off before they even begin. Smart African women are leading research teams, advising governments, and increasingly assuming positions of power. Young African female scientists can increasingly find the role models they need. More women need to step up to the call to leadership, supported by their partners and the broader society. Boards and expert committees must insist on gender balance.

Acknowledgments

F.H.A.O. is supported by a Sofja Kovalevskaja Award from the Alexander von Humboldt Foundation (3.2 - 1184811 - KEN - SKP) and an EDCTP Senior Fellowship (TMA 2015 SF1001) which is part of the EDCTP2 programme supported by the European Union.

Declaration of Interests

The authors declare no competing financial or personal interests.

Resources

[i] www.kemri.org/
[ii] www.waccbip.org/
[iii] www.strathmore.edu/
[iv] www.iavi.org/
[v] www.worldbank.org/
[vi] www.ukri.org/
[vii] https://europa.eu/european-union/about-eu/funding-grants_en
[viii] www.edctp.org/
[ix] www.nih.gov/grants-funding
[x] www.aasciences.africa/aesa/programmes/developing-excellence-leadership-training-and-science-africa-deltas-africa
[xi] www.aasciences.africa/
[xii] www.usaid.gov/

[1] Centre for Infectious Diseases, Heidelberg University Hospital, Heidelberg, Germany
[2] KEMRI-Wellcome Trust Research Programme, Centre for Geographic Medicine Research-, Coast, Kilifi, Kenya
[3] International AIDS Vaccine Initiative (IAVI), Nairobi, Kenya

*Correspondence:
fosier@kemri-wellcome.org (F.H.A. Osier).

https://doi.org/10.1016/j.pt.2021.01.004

References

1. Ballabeni, A. *et al.* (2016) Time to tackle the incumbency advantage in science: A survey of scientists shows strong support for funding policies that would distribute funds more evenly among laboratories and thereby benefit new and smaller research groups. *EMBO Rep.* 17, 1254–1256
2. Daniels, R.J. (2015) A generation at risk: young investigators and the future of the biomedical workforce. *Proc. Natl. Acad. Sci. U. S. A.* 112, 313–318
3. Liani, M.L. *et al.* (2020) Understanding intersecting gender inequities in academic scientific research career progression in sub-Saharan Africa. *Int. J. Gend. Sci. Technol.* 12, 27
4. Chrousos, G.P. and Mentis, A.A. (2020) Imposter syndrome threatens diversity. *Science* 367, 749–750
5. Somerville, L.H. and Gruber, J. (2020) Three trouble spots facing women in science – and how we can tackle them. *Science – Letters to Young Scientists* Published online October 16, 2020. https://doi.org/10.1126/science.caredit.abf3010
6. Fathima, F.N. *et al.* (2020) Challenges and coping strategies faced by female scientists – A multicentric cross sectional study. *PLoS One* 15, e0238635
7. Boussemart, L. (2016) Woman, mother, and scientist: Aiming to fulfill a career in research while maintaining a 'good-enough' work-life balance. *Int. J. Womens Dermatol.* 2, 74–76
8. Mackay, F. (2020) Not from Venus, not from Mars – all equally superstars. *Nat. Immunol.* 21, 238
9. Gautier, L. *et al.* (2018) Deconstructing the notion of 'global health research partnerships' across Northern and African contexts. *BMC Med. Ethics* 19, 49
10. Roper, R.L. (2019) Does gender bias still affect women in science? *Microbiol. Mol. Biol. Rev.* 83, e00018-19
11. Casad, B.J. *et al.* (2021) Gender inequality in academia: Problems and solutions for women faculty in STEM. *J. Neurosci. Res.* 99, 13–23
12. Langin, K. (2019) How scientists are fighting against gender bias in conference speaker lineups. *Science – Workplace Diversity* Published online February 11, 2019. https://doi.org/10.1126/science.caredit.aaw9742
13. Witteman, H.O. *et al.* (2019) Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *Lancet* 393, 531–540
14. Hinton, D.P.R. (2000) *Stereotypes, Cognition and Culture*, Psychology Press
15. Hinton, P. (2017) Implicit stereotypes and the predictive brain: cognition and culture in 'biased' person perception. *Palgrave Commun.* 3, 17086

Forum

iParasitology: Mining the Internet to Test Parasitological Hypotheses

Robert Poulin,[1,3,*]
Jerusha Bennett,[1,3]
Antoine Filion,[1,3]
Upendra Raj Bhattarai,[2]
Xuhong Chai,[1]
Daniela de Angeli Dutra,[1]
Erica Donlon,[1]
Jean-François Doherty,[1]
Fátima Jorge,[1] Marin Milotic,[1]
Eunji Park,[1]
Amandine Sabadel,[1] and
Leighton J. Thomas[1]

Digital data (internet queries, page views, social media posts, images) are accumulating online at increasing rates. Tools for compiling these data and extracting their metadata are now readily available. We highlight the possibilities and limitations of internet data to reveal patterns in host–parasite interactions and encourage parasitologists to embrace iParasitology.

The Rise of Online Digital Data

In recent years, internet searches have become the primary source of information for everyone, while various online repositories have amassed huge amounts of digital data (i.e., text, photos, and videos) that are publicly available on platforms like webpages and social media. Billions of people are indirectly acting as recorders of information about the natural world, sometimes unknowingly taking part in a global citizen science programme. Ecologists [1,2] and public health scientists [3,4] have recently begun to exploit these data sources, which were often generated for other purposes and represent a kind of 'passive crowdsourcing'. For example, photos uploaded to the internet by scuba divers have served to better delimit the geographical range of the black spot syndrome caused by trematode infections in Caribbean reef fishes [5]. Similarly, data from internet search activity for 'West Nile virus' reliably captured the seasonal peaks observed in actual cases reported to the US Centers for Disease Control and Prevention [6].

Here, inspired from iEcology [1], we introduce iParasitology, or internet parasitology (Figure 1). We define iParasitology as the study of patterns and processes in parasitology using online data stored digitally, publicly available, and often but not always generated for other purposes. We propose ways in which parasitologists could make use of online data to test

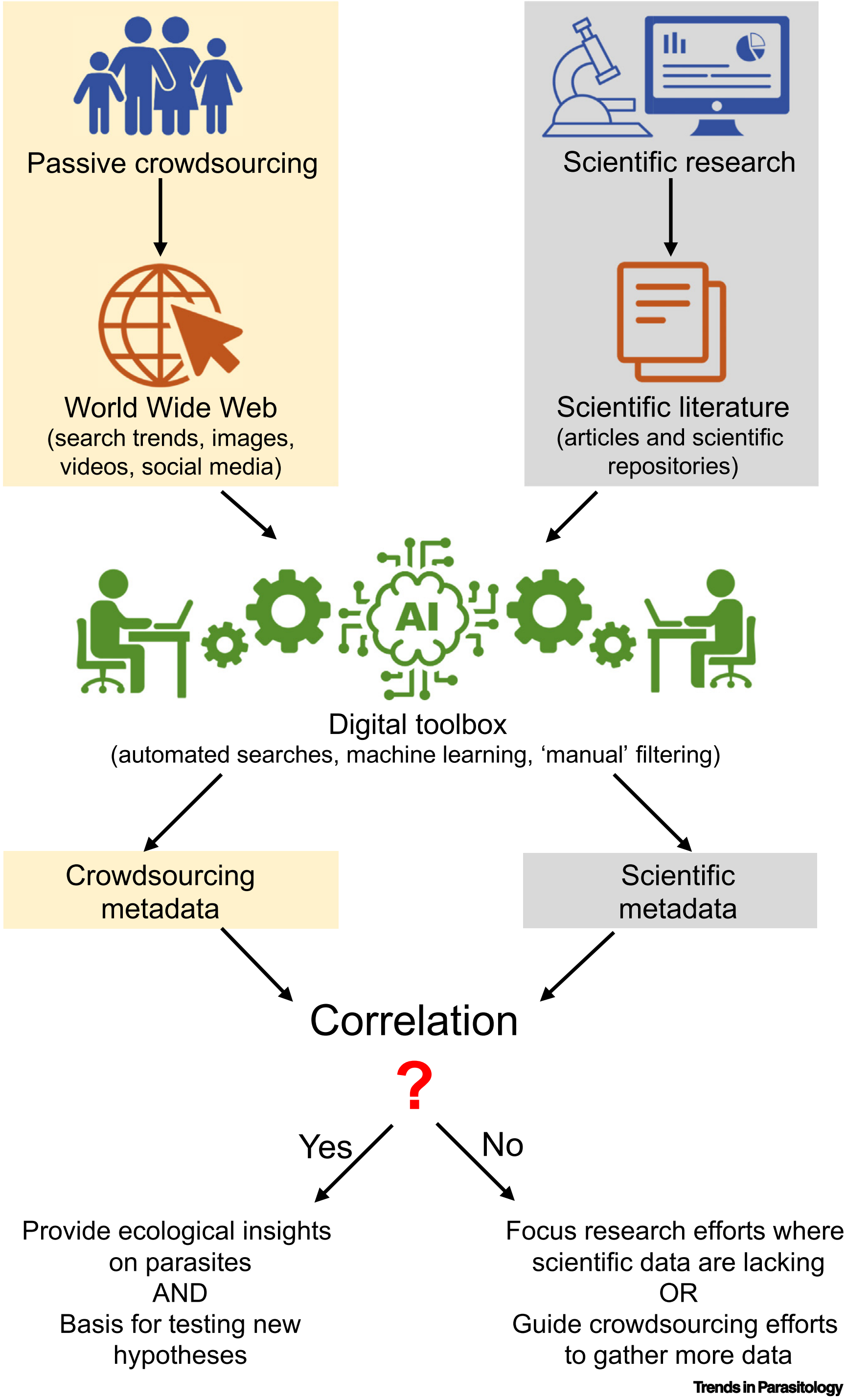

Passive crowdsourcing
Scientific research
World Wide Web
(search trends, images, videos, social media)
Scientific literature
(articles and scientific repositories)
AI
Digital toolbox
(automated searches, machine learning, 'manual' filtering)
Crowdsourcing metadata
Scientific metadata
Correlation
?
Yes
No
Provide ecological insights on parasites
AND
Basis for testing new hypotheses
Focus research efforts where scientific data are lacking
OR
Guide crowdsourcing efforts to gather more data
Trends in Parasitology

Figure 1. Conceptual Flowchart of iParasitology. Data acquisition (blue) is compiled into different resources (orange), depending on the source of information, that is, the public at large or the scientific community. Metadata can be extracted using powerful digital algorithms or filtered manually (green). Comparative analyses between different metadata can determine their correlation. Positive correlations between crowdsourced and scientific datasets not only validate the use of iParasitology but may also provide novel parasitological insights and drive hypothesis testing. Conversely, no correlation could focus research or crowdsourcing efforts where data are poor or lacking.

hypotheses regarding temporal dynamics of infection, geographical distributions of parasites, and the frequency of various host–parasite associations. We discuss some potential applications, identify some powerful tools that can facilitate the use of massive online data sources, and discuss some of the limitations associated with these data. Throughout, we emphasise the need to ground-truth patterns generated by online data with traditional scientific data. If and when the former are validated by the latter, harnessing the full breadth of online data would open up scales of study currently out of reach to most parasitological researchers.

Mining Data from Internet Search Activity

People's curiosity and their quest for answers and knowledge on the internet generate useful data. Internet search activity and information uploaded and shared in social media can be used to study patterns in parasitology through passive crowdsourcing. These data can be obtained and compiled from various search engines (e.g., Google, Bing, Baidu), online encyclopaedias (e.g., Wikipedia and Encyclopaedia Britannica), and social media (e.g., Twitter, Facebook). All of these are in the top 15 most popular websites in the world[i], with Google being at number one. Google Trends[ii] is a free tool that shows real-time global search activity, with a user-friendly interface that allows spatiotemporal data to be graphically displayed. With metadata on where and when each Google search was made, the user can investigate patterns in parasitology ranging from disease monitoring [3,6,7] to predicting outbreaks [4]. For example, Google Trends data have been used to predict tick-borne encephalitis; the authors found a significant correlation between Google search trends and weekly case numbers reported by clinicians [8]. As long as one corrects the absolute number of searches on a topic for the overall growth in internet search volume, Google Trends data hold much promise for iParasitology. To date, this spontaneous citizen science has been mostly ignored for its potential to inform research in parasitology.

With 14–16 billion page views per month[iii], Wikipedia is one of the most frequently visited websites on the internet[i]. One may obtain data for when and where each page was accessed, providing useful data to examine infection patterns. For example, data on the frequency of Wikipedia page views have been used to study seasonal trends in influenza [9] and other human diseases [10]. These data capture what people experience and might therefore reflect real-world patterns. Wikipedia page views could thus be used to explore temporal patterns in the incidence of various parasites infecting animals, including occurrences in host species or localities previously not recorded in the scientific literature.

Lastly, social media have become hugely popular and also provide a source of information to examine patterns in parasitology. Twitter provides an application programming interface (API) allowing access to data on the frequency of tweets on various topics, which may be processed either manually or through machine-learning algorithms. For instance, tweets relating to coronavirus disease 2019 (COVID-19) symptoms have been explored as an approach to assess disease severity [11]. Importantly, whether data are extracted from tweets or browser queries, they must be validated by comparison with data acquired by traditional and rigorous scientific means. For any given question about host–parasite interactions, demonstrating that large-scale patterns derived from internet data are congruent with those obtained by scientific studies on smaller scales would go a long way to establish the validity of iParasitology as an alternative approach.

Parasites in Pictures: Internet Photo Data

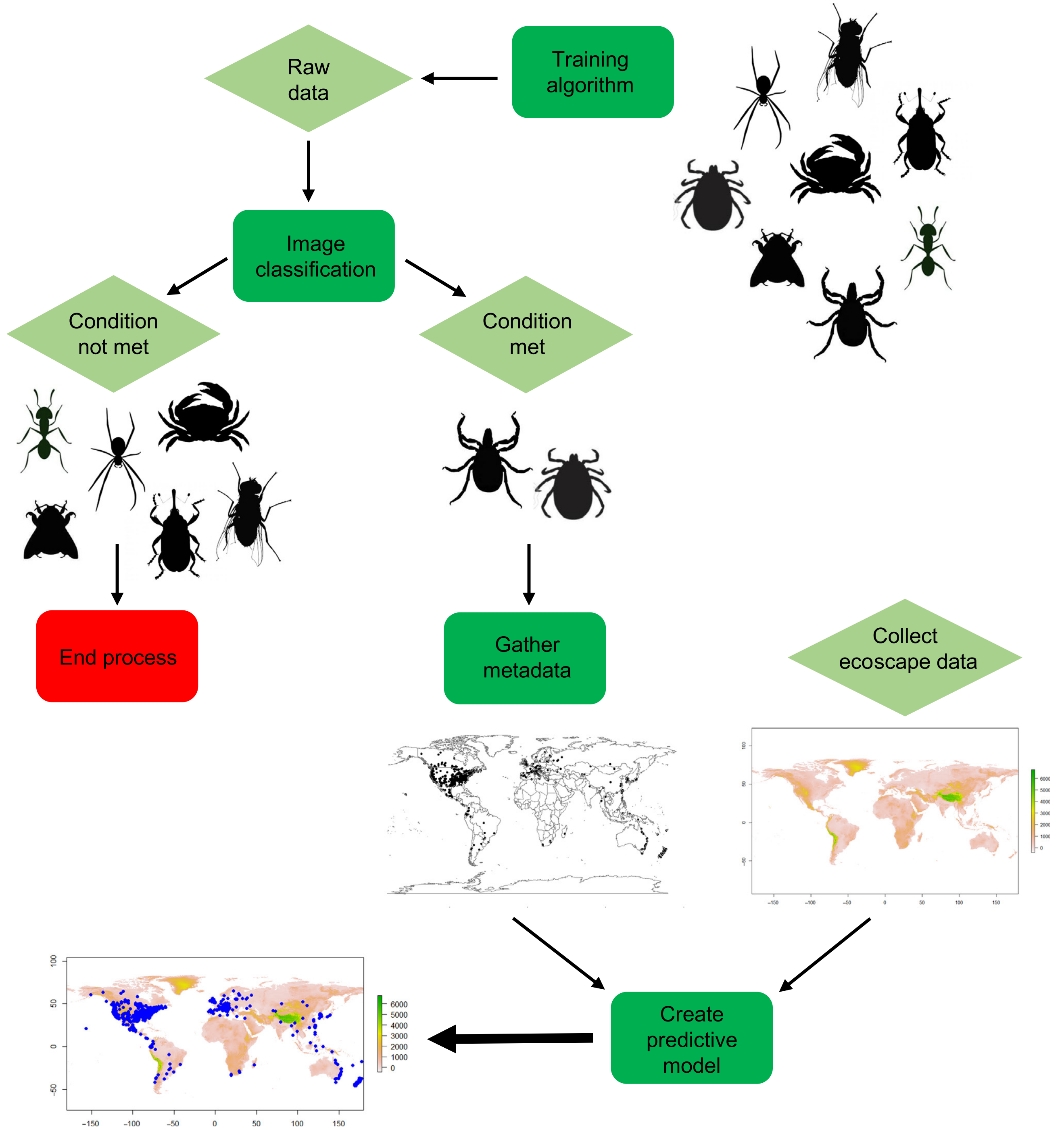
Image and video data are constantly being uploaded worldwide, and researchers can exploit these public data to answer scientific questions. Since the photos themselves only serve to generate information but are not republished, copyright issues do not apply. Photos and videos can provide information regarding host and parasite species identities, interactions, behaviours, and overall ecosystem health. Beyond the image itself, they can hold large amounts of metadata, including where and when they were taken. These resources have already provided some important contributions to species conservation [12], health science [13], and ecology [2].

Many different online platforms can provide photo and video data, such as Google Images[iv] and YouTube[v], the largest search engine of online images and the largest repository of videos[i], respectively. Flickr[vi] and Instagram[vii] are also major photo repositories and therefore important data sources. In addition, there are more specialised databases such as iNaturalist[viii] which gathers scientific data collected by citizens. To date, iNaturalist has more than 53 million photo observations of more than 30 000 species worldwide. While sources such as YouTube and Google Images have larger user bases and associated data, they require more data collation effort than specialised sources such as iNaturalist, where pictures are uploaded for the purpose of species identification. For example, a simple search for ‘tick’ on iNaturalist yields more than 18 000 observations, whereas a Google Images search provides 8800 images, including many irrelevant observations. These platforms differ in their popularity, purpose, accessibility, and associated metadata, providing complementary sources of image data. As for internet search activity, the data can be easily gathered via manual searches, whereas automated methods, based on coding, may require more data curation.

While photo and video resources have been adopted in several other fields [1], they have rarely been used by parasitologists. Provided they are cross-validated through comparison with data obtained by traditional science, incorporating image data into parasitological research could provide information pertaining to the taxonomic identity of parasites and their hosts, stage and status of infection, distribution, and seasonality. As mentioned earlier [5], Google Images searches have been used to characterise spatial and temporal infection patterns in reef fishes. These resources are freely available worldwide with no need to sample, handle, or sacrifice animals. Information collated from online image repositories can validate existing scientific knowledge, as well as expand it by providing new host or locality records for certain parasites. A picture is worth a thousand words, indeed.

Toolbox for Compiling and Analysing Internet Data

The past few years have seen a massive increase in the use of technological tools to acquire open-access data. Compared with traditional methods of data gathering that are often time consuming and require great effort (e.g., manual compiling) to generate adequate datasets, open-source software such as R provides end-users with a structured environment in which to extract, visualize, and analyse large volumes of data. For example, accurate measurement of landscape heterogeneity using MODIS satellite imaging[ix,x] and other climatic variables can now be easily extracted for any geographic location around the globe at the click of a mouse. In turn, these variables can be modelled using species-occurrence probability models, allowing the end-user to extract complex



Trends in Parasitology

Figure 2. Flowchart for the Analysis of Crowdsourced Image Data. In this example of one possible type of iParasitology study, photographic data are collected from passive crowdsourcing (e.g., Google Images search) using 'ticks' as keyword. In turn, the end user creates a training dataset based on distinctive criteria, for example, body size, number of legs, etc., with associated labels for each image using a deep-learning framework. After the model is validated to the satisfaction of the end-user, it is then applied to the prediction dataset, which classifies the images as meeting the conditions of the trained dataset or not. Afterward, for each image meeting the conditions set by the user, useful R packages (see main text) can be used to extract metadata from large numbers of images or videos, allowing analysis of geographical occurrence patterns or seasonal trends in tick abundance. Going further, predictive modelling packages such as Maxlike[xvii] can then be used to create species-occurrence probabilities based on climatic conditions at any given set of coordinates.

Box 1. Tips for iParasitology

(i) Before proceeding further, clearly define the goals and scope of the research, and ascertain that these can be achieved in principle through iParasitology.

(ii) Focusing on a smaller, more regional scale will help to standardise data, by minimising variation in socio-economic and educational levels, access to technology and internet, and languages used.

(iii) Conduct searches in a language appropriate for the target region or country, and not strictly in English.

(iv) Widen search terms, and where possible include all common names for the focal parasite, not just its scientific name.

(v) To correct for variable data quality, assign a reliability or validity score to each record. For instance, when using machine learning to recognise relevant images, a reliability score could be assigned to each image retained based on its resolution (number of pixels) or size.

(vi) Get all records (or a representative subset) validated by an expert. For instance, ask a taxonomist to confirm the identity of parasites seen in images retained by an automated search.

(vii) To reduce temporal and spatial autocorrelations and data nonindependence (e.g., same person uploading multiple photos of the same phenomenon over multiple days), integrate the data over a longer period of time, and/or use appropriate statistical methods.

(viii) Reduce error and bias in data from passive crowdsourcing (nonrandom sampling, false records, etc.) using occupancy models[xviii] that include adjustments for data features that violate statistical assumptions.

interactions between parasite distribution patterns and their surrounding ecoscape. This is one of many situations where a novel approach to big parasitological data allows extrapolation of complex patterns that a traditional method would not easily achieve.

Additionally, the increase in computational power now allows automated data collection and analysis. For instance, deep-learning algorithms (a type of machine learning) are now becoming more user-friendly and proving to be powerful tools with applications in many fields of research [14], with an error rate close to human visual error [15]. Computer vision (e.g., deep-learning algorithms using features of images for classification) can be trained to recognise subsets of images of interest (i.e., those showing a particular species) out of a huge number of images, and rapidly categorise large amounts of digital material (Figure 2). Useful R packages such a gtrendsR[xi], pageviews[xii], rtweet[xiii], photosearcher[xiv], ggmap[xv], or tuber[xvi] can extract metadata from Google Trends, Wikipedia, Twitter, or from large numbers of images or videos, and easily generate datasets on either geographical occurrences or seasonal trends in parasite abundance. This allows researchers to work with massive amounts of data that would otherwise be very difficult to analyse [15]. With technological tools equipped with a user-friendly interface becoming increasingly available, we encourage parasitologists to consider such approaches for mass data gathering and analysis from online sources.

Limitations of Internet Data

Despite its potential, there are several limitations and biases associated with the use of internet-based data for iParasitology. They depend on the question asked (e.g., occurrence or seasonality of a given parasite), and on the data source explored, whether photos, text analysis, website statistics (number of visits, search trends, etc.), media coverage, or a combination. Often, these limitations make it impossible to achieve worldwide coverage.

First, the availability of information is highly unequal amongst countries, a digital divide that causes massive data gaps in some regions of the world. These are due to many variables, including: spatiotemporal variation in population density, in socioeconomic or education levels; accessibility to technology or the internet; universality of smartphone usage. Additionally, social media use can vary among countries, some having preferred platforms while others experience heavy restrictions on social media usage. Second, beyond these differences among countries, language is the second main source of bias. Limiting search terms to a single language like English is highly restrictive, missing out on potentially relevant data tagged or written in other languages. Third, not all data found on the internet can be trusted. Incorrect captions, poor descriptions, and false entries can all generate unreliable search results. Fourth, heterogeneous photo quality can affect image searches made using machine-learning approaches. Low-quality images, or images that have been heavily altered or compressed, might not be picked up by the algorithm. Fifth, from a statistical perspective, the data may suffer from nonindependence, if multiple entries for a particular parasite were made by the same user, or if a photo or text has been uploaded (copied) in more than one repository. Sixth, access to some data, such as the number of views of specific webpages, is not always available directly; one might need to obtain access from the webmaster or perhaps even pay a fee.

Furthermore, whereas the above limitations are broadly applicable to citizen science in general, some apply specifically to iParasitology. Lay people will likely only encounter, notice, and report the most visible parasite species or life stages; for instance, online data will probably accrue more readily on ectoparasites than endoparasites. The lack of clear distinguishing morphological features complicates the identification of many parasites, even at a coarse taxonomic level. Finally, people are generally more interested in what happens to them and their pets, thus biasing parasite reporting online toward infections in humans and domestic animals like cats and dogs, rather than infections of wildlife. Nevertheless, some of the limitations associated with internet data can be overcome (Box 1), allowing parasitologists to fully harness this massive data source.

Concluding Remarks

Online data, unknowingly contributed by anyone with access to the internet, are recorded on a scale that exceeds what any funded science project could achieve through traditional data collection methods. Some early studies provide a proof of concept that these data are potentially useful resources for parasitologists [5]. Often, iParasitology data are likely to be of a qualitative rather than quantitative nature; for instance, they may document the presence of a parasite in an area, rather than its numerical abundance. Nevertheless, occurrence records from iParasitology sources may complement, rather than duplicate, those obtained using traditional scientific approaches. Many highly visible and intriguing parasites have generated much public interest, leading to internet searches and image uploads; these would be great candidates for hypothesis-testing based on online data. They include parasites of wildlife (e.g., isopod and copepod ectoparasites of fish; mermithid nematodes and nematomorphs in insects; etc.), domestic animals (e.g., ticks or *Toxocara* nematodes in cats and dogs), and even humans (e.g., headlice in schoolchildren). In addition, online data could be used to monitor changes in host distribution or abundance, or in human activities and societal behaviour, that may contribute to infection risk and the spread of parasites. The necessary tools to implement iParasitology are now freely available, and we hope this forum article will stimulate parasitologists to make full use of the treasure trove of data that is the internet.

Acknowledgments

This article was prepared during a working retreat hosted by the New Zealand Marine Studies Centre (Portobello) and the Quarantine Island Lodge (Otago Harbour). We thank Rob Lewis for his time and invaluable assistance, the Emerson's Brewery, Dunedin, for generous liquid sponsorship, and two anonymous reviewers for useful and constructive comments.

Declaration of Interests

The authors declare no conflict of interest.

Resources

[i] https://en.wikipedia.org/wiki/List_of_most_popular_websites
[ii] www.google.com/trends
[iii] https://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm
[iv] https://images.google.com
[v] www.youtube.com
[vi] www.flickr.com
[vii] www.instagram.com
[viii] www.inaturalist.org
[ix] www.earthenv.org/
[x] www.worldclim.org/
[xi] https://cran.r-project.org/web/packages/gtrendsR/index.html
[xii] https://cran.r-project.org/web/packages/pageviews/index.html
[xiii] https://cran.r-project.org/web/packages/rtweet/index.html
[xiv] https://github.com/ropensci/photosearcher
[xv] https://cran.r-project.org/web/packages/ggmap/index.html
[xvi] https://cran.r-project.org/web/packages/tuber/index.html
[xvii] https://cran.r-project.org/web/packages/maxlike/index.html
[xviii] https://rdrr.io/cran/unmarked/man/unmarked-package.html

[1] Department of Zoology, University of Otago, P.O. Box 56, Dunedin, New Zealand
[2] Department of Anatomy, University of Otago, P.O. Box 56, Dunedin, New Zealand
[3] These authors contributed equally to this article

*Correspondence:
robert.poulin@otago.ac.nz (R. Poulin).

https://doi.org/10.1016/j.pt.2021.01.003

References

1. Jarić, I. *et al.* (2020) iEcology: harnessing large online resources to generate ecological insights. *Trends Ecol. Evol.* 35, 630–639
2. Mikula, P. *et al.* (2018) Large-scale assessment of commensalistic–mutualistic associations between African birds and herbivorous mammals using internet photos. *PeerJ* 6, e4520
3. Ning, S. *et al.* (2019) Accurate regional influenza epidemics tracking using internet search data. *Sci. Rep.* 9, 5238
4. Aiello, A.E. *et al.* (2020) Social media- and internet-based disease surveillance for public health. *Annu. Rev. Publ. Health* 41, 101–118
5. Elmer, F. *et al.* (2019) Black spot syndrome in reef fishes: using archival imagery and field surveys to characterize spatial and temporal distribution in the Caribbean. *Coral Reefs* 38, 1303–1315
6. Carneiro, H.A. and Mylonakis, E. (2009) Google Trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin. Infect. Dis.* 49, 1557–1564
7. Milinovich, G.J. *et al.* (2014) Using internet search queries for infectious disease surveillance: screening diseases for suitability. *BMC Infect. Dis.* 14, 690
8. Sulyok, M. *et al.* (2020) Predicting tick-borne encephalitis using Google Trends. *Ticks Tick Borne Dis.* 11, 101306
9. Hickmann, K.S. *et al.* (2015) Forecasting the 2013–2014 influenza season using wikipedia. *PLoS Comput. Biol.* 11, e1004239
10. Vilain, P. *et al.* (2017) Wikipedia: a tool to monitor seasonal diseases trends? *Online J. Publ. Health Inform.* 9, e052
11. Mackey, T. *et al.* (2020) Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: retrospective big data infoveillance study. *JMIR Publ. Health Surveill.* 6, e19509
12. Otsuka, R. and Yamakoshi, G. (2020) Analyzing the popularity of YouTube videos that violate mountain gorilla tourism regulations. *PLoS ONE* 15, e0232085
13. van Heerden, A. and Young, S. (2020) Use of social media big data as a novel HIV surveillance tool in South Africa. *PLoS ONE* 15, e0239304
14. Christin, S. *et al.* (2019) Applications for deep learning in ecology. *Methods Ecol. Evol.* 10, 1632–1644
15. Wäldchen, J. and Mäder, P. (2018) Machine learning for image-based species identification. *Methods Ecol. Evol.* 9, 2216–2225