Toward a Frictionless Data Future

PRESENTED BY

Dan Fowler
Developer Advocate
daniel.fowler@okfn.org (@danfowler & @okfnlabs)
Who we are

International non-profit founded in 2004

- **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.
"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:

- 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 Package

- Spreadsheets
- Databases
- Stats Tools
- Scripts
- Devices
- Your App, Tool...

---

Tools and Plugins

- Spreadsheets Excel / Gdocs
- Databases e.g. Postgres
- Stats Tools e.g. R
- Visualizer
- Analyzer e.g. Hadoop
- Your App, Tool...

Data Package

- Data Creator
- Packager
- Consumer

---

Tools and Plugins
Data Containerization
<table>
<thead>
<tr>
<th>Date</th>
<th>Age</th>
<th>Primary Treatment Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1/2012</td>
<td>29</td>
<td>PARTIAL</td>
</tr>
<tr>
<td>4/5/2012</td>
<td>32</td>
<td>COMPLETE</td>
</tr>
<tr>
<td>6/2/2012</td>
<td>18</td>
<td>PARTIAL</td>
</tr>
<tr>
<td>8/12/2012</td>
<td>20</td>
<td>PARTIAL</td>
</tr>
<tr>
<td>19/1/2012</td>
<td>25</td>
<td>COMPLETE</td>
</tr>
</tbody>
</table>
**Data (Packaged)**

```
{
    "name": "my-data",
    "title": "My Data",
    "sources": ["http://example.com/data.csv"],
    "license": "ODC-PDDL-1.0",
    "resources": [{
        "data": "data.csv",
        "format": "csv",
        "dialect": {
            "delimiter": ",",
            "header": true
        },
        "schema": {
            ...
        }
    }]
}
```
"schema": { "fields": [ { "name": "Date", "type": "date", "format": "fmt:%d/%m/%Y" }, { "name": "Age", "type": "integer", "constraints": { "minimum": 0, "required": "true" }, "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/
Benefits of Approach
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

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

- Standards (a few, simple ones)

- Tools (primarily for integration)

- Documentation

- Datasets

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.
Platform Integrations

\[ y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it} \]
Partners
Partners
Seeking Pilot and Technical Partnerships

- Project overview: [http://frictionlessdata.io](http://frictionlessdata.io)
- Partners: [http://frictionlessdata.io/partners/](http://frictionlessdata.io/partners/)
Standards
- Standards (a few, simple ones)
- Tools (primarily for integration)
- Documentation
- Datasets

Questions?

http://www.flickr.com/photos/photohome_uk/1494590209/
Narrow Focus: Onloading / Offloading

Collect Data → Clean & Import → Gov Stat Site → Import to your tool → Analysis

Load / Unload
1. **Narrow Focus:** One small part of the data analysis workflow

2. **Web Oriented:** Uses web-native formats

3. **Focus on Existing Tools:** Maximum impact if you are already using R, Python Pandas, or Excel plus graceful degradation (CSV widely supported)

4. **Open:** Individual scientific communities can reuse what we build, and match to their needs

5. **Simple:** Use most basic formats. Add the minimum, do the least needed.
"Friction"

- Hard to discover or find.
- Not open - access is restricted, or charged for.
- Structure is poor and/or needs significant manipulation to be usable so even when you do find data, it is usually a manual process to connect it with your tool.
- No standardized schema and different data sources are hard to compare or integrate
- Integration with existing tooling is poor or non-existent so, for example, extensive time must be spent manually transferring data from one tool to another
Fiscal Data Package

- We extend the base Data Package to support modelling high-level concepts related to budgets like financial amounts, budget classifications, etc.
- [http://next.openspending.org](http://next.openspending.org)