A rescue strategy for threatened biodiversity data

Anton Güntsch & Walter G. Berendsohn

Freie Universität Berlin, Botanic Garden and Botanical Museum Berlin-Dahlem
Königin-Luise-Str. 6-8, 14195 Berlin, Germany

EMail: BiodiversityInformatics@bgbm.org

ABSTRACT

Large volumes of valuable biodiversity-related data are lost every day, because they are not linked to an institutional data curation strategy ensuring their long-term availability and accessibility. Rather, biodiversity data are often generated in the course of scientific studies and projects and remain with the authors, who usually do not consider data a primary intellectual product which needs a preservation strategy comparable to the treatment of publications. As a consequence, biodiversity data are often insufficiently documented, lack a sound backup and archiving strategy, and are stored on media and software systems which can get quickly outdated and unreadable.

The Biodiversity Informatics group at the Botanic Garden and Botanical Museum Berlin-Dahlem (www.bgbm.org/biodivinf/) carries out a project developing workflows and software supporting the rescue of threatened biodiversity data with a focus on efficiency and cost-effectiveness. The system uses database “wrapper” technology developed by the European BioCASE initiative (www.biocase.org), which is normally used to link up primary data to international biodiversity information networks such as GBIF (www.gbif.org). In the context of data rescue workflows, we use the very same software to create exports of structurally highly heterogeneous databases into standardized XML-based biodiversity data formats, which are widely accepted and well described and understood. The data are then stored in an XML-database and described with further metadata. Using a native XML-store allows us to accommodate data according to different schemes without the need for a full specification of the required data structures.

Additional topics to be addressed by the project include the development of generic services and User-interfaces to the archive as well as data quality software for semi-automatic correction of malformed or erroneous data to be stored in the repository. The project cooperates with the CODATA ‘Data At Risk’ Task Group (DARTG, http://ils.unc.edu/~janeg/dartg/), which is concerned with threatened data across scientific disciplines.

Keywords: primary biodiversity data, data rescue workflow, XML-database