Implementing Nestedness in Darwin Core: An epidemiology case study

Francesca Jaroszynska[‡], Guillaume Body[‡], Sophie Pamerlon[§,|], Anne-Sophie Archambeau[¶,|]

‡ OFB (Office français de la biodiversité), Pérols, France
§ OFB (Office français de la biodiversité), Paris, France
| GBIF France (Global Biodiversity Information Facility), Paris, France
¶ IRD (Institut de recherche pour le développement), Paris, France

Corresponding author: Francesca Jaroszynska (francesca.jaroszynska@ofb.gouv.fr)

Abstract

Wildlife diseases have an impact on biodiversity, the economy, and public health. Better knowledge of disease patterns would be to the benefit of conservation measures, livestock production and, thus, ultimately human health. Nevertheless, disease surveillance systems operate mostly at a national scale, with incompatible data structures, inhibiting effective data sharing and thus rapid transnational responses to disease outbreak. As the risk of disease to biodiversity, the economy and to public health increases with climate change, land-use change and trade, the necessity for a common data standard to improve data sharing of surveillance efforts is greater than ever to enable transnational proactive and reactive measures to be taken.

To address these large issues, a consortium for the European Food and Safety Authority (EFSA; the Enetwild consortium) was formed to collect existing data on wildlife host abundance and distribution in Europe. Alongside data on their associated pathogens, the consortium is attempting to develop data models to aggregate the host data according to the Darwin Core*[1] standard.

However, the complexity of zoonotic disease and wildlife distribution data is not easily captured in the current version of the Darwin Core standard, often comprising complex data structures with species interactions and partial information. Firstly, zoonotic disease data consist of observations of both the host and the pathogen, whereby each positive case of a disease is associated with the observation of the host species. Secondly, wildlife host distribution data frequently contain detailed information for only some individuals of a group when multiple individuals of the host species are observed simultaneously, such as life stage and sex. In order to capture these types of interactions and subsetting, we need to be able to handle the hierarchical structuring of these data.

In an attempt to resolve these issues, a data model in line with the Darwin Core standard was initially developed for wildlife host population data (Enetwild Consortium et al. 2020). Here, we propose a new data model for data on the hosts' pathogens, and use both data

models to demonstrate the model efficacy for complex hierarchical data structures. The epidemiology data model is structured around the existence of a primary occurrence of the host species observation, which may consist of one or many individuals. In order to associate the presence (or absence) of the pathogen to the host, or to provide details on the host group composition, the data model allows child occurrences to relate to the primary, or parent, occurrence at a particular event (Table 1). We propose the implementation of a hierarchical structure of the occurrence extension to allow these two phenomena to be effectively modelled. This hierarchical structuring relies on the adoption of a new term, the 'parentOccurrenceID', whereby each 'parent' observation of the host species can be associated with multiple 'child' observations. To this end, we propose the introduction of the parentOccurrenceID term (currently under discussion on GitHub).

Using the Darwin Core standard, we propose a data model that would allow effective harmonisation of zoonotic disease data, demonstrating that harmonisation of disease surveillance data in Europe is possible. Finding solutions to capturing complex hierarchical biotic interactions in Darwin Core is currently underway at GBIF*[2], [3], although relying on a separate relationship table. However, we advocate for the introduction of the new parentOccurrenceID thanks to its simplicity and very general applicability, being adaptable to any group occurrence where detailed or partial information is available, or to hierarchical interaction relationships such as between hosts and pathogens. We believe the employment of the new term would be complementary to the current GBIF developments, and of benefit to many Darwin Core users where details apply to differing levels of hierarchical occurrences.

Keywords

wildlife management, nested relationship, zoonotic disease surveillance

Presenting author

Francesca Jaroszynska

Presented at

TDWG 2022

Conflicts of interest

The authors declare no conflicts of interest.

References

- Enetwild Consortium, Body G, Mousset M, Chevallier E, Scandura M, Pamerlon S, Blanco-Aguiar JA, Vicente J (2020) Applying the Darwin core standard to the monitoring of wildlife species, their management and estimated records. EFSA Supporting Publications 17 (4). https://doi.org/10.2903/sp.efsa.2020.en-1841

Endnotes

*1 https://www.tdwg.org/standards/dwc/

*2 https://docs.google.com/document/d/1jzb54GbAkB_TOFIjWof5BW6gn1ujSnXXJQu6aZgFAB4/edit?usp=sharing

*3 https://www.gbif.org/fr/composition/HjlTr705BctcnaZkcjRJq/data-model

Table 1.

Hierarchical structuring of the Occurrence extension, accommodating the presence of the host species (*Sus scrofa*) and the pathogen (African Swine Fever Virus).

occurrenceID	parentOccurrenceID	scientificName	individualCount	occurrenceStatus	lifeStage	sex
roadkill1		*Sus scrofa*	11			
roadkill1-1	roadkill1	*Sus scrofa*	3		adult	female
roadkill1-2	roadkill1	*Sus scrofa*	5		adult	male
roadkill1-3	roadkill1	*Sus scrofa*	3		undetermined	undetermined
roadkill1-1-1	roadkill1-1	African Swine Fever Virus		present		