Frictionless Data: Making Research Data Quality Visible

Dan Fowler
daniel.fowler@okfn.org (@danfowler/@okfnlabs)

PRESENTED BY

Jo Barratt
jo.barratt@okfn.org (@jobarratt)

AT
IDCC17
Who we are

International non-profit founded in 2004 by Rufus Pollock

Vision

A world where open knowledge is ubiquitous, enabling citizens and organizations to create insights that drive change on global and local challenges, combat injustice and inequality and hold governments and corporations to account.

Mission

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.
Tabular Data Package

http://frictionlessdata.io/guides/data-package/
http://frictionlessdata.io/guides/json-table-schema/
Similar Initiatives
Pilots
Data Management for TEDDINET

Active Data Biology at Pacific Northwest National Laboratory

GitHub: https://github.com/frictionlessdata/pilot-pnnl
Good Tables

Good Tables is a free hosted service for validating tabular data. Find out more

Let's get started

Add a data table (required)

Choose URL or File to add a data source by your preferred method.

Want to see an example? Click on one of the two options below followed by the green 'Validate' button.

- Example: Invalid Structure
- Example: Invalid Schema (also sets a test schema to validate against)
### Job

<table>
<thead>
<tr>
<th>Job Id</th>
<th>Status</th>
<th>Valid</th>
<th>Errors</th>
<th>Tables</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>3d9d13cd-eecf-48b3-b5f6-ba8c3552bf72</td>
<td>success</td>
<td>True</td>
<td>0</td>
<td>1</td>
<td>0.757</td>
</tr>
</tbody>
</table>

### Files

frictionlessdata/ADB-User-Study/4b786a297d1ac3343fa00d7af918e5dbe7428c38/metadata.tsv

<table>
<thead>
<tr>
<th>Valid</th>
<th>Errors</th>
<th>Rows</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0</td>
<td>177</td>
<td>0.556</td>
</tr>
</tbody>
</table>
**Job**

<table>
<thead>
<tr>
<th>Job Id</th>
<th>Status</th>
<th>Valid</th>
<th>Errors</th>
<th>Tables</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>21ff2708-4c66-47af-8406-fcb8ab3a6aa1</td>
<td>failure</td>
<td>False</td>
<td>56</td>
<td>1</td>
<td>0.741</td>
</tr>
</tbody>
</table>

**Files**

`frictionlessdata/ADB-User-Study/5ae9f3bf6903d9e284d145438c2aad9ec688778d/metadata.tsv`

<table>
<thead>
<tr>
<th>Valid</th>
<th>Errors</th>
<th>Rows</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>False</td>
<td>56</td>
<td>177</td>
<td>0.557</td>
</tr>
</tbody>
</table>

**Errors**

<table>
<thead>
<tr>
<th>Row</th>
<th>Column</th>
<th>Code</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>non-castable-value</td>
<td>Row 2 has non castable value in column 2 (type: number, format: default)</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>non-castable-value</td>
<td>Row 2 has non castable value in column 4 (type: boolean, format: default)</td>
</tr>
</tbody>
</table>
UK Spend Publishing Dashboard

How are government departments living up to their commitments to publish timely, high quality data on spending?

Data Quality Dashboard

<table>
<thead>
<tr>
<th>Title</th>
<th>Type</th>
<th>Homepage</th>
<th>Email</th>
<th>Correct</th>
<th>Score (all time)</th>
<th>Last file</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabinet Office</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td></td>
<td>50 %</td>
<td>August 2014</td>
<td>90 %</td>
</tr>
<tr>
<td>Department for Business, Innovation and Skills</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td></td>
<td>41 %</td>
<td>February 2015</td>
<td>90 %</td>
</tr>
<tr>
<td>Department of Health</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td></td>
<td>56 %</td>
<td>July 2015</td>
<td>90 %</td>
</tr>
<tr>
<td>Foreign and Commonwealth Office</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td></td>
<td>36 %</td>
<td>April 2015</td>
<td>90 %</td>
</tr>
<tr>
<td>Ministry of Defence</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td></td>
<td>59 %</td>
<td>June 2015</td>
<td>90 %</td>
</tr>
<tr>
<td>Department for Communities and Local Government</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td></td>
<td>27 %</td>
<td>April 2012</td>
<td>70 %</td>
</tr>
<tr>
<td>Department for Environment, Food and Rural Affairs</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td>10</td>
<td>57 %</td>
<td>June 2015</td>
<td>60 %</td>
</tr>
<tr>
<td>Scotland Office</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td>5</td>
<td>57 %</td>
<td>October 2012</td>
<td>30 %</td>
</tr>
<tr>
<td>Department for Work and Pensions</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td></td>
<td>45 %</td>
<td>June 2015</td>
<td>20 %</td>
</tr>
<tr>
<td>Attorney General’s Office</td>
<td>Ministerial department</td>
<td></td>
<td></td>
<td></td>
<td>0 %</td>
<td>March 2010</td>
<td>0 %</td>
</tr>
</tbody>
</table>
Case Studies
open-power-system-data.org
Technical Work
Specifications v1.0

http://specs.frictionlessdata.io/

- The removal of many examples of unclear/ambiguous wording making it hard to implement specifications in certain corner cases.
- The delivery of around 20 core fixes and enhancements targeted for version 1.0
- A redesign of how we write the specifications to make it easy to generate RFC versions for the IETF
- The addition of a new specification for “Data Resource”, which was previously only specified as part of “Data Package”. The distinct Data Resource specification addresses an extremely common use case / pattern that we have seen in our own work and that of others
- The renaming of JSON Table Schema to Table Schema, and the simplification of some important aspects of that specification
Internal Implementations

Core libraries
- https://github.com/frictionlessdata/dapackage-py
- https://github.com/frictionlessdata/dapackage-js
- https://github.com/frictionlessdata/jsonschema-py
- https://github.com/frictionlessdata/jsonschema-js

Additional libraries
- https://github.com/frictionlessdata/goodtables-py
- https://github.com/ckan/ckanext-datapackager
- https://github.com/frictionlessdata/jsonschema-sql-py
- https://github.com/frictionlessdata/jsonschema-bigquery-py
Conclusion
● Project website: http://frictionlessdata.io/
● Specifications: http://specs.frictionlessdata.io/
● GitHub: https://github.com/frictionlessdata/
● User Stories: http://frictionlessdata.io/user-stories/
● Follow @okfnlabs on Twitter (#frictionlessdata)