A Pre-ingestion Framework for Darwin Core Archives

Mahmoud Sadeghi[‡], Patricia Koh[‡], Peggy Newman[§]

‡ Atlas of Living Australia, Canberra, Australia
§ Atlas of Living Australia, Melbourne, Australia

Corresponding author: Peggy Newman (Peggy.Newman@csiro.au)

Abstract

The Atlas of Living Australia's (ALA) Pre-ingestion Framework is our alternative to managing datasets via the Global Biodiversity Information Facility's (GBIF) Integrated Publishing Toolkit (IPT). The framework uses a system-agnostic Python codebase to create and update Darwin Core archives: building an archive from a core and extension csv files, merging two archives together, deleting records and identifying duplicates based on the identifiers. The framework dynamicly supports current Darwin Core and GBIF namespace terms.

Previously, this functionality was handled internally by a Java-based biocache-store ingestion application. While flexible and easy to call, this black box approach to data management created challenges like removing problem records and tracking and verifying data sources. Last year, as the ALA merged our ingestion codebase with GBIF's pipelines and upgraded our data store infrastructure, we took the opportunity to manage our source data exclusively as full Darwin Core archives, rather than partial text files or spreadsheets. Consequently, the Python-based framework consolidates a lot of work previously managed using a range of methodologies and technologies including Talend, Java and unix based scripting. Alongside the Darwin Core archive manipulation tools, it has handlers for harvesting data from secure external web services, web hosts or file servers.

The standardised approach to data loading paves the way for improved automation and workflow. The work has the potential to become an open source project to share with the Living Atlas and biodiversity informatics communities.

Keywords

Python, data management

©
.

Presenting author

Mahmoud Sadeghi

Presented at

TDWG 2022

Conflicts of interest