
















Digital data sources and methods for conservation culturomics

Ricardo A. Correia ^{1,2,3,4*} Richard Ladle ^{4,5} Ivan Jarić ^{6,7} Ana C. M. Malhado ⁴
John C. Mittermeier ⁸ Uri Roll ⁹ Andrea Soriano-Redondo ^{5,10} Diogo Veríssimo ^{11,12,13}
Christoph Fink ^{1,2} Anna Hausmann ^{1,2} Jhonatan Guedes-Santos ⁴ Reut Vardi ¹⁴
and Enrico Di Minin ^{1,2,15}

¹Department of Geosciences and Geography, Helsinki Lab of Interdisciplinary Conservation Science, University of Helsinki, Helsinki 00014, Finland

²Helsinki Institute of Sustainability Science (HELSUS), University of Helsinki, Helsinki 00014, Finland

³CESAM - Centre for Environmental and Marine Studies, University of Aveiro, Campus Universitário de Santiago, Aveiro 3910-193, Portugal

⁴Institute of Biological and Health Sciences, Federal University of Alagoas, Maceió 57072-900, Brazil

⁵CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Laboratório Associado, Universidade do Porto, Porto 4485-661, Portugal

⁶Biology Centre of the Czech Academy of Sciences, Institute of Hydrobiology, České Budějovice 37005, Czech Republic

⁷Department of Ecosystem Biology, Faculty of Science, University of South Bohemia, České Budějovice 37005, Czech Republic

⁸School of Geography and the Environment, University of Oxford, Oxford OX1 3QY, U.K.

⁹Mitrani Department of Desert Ecology, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion 8499000, Israel

¹⁰CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Laboratório Associado, Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa 1349-017, Portugal

¹¹Department of Zoology, University of Oxford, Oxford OX1 3SZ, U.K.

¹²Oxford Martin School, University of Oxford, Oxford OX1 3BD, U.K.

¹³San Diego Zoo Institute for Conservation Research, Escondido, CA 92027, U.S.A.

¹⁴The Albert Katz International School for Desert Studies, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion Durban 8499000, Israel

¹⁵School of Life Sciences, University of KwaZulu-Natal, Durban 4041, South Africa

Abstract: Ongoing loss of biological diversity is primarily the result of unsustainable human behavior. Thus, the long-term success of biodiversity conservation depends on a thorough understanding of human–nature interactions. Such interactions are ubiquitous but vary greatly in time and space and are difficult to monitor efficiently at large spatial scales. However, the Information Age also provides new opportunities to better understand human–nature interactions because many aspects of daily life are recorded in a variety of digital formats. The emerging field of conservation culturomics aims to take advantage of digital data sources and methods to study human–nature interactions and thus to provide new tools for studying conservation at relevant temporal and spatial scales. Nevertheless, technical challenges associated with the identification, access, and analysis of relevant data hamper the wider adoption of culturomics methods. To help overcome these barriers, we propose a conservation culturomics research framework that addresses data acquisition, analysis, and inherent biases. The main sources of culturomic data include web pages, social media, and other digital platforms from which metrics of content and engagement can be obtained. Obtaining raw data from these platforms is usually desirable but requires careful consideration of how to access, store, and prepare the data for analysis. Methods for data analysis include network

*email rabc85@gmail.com

Article impact statement: Guidelines for overcoming challenges associated with digital data collection and analysis can advance conservation culturomics applications.

Paper submitted January 31, 2020; revised manuscript accepted June 5, 2020.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

approaches to explore connections between topics, time-series analysis for temporal data, and spatial modeling to highlight spatial patterns. Outstanding challenges associated with culturomics research include issues of interdisciplinarity, ethics, data biases, and validation. The practical guidance we offer will help conservation researchers and practitioners identify and obtain the necessary data and carry out appropriate analyses for their specific questions, thus facilitating the wider adoption of culturomics approaches for conservation applications.

Keywords: data-driven science, digital content, digital methods, human–nature interactions, research framework

Fuentes de Información Digital y Métodos para la Culturomía de la Conservación

Resumen: La continua pérdida de biodiversidad es el resultado principal del comportamiento humano insostenible. Por esto, el éxito a largo plazo de la conservación de la biodiversidad depende de una comprensión exhaustiva de las interacciones humano-naturaleza. Dichas interacciones son ubicuas pero varían enormemente en el tiempo y el espacio, lo que dificulta su monitoreo eficiente a escalas espaciales amplias. Sin embargo, la Era de la Información también nos proporciona nuevas oportunidades para comprender de mejor manera las interacciones humano-naturaleza pues muchos aspectos de la vida diaria quedan registrados en una variedad de formatos digitales. El campo emergente de la culturomía de la conservación busca aprovechar los recursos y los métodos digitales para estudiar las interacciones humano-naturaleza y así proporcionar nuevas herramientas para el estudio de la conservación a escalas temporales y espaciales relevantes. No obstante, las dificultades técnicas asociadas con la identificación, acceso y análisis de la información relevante obstaculizan la adopción más amplia de los métodos de la culturomía. Para ayudar a superar estas barreras proponemos un marco de trabajo de investigación de culturomía de la conservación que aborde la obtención de datos, el análisis y los sesgos inherentes. Entre las principales fuentes de datos sobre culturomía se incluyen las páginas web, las redes sociales y otras plataformas digitales a partir de las cuales se pueden obtener medidas del contenido y la participación. Normalmente se busca obtener datos crudos a partir de este tipo de plataformas, pero esto requiere que se tengan en consideración las vías de acceso, el almacenaje y la preparación de la información para su posterior análisis. Los métodos para el análisis de datos incluyen análisis de redes para explorar las conexiones entre los temas, el análisis de series de tiempo para los datos temporales y el modelado espacial para resaltar los patrones espaciales. Los desafíos sobresalientes asociados a la investigación en culturomía incluyen temas de interdisciplinaria, ética, sesgos de datos y validación. La orientación práctica que ofrecemos ayudará a los investigadores y practicantes de la conservación a identificar y obtener los datos necesarios. También les ayudará a realizar análisis apropiados para responder a sus preguntas específicas, facilitando así la adopción más amplia de las estrategias de culturomía para su aplicación en la conservación.

Palabras Clave: ciencia guiada por datos, contenido digital, interacciones humano-naturaleza, marco de trabajo de investigación, métodos digitales

摘要: 生物多样性的持续丧失主要是人类不可持续行为造成的结果。因此,生物多样性保护的长期成功取决于对人与自然相互作用的深入理解。这种互作普遍存在但有很大的时空差异,难以在大空间尺度上有效监测。然而,信息时代为更好地理解人与自然的互动提供了新的机会,因为日常生活的许多方面都已有数字化的记录。保护文化组学是一个新兴的研究领域,旨在利用数字数据资源和方法来研究人类与自然的互动,从而为相应时空尺度的保护研究提供新的工具。然而,与相关数据的识别、获取和分析相关的技术挑战阻碍了文化组学方法的广泛应用。为了克服这些阻碍,我们构建了一个保护文化组学研究框架,以解决数据获取、分析和固有偏倚的问题。文化组学数据的主要来源包括网页、社交媒体和其它数字平台,我们可以从中获得衡量内容和参与度的指标。从这些平台获取原始数据通常是可行的,但需要仔细考虑如何访问、存储和准备数据以进行分析。数据分析的方法包括探讨主题之间联系的网络方法,针对时间数据的时间序列分析,以及强调空间格局的空间建模。与文化组学研究相关的主要挑战包括跨学科、伦理、数据偏倚和验证等问题。我们提供的实际指导将帮助保护科学的研究者和实践者识别和获取必要的数据,并对具体问题进行分析,从而促进文化组学方法在保护中获得更广泛的应用。【翻译:胡怡思; 审校:聂永刚】

关键词: 数据驱动的科学, 数字内容, 数字方法, 人与自然的互动, 研究框架

Introduction

Information and communication technologies have revolutionized the modern world and more than half of the world's population is now connected to the internet (International Telecommunication Union 2020). Increasing engagement with these technologies across

the world has transformed the digital realm into a vast repository of information on the lives of billions of people. Consequently, the information generated as part of this digital revolution can produce actionable insights regarding human–nature interactions (Di Minin et al. 2015; Ladle et al. 2016). Such information is of great value to conservation science and practice

because the success of conservation efforts partially depends on understanding human interest, values, ideas, and behaviors toward nature (Schultz 2011; Bennett et al. 2017).

Conservation culturomics aims to analyze the digital data generated by people to provide novel insights on human–nature interactions for conservation (Ladle et al. 2016). The most common application of conservation culturomics so far has been to explore temporal and spatial dynamics of public interest in conservation-related topics. Multiple studies have assessed these dynamics over long periods (e.g., Funk & Rusowsky 2014; Proulx et al. 2014; Mittermeier et al. 2019; Troumbis 2019) or in response to specific events such as conservation interventions, news, movies, and nature documentaries (e.g., Papworth et al. 2015; Soriano-Redondo et al. 2017; Fernández-Bellon & Kane 2020; Verissimo et al. 2020). Other common applications of culturomics to conservation include identifying culturally salient species and sites (e.g., Roll et al. 2016; Correia et al. 2018b; Ladle et al. 2019) and investigating preferences for nature-based recreation (e.g., Hausmann et al. 2018; Monkman et al. 2018b; Sbragaglia et al. 2019). There are also a number of other topics in which the use of culturomics approaches for conservation is now developing rapidly, including biological invasions (e.g., Fukano & Soga 2019; Jarić et al. 2021), illegal wildlife trade (Hinsley et al. 2016; Di Minin et al. 2018; Di Minin et al. 2019), and human–wildlife conflict (Miranda et al. 2016).

Although the uptake of culturomics approaches in conservation has been rapid, there are still many practical and technical challenges hindering its broader use. To choose among the multiple digital data sources and methods available, conservation researchers and practitioners must be able to assess the characteristics of the data, determine how to obtain it, and identify which metrics and methods are best suited for the intended analyses. To help guide these decisions and the design of conservation culturomics analyses, we developed a framework for conservation culturomics research based on our experiences carrying out this type of research. We compiled an overview of the framework, considered the main data sources, methods, and challenges associated with conservation culturomics research and identified additional actions necessary to advance the field.

Conservation Culturomics Research Framework

Culturomics research is influenced by emerging data-intensive scientific paradigms. Data-driven science draws heavily on statistical exploration and data-mining techniques to help identify questions and hypotheses worthy of further inquiry (Kitchin 2014). As such, culturomics research is typically a highly iterative process in which decisions regarding the research scope, data character-

istics, accessibility, and analytical methods are revisited frequently. We summarized the main stages of this decision-making process in the conservation culturomics research framework (Fig. 1).

Culturomics Content as Digital Corpora

Digital content for culturomics can be obtained from multiple sources, may include 1 or more data formats (e.g., text, images, and videos), and often varies in metadata availability (e.g., associated temporal and spatial data). These characteristics can pose challenges for researchers when selecting and compiling data for analyses. It is, therefore, useful to think of digital content for culturomics analyses in terms of collections of items, such as web pages, books, or social-network posts that can be used to generate structured data sets for subsequent analysis. In the culturomics literature, such collections are often referred to as *corpora*. In the context of conservation culturomics, the broader definition of *corpora* – collections of knowledge or evidence (Merriam-Webster 2020) – is best suited to account for collections of both textual and nontextual data types, such as images and videos (Michel et al. 2011; Sherren et al. 2017). In other words, any set of texts, images, videos, songs, paintings, or other products of human culture from which a structured data set can be derived for analysis represent potential corpora for conservation culturomics (e.g., Ladle et al. 2017).

There are 2 key dimensions of digital corpora that are relevant to conservation culturomics. One refers to the content featured in the elements composing each corpus. This is the original scope of culturomics analyses (Michel et al. 2011) and generally focuses on what is represented in the corpus and the context of such representation. The other dimension refers to engagement with the elements that compose the corpus and focuses on assessing interactions with elements of the corpus, including searches, views, comments, and shares. The relevance of assessing patterns of engagement for culturomics analyses is that the access, dissemination, and discussion of digital content can be important drivers of cultural dynamics and evolution (Acerbi 2019), including human–nature interactions. To support the selection of corpora for analysis, we examined the main characteristics of commonly used corpora (Table 1) and their potential applications in conservation. However, there are numerous other corpora of potential relevance to conservation (e.g., sets of audio records, maps, reports, children’s books, etc.), and we encourage researchers to explore beyond those outlined here.

Web Pages

Most digital content on the internet is available through web pages, so they can be considered the

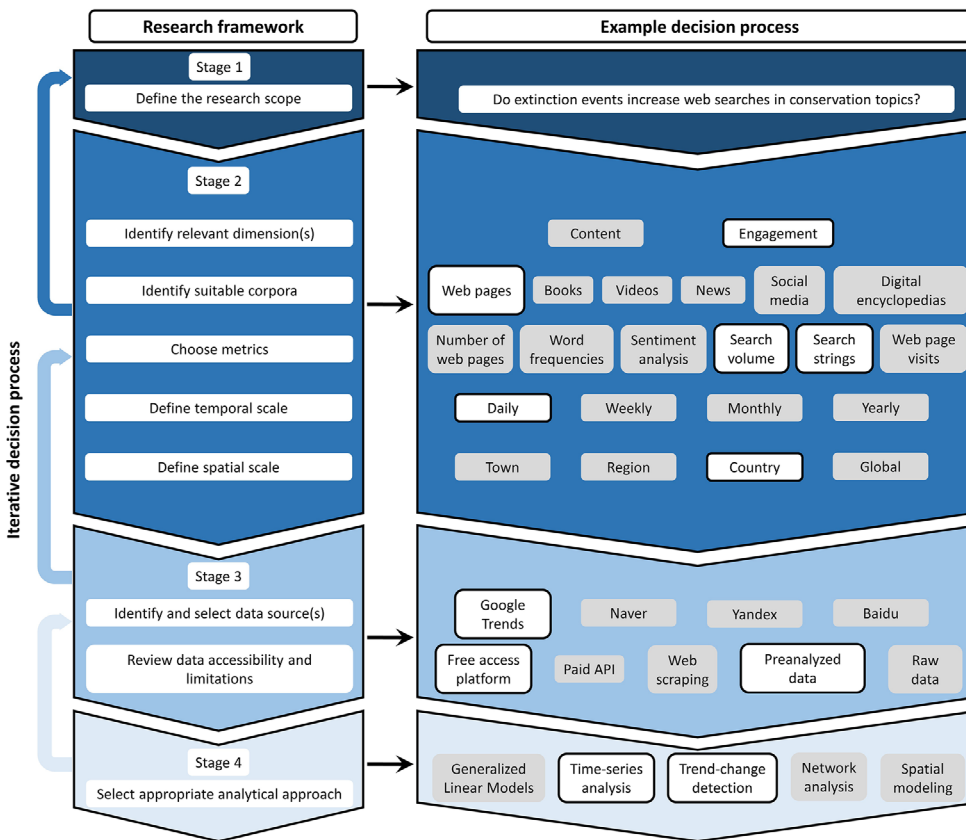


Figure 1. Four key stages of the iterative process of conservation culturomics research design: define the research scope (stage 1), select the relevant corpora for analysis (stage 2), identify possible data sources and data extraction (stage 3), and analyze data (stage 4).

quintessential corpus for culturomics analyses. Web pages are text documents available on the internet, and their information set often contains other types of non-textual data, such as images, audio, and video (Table 1). We focused on sets of web pages as the corpus used for analysis, but individual websites or platforms dedicated to specific content, such as Wikipedia or YouTube, can also be used as individual corpora (see below). Search engines, such as Google, Yahoo, and Bing, specialize in crawling and indexing large volumes of web content and provide a good starting point to access data on web-page content and engagement.

The number and content of web pages can be used to quantify the cultural salience of species or places of conservation importance (Correia et al. 2016, 2018b; Ladle et al. 2019) and or to assess the overlap between societal and scientific interest in conservation topics (Jarić et al. 2019). Web pages located on the deep web or the dark web may also represent useful content for culturomics analyses, although they are often more difficult to access (but see Harrison et al. [2016] and Hayes et al. [2018]). Engagement with internet web pages through web searches (e.g., Google Trends and Naver Trends) or web-page visitation (e.g., Google Analytics and Bing Webmaster Tools) can also be used in conservation research to explore public reactions to conservation interventions (e.g., Do et al. 2015; Soriano-Redondo et al. 2017).

Book Collections

Books have been used as a medium to record and transmit information for centuries. Their contents, including text and images (Table 1), are increasingly being digitized and made available through the internet. This process has facilitated the computational analysis of book contents for a range of purposes and was the genesis of culturomics analyses (Michel et al. 2011). The Google Books project, for example, has digitized over 5 million books whose content is accessible to researchers in pre-analyzed format through Google Ngram Viewer (<https://books.google.com/ngrams>). Other platforms, such as Project Gutenberg (<https://www.gutenberg.org/>), may also be used for analysis.

Book contents can be used to assess historical trends of interest in environmental and conservation topics (Richards 2013), the evolution of human connection and disconnection with nature (Kesebir & Kesebir 2017), and identify popular species through time (Stergiou 2017). Engagement with books has not yet been well explored for conservation purposes but holds great potential. Platforms, such as Goodreads (<https://www.goodreads.com/>) or WorldCat (<https://www.worldcat.org/>), compile information about book reviews and book availability in libraries and may provide a basis for initial analyses in this area.

Table 1. Commonly used corpora in conservation culturomics research and their associated metrics, spatial and temporal data sources, and possible access platforms.

<i>Corpus</i>	<i>Corpus dimension</i>	<i>Example metrics</i>	<i>Temporal data sources</i>	<i>Spatial data sources</i>	<i>Possible data access platforms</i>	<i>Reference example</i>
Web pages	content	number of web pages frequency of representation (words, images, etc.)	web page creation date	web hosting IP location	Google Search Microsoft Bing WordCounter	Funk & Rusowsky 2014 Do et al. 2015 Soriano-Redondo et al. 2017
	engagement	number of web page visits number of internet searches	time of web page visit time of internet search	internet IP location	Google Analytics Google Trends Naver Trends	Correia et al. 2018b Jarić et al. 2019 Ladle et al. 2019
News items	content	number of news items news sentiment polarity	news publication date	country of news publisher	Google News Webhose.io	Papworth et al. 2015 Megias et al. 2017 Brackowski et al. 2018
	engagement	number of news searches number of news comments	time of news search time of comment publication	internet IP location country of news publisher	Google Trends New York Times API The Guardian API	Fink et al. 2020 Francis et al. 2019
Social network posts	content	number of posts sentiment of posts	time of post publication	geolocation of post user profile	Twitter Facebook Instagram	Roberge 2014 Daume & Galaz 2016 Hausmann et al. 2018
	engagement	number of replies number of likes	time of post share time of engagement	geolocation of post user profile	Twitter Facebook Instagram	Kidd et al. 2018 Fink et al. 2020
Video-sharing platforms	content	number of videos frequency of representation (words, entities, etc.)	time of video publication	channel profile	YouTube Vimeo	El Bizri et al. 2015 Sbragaglia et al. 2019 Measey et al. 2019
	engagement	number of comments number of views	time of comment publication time of viewing	channel profile user profile	YouTube TubeBuddy Vimeo	
Books	content	number of books frequency of representation (words, images, etc.)	date of publication	country of publication text language	Google Books WorldCat Google Ngram	Richards 2013 Kesebir & Kesebir 2017 Stergiou 2017
	engagement	book review rating number of books in libraries	review date time of query	user profile location of library	WorldCat	
Digital encyclopedia	content	number of articles frequency of representation (words, images, etc.)	date of article creation	language edition geolocated articles editor location	Wikipedia Encyclopaedia Britannica Everipedia	Roll et al. 2016 Fernández-Bellon & Kane 2020 Mittermeier et al. 2019
	engagement	number of page views number of edits	page viewing time time of editing	language edition	Wikipedia Everipedia	

Video-Sharing Platforms

Video repositories provide another fertile source of data for conservation culturomics. Both the amount of video uploads and the time spent watching video content

online have vastly increased recently and are likely to continue to grow in coming years (Cisco 2018). Video-sharing platforms, such as YouTube or Vimeo, allow researchers to explore aspects of digital (or digitized) video

corpora for culturomics research, including video metadata and engagement based on views and likes (Table 1).

The content of online videos can be used in conservation to explore illegal activities (El Bizri et al. 2015), assess how different recreation practices affect threatened species (Sbragaglia et al. 2019), and characterize human-wildlife conflict (Miranda et al. 2016). Conservation research based on video corpora can also take advantage of data on video engagement, drawing from the social-networking capabilities of many video-sharing platforms to explore views, likes, and online comments.

News Media

News items are a particularly interesting source of data for culturomics because they are often produced in real time, unlike other cultural products, such as books, that lag real-world events (Schwartz 2011). Ongoing efforts to digitize historical periodicals are making large amounts of news items, including text and image data, available for culturomics analysis (e.g., Lansdall-Welfare et al. 2017). Meanwhile, news media have a growing presence online; sound and video recordings are becoming increasingly prominent alongside text and images (Table 1). News-aggregating platforms, such as GDELT (<https://www.gdeltproject.org/>) and Webhose (<https://webhose.io/>), provide compilations of recent news items from across the globe. News from specific media outlets, including *The New York Times* (<https://developer.nytimes.com/>) and *The Guardian* (<https://open-platform.theguardian.com/>), may also be accessible using application programming interfaces (APIs) provided by these platforms.

Online news can be used in conservation research to understand the impact of how specific conservation actions are communicated in the media (Brackowski et al. 2018), assess how the media attention given to conservation compares with other topics (Veríssimo et al. 2014), and evaluate changing perceptions of what constitutes newsworthy wildlife events over time (Francis et al. 2019). Engagement with online news can also be used to explore the role of news media in linking conservation research to social media (Papworth et al. 2015) and to evaluate the sentiment of responses to news reports of charismatic species (Fink et al. 2020).

Social Networks

Data from online social-networking platforms have been used widely in the scientific literature to explore aspects of human culture that relate to environmental and nature conservation topics (Ghermandi & Sinclair 2019; Toivonen et al. 2019). Social-networking data are usually available from dedicated social-media platforms. Twitter, Facebook, Instagram, TikTok, and Sina Weibo are among

the most popular worldwide (Statista 2019). However, platforms that specialize in other services (e.g., video and image sharing, news reporting, and blog hosting), such as Flickr, YouTube, and Blogger, can also provide social-networking features. Social-media data are usually composed of text, images, videos, or a combination of these (Table 1) and can be used for a wide range of potential applications in conservation that extend beyond culturomics research (Di Minin et al. 2015).

Data pertaining to both social media content and engagement can be used for a wide range of conservation purposes. These include analyzing species' popularity and associated sentiment (e.g., Roberge 2014; Kidd et al. 2018; Fink et al. 2020), monitoring wildlife trade online (e.g., Hinsley et al. 2016; Di Minin et al. 2019), studying the emergence of digital citizen science communities (Daume & Galaz 2016), and assessing nature-based recreational preferences (e.g., Hausmann et al. 2018; Monkman et al. 2018b).

Digital Encyclopedias

Encyclopedias are reference works that aim to compile human knowledge and, as such, are prime material for exploring aspects of human culture. Although several digital encyclopedias have emerged since the World Wide Web became publicly available, Wikipedia is the most widely used. Wikipedia is a free online encyclopedia curated by volunteers and currently composed of over 50 million entries in approximately 300 languages (Wikimedia 2020). Each Wikipedia entry contains text data describing the topic being addressed and may also feature a combination of image, video, and audio data that are freely available to anyone (Table 1). Data on public engagement with Wikipedia content is also openly available, including information on page views and edits. Because of these characteristics, Wikipedia data have been used widely in scientific research (Schroeder & Taylor 2015). However, other digital encyclopedias, such as Encyclopaedia Britannica (<https://www.britannica.com/>) or Everipedia (<https://everipedia.org/>), also represent potential corpora for culturomics research.

Data from digital encyclopedias can be used to explore various conservation issues, including the popularity of threatened species (Roll et al. 2016), the effect of nature documentaries on public interest toward featured species (Fernández-Bellon & Kane 2020), and seasonal dynamics of public interest in nature (Mittermeier et al. 2019; Vardi et al. 2021).

Temporal and Spatial Dimensions of Digital Corpora

Besides the content of digital corpora, many conservation applications can also benefit from associated temporal and spatial data (Fig. 1). Some questions can be

answered using spatially and temporally aggregated data (e.g., Jarić et al. 2019; Ladle et al. 2019), and many applications require detailed data on these dimensions. In such cases, researchers must consider that the sources of temporal and spatial information, and thus the coverage and resolution of the data, are highly specific to each corpus.

It is possible to obtain temporal data from different sources depending on the corpus (Table 1). Metadata often contains a timestamp for the date of creation, publication, or engagement. Content originally generated through digital platforms usually features temporal data at very fine resolutions – from minutes to seconds in the case of web searches or social media posts. For digitized corpora, such as books, only the year of publication may be available. Temporal coverage can also vary greatly; data derived purely from digital platforms usually cover only the period since the platform was created, but corpora emerging from digitization efforts, such as books, may span several decades or even centuries.

Spatial information may be immediately available if the corpus contains geolocated data or it may be derived from multiple other sources, such as internet provider location, user profiles, or the language or content of the corpus. Sources of spatial data vary among corpora (Table 1), and the resolution of derived data also varies among sources. Spatial coordinates obtained from geolocated data provide the highest spatial resolution (although not always precise at fine scale [Toivonen et al. 2019]), but such information is only available for a small fraction of all digital content. In contrast, data from some corpora may only be possible to map at the country or region level based on language groups or time zones (e.g., Mittermeier et al. 2019; Fink et al. 2020). Similarly, some digital corpora have near global coverage if the source platforms are used across the globe (e.g., Google Search Engine), whereas others may be used predominantly in certain countries or regions (e.g., Sina Weibo and Naver) or be subject to local access restrictions (e.g., countries blocking access to Wikipedia or Facebook).

Data Sources and Access

There are multiple ways to access digital data and often more than 1 platform offers access to the same corpora (Table 1). Hence, data may be compiled from a single (e.g., a news aggregator) or multiple sources (e.g., multiple individual news outlets). Data can be accessible through dedicated API services (e.g., YouTube API and Twitter API) and online interfaces for data access (e.g., Google Trends and GDELT). A list of relevant data sources is in Appendix S1. Some sources offer free access but may provide data in preanalyzed format (e.g., Google Trends and Google Ngram Viewer) or restrict the amount and type of data accessible (e.g., YouTube API and Twit-

ter API). Other services may charge for access but in exchange provide wider access to data (e.g., Webhose, DataStreamer, and DiffBot). It may also be possible to obtain raw data by scraping it directly from the web page if permitted—it is crucial to consult the Terms of Service and the robots.txt file (which contains instructions on what sections of each website can be crawled) prior to scraping. Data collection often requires good knowledge of web architecture and programming languages, such as R or Python, which can represent an initial barrier for conservation researchers wanting to engage with culturomics. Many books and online courses on programming languages and API architecture (e.g., RESTful APIs) can help overcome this initial barrier, but the development of data aggregation and access platforms with online user interfaces geared toward researchers (e.g., see <https://netlytic.org/index.php>) might facilitate this process even further.

Collecting raw data can often lead to large and unstructured data sets, so researchers should also consider how to store data and whether to subdivide it before storage. For example, in text corpora, it may be possible to identify and filter out homonyms (e.g., instances of the word *jaguar* that refer to the car brand rather than the animal) that are not relevant to the research focus before analysis (Roll et al. 2018). These limitations may force researchers to find a balance between their research budget, ease of data access and storage, and the type of data available when selecting which data to obtain. In extreme cases, the desired data may be inaccessible, access may cease, or the scale and content of the data may change during the project, requiring earlier decisions regarding the research design to be reassessed (Fig. 1). A good example of this problem is the social-networking platform Instagram, which substantially restricted access to public data at the end of 2018, following the Cambridge Analytica scandal (Bruns 2019). Other common instances that may disrupt research include adjustments to APIs and changes to data indexing and preprocessing procedures.

Extracting Metrics and Preparing Data for Analysis

Several metrics can be used in culturomics research for quantitative analysis of content and engagement with digital corpora (Table 2), both in absolute (count) or relative (frequency) terms. These metrics may be obtained from elements of corpora content and engagement, including volume, context, and interest. Many are also readily available to researchers in raw or preanalyzed formats including, for example, Wikipedia page edits, internet search volume, or YouTube video comments. However, metrics relating specifically to the content of digital corpora often need to be extracted from the corpus after data are collected. Recent advances in machine learning methods, namely, in computer vision and

Table 2. Examples of metrics of corpus content and engagement commonly used in culturomics research.

Metrics	Description	Example metrics
Volume	absolute number of items (e.g., web pages, videos, and news items) that constitute the corpus	number of web pages number of videos
Frequency	relative frequency with which entities and concepts are represented within items of the corpus	word frequency in texts frequency of entity representation in images
Sentiment	polarity, intensity, or type of sentiments and emotions expressed in the corpus or in engagements with the corpus	emotions expressed in news images sentiment polarity of social media posts
Context	items associated with an element in the corpus	word or topic associations temporal or geographic context in which an element appears
Interest	number or proportion of searches, shares, and likes associated with the corpus	number of internet searches number of social media shares
Discussion	number of discussions, comments, or edits to elements of the corpus	number of comments to news number of users editing digital encyclopedias

a) Natural language processing

1 - Original text

“Alarming IUCN report found that 58% of Europe's endemic tree species are now at risk of extinction”

2 - Sentiment analysis

Alarming : **NEGATIVE** IUCN report found that 58% of Europe's endemic tree species are now at risk : **NEGATIVE** of extinction

Sentiment class	Score
Positive	0
Negative	-2
Total	-2

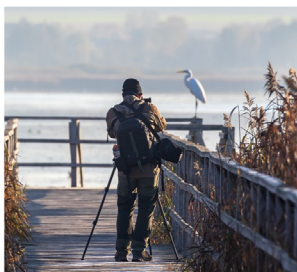
3 - Named entity recognition

Alarming IUCN : **ORGANIZATION** report found that 58% of Europe : **PLACE** Endemic : **CONCEPT** Tree : **CONCEPT** Species : **CONCEPT** are now at risk of Extinction : **CONCEPT**

Entity type	Count
Organization	1
Place	1
Concept	4

b) Computer vision

4 - Original image



5 - Instance segmentation



Items	Count
Bird	1
Person	1
Railing	2
Walkway	1

Figure 2. Applications of (a) natural language processing and (b) computer vision to quantification of corpora elements from text and images.

natural language processing, have greatly facilitated content analysis for large volumes of texts and images (e.g., Di Minin et al. 2018; Toivonen et al. 2019). Using natural language processing approaches, such as named entity recognition or sentiment analysis, allows the extraction of quantitative information on entities mentioned in texts and the sentiments expressed in relation to them (Fig. 2b & 2c). Similarly, using computer

vision algorithms (Fig. 2e) permits the identification and quantification of elements and sentiments expressed in images (Do 2019; Väisänen et al. 2021). Similar methods are being developed for sound and video data and are likely to become widespread in the near future, thus facilitating the large-scale analyses of these data formats (e.g., Kabra et al. 2012; Priyadarshani et al. 2018).

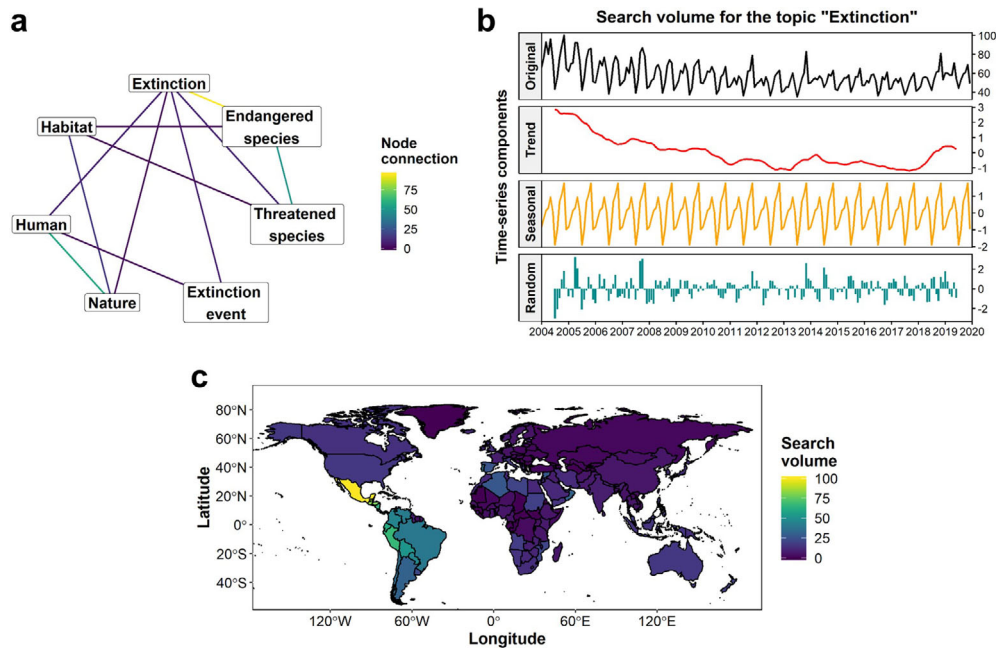


Figure 3. Examples of culturomics data analysis based on (a) network analysis, (b) time-series decomposition, and (c) choropleth maps. The examples are based on data obtained from Google Trends for the topic extinction (interest over time, interest by region, and top related topics) to demonstrate the range of analytical options available for data extracted from a single corpus.

Methods for Analyzing Culturomics Data

Culturomics analyses usually draw on a wide range of statistical methods to describe, classify, and make inferences from corpus metrics. Descriptive statistics can be used to summarize the main features of the data, often supported by graphical methods, such as histograms, box plots, and scatter plots (Heumann et al. 2016). Topic modeling is also a useful method for text corpora that may be used in a preliminary stage to identify relevant data for further analysis and storage or as the focus of analysis to identify the core topics in the corpus (Westgate et al. 2015). Network analysis can be used to analyze the co-occurrence patterns of entities in the corpus (Jarić et al. 2021) or be combined with topic modeling to explore connections between key topics (Fig. 3a). Regression-analysis methods, from generalized linear models to machine learning models (Ciaburro 2018), can help researchers make inferences on how culturomics metrics relate to other variables of interest, which usually include biological, social, cultural, and geographical factors. These methods can be used to identify traits driving species popularity in the public eye (e.g., Roll et al. 2016; Ladle et al. 2019; Vardi et al. 2021) or landscape factors associated with public preferences for protected areas (e.g., Hausmann et al. 2017; Correia et al. 2018b).

In temporal analyses of culturomics data, time series plots can help researchers explore and visualize tem-

poral patterns in the data. Long time-series data are usually composed of long-term trends, short-term cyclical and seasonal variation, and a random component (Fig. 3b). Methods to decompose time-series data to its multiple elements can be used to explore each component in detail (Cowpertwait & Metcalfe 2009). Auto-correlation functions can be used to explore the existence of temporal dependence in the time-series data, and cross-correlation functions are useful for exploring the relationship between multiple time series of interest. These functions can be used, for instance, to explore the temporal relationship between online news and public interest in conservation topics and charismatic species (e.g., Nghiem et al. 2016; Fink et al. 2020). Autoregressive integrated moving average models are commonly used to explore time-series data, but there are a range of other modeling approaches available, including machine-learning-based methods (Cowpertwait & Metcalfe 2009). Approaches based on change-point detection can identify shifts in temporal trends over time and be used to explore the role of specific events in influencing public interest trends toward conservation topics (e.g., Correia et al. 2019a). In such cases, adopting a counterfactual approach to the analysis may help generate more robust inferences from the data (e.g., Verissimo et al. 2020).

Spatial analyses of culturomics data commonly involve the use of geographical information systems to visualize and map location- and area-based data. Point-location data can be plotted directly on a map

or summarized over relevant spatial areas with point-density analysis and plotted using heat maps (Shook et al. 2012). Area- or region-based data can also be mapped onto the relevant spatial units with choropleth, cartogram, or proportional symbol maps (Fig. 3c). Bivariate maps can be used when the analysis focuses on more than 1 spatial variable (e.g., Archibald & Butt 2018). In such cases, spatial modeling methods ranging from geographically weighted regression, to generalized additive models, to Bayesian spatial modeling (Blangiardo & Cameletti 2015) may be used to statistically explore the relationship between variables.

Challenges for Culturomics Analyses

Interdisciplinarity

There are several recognized applications of culturomics to conservation that have not been well explored and many more applications that are likely to emerge in the coming years. These include recognizing conservation-oriented constituencies and promoting public understanding of conservation issues (Ladle et al. 2016). For example, exploring and shaping the evolution of conservation culture (Lennox et al. 2020), particularly outside of the academic environment, may be a subject worthy of further exploration with culturomics approaches. In-depth exploration of these topics in conservation would greatly benefit from expertise in areas such as cultural evolution, digital humanities, media studies, social marketing, linguistics, and psychology. This clearly illustrates the interdisciplinary nature of many conservation culturomics projects and highlights the necessity for collaborations to enhance the reach, scope, and impact of future conservation applications.

Ethical Issues

The wealth of digital data available on the internet can be a source of potentially sensitive information, and researchers need to carefully consider the ethical implications of using such information. There are numerous examples of recorded illegal activity in digital content, including illegal wildlife trade and illegal hunting (e.g., El Bizri et al. 2015; Hinsley et al. 2016; Di Minin et al. 2018). Much digital data contain personal information, including names, locations, or photographs, that could conceivably be used to directly identify specific individuals. Indirect identifiers, including workplace, occupation, and residence, may also be available and can allow the extrapolation of direct personal information even when it has been obscured (Monkman et al. 2018a). The use of digital data for research is usually permitted in legal frameworks, especially if such data are publicly available, but the usage of personal information is more

sensitive and subject to specific national or regional legislation (e.g., the European Union's General Data Protection Regulation; see Di Minin et al. 2021). When personal data are collected, researchers are required to respect users' privacy by anonymizing or pseudonymizing the data during or immediately after collection (Monkman et al. 2018a). Researchers should also consider which data to publish, including indirect identifiers, in order to protect the personal identity of research subjects (Monkman et al. 2018a). The same rationale applies to sensitive information about threatened species (Lindenmayer & Scheele 2017), such as identity, date, and location, especially when these data are recent and not easily accessible.

Inherent Biases in the Data

Although more than half of the human population is now connected to the internet, access and participation in the digital realm can differ significantly within and between regions, including many regions of the world where conservation is a priority. Gender, age, education, and other socioeconomic, cultural, political, and geographical factors are often important drivers of data availability and representativeness (Graham et al. 2015). For instance, traditional and indigenous people play a critical role in biodiversity conservation (Kohler & Brondizio 2017), but their interactions with nature are frequently underrepresented in digital data. These biases are likely to be similar to those associated with biological recording and citizen science (e.g., Geldmann et al. 2016; Correia et al. 2019b), which conservation researchers are more familiar with. Existing solutions to account for biases in these research areas may be used to inform conservation culturomics research and other emerging areas of inquiry drawing from similar data sources (Jarić et al. 2020). Furthermore, existing biases should not discourage the use of digital data but rather spur the development of methods that can generate inferences from multiple data sources (e.g., Vieira et al. 2018). This will allow scarce research resources to be redirected toward obtaining data from less represented populations to ensure that all relevant views are considered.

Data Validation

Given the large volume of available digital data and its potential to generate quick and large-scale insights on conservation issues, it may be tempting to proceed through data gathering and analysis without considering the need for validation. However, the algorithms used to sample or generate the data for analyses are not always transparent, which can make it difficult to identify the main driver of observed patterns (Ficetola 2013; Correia 2019). Language complexity, including synonyms, homonyms, negation, and sarcasm, can also

introduce additional noise and may require careful data sampling, evaluation, and filtering prior to analysis (e.g., Correia et al. 2017; Correia et al. 2018a; Roll et al. 2018). There are also increasing volumes of digital content generated by automatized bots, which may be present in the data and influence analytical outcomes. Therefore, data and results validation are key aspects of any conservation culturomics project. Ideally, results obtained using digital data should be validated with data from independent nondigital data sources (e.g., Hausmann et al. 2018; Verissimo et al. 2020), but that is not always possible due to a lack of suitable independent data. One alternative is to use data from multiple sources to ensure that the results returned by different corpora agree (e.g., Cooper et al. 2019; Jarić et al. 2019; Vardi et al. 2021), a process that can be considered a form of triangulation with digital data (Leckner & Severson 2019). Similarly, it may be possible to obtain robust inferences from corpora composed of multiple data types (e.g., text, image, sound, etc.) by combining their analyses. Accounting for multimodality (i.e., the presence of more than 1 data type) is an emerging topic of research in automated content analysis (Ramachandram & Taylor 2017) that is likely to be of relevance for conservation culturomics.

Data Sharing and Standards

The dynamic nature of digital data sources poses an important challenge for conservation culturomics. Changes in data access can affect ongoing projects, as highlighted above, and can also prevent the results of culturomics research from being reproduced (Troumbis 2019) if the original data used in the study cannot be recovered. Access to unstructured data sets generated by scraping online resources can be particularly volatile because web pages may change frequently, but even APIs and dedicated data-access platforms are updated regularly. Researchers can prevent this problem by sharing data and code in open repositories whenever possible (some data sources do not allow the redistribution of original data). One way to stimulate such efforts is to develop standards for culturomics data sharing that are applicable to multiple types of data, similar to efforts developed for biodiversity data (e.g., Wiczorek et al. 2012).

Advancing Conservation Culturomics

Conservation culturomics is likely to advance rapidly in the coming years. Our overview of the culturomics analytical framework from research planning, to data acquisition, to data analysis aims to support further developments and applications in conservation. Overcoming practical challenges is only the first step in advancing conservation culturomics, however. At least 3 additional actions are needed to ensure such developments

have real impact on conservation. First, culturomics techniques and other emerging digital applications of relevance to conservation (e.g., Jarić et al. 2020) should be included in conservation education curricula. Allowing new generations of conservationists to become familiar with digital methods will greatly facilitate their widespread adoption and application. Second, the uptake of conservation culturomics needs to extend beyond the academic realm. Nongovernmental organizations, conservation managers, and decision makers need to be stimulated to engage with culturomics techniques to ensure that their potential benefits and impact on conservation are fully realized. Finally, establishing partnerships with the private sector will be essential to draw upon the wider universe of digital data available as a result of the digital revolution. Data from several digital and tech companies are available for purchase, but this may not be viable for conservation institutions with already scarce resources. Developing corporate social-responsibility projects focusing on providing data access for conservation goals may be a suitable way forward. Digital corpora and internet participation will continue to expand, bringing new opportunities and increasing the power of culturomics research to identify patterns in human-nature interactions that are relevant to conservation.

Acknowledgments

R.A.C. thanks the Helsinki Institute of Sustainability Science (HELSUS) and the University of Helsinki for funding to E.D.M. R.L. and A.C.M.M. are funded by the Brazilian National Council for Scientific and Technological Development, CNPq (#309980/2018-6, #309879/2019-1, and #400325/2014-4). I.J. was supported by the J. E. Purkyně Fellowship of the Czech Academy of Sciences. U.R. is partially supported by GIF Research Grant number I-2519-119.4/2019. C.F. thanks the University of Helsinki for a grant to E.D.M. E.D.M. and A.H. thank the European Research Council (ERC) for funding under the European Union's Horizon 2020 research and innovation program (grant agreement 802933). P. Jepson and 3 anonymous reviewers provided insightful feedback on earlier drafts of the manuscript. This manuscript also benefitted from discussions within the Society for Conservation Biology (SCB) Conservation Culturomics Working Group.

Supporting Information

Additional information is available online in the Supporting Information section at the end of the online article. The authors are solely responsible for the content and functionality of these materials. Queries (other than

absence of the material) should be directed to the corresponding author.

Literature Cited

- Acerbi A. 2019. Cultural evolution in the digital age. Oxford University Press, Oxford, United Kingdom.
- Archibald CL, Butt N. 2018. Using Google search data to inform global climate change adaptation policy. *Climatic Change* **150**:447–456.
- Bennett NJ, et al. 2017. Conservation social science: understanding and integrating human dimensions to improve conservation. *Biological Conservation* **205**:93–108.
- Blangiardo M, Cameletti M. 2015. Spatial and spatio-temporal Bayesian models with R — INLA. Wiley, Chichester, United Kingdom.
- Braczkowski A, et al. 2018. Reach and messages of the world's largest ivory burn. *Conservation Biology* **32**:765–773.
- Bruns A. 2019. After the 'APocalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society* **22**:1544–1566.
- Ciaburro G. 2018. Regression analysis with R. Packt Publishing, Birmingham, United Kingdom.
- Cisco. 2018. Cisco Visual Networking Index: forecast and trends, 2017–2022 white paper, Available from <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html> (accessed October 2020).
- Cooper MW, Di Minin E, Hausmann A, Qin S, Schwartz AJ, Correia RA. 2019. Developing a global indicator for Aichi Target 1 by merging online data sources to measure biodiversity awareness and engagement. *Biological Conservation* **230**:29–36.
- Correia RA. 2019. Google trends data need validation: comment on Durmuşoğlu (2017). *Human and Ecological Risk Assessment: An International Journal* **25**:787–790.
- Correia RA, Di Minin E, Jarić I, Jepson P, Ladle R, Mittermeier J, Roll U, Soriano-Redondo A, Verissimo D. 2019a. 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. 2018a. Nomenclature instability in species culturomic assessments: why synonyms matter. *Ecological Indicators* **90**:74–78.
- Correia RA, Jepson P, Malhado ACM, Ladle RJ. 2016. Familiarity breeds content: assessing bird species popularity with culturomics. *PeerJ* **4**:e1728.
- Correia RA, Jepson P, Malhado ACM, Ladle RJ. 2017. Internet scientific name frequency as an indicator of cultural salience of biodiversity. *Ecological Indicators* **78**:549–555.
- Correia RA, Jepson P, Malhado ACM, Ladle RJ. 2018b. Culturomic assessment of Brazilian protected areas: exploring a novel index of protected area visibility. *Ecological Indicators* **85**:165–171.
- Correia RA, Ruete A, Stropp J, Malhado ACM, dos Santos JW, Lessa T, Alves JA, Ladle RJ. 2019b. Using ignorance scores to explore biodiversity recording effort for multiple taxa in the Caatinga. *Ecological Indicators* **106**:105539.
- Cowpertwait PSP, Metcalfe AV. 2009. Introductory time series with R. Springer, New York.
- Daume S, Galaz V. 2016. Anyone know what species this is? – Twitter conversations as embryonic citizen science communities. *PLOS ONE* **11** (e0151387).
- Di Minin E, Fink C, Hausmann A, Kremer J, Kulkarni R. 2021. How to address data privacy concerns when using social media data in conservation science. *Conservation Biology*.
- Di Minin E, Fink C, Hiippala T, Tenkanen H. 2019. A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology* **33**:210–213.
- Di Minin E, Fink C, Tenkanen H, Hiippala T. 2018. Machine learning for tracking illegal wildlife trade on social media. *Nature Ecology & Evolution* **2**:406–407.
- Di Minin E, Tenkanen H, Toivonen T. 2015. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science* **3**:63.
- Do Y. 2019. Valuating aesthetic benefits of cultural ecosystem services using conservation culturomics. *Ecosystem Services* **36**:100894.
- Do Y, Kim JY, Lineman M, Kim DK, Joo GJ. 2015. Using internet search behavior to assess public awareness of protected wetlands. *Conservation Biology* **29**:271–279.
- El Bizri HR, Morcatty TQ, Lima JJS, Valsecchi J. 2015. The thrill of the chase: uncovering illegal sport hunting in Brazil through YouTube posts. *Ecology and Society* **20**:30.
- Fernández-Bellón D, Kane A. 2020. Natural history films raise species awareness—a big data approach. *Conservation Letters* **13**: e12678.
- Ficetola GF. 2013. Is interest toward the environment really declining? The complexity of analysing trends using internet search data. *Biodiversity and Conservation* **22**:2983–2988.
- Fink C, Hausmann A, Di Minin E. 2020. Online sentiment towards iconic species. *Biological Conservation* **241**:108289.
- Francis FT, et al. 2019. Shifting headlines? Size trends of newsworthy fishes *PeerJ* **7**:e6395.
- Fukano Y, Soga M. 2019. Spatio-temporal dynamics and drivers of public interest in invasive alien species. *Biological Invasions* **21**:3521–3532.
- Funk SM, Rusowsky D. 2014. The importance of cultural knowledge and scale for analysing internet search data as a proxy for public interest toward the environment. *Biodiversity and Conservation* **23**:3101–3112.
- Geldmann J, Heilmann-Clausen J, Holm TE, Levinsky I, Markussen B, Olsen K, Rahbek C, Tøttrup AP, Leung B. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions* **22**:1139–1149.
- Ghermandi A, Sinclair M. 2019. Passive crowdsourcing of social media in environmental research: a systematic map. *Global Environmental Change* **55**:36–47.
- Graham M, De Sabbata S, Zook MA. 2015. Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment* **2**:88–105.
- Harrison JR, Roberts DL, Hernandez-Castro J. 2016. Assessing the extent and nature of wildlife trade on the dark web. *Conservation Biology* **30**:900–904.
- Hausmann A, Toivonen T, Heikinheimo V, Tenkanen H, Slotow R, Di Minin E. 2017. Social media reveal that charismatic species are not the main attractor of ecotourists to sub-Saharan protected areas. *Scientific Reports* **7**:763.
- Hausmann A, Toivonen T, Slotow R, Tenkanen H, Moilanen A, Heikinheimo V, Di Minin E. 2018. Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters* **11**:e12343.
- Hayes D, Cappa F, Cardon J. 2018. A framework for more effective dark web marketplace investigations. *Information* **9**:186.
- Heumann C, Schomaker M, Shalabh. 2016. Introduction to statistics and data analysis: with exercises, solutions and applications in R. Springer, New York.
- Hinsley A, Lee TE, Harrison JR, Roberts DL. 2016. Estimating the extent and structure of trade in horticultural orchids via social media. *Conservation Biology* **30**:1038–1047.
- International Telecommunication Union. 2020. Individuals using the Internet, 2005–2019. Available from <https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx> (accessed January 2020).
- Jarić I, et al. 2020. iEcology: harnessing large online resources to generate ecological insights. *Trends in Ecology & Evolution* **35**:630–639.

- Jarić I, et al. 2021. Invasion culturomics and iEcology approaches to better understand biological invasions. *Conservation Biology*.
- Jarić I, Correia RA, Roberts DL, Gessner J, Meinard Y, Courchamp F. 2019. On the overlap between scientific and societal taxonomic attentions — insights for conservation. *Science of the Total Environment* **648**:772–778.
- Kabra M, Robie AA, Rivera-Alba M, Branson S, Branson K. 2012. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature Methods* **10**:64–67.
- Kesebir S, Kesebir P. 2017. A growing disconnection from nature is evident in cultural products. *Perspectives on Psychological Science* **12**:258–269.
- Kidd LR, Gregg EA, Bekessy SA, Robinson JA, Garrard GE. 2018. Tweeting for their lives: visibility of threatened species on Twitter. *Journal for Nature Conservation* **46**:106–109.
- Kitchin R. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society* **1**. <https://doi.org/10.1177/2053951714528481>.
- Kohler F, Brondizio ES. 2017. Considering the needs of indigenous and local populations in conservation programs. *Conservation Biology* **31**:245–251.
- 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**:270–276.
- Ladle RJ, Jepson P, Correia RA, Malhado ACM. 2017. The power and the promise of culturomics. *Frontiers in Ecology and the Environment* **15**:290–291.
- Ladle RJ, Jepson P, Correia RA, Malhado ACM, Gould R. 2019. A culturomics approach to quantifying the salience of species on the global internet. *People and Nature* **1**:524–532.
- Lansdall-Welfare T, Sudhahar S, Thompson J, Lewis J, Cristianini N. 2017. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences of the United States of America* **114**:E457–E465.
- Leckner S, Severson P. 2019. Exploring the meaning problem of big and small data through digital method triangulation. *Nordicom Review* **40**:79–94.
- Lennox RJ, Verissimo D, Twardek WM, Davis CR, Jarić I. 2020. Sentiment analysis as a measure of conservation culture in scientific literature. *Conservation Biology* **34**:462–471.
- Lindenmayer D, Scheele B. 2017. Do not publish. *Science* **356**:800–801.
- Measey J, Basson A, Rebelo AD, Nunes AL, Vimercati G, Louw M, Mohanty NP. 2019. Why have a pet amphibian? Insights from YouTube. *Frontiers in Ecology and Evolution* **7**:52.
- Megias DA, Anderson SC, Smith RJ, Verissimo D. 2017. Investigating the impact of media on demand for wildlife: A case study of Harry Potter and the UK trade in owls. *Plos One* **12**:e0182368.
- Merriam-Webster. 2020. Dictionary: corpus. Available from <https://www.merriam-webster.com/dictionary/corpus> (accessed January 2020).
- Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J. 2011. Quantitative analysis of culture using millions of digitized books. *Science* **331**:176–182.
- Miranda EBP, Ribeiro RP, Strüssmann C. 2016. The ecology of human-anaconda conflict: a study using internet videos. *Tropical Conservation Science* **9**:43–77.
- 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.
- Monkman GG, Kaiser M, Hyder K. 2018a. The ethics of using social media in fisheries research. *Reviews in Fisheries Science & Aquaculture* **26**:235–242.
- Monkman GG, Kaiser MJ, Hyder K. 2018b. Text and data mining of social media to map wildlife recreation activity. *Biological Conservation* **228**:89–99.
- Nghiem LTP, Papworth SK, Lim FKS, Carrasco LR. 2016. Analysis of the capacity of Google Trends to measure interest in conservation topics and the role of online news. *PLOS ONE* **11** (e0152802). <https://doi.org/10.1371/journal.pone.0152802>.
- Papworth SK, Nghiem TPL, Chimalakonda D, Posa MRC, Wijedasa LS, Bickford D, Carrasco LR. 2015. Quantifying the role of online news in linking conservation research to Facebook and Twitter. *Conservation Biology* **29**:825–833.
- Priyadarshani N, Marsland S, Castro I. 2018. Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology* **49**:jav-01447.
- Proulx R, Massicotte P, Pepino M. 2014. Googling trends in conservation biology. *Conservation Biology* **28**:44–51.
- Ramachandram D, Taylor GW. 2017. Deep multimodal learning: a survey on recent advances and trends. *IEEE Signal Processing Magazine* **34**:96–108.
- Richards DR. 2013. The content of historical books as an indicator of past interest in environmental issues. *Biodiversity and Conservation* **22**:2795–2803.
- Roberge J-M. 2014. Using data from online social networks in conservation science: which species engage people the most on Twitter? *Biodiversity and Conservation* **23**:715–726.
- Roll U, Correia RA, Berger-Tal O. 2018. Using machine learning to disentangle homonyms in large text corpora. *Conservation Biology* **32**:716–724.
- 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**:42–50.
- Sbragaglia V, Correia RA, Coco S, Arlinghaus R, Schmidt J. 2019. Data mining on YouTube reveals fisher group-specific harvesting patterns and social engagement in recreational anglers and spearfishers. *ICES Journal of Marine Science* <https://doi.org/10.1093/icesjms/fsz100>.
- Schroeder R, Taylor L. 2015. Big data and Wikipedia research: social science knowledge across disciplinary divides. *Information, Communication & Society* **18**:1039–1056.
- Schultz PW. 2011. Conservation means behavior. *Conservation Biology* **25**:1080–1083.
- Schwartz T. 2011. Culturomics: periodicals gauge culture's pulse. *Science* **332**:35–36.
- Sherrin K, Smit M, Holmlund M, Parkins JR, Chen Y. 2017. Conservation culturomics should include images and a wider range of scholars. *Frontiers in Ecology and the Environment* **15**:289–290.
- Shook E, Leataru K, Cao G, Padmanabhan A, Wang S. 2012. Happy or not: generating topic-based emotional heatmaps for culturomics using CyberGIS. Pages 1–6 in 2012 IEEE 8th International Conference on E-Science.
- Soriano-Redondo A, Bearhop S, Lock L, Votier SC, Hilton GM. 2017. Internet-based monitoring of public perception of conservation. *Biological Conservation* **206**:304–309.
- Statista. 2019. Most popular social networks worldwide as of October 2019, ranked by number of active users. Available from <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> (accessed January 2020).
- Stergiou KI. 2017. The most famous fish: human relationships with fish as inferred from the corpus of online English books (1800–2000). *Ethics in Science and Environmental Politics* **17**:9–18.
- Toivonen T, Heikinheimo V, Fink C, Hausmann A, Hiippala T, Järvi O, Tenkanen H, Di Minin E. 2019. Social media data for conservation science: a methodological overview. *Biological Conservation* **233**:298–315.
- Troumbis AY. 2019. The time and timing components of conservation culturomics cycles and scenarios of public interest in the Google era. *Biodiversity and Conservation* **28**:1717–1727.
- Väisänen T, Heikinheimo V, Hiippala T, Toivonen T. 2021. Discovering human–nature interactions in national parks using social media photographs and computer vision. *Conservation Biology*.

- Vardi R, Mittermeier JC, Roll U. 2021. Combining culturomic sources to uncover trends in popularity and seasonal interest in plants. *Conservation Biology*.
- Veríssimo D, Anderson S, Tlustý M. 2020. Did the movie *Finding Dory* increase demand for blue tang fish? *Ambio* 49:903–911.
- Veríssimo D, MacMillan DC, Smith RJ, Crees J, Davies ZG. 2014. Has climate change taken prominence over biodiversity conservation? *Bioscience* 64:625–629.
- Vieira FAS, Bragagnolo C, Correia RA, Malhado ACM, Ladle RJ. 2018. A salience index for integrating multiple user perspectives in cultural ecosystem service assessments. *Ecosystem Services* 32:182–192.
- Westgate MJ, Barton PS, Pierson JC, Lindenmayer DB. 2015. Text analysis tools for identification of emerging topics and research gaps in conservation science. *Conservation Biology* 29:1606–1614.
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. 2012. Darwin Core: an evolving community-developed biodiversity data standard. *PLOS ONE* 7 (e29715). <https://doi.org/10.1371/journal.pone.0029715>.
- Wikimedia. 2020. List of Wikipedias. Available from https://meta.wikimedia.org/wiki/List_of_Wikipedias (accessed January 2020).

