Open source tools and workflows for repository management

Irene Steves, Jesse Goldstein, Matt Jones & the NSF Arctic Data Center Team
NSF Arctic Data Center, National Center for Ecological Analysis & Synthesis, University of California, Santa Barbara

NSF Award Number 1546024
https://arcticdata.io
https://github.com/nceas
@arcticdatactr
@nceas

The Arctic Data Center is geographically focused rather than domain specific, which fosters cross-disciplinary data discovery and synthesis. However, developing guidelines and support for heterogeneous data and metadata across social, physical, and biological sciences can be challenging.

Open source tools and workflows help us collaborate as a team to address some of these issues.

Data processing

Edit metadata
The Arctic Data Center stores scientific metadata using Ecological Metadata Language (EML). We make scripted edits to metadata to provide a reproducible record of the changes, as well as to add sections not yet supported by our web-based editor.

Interact with repositories
We upload and download metadata, data objects, and data packages (groups of data objects and their associated metadata) through R. We also use R to copy data packages between repositories in the DataONE network.

Batch-process data packages
Scripted changes to EML allow us to programmatically process groups of related data packages. For example, measurements taken at different sites or across several years may have metadata records that vary only in geographic or temporal coverage.

Resource development

Challenges
- Heterogeneous data
- Continually increasing data quality standards
- Concurrent development of software and R tools

Build tools
We have written R functions to (a) process data packages more efficiently, including a Shiny app for attribute metadata; (b) perform quality assurance on data packages; and (c) improve data use and synthesis.

Share code
GitHub streamlines the process of sharing scripts and R functions within the Support Team, as well as with synthesis scientists using the repository. Versioned and transparent code enhances remote collaboration.

Create training materials
For our training materials, we compile RMarkdown files into an online book format (bookdown), integrating code chunks with written explanations to develop compelling and reproducible examples.

Scientists submit data and metadata to the Arctic Data Center, which the Support Team reviews for quality and completeness.