BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25608
doi: 10.3897/biss.2.25608

OPEN ACCESS

Conference Abstract

Data Quality Task Group 2: Tests and Assertions

Lee Belbin‡, Arthur Chapman§, John Wieczorek|, Paula F Zermoglio¶, Alex B Thompson#, Paul J Morris¤

‡ Blatant Fabrications Pty Ltd, Carlton, Australia
§ Australian Biodiversity Information Services, Ballan, Australia
| Museum of Vertebrate Zoology, University of California, Berkeley, United States of America
¶ Instituto de Ecología, Genética y Evolución de Buenos Aires (IEGEBA-CONICET), University of Buenos Aires, Buenos Aires, Argentina
# Apple, Tampa, United States of America
¤ Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States of America

Corresponding author: Lee Belbin (leebelbin@gmail.com)

Received: 08 Apr 2018 | Published: 18 May 2018

Citation: Belbin L, Chapman A, Wieczorek J, Zermoglio P, Thompson A, Morris P (2018) Data Quality Task Group 2: Tests and Assertions. Biodiversity Information Science and Standards 2: e25608. https://doi.org/10.3897/biss.2.25608

Abstract

Task Group 2 of the TDWG Data Quality Interest Group aims to provide a standard suite of tests and resulting assertions that can assist with filtering occurrence records for as many applications as possible. Currently 'data aggregators' such as the Global Biodiversity Information Facility (GBIF), the Atlas of Living Australia (ALA) and iDigBio run their own suite of tests over records received and report the results of these tests (the assertions): there is, however, no standard reporting mechanisms. We reasoned that the availability of an internationally agreed set of tests would encourage implementations by the aggregators, and at the data sources (museums, herbaria and others) so that issues could be detected and corrected early in the process.

All the tests are limited to Darwin Core terms. The ~95 tests refined from over 250 in use around the world, were classified into four output types: validations, notifications, amendments and measures. Validations test one of more Darwin Core terms, for example, that dwc:decimalLatitude is in a valid range (i.e. between -90 and +90 inclusive). Notifications report a status that a user of the record should know about, for example, if there is a user-annotation associated with the record. Amendments are made to one or more Darwin Core terms when the information across the record can be improved, for

example, if there is no value for dwc:scientificName, it can be filled in from a valid dwc:taxonID. Measures report values that may be useful for assessing the overall quality of a record, for example, the number of validation tests passed.

Evaluation of the tests was complex and time-consuming, but the important parameters of each test have been consistently documented. Each test has a globally unique identifier, a label, an output type, a resource type, the Darwin Core terms used, a description, a dimension (from the Framework on Data Quality from TG1), an example, references, implementations (if any), test-prerequisites and notes. For each test, generic code is being written that should be easy for institutions to implement – be they aggregators or data custodians.

A valuable product of the work of TG2 has been a set of general principles. One example is “Darwin Core terms are either:

1. literal verbatim (e.g., dwc:verbatimLocality) and cannot be assumed capable of validation,
2. open-ended (e.g., dwc:behavior) and cannot be assumed capable of validation, or
3. bounded by an agreed vocabulary or extents, and therefore capable of validation (e.g., dwc:countryCode)”.

Another is “criteria for including tests is that they are informative, relatively simple to implement, mandatory for amendments and have power in that they will not likely result in 0% or 100% of all record hits.” A third: “Do not ascribe precision where it is unknown.”

GBIF, the ALA and iDigBio have committed to implementing the tests once they have been finalized. We are confident that many museums and herbaria will also implement the tests over time. We anticipate that demonstration code and a test dataset that will validate the code will be available on project completion.

Keywords

Quality, data, fitness, test, assertion,validation, report, ALA, GBIF, iDigBio

Presenting author

Lee Belbin

Acknowledgements

We gratefully acknowledge the support of the Atlas of Living Australia, iDigBio and Kurator.