Tackling Data Quality Challenges in the Finnish Biodiversity Information Facility (FinBIF)

Kari M Lahti[‡], Mikko Heikkinen[‡], Aino Juslén[‡], Leif Schulman[§,|]

‡ Finnish Museum of Natural History, Helsinki University, Finland
§ Finnish Environment Institute, Helsinki, Finland
| University of Helsinki, Helsinki, Finland

Corresponding author: Kari M Lahti (kari.lahti@helsinki.fi)

Abstract

The Finnish Biodiversity Information Facility (FinBIF) Research Infrastructure (Schulman et al. 2021) is a national service with a broad coverage of the components of biodiversity informatics (Bingham et al. 2017). Data flows are managed under a single information technology (IT) architecture. Services are available in a single, branded on-line portal. Data are collated from all relevant sources e.g., research institutes, scientific collections, public authorities and citizen science projects, whose data represent a major contribution. The challenge is to analyse, classify and share good quality data in a way that the user understands its utility.

Need for quality data

The philosophy of FinBIF is that all observation records are important, and that all data are assessed for quality and able to be annotated. The challenge is that, in practice, many users desire data with 100% reliability. In our experience, most user concerns about data quality are related to citizen science data. Researchers are usually able to manage raw data to serve their purposes. However, decision-making authorities often have less capacity to analyse the data and thus require data that can be used instantly. Therefore, we need tools to provide users the data that are the most relevant and reliable for their specific use. For all users, standardized metadata (information about datasets) are key, when the user has doubts about the fitness-for-use of a particular dataset. There is also a need to provide data in different formats to serve various users. Finally, the service has to be machine-actionable (using an application programming interface (API) and R-package) as well as human-accessible for viewing and downloading data.

Quality assignment

FinBIF data accuracy varies significantly within and between datasets, and observers. Two quality-based classifications suitable for filtering are therefore applied. The **dataset origin filter** is based on the quality of a whole dataset (e.g. citizen science project) and includes three broad classes assigned with an appropriate quality label: Datasets by Professionals,

by Specialists and by Citizen Scientists. The **observation reliability filter** is based on a single observation and on annotations by FinBIF users. This classification includes Expert verified, Community verified, Unassessed (default for all records), Uncertain, and Erroneous. The dataset origin does not necessarily determine the quality of the individual records in it. Observations made by citizen scientists are often accurate, while there may be errors in the professionally collected data. Records are frequently subject to annotation, which raises their quality over time (e.g., iNaturalist). Naturally, evidence (e.g., media, detailed descriptions, specimens) is needed for reliable identification.

Annotating data

When observations are compiled at FinBIF's portal (Laji.fi), they are initially "Unassessed" (unless they have otherwise been assessed at the original source). When annotating occurrences, volunteers can make various entries using the tools provided. The aim of the commentary is to improve the quality of the observation data. Annotators are divided into two categories with two different roles:

1. As a **basic user**, anyone who has logged in at Laji.fi can make comments or tag observations for review by experts.
2. Users defined as **experts** have wider rights than basic users and their comments carry more weight. The most desired actions of expert users are to classify observations into confidence levels or to give them new or refined identifications.

Information about new comments passes to the observer if the observation is recorded by using the FinBIF Observation Management System "Notebook". However, comments cannot yet be automatically forwarded e.g., to the primary data management systems at the original source.

Annotations add extra indications of quality. They do not replace or delete the original information. Nevertheless, annotations can change a record's taxonomic identification, and by default, a record will be handled based on its latest identification.

R-package for researchers and Public Authority Portal (PAP) for decision makers

FinBIF has produced an R programming language interface to its API, which makes the publicly available data in FinBIF accessible from within R. For authorities, the PAP offers direct access to all available species information to authorised users, including sensitive and restricted-use data.

Keywords

citizen science, biodiversity data quality, research infrastructure

Presenting author

Kari M Lahti

Presented at

TDWG 2021

Conflicts of interest

References

- Bingham H, Doudin M, Weatherdon L, Despot-Belmonte K, Wetzel F, Groom Q, Lewis E, Regan E, Appeltans W, Güntsch A, Mergen P, Agosti D, Penev L, Hoffmann A, Saarenmaa H, Geller G, Kim K, Kim H, Archambeau A, Häuser C, Schmeller D, Geijzendorffer I, García Camacho A, Guerra C, Robertson T, Runnel V, Valland N, Martin C (2017) The Biodiversity Informatics Landscape: Elements, Connections and Opportunities. Research Ideas and Outcomes 3: e14059. https://doi.org/10.3897/rio.3.e14059
- Schulman L, Lahti K, Piirainen E, Heikkinen M, Raitio O, Juslén A (2021) The Finnish Biodiversity Information Facility as a best-practice model for biodiversity data infrastructures. Scientific data 8 (1): 137. https://doi.org/10.1038/s41597-021-00919-6