

Towards a Data Archiving Infrastructure for Primary Biodiversity Information

Anton Güntsch, Birgit Gemeinholzer & Walter G. Berendsohn

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

E-Mail: a.guentsch@bgbm.org

ABSTRACT

Free and stable access to primary biodiversity data is essential for biodiversity sciences and the efficiency of the associated scientific workflows. Over the last 15 years, international initiatives such as GBIF (Global Biodiversity Information Facility, www.gbif.org) and BioCASE (Biological Collection Access Service for Europe, www.biocase.org) have designed and implemented a world wide data network, which gives instant access to hundreds of distributed primary biological data providers, initially with a focus on occurrence data documenting the presence of an organism at a given point in time and location. Today, the global network provides about 175 million records in 7500 datasets from 285 data providers (June 2009). With the incorporation of biodiversity monitoring data, these numbers will rapidly grow over the next few years.

The distributed nature has several advantages over a centralized repository. Most importantly, it ensures that information providers have always full control over their data, which contributed significantly to the acceptance of the new infrastructure. However, a downside is the absence of stable mechanisms for long-term preservation of data held by smaller and technically less experienced providers. As a consequence, numerous important data sources are at risk and will disappear if no measures are taken to save them.

We therefore propose to complement the existing biodiversity data infrastructures with a data archiving system providing at least the following services:

- **Transformation** of legacy data into well-described data formats and **upload** into the archive.
- **Metadata** capture for rich descriptions of primary data as well as technical parameters supporting access.
- **Migration** service supporting data transformations into contemporary formats.
- **Software repository** for tools necessary to understand or process particular data sets.
- **Access** service ensuring the integration of archived data in the biodiversity information architecture and beyond.

Several national and international initiatives are now underway to establish the biodiversity data archive. We believe that the new facility will greatly contribute to the stability of the global information network and help to answer the challenging scientific questions associated with our changing biodiversity.