Toward a Frictionless Data Future

PRESENTED BY

Dan Fowler

Developer Advocate

daniel.fowler@okfn.org (@danfowler & @okfnlabs)

OPEN KNOWLEDGE INTERNATIONAL

FRICTIONLESS DATA
STANDARDS AND TOOLING

Who we are

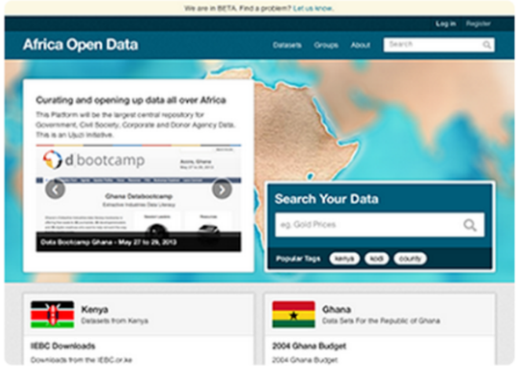
International non-profit founded in 2004

OPEN KNOWLEDGE
INTERNATIONAL

- **Vision**
  - A world where **open knowledge is ubiquitous**
- **Mission**
  - **Open up** all essential public-interest information
  - **Build communities, tools, and skills** to empower individuals to use open information to drive change

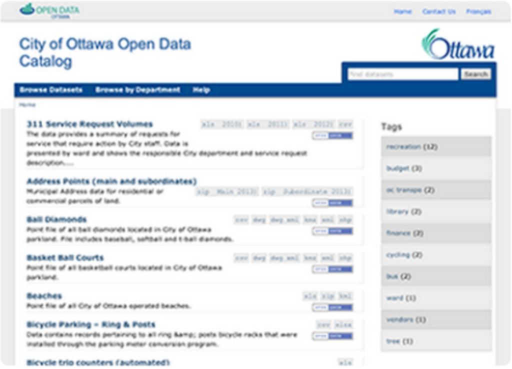
CKAN -> Frictionless Data

- Our work on Frictionless Data is borne out of what we’ve learnt from building and deploying CKAN
- Working with publishers and data publication flows all around the world has highlighted the “friction” involved in working with data



Africa Open Data

Buenos Aires Data



City of Ottawa Open Data Catalog
Ottawa
311 Service Request Volumes
Address Points (main and subordinates)
Ball Diamonds
Basket Ball Courts
Beaches
Bicycle Parking - King & Posts
Tags

City of Ottawa Open

"Tidy" Data

Framework for prepping data for analysis

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data *(Dasu and Johnson 2003). Data preparation is not just a first step, but must be repeated many over the course of analysis as new problems come to light or new data is collected. [...] Part of the challenge is the breadth of activities it encompasses: from outlier checking, to date parsing, to missing value imputation.*

Hadley Wickham, Tidy Data (2014)
https://www.jstatsoft.org/article/view/v059i10

FRICTIONLESS DATA
STANDARDS AND TOOLING

Frictionless Data:

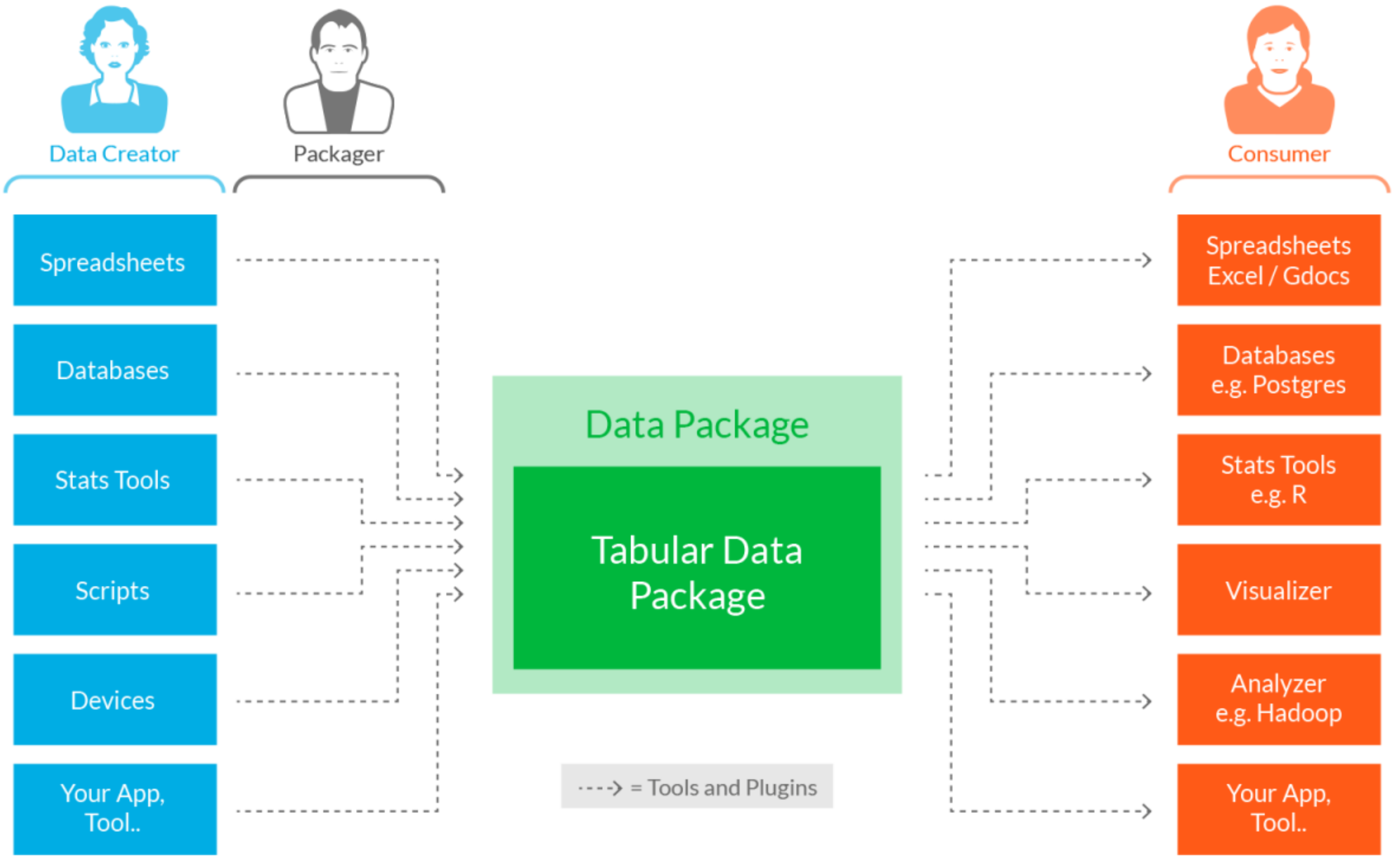
- Lightweight extensible **specifications** for “packaging” datasets

  Current focus: **Tabular Data**
- **Integrations** for loading datasets into tools and platforms relevant to researchers

Aims:

- Introduce a significant, *measurable* improvement in how research data is shared, consumed, and analyzed.
- Make it easier to maintain and improve data **quality**.

Integrations



Data Creator

Packager

Consumer

Spreadsheets

Databases

Stats Tools

Scripts

Devices

Your App, Tool..

Data Package

Tabular Data Package

----> = Tools and Plugins

Spreadsheets Excel / Gdocs

Databases e.g. Postgres

Stats Tools e.g. R

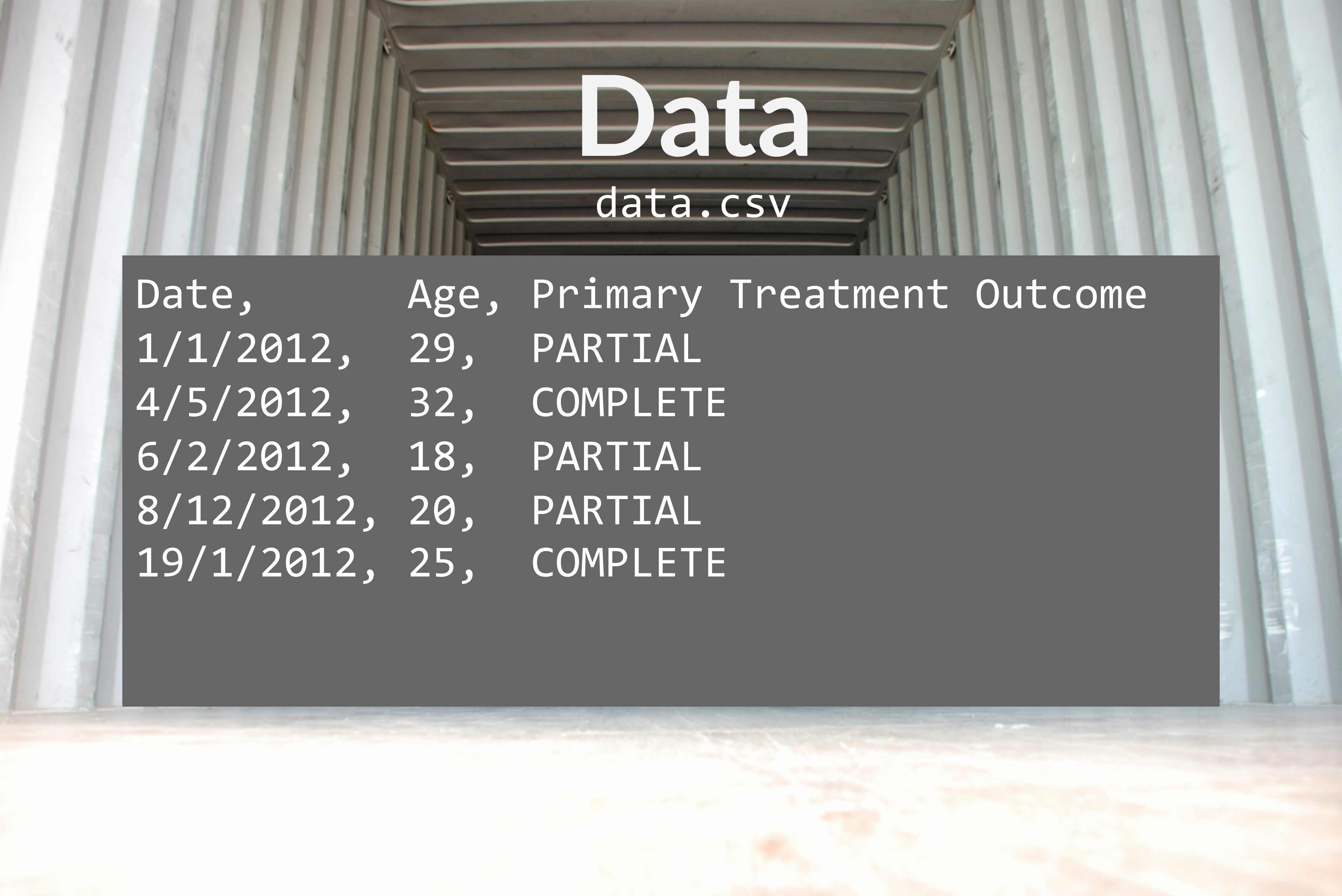
Visualizer

Analyzer e.g. Hadoop

Your App, Tool..

Data Containerization

http://www.flickr.com/photos/photohome_uk/1494590209/



Data

data.csv

```
Date,       Age, Primary Treatment Outcome
1/1/2012,   29,  PARTIAL
4/5/2012,   32,  COMPLETE
6/2/2012,   18,  PARTIAL
8/12/2012,  20,  PARTIAL
19/1/2012,  25,  COMPLETE
```

Data (Packaged)

datapackage.json

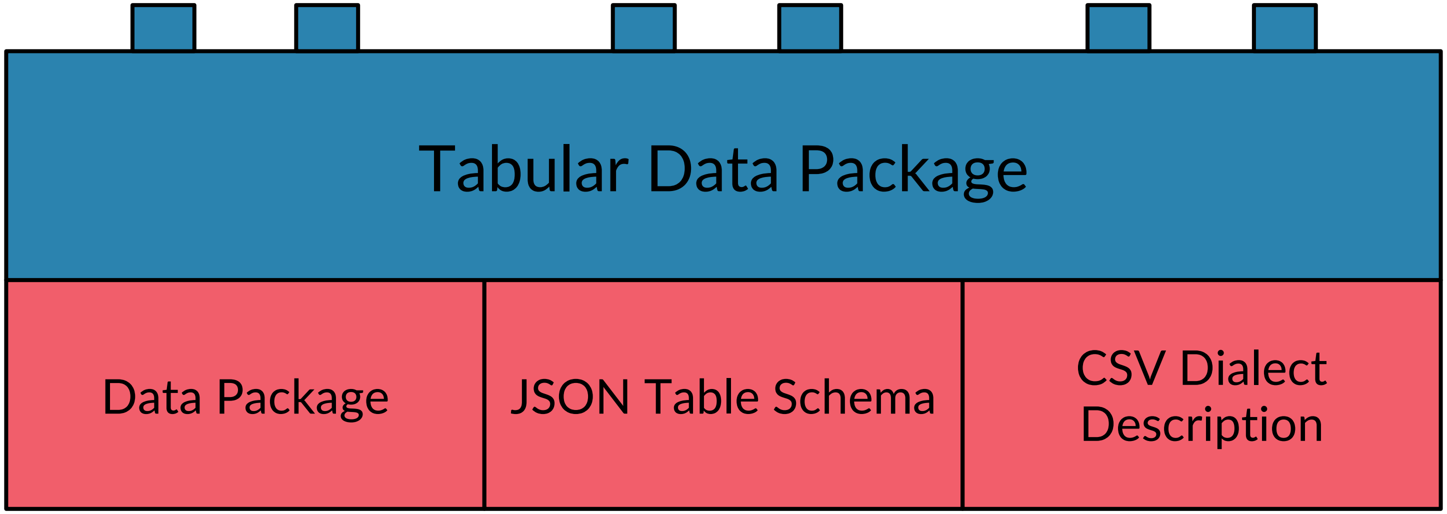
```
{
  "name": "my-data",
  "title": "My Data",
  "sources": ["http://example.com/data.csv"],
  "license" : "ODC-PDDL-1.0",
  "resources": [{
    "path": "data.csv",
    "format": "csv",
    "dialect": {
      "delimiter": ",",
      "header": true
    },
    "schema": {
      ...
    }
  }]
}
```

OPEN KNOWLEDGE

JSON Table Schema

```
"schema": {
  "fields": [
    {
      "name":"Date", "type":"date", "format":"fmt:%d/%m/%Y",
      "constraints": {"required": "true"}
    },
    {
      "name":"Age", "type":"integer",
      "missingValue": "NULL"
    },
    {
      "name":"Primary Therapy Outcome",
      "description": "Whether primary treatment success",
      "type": "string",
      "constraints": {"enum": ["PARTIAL","COMPLETE"]}
    }
  ]
}
```

Tabular Data Package



http://specs.frictionlessdata.io/csv-dialect/

http://specs.frictionlessdata.io/data-packages/

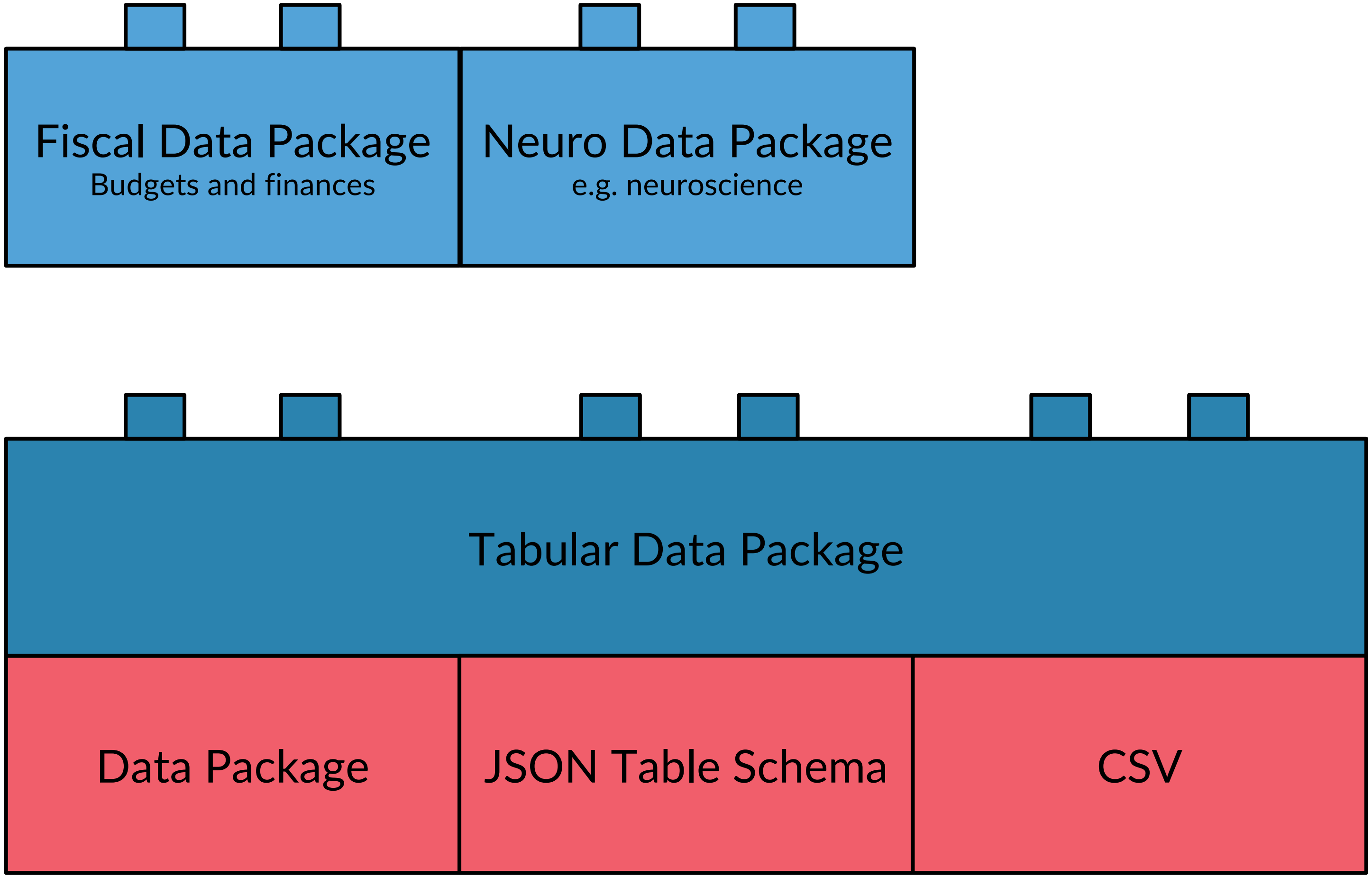
http://specs.frictionlessdata.io/json-table-schema/

http://specs.frictionlessdata.io/tabular-data-package/

Benefits of Approach

http://www.flickr.com/photos/photohome_uk/1494590209/

Focus on Extensibility



Fiscal Data Package
Budgets and finances

Neuro Data Package
e.g. neuroscience

Tabular Data Package

Data Package

JSON Table Schema

CSV

MEXICO

1st COUNTRY TO FORMALLY ADOPT THE

OPEN FISCAL DATA PACKAGE

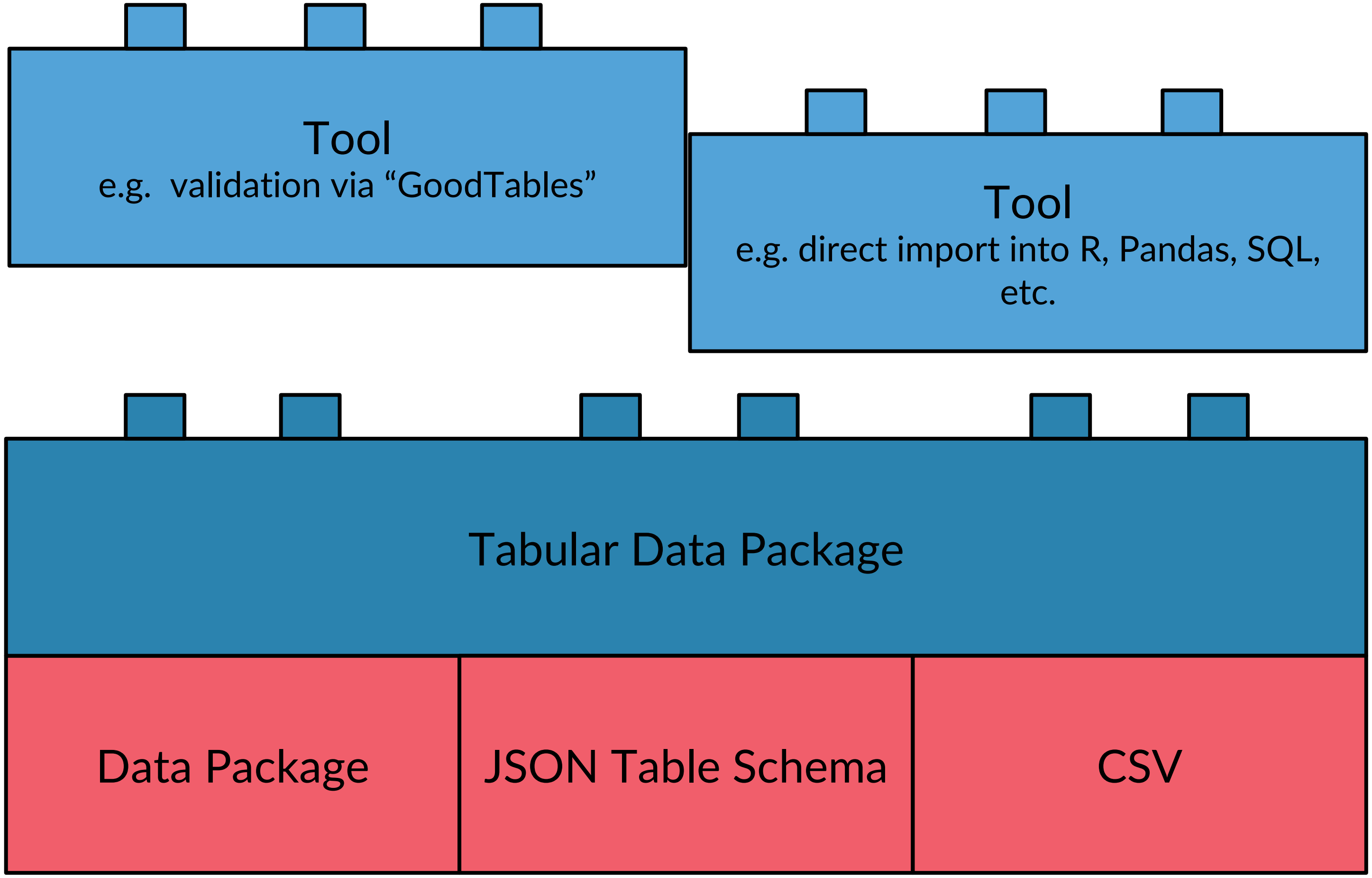
SHCP
SECRETARÍA DE HACIENDA Y CRÉDITO PÚBLICO
ESTADOS UNIDOS MEXICANOS
GIFT
GLOBAL INITIATIVE FOR FISCAL TRANSPARENCY
TRANSPARENCIA PRESUPUESTARIA
Observatorio del gasto
OPEN KNOWLEDGE INTERNATIONAL

http://www.transparenciapresupuestaria.gob.mx/es/PTP/datos_presupuestarios_abiertos

Tools

http://www.flickr.com/photos/photohome_uk/1494590209/

Solid Foundation for Tools



Tool
e.g. validation via “GoodTables”

Tool
e.g. direct import into R, Pandas, SQL, etc.

Tabular Data Package

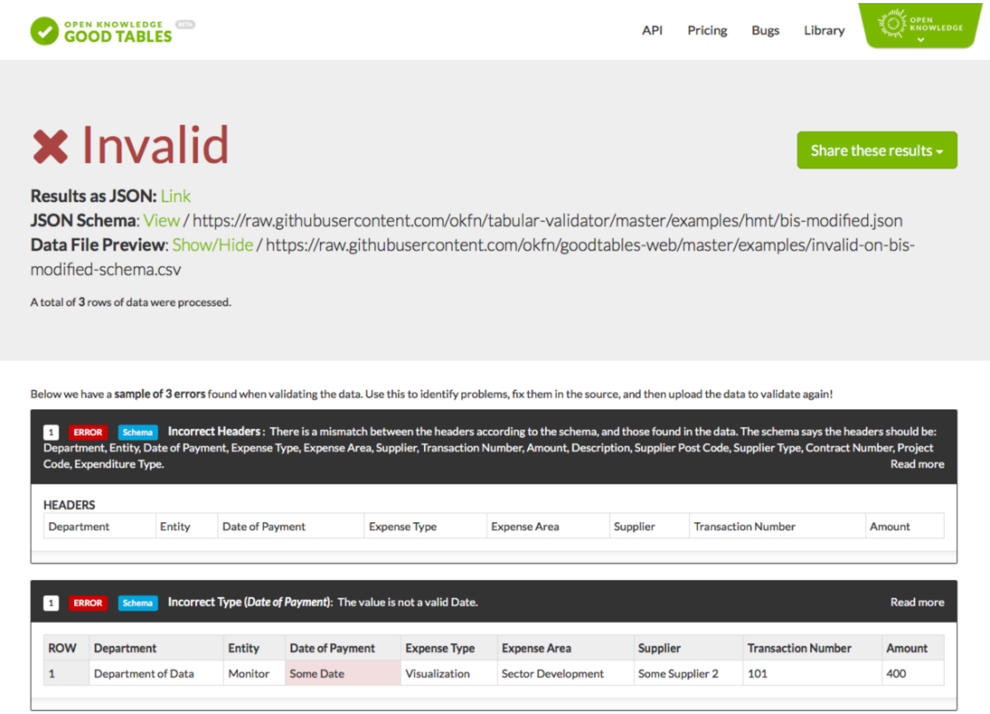
Data Package

JSON Table Schema

CSV

Good Tables

- GoodTables is a tool built by Open Knowledge International for data source validation.
- Once you have defined a schema using JSON Table Schema, you can make sure your CSV file is valid against the schema



✖ Invalid

Results as JSON: Link
JSON Schema: View / https://raw.githubusercontent.com/okfn/tabular-validator/master/examples/hmt/bis-modified.json
Data File Preview: Show/Hide / https://raw.githubusercontent.com/okfn/goodtables-web/master/examples/invalid-on-bis-modified-schema.csv

A total of **3** rows of data were processed.

Below we have a **sample of 3 errors** found when validating the data. Use this to identify problems, fix them in the source, and then upload the data to validate again!

1 ERROR Schema **Incorrect Headers**: There is a mismatch between the headers according to the schema, and those found in the data. The schema says the headers should be: Department, Entity, Date of Payment, Expense Type, Expense Area, Supplier, Transaction Number, Amount, Description, Supplier Post Code, Supplier Type, Contract Number, Project Code, Expenditure Type. Read more

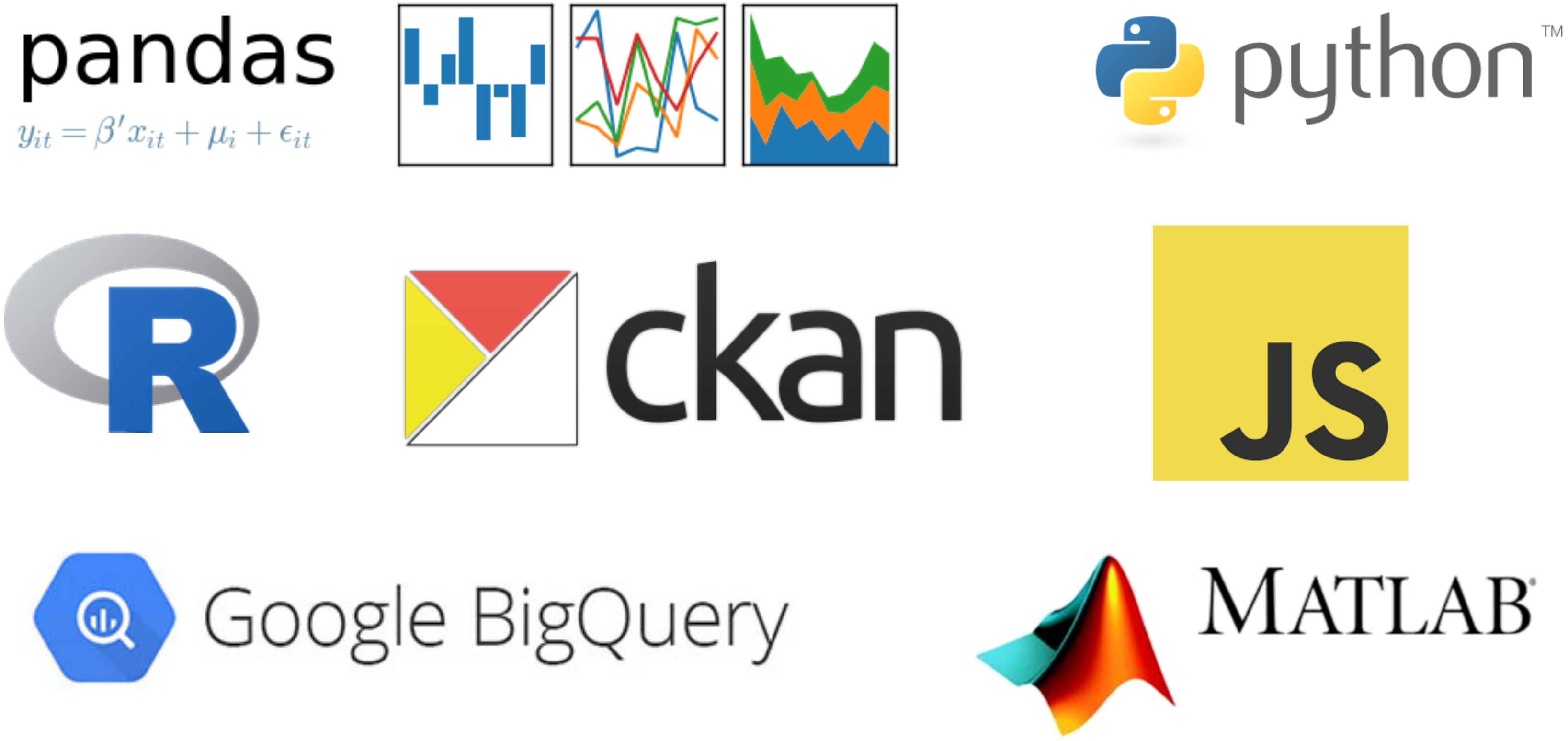
HEADERS

Department	Entity	Date of Payment	Expense Type	Expense Area	Supplier	Transaction Number	Amount

1 ERROR Schema **Incorrect Type (*Date of Payment*)**: The value is not a valid Date. Read more

ROW	Department	Entity	Date of Payment	Expense Type	Expense Area	Supplier	Transaction Number	Amount
1	Department of Data	Monitor	Some Date	Visualization	Sector Development	Some Supplier 2	101	400

Platform Integrations



Partners

©photohome.co.uk

http://www.flickr.com/photos/photohome_uk/1494590209/

Partners

Western Pennsylvania Regional Data Center

DM4T

EPSRC Data Management for TEDDINET using Semantic Technologies

R OpenSci

Dataship

Pacific Northwest NATIONAL LABORATORY

UNIVERSITY OF CAMBRIDGE

FRICTIONLESS DATA
STANDARDS AND TOOLING

Seeking Pilot and Technical Partnerships

- Project overview: http://frictionlessdata.io
- Partners: http://frictionlessdata.io/partners/
- Case Studies: http://frictionlessdata.io/case-studies/
- Specifications: http://specs.frictionlessdata.io/
- **Newsletter**: http://frictionlessdata.io/get-involved/#newsletter

Questions?

©photohome.co.uk

http://www.flickr.com/photos/photohome_uk/1494590209/