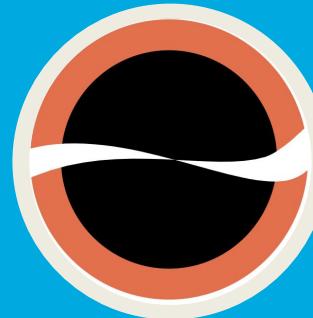


Frictionless Data for Reproducible Biology



Lilly Winfree, PhD
lilly.winfree@okfn.org . [@lilscientista](https://twitter.com/lilscientista)
[Frictionlessdata.io](https://frictionlessdata.io) .
github.com/frictionlessdata

SciPy 2020

Socially-
distanced
“Hello!”
from
Austin!



Science has a Data Management problem



Kate Laskowski
@KateLaskowski

I'm starting the year off with something I didn't expect to ever do: I'm retracting a paper. I recently discovered major problems in the raw data associated with it and so the results shouldn't be trusted.

[journals.uchicago.edu/toc/an/current...](https://journals.uchicago.edu/toc/an/current)

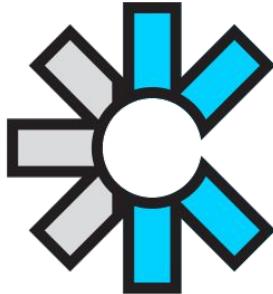
Retraction

The authors hereby retract the article “Individual and Group Performance Suffers from Social Niche Disruption,” published in the June 2016 issue (pp. 776–785) of *The American Naturalist*. After receiving a question from a reader about the publicly available data, the authors noticed irregularities in the raw data, which were collected in the laboratory of the third author. Unfortunately, the anomalies in the raw data



Open Knowledge
Foundation

<https://laskowskilab.faculty.ucdavis.edu/2020/01/29/retractions/>



Open Knowledge Foundation

For a Fair, Free, and Open Future: An open world, where all non-personal information is open, free, for everyone to use, build on and share, and creators and innovators are recognised and rewarded.

Build communities, tools and skills to empower individuals and organizations to use open information to create insights that drive change.

What is “Friction” in Data?

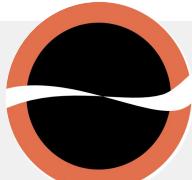
What does this column name mean?



Checking data quality

How was this analysis done?

Who created this data?

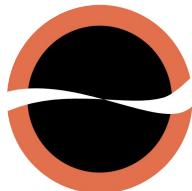


What is Frictionless Data?



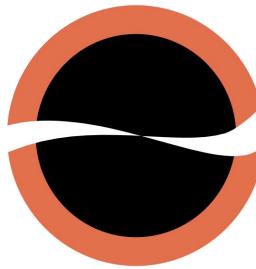
- Specifications for data & metadata interoperability
- Plus a collection of open source software libraries
- & a range of best practices for data management
- Platform agnostic interoperability

How can researchers use Frictionless Data?



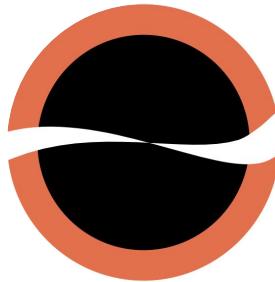
Frictionless Data for Reproducible Research

Removing the “friction” in *research* data to move from data to insight faster



Open source & community focused:
<https://github.com/frictionlessdata>

Frictionless Data for Reproducible Research



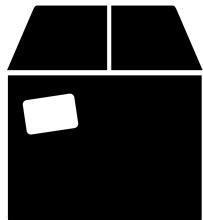
Alfred P. Sloan
FOUNDATION

- Fellows

- Tool Fund

- Pilots

3 Biology Frictionless Data Use Cases



Data Packages
+ Pipelines



Goodtables +
Table Schema

create.frictionlessdata.io

<https://github.com/frictionlessdata/datapackage-py>

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



Frictionless
Specifications



Lily Zhao,
Graduate
student
UC Santa
Barbara



Dr. Philippe
Rocca-Serra

BCO-DMO: Messy data → clean data → hosted data



How do we make this process reproducible??

`data.csv`

```
1 core,Depth,pH,alkalinity,Nitrate,Chlorinity,Ca,B,  
2 2014 bottom water,,7.92,2.32,21.1,544.9,10.17,413  
3 J2-733-PC 1,2,7.68,2.08,22.3,546.2,9.69,524,<0.1,  
4 J2-733-PC 1,6,7.69,2.11,23.8,546.2,9.64,535,0.4,<  
5 J2-733-PC 1,10,7.71,2.2,25.1,545.2,9.59,533,<0.1,  
6 J2-733-PC 1,13,7.7,2.22,25.8,547.2,9.62,531,0.2,<  
7 J2-733-PC 1,16,7.69,2.22,24.7,544.6,9.67,529,0.5,  
8 J2-733-PC 1,18,7.71,2.22,24.6,546.6,9.67,525,0.2,  
9 J2-733-PC 1,2,7.69,2.13,21.9,548.5,9.72,528,0.3,<  
10 J2-733-PC 1,7.7,2.17,24.6,543.9,9.65,536,<0.1,<  
11 J2-733-PC 1,11,7.73,2.18,25.5,546.2,9.62,532,<0.1,  
12 J2-733-PC 1,15,7.7,2.16,26.1,544.2,9.6,530,0.3,<  
13 J2-733-PC 1,18,7.72,2.14,25.7,545.9,9.64,519,0.4,  
14 J2-733-PC 1,20,7.7,2.16,25.4,546.3,9.62,527,0.2,<  
15 J2-733-PC 1,22,7.72,2.16,25.2,9.63,525,<0.1,<0.1,  
16 J2-733-PC 1,3,7.67,2.05,23,547.3,9.7,521,<0.1,<0.1,  
17 J2-733-PC 1,6,7.69,2.1,23.8,545.1,,516,0.2,<0.1,9,  
18 J2-733-PC 1,8,7.71,2.11,24.5,544.9,9.66,516,0.2,<0.1,  
19 J2-733-PC 1,10,7.75,2.13,25.1,544.2,9.64,517,0.2,  
20 J2-733-PC 1,12,7.71,2.13,25.3,544.9,9.61,514,0.1,<0.1
```

landing page

The landing page displays the following information:

- Dataset:** Water Chemistry
- Data:** Map, Data
- Spatial Extent:** N 22.82078 E-46.11082 S 22.8204 W-46.11082
- Temporal Extent:** 2014-04-11
- Parameters:** 1,415
- People:** 2,664
- Affiliations:** 583
- Funding:** 93
- Awards:** 1,966
- Principal Investigator:** Three kids are jumping in the train - Three kids are jumping in! - Three kids are jumping in the basket
- BCO-DMO Data Manager:** Shannon Rauch (Woods Hole Oceanographic Institution, WHOI BCO-DMO)
- Version Date:** 2019-04-11
- Restricted:** No
- Validated:** Yes



Open Knowledge
Foundation

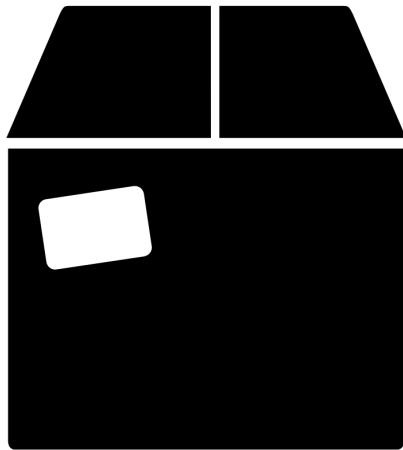
Slide modified from Amber York, BCO-DMO
<https://zenodo.org/record/2687557>

BCO-DMO

BCO-DMO data goal: use data packages

Data Package

Data
e.g.,
experiment.csv



Metadata
(+ optional
schema)

<https://specs.frictionlessdata.io/data-package/>

<https://github.com/frictionlessdata/datapackage-py>

BCO-DMO data goal: keep rich metadata

Temp	Cond	Salinity	pH	Cell_counts	
15.23	29.34	22.82	nd	4.53E+6	raw data
17.99	30.94	22.59	8.01	1.96E+6	
15.24	20.95	15.77	8.26	6.62E+6	
22.02	39.66	27.04	8.13	1.93E+6	
28.87	19.1	Temp		Temperature measured via YSI at the sampling site	degrees Celsius
29.03	15.6				
29.23	48.7				
28.98	39.9	Cond		Conductivity measured via YSI at the sampling site	ms/cm
23.76	42.9				
18.61	21.5	Salinity		Calculated salinity based on the conductivity measurement from the YSI	unitless
https://www.bco-dmo.org/d9					
metadata		pH		pH of the seawater at the sampling site	unitless
		Cell_counts		Number of cells passing through a 2.7 μm filter (Whatman GF/D) determined via flow cytometry	cells/mL

Data Package Creator: Create.frictionlessdata.io

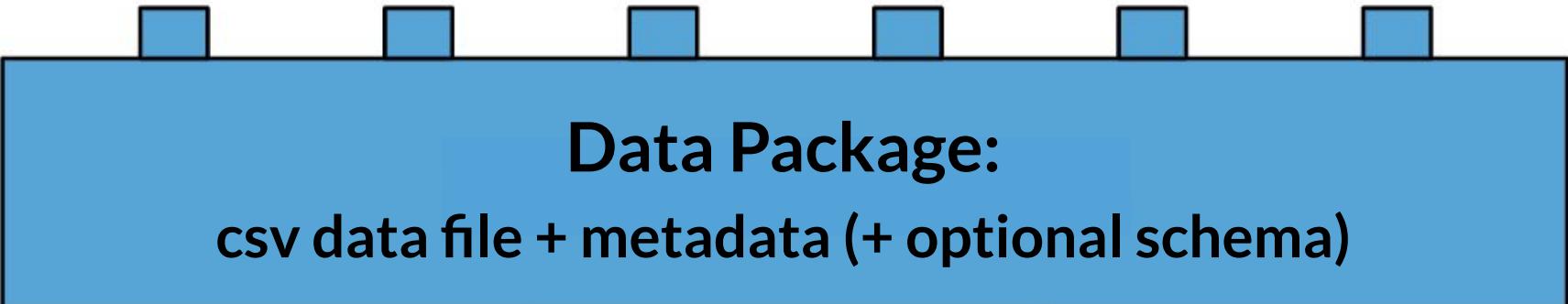
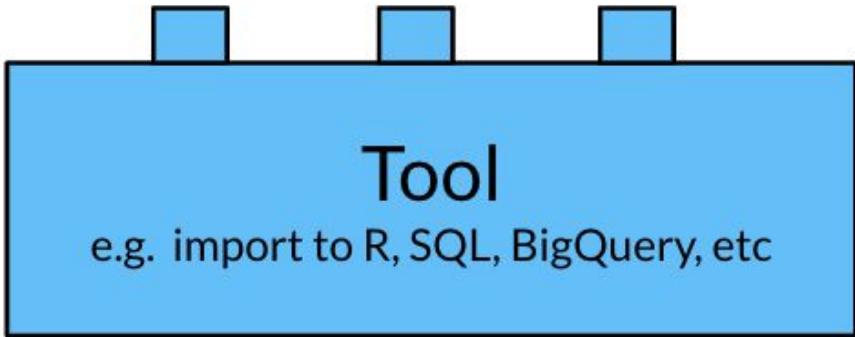
Name resc Path sar11.csv Load

Resource	Title	Description	Type	Format
Site	Site		string	default
Lat	Lat		integer	default
Lon	Lon		string	default
Date	Date		integer	default
Time_Central	Time_Central		string	default
Offset	Offset		string	default

▶ Preview

```
{  
  "profile": "tabular-data-package",  
  "resources": [  
    {  
      "name": "resource1",  
      "path": "sar11.csv",  
      "profile": "tabular-data-resource",  
      "schema": {  
        "fields": [  
          {  
            "name": "Site",  
            "type": "string",  
            "format": "default",  
            "title": "Site"  
          },  
          {  
            "name": "Lat",  
            "type": "integer",  
            "format": "default",  
            "title": "Lat"  
          },  
          {  
            "name": "Lon",  
            "type": "string",  
            "format": "default",  
            "title": "Lon"  
          },  
          {  
            "name": "Date",  
            "type": "integer",  
            "format": "default",  
            "title": "Date"  
          }  
        ]  
      }  
    }  
  ]  
}
```

Packaged data is useful data



BCO-DMO goal: use reproducible data pipelines

Data Package Pipelines (DPP): data processing pipelines

- Python framework for declarative processing of tabular data
- Standardize data processing steps
 - e.g. joins, find and replace, add/remove columns, unpivot
- Can write custom processors in python
- Pipelines are defined in `pipeline-spec.yaml` files
 - Specifies processors + execution parameters → reproducibility!
- Generates a single data package as its output

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

Frictionless, reproducible research @ BCO-DMO

	A	B	C
1	Site Code	Site Code	Deployment Dates
2	1	Dittlif Point	6/1/16 - 3/22/17
3			3/27/17 - 6/22/17
4	2	Cocoloba Cay	5/29/16 - 3/22/17
5			3/27/17 - 7/11/17
6			5/29/16 - 10/22/16
7	3	Joel's Shoal	11/10/16 - 3/22/17
8			3/28/16 - 7/11/17
			5/29/16 - 10/22/16



pipeline-spec.yaml

```
- run: join
parameters:
  source:
    name: world_population
    key: ["country_code"]
    delete: yes
  target:
    name: country_gdp_2015
    key: ["CC"]
  fields:
    population:
      name: "census_2015"
  full: true
```

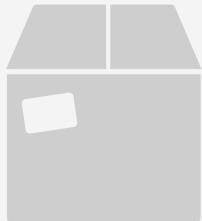
datapackage.json

```
1 {
2   "bytes": 24061,
3   "count_of_rows": 433,
4   "hash": "c3aaad307223086fa611c40f9ab8ae100",
5   "name": "_",
6   "resources": [
7     {
8       "bytes": 24061,
9       "count_of_rows": 433,
10      "dialect": {
11        "delimiter": ",",
12        "doubleQuote": true,
13        "lineTerminator": "\r\n",
14        "quoteChar": "\"",
15        "skipInitialSpace": false
16      },
17      "dpp:streamedFrom": "http://datadocs.bco-dmo.org/docs/TestProject/data",
18      "encoding": "utf-8",
19      "format": "csv",
20      "hash": "405e348a5bb172c191abbe8d5a72880b",
21      "headers": 1,
22      "name": "mcmurdo_epifauna",
23      "path": "data/mcmurdo_epifauna.csv",
24      "schema": {
25        "fields": [
26          {
27            "decimalChar": ".",
28            "groupChar": ",",
29            "name": "year"
30          }
31        ]
32      }
33    }
34  ]
35}
```

data.csv

```
1 Core,Depth,pH,alkalinity,Nitrate,Chlorinity,Ca,B,
2 2014 bottom water,,7.92,2.32,21.1,544.9,10.17,413
3 J2-733-PC 1,2,7.68,2.08,22.3,546.2,9.69,524,<0.1,
4 J2-733-PC 1,6,7.69,2.11,23.8,546.2,9.64,535,0.4,<
5 J2-733-PC 1,10,7.71,2.2,25.1,545.2,9.59,533,<0.1,
6 J2-733-PC 1,13,7.7,2.22,25.8,547.2,9.62,531,0.2,<
7 J2-733-PC 1,16,7.69,2.22,24.7,544.6,9.67,529,0.5,
8 J2-733-PC 1,18,7.71,2.22,24.6,546.6,9.67,525,0.2,
9 J2-733-PC 2,2,7.69,2.13,21.9,548.5,9.72,528,0.3,<
10 J2-733-PC 2,7,7.7,2.17,24.6,543.9,9.65,536,<0.1,<
11 J2-733-PC 2,11,7.73,2.18,25.5,546.2,9.62,532,<0.1
12 J2-733-PC 2,15,7.7,2.16,26.1,544.2,9.6,530,0.3,<0.1
13 J2-733-PC 2,18,7.72,2.14,25.7,545.9,9.64,519,0.4,
14 J2-733-PC 2,20,7.7,2.16,25.4,546.3,9.62,527,0.2,<
15 J2-733-PC 2,22,7.72,2.16,25.2,,9.63,525,<0.1,<0.1
16 J2-733-PC 4,3,7.67,2.05,23,547.3,9.7,521,<0.1,<0.1
17 J2-733-PC 4,6,7.69,2.1,23.8,545.1,,516,0.2,<0.1,9
18 J2-733-PC 4,8,7.71,2.11,24.5,544,9.66,516,0.2,<0.2
19 J2-733-PC 4,10,7.75,2.13,25.1,544.2,9.64,517,0.2,
20 J2-733-PC 4,12,7.71,2.13,25.3,544,9.61,514,0.1,<0.1
```

3 Biology Frictionless Data Use Cases



Data Packages
+ Pipelines



Goodtables +
Table Schema



Lily Zhao,
Graduate
student
UC Santa
Barbara



try.goodtables.io

<https://github.com/frictionlessdata/goodtables-py>

<https://github.com/frictionlessdata/tableschema-py>

A data validation tale...

```
{  
  "name": "group",  
  "type": "string",  
  "format": "default",  
  "description": "whether respondent is a scientist or resident",  
  "title": "group"  
},  
{  
  "name": "age_bracket",  
  "type": "string",  
  "format": "default",  
  "title": "age bracket",  
  "description": "age bracket respondent belongs to"  
},  
{  
  "name": "age_range",  
  "type": "number",  
  "format": "default",  
  "title": "age range",  
  "description": "age range respondent belong to"  
},  
https://github.com/frictionlessdata/tableschema-py
```



group	age_bracket	age_range	gender
scientist	25 to 35	under 35	F
scientist	25 to 35	under 35	F

Data from Lily Zhao, FD Fellow 2019/2020:
https://github.com/lilyzzhao/resident-scientist-data/blob/master/resident_researcher_data.csv

Goodtables: validate your data with a schema

157 x

Type or Format Error

The value does not match the schema type and format for this field.

How it could be resolved:

- If this value is not correct, update the value.
- If this value is correct, adjust the type and/or format.
- To ignore the error, disable the `type-or-format-error` check in `goodtables.yml`. In the values will be ignored.

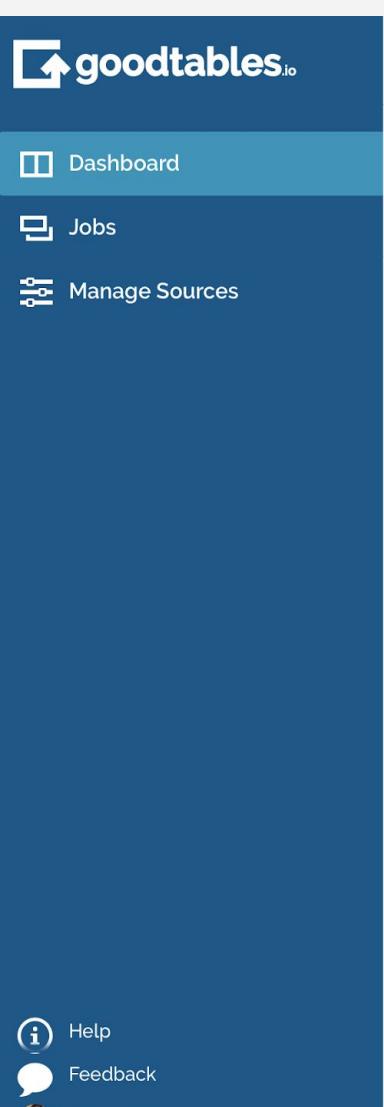
The full list of error messages:

The value "under 35" in row 2 and column 5 is not type "number" and format "default"

1	document_name	role	group	age_bracket	age_range	gender	row_index
2	S16	scientist	scientist	25 to 35	under 35	F	



Continuous data validation: <http://goodtables.io>



The sidebar contains the Goodtables logo, navigation links for Dashboard, Jobs, Manage Sources, Help, and Feedback, and a feedback icon.

- Dashboard
- Jobs
- Manage Sources
- Help
- Feedback

Action required

frictionlessdata/ckanext-wprdcpilot

11 minutes ago - #1 (2028c1) - ERROR

frictionlessdata/pilot-causanatura

9 months ago - #7 (79c0a9)

1	ENTIDAD_FEDERATIVA	CLAVE_SITIO_DESEM
269	BAJA CALIFORNIA	02A
424	BAJA CALIFORNIA	02A
450	BAJA CALIFORNIA	02A
639	BAJA CALIFORNIA	02A

datos/Produccion/2014-2015/AvisosArriboNacional2014.csv

1	ENTIDAD_FEDERATIVA	CLAVE_SITIO_DESEM
425	BAJA CALIFORNIA	201000
718	BAJA CALIFORNIA	201000
784	BAJA CALIFORNIA	201000
965	BAJA CALIFORNIA	201000

datos/Produccion/2014-2015/AvisosArriboNacional2015.csv

1					
3					
24					
25					
26					

datos/inspeccion/PRESUPUESTO DGIV 2014-2015 (ANEXO 1).ods

1					
2					
42					
43					

datos/inspeccion/RESULTADOS 2014-2015 (ANEXO 3).ods

1	AÑO	ESTADO	MUNICIPIO
201	2012	BAJA CALIFORNIA	ENSENADA

1	AÑO	ESTADO	MUNICIPIO
178	2012	SINALOA	AHOME

Jobs

github/frictionlessdata/ckanext-wprdcpilot

11 minutes ago

#1 (2028c1)

github/frictionlessdata/example-data-packages

14 minutes ago

#26 (451323)

github/serahrono/color_codes

14 days ago

#11 (74dae2)

github/frictionlessdata/pilot-causanatura

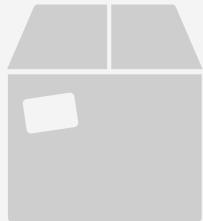
9 months ago

#7 (79c0a9)



Open Knowledge
Foundation

3 Biology Frictionless Data Use Cases



Data Packages
+ Pipelines



Goodtables +
Table Schema

{ } Frictionless
Specifications



Lily Zhao,
Graduate
student
UC Santa
Barbara



Dr. Philippe
Rocca-Serra

FAIR + Frictionless Rose Data

Letter | OPEN | Published: 30 April 2018

The Rosa genome provides new insights into the domestication of modern roses

Olivier Raymond, Jérôme Gouzy, [...] Mohammed Bendahmane

Protocol information

repeated 3 times
fully opened flowers : 1 g petals in 2 mL hexane containing 5 mg/L camphor
injection of 2 microl, split 2:1, DB-5 gas chromatography column for GC-MS (Agilent 6890 with helium)
3 min 40°C, 1.5°C/min until 180°C, 6°C/min until 240°C
ionizing voltage 70eV, mass scan rate 35-450 m/z, mass scan rate 2.94/s
Data analysis made with a threshold of 15 to detect major peaks between 4 and 80 min (except farnesyl acetates after 80 min)
Long chain hydrocarbons are not presented
Compounds determined with Wiley database, NIST online database, purified standard and bibliography
Results in microg/g fresh weight of petals
For Rosa chinensis 'Old Blush', sepals and stamens were analyzed in addition to petals.

Table of actual data

Compound	Average	Standard Error	Average
hexan-2-ol			
hexanal	4.95	0.59	0.90
(E)-2-hexenal	57.62	7.34	9.41
(Z)-3-hexen-1-ol	7.64	0.63	5.39
(E)-2-hexen-1-ol	1.79	0.98	
hexan-1-ol			
nonane	2.09	0.07	
alpha-pinene			
benzaldehyde			



Extending Table Schema: Rose Data

489 lines (489 sloc)

Search this file...

chemical_name

hexan-2-ol

hexanal

(E)-2-hexenal

(Z)-3-hexen-1-ol

(E)-2-hexen-1-ol

hexan-1-ol

nonane

alpha-pinene

benzaldehyde

beta-myrcene

(Z)-3-hexenyl acetate

hexyl acetate

(E)-hexenyl acetate

(+/-)-limonene

benzylalcohol

<https://github.com/processor/data/processed/r018-treatment-group-n>

```
{  
    "name": "sample_size",  
    "title": "sample size",  
    "description": "statistical sample size is a count evaluating the number of individual experimental units in a",  
    "format": "default",  
    "type": "integer",  
    "rdfType": "http://purl.obolibrary.org/obo/STATO_0000088",  
    "constraints": {  
        "required": false  
    },  
    {  
        "name": "sample_mean",  
        "title": "sample mean",  
        "description": "the sample mean is a measure of dispersion of the observations made on the sample and provides",  
        "format": "default",  
        "type": "number",  
        "rdfType": "http://purl.obolibrary.org/obo/STATO_0000401",  
        "constraints": {"required": true}  
    },  
    {  
        "name": "unit",  
        "title": "unit",  
        "description": "the unit associated with the sample mean",  
        "format": "default",  
        "type": "string",  
        "rdfType": "",  
        "constraints": {"required": false}  
    },  
    {  
        "name": "sem",  
        "title": "standard error of the mean",  
        "description": "The standard error of the mean (SEM) is data item denoting the standard deviation of the sample",  
        "format": "default",  
        "type": "number",  
        "rdfType": "http://purl.obolibrary.org/obo/STATO_0000037",  
        "constraints": {"required": false}  
    },  
}
```



<https://github.com/ISA-tools/stato>

ata_integration.csv",

l compound",

y is meant to hold inform

197",

\(\backslash\)\\\\[,]+)\\$/ig"

018ng-notebook/blob/e-aroma-data-integrati

Publish Rose DataPackages on Zenodo

zenodo

Search

Upload Communities

proccaserra@gmail.com

July 9, 2019

Lesson Open Access

Experimental design-driven FAIRification of data matrices: example of a principled approach

Philippe Rocca-Serra; Susanna Assunta Sansone

We outline a principled approach to data FAIRification rooted in the notions of study design. This is an example of retrospective data FAIRification, using as a metabolomics dataset associated to a published in a journal article.

SUMMARY

- Our first data source: article by Raymond et al. Nat Genet. 50:772-777 (2018) <https://doi.org/10.1038/s41588-018-0110-3>; this is targeted metabolite profiling study of strain-related chemical signatures of the rose fragrance; the biological materials was selected to allow a comparison between parts of the plant, and across cultivars in the same tissue type.
 - Our starting point: their human-understandable data in the supplementary table https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-018-0110-3/MediaObjects/41588_2018_110_MOESM3_ESM.zip, containing the mean concentrations of 61 metabolites measured in three different parts of the rose flower, in six distinct genotypes.
- Our second data source: article by Magnard et al. Science.Jul 3,349(6243):81-3 (2015) <https://doi.org/10.1126/science.aab0696>; this is early work of the same group author of the first data source.
- Our approach: we performed a retrospective curation and re-annotation of the data matrices, disambiguating of the experimental design, using community, open interoperability standards from FAIRsharing (<https://fairsharing.org>); we focused on the clarity of the statistical results to ensure reusability and reproducibility of the analytical workshop by humans and machines. The FAIRification steps for the first data source are documented in the sections below; the same steps were applied to the second data source to assess inter-experiment agreement, as both studies used the same varieties of rose and plant parts.
- Our results: semantically-anchored data matrices served as Linked Data, deposited in public archives (Zenodo and MetaboLights), and consumable by software agents for queries like "Retrieve study predictor variables and their levels" and "What is sample size used to compute the means?" to support study results review and assessment.
- It is associated to the following project: <https://github.com/proccaserra/rose2018ng-notebook> with all the necessary information, executable code and tutorials in the form of Jupyter notebooks.

0 views 0 downloads See more details...

Indexed in

OpenAIRE

Publication date: July 9, 2019

DOI: [10.5281/zenodo.3274257](https://doi.org/10.5281/zenodo.3274257)

Keyword(s): Design of Experiment, FAIR, data matrix, metabolomics, Tabular Data Package, JSON Data Package, ISA format, CHEBI ontology, STATO ontology

Grants:

European Commission:

- PhenoMeNal - PhenoMeNal: A comprehensive and standardised e-infrastructure for analysing medical metabolic phenotype data (654241)



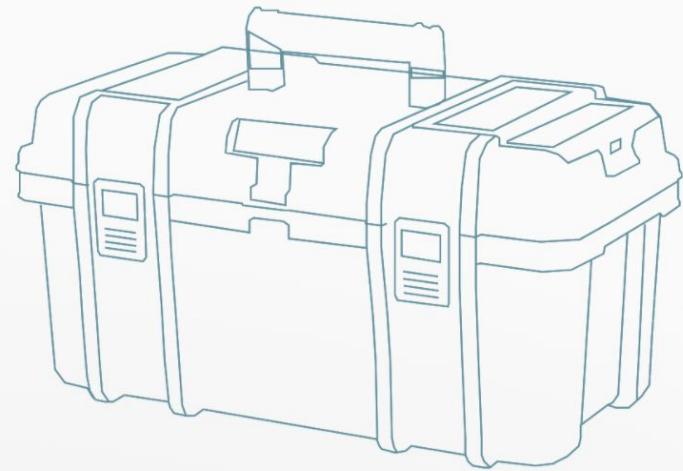
More info:
<https://vimeo.com/423719773>

<https://doi.org/10.5281/zenodo.3274257>

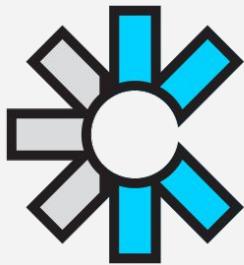
Want to try it yourself?

The progressive data toolkit

Frictionless Data is a progressive, incrementally adoptable open-source toolkit that brings simplicity and gracefulness to the data experience - whether you're wrangling a CSV or engineering complex pipelines with gigabytes.

[Why Frictionless Data?](#)[Get Started](#)

frictionlessdata.io/guide/



Open Knowledge
Foundation



<http://github.com/frictionlessdata/>

<https://discord.com/invite/j9DNFNw>

[youtube.com/user/openknowledgefdn](https://www.youtube.com/user/openknowledgefdn)

<https://frictionlessdata.io/guide/>

Twitter: [@frictionlessd8a](https://twitter.com/@frictionlessd8a)

Thank you!

Join our
community!



Did you record the
metadata?





BCO-DMO data is MESSY

Table . Pore water chemical concentration data and location.

Some Mn and Fe values are below detection and are listed as 0.

Core Depth cm pH alkalinity mmol/kg titration

METHOD electrode

2014 bottom water 7.92 2.32

J2-733-PC 1 - Did not hit bottom (18:24) and was positioned next to PC 2. 22.0

2 7.68 2.08
6 7.69 2.11
10 7.71 2.20
13 7.7 2.22
16 7.69 2.22
18 7.71 2.22

J2-733-PC 2 - Did not hit bottom (18:27) and was positioned next to PC 1. 22.0

2 7.69 2.13
7 7.70 21.9
11 7.73 548.
15 7.70
18 7.72
20 7.70
22 7.72

J2-733-PC 4 - Hit bottom (18:01) and was positioned next to PC 3. 22.0

3 7.67
6 7.69
8 7.71
11 7.75
13 7.71
15 7.73

Site Code Site Code Deployment Dates

1 Dittli Point 6/1/16 - 3/22/17
2 Cocoloba Cay 3/27/17 - 6/22/17
3 Joel's Shoal 5/29/16 - 3/22/17
4 White Point 3/27/17 - 7/11/17
5 Europa Bay 5/29/16 - 10/22/16
6 11/10/16 - 3/22/17
7 3 3/28/16 - 7/11/17
8 5/29/16 - 10/21/16
9 10/23/16 - 3/23/17
10 11 5 5/29/16 - 10/21/16
11 12 12/23/16 - 12/12/16
12 13 5/29/16 - 10/21/16
13 14 6 Tektite 11/5/16 - 3/20/17
14 15 3/28/17 - 6/14/17
15 16 17 7 Yab
16 18
17 19
18 20 8 Bo
19 21
20 22
21 23 9 Ra
22 24
23 25
24 26 S Re
25 27
26 28

Cnidaria species: Moon Jellyfish (*Aurelia aurita*)

Individual: 1

Video ID	Time (minute)	E	D	C.T.	C.O.
Clip015	4.06			X	
	13.04	X			
	26.23			X	
	30.17			X	
	45.09			X	
	134.06			X	
Clip016	10.06			X	
	27.45			X	
	32.22	X			
	50.13			X	
Clip017	10.06			X	
	27.2			X	
	25.45	X			
	30.24				
	36.02				
	39.15	XX			
	43.19				
	1.06.27				

Open Knowledge Foundation

Slide modified from Amber York, BCO-DMO
<https://zenodo.org/record/2687557>

@lilscientista

“Friction” in BCO-DMO data

BCO-DMO data managers:

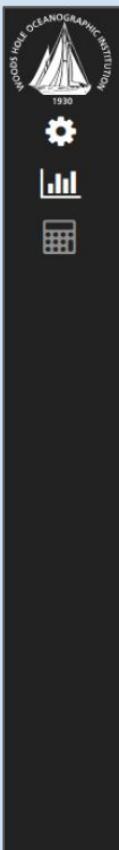
- Add spatio-temporal context in standard formats
 - date/time (ISO 8601), timezones
 - lat, lon
 - depth
- Correct quality issues
 - Inconsistent formatting
 - corrupt data characters
 - data gaps
 - invalid species names
 - typos
- Reformat for reusability

How does Frictionless Data Package Pipelines help?

- Created a web tool for data managers to clean the data
- Reduced dataset processing time
- Removed barrier of programming ability
- Documented a reproducible workflow
- Created packaged data to be published for others to (re)use



BCO-DMO pipeline tool



usecase_746395_PierCTD

- 1 ▾ Load + ➜ ✖
- 2 ▾ Round field + ➜ ✖
- 3 ▾ Find and replace + ➜ ✖
- 4 ▾ Add a computed field + ➜ ✖
- 5 ▾ Convert date + ➜ ✖
- 6 ▾ Rename fields + ➜ ✖
- 7 ▾ Delete fields + ➜ ✖
- 8 ▾ Add a computed field + ➜ ✖
- 9 ▾ Add a computed field + ➜ ✖
- 10 ▾ Reorder fields + ➜ ✖
- 11 ▾ Set types ↴ + ➜ ✖

3

Find and replace

Processor: Find and replace

Resource: ctd

Field: Date
Time
conductivity
temperature

Find pattern: `^(\d+:\d+)$`

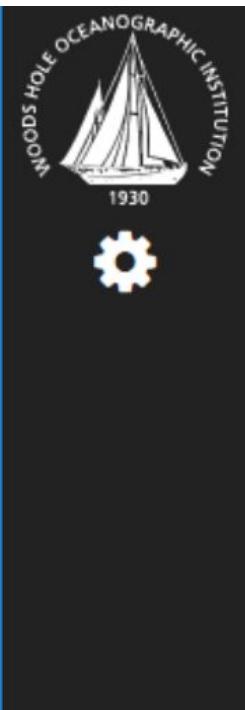
Replace pattern: `\1:00`

Notes: Fix inconsistent time format (some didn't have seconds).

pipeline-spec.yaml

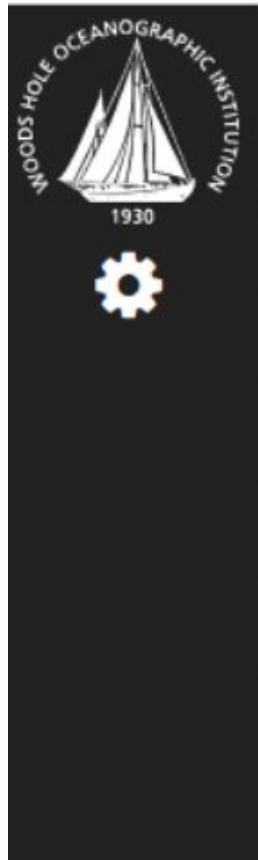
```
- run: find_replace
  bcodmo_notes: Fix inconsistent time format
  (some didn't have seconds).
  cache: true
  parameters:
    fields:
      - name: Time
        patterns:
          - {find: '^(\d+:\d+)$', replace: '\1:00'}
  resources: [ctd]
```

Example pipeline processor step: change date format



#	Date	Time	conduct	tempera	pressure	dissolve
	string	string	string	string	string	string
1	4/16/18	9:51:48	42.13	15.86	1.18	98.17
2	4/16/18	9:51:48	42.13	15.85	1.15	98.18
3	4/16/18	9:51:48	42.13	15.85	1.14	98.19
4	4/16/18	9:51:49	42.13	15.84	1.12	98.21
5	4/16/18	9:51:49	42.13	15.84	1.09	98.19

Example pipeline processor step: change date format



#	ISO_DateTime_UTC	Date_local	Time_local
	datetime	date	time
1	2018-04-16T16:51:48	2018-04-16	09:51:48
2	2018-04-16T16:51:48	2018-04-16	09:51:48
3	2018-04-16T16:51:48	2018-04-16	09:51:48
4	2018-04-16T16:51:49	2018-04-16	09:51:49
5	2018-04-16T16:51:49	2018-04-16	09:51:49
6	2018-04-16T16:51:49	2018-04-16	09:51:49



OPEN KNOWLEDGE FOUNDATION



PASSIONATE TEAM

Passionate about openness. Using advocacy, technology and training to unlock information and enable people to create and share knowledge.



GLOBAL NETWORK

Meet, campaign, learn, innovate, share, train, create, support, explore: some of the ways you can help open up knowledge for everyone. Join us.



DIVERSE PROJECTS

Through our projects, research and collaborations, we explore niche areas of data, and ways in which it can be used to empower people around the world.

@okfn

@lilscientista

BCO-DMO Pilot next steps

- Release of an open-source community version of the BCO-DMO pipeline UI, custom processors, & statistics calculator
- Allow the public to re-run pipelines, or build upon existing pipelines
- Validation and QA/QC using goodtables

Learn more about the great work BCO-DMO is doing:

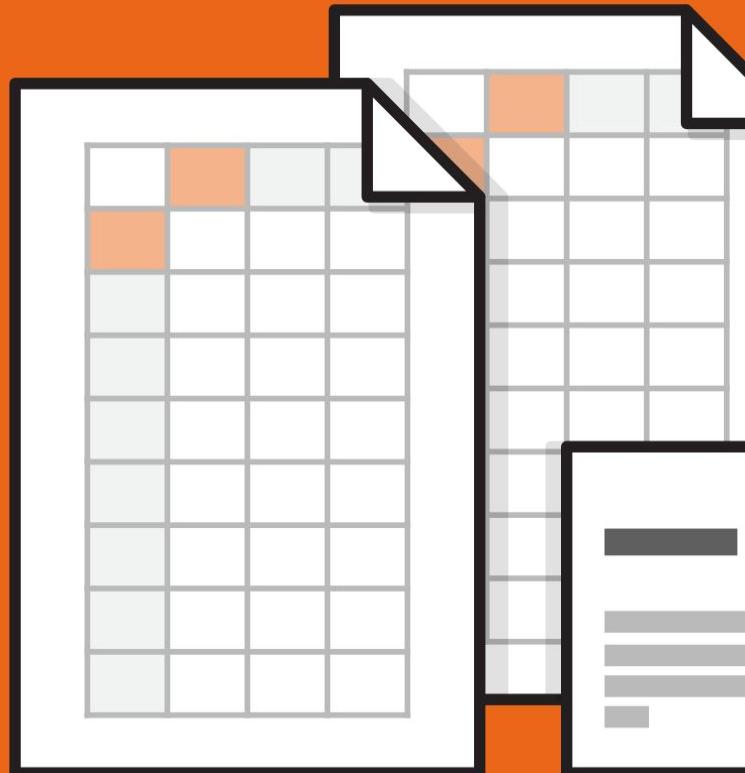
<https://bco-dmo.org>; @BCO-DMO

<https://github.com/BCODMO>

FRictionless DATA

Field guide

Data Package Creator



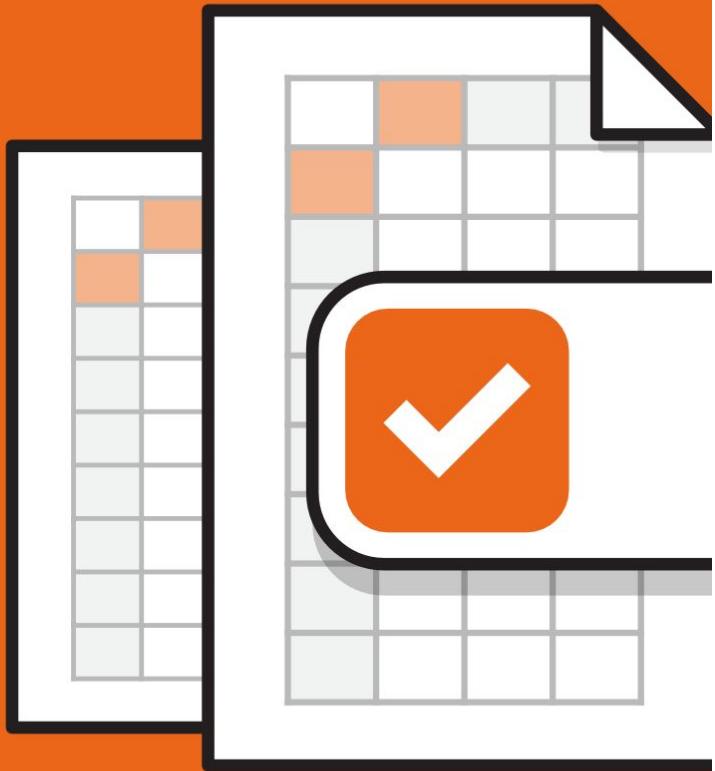
Create.frictionlessdata.io

Data you can play with: <https://bit.ly/2BhdOj2>

FRictionless DATA

Field guide

Validating data with **try.goodtables**



[Try.goodtables.io](https://try.goodtables.io)

<https://bit.ly/2P6oogW> Or lots of example data:

<https://github.com/frictionlessdata/goodtables-py/tree/master/data>

Continuous data validation: <http://goodtables.io>



Sign in with GitHub



repository/my-data



a minute ago