VIRTUAL FRICTIONLESS DATA WORKSHOP

Link to Slides:

Frictionless Fellows
Frictionlessdata.io

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.

@okfn https://okfn.org @lilscientista

Icebreaker!

If you could spend one day with any cartoon character, who would it be?

Introduction to the programme

- What is the Frictionless Data Project?
  - Open source
  - Open and reproducible research
  - Data science
- What is the Frictionless Data Fellowship?
  - Tools & specifications
  - Coding, writing, capacity building, discussions

Introduction of fellows

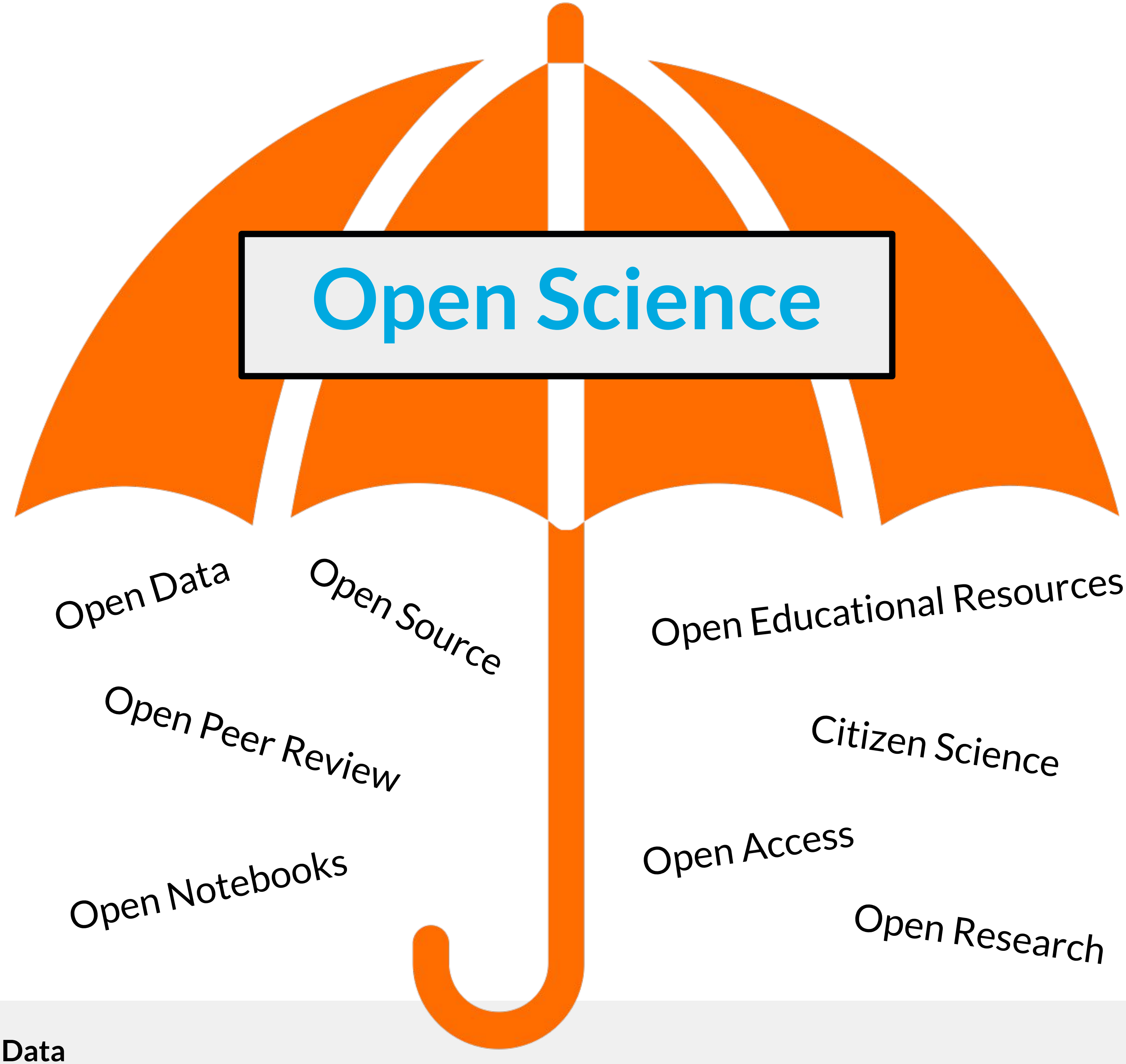
- Lindsay Gypin: FD Fellow | she/her | Data Librarian | USA | @menacegypin
- Kevin Kidambasi: FD Fellow | He/Him | Biochemist | *icipe*/Kenya
- Melvin: FD Fellow | she/her| Soil scientist | APNI | Kenya
- Guo-Qiang Zhang: FD Fellow | he/him | PhD Student | Sweden

Objectives of the workshop

- Introduce frictionless data tools
  - Data Package Creator
  - Goodtables
- Communicate the importance of good data practices and metadata

Expected Outcomes

- Be able to create a Data Package using the web app
- Be able to create a schema file and use it to validate a data set using trygoodtables.io
- Be able to handle common errors in FD workflows



Open Science

Open Data

Open Source

Open Educational Resources

Open Peer Review

Citizen Science

Open Notebooks

Open Access

Open Research

Open Science

- Transparent and inclusive scientific process
- Supports reproducibility *(Reproducible Research)*
- Science belongs to everyone

FAIR Data Principles

FINDABLE

ACCESSIBLE

INTEROPERABLE

REUSABLE

The Dataset we are using today

- Vector-borne diseases poses a serious global health burden

- Over 17% of the global tropical infectious disease burden is vector-borne e.g Trypanosomiasis, Malaria, Leishmaniasis

- Accounts for about 1 million human deaths annually

Some of the vectors of diseases

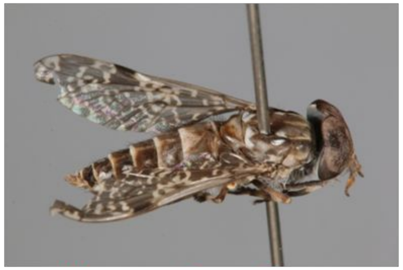
Tabanus

Chrysops

Stomoxys calcitrans

Hippobosca

Tsetse fly



Haematopota

Mosquito

Tick

Sand fly

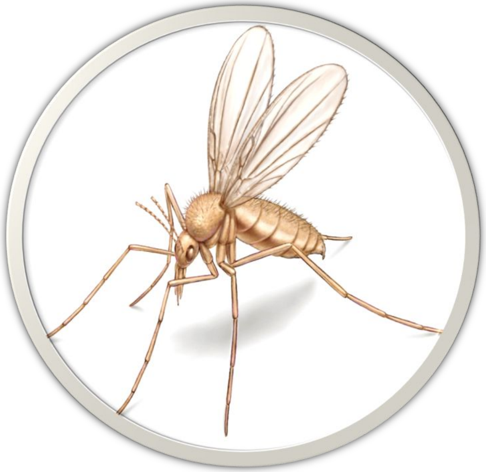
Transmission of Leishmaniasis by Sand flies

- Leishmaniasis is transmitted by infected female sand fly during blood feeding

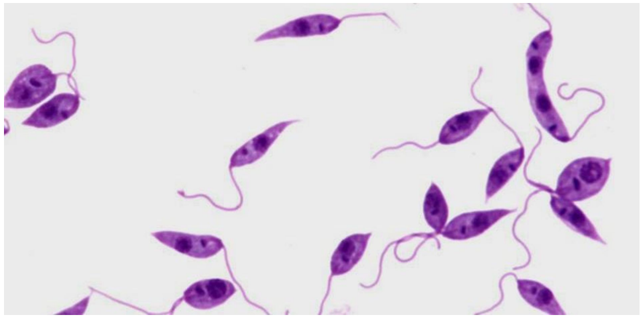
- Over 0.7 million leishmaniasis cases of infections are reported annually

- About 350 million people are at risk of contracting the disease across the 98 countries

Transmission of Leishmaniasis by Sand flies



Sandfly



Leishmania parasites

- Study carried out in the northern Kenya identified different sand fly spp. infected with *Leishmania* parasites

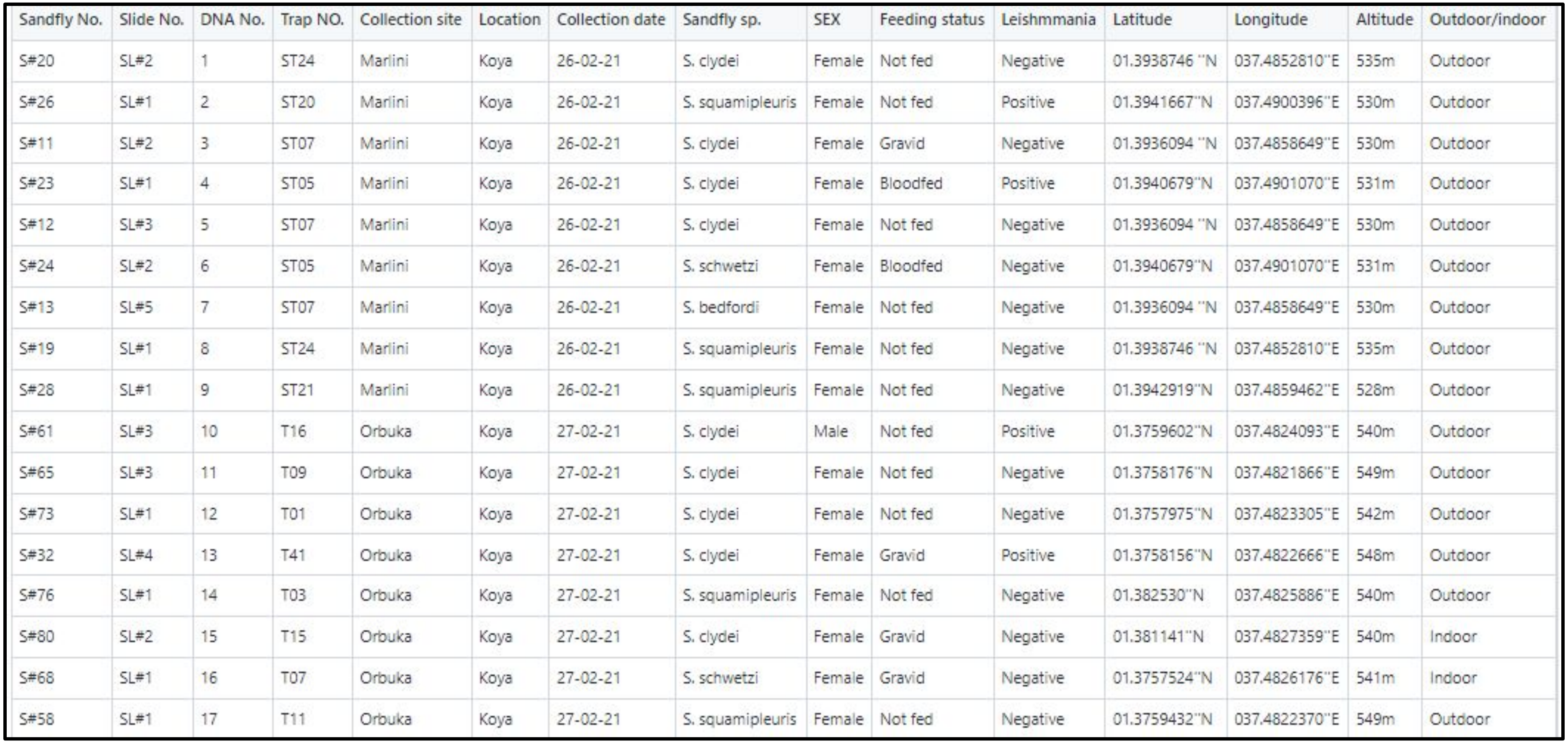
- The findings important in guiding targeted control programs

Metadata

- Metadata - information about the data
- Info like: author, date and site of collection, licences, data collection tools and keywords to describe the data
- Makes data easy to use and preserved in a format accessible and reusable by others
- My tabular data metadata - the date of sample collection, location and GPS coordinates, sample ID as well as sand fly sex and infection status as key metadata

Metadata

Sample dataset

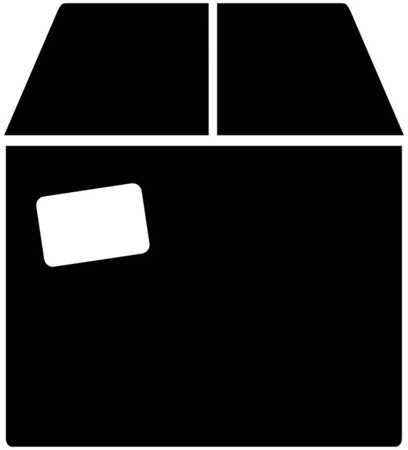


Sandfly No.	Slide No.	DNA No.	Trap NO.	Collection site	Location	Collection date	Sandfly sp.	SEX	Feeding status	Leishmmania	Latitude	Longitude	Altitude	Outdoor/indoor
S#20	SL#2	1	ST24	Marlini	Koya	26-02-21	S. clydei	Female	Not fed	Negative	01.3938746 "N	037.4852810"E	535m	Outdoor
S#26	SL#1	2	ST20	Marlini	Koya	26-02-21	S. squamipleuris	Female	Not fed	Positive	01.3941667"N	037.4900396"E	530m	Outdoor
S#11	SL#2	3	ST07	Marlini	Koya	26-02-21	S. clydei	Female	Gravid	Negative	01.3936094 "N	037.4858649"E	530m	Outdoor
S#23	SL#1	4	ST05	Marlini	Koya	26-02-21	S. clydei	Female	Bloodfed	Positive	01.3940679"N	037.4901070"E	531m	Outdoor
S#12	SL#3	5	ST07	Marlini	Koya	26-02-21	S. clydei	Female	Not fed	Negative	01.3936094 "N	037.4858649"E	530m	Outdoor
S#24	SL#2	6	ST05	Marlini	Koya	26-02-21	S. schwetzi	Female	Bloodfed	Negative	01.3940679"N	037.4901070"E	531m	Outdoor
S#13	SL#5	7	ST07	Marlini	Koya	26-02-21	S. bedfordi	Female	Not fed	Negative	01.3936094 "N	037.4858649"E	530m	Outdoor
S#19	SL#1	8	ST24	Marlini	Koya	26-02-21	S. squamipleuris	Female	Not fed	Negative	01.3938746 "N	037.4852810"E	535m	Outdoor
S#28	SL#1	9	ST21	Marlini	Koya	26-02-21	S. squamipleuris	Female	Not fed	Negative	01.3942919"N	037.4859462"E	528m	Outdoor
S#61	SL#3	10	T16	Orbuka	Koya	27-02-21	S. clydei	Male	Not fed	Positive	01.3759602"N	037.4824093"E	540m	Outdoor
S#65	SL#3	11	T09	Orbuka	Koya	27-02-21	S. clydei	Female	Not fed	Negative	01.3758176"N	037.4821866"E	549m	Outdoor
S#73	SL#1	12	T01	Orbuka	Koya	27-02-21	S. clydei	Female	Not fed	Negative	01.3757975"N	037.4823305"E	542m	Outdoor
S#32	SL#4	13	T41	Orbuka	Koya	27-02-21	S. clydei	Female	Gravid	Positive	01.3758156"N	037.4822666"E	548m	Outdoor
S#76	SL#1	14	T03	Orbuka	Koya	27-02-21	S. squamipleuris	Female	Not fed	Negative	01.382530"N	037.4825886"E	540m	Outdoor
S#80	SL#2	15	T15	Orbuka	Koya	27-02-21	S. clydei	Female	Gravid	Negative	01.381141"N	037.4827359"E	540m	Indoor
S#68	SL#1	16	T07	Orbuka	Koya	27-02-21	S. schwetzi	Female	Gravid	Negative	01.3757524"N	037.4826176"E	541m	Indoor
S#58	SL#1	17	T11	Orbuka	Koya	27-02-21	S. squamipleuris	Female	Not fed	Negative	01.3759432"N	037.4822370"E	549m	Outdoor

Introduction to Data Package

What is a data package?

- Blueprint for data structure
- Metadata: data about the data
- Schema: data organization



Introduction to Data

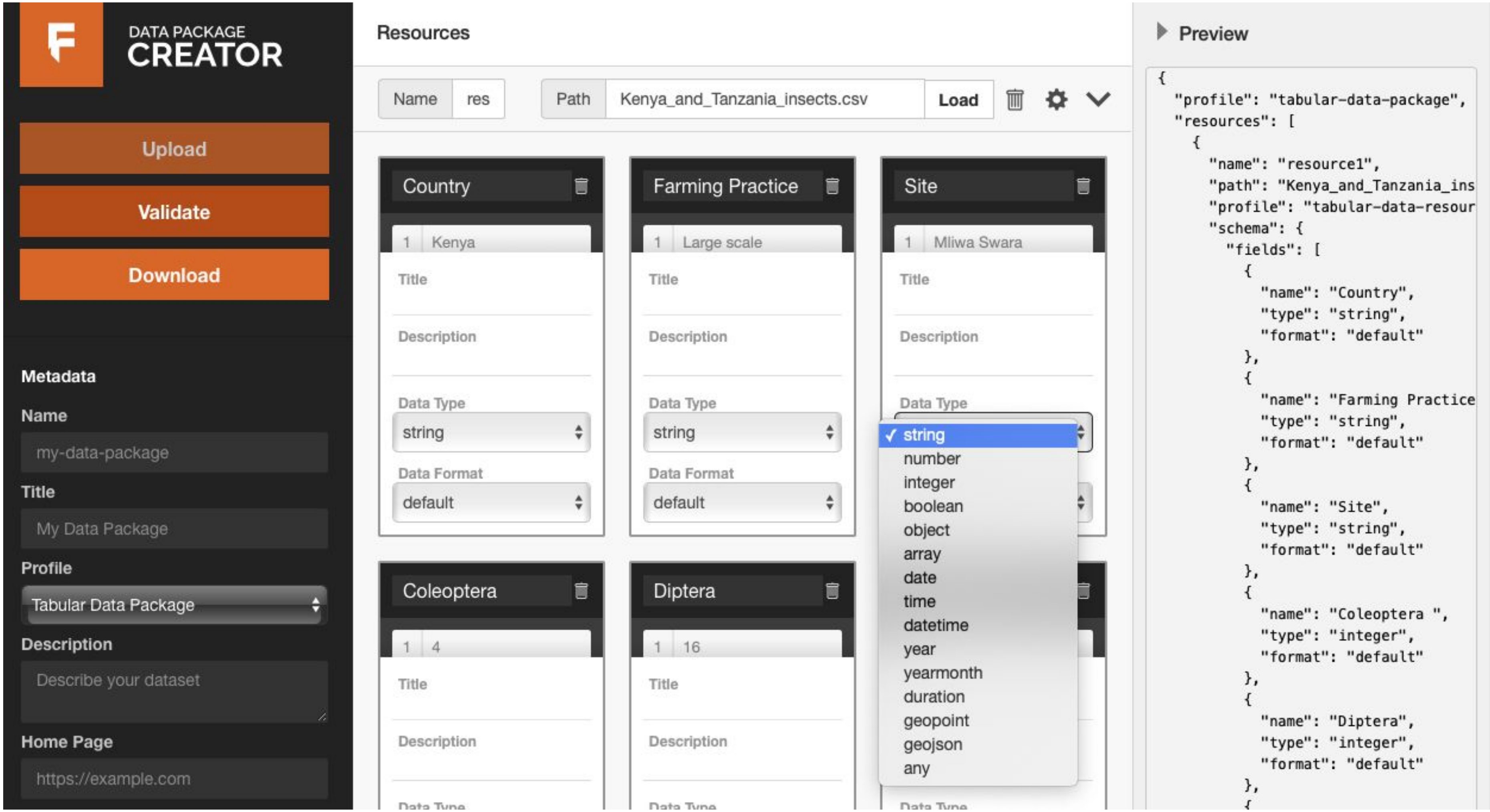
Why a data package?

- Transporting and reusing data
- Reduce friction in data workflows
- Reproducible research
- Validation

Introduction to Data Package ...

Creating a data package

- Web browser tool: Data Package Creator
- Advanced: programmatic interfaces available- CLI, Python, R etc
- Output: *.json (format)



Metadata

Name: my-data-package

Title: My Data Package

Profile: Tabular Data Package

Description: Describe your dataset

Home Page: https://example.com

Name: res

Path: Kenya_and_Tanzania_insects.csv

Country	Farming Practice	Site
1 Kenya	1 Large scale	1 Mliwa Swara
Title	Title	Title
Description	Description	Description
Data Type: string	Data Type: string	Data Type: ✓ string
Data Format: default	Data Format: default	

Coleoptera	Diptera
1 4	1 16
Title	Title
Description	Description
Data Type	Data Type

```
{
  "profile": "tabular-data-package",
  "resources": [
    {
      "name": "resource1",
      "path": "Kenya_and_Tanzania_ins
      "profile": "tabular-data-resour
      "schema": {
        "fields": [
          {
            "name": "Country",
            "type": "string",
            "format": "default"
          },
          {
            "name": "Farming Practice
            "type": "string",
            "format": "default"
          },
          {
            "name": "Site",
            "type": "string",
            "format": "default"
          },
          {
            "name": "Coleoptera ",
            "type": "integer",
            "format": "default"
          },
          {
            "name": "Diptera",
            "type": "integer",
            "format": "default"
          },
          {
```

DATA PACKAGE CREATOR

Hands On Time

https://create.frictionlessdata.io/

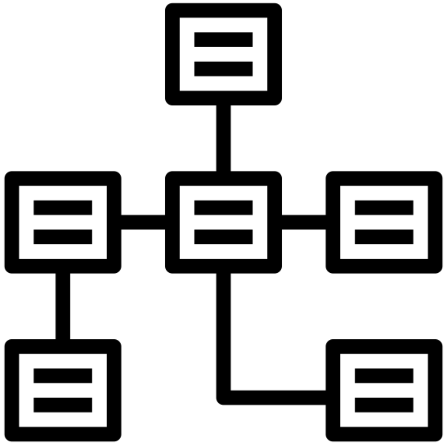
DATA

How is your data package useful?

- Precise data archiving format
- Easily shareable
- Data reproducibility
- Findable, Accessible, Interoperable, Reusable (FAIR)
- Data integrity validation

Questions

Introducing goodtables.io



The Frictionless datapackage helps make datasets easier to share across systems and file types.

But how do you check the quality of one data set?

What is Goodtables?

Goodtables is a free, open-source tool that helps to validate data

It helps to identify errors both in the structure and content of a data set

Using goodtables.io

Validation checks & errors:

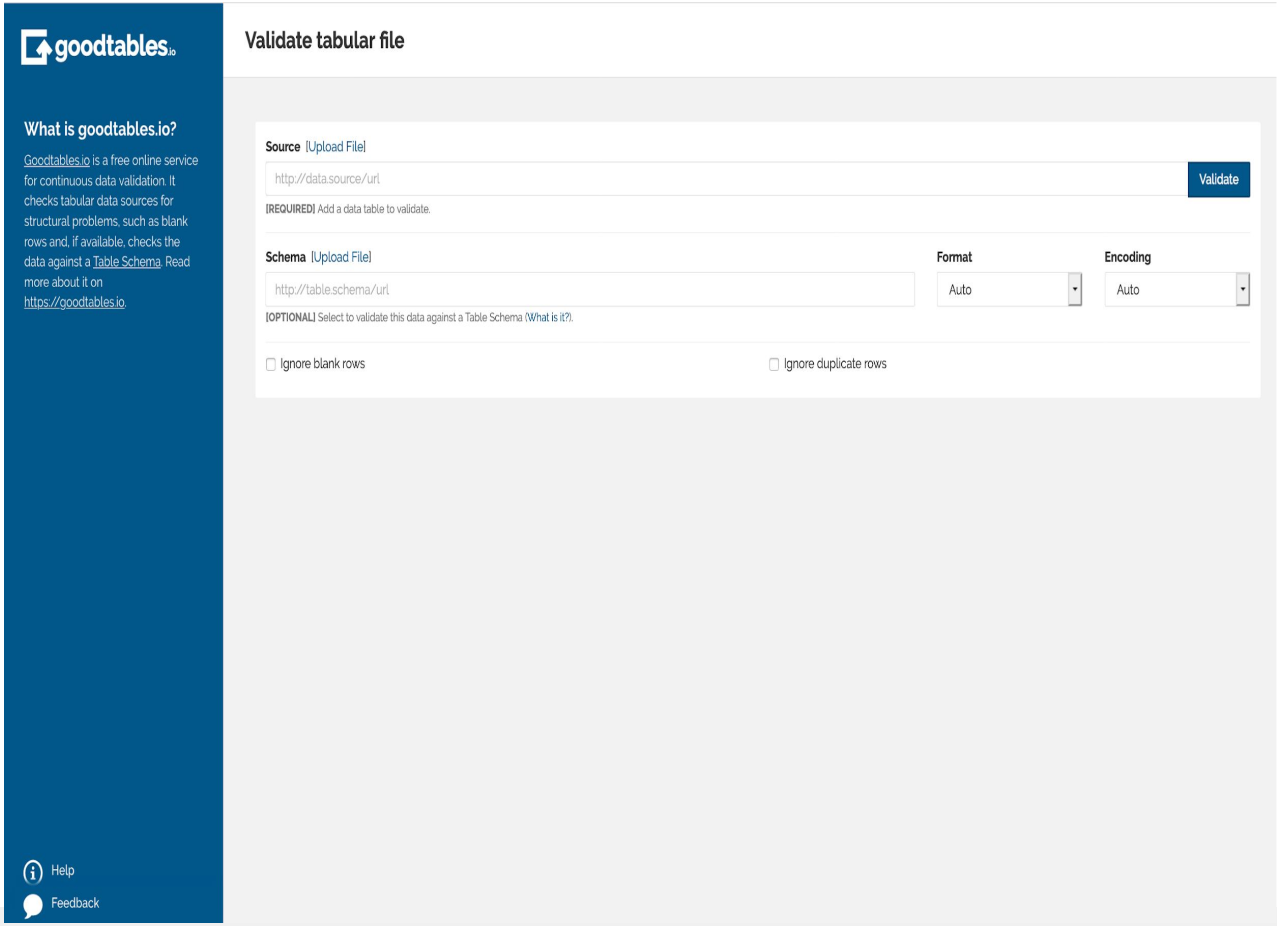
- **Structural checks**: Ensure that there are no empty rows, no blank headers, etc.
- **Content checks**: Ensure that values have the correct types, that their format is valid, that they respect constraints

Introducing goodtables.io

1. One time data validation via web tool, command line
2. Continuous validation for data hosted in GitHub or other open repositories (Amazon S3)

Using goodtables.io online

Web version: try.goodtables.io/



goodtables.io

Validate tabular file

What is goodtables.io?

Goodtables.io is a free online service for continuous data validation. It checks tabular data sources for structural problems, such as blank rows and, if available, checks the data against a Table Schema. Read more about it on https://goodtables.io.

Source [Upload File]

http://data.source/url

Validate

[REQUIRED] Add a data table to validate.

Schema [Upload File]

http://table.schema/url

[OPTIONAL] Select to validate this data against a Table Schema (What is it?).

Format: Auto

Encoding: Auto

□ Ignore blank rows

□ Ignore duplicate rows

Goodtables

How to use

1. Upload resource: structural check
    - link to file
    - file directly
2. Schema: content check

Read more at: https://docs.goodtables.io/

goodtables.io

Hands On Time

http://try.goodtables.io/

DATA

(optional Schema file)

Questions

Summary & Conclusions

Open Science

- Open Data
- Open Methodology
- Open Source
- Open Access
- Open Peer Review
- Open Educational Resources

A Community-Sourced Glossary of Open Scholarship Terms.
https://forrt.org/glossary/open-science/

Open Science

- Transparent
- Credible
- Reproducible
- Accessible
- Beneficial to everyone

A Community-Sourced Glossary of Open Scholarship Terms.
https://forrt.org/glossary/open-science/

Open Science

- **Open Data**
- Open Methodology
- Open Source
- Open Access
- Open Peer Review
- Open Educational Resources

A Community-Sourced Glossary of Open Scholarship Terms.
https://forrt.org/glossary/open-science/

FAIR Data

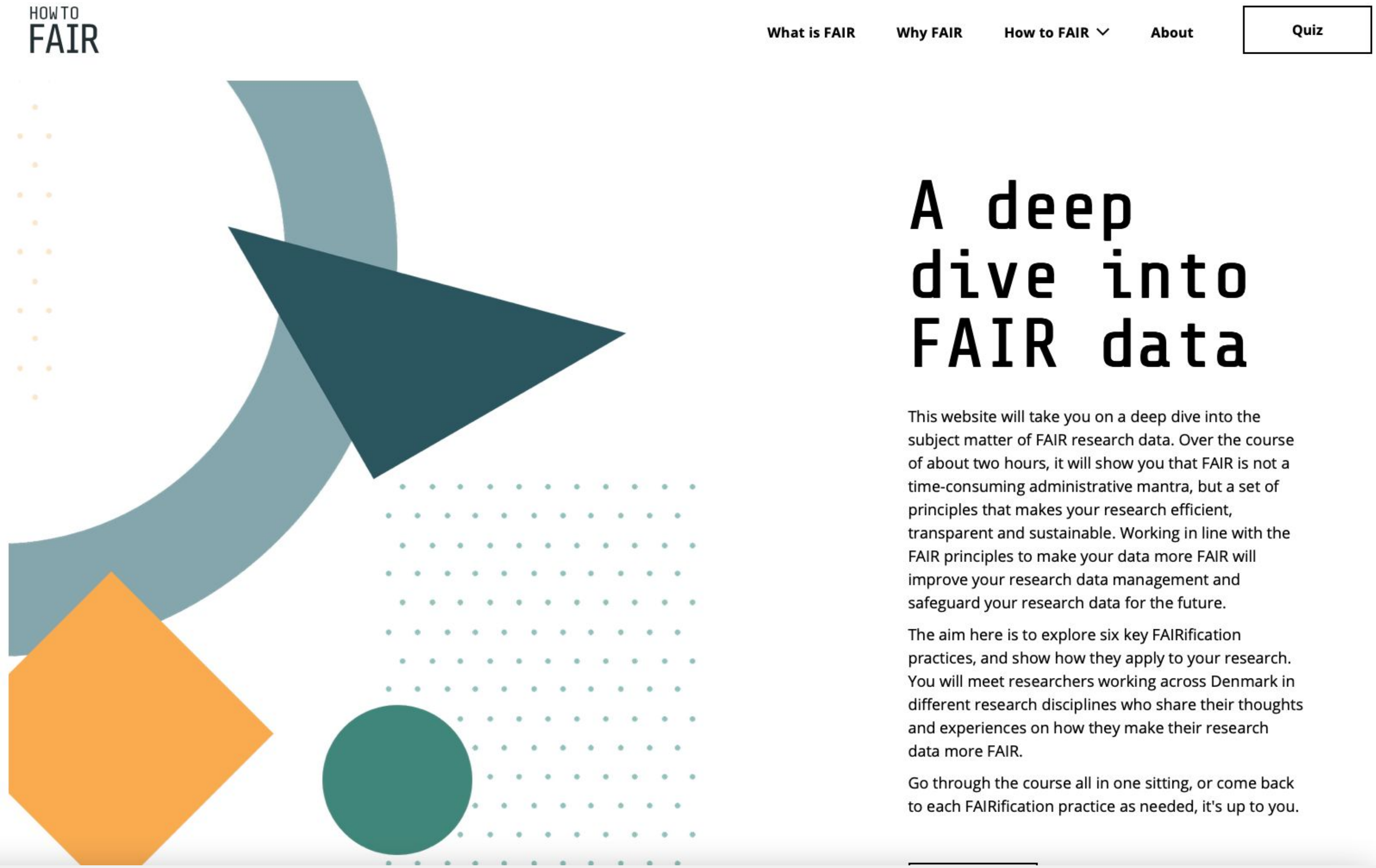
PERSISTANT
FINDABLE
MEANINGFUL INTERACTION
ACCESSIBLE
FAIR
PRINCIPLES
INTEROPERABLE
FULL DISCLOSURE
REUSABLE
Scriberia

- Findability
- Accessibility
- Interoperability
- Reusability

Frictionless Data Fellows

The Turing Way: A Handbook for Reproducible Data Science (Version v0.0.4). Zenodo. https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-fair.html?highlight=fair%20principles

How to FAIR



A deep dive into FAIR data

This website will take you on a deep dive into the subject matter of FAIR research data. Over the course of about two hours, it will show you that FAIR is not a time-consuming administrative mantra, but a set of principles that makes your research efficient, transparent and sustainable. Working in line with the FAIR principles to make your data more FAIR will improve your research data management and safeguard your research data for the future.

The aim here is to explore six key FAIRification practices, and show how they apply to your research. You will meet researchers working across Denmark in different research disciplines who share their thoughts and experiences on how they make their research data more FAIR.

Go through the course all in one sitting, or come back to each FAIRification practice as needed, it's up to you.

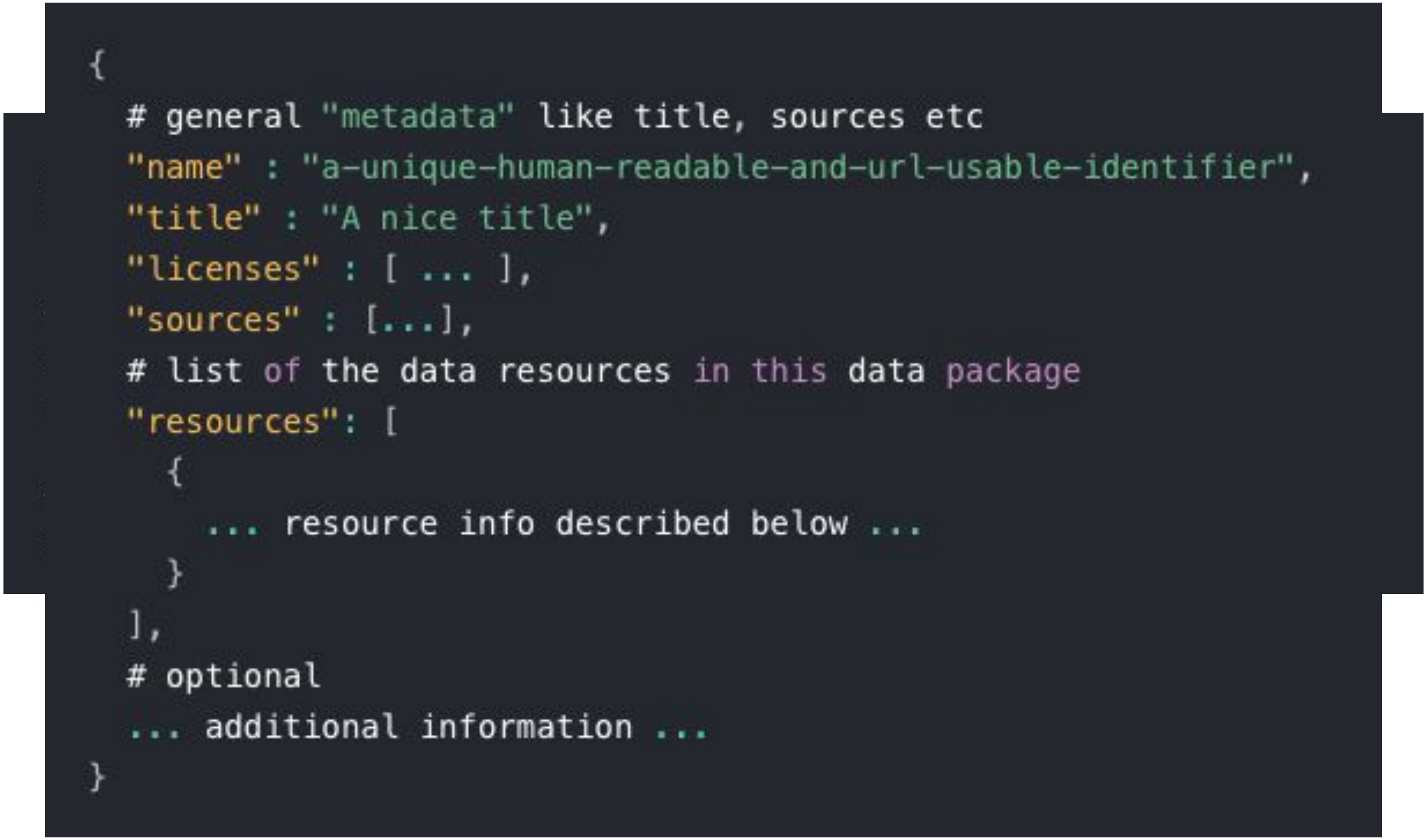
How to FAIR: https://howtofair.dk

Data Package

- A Data Package is a simple container format used to describe and package a collection of data
- Contains a descriptor (including metadata and schema)
- May includes data, analysis code, etc.
- FAIR

Frictionless Standards. Data Package:
https://specs.frictionlessdata.io/data-package/#language

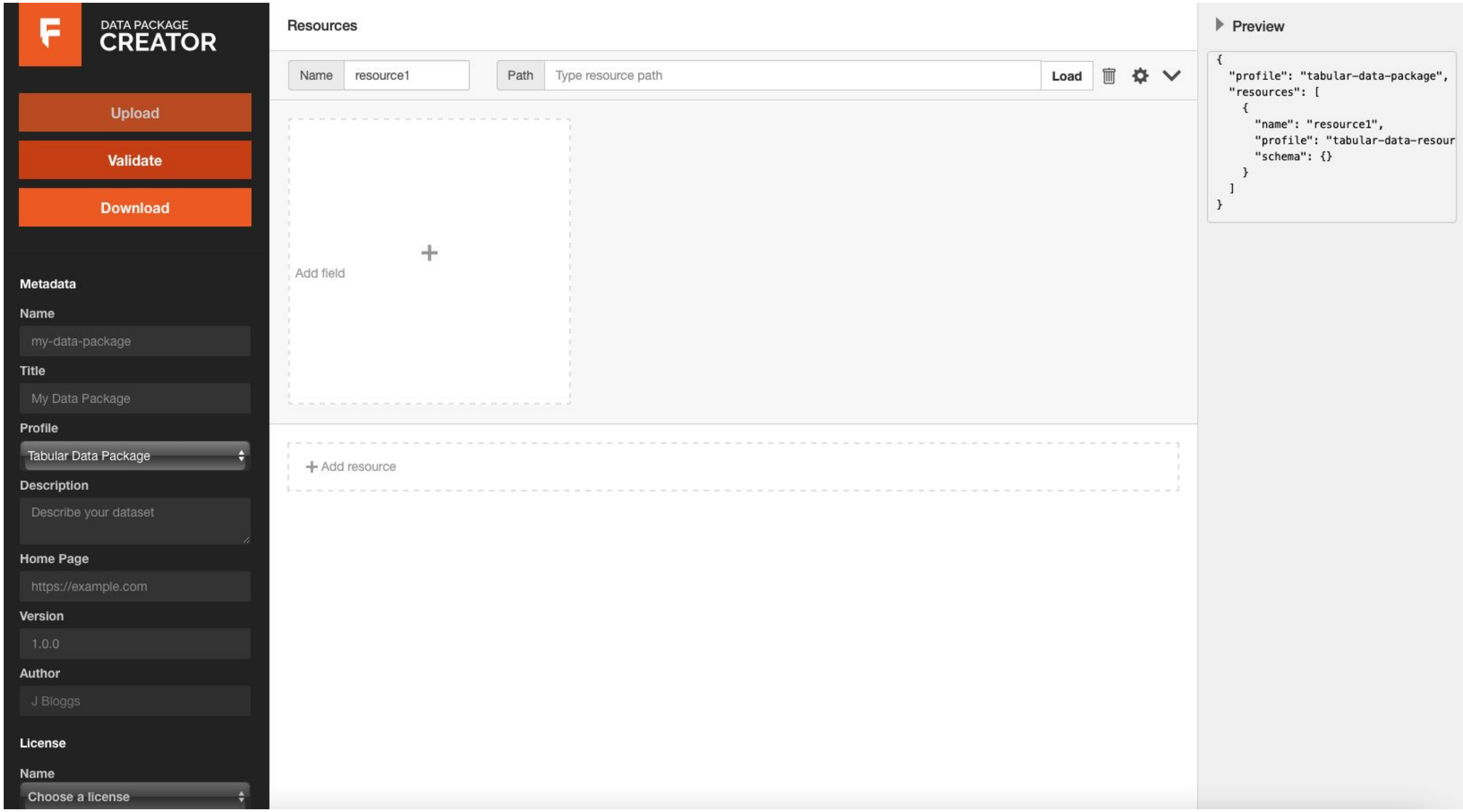
Data Package



```
{
  # general "metadata" like title, sources etc
  "name" : "a-unique-human-readable-and-url-usable-identifier",
  "title" : "A nice title",
  "licenses" : [ ... ],
  "sources" : [...],
  # list of the data resources in this data package
  "resources": [
    {
      ... resource info described below ...
    }
  ],
  # optional
  ... additional information ...
}
```

Frictionless Standards. Data Package: https://specs.frictionlessdata.io/data-package/#language

Data Package Creator



Metadata

Name: my-data-package

Title: My Data Package

Profile: Tabular Data Package

Description: Describe your dataset

Home Page: https://example.com

Version: 1.0.0

Author: J Bloggs

License

Name: Choose a license

Name: resource1 Path: Type resource path

```
{
  "profile": "tabular-data-package",
  "resources": [
    {
      "name": "resource1",
      "profile": "tabular-data-resour
      "schema": {}
    }
  ]
}
```

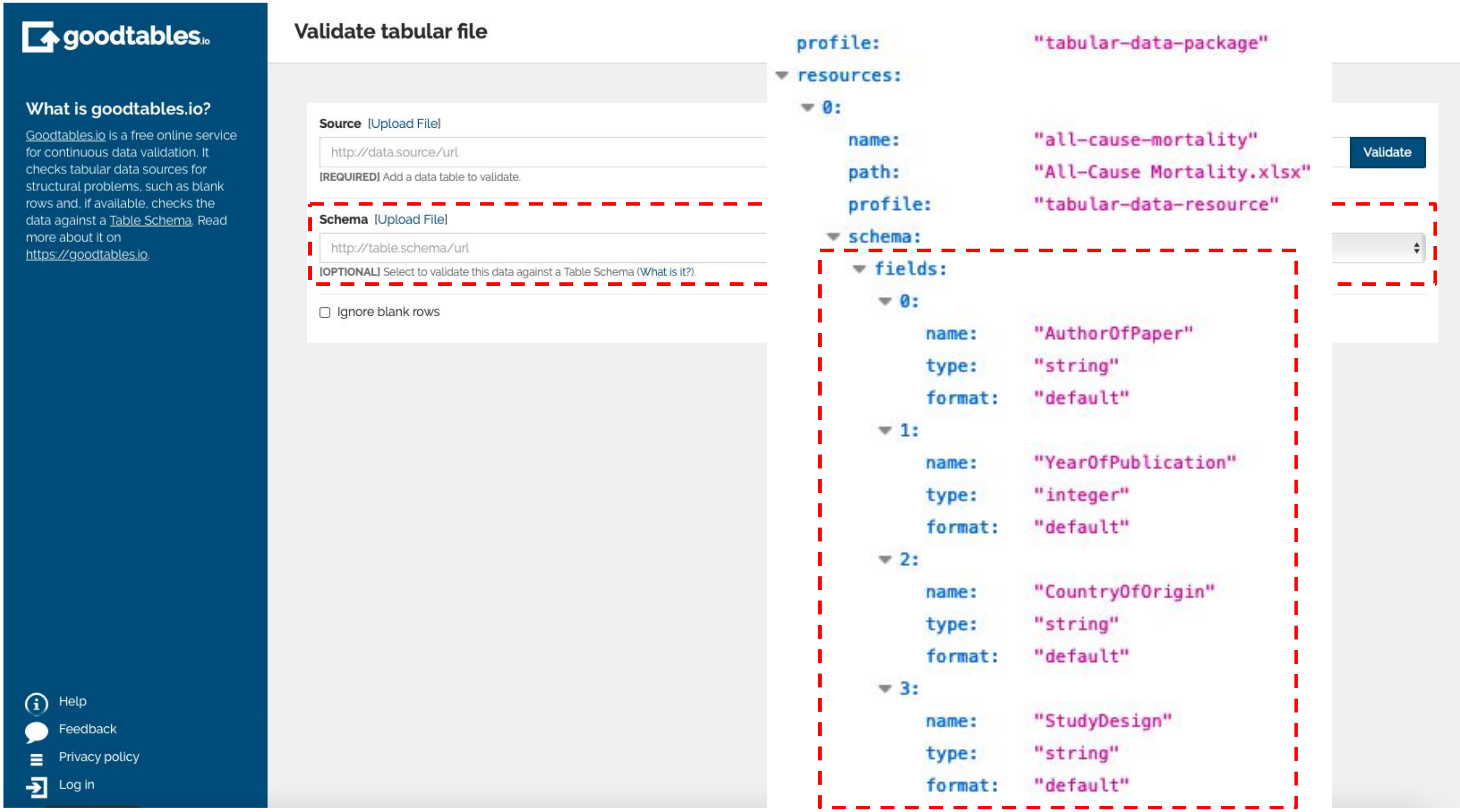
Data Package Creator:
https://create.frictionlessdata.io

Data Validation

- Validate with or without a schema
- One-time validation or continuous validation
- Validate data before sharing your data or using it for analysis

Frictionless Data. Validated Tabular Data: https://frictionlessdata.io/blog/2018/07/16/validated-tabular-data/#one-time-data-validation-with-try-goodtables-io

Goodtables Tool



Validate tabular file

What is goodtables.io?

Goodtables.io is a free online service for continuous data validation. It checks tabular data sources for structural problems, such as blank rows and, if available, checks the data against a Table Schema. Read more about it on https://goodtables.io.

Source [Upload File]

http://data.source/url

[REQUIRED] Add a data table to validate.

Schema [Upload File]

http://table.schema/url

[OPTIONAL] Select to validate this data against a Table Schema (What is it?).

□ Ignore blank rows

```
  profile:          "tabular-data-package"
▼ resources:
  ▼ 0:
      name:         "all-cause-mortality"
      path:         "All-Cause Mortality.xlsx"
      profile:      "tabular-data-resource"
    ▼ schema:
      ▼ fields:
        ▼ 0:
            name:   "AuthorOfPaper"
            type:   "string"
            format: "default"
        ▼ 1:
            name:   "YearOfPublication"
            type:   "integer"
            format: "default"
        ▼ 2:
            name:   "CountryOfOrigin"
            type:   "string"
            format: "default"
        ▼ 3:
            name:   "StudyDesign"
            type:   "string"
            format: "default"
```

Goodtables Tool:
https://try.goodtables.io

http://fellows.frictionlessdata.io/

https://framework.frictionlessdata.io/

Github: http://github.com/frictionlessdata/

Slack: https://join.slack.com/t/frictionlessdata/shared_invite/zt-17kpbffnm-tRfDW_wJgOw8tJVLvZTrBg

Youtube: youtube.com/user/openknowledgefdn

Twitter: @frictionlessd8a

Thank you!

Join our community!