Frictionless Data for Reproducible Research

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

FOSDEM - Feb 2020

http://bit.ly/FDfosdem

Science has a reproducibility 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.

Open up all essential, public interest information and see it used to create insight that drives positive change

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

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

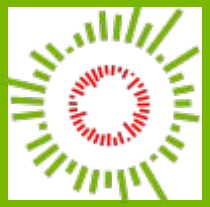
Frictionless Data for Reproducible Research

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

What is Frictionless Data?

Data cleaning

What's the license for this data?

What does this data value mean?

Checking data quality

Can I use this Excel data in Python?

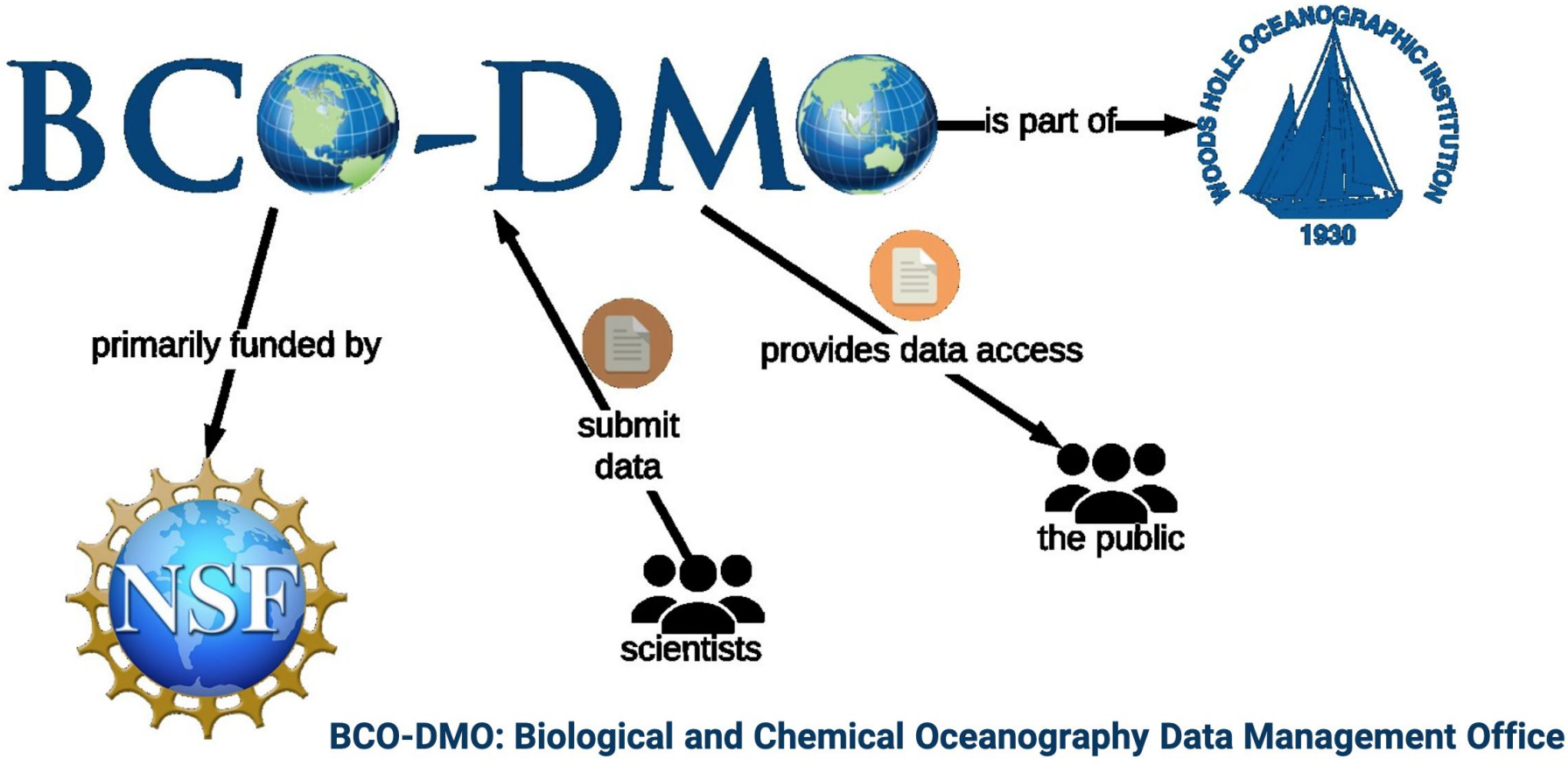
Who created this data?

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

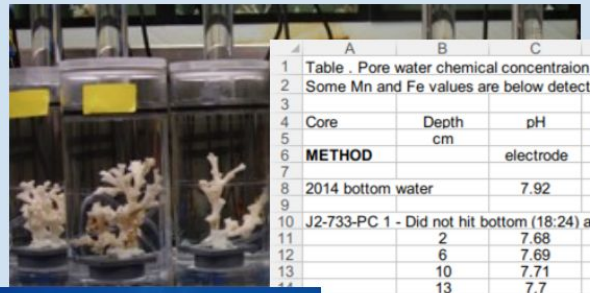
How can researchers (and any data wrangler) use Frictionless Data?

Pilot: BCO-DMO



BCO-DMO Pilot team: Amber York, Conrad Schloer, Adam Shepherd, Danie Kinkade

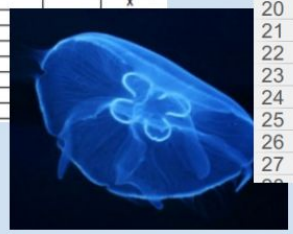
BCO-DMO data is MESSY



	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			3/28/17 - 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				



OPEN KNOWLEDGE

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

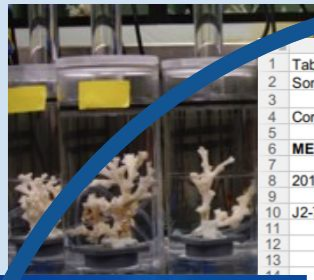
@lilscientista

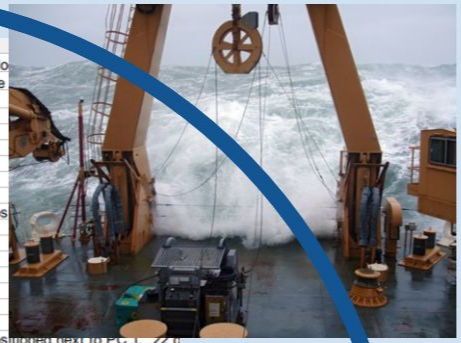
DPP helps BCO-DMO data managers

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

Messy data → clean data → hosted data



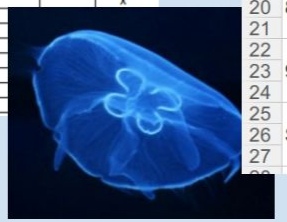


	A	B	C	D
1	Table . Pore water chemical concentraion data and lo			
2	Some Mn and Fe values are below detection and are			
3				
4	Core	Depth	pH	alkalinity
5		cm		mmol/kg
6	METHOD		electrode	titration
7				
8	2014 bottom water		7.92	2.32
9				
10	J2-733-PC 1 - Did not hit bottom (18:24) and was pos			
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

					2	7.69	2.13
					7	7.70	
					11	7.73	
					15	7.70	
					18	7.72	
					20	7.70	
					22	7.72	
					3	7.67	
					6	7.69	
					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			
16			
17	7	Ya	
18			
19			
20	8	Bo	
21			
22			
23	9	Ra	
24			
25			
26	S	Re	
27			

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				



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,
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,<
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
```

Dataset landing page

OPEN KNOWLEDGE

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

@lilscientista


Did you record the
metadata?

Keep track of your metadata

Figure 2	Genotype	Mean	Std. Error	SEM	N (animals/hemisegment)	p-value
	DIP-α-GAL4/+>EGFP	0.767	0.427	0.0551	12/60	n/a
	DIP-α-GAL4>EGFP	0	0	0	10/30	<0.0001

raw data

metadata
In JSON

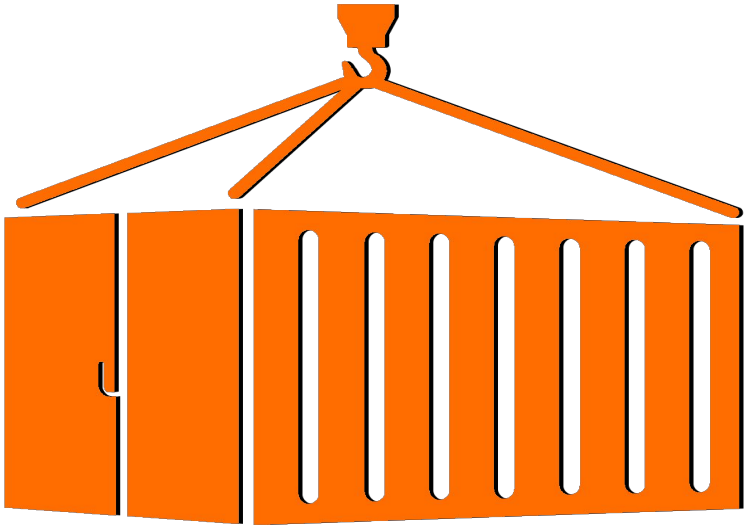

```
"resources": [
    {
        "mediatype": "text/csv",
        "name": "elife",
        "profile": "tabular-data-resource",
        "schema": {
            "missingValues": [
                ""
            ],
            "fields": [
                {
                    "type": "string",
                    "format": "default",
                    "name": "Genotype"
                },
                {
                    "type": "number",
                    "format": "default",
                    "name": "Mean"
                },
                {
                    "type": "number",
                    "format": "default",
                    "name": "Std. Error"
```


create.frictionless
data.io

Package your data with its metadata

Data Package

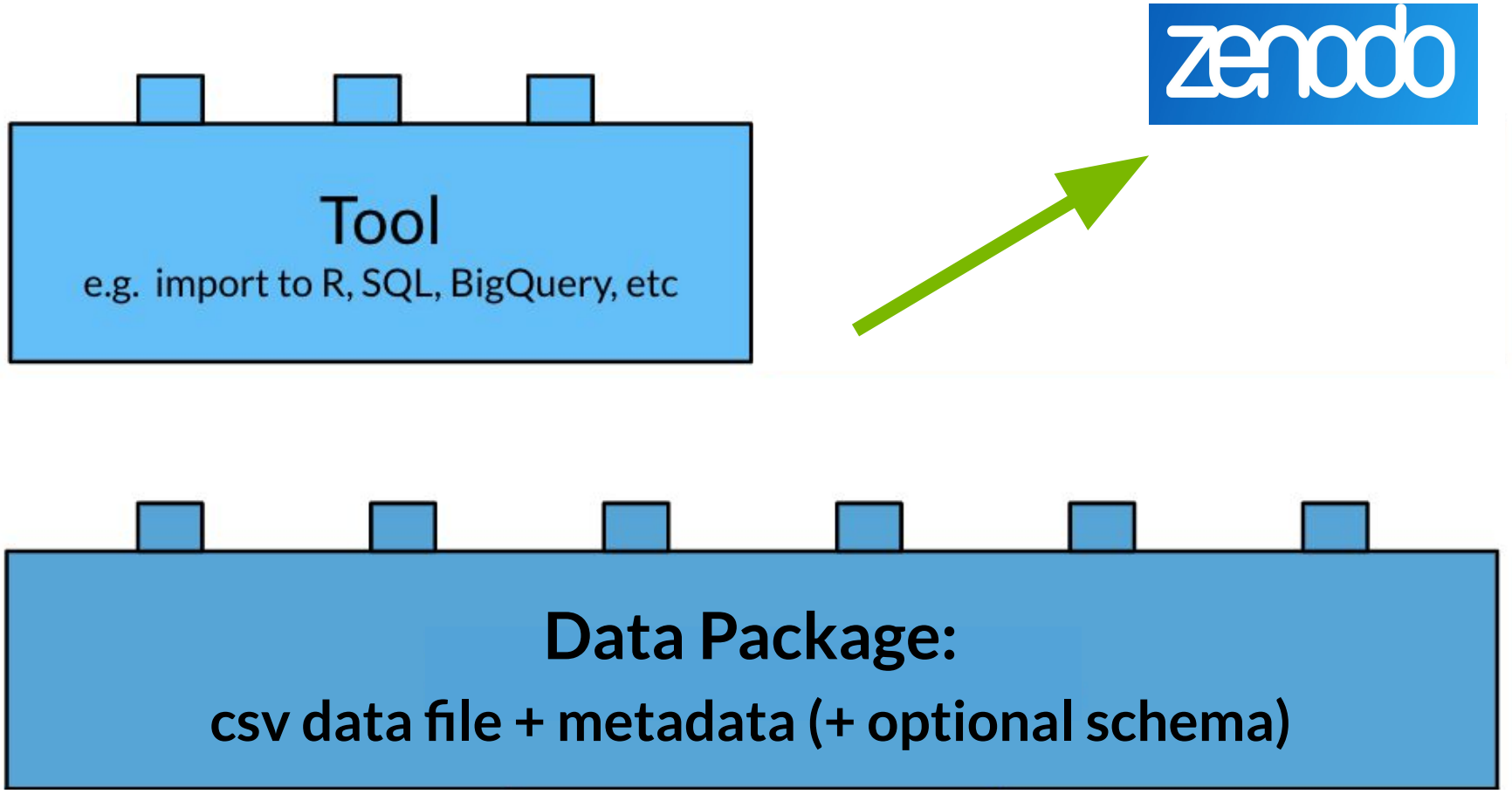
Data
e.g.,
experiment.csv
+
Metadata
(+ optional
schema)



Datapackage
software libraries

create.frictionless
data.io

Packaged data is useful data



Create a schema + validate your data

Correspondence | Open Access | Published: 23 June 2004

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Lineh
Barrett & John N Weinstein

BMC Bioinformatics **5**, Article number: 80 (2004) | Cite this article

```
$ goodtables data/color_codes.csv

DATASET
=======
{'error-count': 0,
 'preset': 'nested',
 'table-count': 1,
 'time': 0.107,
 'valid': True}
```

try.goodtables.io

Goodtables software libraries

Tableschema software libraries

Use FD in your data pipeline

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

Open source software for Frictionless Data

tableschema-js
A library for working with Table Schema in Javascript.

Repository

goodtables-py
Validate and process tabular data in Python.

Repository

datapackage-py
A Python library for working with Data Packages.

Repository

datapackage-js
Official JavaScript library for Data Packages in Node and the browser.

Repository

tableschema-py
A Python library for working with Table Schema.

Repository

tabulator-py
Consistent interface for stream reading and writing tabular data (csv/xls/json/etc).

Repository

datapackage-java
A Java library for working with Data Packages.

Repository

datapackage-r
An R library for working with Data Packages.

Repository

tableschema-java
A Java library for working with Table Schema.

Repository

tableschema-pandas-py
Table Schema to Pandas module for

tableschema-clj
A Clojure library for working with Table

datapackage-go

OPEN KNOWLEDGE

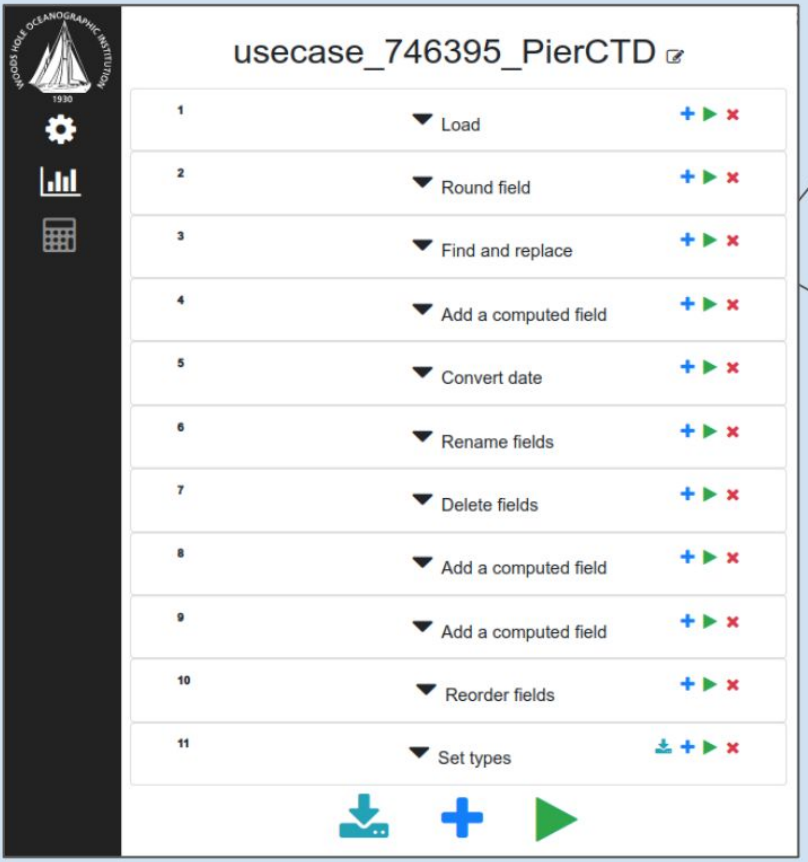
https://frictionlessdata.io/software/
https://github.com/frictionlessdata

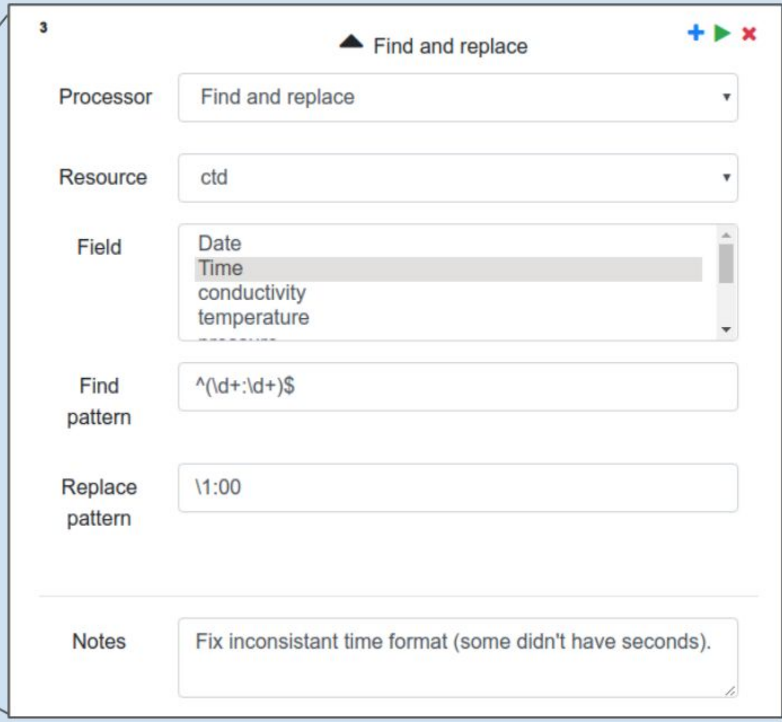

F
Did you record the
metadata?

How does Data Package Pipelines help?

- Give data managers a more immersive experience + improved UI
- Reduce dataset processing time
- Remove barrier of programming ability
- Avoid hand-writing pipeline-spec.yaml or python scripts
  - eliminate syntax issues
  - reduce repetitive tasks
- Add custom metadata to pipeline
- Add capabilities not currently in datapackage-pipeline with custom processors

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

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

Pipeline-spec.yaml + data.csv + datapackage.json

Datapackage pipelines produce reproducible research

	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

```
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

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

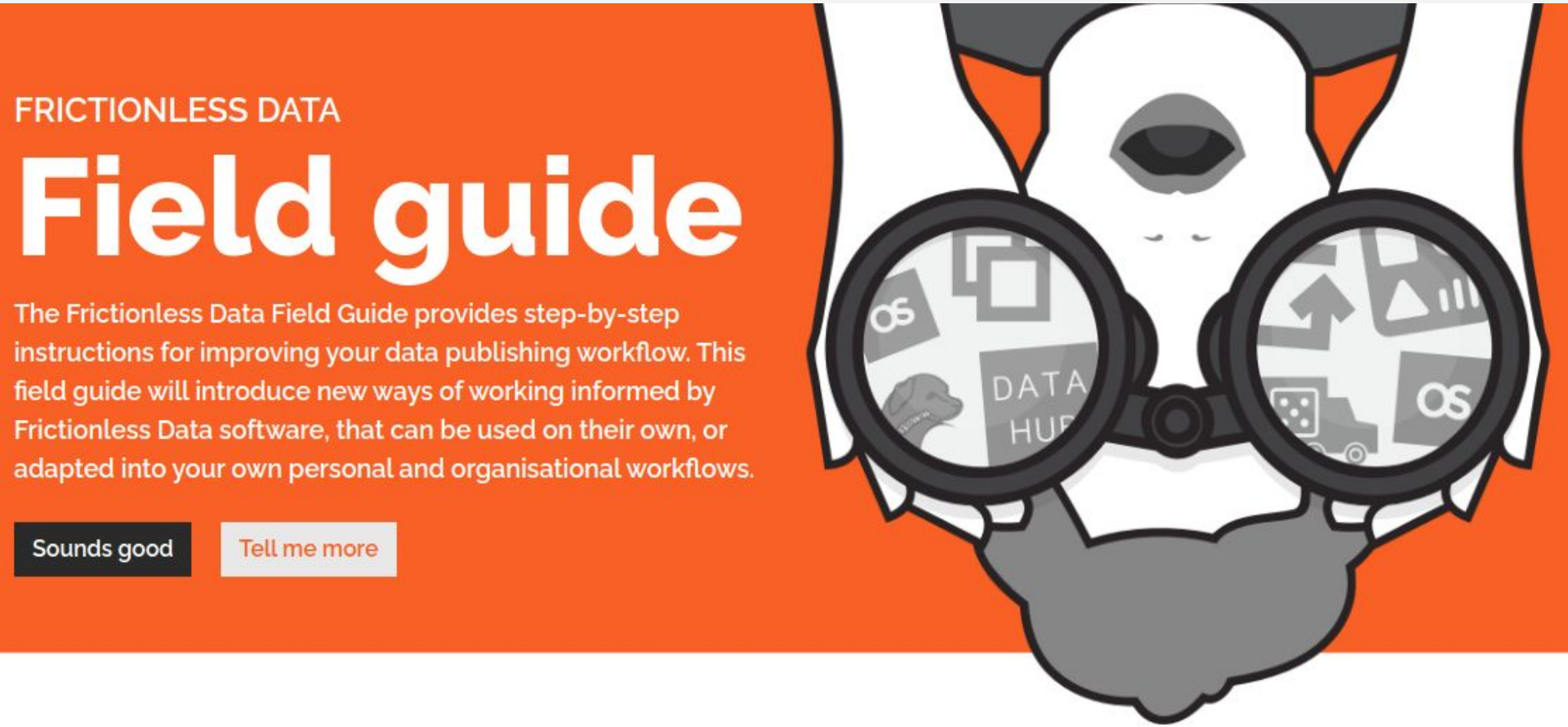
@lilscientista

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

Want to try it yourself?

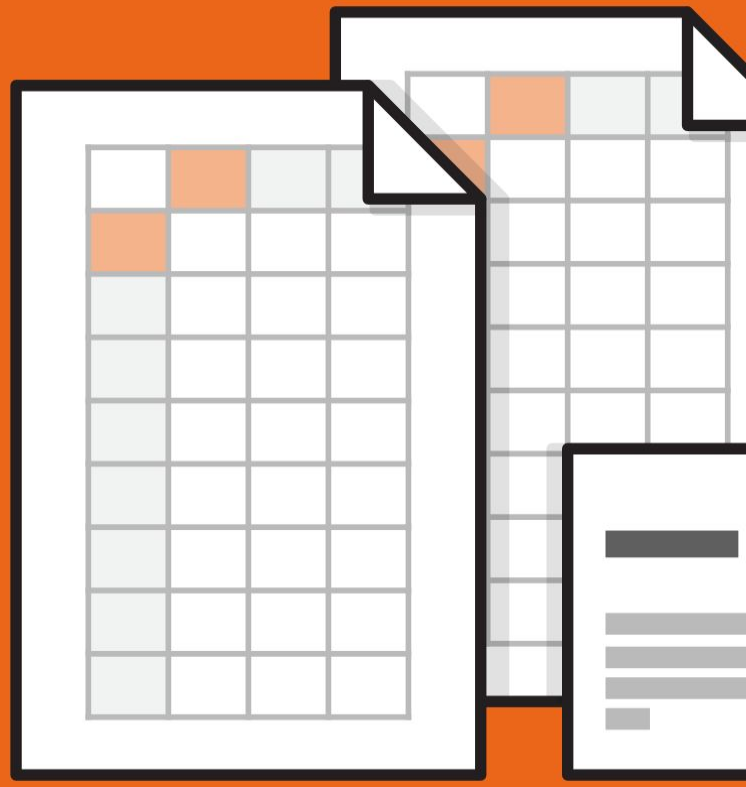


frictionlessdata.io/field-guide/

FRICTIONLESS DATA

Field guide

Data Package Creator



Create.frictionlessdata.io

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

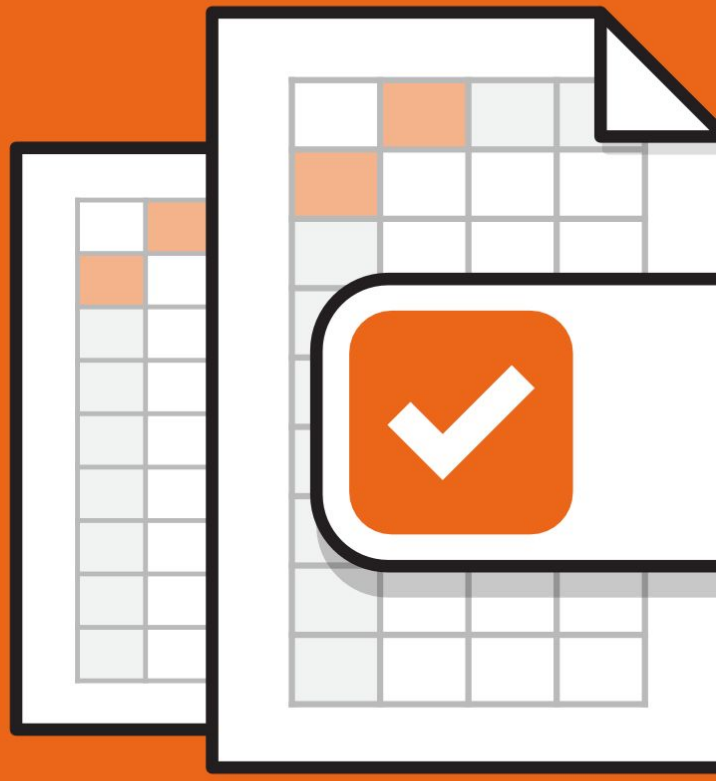
OPEN KNOWLEDGE

@lilscientista

FRICTIONLESS DATA

Field guide

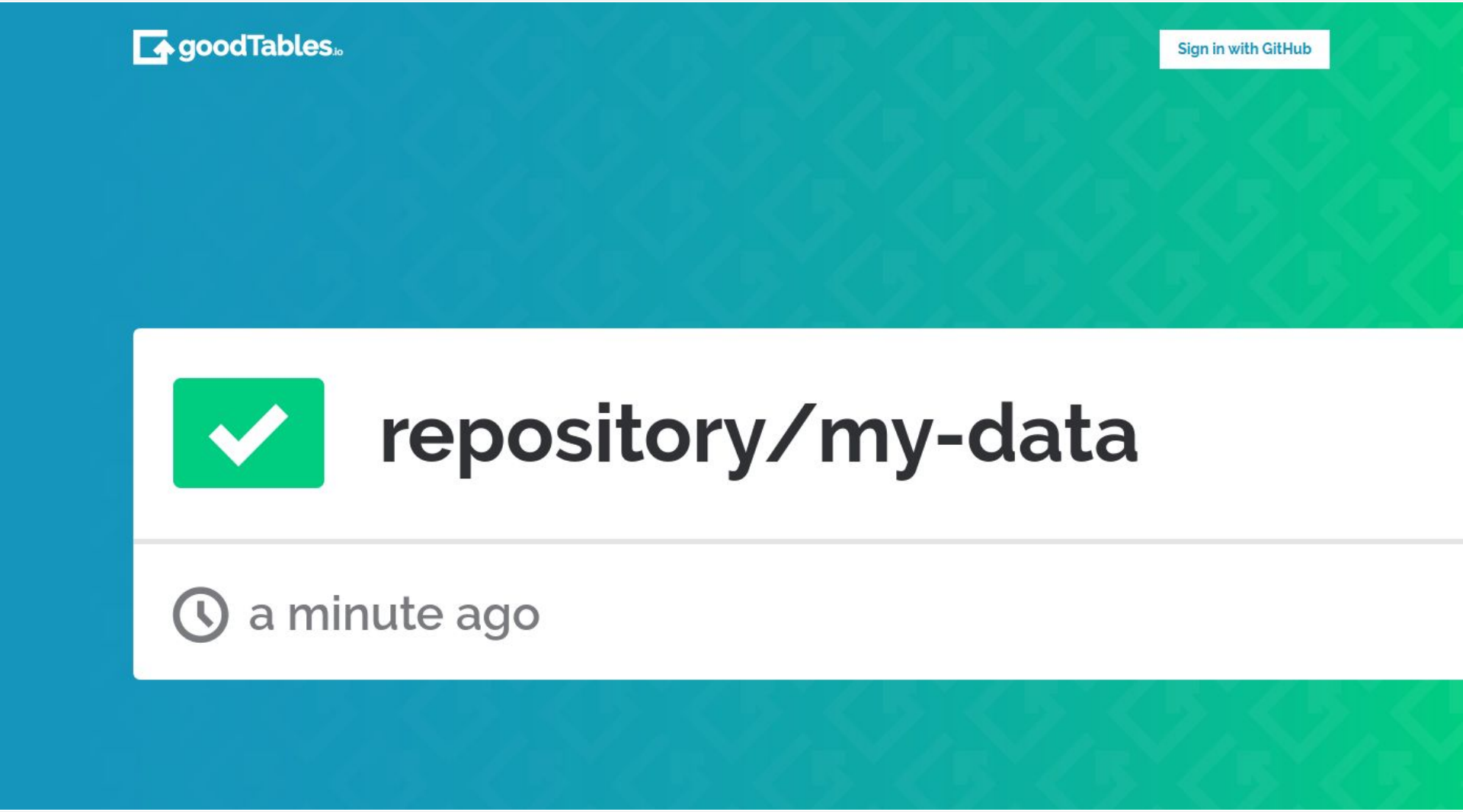
Validating data with try.goodtables

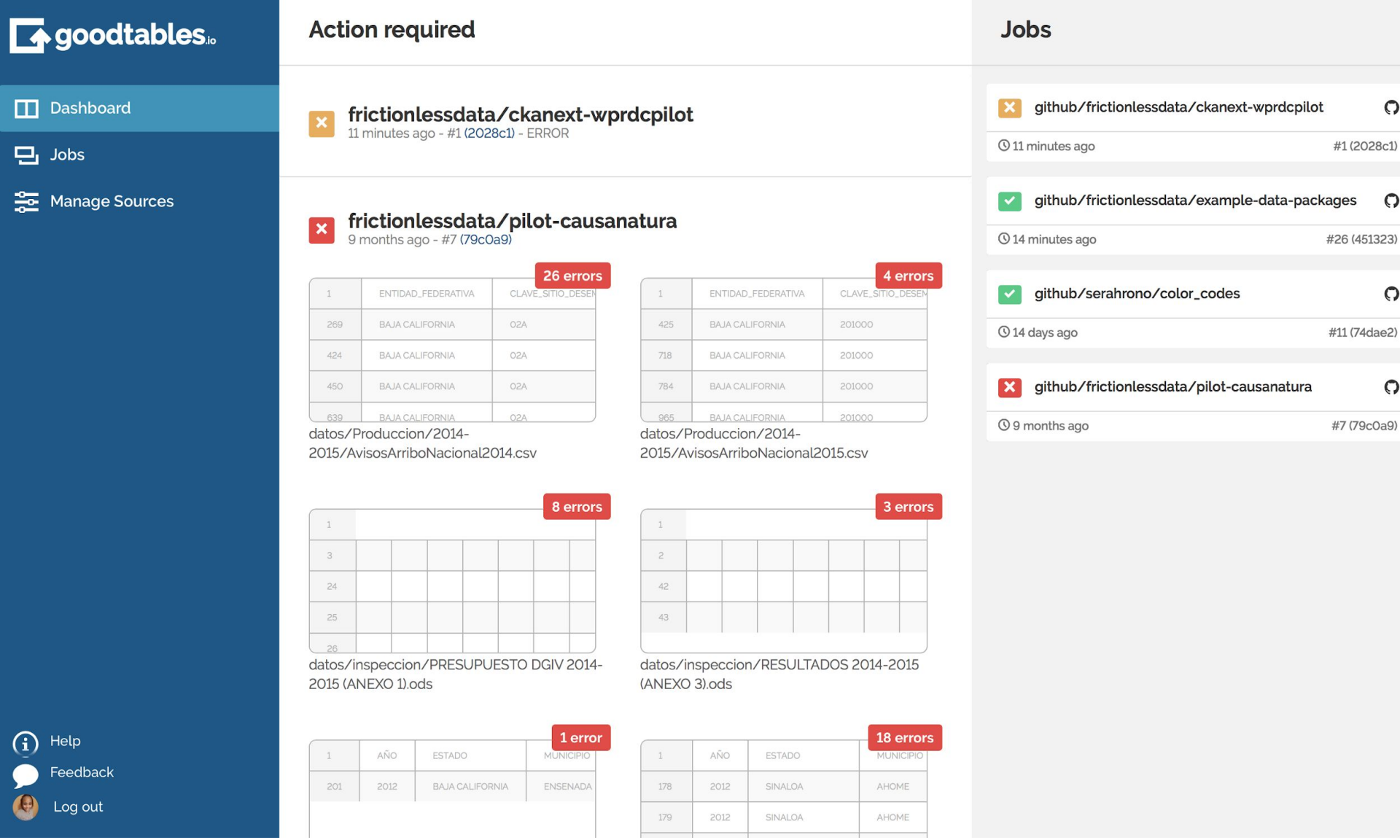


Try.goodtables.io

https://bit.ly/2P6oogW Or lots of example data:
https://github.com/frictionlessdata/goodtables-py/tree/master/data

Continuous data validation: http://goodtables.io





http://goodtables.io/

OPEN KNOWLEDGE

@lilscientista

OPEN KNOWLEDGE FOUNDATION

http://bit.ly/FDfosdem

http://github.com/frictionlessdata/

discuss.okfn.org/c/frictionless-data

gitter.im/frictionlessdata/chat

youtube.com/user/openknowledgefdn

https://frictionlessdata.io/field-guide/

Twitter: @okfn . @okfnlabs

Thank you!

Join our community!