Frictionless Data for Reproducible Biology

Lilly Winfree, PhD
lilly.winfree@okfn.org . @lilscientista
Frictionlessdata.io .
github.com/frictionlessdata

SciPy 2020

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...

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

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?

How was this analysis done?

Checking data quality

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?

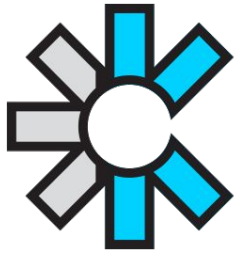
Frictionless Data for Reproducible Research

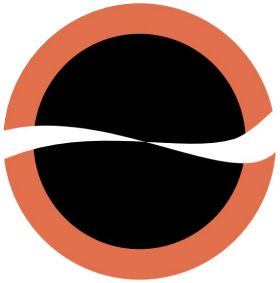
Removing the “friction” in *research* data to move from data to insight faster




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

Frictionless Data for Reproducible Research

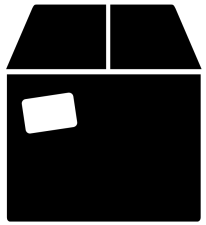




Alfred P. Sloan
FOUNDATION

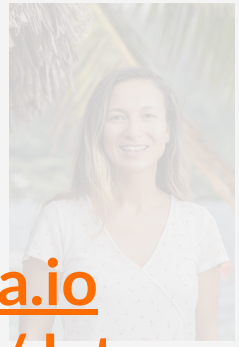
- Fellows
- Tool Fund
- Pilots

3 Biology Frictionless Data Use Cases



Data Packages
+ Pipelines

BCO-DMO



Lily Zhao,
Graduate
student
UC Santa
Barbara

Goodtables +
Table Schema

create.frictionlessdata.io

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

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

{ }
Frictionless
Specifications

Dr. Philippe
Rocca-Serra

Open Knowledge
Foundation

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

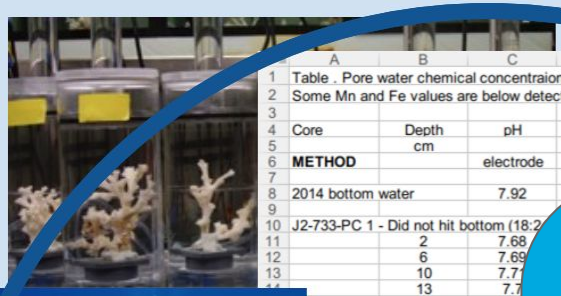
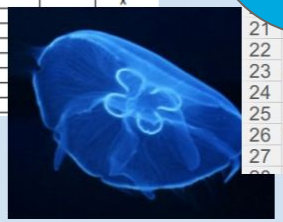


Table . Pore water chemical concentraion data and lo
Some Mn and Fe values are below detection and are

Core	Depth cm	pH	alkalinity mmol/kg
METHOD		electrode	titration
2014 bottom water		7.92	
J2-733-PC 1 - Did not hit bottom (18:2			
	2	7.68	
	6	7.69	
	10	7.7	
	13	7.7	
	16	7.6	
	18	7.7	
-733-PC 2 - Did not hit bottom (1			
	2	7.6	
	7	7.7	
	11	7.7	
	15	7.7	
	18	7.7	
	20	7.7	
	22	7.7	
-733-PC 4 - Hit bottom (18:01) ar			
	3	7.6	
	6	7.6	
	8	7.7	

Video ID	Time (minute)	E	D	C.T.	C.O.
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	
	1.34.06				x
Clip	10.06				x
	27.45				x
	32.22	x			
	50.13				x
Clip017	10.06				x
	27.2				
	25.45	x			
	30.24				
	36.02				
	39.15	xx			
	43.19				
	1.06.27				

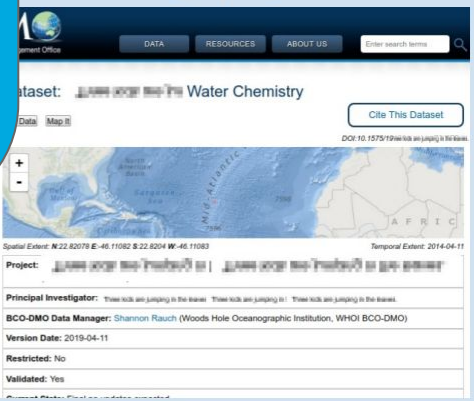


How do we make this process reproducible??

data.csv

```
Core,Depth,pH,alkalinity,Nitrate,Chlorinity,Ca,B,
2014 bottom water,,7.92,2.32,21.1,544.9,10.17,413
J2-733-PC 1,2,7.68,2.08,22.3,546.2,9.69,524,<0.1,
J2-733-PC 1,6,7.69,2.11,23.8,546.2,9.64,535,0.4,<
J2-733-PC 1,10,7.71,2.2,25.1,545.2,9.59,533,<0.1,
PC 1,13,7.7,2.22,25.8,547.2,9.62,531,0.2,<
1,16,7.69,2.22,24.7,544.6,9.67,529,0.5,
,18,7.71,2.22,24.6,546.6,9.67,525,0.2,
2,7.69,2.13,21.9,548.5,9.72,528,0.3,<
7,7.7,2.17,24.6,543.9,9.65,536,<0.1,<
11,7.73,2.18,25.5,546.2,9.62,532,<0.1
15,7.7,2.16,26.1,544.2,9.6,530,0.3,<0
18,7.72,2.14,25.7,545.9,9.64,519,0.4,
20,7.7,2.16,25.4,546.3,9.62,527,0.2,<
22,7.72,2.16,25.2,,9.63,525,<0.1,<0.1
3,7.67,2.05,23,547.3,9.7,521,<0.1,<0.
6,7.69,2.1,23.8,545.1,,516,0.2,<0.1,5
8,7.71,2.11,24.5,544,9.66,516,0.2,<0.
10,7.75,2.13,25.1,544.2,9.64,517,0.2,
12,7.71,2.13,25.3,544,9.61,514,0.1,<0
```

landing page



Open Knowledge Foundation

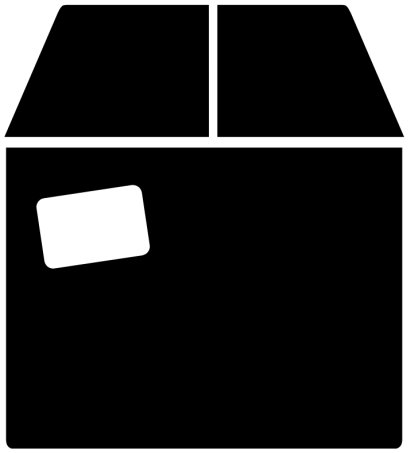
Slide modified from Amber York, BCO-DMO
https://zenodo.org/record/2687557

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
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			
29.03	15.6			
29.23	48.7			
28.98	39.9			
23.76	42.9			
18.61	21.5			

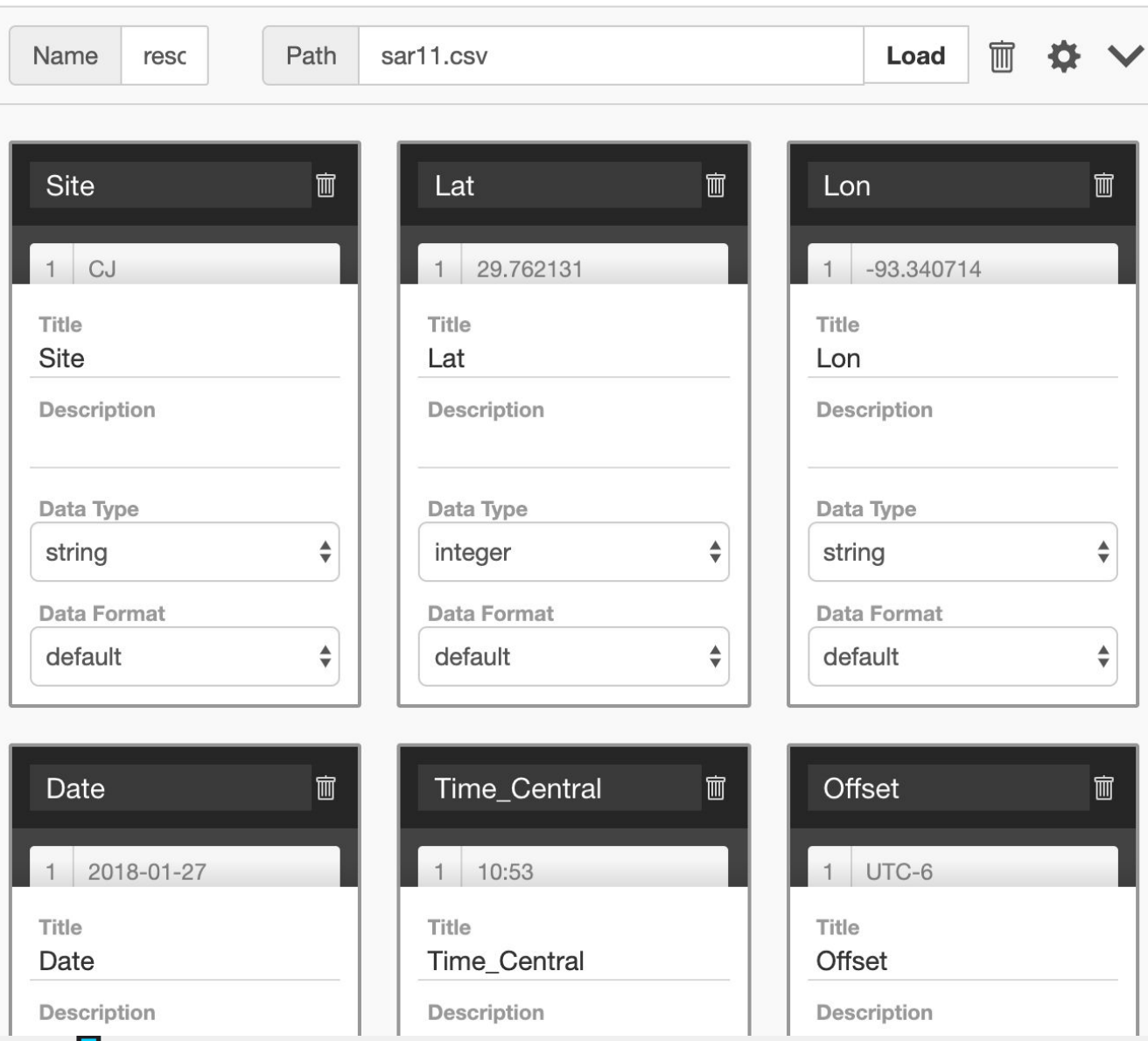
raw data

https://www.bco-dmo.org/d
9

metadata

Temp	Temperature measured via YSI at the sampling site	degrees Celsius
Cond	Conductivity measured via YSI at the sampling site	ms/cm
Salinity	Calculated salinity based on the conductivity measurement from the YSI	unitless
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



Preview

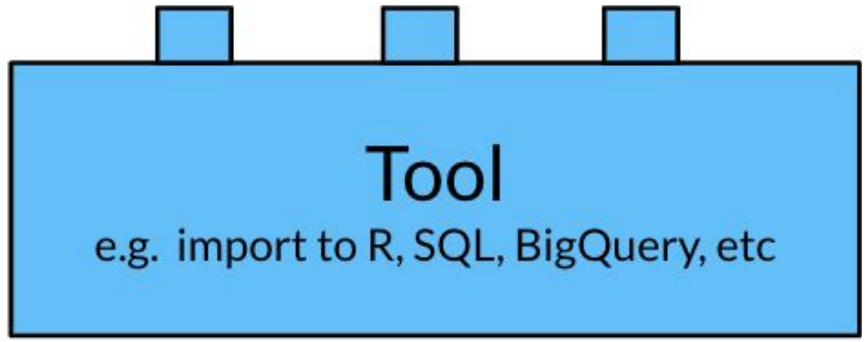

```
{
  "profile": "tabular-data-package",
  "resources": [
    {
      "name": "resource1",
      "path": "sar11.csv",
      "profile": "tabular-data-resou
      "schema": {
        "fields": [
          {
            "name": "Site",
            "type": "string",
            "format": "default",
            "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"
```

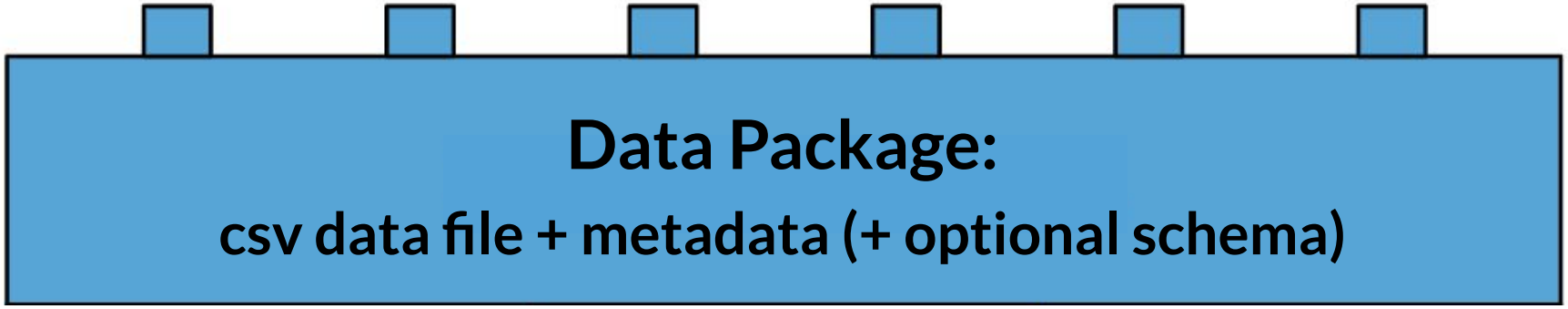

Open Knowledge Foundation

Python library:
https://github.com/frictionlessdata/datapackage-py

https://www.bco-dmo.org/dataset/745449

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	3	Joel's Shoal	5/29/16 - 10/22/16
7			11/10/16 - 3/22/17
8			3/28/16 - 7/11/17

pipeline-spec.yaml

```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


```json
{
  "bytes": 24061,
  "count_of_rows": 433,
  "hash": "c3aaa307223086fa611c40f9ab8ae100",
  "name": "_",
  "resources": [
    {
      "bytes": 24061,
      "count_of_rows": 433,
      "dialect": {
        "delimiter": ",",
        "doubleQuote": true,
        "lineTerminator": "\r\n",
        "quoteChar": "\"",
        "skipInitialSpace": false
      },
      "dpp:streamedFrom": "http://datadocs.bco-dmo.org/docs/TestProject/data_
      "encoding": "utf-8",
      "format": "csv",
      "hash": "405e348a5bb172c191abbe8d5a72880b",
      "headers": 1,
      "name": "mcmurdo_epifauna",
      "path": "data/mcmurdo_epifauna.csv",
      "schema": {
        "fields": [
          {
            "decimalChar": ".",
            "groupChar": "",
            "name": "year",
```


data.csv


```csv
Core,Depth,pH,alkalinity,Nitrate,Chlorinity,Ca,B,
2014 bottom water,,7.92,2.32,21.1,544.9,10.17,413
J2-733-PC 1,2,7.68,2.08,22.3,546.2,9.69,524,<0.1,
J2-733-PC 1,6,7.69,2.11,23.8,546.2,9.64,535,0.4,<
J2-733-PC 1,10,7.71,2.2,25.1,545.2,9.59,533,<0.1,
J2-733-PC 1,13,7.7,2.22,25.8,547.2,9.62,531,0.2,<
J2-733-PC 1,16,7.69,2.22,24.7,544.6,9.67,529,0.5,
J2-733-PC 1,18,7.71,2.22,24.6,546.6,9.67,525,0.2,
J2-733-PC 2,2,7.69,2.13,21.9,548.5,9.72,528,0.3,<
J2-733-PC 2,7,7.7,2.17,24.6,543.9,9.65,536,<0.1,<
J2-733-PC 2,11,7.73,2.18,25.5,546.2,9.62,532,<0.1
J2-733-PC 2,15,7.7,2.16,26.1,544.2,9.6,530,0.3,<0
J2-733-PC 2,18,7.72,2.14,25.7,545.9,9.64,519,0.4,
J2-733-PC 2,20,7.7,2.16,25.4,546.3,9.62,527,0.2,<
J2-733-PC 2,22,7.72,2.16,25.2,,9.63,525,<0.1,<0.1
J2-733-PC 4,3,7.67,2.05,23,547.3,9.7,521,<0.1,<0.
J2-733-PC 4,6,7.69,2.1,23.8,545.1,,516,0.2,<0.1,9
J2-733-PC 4,8,7.71,2.11,24.5,544,9.66,516,0.2,<0.
J2-733-PC 4,10,7.75,2.13,25.1,544.2,9.64,517,0.2,
J2-733-PC 4,12,7.71,2.13,25.3,544,9.61,514,0.1,<0
```


Open Knowledge Foundation

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

@lilscientista

3 Biology Frictionless Data Use Cases

Data Packages
+ Pipelines

BCO-DMO

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

Open Knowledge
Foundation

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"
},
```

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

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

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 th values will be ignored.

The full list of error messages:

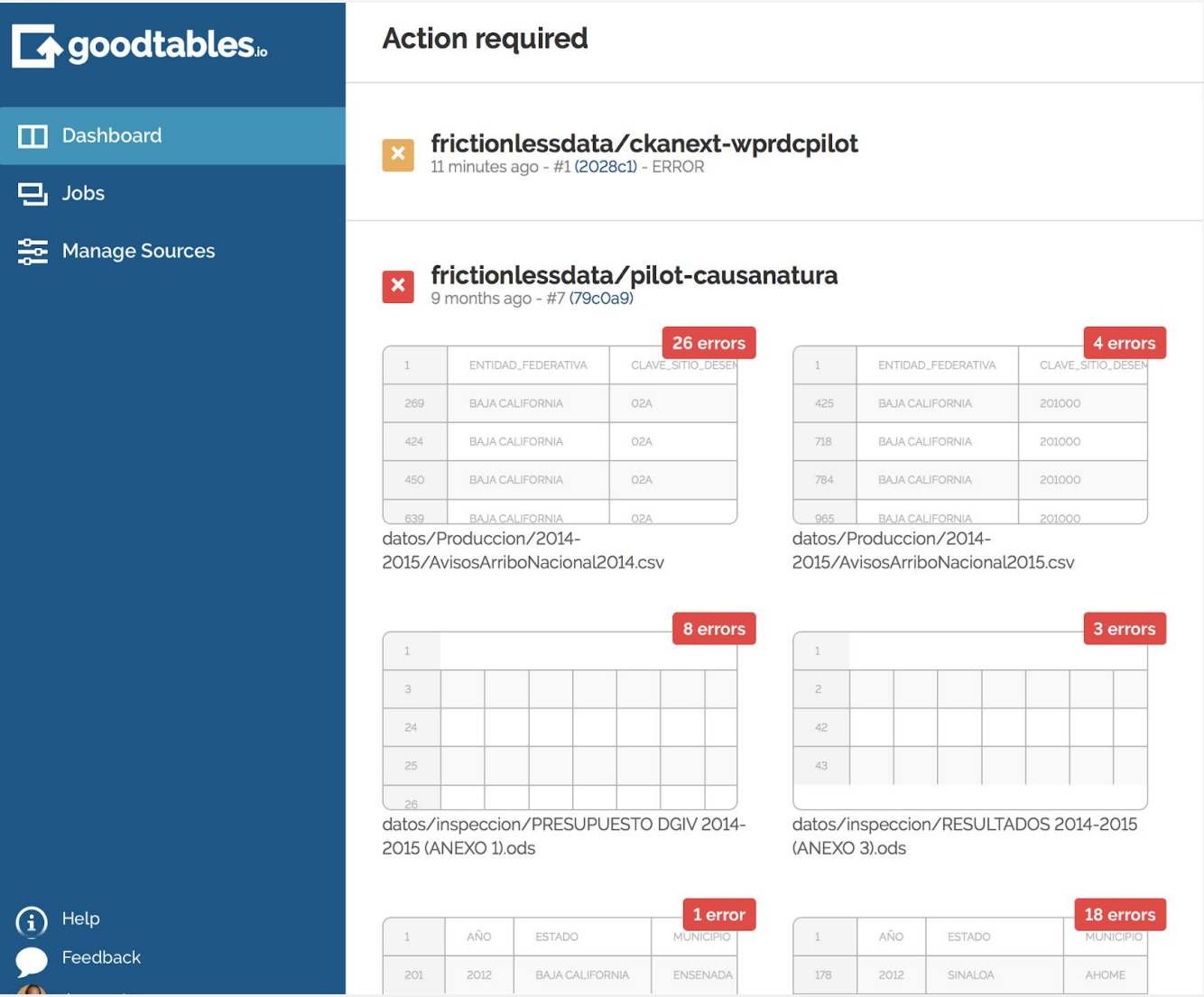
```
The value "under 35" in row 2 and column 5 is not type "number" and format "default"
```

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

Open Knowledge Foundation

try.goodtables.io
https://github.com/frictionlessdata/goodtables-py
https://github.com/frictionlessdata/tableschema-py

Continuous data validation: http://goodtables.io



Open Knowledge Foundation

3 Biology Frictionless Data Use Cases

Data Packages
+ Pipelines

BCO-DMO

Goodtables +
Table Schema

Lily Zhao,
Graduate
student
UC Santa
Barbara

{ } Frictionless
Specifications

Dr. Philippe
Rocca-Serra

Open Knowledge
Foundation

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

Supplementary Data 3

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
ionosing 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 databasee, NIST online database, purified standard and bibliography
Results in microg/g fresh weight of petals
For Rosa chinensis 'Old Blush', sepals ans stamens were analyzed in addition to petals.

Table of actual data

	R. chinensis 'Old Blush' sepals		*R. chinensis* 'Old Blush' stamen
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			

isatools
isa-tools.org

Open Knowledge Foundation

Philippe Rocca-Serra, PhD [0000-0001-9853-5668]
Data Readiness Group, Oxford e-Research Centre

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

```
{
    "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}
```

stato
statistics ontology

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

ata_integration.csv",

l compound",

y is meant to hold inform

197",

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

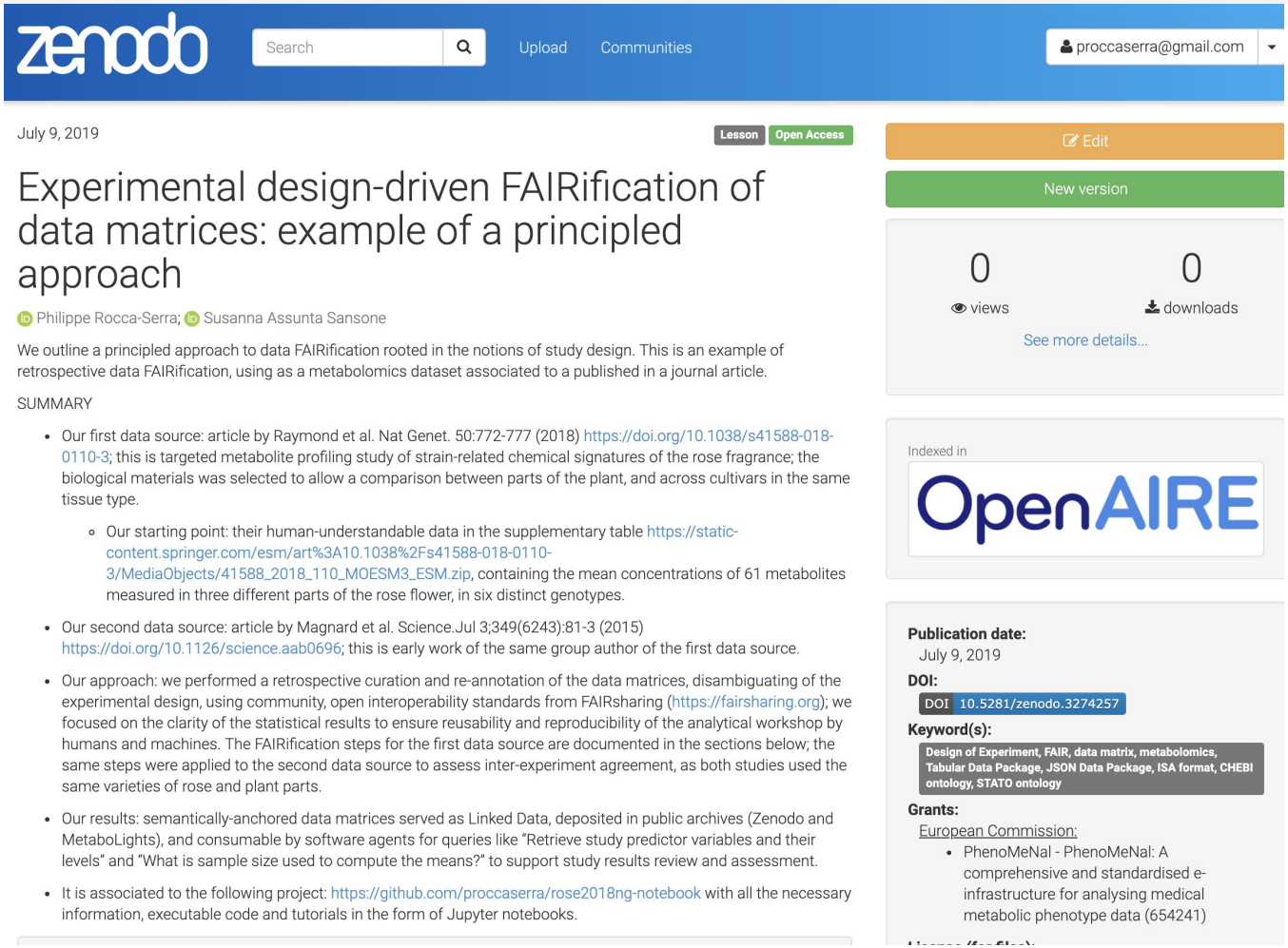
https://github.com/proc
aster/data/processed/r
018-treatment-group-m

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

Open Knowledge Foundation

Philippe Rocca-Serra, PhD [0000-0001-9853-5668]
Data Readiness Group, Oxford e-Research Centre

Publish Rose DataPackages on Zenodo



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

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

Open Knowledge Foundation

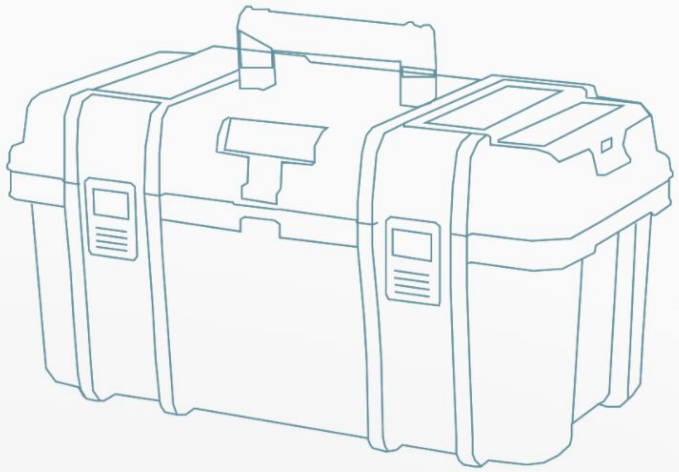
Philippe Rocca-Serra, PhD [0000-0001-9853-5668]
Data Readiness Group, Oxford e-Research Centre

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

DATOPIAN

http://github.com/frictionlessdata/

https://discord.com/invite/j9DNFNw

youtube.com/user/openknowledgefdn

https://frictionlessdata.io/guide/

Twitter: @frictionlessd8a

Thank you!

Join our community!

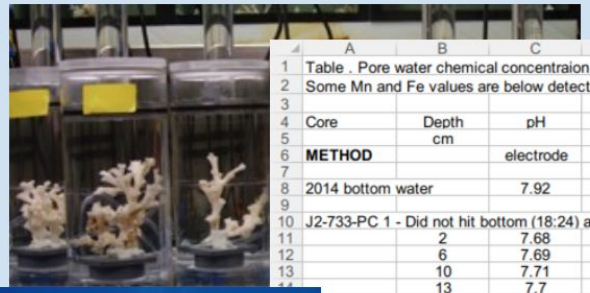

Did you record the
metadata?

Open Knowledge Foundation

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

@lilscientista

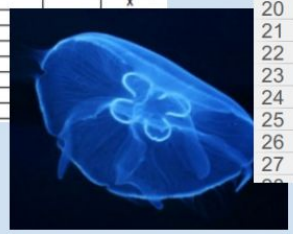
BCO-DMO data is MESSY



	A	B	C	D
1	Table . Pore water chemical concentraion data and lo			
2	Some Mn and Fe values are below detection and are			
3				
4	Core	Depth	pH	alkalinity
5		cm		mmol/kg
6	METHOD		electrode	titration
7				
8	2014 bottom water		7.92	2.32
9				
10	J2-733-PC 1 - Did not hit bottom (18:24) and was pos			
11		2	7.68	2.08
12		6	7.69	2.11
13		10	7.71	2.20
		13	7.7	2.22
		16	7.69	2.22
		18	7.71	2.22
	-733-PC 2 - Did not hit bottom (18:27) and was positioned next to PC 1, 22 d			
		2	7.69	2.13
		7	7.70	
		11	7.73	
		15	7.70	
		18	7.72	
		20	7.70	
		22	7.72	
	-733-PC 4 - Hit bottom (18:01) and wa			
		3	7.67	
29		6	7.69	
30		8	7.71	
			7.75	
			7.71	
			7.73	
			7.71	
			7.71	

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

C	D	E	F	G	H
Video ID	Time (minute)	E	D	C.T.	C.O.
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	
	1.34.06				X
Clip016	10.06				X
	27.45				X
	32.22	X			
	50.13				X
Clip017	10.06				X
	27.2				
	25.45	X			
	30.24				
	36.02				
	39.15	XX			
	43.19				
	1.06.27				



“Friction” in BCO-DMO data

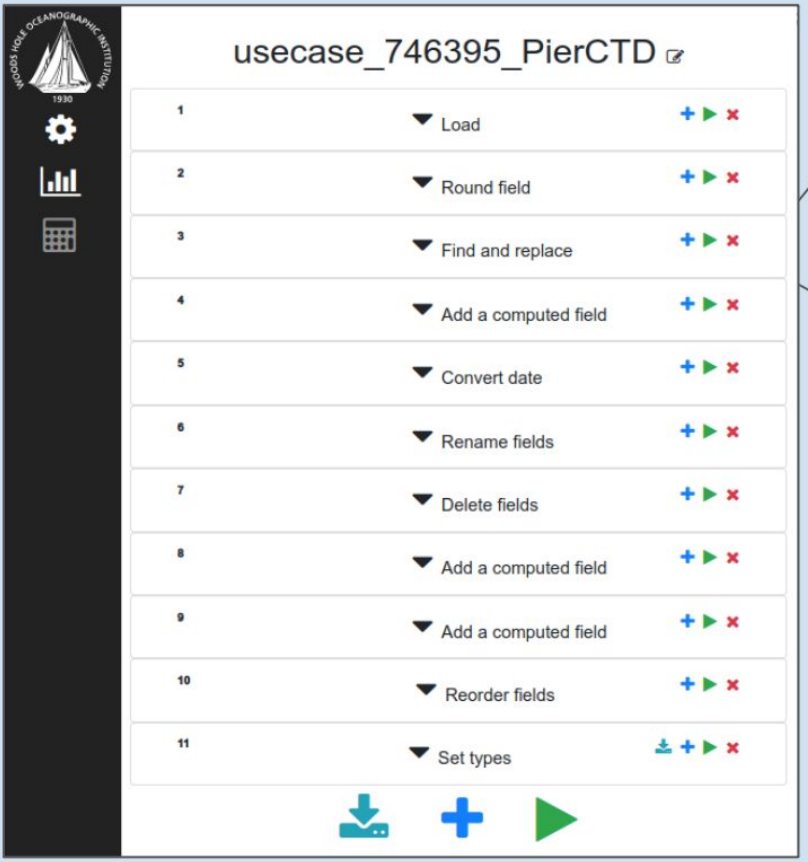
BCO-DMO data managers:

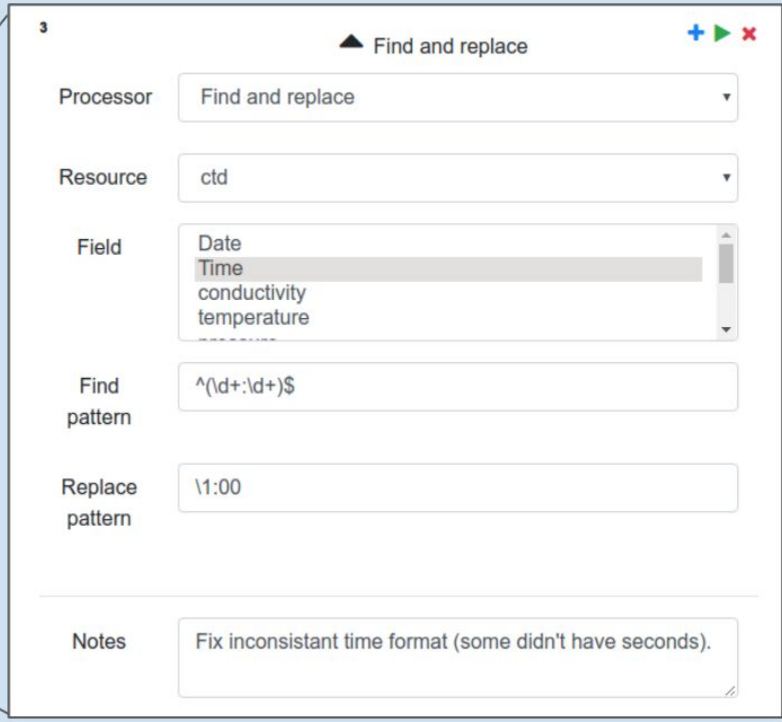
- Add spatio-temporal context in standard formats
  - date/time (ISO 8601), timezones
  - lat, lon
  - depth
- Correct quality issues
  - Inconsistent formatting
  - corrupt data characters
  - data gaps
  - invalid species names
  - typos
- Reformat for reusability

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





pipeline-spec.yaml

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

Open Knowledge Foundation

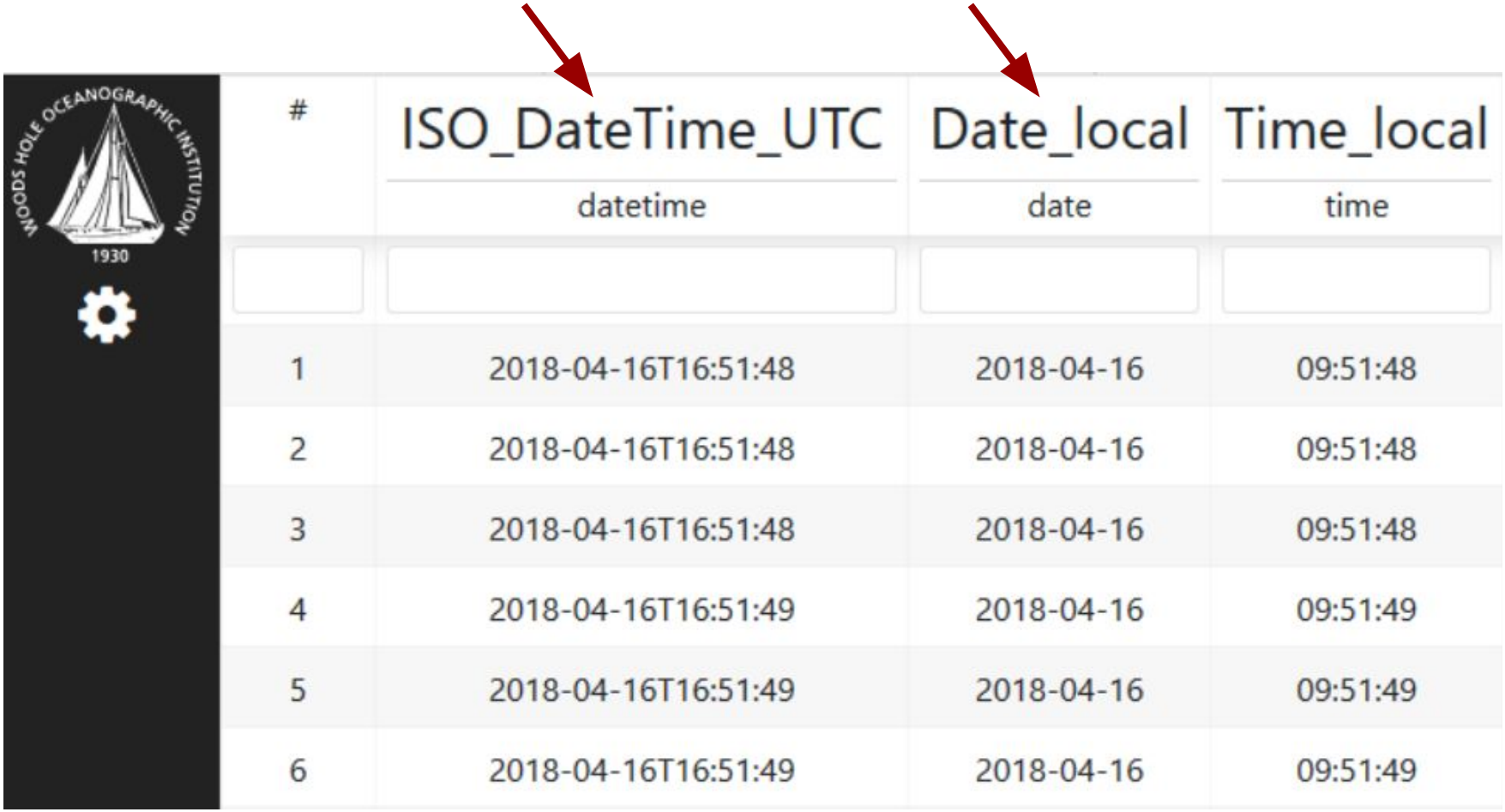
Slide modified from Amber York, BCO-DMO
https://zenodo.org/record/2687557

@lilscientista

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:

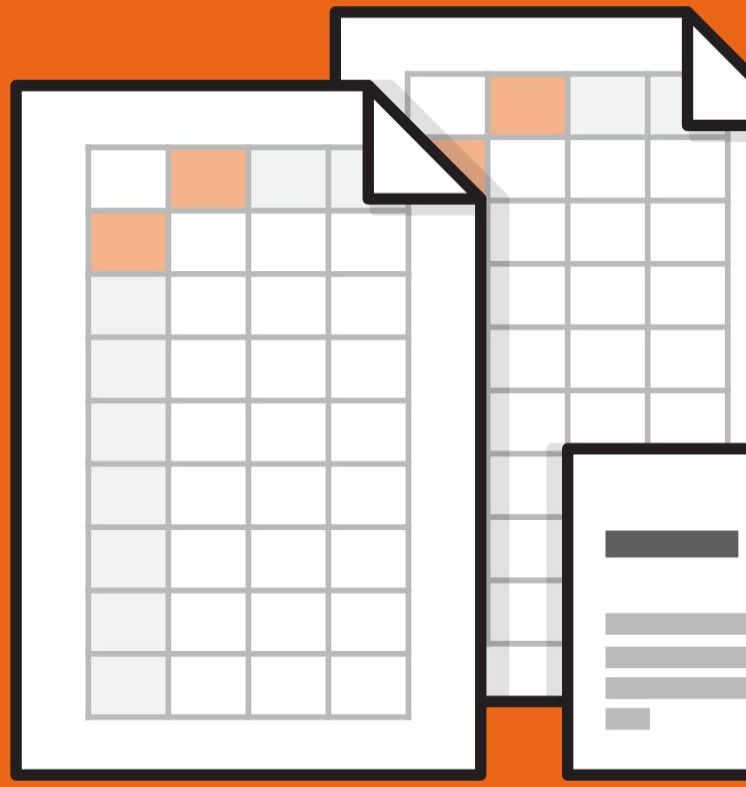
https://bco-dmo.org; @BCO-DMO

https://github.com/BCODMO

FRICTIONLESS DATA

Field guide

Data Package Creator



Create.frictionlessdata.io

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

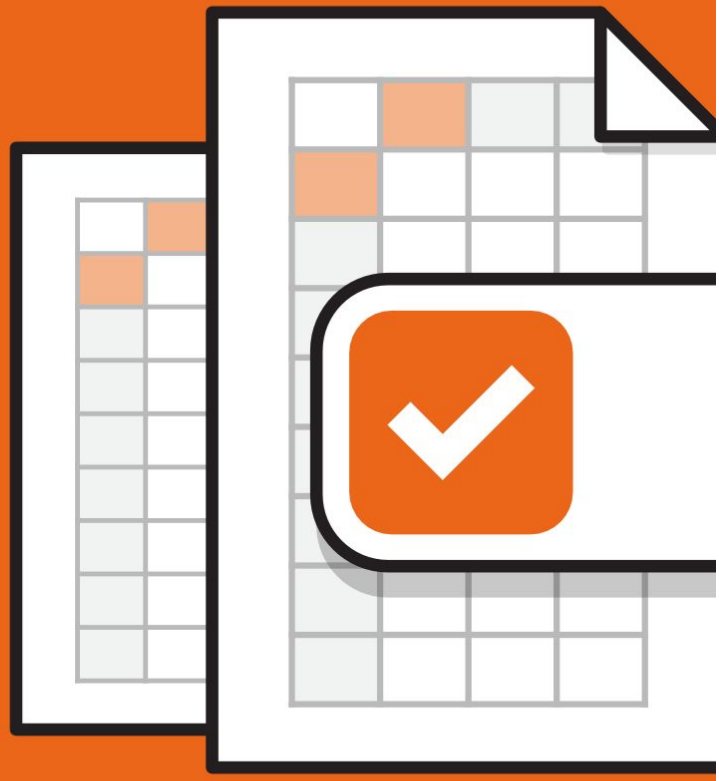
Open Knowledge Foundation

@lilscientista

FRICTIONLESS DATA

Field guide

Validating data with try.goodtables



Try.goodtables.io

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

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

Open Knowledge Foundation

@lilscientista

Continuous data validation: http://goodtables.io