



Shortening the “long tail” of “dark data”: Automated export of small data from Biodiversity Data Journal to GBIF and EOL through Darwin Core Archive

Markus Döring¹, Tim Robertson¹, Teodor Georgiev², Jordan Biserkov², Pavel Stoev²,
Patrick Leary³, Katja Schulz³, Cynthia Parr³, Guido Sautter⁴, Donat Agosti⁴, Lyubomir Penev²

¹ GBIF, Copenhagen, Denmark
² Pensoft Publishers, Sofia, Bulgaria
³ Encyclopedia of Life, Woods Hole, USA
⁴ Plazi, Bern, Switzerland

According to a recent study, 80 % of scientific data are lost within two decades, at an alarming rate (Vines et al. 2014). The majority of biodiversity data is usually not indexed or properly preserved and form the so called “long tail” of “dark data”, which therefore is more likely to remain invisible to scientists and eventually lost (Heidorn 2008). The bulk of “dark” biodiversity data is constituted by small and scattered datasets, especially species occurrences, published in various literature sources or in grey literature.

The Biodiversity Data Journal uses a novel workflow that allows for upfront markup of occurrence data via its own article authoring tool, the Pensoft Writing Tool. Occurrence data, if compliant to Darwin Core, can be imported into the manuscript in a human-readable format and downloaded from the published article as a CSV file per each separate taxon treatment. In addition, all occurrence data published within an article, are automatically exported on the day of publication into a Darwin Core Archive.

The Darwin Core Archive is then registered with GBIF through their RESTful API so it becomes immediately visible and the data subsequently indexed by the GBIF data portal. The whole process happens on the day of publication, saving in this way a great deal of effort to markup occurrences and export these to aggregators. The same DwC archive is used by Encyclopedia of Life (EOL) to harvest richer data, such as descriptions, images, bibliographies, etc.

The workflow was successfully tested with the paper of Miller et al. (2014) which is remarkable also as probably the fastest ever description and peer-reviewed publication of a new species. It took less than 30 days from discovery in the field station of Malaysian Borneo to publication and data sharing in GBIF and EOL.

The elaboration of the workflow from BDJ and Plazi to GBIF and EOL through Darwin Core Archive was supported by the EU-funded project EU BON, grant agreement No 308454.

Literature

Heidorn PB (2008) Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2) Fall 2008. Institutional Repositories: Institutional Repositories: Current State and Future. Edited by Sarah Sheeves and Melissa Cragin. (<http://hdl.handle.net/2142/9127>).

Miller J, Schilthuizen M, Burmester J, van der Graaf L, Merckx V, Jocqué M, Kessler P, Fayle T, Breeschoten T, Broeren R, Bouman R, Chua W, Feijen F, Fermont T, Groen K, Groen M, Kil N, de Laat H, Moerland M, Moncoquet C, Panjang E, Philip A, Roca-Eriksen R, Rooduijn B, van Santen M, Swakman V, Evans M, Evans L, Love K, Joscelyne S, Tober A, Wilson H, Ambu L, Goossens B (2014) Dispatch from the field: ecology of micro web-building spiders with description of a new species. Biodiversity Data Journal 2: e1076. DOI: 10.3897/BDJ.2.e1076

Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore J-S, Renaut S, Rennison DJ (2014) The availability of research data declines rapidly with article age. Current biology 24(1): 94-97. doi: <http://dx.doi.org/10.1016/j.cub.2013.11.014>.

