BCO-DMO
Biological & Chemical Oceanography Data Management Office

IN33C-0139
13 Dec 2017
Fall AGU

The Frictionless Data Package: Data Containerization for Automated Scientific Workflows

Adam Shepherd, Woods Hole Oceanographic Institution | **Douglas Fils**, Consortium for Ocean Leadership | **Danie Kinkade**, Woods Hole Oceanographic Institution | **Mak Saito**, Woods Hole Oceanographic Institution

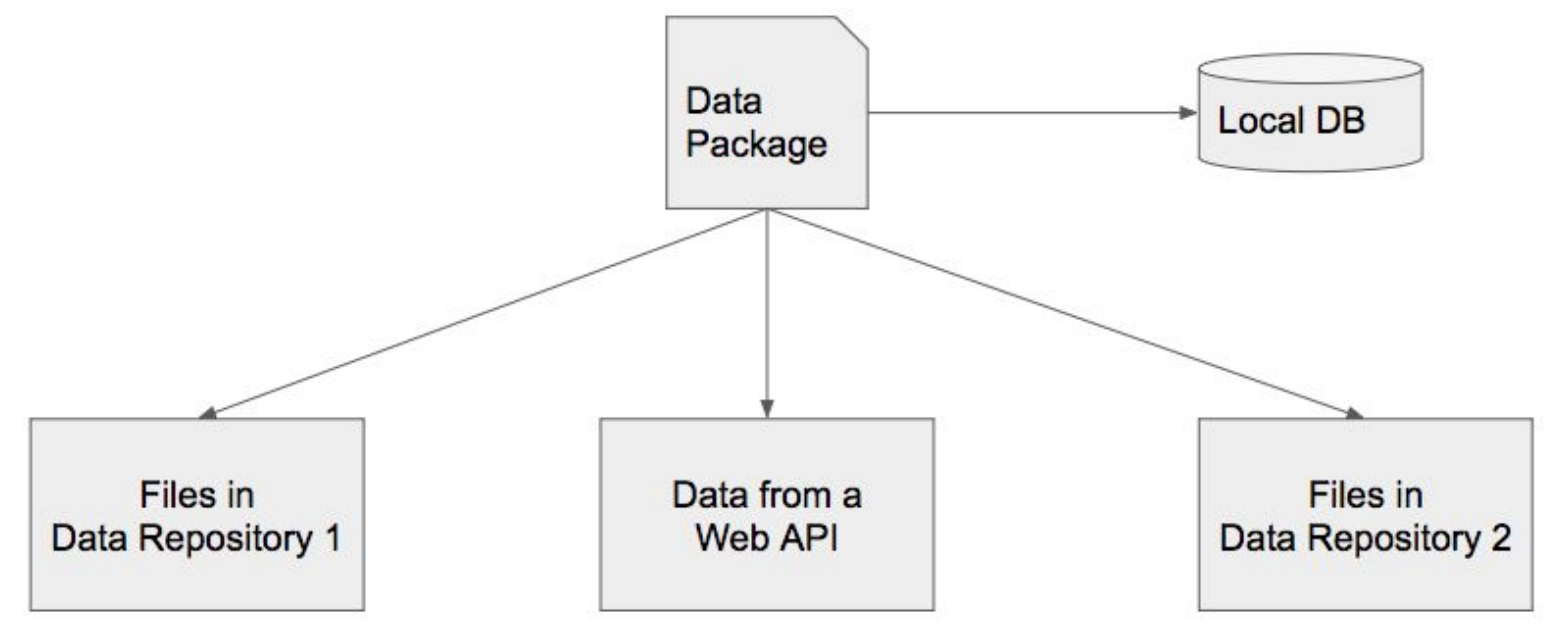
Abstract

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 human could assess fitness for purpose by reading loosely coupled, unstructured information from web pages and documentation. In attempts to shorten the time to science and access data resources across may disciplines, expectations for machines to mediate the process of discovery and access is challenging data repository infrastructure. 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. Presently, 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, analysis and visualization.

What is Data Containerization?

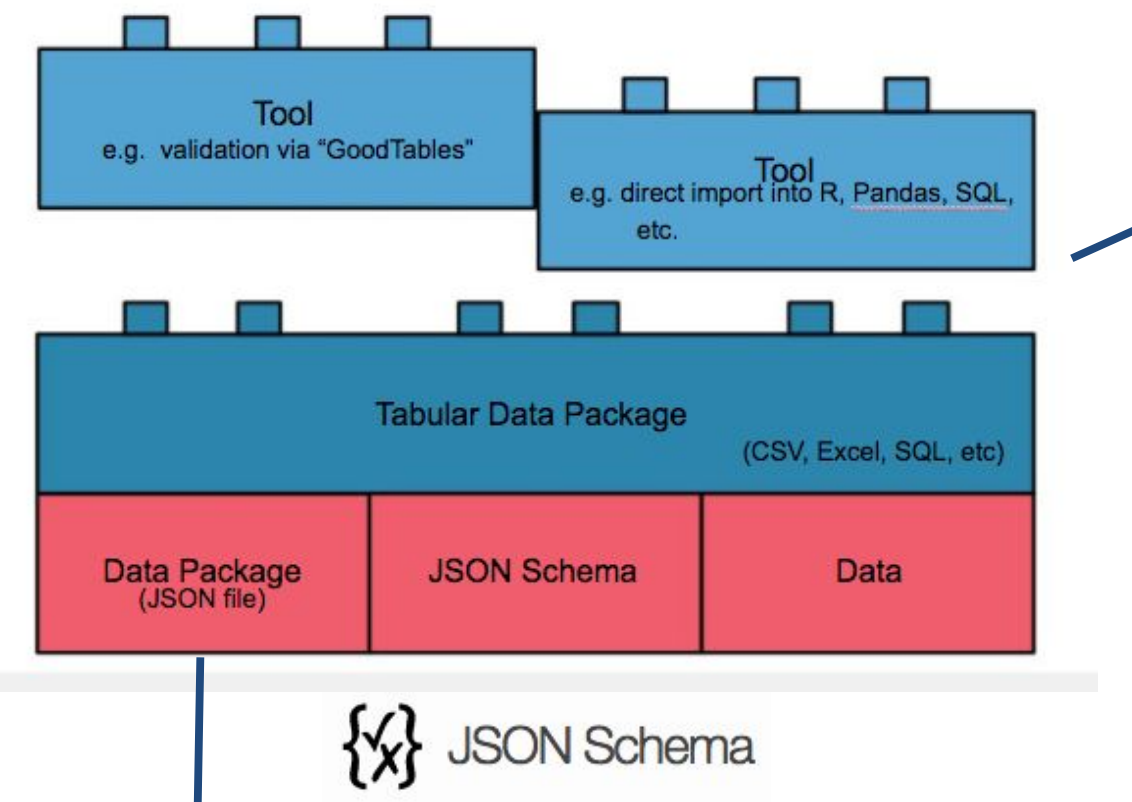
Problem: Data can be distributed across multiple locations - databases, files on a server, files on the web, or available from web APIs, etc.

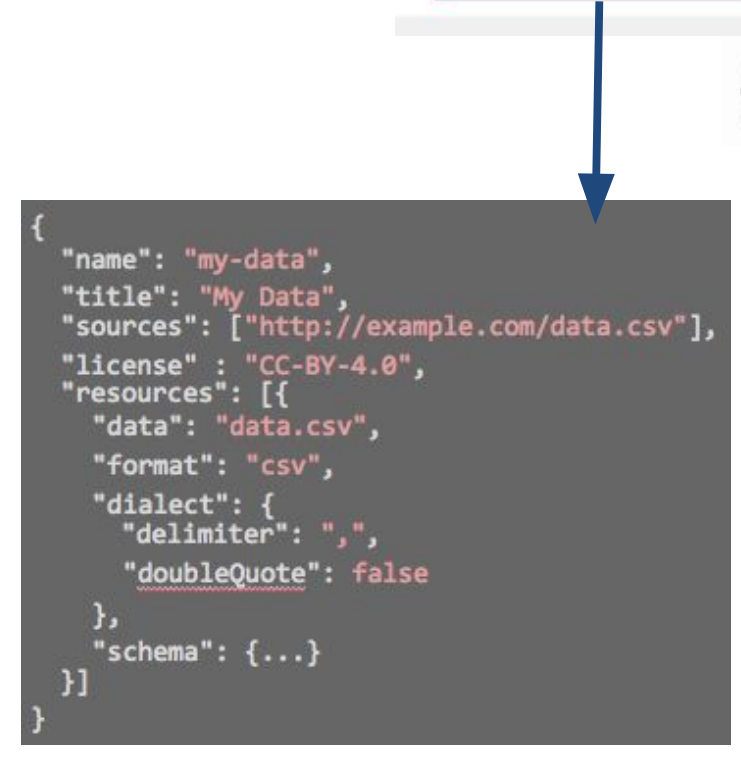


Problem: Data can be large so that traditional packages (TAR, ZIP, BagIt) are inefficient transports.

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

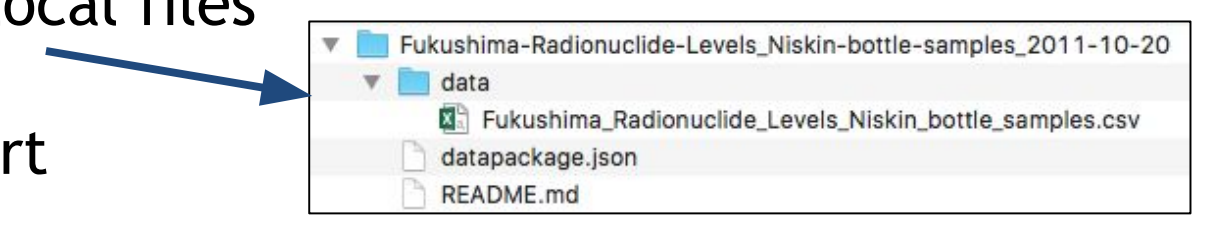
The Frictionless Data Package is a set of extendible, lightweight formats for packaging data and metadata.





```
{
  "name": "my-data",
  "title": "My Data",
  "sources": ["http://example.com/data.csv"],
  "license" : "CC-BY-4.0",
  "resources": [{
    "data": "data.csv",
    "format": "csv",
    "dialect": {
      "delimiter": ",",
      "doubleQuote": false
    },
    "schema": {...}
  }]
}
```

datapackage.json
- can be added to traditional packaging formats (BagIt, TAR, ZIP) for describing local files
- can point to data resources that aren't local to the package or file system
- for data accessible by URL, can be the only file needed to be passed in transport



```
▼ Fukushima-Radionuclide-Levels_Niskin-bottle-samples_2011-10-20
  ▼ data
      Fukushima_Radionuclide_Levels_Niskin_bottle_samples.csv
    datapackage.json
    README.md
```

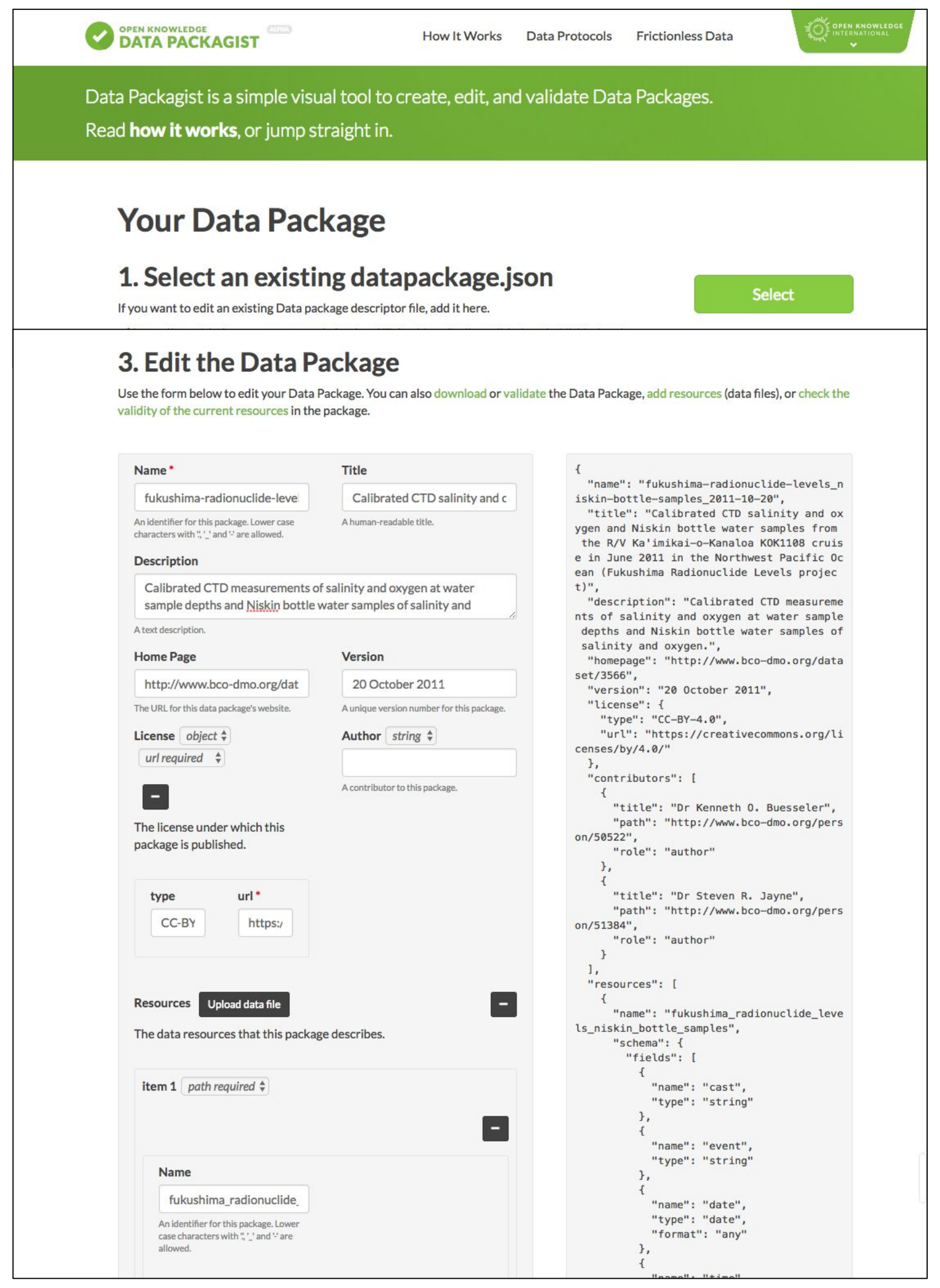
BCO-DMO wants:

- Simpler, expedited data ingest for submitters
- Data transformation that captures provenance
- Continuous integration testing of data holdings

Simpler, expedited data ingest for submitters

DataPackagist - A web service for creating Data Packages.

https://github.com/frictionlessdata/datapackagist



OPEN KNOWLEDGE DATA PACKAGIST | How It Works | Data Protocols | Frictionless Data | OPEN KNOWLEDGE INTERNATIONAL

Data Packagist is a simple visual tool to create, edit, and validate Data Packages.
Read **how it works**, or jump straight in.

Your Data Package

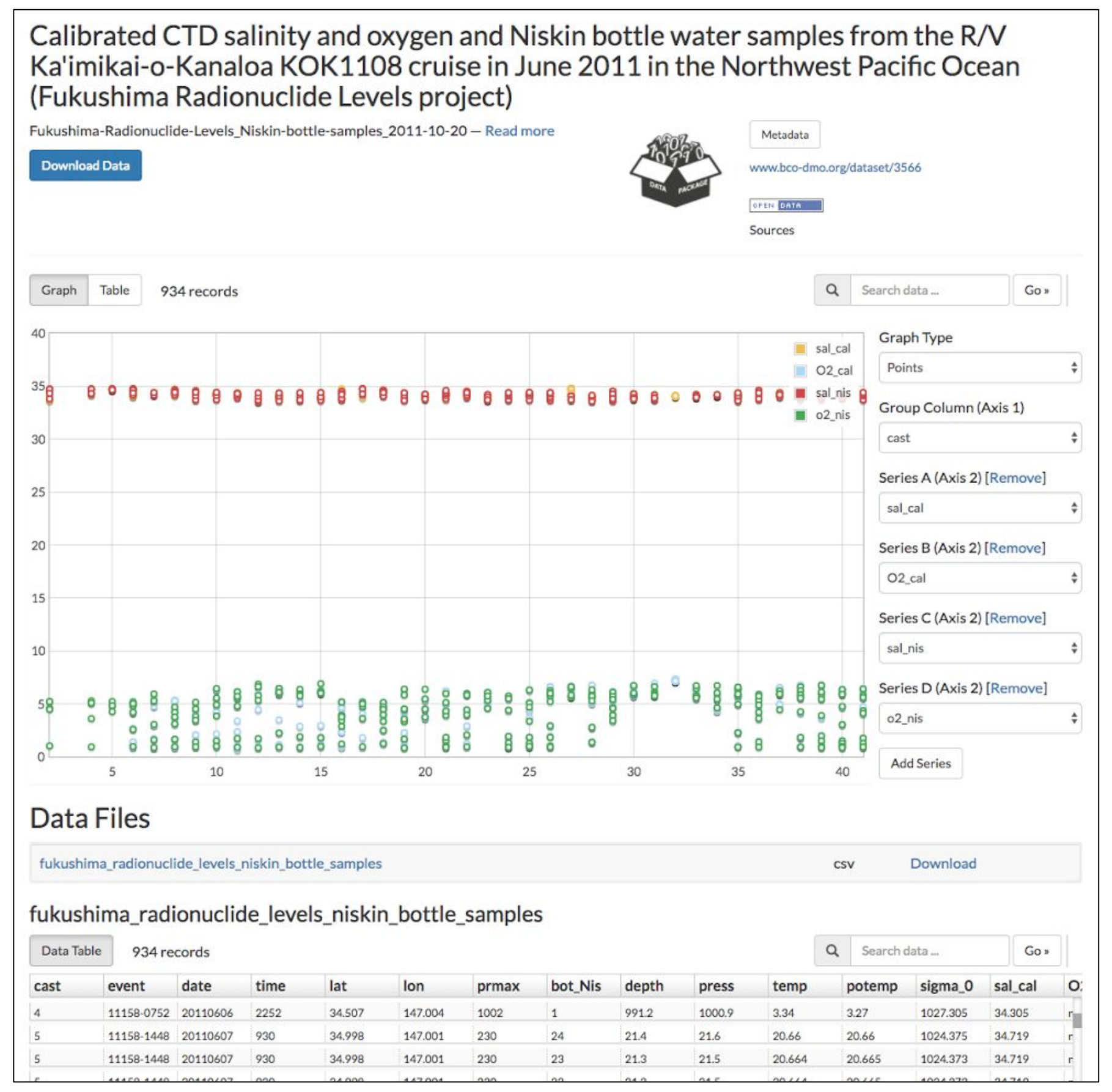
1. Select an existing datapackage.json — Select
If you want to edit an existing Data package descriptor file, add it here.

3. Edit the Data Package
Use the form below to edit your Data Package. You can also download or validate the Data Package, add resources (data files), or check the validity of the current resources in the package.

Name*: fukushima-radionuclide-leve
An identifier for this package. Lower case characters with '.', '_' and '-' are allowed.
Title: Calibrated CTD salinity and c
A human-readable title.
Description: Calibrated CTD measurements of salinity and oxygen at water sample depths and Niskin bottle water samples of salinity and
A text description.
Home Page: http://www.bco-dmo.org/dat
The URL for this data package's website.
Version: 20 October 2011
A unique version number for this package.
License: object — url required
The license under which this package is published.
type: CC-BY — url*: https:/
Author: string
A contributor to this package.
Resources — Upload data file
The data resources that this package describes.
item 1 — path required
Name: fukushima_radionuclide_
An identifier for this package. Lower case characters with '.', '_' and '-' are allowed.

```
{
  "name": "fukushima-radionuclide-levels_n
iskin-bottle-samples_2011-10-20",
  "title": "Calibrated CTD salinity and ox
ygen and Niskin bottle water samples from
 the R/V Ka'imikai-o-Kanaloa KOK1108 cruis
e in June 2011 in the Northwest Pacific Oc
ean (Fukushima Radionuclide Levels projec
t)",
  "description": "Calibrated CTD measureme
nts of salinity and oxygen at water sample
 depths and Niskin bottle water samples of
 salinity and oxygen.",
  "homepage": "http://www.bco-dmo.org/data
set/3566",
  "version": "20 October 2011",
  "license": {
    "type": "CC-BY-4.0",
    "url": "https://creativecommons.org/li
censes/by/4.0/"
  },
  "contributors": [
    {
      "title": "Dr Kenneth O. Buesseler",
      "path": "http://www.bco-dmo.org/pers
on/50522",
      "role": "author"
    },
    {
      "title": "Dr Steven R. Jayne",
      "path": "http://www.bco-dmo.org/pers
on/51384",
      "role": "author"
    }
  ],
  "resources": [
    {
      "name": "fukushima_radionuclide_leve
ls_niskin_bottle_samples",
      "schema": {
        "fields": [
          {
            "name": "cast",
            "type": "string"
          },
          {
            "name": "event",
            "type": "string"
          },
          {
            "name": "date",
            "type": "date",
            "format": "any"
          },
          {
```

Data Submitters: login, describe submission, get immediate feedback



Calibrated CTD salinity and oxygen and Niskin bottle water samples from the R/V Ka'imikai-o-Kanaloa KOK1108 cruise in June 2011 in the Northwest Pacific Ocean (Fukushima Radionuclide Levels project)

Fukushima-Radionuclide-Levels_Niskin-bottle-samples_2011-10-20 — Read more

Download Data | Metadata: www.bco-dmo.org/dataset/3566 | Sources

Graph | Table | 934 records | Search data ... | Go »

Graph Type: Points
Group Column (Axis 1): cast
Series A (Axis 2) [Remove]: sal_cal
Series B (Axis 2) [Remove]: O2_cal
Series C (Axis 2) [Remove]: sal_nis
Series D (Axis 2) [Remove]: o2_nis
Add Series

Data Files

fukushima_radionuclide_levels_niskin_bottle_samples	csv	Download

fukushima_radionuclide_levels_niskin_bottle_samples

Data Table | 934 records | Search data ... | Go »

cast	event	date	time	lat	lon	prmax	bot_Nis	depth	press	temp	potemp	sigma_0	sal_cal
4	11158-0752	20110606	2252	34.507	147.004	1002	1	991.2	1000.9	3.34	3.27	1027.305	34.305
5	11158-1448	20110607	930	34.998	147.001	230	24	21.4	21.6	20.66	20.66	1024.375	34.719
5	11158-1448	20110607	930	34.998	147.001	230	23	21.3	21.5	20.664	20.665	1024.373	34.719

Data Package View Tool by Open Knowledge Foundation
http://data.okfn.org/tools/view?url=https%3A%2F%2Fraw.githubusercontent.com%2Fashepherd%2FFukushima-Radionuclide-Levels_Niskin-bottle-samples_2011-10-20%2Fmaster%2Fdatapackage.json

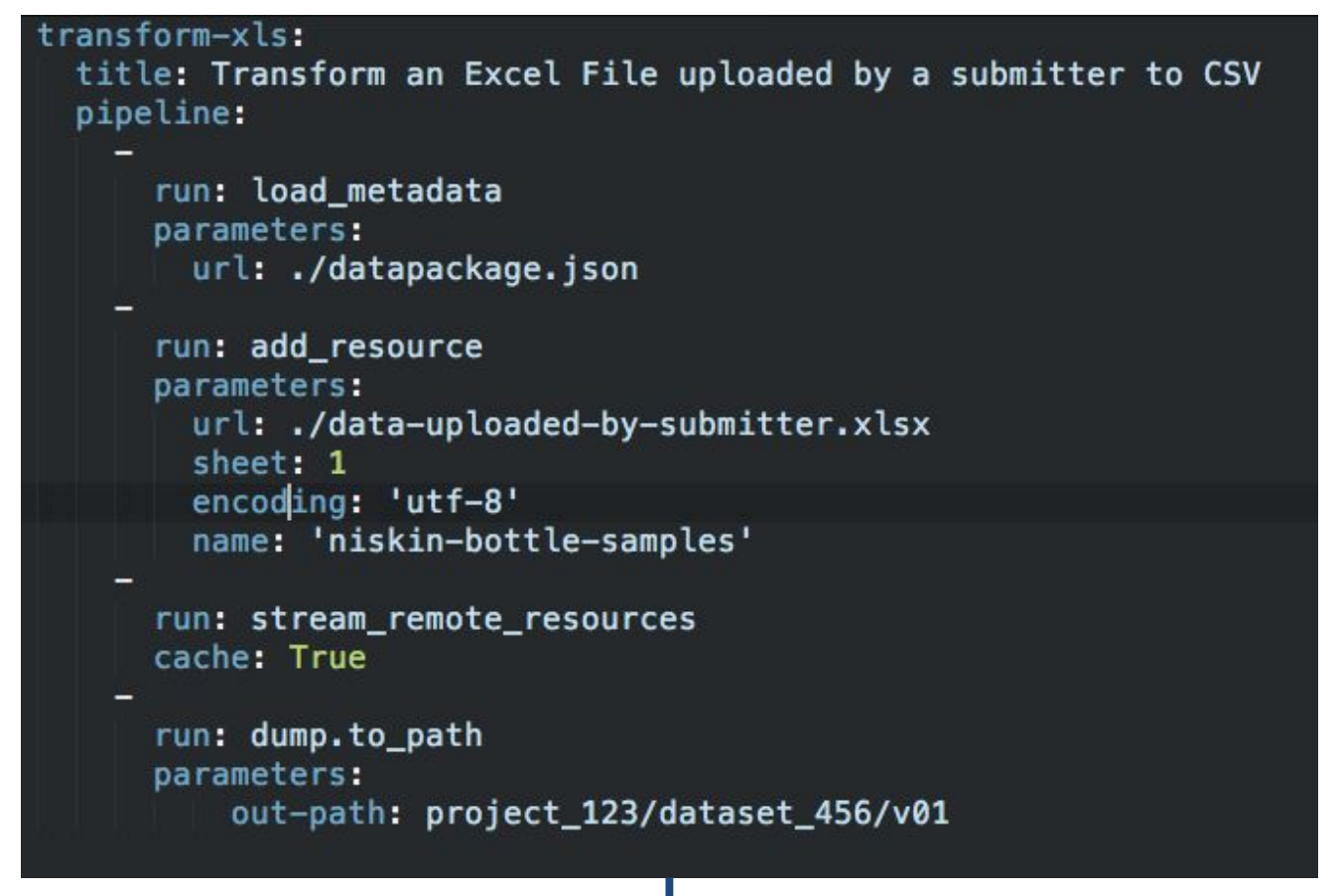
What about the work of a data manager?

Data Package Pipelines - Framework for processing data packages in pipelines of modular components.

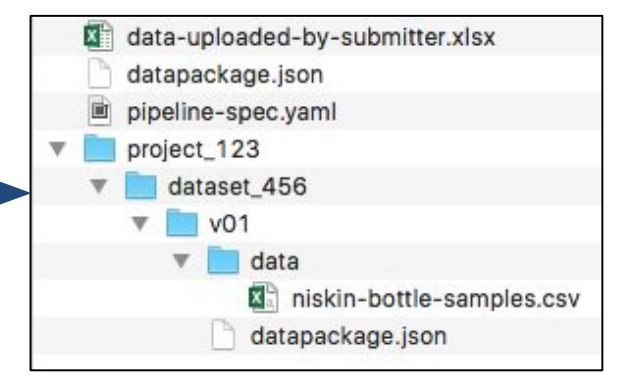
https://github.com/frictionlessdata/datapackage-pipelines

- 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



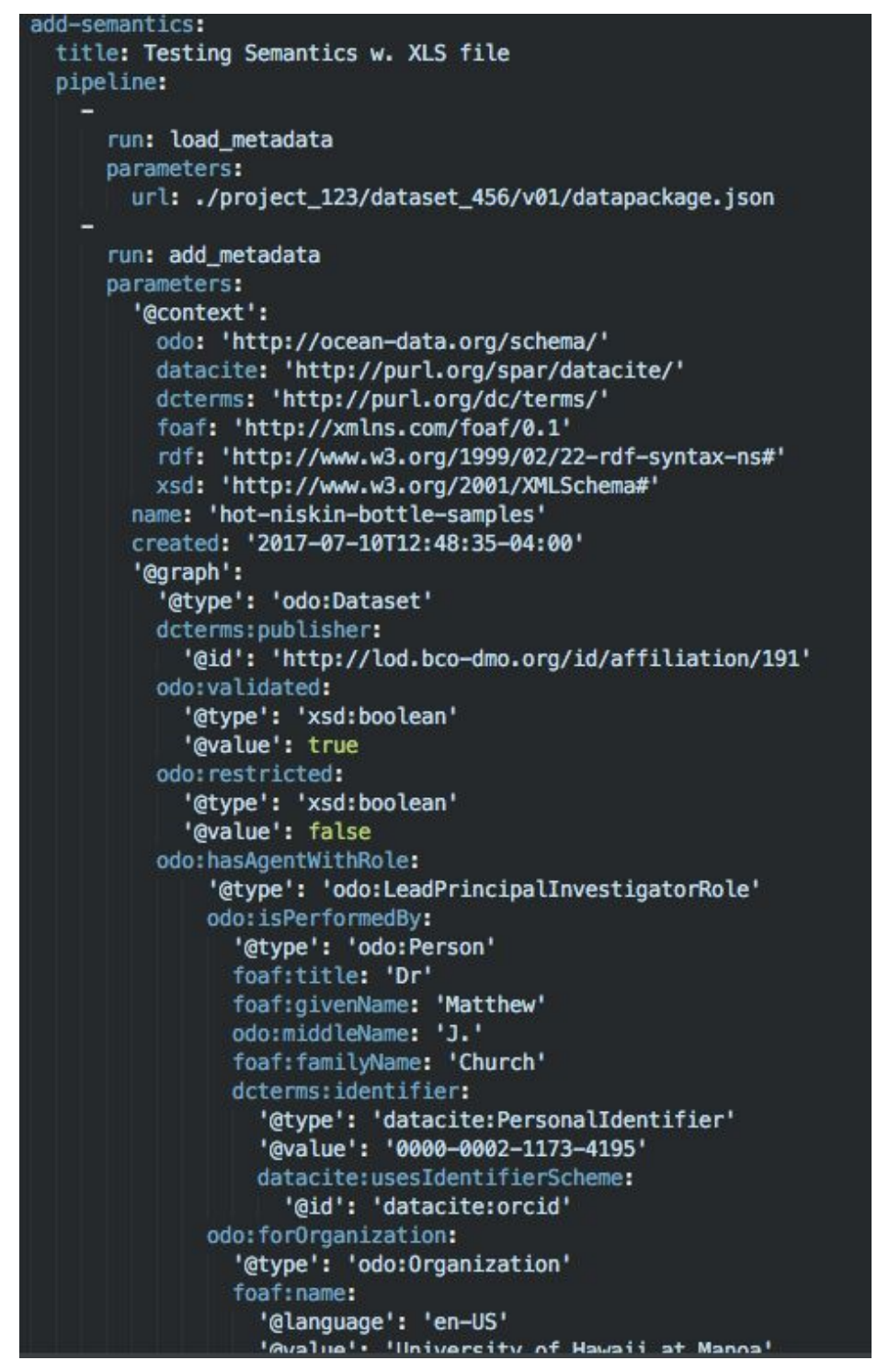
```
transform-xls:
  title: Transform an Excel File uploaded by a submitter to CSV
  pipeline:
    -
      run: load_metadata
      parameters:
        url: ./datapackage.json
    -
      run: add_resource
      parameters:
        url: ./data-uploaded-by-submitter.xlsx
        sheet: 1
        encoding: 'utf-8'
        name: 'niskin-bottle-samples'
    -
      run: stream_remote_resources
      cache: True
    -
      run: dump.to_path
      parameters:
          out-path: project_123/dataset_456/v01
```



```
   data-uploaded-by-submitter.xlsx
   datapackage.json
   pipeline-spec.yaml
▼ project_123
  ▼ dataset_456
    ▼ v01
      ▼ data
          niskin-bottle-samples.csv
        datapackage.json
```

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

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



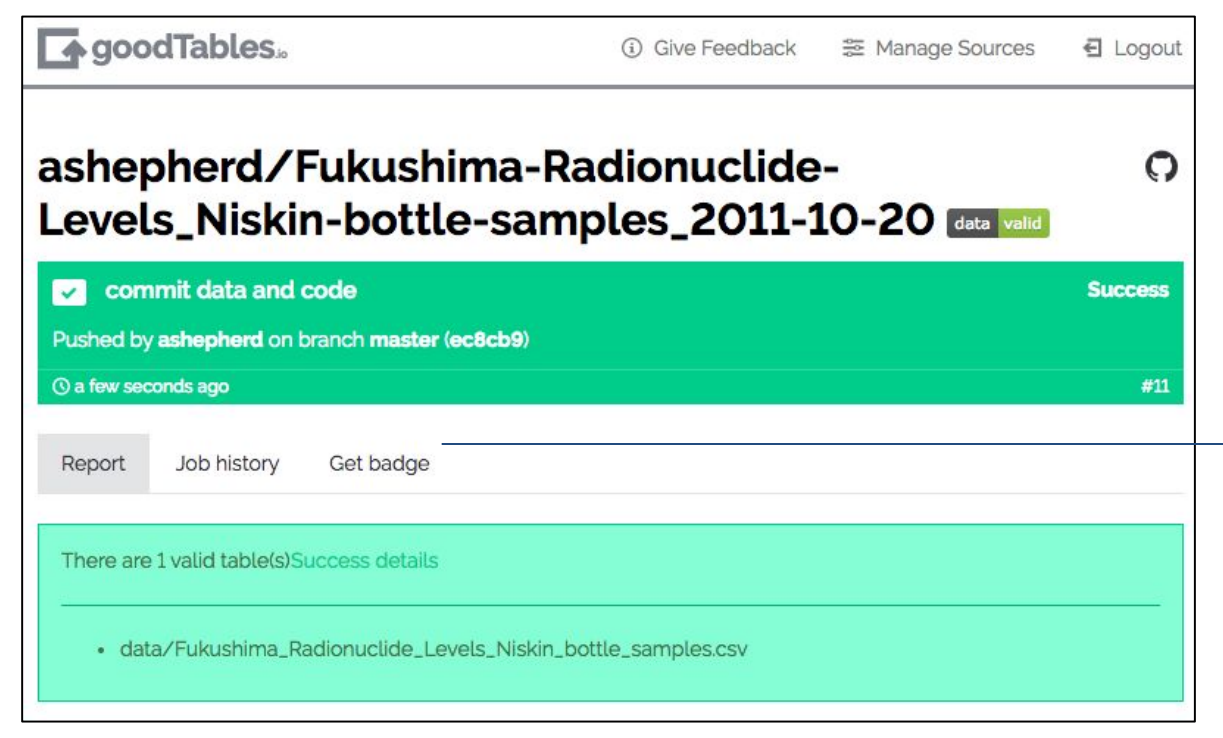
```
add-semantics:
  title: Testing Semantics w. XLS file
  pipeline:
    -
      run: load_metadata
      parameters:
        url: ./project_123/dataset_456/v01/datapackage.json
    -
      run: add_metadata
      parameters:
        '@context':
          odo: 'http://ocean-data.org/schema/'
          datacite: 'http://purl.org/spar/datacite/'
          dcterms: 'http://purl.org/dc/terms/'
          foaf: 'http://xmlns.com/foaf/0.1'
          rdf: 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'
          xsd: 'http://www.w3.org/2001/XMLSchema#'
        name: 'hot-niskin-bottle-samples'
        created: '2017-07-10T12:48:35-04:00'
        '@graph':
          '@type': 'odo:Dataset'
          dcterms:publisher:
            '@id': 'http://lod.bco-dmo.org/id/affiliation/191'
          odo:validated:
            '@type': 'xsd:boolean'
            '@value': true
          odo:restricted:
            '@type': 'xsd:boolean'
            '@value': false
          odo:hasAgentWithRole:
              '@type': 'odo:LeadPrincipalInvestigatorRole'
              odo:isPerformedBy:
                '@type': 'odo:Person'
                foaf:title: 'Dr'
                foaf:givenName: 'Matthew'
                odo:middleName: 'J.'
                foaf:familyName: 'Church'
                dcterms:identifier:
                  '@type': 'datacite:PersonalIdentifier'
                  '@value': '0000-0002-1173-4195'
                  datacite:usesIdentifierScheme:
                    '@id': 'datacite:orcid'
              odo:forOrganization:
                '@type': 'odo:Organization'
                foaf:name:
                  '@language': 'en-US'
                  '@value': 'University of Hawaii at Manoa'
```

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



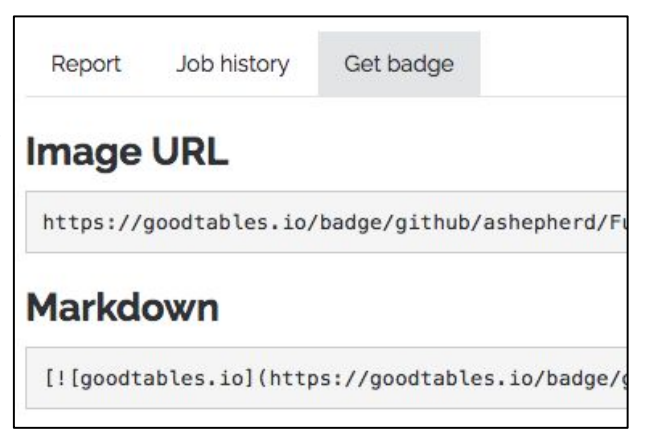
goodTables.io | Give Feedback | Manage Sources | Logout

ashepherd/Fukushima-Radionuclide-Levels_Niskin-bottle-samples_2011-10-20 data valid

☑ commit data and code — Success
Pushed by **ashepherd** on branch **master (ec8cb9)**
a few seconds ago — #11

Report | Job history | Get badge

There are 1 valid table(s)Success details
- data/Fukushima_Radionuclide_Levels_Niskin_bottle_samples.csv



Report | Job history | Get badge

Image URL
https://goodtables.io/badge/github/ashepherd/Fu

Markdown
[![goodtables.io](https://goodtables.io/badge/g

funding provided by National Science Foundation

Award# 1435578 Award# 1639714

for specs: https://frictionlessdata.io/
for tooling: https://frictionlessdata.io/software/

OPEN KNOWLEDGE INTERNATIONAL — With support from Alfred P. Sloan Foundation, Google.org, ODI open data institute

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