BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25657
doi: 10.3897/biss.2.25657

OPEN ACCESS

Conference Abstract

The Online Pollen Catalogs Network (RCPol) data quality assurance system

Allan Koch Veiga‡, Antonio Mauro Saraiva‡, Cláudia Inês da Silva‡

‡ University of São Paulo, São Paulo, Brazil

Corresponding author: Allan Koch Veiga (allan.kv@gmail.com), Antonio Mauro Saraiva (saraiva@usp.br)

Received: 09 Apr 2018 | Published: 18 May 2018

Citation: Veiga A, Saraiva A, da Silva C (2018) The Online Pollen Catalogs Network (RCPol) data quality assurance system. Biodiversity Information Science and Standards 2: e25657. https://doi.org/10.3897/biss.2.25657

Abstract

The Online Pollen Catalogs Network (RCPol) (http://rcpol.org.br) was conceived to promote interaction among researchers and the integration of data from pollen collections, herbaria and bee collections. In order to structure RCPol work, researchers and collaborators have organized information on Palynology in four branches: palynoecology, paleopalynology, palynotaxonomy and spores. This information is collaboratively digitized and managed using standardized Google Spreadsheets. These datasets are assessed by the RCPol palynology experts and when a dataset is compliant with the RCPol data quality policy, it is published to http://chaves.rcpol.org.br.

Data quality assessment used to be performed manually by the experts and was time-consuming and inconsistent in detecting data quality problemas such as incomplete and inconsistent information. In order to support data quality assessment in a more automated and effective way, we are developing a data quality tool which implements a series of mechanisms to measure, validate and improve completeness, consistency, conformity, accessibility and uniqueness of data, prior to a manual expert assessment. The system was designed according to the conceptual framework proposed by Task Group 1 of the Biodiversity Data Quality Interest Group Veiga et al. 2017. For each sheet in the Google Spreadsheet, the system generates a set of assertions of measures, validations and amendments for the records (rows) and datasets (sheets), according to a profile defined for RCPol. The profile follows the policies of data quality measurement, validation and enhancement. The data quality measurement policy encompassess the dimensions of completeness, consistency, conformity, accessibility and uniqueness. RCPol uses a quality assurance approach: only data that are compliant with all the quality requirements are published in the system. Therefore, its data quality validation policy only considers datasets with 100% completeness, consistency, conformity, accessibility and uniqueness. In order to improve the quality in each relevant dimension, a set of enhancements was defined in the data quality enhancement policy. Based on this RCPol profile, the system is able to generate reports that contain measures, validations and amendments assertions with the method and tool used to generate the assertion. This web-based system can be tested at http://chaves.rcpol.org.br/admin/data-quality with the dataset https://docs.google.com/spreadsheets/u/1/d/1gH0aa2qqnAgfAixGom3Gnx6Qp_91ZvWhUHPb_QeoIreQ. This system is able to assure that only data compliant with the data quality profile defined by RCPol are fit for use and can be published.

This system contributes significantly to decreasing the workload of the experts. Some data may still contain values that cannot be easily automatically assessed, e.g. validate if the content of an image matches the respective scientific name, so expert manual assessment remains necessary. After the system reports that data are compliant with the profile, a manual assessment must be performed by the experts, using the data quality report as support, and only after that will the data be published. The next steps include archival of the data quality reports in a database, improving the web interface to enable searching and sorting of assertions, and to provide a machine readable interface for the data quality reports.

Keywords

data quality, quality assurance, conceptual framework, data quality profile, data quality report

Presenting author

Allan Koch Veiga

References

- Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D, Robertson TJ (2017) A conceptual framework for quality assessment and management of biodiversity data. PLOS ONE 12 (6): e0178731. https://doi.org/10.1371/journal.pone.0178731