BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e26104
doi: 10.3897/biss.2.26104

OPEN ACCESS

Conference Abstract

Darwin Core Spatial Processor (DwCSP): a Fast Biodiversity Occurrences Curator

Julien Troudet‡, Fred Legendre‡, Régine Vignes-Lebbe‡

‡ Institut Systématique Evolution Biodiversité (ISYEB), Sorbonne Université, MNHN, CNRS, EPHE, Paris, France

Corresponding author: Julien Troudet (jtroudet@mnhn.fr)

Received: 23 Apr 2018 | Published: 22 May 2018

Citation: Troudet J, Legendre F, Vignes-Lebbe R (2018) Darwin Core Spatial Processor (DwCSP): a Fast Biodiversity Occurrences Curator. Biodiversity Information Science and Standards 2: e26104. https://doi.org/10.3897/biss.2.26104

Abstract

Primary biodiversity data, or occurrence data, are being produced at an increasing rate and are used in numerous studies (Hampton et al. 2013, La Salle et al. 2016). This data avalanche is a remarkable opportunity but it comes with hurdles. First, available software solutions are rare for very large datasets and those solutions often require significant computer skills (Gaiji et al. 2013), while most biologists are not formally trained in bioinformatics (List et al. 2017). Second, large datasets are heterogeneous because they come from different producers and they can contain erroneous data (Gaiji et al. 2013). Hence, they need to be curated. In this context, we developed a biodiversity occurrence curator designed to quickly handle large amounts of data through a simple interface: the Darwin Core Spatial Processor (DwCSP). DwCSP does not require the installation or use of third-party software and has a simple graphical user interface that requires no computer knowledge. DwCSP allows for the data enrichment of biodiversity occurrences and also ensures data quality through outlier detection. For example, the software can enrich a tabulated occurrence file (Darwin Core for instance) with spatial data from polygon files (e.g., Esri shapefile) or a Rasters file (geotiff). The speed of the enriching procedures is ensured through multithreading and optimized spatial access methods (R-Tree indexes). DwCSP can also detect and tag outliers based on their geographic coordinates or environmental variables. The first type of outlier detection uses a computed distance between the occurrence and its nearest neighbors, whereas the second type uses a

Mahalanobis distance (Mahalanobis 1936). One hundred thousand occurrences can be processed by DwCSP in less than 20 minutes and another test on forty million occurrences was completed in a few days on a recent personal computer. DwCSP has an English interface including documentation and will be available as a stand-alone Java Archive (JAR) executable that works on all computers having a Java environment (version 1.8 and onward).

Keywords

biodiversity occurrences, software, curration, spatial data, outliers

Presenting author

Régine Vignes-Lebbe, http://www.infosyslab.fr/?q=fr/equipe/rvl

References

- Gaiji S, Chavan V, Ariño A, Otegui J, Hobern D, Sood R, Robles E (2013) Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. Biodiversity Informatics 8 (2): 94-172. https://doi.org/10.17161/bi.v8i2.4124
- Hampton S, Strasser C, Tewksbury J, Gram W, Budden A, Batcheller A, Duke C, Porter J (2013) Big data and the future of ecology. Frontiers in Ecology & Environment 11: 156-162. https://doi.org/10.1890/120103
- La Salle J, Williams K, Moritz C (2016) Biodiversity analysis in the digital era. Philosophical Transactions of the Royal Society B: Biological Sciences 371 (1702): 20150337. https://doi.org/10.1098/rstb.2015.0337
- List M, Ebert P, Albrecht F (2017) Ten Simple Rules for Developing Usable Software in Computational Biology. PLOS Computational Biology 13 (1): e1005265. https://doi.org/10.1371/journal.pcbi.1005265
- Mahalanobis P (1936) On the generalized distance in statistics. Proceedings of the National Institute of Sciiences of India (Calcutta) 49-55.