Corrected data re-harvested: curating literature in the era of networked biodiversity informatics

Jeremy A. Miller[‡,§], Teodor Georgiev[|], Pavel Stoev[¶], Guido Sautter[§], Lyubomir Penev[#]

‡ Naturalis Biodiversity Center, Leiden, Netherlands
§ www.Plazi.org, Bern, Switzerland
| Pensoft Publishers, Sofia, Bulgaria
¶ National Museum of Natural History and Pensoft Publishers, Sofia, Bulgaria
# Institute of Biodiversity & Ecosystem Research, Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria

Corresponding author:

Academic editor: Donat Agosti

Science makes progress through a constant process of re-evaluation. Revision and error correction are inevitable and generally healthy for the advancement of science. In biodiversity literature, re-evaluation of earlier work can lead to new conclusions, such as a revised taxonomic determination. When significant errors are discovered, conscientious authors may correct the record by publishing an erratum or corrigendum.

Aggregated global biodiversity data is an increasingly powerful resource supporting research, conservation, policy, and public bioliteracy (Hardisty et al. 2013, Arzberger et al. 2004). Along with databases devoted to specimen collections and observation records, literature is an integral part of the biodiversity informatics ecosystem (Miller et al. 2012, Penev et al. 2012, Penev et al. 2011a, Penev et al. 2011b). Pensoft journals pioneered the routine distribution of primary specimen data from publications to a collection of online resources, including the Global Biodiversity Information Facility (GBIF) and the Encyclopedia of Life (EOL) (Penev et al. 2009, Penev et al. 2008, Penev et al. 2010, Smith et al. 2013, Chavan and Penev 2011, Penev et al. 2012, Faulwetter et al. 2014). In the era of digital biodiversity informatics, maintaining data quality presents new challenges. In the realm of corrected taxonomic literature, we argue the objective should be to amend the structured digital record so that the correct information appears on resources like GBIF and the disavowed data are expunged. At the same time, good publishing practice requires that the original document and associated data remain part of the permanent scientific record.

A recent paper on central European spiders included a number of taxonomic errors (Čandek et al. 2013). In a corrigendum published in this issue (Čandek et al. 2015), the authors duly correct the record. Data from the original publication have already been harvested by online resources including GBIF. To guarantee that the data is corrected not only in the scientific literature but also in GBIF, the Darwin Core Archive (DwC-A) file (which is the vehicle for distributing content to a collection of online resources; GBIF 2010, Wieczorek et al. 2012) has been updated and submitted to GBIF. The supplier (Pensoft) needs to trigger a re-indexing through the API (Application Programming Interface, a set of protocols that, in this context, is used to share data between software applications) and the content will be added to the indexing queue. Normally it takes few hours to be indexed (Markus Döring, GBIF senior software developer, pers. comm.). However, the original DwC-A file remains available for users to download from the journal web site. The original and corrected data files are clearly labeled as such and visible alongside the original publication. A link landing at the corrigendum will be added to the original publication metadata to facilitate its discoverability. In addition, the XML data file from the original article has been retained on the servers of Plazi, but the XML tags have been amended to render them no longer exposed for harvest. A modified XML document combining the original data with all corrections specified in the corrigendum (i.e., a single corrected document) has been made available as a supplementary document linked to the corrigendum, and will be uploaded to Plazi upon publication of the corrigendum. This will present the corrected data in XML form, permiting the export of treatments and data to various aggregators (Penev et al. 2012).

This demonstrates a small but important step toward insuring high data quality in the era of growing online networks of biodiversity data. The power of structured biodiversity data aggregated from many sources and freely available online is becoming increasingly valuable to a range of traditional and nontraditional data consumers (Moritz et al. 2011, Arzberger et al. 2004). It is in the interest of the general community and publishers in particular to insure that data are of the highest possible standard.

As large aggregations of data become increasingly important in myriad scientific disciplines, warnings are being sounded that the Achilles' heel of these otherwise promising enterprises is data quality. Big data need robust curatorial mechanisms to assure accuracy and reliability so that the promise of these great collaborative efforts is not squandered (Leonelli 2014, Mesibov 2013, Thessen and Patterson 2011, Hjarding et al. 2014, Belbin et al. 2013). An emerging solution is aimed at collections data from natural history research institutions, a major class of data suppliers to GBIF (Berendsohn et al. 2010, Robertson et al. 2014). The idea is to provide a mechanism for users to flag suspicious records and make possible errors known to data providers (who have the power to check and correct errors) and the broader user community (Wang et al. 2009, Tschöpe et al. 2013, Morris et al. 2013). Wide online access to primary biodiversity data through aggregating databases like GBIF facilitate unprecedented power for data comparison and scrutiny, well beyond what is possible with unnetworked collections databases and literature published on paper without structured digital data. Errors are inevitable in any field, but science is a self-correcting process. The path forward toward well-curated, accessible, aggregated biodiversity data can be accomplished with the participation of the whole community, including publishers, authors, institutional collections personnel, and end users.

Acknowledgements

Development of the data publishing toolkit was supported by EU BON (Building the European Biodiversity Observation Network), an FP-7 (European Union Seventh Framework Programme, 2007-2013) grant (No 308454). Thanks to Markus Döring (GBIF senior software developer) for answering our questions about the GBIF workflow.

References

- Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, Moorman D, Uhlir P, Wouters P (2004) Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Data Science Journal 3: 135-152. https://doi.org/10.2481/dsj.3.135
- Belbin L, Daly J, Hirsch T, Hobern D, LaSalle J (2013) A specialist's audit of aggregated occurrence records: An 'aggregator's' perspective. ZooKeys 305: 67-76. https://doi.org/10.3897/zookeys.305.5438
- Berendsohn W, Chavan V, Macklin J (2010) Recommendations of the GBIF task group on the global strategy and action plan for the mobilization of natural history collections data. Biodiversity Informatics 7: 67-71.
- Čandek K, Gregorič M, Kostanjšek R, Frick H, Kropf C, Kuntner M (2013) Targeting a portion of central European spider diversity for permanent preservation. Biodiversity Data Journal 1: e980. https://doi.org/10.3897/bdj.1.e980
- Čandek K, Gregorič M, Kostanjšek R, Frick H, Kropf C, Kuntner M (2015) Corrigendum: Targeting a portion of central European spider diversity for permanent preservation. Biodiversity Data Journal 3: e4301. https://doi.org/10.3897/BDJ.3.e4301
- Chavan V, Penev L (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics 12: S2. https://doi.org/10.1186/1471-2105-12-S15-S2
- Faulwetter S, Markantonatou V, Pavloudi C, Papageorgiou N, Keklikoglou K, Chatzinikolaou E, Pafilis E, Chatzigeorgiou G, Vasileiadou K, Dailianis T, Fanini L, Koulouri P, Arvanitidis C (2014) Polytraits: A database on biological traits of marine polychaetes. Biodiversity Data Journal 2: e1024. https://doi.org/10.3897/bdj.2.e1024
- GBIF (2010) Darwin Core Archives – How-to Guide, version 1, released on 1 March 2011, (contributed by Remsen D, Braak, K, Döring M, Robertson, T). Global Biodiversity Information Facility, Copenhagen, 21 pp. URL: http://links.gbif.org/gbif_dwca_how_to_guide_v1
- Hardisty A, Roberts D, The Biodiversity Informatics Community (2013) A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology 13: 16. https://doi.org/10.1186/1472-6785-13-16
- Hjarding A, Tolley K, Burgess N (2014) Red List assessments of East African chameleons: a case study of why we need experts. Oryx FirstView: 1-7. https://doi.org/10.1017/s0030605313001427

- Leonelli S (2014) What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society 1 (1): 1-11. https://doi.org/10.1177/2053951714534395
- Mesibov R (2013) A specialist's audit of aggregated occurrence records. ZooKeys 293: 1-18. https://doi.org/10.3897/zookeys.293.5111
- Miller J, Dikow T, Agosti D, Sautter G, Catapano T, Penev L, Zhang Z, Pentcheff D, Pyle R, Blum S, Parr C, Freeland C, Garnett T, Ford LS, Muller B, Smith L, Strader G, Georgiev T, Bénichou L (2012) From taxonomic literature to cybertaxonomic content. BMC Biology 10 (1): 87. https://doi.org/10.1186/1741-7007-10-87
- Moritz T, Krishnan S, Roberts D, Ingwersen P, Agosti D, Penev L, Cockerill M, Chavan V (2011) Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. BMC Bioinformatics 12: S1. https://doi.org/10.1186/1471-2105-12-s15-s1
- Morris R, Dou L, Hanken J, Kelly M, Lowery D, LudÃ¤scher B, Macklin J, Morris P (2013) Semantic Annotation of Mutable Data. PLoS ONE 8 (11): e76093. https://doi.org/10.1371/journal.pone.0076093
- Penev L, Catapano T, Agosti D, Georgiev T, Sautter G, Stoev P (2012) Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher. Journal Article Tag Suite Conference (JATS-Con). Proceedings 2012 [Internet], Bethesda (MD). National Center for Biotechnology Information (US) URL: http://www.ncbi.nlm.nih.gov/books/NBK100351/
- Penev L, Erwin T, Miller J, Chavan V, Moritz T, Griswold C (2009) Publication and dissemination of datasets in taxonomy: ZooKeys working example. ZooKeys 11: 1-8. https://doi.org/10.3897/zookeys.11.210
- Penev L, Lyal C, Weitzman A, Morse D, King D, Sautter G, Georgiev T, Morris R, Catapano T, Agosti D (2011a) XML schemas and mark-up practices of taxonomic literature. ZooKeys 150: 89-116. https://doi.org/10.3897/zookeys.150.2213
- Penev L, Hagedorn G, Mietchen D, Georgiev T, Stoev P, Sautter G, Agosti D, Plank A, Balke M, Hendrich L, Erwin T (2011b) Interlinking journal and wiki publications through joint citation: Working examples from ZooKeys and Plazi on Species-ID. ZooKeys 90: 1-12. https://doi.org/10.3897/zookeys.90.1369
- Penev L, Erwin T, Thompson FC, Sues H, Engel M, Agosti D, Pyle R, Ivie M, Assmann T, Henry T, Miller J, Ananjeva N, Casale A, Lourenco W, Golovatch S, Fagerholm H, Taiti S, Alonso-Zarazaga M, Nieukerken Ev (2008) ZooKeys, unlocking Earth's incredible biodiversity and building a sustainable bridge into the public domain: From "print-based" to "web-based" taxonomy, systematics, and natural history. ZooKeys Editorial Opening Paper. ZooKeys 1: 1-7. https://doi.org/10.3897/zookeys.1.11
- Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith V, Brake I, Ryrcroft S, Scott B, Johnson N, Morris R, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson F, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1-16. https://doi.org/10.3897/zookeys.50.538
- Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE 9 (8): e102623. https://doi.org/10.1371/journal.pone.0102623

- Smith V, Georgiev T, Stoev P, Biserkov J, Miller J, Livermore L, Baker E, Mietchen D, Couvreur TP, Mueller G, Dikow T, Helgen K, Frank J, Agosti D, Roberts D, Penev L (2013) Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal. Biodiversity Data Journal 1: e995. https://doi.org/10.3897/BDJ.1.e995
- Thessen A, Patterson D (2011) Data issues in the life sciences. ZooKeys 150: 15-51. https://doi.org/10.3897/zookeys.150.1766
- Tschöpe O, Macklin J, Morris R, Suhrbier L, Berendsohn W (2013) Annotating biodiversity data via the Internet. Taxon 62 (6): 1248-1258. https://doi.org/10.12705/626.4
- Wang Z, Dong H, Kelly M, Macklin J, Morris P, Morris R (2009) 2009 WRI World Congress Computer Science and Information Engineering. 3. 2009 WRI World Congress on Computer Science and Information Engineering, Los Alamitos, California, 731-735 pp. https://doi.org/10.1109/csie.2009.948
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1): e29715. https://doi.org/10.1371/journal.pone.0029715