Using Wikipedia to measure public interest in biodiversity and conservation

John C. Mittermeier, Ricardo Correia, Rich Grenyer, Tuuli Toivonen and Uri Roll

1School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, U.K.
2American Bird Conservancy, 4301 Connecticut Avenue NW, Washington, DC, 20008, U.S.A.
3The Digital Geography Lab, Department of Geosciences and Geography, University of Helsinki, Helsinki, 00014, Finland
4Helsinki Lab of Interdisciplinary Science (HELICS), University of Helsinki, Helsinki, 00014, Finland
5Mitrani Department of Desert Ecology, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, 8499000, Israel

Abstract: The recent growth of online big data offers opportunities for rapid and inexpensive measurement of public interest. Conservation culturomics is an emerging research area that uses online data to study human-nature relationships for conservation. Methods for conservation culturomics, though promising, are still being developed and refined. We considered the potential of Wikipedia, the online encyclopedia, as a resource for conservation culturomics and outlined methods for using Wikipedia data in conservation. Wikipedia’s large size, widespread use, underlying data structure, and open access to both its content and usage analytics make it well suited to conservation culturomics research. Limitations of Wikipedia data include the lack of location information associated with some metadata and limited information on the motivations of many users. Seven methodological steps to consider when using Wikipedia data in conservation include metadata selection, temporality, taxonomy, language representation, Wikipedia geography, physical and biological geography, and comparative metrics. Each of these methodological decisions can affect measures of online interest. As a case study, we explored these themes by analyzing 757 million Wikipedia page views associated with the Wikipedia pages for 10,099 species of birds across 251 Wikipedia language editions. We found that Wikipedia data have the potential to generate insight for conservation and are particularly useful for quantifying patterns of public interest at large scales.

Keywords: bird conservation, conservation culturomics, flagship species, online encyclopedias, public engagement, Wikipedia

La Wikipedia como Instrumento de Medición del Interés Público por la Biodiversidad y la Conservación

Resumen: El crecimiento reciente de los datos masivos en línea ofrece oportunidades para la medición rápida y asequible del interés público. La culturomia de la conservación es un área emergente de investigación que utiliza la información en línea para estudiar las relaciones entre el humano y la naturaleza y usarlas para la conservación. Los métodos de conservación basados en culturomia, aunque prometedores, todavía están siendo desarrollados y refinados. Consideramos el potencial de Wikipedia, la enciclopedia en línea, como recurso para la culturomia de la conservación y los métodos para usar sus datos en la conservación. El gran tamaño de Wikipedia, su uso extenso, estructura subyacente de datos y acceso abierto tanto a su contenido como a sus análisis de uso hacen que sea muy adecuada para usarse en la investigación de culturomia de la conservación. Las limitantes de usar la información de Wikipedia incluyen la falta de ubicación de la información asociada con algunos metadatos y la información limitada sobre los motivos de muchos usuarios. Hay siete pasos metodológicos a considerar cuando se usa la información de Wikipedia para la conservación: la selección de metadatos, temporalidad, taxonomía, representación del idioma, geografía de la Wikipedia, geografía física y biológica y medidas comparativas. Cada una de estas decisiones metodológicas puede afectar a las medidas del interés en línea. Como estudio de caso, exploramos estos temas analizando 757 millones de vistas de páginas en Wikipedia para las páginas sobre 10,099 species of birds.
species of aves a través de 251 ediciones de Wikipedia en idiomas diferentes. Encontramos que la información de Wikipedia fue particularmente útil para cuantificar los patrones de interés público a grandes escalas y tiene el potencial para generar conocimiento para la conservación.

**Palabras Clave:** conservación de aves, culturomia de la conservación, enciclopedias en línea, especie bandera, participación pública, Wikipedia

**摘要:** 在线大数据的增长为快速简便地衡量公众兴趣提供了机会。保护文化组学是一个新兴的研究领域，利用在线数据来研究保护中人与自然之间的关系。保护文化组学的方法虽然很有前景，但仍在开发和完善当中。本研究探索了将在线百科全书——维基百科用作保护文化组学资源的潜力以及将维基百科数据用于保护的方法。维基百科庞大的规模、广泛的应用、底层数据结构，以及对其内容和使用分析的开放访问，使其非常适合于保护文化组学研究。而维基百科数据的局限性包括缺少某些元数据的位置信息，以及许多用户在动机的信息有限。在使用维基百科数据进行保护时需要考虑七个方法步骤，包括元数据选择、时间性、生物分类、语言代表性、维基百科地理、物理和生物地理，以及比较指标。对这些方法的选择会影响对在线兴趣的衡量。作为案例研究，我们还分析了维基百科251种语言版本中10,099种鸟类7.57亿次的浏览情况来探究这些主题。我们发现，维基百科的数据对于大尺度量化公众兴趣格局有很大帮助，并且有产生保护见解的潜力。【翻译：胡怡思 审校：聂永刚】

**关键词:** 鸟类保护, 保护文化组学, 旗舰种, 在线百科, 公众参与, 维基百科

**Introduction**

The importance of assessing public interest in biodiversity has been recognized by conservationists for decades (e.g., Manfredo 1989). However, measuring public interest across large numbers of people using traditional methodologies is expensive, time consuming, and frequently infeasible. Recently, new digital data archives have enabled quantitative comparisons at scales that were unimaginable only a few years ago, and these digital big data can often be analyzed rapidly and inexpensively. In addition to offering opportunities, digital big data also present significant methodological and interpretative challenges (Kitchin 2014).

Conservation culturomics is an emerging research area in which digital data are used to study human–nature interactions, including public interest in nature and conservation (Ladle et al. 2016). Previous researchers have used conservation culturomic methods to compare public interest in aspects of biodiversity (e.g., Correia et al. 2016; Roll et al. 2016). Although these approaches are promising, methods for conducting culturomic analyses in conservation are still being developed (Ladle et al. 2016; Sutherland et al. 2018; Correia et al. 2019; Toivonen et al. 2019).

A variety of digital data sets can be used in conservation culturomics, each enabling investigations of different content and forms of engagement with nature (Correia et al. 2021). Wikipedia, the online encyclopedia, has several features that make it particularly useful for comparing aspects of public interest at large scales. It is extremely popular. As of 2019, Wikipedia is the 10th most-visited site on the internet (Alexa 2019), and it receives upwards of 16 billion page views per month across its associated projects (Zachte 2019). Wikipedia has wide cultural, geographical, and thematic coverage; it currently includes 310 language editions and over 220 million total pages (Wikipedia 2020a). These include thousands of pages for biodiversity-related topics. Wikipedia has an organized structure that allows for comparisons across large numbers of topics and languages within the encyclopedia and frequently links to outside data structures, such as structured taxonomies. Wikipedia is fully open access with raw data freely available to researchers, and its terms of access are stable and community driven. Of the 10 most visited sites on the internet in 2019, Wikipedia is the only one to allow this open access (Alexa 2019). Wikipedia is the subject of a growing body of existing research that explores its content (Messner & DiStaso 2013; Samoilenko & Yasseri 2014), contributor demographics (Wilson 2014), and user dynamics (Yasseri et al. 2012, 2014). Previous researchers have used Wikipedia to quantitatively compare the fame and cultural impact of individual people (Skiena & Ward 2014; Yu et al. 2016) and established a precedent that Wikipedia data can be used to measure aspects of public interest in conservation (Roll et al. 2016; Mittemeier et al. 2019).

We devised methods for using Wikipedia data to quantitatively assess public interest in conservation. As a case study, we used Wikipedia to compare interest in 10,099 bird species across 251 different languages. We hope our method will facilitate the use of Wikipedia and other culturomic resources in conservation research.

**Methods**

We identified 7 methodological considerations for the use of Wikipedia data to compare public interest in the context of conservation (Table 1).
Table 1. A methodological framework for using Wikipedia data for conservation culturomics research.

| Research step | Action | Consider |
|---------------|--------|----------|
| 1 Metadata selection. What online interactions are of interest? | Select metadata type. | Motivations behind some metadata types can be hard to ascertain. Metadata vary in quantity and in the influence of bots. Aggregating different types of metadata may not make sense. |
| 2 Temporal variation. What is the relevant time frame? | Identify appropriate time frames. | Aspects of the data structure may limit availability (e.g., page views were redefined in 2015). Seasonal patterns and brief spikes in activity can influence results. Wikipedia is constantly increasing and revising its content. |
| 3 Taxonomy. What entities should be included? | Select taxonomy and consider limitations of the taxonomic choice. | Taxonomic lists vary in their degree of integration with Wikipedia. Taxonomic differences can influence results. Activity in Wikipedia may not align with taxonomic units. |
| 4 Language representation. What languages should be included? | Select Wikipedia language editions. | There is huge variation in the size and usage of language editions. People may interact with Wikipedia pages in languages other than their spoken language. |
| 5 Wikipedia geography. What is the distribution of languages and users? | Review the distribution of selected languages and their users. | Wikipedia use is biased toward Europe and North America. The distribution of a language’s Wikipedia users may differ from the distribution of its speakers. Some language editions have more clearly defined geographic distributions than others. |
| 6 Physical and biological geography. What is the distribution of the entities being compared? | Review the distribution of selected entities and assess the influence of geographic overlap. | People are often more interested in things that are local. Entities that overlap with the distribution of Wikipedia users are likely to be overrepresented. |
| 7 Comparative metrics. What metrics should be used to aggregate data from multiple languages or metadata types? | Identify appropriate metrics if data from multiple languages or metadata types are being used. | Decisions to scale or not scale data in comparative metrics can affect results. Appropriate scaling methods will vary depending on the research question. |

**Metadata Selection**

Wikipedia pages have a wide array of metadata that can be quantified and compared across pages. As of 2019, each Wikipedia page includes over 30 attributes relating to its size, edit history, edit frequency, and links to other pages in Wikipedia (Wikimedia 2020a). These metadata reflect distinct forms of online engagement and different communities of users. For example, the editing and writing of Wikipedia articles is content generation, and the viewing of Wikipedia pages is content consumption (Correia et al. 2021). These edits and page views may be more comparable to information in other data sources than they are to one another (e.g., Wikipedia page views could be compared with article reads of online newspapers, another form of content consumption). Metadata also vary in the quantity of interactions they contain. For example, there were 890 million edits versus 480 billion page views to English Wikipedia from 2016 to 2020 (Wikimedia 2020b). Some metadata can be influenced by the activity of bots, automated programs that edit and contribute data to Wikipedia, and the ability of researchers to remove bot activity varies between metadata types. As a result of these differences, Wikipedia metadata accrue content differently and generate different measures of online interest. Ultimately, certain types of metadata will be useful for answering particular questions, such as edit histories being reflective of controversial pages (Yasseri et al. 2012).
Wikipedia page views have advantages over other metadata in measuring public interest. They reflect a distinct type of interaction (seeking information about a subject). They capture the actions of the widest community of users and contain the largest quantity of user interactions (e.g., 480 billion page views vs. 890 million edits). Page views also have a published precedent for being used to compare cultural relevance and public interest (e.g., Correia et al. 2018). Wikipedia’s reclassification of its page views in 2015 allows for differentiation of user as opposed to bot-generated views. In some cases, it may still be possible to manipulate Wikipedia page views with automated programs, but these are rare (Wikipedia 2020b). Wikipedia page views also have limitations. They do not capture sentiment (i.e., it is not possible to distinguish whether a viewer reached a page because they felt positively or negatively about a subject) or reflect total unique visitors (many page views may be generated by repeated visits from a single user). Due to Wikipedia’s privacy policy, page views do not include precise location data, making it difficult to identify viewers’ locations. Summary data that provide the proportion of page views to each Wikipedia language by country are available (Zachte 2020), and a page’s language can be used as a coarse proxy for its geography (Generous et al. 2014; Mittermeier et al. 2019). This approach has important limitations, however, and is not useful for assessing patterns at fine geographic scales.

Temporal Variation

As an open-access, user-generated resource, Wikipedia is constantly being updated and revised. As of December 2019, Wikipedia had received 45 million edits and was gaining 80 GB of content per month (Wikimedia 2020b). To be reproducible, analyses should identify the date a list of pages was accessed as well as the time frames over which the metadata associated with those pages was collected. In addition to its overall patterns of growth, Wikipedia activity can follow seasonal patterns (Mittermeier et al. 2019) and undergo short bursts of attention due to events, such as the release of a popular film or the death of a prominent public figure (Wikipedia 2020c). Although identifying seasonality or short bursts of interest will be relevant for some questions (e.g., assessing the impact of a publicity campaign or conservation debate), in other cases, researchers may want to identify topics that attract consistent attention. This can be done by extracting data over long, preferably multiyear periods or by using robust statistical measures.

Taxonomy

Which entities are included in a study can influence the outcome and accuracy of comparisons made with online data (Correia et al. 2018). These taxonomic decisions are important in Wikipedia, where public interest may not neatly match taxonomic boundaries. For example, interest in some groups of plants and animals may be higher at the subspecies or family level than at the species level, despite the latter being most frequently used in conservation planning. Taxonomies also differ in their degree of integration with Wikipedia and Wikidata (Wikipedia’s underlying structured database). Querying Wikidata for all entities marked with a Global Biodiversity Information Facility ID (Wikidata identifier P846) returned 2,153,907 entities as of June 2020. Queries for all entities tagged with an Integrated Taxonomic Information System identifier (Wikidata identifier P815) and an Encyclopedia of Life identifier (P830) on the same date returned 568,986 and 1,093,306 entities, respectively. In addition to these global taxonomies of organisms, Wikipedia and Wikidata include lists for specific taxonomic groups (e.g., from eBird and BirdLife International, FishBase, and Plants of the World) and regions (e.g., the Finnish Biodiversity Information Facility’s Species List, the New Zealand Organisms Register, and the Flora of North America). There are also taxonomies of geographic areas (the U.S. National Park System, UNESCO [United Nations Educational, Scientific and Cultural Organization] World Heritage Sites) and concepts (JSTOR topics, the UNESCO thesaurus). Different taxonomies will be appropriate for different research questions. From a methodological standpoint, however, taxonomic choices should be explicitly stated and justified.

Language Representation

Wikipedia currently contains more than 300 different language editions. These vary dramatically in the number of articles they contain: the smallest language editions have 0 articles (i.e., only an introductory main page), whereas English, the largest edition, has more than 6 million articles (Wikipedia 2020a). Language editions also vary in how frequently they are viewed and edited. The 10 most-viewed language editions account for approximately 88% of all page views (Zachte 2019). English alone receives 49% of all Wikipedia page views and 25% of Wikipedia edits (Zachte 2019; Wikipedia 2020a). It is worth noting that many English-language page views originate from countries where English is not the primary spoken language and thus many viewers to English Wikipedia probably do not speak English as their first language (Zachte 2020). Given these linguistic inequalities, multilanguage comparisons that do not adjust for variations in the size and usage of Wikipedia languages will be strongly influenced by a small subset of languages, in particular English. It is also important to keep in mind that the languages represented in Wikipedia are <5% of the over 7000 recognized languages currently spoken (Eberhard et al. 2019).
Wikipedia Geography

Because Wikipedia does not include location information with all of its metadata, language provides the best surrogate for geographies of Wikipedia use. The effectiveness of this surrogacy varies. For more geographically constrained languages, such as those in northern or eastern Europe, language is a more reliable proxy for geography than it is for widely spoken languages, such as English or Spanish. Languages also reflect Wikipedia’s geographical bias. The diversity of spoken languages is highest in Africa and Asia (Eberhard et al. 2019), but the majority of Wikipedia languages are European. With the exception of China, which has intermittently blocked access to Wikipedia along with several other internet platforms (Wikipedia 2020d), this pattern mirrors the distribution of global internet access (Graham 2014). For multilingual comparisons, these uneven linguistic representations and inequalities in internet access need to be taken into account. If each language in Wikipedia is given equal weight, the results will be strongly influenced by the geographic distribution of the languages and thus skewed toward Europe.

Physical and Biological Geography

In addition to the geography of Wikipedia languages, the distribution of the entities being compared can affect interest in Wikipedia. Although certain entities attract widespread global attention, people are often more interested in local issues and entities (e.g., Correia et al. 2016). In Wikipedia, this is true for historical figures (Yu et al. 2016) and certain aspects of biodiversity (Roll et al. 2016). As a result, metrics that do not weight for geographic distribution will emphasize entities that co-occur with the distribution of Wikipedia languages. For example, page views for the common European viper (Vipera berus) tend to be higher in languages whose Wikipedia page views primarily originate from countries within the viper’s geographic distribution, such as German, and lower in languages that do not, such as Japanese (Roll et al. 2016). Thus, the fact that the viper has one of the most-viewed reptile pages in Wikipedia overall (Roll et al. 2016) is partially due to its geographic distribution across many European countries (and languages).

Comparative Metrics

Once the appropriate metadata types have been selected and biases related to taxonomy, language, and geography considered, it is important to consider how to appropriately scale Wikipedia data. For assessments based on a single language edition and single metadata type, a simple sum total may be a suitable metric (i.e., the English language page with the most page views). In situations where it is important to adjust for the influence of outliers, such as when identifying entities that attract consistent interest over time (as opposed to brief spikes in attention), robust statistical methods that account for outliers can be used (e.g., Jurečková et al. 2019). Even in these cases, it is important to consider the overall quantity of the metadata type in the language edition (or in Wikipedia as a whole). This is especially true for time series; a positive trend in page views, for example, could result from a general increase in Wikipedia usage rather than growing interest in a particular topic. The choice of comparative metric becomes more complex when combining data from multiple language editions or metadata types. In these cases, it is important to consider carefully what adjustments should be made to account for differences in the overall size and usage of the language editions or metadata types. Yu et al. (2016) propose methods to scale Wikipedia page view data from different language editions to measure the cultural impact of people. However, these methods may not be appropriate for entities that are more commonly of interest to conservationists, such as biological organisms or geographic areas.

Case Study of Bird Species with the Highest Interest in Wikipedia

As a case study of these methods, we used Wikipedia to explore public interest in bird species. We examined each of methodological choices above in the context of this case study. We used the Wikidata Query Service (Wikidata 2020) to extract a list of Wikidata entities for the case study and scraped the associated Wikipedia sitelinks (Wickham 2019).

We obtained Wikipedia page views for bird species from all language editions in Wikipedia. We limited our page views to human users (removing bot-generated views) and obtained views from desktop and mobile sources (Keyes & Lewis, 2020). We filtered our results to include only page views to Wikipedia language editions and excluded views to other Wikimedia projects (e.g., Wikibooks, Wikiquote, and Wikispecies). For pages in English Wikipedia, we scraped 12 additional metadata attributes: size of the article in bytes, total words, links to the page, links from the page, number of references, number of editors, number of edits, average monthly edits, average edits per user, number edits made by the most active Wikipedia editors, article age, and number of page watchers (Wikipedia 2020a).

We specified the date that our list of Wikipedia pages was extracted from Wikidata (29 April 2019) and obtained metadata for all pages in our data set concomitantly to facilitate comparability. Page views were obtained from 1 July 2015 to 1 May 2019. To minimize the effect of seasonal variations and short spikes in interest, we selected a multiyear time series of page views. We calculated a robust measure of mean daily page views...
(with Tukey’s biweight [Bunn et al. 2018]) as well as the sum total of page views for each page.

We used the Clements Checklist of Birds of the World (Clements et al. 2018) to identify bird species in Wikidata by obtaining all entities labeled with an eBird taxon ID (Wikidata property: P3444). At the time of our study, this taxonomy had a higher degree of integration with Wikidata than other taxonomies that we tested (e.g., Avibase and BirdLife). To explore how this choice could influence our results, we also downloaded page view data for some species that other taxonomies treated differently from Clements. For example, Clements considers the Barn Owl to be 1 species, whereas other taxonomies identify it as 3 species (e.g., Gill & Donsker 2019). Wikipedia has pages for both the more inclusive species (Barn Owl [Tyto alba]) and the 3 split species (Eastern Barn Owl [Tyto javanica], Western Barn Owl [Tyto alba], and American Barn Owl [Tyto furcata]) and for the barn owl family (Barn-owl [Tytonidae]). Within the Clements list, we restricted our analyses to pages for species because they are the most frequently used taxonomic unit in biodiversity assessments. This choice could lead to some birds being underrepresented due to recent taxonomic revisions or public interest coalescing at other levels of their taxonomic hierarchy.

We included data from all Wikipedia language editions that had pages for bird species that met our taxonomic criteria. To identify species that attract high interest across a range of languages, we scaled page views for each species by the total number of bird page views in a language in 2 of our comparative metrics (details below).

We did not adjust our results for the geographic distribution of languages in our data set. Thus, our results represent the views of an internet-using public that is primarily located in Europe, North America, and parts of east and south Asia.

To identify species that attract interest beyond the geographic area where they occur, we looked at large-scale patterns of geographic overlap between the breeding distributions of birds and the geographic location of countries associated with languages in Wikipedia. For bird distributions, we obtained the “general region” of a bird’s breeding distribution from Gill and Donsker (2019). For language distributions, we obtained a list of the countries where a language was spoken from the CIA World Factbook and Glottolog (Central Intelligence Agency 2019; Hammarstrom et al. 2019) and defined the language’s distribution as all of the countries where it was listed. To explore patterns of overlap, we categorized countries into the same general regions as bird species (Appendix S1). This approach is limited in that it relies on very large geographic units: species and countries were classified as present or absent in a geographic region regardless of the size of a species’ range or the land area of the country. Furthermore, there can be important differences between where languages are spoken and where their Wikipedia page views come from (Zachte 2020). Despite these limitations, this method provided insight into the influence of biogeographic patterns at large scales.

We compared 3 metrics for calculating online interest in bird species across multiple languages: sum page views (sum), page views scaled by language (language scaled), and page views scaled by distribution and language (distribution-language scaled). Sum was calculated simply as the total page views that a species received across all of the language editions it appeared in. For language scaled, we scaled the page views for each page in a language by the total page views in that language (Wickham & Seidel 2020) and summed the scaled views for each species across all of the languages that the species appeared in. For distribution-language scaled, we used the same scaled page views as language scaled but only counted page views in language editions located outside of the general region of a bird’s breeding distribution.

We explored the resulting lists of species to gain insight into the relationships between languages and among metadata types. Similarity between lists of all species in a given language or metadata was assessed using Spearman’s rank correlation and Euclidean distance matrices with data scaled to a mean of 0 (SD 1). Distance matrices were visualized with agglomerative hierarchical clustering dendrograms based on Ward’s minimum variance method (Maechler et al. 2019). To investigate factors correlating with high online interest, we manually compared the most-viewed bird species in languages and in each of our multilanguage metrics and assessed these for patterns of geographic overlap and trends relating to body size, regional popularity, and presence in the pet trade. All data analyses and visualizations were done in R (R Core Team 2019).

Results

Our initial Wikidata query returned 12,855 entities tagged with an eBird taxon ID, 99.8% of which matched the Clements world list. After filtering to the species category of Clements, we were left with 10,099 bird species that had a page in at least 1 Wikipedia language edition (95.4% of the species in Clements version 2018). We obtained distribution data from Gill and Donsker (2019) for 9,861 of these species.

Our page view data set included 199,699 pages with nearly 757 million page views across 251 Wikipedia language editions. The distribution of page views across languages was highly uneven (views per language edition: 1–290 million, mean 3.01 million [SD 19.7]). English was the largest language edition and accounted for 38.3% of all page views. Together, the 10 most-viewed languages (English, German, Spanish, Russian, Japanese, French,
Table 2. Bird species with the most metadata in English Wikipedia based on 4 different metadata types: page views, number of page editors, number of page watchers (i.e., users actively monitoring changes to the page), and number of words on the page.

| Species            | Page views | Species          | Editors |
|--------------------|------------|------------------|---------|
| Dodo               | 4,422,446  | Dodo             | 2364    |
| Bald Eagle         | 4,040,462  | Common Ostrich   | 2355    |
| Peregrine Falcon   | 3,363,313  | Peregrine Falcon | 2142    |
| Emu                | 2,738,679  | Bald Eagle       | 2132    |
| Golden Eagle       | 2,537,322  | Emperor Penguin  | 1404    |
| Osprey             | 2,368,818  | Canada Goose     | 1403    |
| Common Raven       | 1,753,280  | Snowy Owl        | 1400    |
| Harpy Eagle        | 1,751,775  | Budgerigar       | 1335    |
| Canada Goose       | 1,734,665  | Mallard          | 1275    |
| Emperor Penguin    | 1,664,554  | Emu              | 1142    |

| Species            | Page watchers | Species          | Words   |
|--------------------|---------------|------------------|---------|
| Common Quail       | 25,801        | White-tailed Eagle | 20,423   |
| Red-shouldered Hawk| 19,667        | Red-tailed Hawk  | 18,510   |
| Bufflehead         | 18,389        | Northern Goshawk | 18,273   |
| Red-headed Woodpecker| 18,228    | Passenger Pigeon | 13,240   |
| Asian Koel         | 18,093        | Common Buzzard   | 13,081   |
| Sharp-shinned Hawk | 17,310        | Great Horned Owl | 11,547   |
| Indigo Bunting     | 16,512        | Martial Eagle    | 11,236   |
| Common Merganser   | 15,884        | Bonelli's Eagle  | 10,631   |
| Rose-breasted Grosbeak| 14,808   | Dodo             | 10,486   |
| Black-billed Magpie| 14,691        | Eastern Imperial Eagle | 9867   |

Figure 1. Relationships between rankings of English Wikipedia pages for birds based on different Wikipedia metadata types: (left) dendrogram based on agglomerative hierarchical clustering with Ward’s minimum variance method and (right) correlation between metadata types.

Polish, Dutch, Italian, and Portuguese) accounted for 81.3% of page views.

Rankings derived from different metadata types highlighted different bird species (Table 2). For example, in English Wikipedia, the page for Dodo (*Raphus cucullatus*) received the most page views and edits, whereas the page for Common Quail (*Coturnix coturnix*) had the most page watchers. Page for raptors featured prominently among the English pages with the most words and the White-tailed Eagle (*Haliaeetus albicilla*) had the most words overall. These rankings correlated positively with another but varied in their similarity (Spearman’s $\rho = 0.12–0.99$, mean 0.64 [SD 0.18]) (Fig. 1). Metadata related to some aspects of edits (number of editors, total edits, and average monthly edits), and page views clustered together, as did metadata related to the quantity of information on a page (page size, total words, total references, and links from the page). Other metadata, such as the number of links to a page and the average edits per user, were less correlated.

Bird species differed in how they accumulated page views over time (i.e., their page view time series). These
Figure 2. Time series of English Wikipedia page views for selected bird pages from 2015 to 2019: (top) 3 species with comparable page-view totals but different temporal patterns (Olive-sided Flycatcher 51,900 total page views; Purple-rumped Sunbird 44,900 page views; and Red-shouldered Vanga 55,200 page views) and (bottom) page views for different taxonomic divisions of barn owls (American Barn Owl 24,900 total page views; Barn Owl 1.35 million page views; barn-owl [family] 222,000 page views; Eastern Barn Owl 18,300 page views; and Western Barn Owl 17,900 page views).

The birds that received the most page views varied between Wikipedia language editions (Table 3). Some patterns were apparent in these differences. For example, 5 of the 10 most-viewed birds in Persian Wikipedia are species that are farmed (e.g., Common Ostrich [Struthio camelus]) or kept as pets (Lennox & Harrison 2006). Meanwhile, none of the top 10 species in Finnish Wikipedia feature in the pet trade. In many language editions, species that occur in the wild in the country responsible for most of the language’s page views were strongly represented among the most-viewed pages. In the cluster analysis, languages spoken in countries located in the same geographic region often grouped together (Fig. 3).

The 3 comparative metrics we used to explore Wikipedia page views for birds across language editions (sum, language scaled, and distribution-language scaled) correlated positively with one another. Sum and language scaled produced the most similar rankings (Spearman’s ρ 0.83), whereas sum and distribution-language scaled were the most different (Spearman’s ρ 0.57). Despite these positive correlations, only 2 species appeared among the top 20 species in all 3 rankings: Common Ostrich and Budgerigar [Melopsittacus undulatus]. Reviewing the highest ranking species according to each metric revealed that sum shared many similarities with English Wikipedia, language scaled highlighted more species native to Europe, and distribution-language scaled emphasized species that feature in the pet trade or are farmed (Fig. 3).

Discussion

Several types of Wikipedia metadata can be used for conservation applications. We found that Wikipedia page views were a particularly effective resource for measuring public interest at large scales. Although Wikipedia data can capture the online actions of many people, they do not represent all of the public constituencies relevant to conservation. Furthermore, page views do not explain the causes of online interest. In the case of species, high page view totals can equally result from a successful conservation awareness campaign as from people disliking a species or wanting to catch or hunt it. Page views can also be driven by factors other than direct interactions. The Dodo, an extinct species, received the most page views overall in our data set. This could be related to the Dodo’s role as an icon of extinction, but it may also result from the bird’s prominence in English literature and figures of speech, as well as the title of a well-known website (thedodo.com). Thus, although Wikipedia metrics can contribute valuable perspectives for conservationists, additional information and cultural awareness is necessary to contextualize them before they are applied to designing conservation policy.
Bird species with the most page views in 4 representative Wikipedia language editions (Finnish, Japanese, Persian, and Ukrainian).

| Finnish          | Japanese                |
|------------------|-------------------------|
| Species          | Page views   | Species          | Page views   |
| Whooper Swan     | 253,644       | Ural Owl        | 877,105      |
| White-tailed Eagle | 201,860     | Shoebill        | 799,946      |
| Eurasian Blackbird | 189,905      | Bull-headed Shrike | 791,708     |
| Western Capercaillie | 173,859    | Barn Swallow    | 778,509      |
| Golden Eagle     | 152,105       | Crested Ibis    | 766,748      |
| Eurasian Eagle-Owl | 151,402      | Eurasian Tree Sparrow | 743,799    |
| Osprey           | 142,068       | Japanese Bush Warbler | 720,551     |
| Great Tit        | 138,571       | Oriental Stork  | 624,624      |
| Eurasian Hoopoe  | 136,887       | White-cheeked Starling | 607,717     |
| Common Crane     | 131,865       | Japanese White-eye | 565,931      |
| Budgerigar       | 452,676       | White Stork     | 194,600      |
| Cockatiel        | 302,851       | Great Spotted Woodpecker | 169,705     |
| Common Ostrich   | 225,851       | Common Cuckoo  | 85,489       |
| Gray Parrot      | 169,764       | European Starling | 76,723      |
| Rosy-faced Lovebird | 129,435     | Eurasian Bullfinch | 67,387      |
| Bearded Vulture  | 111,848       | Common Raven    | 64,233       |
| White-eared Bulbul | 109,328      | Common Ostrich  | 58,220       |
| Rose-ringed Parakeet | 105,267    | Red Crossbill   | 54,195       |
| Eurasian Hoopoe  | 102,385       | Golden Eagle    | 53,764       |
| Ring-necked Pheasant | 99,347      | Peregrine Falcon | 52,699       |

*a Species that occur in the wild in the country responsible for the majority of the Wikipedia edition’s page views (corresponding countries: Finland, Japan, Iran, and Ukraine) (occurrence data from eBird.org).

*b Species that are farmed or commonly kept as pets.

At the scale of individual languages, Wikipedia page views can provide measures of public interest within a particular linguistic context. Sum totals and robust daily means of page views produced similar results across large numbers of species in our data set, but generated differences relevant for comparisons between smaller subsets of entities (e.g., the species in Fig. 2). Overall, sum totals may be better in situations that include pages with low page views (e.g., <1 view per day), whereas robust means may be more effective for assessing consistent interest between pages with relatively high page views. The geographic grouping of languages in our cluster analysis demonstrated the importance of geography (both biological and cultural) in online interest in species.

When a high proportion of Wikipedia page views originate from a single country or region, page views can reveal public interest within a geographic area. The geographic resolution of Wikipedia page views limits their effectiveness at geographic scales smaller than countries, however, and even in instances where most page views originate from a single country, it is impossible to tell which parts of that country contribute the page views. Despite these limitations, entities that attract high levels of attention in a Wikipedia language could be candidates for conservation flagships in the country or countries associated with that language. The most-viewed bird species in several Wikipedia language editions already act as conservation flagships (e.g., the Whooper Swan [Cygnus cygnus] in Finland).

Combining data from multiple Wikipedia language editions or metadata types presents additional methodological challenges. Each of the 3 multilanguage metrics we compared has strengths and drawbacks. By counting all page views equally, the sum metric measures overall interest among the Wikipedia-using public. This could be beneficial for marketing conservation activities: species that attract many page views in Wikipedia may also generate attention and perhaps conservation support in other contexts. However, this type of raw score is strongly influenced by English and underrepresents regions with lower levels of internet penetration and usage. It also does not account for how page views are distributed among languages (i.e., a high number of page views in one language could result in the same sum as low page views in many languages). The language scaled metric highlights species that generate interest across multiple cultural contexts (as represented by languages). By treating all languages equally, however, this metric is biased toward regions with many Wikipedia languages (i.e., Europe) and toward entities that occur in those regions (because they are likely to overlap geographically with more Wikipedia languages). Biases or random patterns present in smaller language editions can also have an outsized influence in this type of metric. The distribution-language scaled metric...
identifies species that generate interest beyond their area of geographic distribution. In some instances, these could be candidates for flagship species that need to be relevant across many cultural and geographic contexts. For birds, several of the highest ranking species according to distribution-language scaled, such as the Emperor Penguin \( \text{Aptenodytes forsteri} \), fulfill this role. By adjusting for the influence of geography, metrics like distribution-language scaled can also be useful for investigating cultural relationships and biological traits that correlate with increased public interest (such as the pet trade). Disadvantages of the distribution-language scaled metric are that it uses less data (page views from languages that overlapped geographically with a species were not included), relies on a very coarse geographic resolution, and inflates the importance of species that overlap in distribution with only a few languages. Antarctica, for example, home of the Emperor Penguin, did not have any languages assigned to it in our data set. Furthermore, the use of Wikipedia languages online does not always align with where those languages are spoken in the real world. Future studies could address these shortcomings by incorporating more precise mapping of species’ distributions together with data from...
culturomic resources with more fine-scale geographic resolution, such as Google Trends (Proulx et al. 2013).

Our results demonstrate the potential of Wikipedia as a resource for assessing public interest in the context of conservation. In addition to species, Wikipedia data could be used to compare interest in protected areas, conservation-relevant concepts, and plants and animals at different levels of the taxonomic hierarchy. Many of the methodological considerations that we highlighted in the context of Wikipedia are also relevant to other digital big data sources, such as Google Trends (Proulx et al. 2013), Twitter, and other social media platforms (Fink et al. 2019; Toivonen et al. 2019), and digital news media (Acerbi et al. 2020). We hope our findings will encourage researchers to engage with these data and further explore their conservation applications.

Acknowledgment

We thank M. Poulter for initial discussions on this idea and feedback on the manuscript.

Supporting Information

Mapping overlap geographic between species and languages (Appendix S1) is available online. The authors are solely responsible for the content and functionality of this material. Queries (other than absence of the material) should be directed to the corresponding author.

Literature Cited

Acerbi A, Kerhoas D, Webber AD, McCabe G, Mittermeier RA, Schwitzer C. 2020. The impact of the “World’s 25 Most Endangered Primates” list on scientific publications and media. Journal for Nature Conservation 54:125794.
Alexa. 2019. The top 500 sites on the web. Available from https://www.alexa.com/topsites (accessed June 2019).
Bunn A, et al. 2018. dplR: dendrochronology program library in R. CRAN.R-project. Available from https://cran.r-project.org/web/packages/dplR/index.html (accessed July 2020).
Central Intelligence Agency (CIA). 2019. The world factbook. CIA, Washington, D.C. Available from https://www.cia.gov/library/publications/the-world-factbook/ (accessed August 2019).
Clements JF, Schulenberg TS, IIff MJ, Roberson D, Fredericks TA, Sullivan BL, Wood CL. 2018. The eBird/Clements checklist of birds of the world. Version 2018. Available from http://www.birds.cornell.edu/clementschecklist/download/ (accessed April 2019).
Correia RA, Ladle R, Jarić I, Malhado Ana CM, Mittermeier JC, Roll U, Soriiano-Redondo A, Verissimo D, Fink C, Hausmann A, Guedes-Santos J, Vardi R, Di Minin E. 2021. Digital data sources and methods for conservation culturomics. Conservation Biology (this issue). https://doi.org/10.1111/cobi.13706.
Correia RA, Di Minin E, Jarić I, Jepson P, Ladle R, Mittermeier J, Roll U, Soriiano-Redondo A, Verissimo D. 2019. Inferring public interest from search engine data requires caution. Frontiers in Ecology and the Environment 17:254–255.
Correia RA, Jarić I, Jepson P, Malhado ACM, Alves JA, Ladle RJ. 2018. Nomenclature instability in species culturomic assessments: why synonyms matter. Ecological Indicators 90:74–78.
Correia RA, Jepson PR, Malhado ACM, Ladle RJ. 2016. Familiarity breeds content: assessing bird species popularity with culturomics. PLoS Life & Environment 4:1–15.
Correia RA, Ladle R, Jarić I, Malhado ACM, Mittermeier JC, Roll U, Soriiano-Redondo A, Verissimo D, Fink C, Hausmann A, Guedes-Santos J, Vardi R, Di Minin E. 2021. Digital data sources and methods for conservation culturomics. Conservation Biology (this issue). https://doi.org/10.1111/cobi.13706.
Eberhard DM, Simons GF, Fennig CD (editors). 2019. Ethnologue: languages of the world. 22nd edition. SIL International, Dallas, Texas.
Fink C, Hausmann A, Di Minin E. 2020. Online sentiment towards iconic species. Biological Conservation 241:108289.
Generous N, Fairchild G, Deshpande A, Valle SYD, Priedhorsky R. 2014. Global disease monitoring and forecasting with Wikipedia. PLoS Computational Biology 10:e1003892.
Gill F, Donsker D (editors). 2019. IOC world bird list. Version 9.1. International Ornithologists Union. Available from https://www.worldbirdnames.org/ (accessed June 2019).
Graham M. 2014. Inequitable distributions in internet geographies: the Global South is gaining access, but lags in local content. Innovations: Technology, Governance. Globalization 9:3–19.
Hammarstam H, Forkel R, Haspelmath M. 2019. Glottolog 3.4. Available from http://glottolog.org/ (accessed August 2019).
Jurečková J, Picek J, Schindler M. 2019. Robust statistical methods with R. CRC Press, Cleveland, Ohio.
Keyes O, Lewis J. 2020. pageviews: An API Client for Wikimedia Traffic Data. Available from https://cran.r-project.org/web/packages/pageviews/index.html (accessed July 2020).
Kitchin R. 2014. The data revolution: big data, open data, data infrastructures and their consequences. SAGE Publications, London.
Ladle RJ, Correia RA, Do Y, Joo GJ, Malhado ACM, Proulx R, Roberge JM, Jepson P. 2016. Conservation culturomics. Frontiers in Ecology and the Environment 14:269–275.
Lennox A, Harrison G. 2006. The companion bird. Pages 29–44 in Harrison G, Lightfoot T, editors. Clinical avian medicine. Volume I. Spix Publishing, Palm Beach, Florida.
Macchler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. 2019. Cluster: cluster analysis basics and extensions. Available from https://cran.r-project.org/web/packages/cluster/cluster.pdf (accessed July 2019).
Manfredo MJ. 1989. Human dimensions of wildlife management. Wildlife Society Bulletin 17:447–449.
Messner M, DiStaso MW. 2013. Wikipedia versus Encyclopedia Britannica: a longitudinal analysis to identify the impact of social media on the standards of knowledge. Mass Communication and Society 16:465–486.
Mittermeier JC, Roll U, Matthews TJ, Grenyer R. 2019. A season for all things: phenological imprints in Wikipedia usage and their relevance to conservation. PLoS Biology 17:e3000146.
Proulx L, Massicotte P, Marc PE. 2013. Googling trends in conservation biology. Conservation Biology 28:44–51.
R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
Roll U, Mittermeier JC, Diaz GI, Novosolov M, Feldman A, Itescu Y, Meiri S, Grenyer R. 2016. Using Wikipedia page views to explore the cultural importance of global reptiles. Biological Conservation 204:12–50.
Samonenko A, Yasseri T. 2014. The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. EPJ Data Science 3:1.
Skiena S, Ward C. 2014. Who’s bigger? Where historical figures really rank. Cambridge University Press, New York.
Sutherland WJ, et al. 2018. A 2018 horizon scan of emerging issues for global conservation and biological diversity. Trends in Ecology & Evolution 33:47–58.
Toivonen T, Heikinheimo V, Fink C, Hausmann A, Hiippala T, Järv O, Tenkanen H, Di Minin E. 2019. Social media data for conservation science: a methodological overview. Biological Conservation 233:298–315.
Wickham H. 2019. rvest: Easily Harvest (scrape) Web Pages. Available from https://cran.r-project.org/web/packages/rvest/index.html (accessed May 2019).
Wickham H, Seidel D. 2020. scales: scale functions for visualization. Available from https://cran.r-project.org/web/packages/scales/index.html (accessed July 2020).
Wikidata. 2020. Wikidata query service. Available from https://query.wikidata.org/ (accessed June 2020).
Wikipedia. 2020a. XTools: feeding your data hunger. Available from https://xtools.wmflabs.org/ (accessed June 2020).
Wikipedia. 2020b. Wikimedia statistics. Available from https://stats.wikimedia.org/#/en.wikipedia.org (accessed June 2020).
Wikipedia. 2020a. List of Wikipedias. Available from https://meta.wikimedia.org/wiki/List_of_Wikipedias (accessed June 2020).
Wikipedia. 2020b. Wikipedia:Pageview statistics. Available from https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics (accessed July 2020).
Wikipedia. 2020c. Wikipedia:Article traffic jumps. Available from https://en.wikipedia.org/wiki/Wikipedia:Article_traffic_jumps (accessed June 2020).
Wikipedia. 2020d. Censorship of Wikipedia. Available from https://en.wikipedia.org/wiki/Censorship_of_Wikipedia (accessed July 2020).
Wilson JL. 2014. Proceed with extreme caution: citation to Wikipedia in light of contributor demographics and content policies. Vanderbilt Journal of Entertainment & Technology Law 16: 857–908.
Yasseri T, Spoerri A, Graham M, Kertész J. 2014. The most controversial topics in Wikipedia: a multilingual and geographical analysis. Pages 25–48 in Fichman P, Hara N (editors). Global Wikipedia: international and cross-cultural issues in collaboration. Scarecrow Press, Lanham, Maryland.
Yasseri T, Sumi R, Rung A, Kornai A, Kertész J. 2012. Dynamics of conflicts in wikipedia. PLOS ONE 7 (e38869). https://doi.org/10.1371/journal.pone.0038869.
Yu AZ, Ronen S, Hu KZ, Hidalgo CA. 2016. Pantheon 1.0, a manually verified dataset of globally famous biographies. Scientific Data 3:150075.
Zachte E. 2019. Page views for Wikimedia, all projects, both sites, normalized. Available from https://stats.wikimedia.org/EN/TablesPageViewsMonthlyAllProjects.htm (accessed July 2020).
Zachte E. 2020. Wikimedia traffic analysis report — page views per Wikipedia Language — Breakdown. Available from https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm (accessed July 2020).