The Frictionless Data Package: Data Containerization for Automated Scientific Workflows

Abstract
As cross-disciplinary geoscience research increasingly relies on machines to discover and access data, one of the critical questions facing data repositories is how data and supporting materials should be packaged for consumption. Traditionally, data repositories have relied on a human’s involvement throughout discovery and access workflows. This challenge is to find ways to deliver data and information in ways that enable machines to make better decisions by enabling them to understand the data and metadata of many data types. Additionally, once machines have recommended a data resource as relevant to an investigator’s needs, the data resource should be easy to integrate into that investigator’s toolkits for analysis and visualization.

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) supports NSF-funded OCE and PLR investigators with their project’s data management needs. These needs involve a number of varying data types some of which require multiple files with differing formats. BCO-DMO has described these data types and the important relationships between the type’s data files through human-readable documentation on web pages. For machines directly accessing data files from BCO-DMO, this documentation could be overlooked and lead to misinterpreting the data. Instead, BCO-DMO is exploring the idea of data containerization, or packaging data and related information for easier transport, interpretation, and use. In researching the landscape of data containerization, the FrictionlessData Data Package (http://frictionlessdata.io/) provides a number of valuable advantages over similar solutions. This presentation will focus on these advantages and how the FrictionlessData Data Package addresses a number of real-world use cases faced for data discovery, access, and analysis and visualization.

Simpler, expedited data ingest for submitters
DataPackagist - A web service for creating Data Packages.
https://github.com/frictionlessdata/datapackage-pipelines

Data Submitters: login, describe submission, get immediate feedback

Problem: Data can be distributed across multiple locations - databases, files on a server, files on the web, or available from web APIs, etc.

The Frictionless Data Package is a set of extendible, lightweight formats for packaging data and metadata.

Problem: Data can be large so that traditional packages (TAR, ZIP, BagIt) are inefficient transports.

Q: How do we package data to handle these various locations and sizes?

Data Package Pipelines - Framework for processing data packages in pipelines of modular components.
https://github.com/frictionlessdata/datapackage-pipelines

- A pipeline has a list of processing steps, and it generates a single data package as its output.
- A pipeline is defined in a declarative way, not in code, stored in a file named pipeline-spec.yaml.
- Data Package Pipelines define some common processors, custom processors can be created.

Data Submissions may need to be converted to open formats

Data Managers can extend the datapackage.json to add semantic markup.

The resulting datapackage.json can then be used to populate repository metadata catalog.

Continuous Integration Testing for Data
Goodtables.io - Continuous data validation as a service.
https://github.com/frictionlessdata/goodtables.io

Because Data Package Pipelines are declarative, pipeline-spec.yaml files are provenance records.

- Pipelines can be re-run to verify that the workflow is reproducible
- Data Packages can be validated
  - Is the datapackage.json well-formed?
  - Does it meet the JSON schema specification?
- Tabular Data Packages have deeper validation
  - Column header checking vs. datapackage.json defined fields
  - Methods for handling missing data values
  - Regex processing of a data cell for conforming to a pattern

for more information on data in context through self-publishing metadata:
IN33B-0116: Open Core Data approaches to exposing facility data to support FAIR principles
Wed Dec 13, 1:40PM - 6:00PM
Poster Hall D-F