It Takes Years for a Good Wine to Mature: Task Group 2 - data quality tests and assertions

Lee Belbin[‡], Arthur Chapman[§], Paul J. Morris[|], John Richard Wieczorek[¶]

‡ Blatant Fabrications Pty Ltd, Hobart, Australia
§ Australian Biodiversity Information Service, Ballan, Australia
| Harvard University, Boston, United States of America
¶ University of California, Berkeley, United States of America

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

Abstract

Data Quality Task Group 2 was established to create a suite of core tests and associated assertions about the 'quality' of biodiversity informatics data (Chapman et al. 2020). The group has been active since January 2017, about four years longer than its four main members would have anticipated. We all thought “How hard could it be?” The answer was “Harder than we thought!” We have invested well over two years full time into this project. There were multiple times over the past five years where we thought we were 95% done, but we were wrong. Were we dumb? I doubt it! The authors (other than the lead author) are highly experienced in biodiversity data quality, Darwin Core and data testing. Neither were we lazy.

Why has it gone so slowly? It is mostly due to the complexity of the task and the inability to meet face-to-face. Zoom just doesn’t cut it for this type of work. We achieved the most at our one face-to-face meeting in Gainesville (Florida) in 2018. Our advances over the past year have come from rounds of feedback between the test specifications, test implementation, development of data for validating the tests and comparison between results from implementations and the expectations of the validation data. There are hopefully useful lessons in this for similar projects.

We now have a solid base where future evolution, such as tests for specific environments, will be made relatively easy. The major components of this project are the 99 tests themselves, the parameters for these tests (see https://github.com/tdwg/bdq/issues/122), a vocabulary of the terms used in the framework and test data for validating implementations of the tests.

We remain focused on what we call core tests: those that provide power in evaluating 'fitness for use', are widely applicable and are relatively easy to implement. The test descriptions we have settled on are:

1. A human readable label (split into a test class, a target Darwin Core term and an 'action');
2. A Globally Unique Identifier for the test (a GUID);
3. A simple English description;
4. Test class from the Fitness-For-Use Framework (Data Quality Task Group 1): Validation, Amendment, Measure or Issue;
5. Resource Type (all of the Core tests operate on a single record);
6. Information Elements (specified as the applicable Darwin Core Class and as a list of specific Darwin Core terms required as inputs for the test);
7. Specification (an explanation of how the test works from an implementation perspective);
8. Data quality dimension (from the Fitness-for-Use Framework);
9. Warning type (ambiguous, amended, incomplete, invalid, issue, report, unlikely);
10. Parameters (options that allow implementations to behave differently in clearly defined ways such as the use of a national species list);
11. Source Authority (external references required by the test);
12. An example;
13. Source (the origin of the test);
14. References;
15. Link to reference implementations;
16. Link to source code and
17. Notes (explanations of subtle or not so subtle aspects of the test).

The composition of the core tests has been stable for over a year. We have generated most of the test data using the template: the applicable test, a unique identifier, input data, expected output data, the response status (e.g., "internal prerequisites not met"), the response result (e.g., "not compliant"), and an optional comment.

What remains to be done? We need to complete the test data, produce normative and non-normative documentation, and transform our work into a TDWG Technical Specification. While TG2 is over 95% complete, we would still welcome anyone who is interested to learn about biodiversity data quality to contribute.

Keywords

specifications, vocabulary, biodiversity data, validation, amendment, report

Presenting author

Lee Belbin

Presented at

TDWG 2022

Acknowledgements

We acknowledge the significant contributions of Paula Zermoglio and Alex Thompson as original TG2 team members. We also value the comments of Deborah Paul and Allan Koch Veiga on our GitHub issues.

Conflicts of interest

References

- Chapman A, Belbin L, Zermoglio P, Wieczorek J, Morris P, Nicholls M, Rees ER, Veiga A, Thompson A, Saraiva A, James S, Gendreau C, Benson A, Schigel D (2020) Developing Standards for Improved Data Quality and for Selecting Fit for Use Biodiversity Data. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.50889