

Oct 23, 2017

Characterizing Biodiversity Data Leaks

DOI

<https://dx.doi.org/10.17504/protocols.io.kebctan>

A. Townsend Peterson¹, Alex Asase², Dora L. Canhos³, Sidnei de Souza³, John Wieczorek⁴

¹University of Kansas; ²University of Ghana; ³Centro de Referência em Informação Ambiental;

⁴University of California at Berkeley



A. Townsend Peterson

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.kebctan>

Protocol Citation: A. Townsend Peterson, Alex Asase, Dora L. Canhos, Sidnei de Souza, John Wieczorek 2017. Characterizing Biodiversity Data Leaks. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.kebctan>

Manuscript citation:

Peterson, A. T., A. Asase, D. L. Canhos, S. de Souza, and J. Wieczorek. In review. Data Leakage and Loss in Biodiversity Informatics. *Biodiversity Data Journal*.

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited



Protocol status: Working

Created: October 23, 2017

Last Modified: February 25, 2018

Protocol Integer ID: 8355

Keywords: biodiversity, data, characterizing biodiversity data leaks biodiversity data, biodiversity data leaks biodiversity data, scale biodiversity data set, scale biodiversity data sets in term, boon to biodiversity science, biodiversity science, data record, utility of the data, leakage, record, proportion of those record, usable for science application

Abstract

Biodiversity data exist in large quantities, which is a boon to biodiversity science. In spite of large numbers of data records being available, however, the proportion of those records that is readily usable for science applications is quite small. The difference between the full number of data records existing *versus* the records that are ready for use is the result of what we call "leakage" of data, in the form of steps that have not been taken or errors that remove the utility of the data. In this contribution, we explore several large-scale biodiversity data sets in terms of why they contain data records that are or are not ready for use.

Guidelines

GBIF Data Precautions: We note that data from the GBIF data portal are subjected to the GBIF taxonomic filtering (Gaiji et al. 2013), although our experience indicates that the GBIF filters apply to species-based searches, but not to database-level or region-based searches, such that the data analyzed herein have not to our knowledge been subjected to these filters. For the Brazilian Virtual Herbarium, names are from Brazilian Flora 2020 (<http://floradobrasil.jbrj.gov.br>) and Catalogue of Life (<http://www.catalogueoflife.org/>), in that order). We did not consider the potential for an expert to review and identify the specimen fully as rescuable, as that step would extend beyond the data to actual handling of the specimen, or at least detailed inspection of images by specialists.

Georeferencing Caveats: We required that georeferences demonstrate the characteristics of best practice, as outlined in the MaNIS georeferencing protocol. These data were considered as complete only when geographic coordinates were accompanied by full metadata, such that information was present in the fields `coordinatePrecision` and `coordinateUncertaintyInMeters`, as this information is crucial to many applications of these data in biodiversity informatics applications, preventing misinterpretation of coarse-resolution coordinates.

- 1 Download datasets from online sources: [VertNet](#) and [Global Biodiversity Information Facility](#).
- 2 Each record is analyzed with respect to completeness and fitness for use in terms of time (i.e., day, month, year, eventDate), taxon, and place (textual geographic information, geographic coordinates), considering 4 categories of completeness and fitness for use: 0 = information missing completely, 1 = information partial, 2 = information incomplete but with sufficient information that it could be “rescued” and brought to completeness (we deemed information as “rescuable” when information can be improved or corrected, such as by georeferencing textual geographic information quantitatively, or by correcting a scientific name that is not a standard name), and 3 = information complete and ready for use. Details for processing each dimension are in the succeeding steps.

Software	
Access	NAME
Windows 10	OS
Microsoft	DEVELOPER
https://products.office.com/en-us/access	SOURCE LINK

- 3 Time: Data were considered partial when information on day, month, year, or their equivalent in eventDate was missing. Data were considered as rescuable when full information appeared to be present in verbatimEventDate, but was not parsed appropriately into day, month, and year, or eventDate.

Software	
Access	NAME
Windows 10	OS
Microsoft	DEVELOPER
https://products.office.com/en-us/access	SOURCE LINK

- 4 Taxon: Data were considered missing if no genus-level information existed. Data were considered partial if identified to genus but not to species. Data were considered rescuable if not a name listed in at least one taxonomic authority. For birds, ornithological authorities checked included Peters (1931-1987), Sibley and Monroe (1990), Clements (2007), and Gill and Donsker (2016). The rescuable/complete distinction was possible only for ornithological data; for plants, no global species names authority lists were available, so we considered all full Latin binomials as complete.

Software	
Access	NAME
Windows 10	OS
Microsoft	DEVELOPER
https://products.office.com/en-us/access	SOURCE LINK

- 5 Place: Data were considered missing when geographic coordinates were lacking, and textual geographic descriptions lacked information more precise than state. Data were considered partial when information was available at the level of county/municipality, but not to the level of a specific locality. Data were considered rescuable when the locality was described fully in textual terms, but geographic coordinates were missing, or when geographic coordinates were not completely documented. We also included data records as rescuable (not complete) in terms of place when the coordinates were inconsistent—e.g., the coordinate information fell in a different country from the country information given in the data record.

Software	
Access	NAME
Windows 10	OS
Microsoft	DEVELOPER
https://products.office.com/en-us/access	SOURCE LINK

- 6 Usability analyses: We considered two scenarios: (1) ecological niche modeling and species distribution modeling require information on place and taxon, and (2) evaluations of inventory completeness require information on time, taxon, and place. To combine information across multiple dimensions, we took the minimum value of the 4-level categorization given above across the two or three dimensions.

Software	
Access	NAME
Windows 10	OS
Microsoft	DEVELOPER
https://products.office.com/en-us/access	SOURCE LINK