Towards Linked Open Molecular Data: Recommendations for researchers, collections, infrastructures and publishers

Gabriele Droege[‡], Ilene Karsch-Mizrachi[§], Katharine Barker[|], Jonathan Coddington[|], Ole Seberg[¶]

‡ Botanic Garden and Botanical Museum Berlin, Berlin, Germany
§ National Institute of Health, Washington D.C., United States of America
| National Museum of Natural History, Smithsonian Institution, Washington, D.C., United States of America
¶ Natural History Museum Denmark, Copenhagen, Denmark

Corresponding author: Gabriele Droege (g.droege@bgbm.org)

Abstract

The variety of molecular methods used to analyze biosamples is continuously increasing, as is the need for the standardized deposition, documentation and citation of both the samples as well as the methods applied to them. Global initiatives such as the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org), Barcode of Life Data System (BOLD, http://www.boldsystems.org), the Global Biodiversity Information Facility (GBIF, http://www.gbif.org) and the Global Genome Biodiversity Network (GGBN, http://www.ggbn.org), in addition to many others, have been working towards standardized access to biological data for many years. Collectively, these biodiversity data management platforms provide a considerable and indispensable infrastructure to the research community. However, cross-linking the massive amounts of protein and DNA sequence data submitted to these databases every year with standardized records of the underlying biological material remains challenging. Best practices for standardized data submissions and data citations are urgently needed.

In the long run, two goals should be achieved above all else:

1. all sequence data should be linked to natural history collections, and
2. biological material that was used for molecular research, especially DNA sequencing, should be deposited and, thus, made accessible in public, well curated collections.

Here we will provide recommendations both for researchers and collections how to cite underlying biological material at INSDC and in publications in a standardized way towards Linked Open Data. We will also address how the global infrastructures and publishers can improve their interoperability.

Keywords

GGBN; Linked Open Data; INSDC, GBIF

Presenting author

Gabriele Droege

Presented at

Biodiversity_Next 2019

Funding program

This work was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Conflicts of interest