Taxonomic Data Quality Control for CoL-China

Congtian Lin[‡,§], Liqiang Ji[‡]

‡ Institute of Zoology, Chinese Academy of Sciences, Beijing, China
§ National Basic Science Data Center, Beijing, China

Corresponding author: Congtian Lin (linct@ioz.ac.cn), Liqiang Ji (ji@ioz.ac.cn)

Abstract

High quality checklists of species are important in biodiversity data to help answer what and how many species are present in a country or a region, and they are often used as backbones in biodiversity databases. Catalogue of Life, China (CoL-China), hosted by the Species 2000 China Node in the Chinese Academy of Sciences, has published 16 annual versions since 2008, and these have been used very widely in China for supporting biodiversity research, conservation decisions and citizen science. Taxonomic data quality is one of the reasons for its popularity, due to a systematic workflow that guarantees quality control. Our goals of quality management are to ensure that the contents of the CoL-China comply with the standard of the global Catalogue of Life (CoL), ensure that data items such as the taxonomy system, accepted names, Chinese names, common names, synonyms, distributions, literature, and data sources are complete and accurate, improve the scientific value and reliability of CoL-China, and ensure the smooth release of each annual version. Several measures were implemented in our quality control workflow:

1. **A professional organization ensures the scientific credentials of CoL-China.** An editorial board, including 31 authoritative scientists as a decision-making body, leads the Species 2000 China Node to establish rules and goals for making and publicizing CoL-China. More than 300 taxonomists from institutes of the Chinese Academy of Sciences are involved in working on different taxon groups like animals, plants etc. There is a working group that is composed of taxonomists and information scientists for managing the procedure of annual checklist production and compiling the checklists of various taxon groups into CoL-China.
2. **Authoritative data sources ensure the quality of CoL-China from the outset.** All selected data sources are from peer-reviewed taxonomic papers, dissertations or mature databases. Selection of each data source is controlled by a specialist in the relevant taxon group. Principles for filtering data sources are as follows:
    1. *Completeness*. At a minimum, the data source should hold a checklist of a family and contain most of the fields required, such as accepted name with authorship, classification and distribution. The fields like common name and reference are optional.

2. *Science*. The data source should be maintained by the professional community, and all the data items should be checked by experts for each species.
3. *Timeliness*. Maintenance of the data source should comply with leading-edge practices for taxonomic research.

3. **A taxonomic data management tool was developed to implement the workflow of data quality.** This is a platform for multi-person collaboration, which allows experts who study the same taxonomic group to work on the same datasets together. It can collect and manage multi-dimensional data of species, meeting the requirements of CoL-China. The tool provides several convenient functions e.g., batch import taxon data, a visualization tool for editing taxonomic trees, and extraction of taxonomic data from labeled free text. An auto checking process with 28 steps (Fig. 1) was implemented in this tool to verify each item for all species.
4. **Artificial intelligence helps improve quality control.** Taxonomic data is in free text for many data sources, and it is easy to make mistakes when retrieving data items manually. We developed an artificial intelligence (AI) tool for extracting distribution data from free text, which significantly helps to promote data quality.
5. **Unique identifiers introduced for each taxon imported** into CoL-China so that taxon concepts can be tracked. This helps control data quality from the original source to the CoL-China database.

Major progress has been made in listing the Chinese known species by CoL-China. But large gaps still exists in some taxon groups especially for insects, which affects the whole quality of CoL-China. Another challenge is how to keep CoL-China up-to-date with new discoveries. As a next step, we will focus on these problems and continue to keep consistent data quality assurance and data quality control mechanisms with CoL.

Keywords

workflow, Catalog of Life, Species 2000

Presenting author

Congtian Lin

Presented at

TDWG 2023

Acknowledgements

We thank Yan Han and Jiangning Wang, members of Institute of Zoology, CAS, for their contributions on collecting data. This job is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences(Grant No. XDA19050202).

Conflicts of interest

The authors have declared that no competing interests exist.

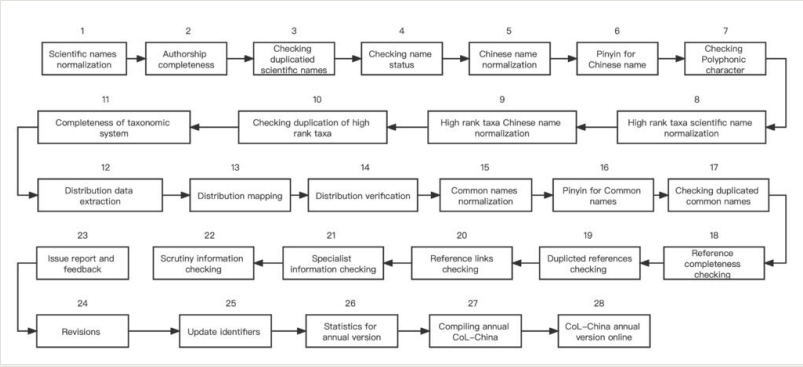

Figure 1.

Taxonomic data quality control workflow.