
409 analysing internet search data as a proxy for public interest toward the environment.
410 Biodivers Conserv, 1-12.
411 Giustini, D., Wright, M.-D., 2009. Twitter: an introduction to microblogging for health
412 librarians. Journal of the Canadian Health Libraries Association (JCHLA) 30, 11-17.
413 Graham, M., De Sabbata, S., Zook, M.A., 2015. Towards a study of information
414 geographies: (im)mutable augmentations and a mapping of the geographies of
415 information. Geo: Geography and Environment 2, 88-105.
416 Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller,
417 A.L., Duke, C.S., Porter, J.H., 2013. Big data and the future of ecology. Front Ecol
418 Environ 11, 156-162.
419 Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D.,
420 Mislove, A., Wilson, C., 2013. Measuring personalization of web search, Proceedings
421 of the Twenty-Second International World Wide Web Conference, Rio de Janeiro,
422 Brazil, pp. 527-538.
423 Internet Live Stats, 2016. http://www.internetlivestats.com/, accessed on 09/05/2016.
424 Isaac, N.J., Mallet, J., Mace, G.M., 2004. Taxonomic inflation: its influence on
425 macroecology and conservation. Trends in Ecology & Evolution 19, 464-469.
426 IUCN, 2015. The IUCN Red List of Threatened Species, Version 2015.
427 <www.iucnredlist.org>. Downloaded on 09 May 2015.

428 Jaric, I., Courchamp, F., Gessner, J., Roberts, D.L., 2016. Data mining in conservation
429 research using Latin and vernacular species names. Peerj 4, e2202.
430 Jepson, P., Barua, M., 2015. A Theory of Flagship Species Action Conservation &
431 Society 12.
432 Kitchin, R., 2013. Big data and human geography: Opportunities, challenged and risks.
433 Dialogues in Human Geography 3, 262-267.
434 Kitchin, R., 2014. Big Data, new epistemologies and paradigm shifts. Big Data &
435 Society 1, 2053951714528481.
436 Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., Mislove, A., 2015. Location,
437 Location, Location: The Impact of Geolocation on Web Search Personalization,
438 Proceedings of the 2015 ACM Conference on Internet Measurement Conference,
439 Tokyo, Japan, pp. 121-127.
440 Ladle, R.J., Correia, R.A., Do, Y., Joo, G.J., Malhado, A.C.M., Proulx, R., Roberge,
441 J.M., Jepson, P., 2016. Conservation Culturomics. Frontiers in Ecology and the
442 Environment In Press.
443 Lepage, D., 2015. Avibase, the World Bird Database. Published online
444 at http://avibase.bsc-eoc.org/.
445 Lorimer, J., 2007. Nonhuman charisma. Environment and Planning D 25, 911-932.
446 Mascia, M.B., Brosius, J.P., Dobson, T.A., Forbes, B.C., Horowitz, L., McKean, M.A.,
447 Turner, N.J., 2003. Conservation and the Social Sciences. Conserv Biol 17, 649-650.
448 Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg,
449 D., Clancy, D., Norvig, P., Orwant, J., 2011. Quantitative analysis of culture using
450 millions of digitized books. Science 331, 176-182.
451 NetMarketShare, 2016. http://www.netmarketshare.com/search-engine-market-
452 share.aspx?qprid=4&qpcustomd=0, accessed on 09/05/2016.

Nghiem, L.T.P., Papworth, S.K., Lim, F.K.S., Carrasco, L.R., 2016. Analysis of the Capacity of Google Trends to Measure Interest in Conservation Topics and the Role of Online News. Plos One 11.

Novacek, M.J., 2008. Engaging the public in biodiversity issues. Proceedings of the National Academy of Sciences 105, 11571-11578.

Roll, U., Mittermeier, J.C., Diaz, G.I., Novosolov, M., Feldman, A., Itescu, Y., Meiri, S., Grenyer, R., 2016. Using Wikipedia page views to explore the cultural importance of global reptiles. Biol Conserv 204, 42-50.

Ronen, S., Goncalves, B., Hu, K.Z., Vespignani, A., Pinker, S., Hidalgo, C.A., 2014. Links that speak: The global language network and its association with global fame. P Natl Acad Sci USA 111, E5616-E5622.

Ruppert, E., 2013. Rethinking empirical social sciences. Dialogues in Human Geography 3, 288-273.

Schuetz, J., Soykan, C.U., Distler, T., Langham, G., 2015. Searching for backyard birds in virtual worlds: Internet queries mirror real species distributions. Biodiversity & Conservation 24, 1147-1154.

Tzanis, G., 2015. Biological and Medical Big Data Mining, in: Khosrow-Pour, M., Clarke, S., Jennex, M.E., Becker, A., Anttiroiko, A.-V. (Eds.), Business Intelligence: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications. Business Science Reference, Hershey, PA, USA, pp. 246-262.

Veríssimo, D., Pongiluppi, T., Santos, M.C.M., Develey, P.F., Fraser, I., Smith, R.J., MacMilan, D.C., 2014. Using a systematic approach to select flagship species for bird conservation. Conservation biology 28, 269-277.

476 Web Technology Surveys,
477 2016. https://w3techs.com/technologies/overview/content_language/all, accessed on
478 08/07/2016.
479 Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag,
480 New York, USA.
481 Żmihorski, M., Dziarska -Pałac, J., Sparks, T.H
482 correlates of the popularity of birds and butterflies in Internet information resources.
483 Oikos 122, 183-190.

484

485 **Tables**

486 Table 1 – List of relevant ‘onyms’, their description and featured examples in our data-
487 sets.

Type of “onym”	Description	Example
Anthoponym	A name of a human being; as reflected in surnames or proper names of people	Robin – *Erithacus rubecula*
Aptronym	A name appropriate to its owner's occupation or physical properties, such as "Goldsmith" or "Longman"	Dipper – *Cinclus cinclus*
Eponym	A botanical, zoological, artwork, or place name that derives from a real or legendary person	Bonelli’s Warbler – *Phylloscopus bonelli*
Homonym	A word that is said or spelled the same way as another word but has a different meaning	Swift – *Apus sp.*
Odonym	A name of a street or road	Blackbird Leys – *Turdus merula*
Synonym	A word or phrase that means exactly or nearly the same as another word or phrase in the same language	Lapwing and peewit – *Vanellus vanellus*
Theronym	A name - especially a product name - that has been derived from the name of an animal	Teal – *Anas crecca*

488

489 Table 2 – Summary statistics of the number of webpages returned by global and country-

490 level scrapes using vernacular and scientific species names.

Search scope	Number of species	Vernacular name webpages		Scientific name webpages	
		Mean	Std. Error	Mean	Std. Error
Australia	954	574.5	38.3	175.6	6.7
Brazil	1 787	502.0	115.6	147.8	6.0
Indonesia	1 582	145.8	52.2	51.1	10.2
Spain	615	832.0	41.1	1 139.1	84.1
Tanzania	1 028	7.0	0.3	4.7	0.1
USA	1 134	11 333.4	2 008.7	2 379.0	102.8
Global	10 423	10 873.1	4 372.7	1 623.9	48.5

491

492 Table 3 – Percentage of outliers classified as vernacular polyonymous species (species
493 with multiple vernacular names), vernacular homonymous species (vernacular names
494 with more than one meaning), taxonomically revised species, super-salient species and
495 other outliers. Significant differences in the occurrence of different outlier categories
496 were found for all countries and at the global level according to Chi-square tests ($p <$
497 0.001).

Search scope	Outlier classification				
	Polyonymous Species	Homonymous Species	Taxonomically Revised Species	Super-salient Species	Other
Australia	55.2%	0.6%	31.8%	2.4%	10.0%
Brazil	10.6%	42.5%	35.8%	3.3%	7.8%
Indonesia	54.5%	4.1%	19.5%	19.5%	2.4%
Spain	38.7%	1.1%	31.2%	0%	29.0%
Tanzania	18.1%	0%	50.0%	11.2%	20.7%
USA	33.3%	11.1%	46.9%	3.7%	5.0%
Global	9.1%	4.8%	48.3%	21.2%	16.6%

498

Figures

Figure 1 – Global relationship between vernacular and scientific species name salience. Axes represent the log number of internet hits.

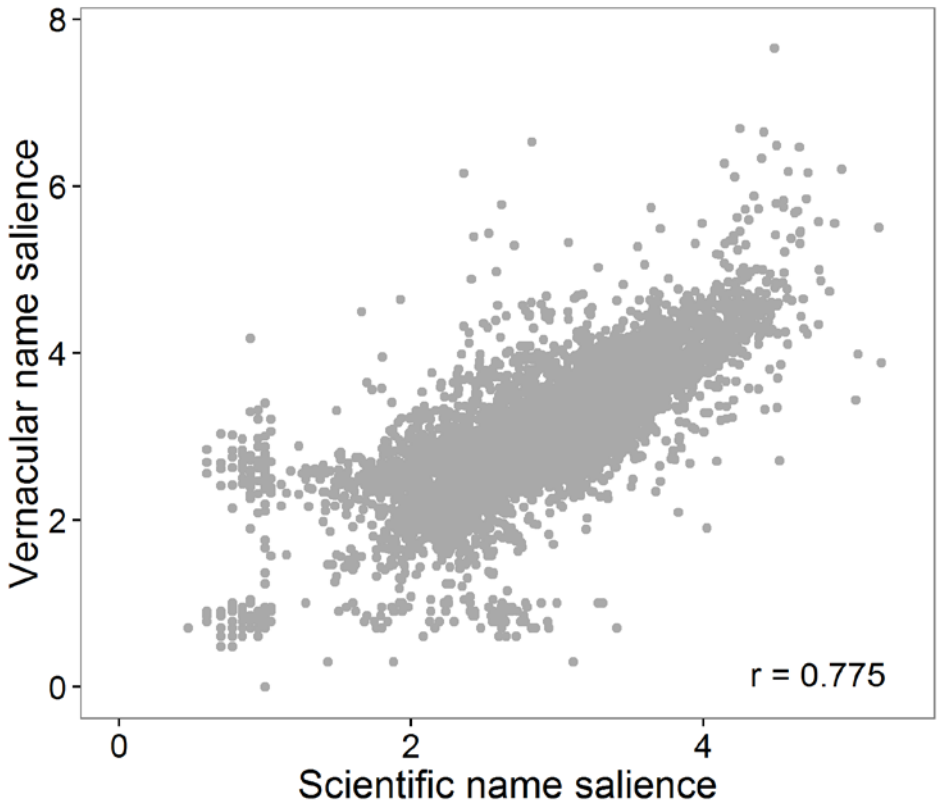


Figure 2 – Country-level relationship between vernacular and scientific species name salience. Axes represent the log number of internet hits.

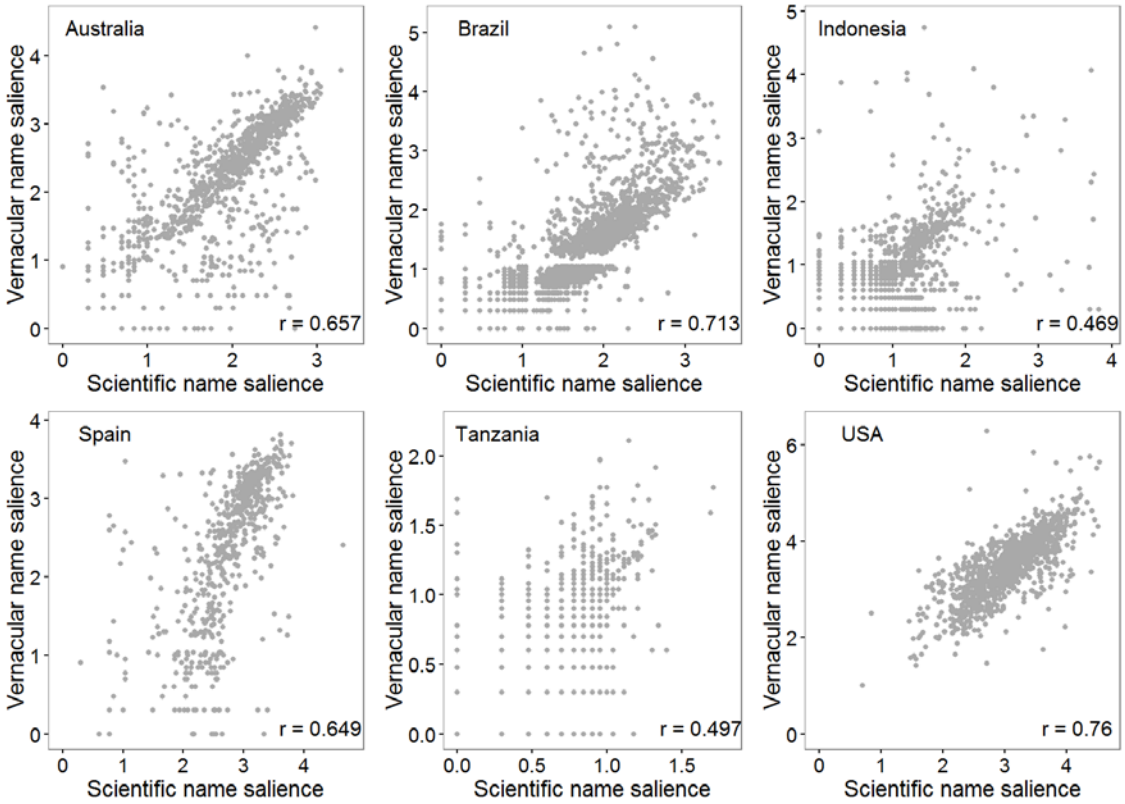
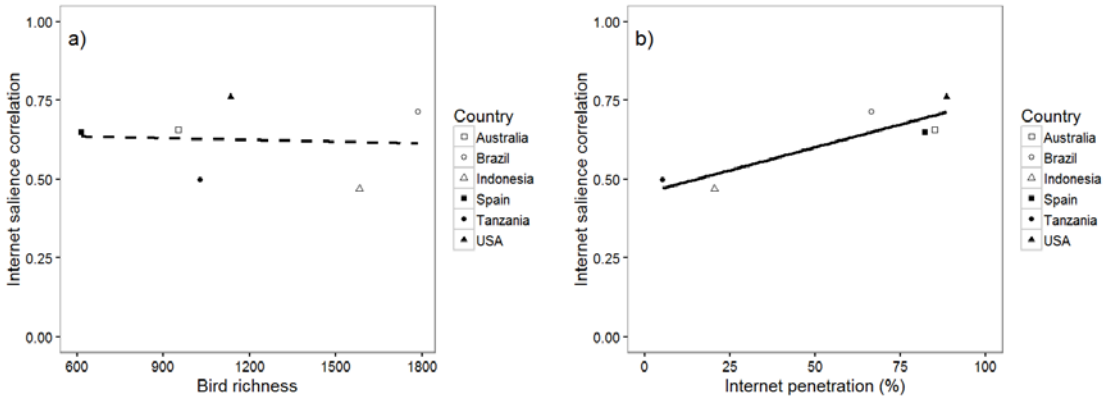
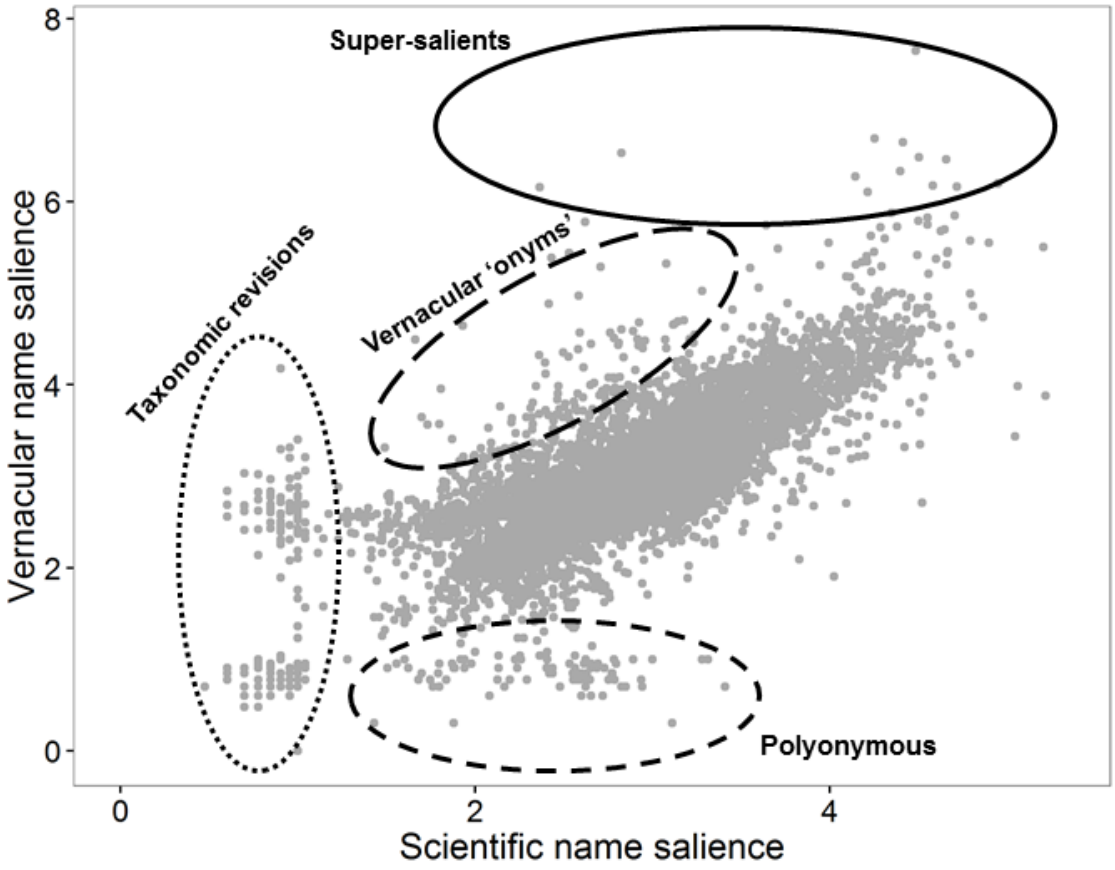


Figure 3 – Association between the strength of the relationship among vernacular and scientific species name internet salience (measured by Pearson's *r* correlation) and two country-level variables: species richness (a) and internet penetration (b).



513 Figure 4 – Representation of the four groups of outliers identified in the global dataset,
514 including polyonymous species (long dashed line), super salient species (solid line),
515 taxonomic name revisions (dotted line) and vernacular name 'onyms' (short dashed line).



516