PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20240
doi: 10.3897/tdwgproceedings.1.20240

OPEN ACCESS

Conference Abstract

Fitness for Use: The BDQIG aims for improved Stability and Consistency

Arthur D. Chapman[‡], Antonio Mauro Saraiva[§], Lee Belbin[|], Allan Veiga[¶], Miles Nicholls[#], Paula F Zermoglio[¤], Paul J. Morris[«], Dmitry Schigel[»], Alexander Thompson[ˆ]

‡ Australian Biodiversity Information Services, Ballan, Victoria, Australia
§ Universidade de São Paulo, São Paulo, Brazil
| Blatant Fabrications Pty Ltd, Carlton, Australia
¶ University of São Paulo, São Paulo, Brazil
# Atlas of Living Australia, Canberra, Australia
¤ Universidad de Buenos Aires, Buenos Aires, Argentina
« Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States of America
» Global Biodiversity Information Facility - Secretariat, Copenhagen Ø, Denmark
ˆ iDigBio, Gainesville, United States of America

Corresponding author: Arthur D. Chapman (biodiv_2@achapman.org)

Received: 13 Aug 2017 | Published: 14 Aug 2017

Citation: Chapman A, Saraiva A, Belbin L, Veiga A, Nicholls M, Zermoglio P, Morris P, Schigel D, Thompson A (2017) Fitness for Use: The BDQIG aims for improved Stability and Consistency. Proceedings of TDWG 1: e20240. https://doi.org/10.3897/tdwgproceedings.1.20240

Abstract

The process of choosing data for a project and then determining what subset of records are suitable for use has become one of the most important concerns for biodiversity researchers in the 21st century. The rise of large data aggregators such as GBIF (Global Biodiversity Information Facility), iDigBio (Integrated Digitized Biocollections), the ALA (Atlas of Living Australia) and its many clones, OBIS (Ocean Biogeographic Information System), SIBBr (Sistema de Informação sobre a Biodiversidade Brasileria), CRIA (Centro de Referência em Informação Ambiental) and many others has made access to large volumes of data easier, but choosing which data are fit for use remains a more difficult task. There has been no consistency between the various aggregators on how best to clean and document the quality – how tests are run, or how annotations are stored and reported. Feedback to data custodians on possible errors has been minimal, inconsistent, and adherence to recommendations and controlled vocabularies (where they exist) has been haphazard to say the least.

The TDWG Data Quality Interest Group is addressing these issues, either alone or in conjunction with other Interest Groups (Annotations, Darwin Core, Invasive Species, Citizen Science and Vocabulary Maintenance) to develop a framework, tests and assertions, use cases and controlled vocabularies. The Interest Group is also working closely with the data aggregators toward consistent implementations. The practical work is being done through five Task Groups. A published framework is leading to a user-friendly Fitness for Use Backbone (FFUB) and data quality profiles by which users can document the quality they need for a project. A standard set of core tests and assertions has been developed around the Darwin Core standard and are currently being tested and integrated into several aggregators. A use case library has been compiled and these cases will lead to themed data quality profiles as part of the FFUB. Two new Task Groups are being established to develop controlled vocabularies to address the inconsistencies in values of at least 40 Darwin Core terms. These inconsistencies make the evaluation of fitness for use far more difficult than achieved by using controlled vocabularies. The first TG is looking at vocabularies generally, while the second is looking at those just pertaining to Invasive Species.

It is not just the aggregators though that are the stakeholders in this work. The data custodians and even the collectors have a vested interest in ensuring their data and metadata are of highest quality and therefore seeing their data used widely. It is only after aggregation that many uses of the data become apparent, and most collectors aren't aware of these uses at the time of collecting. Issues of data quality at the time of collection can later restrict the range of later uses of the data. Feeding back information to the data custodians from users and aggregators on suspect records is essential, and this is where annotations and reporting back on the results of tests conducted by aggregators is important. The project is also generating standard code and test data for the tests and assertions so that data custodians can readily integrate them into their own procedures. It is far cheaper to correct errors at the source than try and rectify them further down the line.

A lot of progress has been made, but we still have a long way to go – join us in making biodiversity data quality a product of which we can all be proud.

Keywords

TDWG BDQIG "Data Quality" "Fitness for Use" Framework "Tests and Assertions" quality vocabularies

Presenting author

Arthur D. Chapman