

Data Analyst - EcoCommons

Are you an experienced data analyst seeking a new and challenging opportunity?
Contact the CSIRO's Atlas of Living Australia today!

Role Overview

- Do you want to apply your data skills to be part of science and research at CSIRO?
- Are you passionate about open-source software and open data?
- Would you like to work on international collaborations?

The Atlas of Living Australia (ALA) is Australia's national biodiversity data aggregator funded under the National Collaborative Research Infrastructure Strategy (NCRIS) and hosted by CSIRO. The ALA is the Australian node of the Global Biodiversity Information Facility (GBIF). Our digital infrastructure is developed in-house to support research activities, government decision-making and community events.

The ALA is a key partner in the 3-year digital innovation program, EcoCommons Australia, designed to tackle complex technical challenges encountered by researchers and decision-makers concerned with biodiversity including ecosystem services, biosecurity, natural resource management and climate-related impacts and responses. This is a multi-institutional program with partners across technical and research fields.

As part of this program the ALA has received Australian Research Data Commons funding (through the Queensland Cyber Infrastructure Foundation) to appoint a Data Analyst to work on data acquisition, transformation, loading, integration and quality assurance for a 1.5 year opportunity based at CSIRO in Canberra. Our team is technically oriented and uses multiple technologies and platforms to explore and manipulate large datasets into a standardised format, which we then ingest through our processing pipeline.

What will you be doing?

As a successful candidate you will develop new and support existing automated jobs to harvest data from a series of data providers including national and international data repositories, ensuring data currency and quality is consistent with expectations. The successful candidate will have knowledge in processing species data (occupancy and/or abundance) as well as data on environmental variables (e.g. rainfall, temperature, soil characteristics etc.). You will need to be effective both as a team member and as a reliable point of contact for data providers. We're looking for strong collaboration and communication skills, and the ability to develop great rapport with stakeholders. Suitable candidates located in Canberra are encouraged to apply.

Duties and Key Result Areas:

- Report to the EcoCommons Program Manager and Technical Lead to build and manage both automated and manual data loading processes and specifically focus on a better integration of ALA-provided data into EcoCommons
- Architect a framework for the data lifecycle: from ingestion to processing to search to outputs in scientific workflows and analysis pipeline
- Assist in providing advice on engineering a pipeline for data ingestion and processing (automate as much as possible) to ensure dataset updates + additions are sustainable by the dev team
- Create guidelines for data management (incl. metadata, updates, criteria for inclusion.)
- Map datasets to required data standards (e.g. Darwin Core, Darwin Event Core, Humboldt Core).
- Implement, deploy, schedule and maintain data load processes.
- Implement quality assurance and verification on datasets to ensure loaded records meet expectation
- Engage professionally with external stakeholders offering technical guidance on data management issues such as data mapping, automation, and loading and ensuring data is useful in models and meets the expectations of providers.
- Contribute to team meetings and planning and review activities
- Contribute work to ALA Data Management Team on spatial layer management (adding, updating, deleting layers according to a prioritised worklist).
- Advocate for open science principles wherever feasible and help align projects and development efforts for the benefit of ALA and EcoCommons.

Example of responsibilities of role:

- Work closely with a Business/Scientific Analyst and a software development team to make sure that seamless access to all relevant biodiversity data into the EcoCommons platform is provided
- Work with R programmer/modeller to ensure data for scientific workflows is available
- Get familiar with the biodiversity landscape and compile a summary of all biodiversity datasets including datasets on environmental variables such as land cover but also climate change projection datasets and global oscillation model data
- Understand the data requirements on the currently integrated datasets and expand these
- Work with partners from the CSIRO Knowledge Network to integrate new data repositories into this catalogue of datasets that EcoCommons collaborate with
- Liaise with data providers for integration of datasets into EcoCommons - especially for datasets that require a licence agreement
- Transform/pre-process accessed datasets from repositories into the format required for visualisation within EcoCommons
- Ensure datasets from CSIRO's Knowledge Network are appropriately linked, and datasets to Knowledge Network have all the relevant metadata
- Help develop metadata and processes to ensure that all platform dataset metadata is accurate and incorporated into outputs from the workflow to facilitate reproducibility

Who are we looking for?

Essential Criteria:

We understand that women and other marginalised groups don't tend to apply for these roles unless they meet all of the criteria, and we recognise that there can be other things that make a candidate a great fit. If you're enthusiastic about working in biodiversity science, have strengths in just some of these areas and a willingness to learn fast, please get in touch.

- Strong knowledge of scripting languages in a command line environment – Python or R
- Experience in both delivering and consuming REST services
- Strong (extract, transform, load) ETL skills with large datasets with a focus on efficiency and scale
- Experience with a variety of open source relational and non-relational databases
- Source code management using git, svn or Bitbucket
- Effective stakeholder engagement and technical liaison skills

Desirable Criteria:

- Experience with geospatial data systems and development
- Experience in processing species data (occupancy and/or abundance) as well as data on environmental variables (e.g. rainfall, temperature, soil characteristics etc.)
- Background or strong interest in biodiversity/ecology/taxonomy
- Enthusiasm and knowledge of open data standards, procedures and policy
- Experience with Darwin Core standard
- Experience with Apache Airflow

Eligibility:

The successful applicant will be required to obtain and provide a National Police Check or equivalent.

To be eligible for this position you must be willing and able to travel interstate occasionally.

How to apply

If you would like to be considered for this role, please provide a cover letter of no more than 2 pages with your attached CV that best describes your interest in this role and addresses the above selection criteria. This is a great opportunity to tell the Australian Atlas of Living about you. Please send your resume and covering letter directly to Peggy Newman via email at Peggy.Newman@csiro.au.