Ecological Indicators 78 (2017) 549–555

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original Articles

Internet scientific name frequency as an indicator of cultural salience of biodiversity

Ricardo A. Correia[a,b,*], Paul Jepson[b], Ana C.M. Malhado[a], Richard J. Ladle[a,b]

[a] Institute of Biological and Health Sciences, Federal University of Alagoas, Av. Lourival Melo Mota, s/n, Tabuleiro do Martins, 57072-90, Maceió, AL, Brazil
[b] School of Geography and the Environment, University of Oxford, Oxford OX1 3QY, United Kingdom

ARTICLE INFO

Article history:
Received 7 November 2016
Received in revised form 29 March 2017
Accepted 29 March 2017
Available online 4 April 2017

Keywords:
Biodiversity
Birds
Culturomics
Internet salience
Public perception
Scientific names
Vernacular names

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

[*] Corresponding author at: Institute of Biological and Health Sciences, Federal University of Alagoas, Av. Lourival Melo Mota, s/n, Tabuleiro do Martins, 57072-90, Maceió, AL, Brazil.
E-mail address: rahc85@gmail.com (R.A. Correia).

http://dx.doi.org/10.1016/j.ecolind.2017.03.052
1470-160X/© 2017 Elsevier Ltd. All rights reserved.

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

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

There are over one billion registered web-sites worldwide (Internet Live Stats, 2016) which most internet users access through search engines such as Google Search or Microsoft Edge. We extracted data from the World Wide Web using Google's Search Engine, which currently claims over 70% of the global search engine market share (NetMarketShare, 2016). Much of Google Search Engine's success is attributable to its personalization algorithms that filter and rank the most relevant search results to the users 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