Data Packages & Frictionless Data
http://data.okfn.org/

PRESENTED BY
Rufus Pollock, @rufuspollock / @okfn
Non-profit founded in 2004, operating globally

Mission to open up all, essential, public interest information and see it used to create insight that drives positive change
The Problem
Goal: Reduce “time to insight”

BUT

Friction in getting, using & sharing data is in the way
Goal: Eliminating Friction
The Analogy: Logistics
Logistics for “Goods” e.g. a Phone

1. Fabricate
2. Truck and Load
3. Ship by Sea
4. Unload & Truck
5. Retail Store

Logistics for Data e.g. Gov Stats

1. Collect Data
2. Clean & Import
3. Gov Stat Site
4. Import to your tool
5. Analysis
Many sources of friction

Data is ...

- 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
Narrow Focus: Onloading / Offloading

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

Load / Unload
1. Employment status of the civilian noninstitutional population, 1940 to date

(Numbers in thousands)

<table>
<thead>
<tr>
<th>Year</th>
<th>Civilian noninstitutional population</th>
<th>Total</th>
<th>Percent of population</th>
<th>Employed</th>
<th>Percent of population</th>
<th>Agriculture</th>
<th>Nonagricultural industries</th>
<th>Number</th>
<th>Percent of labor force</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1940</td>
<td>99,840</td>
<td>55,640</td>
<td>55.7</td>
<td>47,520</td>
<td>47.6</td>
<td>9,540</td>
<td>37,980</td>
<td>8,120</td>
<td>14.6</td>
</tr>
<tr>
<td>1941</td>
<td>99,900</td>
<td>55,910</td>
<td>56.0</td>
<td>50,350</td>
<td>50.4</td>
<td>9,100</td>
<td>41,250</td>
<td>5,560</td>
<td>9.9</td>
</tr>
<tr>
<td>1942</td>
<td>98,640</td>
<td>56,410</td>
<td>57.2</td>
<td>53,750</td>
<td>54.5</td>
<td>9,250</td>
<td>44,500</td>
<td>2,660</td>
<td>4.7</td>
</tr>
<tr>
<td>1943</td>
<td>94,640</td>
<td>55,540</td>
<td>58.7</td>
<td>54,470</td>
<td>57.6</td>
<td>9,080</td>
<td>45,390</td>
<td>1,070</td>
<td>1.9</td>
</tr>
<tr>
<td>1944</td>
<td>93,220</td>
<td>54,630</td>
<td>58.6</td>
<td>53,960</td>
<td>57.9</td>
<td>8,950</td>
<td>45,010</td>
<td>670</td>
<td>1.2</td>
</tr>
<tr>
<td>1945</td>
<td>94,090</td>
<td>53,860</td>
<td>57.2</td>
<td>52,820</td>
<td>56.1</td>
<td>8,580</td>
<td>44,240</td>
<td>1,040</td>
<td>1.9</td>
</tr>
<tr>
<td>1946</td>
<td>103,070</td>
<td>57,520</td>
<td>55.8</td>
<td>55,250</td>
<td>53.6</td>
<td>8,320</td>
<td>46,930</td>
<td>2,270</td>
<td>3.9</td>
</tr>
<tr>
<td>1947</td>
<td>106,018</td>
<td>60,168</td>
<td>56.8</td>
<td>57,812</td>
<td>54.5</td>
<td>8,256</td>
<td>49,557</td>
<td>2,356</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Persons 14 years of age and over

<table>
<thead>
<tr>
<th>Year</th>
<th>Civilian noninstitutional population</th>
<th>Total</th>
<th>Percent of population</th>
<th>Employed</th>
<th>Percent of population</th>
<th>Agriculture</th>
<th>Nonagricultural industries</th>
<th>Number</th>
<th>Percent of labor force</th>
</tr>
</thead>
<tbody>
<tr>
<td>1947</td>
<td>101,827</td>
<td>59,350</td>
<td>58.3</td>
<td>57,038</td>
<td>56.0</td>
<td>7,890</td>
<td>49,148</td>
<td>2,311</td>
<td>3.9</td>
</tr>
<tr>
<td>1948</td>
<td>103,068</td>
<td>60,621</td>
<td>58.8</td>
<td>58,343</td>
<td>56.6</td>
<td>7,629</td>
<td>50,714</td>
<td>2,276</td>
<td>3.8</td>
</tr>
<tr>
<td>1949</td>
<td>103,994</td>
<td>61,286</td>
<td>58.9</td>
<td>57,651</td>
<td>55.4</td>
<td>7,658</td>
<td>49,993</td>
<td>3,637</td>
<td>5.9</td>
</tr>
<tr>
<td>1950</td>
<td>104,995</td>
<td>62,208</td>
<td>59.2</td>
<td>58,918</td>
<td>56.1</td>
<td>7,160</td>
<td>51,758</td>
<td>3,288</td>
<td>5.3</td>
</tr>
<tr>
<td>1951</td>
<td>104,621</td>
<td>62,017</td>
<td>59.2</td>
<td>59,961</td>
<td>57.3</td>
<td>6,726</td>
<td>53,235</td>
<td>2,055</td>
<td>3.3</td>
</tr>
<tr>
<td>1952</td>
<td>105,231</td>
<td>62,138</td>
<td>59.0</td>
<td>60,250</td>
<td>57.3</td>
<td>6,500</td>
<td>53,749</td>
<td>1,883</td>
<td>3.0</td>
</tr>
<tr>
<td>1953</td>
<td>107,056</td>
<td>63,015</td>
<td>58.9</td>
<td>61,179</td>
<td>57.1</td>
<td>6,260</td>
<td>54,919</td>
<td>1,834</td>
<td>2.9</td>
</tr>
<tr>
<td>1954</td>
<td>108,321</td>
<td>63,643</td>
<td>58.8</td>
<td>60,109</td>
<td>55.5</td>
<td>6,205</td>
<td>53,904</td>
<td>3,532</td>
<td>5.5</td>
</tr>
<tr>
<td>1955</td>
<td>109,683</td>
<td>65,023</td>
<td>59.3</td>
<td>62,170</td>
<td>56.7</td>
<td>6,450</td>
<td>55,722</td>
<td>2,852</td>
<td>4.4</td>
</tr>
<tr>
<td>1956</td>
<td>110,954</td>
<td>66,552</td>
<td>60.0</td>
<td>63,799</td>
<td>57.5</td>
<td>6,283</td>
<td>57,514</td>
<td>2,750</td>
<td>4.1</td>
</tr>
<tr>
<td>1957</td>
<td>112,265</td>
<td>66,929</td>
<td>59.6</td>
<td>64,071</td>
<td>57.1</td>
<td>5,947</td>
<td>58,123</td>
<td>2,859</td>
<td>4.3</td>
</tr>
<tr>
<td>1958</td>
<td>113,727</td>
<td>67,639</td>
<td>59.5</td>
<td>63,036</td>
<td>55.4</td>
<td>5,586</td>
<td>57,450</td>
<td>4,602</td>
<td>6.8</td>
</tr>
<tr>
<td>1959</td>
<td>115,329</td>
<td>68,369</td>
<td>59.3</td>
<td>64,630</td>
<td>56.0</td>
<td>5,565</td>
<td>59,065</td>
<td>3,740</td>
<td>5.5</td>
</tr>
</tbody>
</table>
A Solution: Containerization
Cargo loading ~1955
Manual, Slow, Costly
(and Dangerous)
Containerization
The impact of containerisation

<table>
<thead>
<tr>
<th>year</th>
<th>vessel</th>
<th>loadingCostPerTon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1956</td>
<td>SS Ideal-X</td>
<td>0.16</td>
</tr>
<tr>
<td>1956</td>
<td>medium break-bulk</td>
<td>5.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>year</th>
<th>tonsPerManHour</th>
<th>hoursInPort</th>
</tr>
</thead>
<tbody>
<tr>
<td>1959</td>
<td>0.627</td>
<td>504</td>
</tr>
<tr>
<td>1976</td>
<td>4234</td>
<td>18</td>
</tr>
</tbody>
</table>
Data is Shipping Pre-Containerization
Data Packages
http://data.okfn.org
Data Containerization

Standards
- Standards (a few, simple ones)

Tools (primarily for integration)

Documentation

Datasets

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

datapackage.json

{
  "name": "gold-prices",
  "title": "Gold Prices (Monthly in USD)",
  "resources": [{
    "path": "data.csv",
    "format": "csv",
  }],
}
Key Principles

1. Narrow Focus
One (small) part of the data chain.

2. Web Oriented
Build for the web. Use formats that are web "native" (JSON) and work naturally with HTTP (e.g. CSV streaming, CSV plain text nature).

3. User Focused / Existing Tools
Work as easily as possible with existing tools. E.g. everyone has tools to use CSV and it's supported by almost every language.

4. Open
Anyone should be able to freely and openly use and reuse what we build. Our community is open to everyone.

5. Simplicity
Use most basic formats. Add the minimum, do the least needed. Reflected in all of the above.
3 minor "innovations"

- In formats - data package, tabular data package (both based heavily on pre-existing approaches e.g. CommonJS, DublinCore)
- In process - packaging as a distinct step
- Of approach - stick closely to existing tooling - prioritize integration
Data Package

↑

Tabular Data Package

Specialization
[Tabular] Data Packages
<table>
<thead>
<tr>
<th>year</th>
<th>admin1</th>
<th>admin2</th>
<th>admin3</th>
<th>admin4</th>
<th>admin5</th>
<th>admin6</th>
<th>func1</th>
<th>func2</th>
<th>econ1</th>
<th>econ2</th>
<th>fin_source</th>
<th>exp_type</th>
<th>transfer</th>
<th>approved</th>
<th>adjusted</th>
<th>executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>111 Remuneration</td>
<td>111.00 Remuneration of Base Personnel Excluding Base Personnel Excluding</td>
<td>30269300</td>
<td>30269300</td>
<td>27498186</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>112 Mandatory state</td>
<td>112.00 Mandatory state</td>
<td>5564000</td>
<td>5564000</td>
<td>5410121</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Office supplies</td>
<td>113.03 Office supplies, Base Non-Excluding</td>
<td>1557700</td>
<td>1262700</td>
<td>1117470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Office supplies</td>
<td>113.03 Office supplies, Special Non-Excluding</td>
<td>15000</td>
<td>15000</td>
<td>5282.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Books and Drugs</td>
<td>113.10 Books and Drugs Non-Excluding</td>
<td>4400</td>
<td>4400</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Telecommunication</td>
<td>113.11 Telecommunication Non-Excluding</td>
<td>1235760</td>
<td>952600</td>
<td>783776.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Transportation</td>
<td>113.13 Transportation Non-Excluding</td>
<td>130000</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Current</td>
<td>113.17 Current Non-Excluding</td>
<td>650000</td>
<td>35000</td>
<td>21689</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Current repair</td>
<td>113.18 Current repair of Base Non-Excluding</td>
<td>239000</td>
<td>183000</td>
<td>88835.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Use of government</td>
<td>113.20 Use of government Base Non-Excluding</td>
<td>35000</td>
<td>65000</td>
<td>55812</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Trainings</td>
<td>113.21 Trainings Base Non-Excluding</td>
<td>300000</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Printing services</td>
<td>113.22 Printing services Base Non-Excluding</td>
<td>189300</td>
<td>114300</td>
<td>114204</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Representation costs</td>
<td>113.23 Representation costs Base Non-Excluding</td>
<td>161000</td>
<td>810000</td>
<td>335558.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for IT and PC works</td>
<td>113.30 IT and PC works Non-Excluding</td>
<td>120000</td>
<td>20000</td>
<td>14373.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Compensation for</td>
<td>113.38 Compensation for Base Non-Excluding</td>
<td>430500</td>
<td>430500</td>
<td>193208</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Compensation for</td>
<td>113.39 Compensation for Base Non-Excluding</td>
<td>672000</td>
<td>272000</td>
<td>236280</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Payment for Goods and services</td>
<td>113.45 Goods and services Base Non-Excluding</td>
<td>50000</td>
<td>20000</td>
<td>15729.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>114 Duty travels</td>
<td>114.01 Trips within the Base Non-Excluding</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>114 Duty travels</td>
<td>114.02 Trips abroad Base Non-Excluding</td>
<td>2124800</td>
<td>853200</td>
<td>428432.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>116 Employer’s</td>
<td>116.01 Employer’s share of Pay Base Personnel Excluding</td>
<td>848700</td>
<td>848700</td>
<td>813988.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>113 Transfers to</td>
<td>113.20 Other transfers made Base Non-Excluding</td>
<td>6199300</td>
<td>5899300</td>
<td>561770</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>242 Purchase of</td>
<td>242.00 Purchase of fixed Base Capital Excluding</td>
<td>1810100</td>
<td>114000</td>
<td>244052.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>243 Capital repair</td>
<td>243.04 Capital refurbishment Base Capital Excluding</td>
<td>600000</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Parliament</td>
<td>010 Central apparatus</td>
<td>01</td>
<td>01:01</td>
<td>243 Capital repair</td>
<td>243.06 Capital refurbishment Base Capital Excluding</td>
<td>100000</td>
<td>28300</td>
<td>28228</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Directorate</td>
<td>218 Building</td>
<td>01</td>
<td>01:01</td>
<td>111 Remuneration</td>
<td>111.00 Remuneration of Base Personnel Excluding Base Personnel Excluding</td>
<td>2466000</td>
<td>2683500</td>
<td>2683489.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Directorate</td>
<td>218 Building</td>
<td>01</td>
<td>01:01</td>
<td>111 Remuneration</td>
<td>111.00 Remuneration of Special Personnel Excluding Special Personnel Excluding</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Directorate</td>
<td>218 Building</td>
<td>01</td>
<td>01:01</td>
<td>112 Mandatory state</td>
<td>112.00 Mandatory state Base Personnel Excluding</td>
<td>488500</td>
<td>521000</td>
<td>521021.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>01:01</td>
<td>Directorate</td>
<td>218 Building</td>
<td>01</td>
<td>01:01</td>
<td>112 Mandatory state</td>
<td>112.00 Mandatory state Special Personnel Excluding</td>
<td>2300</td>
<td>2300</td>
<td>2300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The best slides may have little or no text at all.
It’s about the tools, stupid!
Tools

- **Publish and manage data packages**
  - Online services and offline libraries to get, create, manipulate and share data packages

- **Data package support in existing tooling**
  - Spreadsheets - Excel, Google Docs ...
  - DaaS - SQL, BigQuery, Redshift, ...
  - End-user tools - R, MatLab, ...
  - Code: Python, Ruby, Java, ...
Data Packagist

Create, edit, and validate Data Packages.

DataPackage

DataPackage is a format and specification for convenient delivery, installation and management of data. This is the base profile for DataPackage.

Name*

my-data-package

A lower-case string with only alphanumeric characters along with '.', '_' or '-' characters.

Title

My Data Package

A human-readable title for this package.

Description

http://datapackagist.okfnlabs.org/
Invalid

A total of 29 rows of data were processed.

Below we have a sample of 2 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. **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 family, Entity, Date, Expense type, Expense area, Supplier, Transaction number, Amount, Description, Supplier postcode, Supplier type, Contract number, Project code, Expenditure type, VAT registration number.

   **HEADERS**
   - Entity
   - Expense Area
   - Payment Date
   - Payment Number
   - Supplier Name
   - Amount
   - NAC Description
   - Service Supplied

2. **Incorrect Type**: The value is not a valid String.

   **Required Field**: Column Entity is a required field, but no value can be found in row 28.

   **ROW**
   - Entity
   - Expense Area
   - Payment Date
   - Payment Number
   - Supplier Name
   - Amount
   - NAC Description
   - Service Supplied
   
   28. Commission on Devolution Spend Incurred by Silk Commission

These results were calculated on Wed Aug 12 at 12:54 PM.


Schema: [View](https://raw.githubusercontent.com/okfn/tabular-validator/master/examples/nmt/spend-publishing-schema.json)

“Data Package Standards”

Data Package

Data Package Tabular Profile (#tabular)
Data Package Geo Profile (#geo)
More ...

Community

Docs, Outreach, Events, ...

Tooling and Integration

Publish and Use [profile specific] Data Packages

Core Tools
- Command Line Tool
- Web Tool
- JS / Python / Ruby libs
- Validation
- Registry

Integration
- R / Excel / ...
- DBs, DaaS (Bigquery etc)
- OpenRefine, Dat etc
- Data API
- ...

Conclusion
Conclusion

- Containerization analogy
  - Generic across disciplines
  - Focused
  - Simple
- Data Package and Tabular Data Package more
  - Zen-like Simplicity
  - User and tool obsessed
- Mature, based on existing e.g. software patterns
  - Rapidly growing set of tools
  - Integration is key
- Active community
  - And we are looking for research groups to collaborate with on further pilots
Appendix
3 minor "innovations"
- In formats - data package, tabular data package (both based heavily on pre-existing approaches e.g. CommonJS, DublinCore)
- In process - packaging as a distinct step
- Of approach - stick closely to existing tooling - prioritize integration
“Standards”

Data Containerization

Connect different tools with a lightweight common interchange format
Standards
- Standards (a few, simple ones)
- Tools (primarily for integration)
- Documentation
- Datasets

3 minor "innovations"
- In formats - data package, tabular data package (both based heavily on pre-existing approaches e.g. CommonJS, DublinCore)
- In process - packaging as a distinct step
- Of approach - stick closely to existing tooling - prioritize integration
Data package: minimal wrapping, agnostic about the data its "packaging", designed for extension. Very flexible: can be used as a transport for any kind of data BUT limits integration and tooling. [http://data.okfn.org/standards/data-package/](http://data.okfn.org/standards/data-package/)

Tools

http://data.okfn.org/tools/

• Publish and manage data packages
  ○ Online services and offline libraries to get, create, manipulate and share data packages

• Data package support in existing tooling
  ○ Spreadsheets - Excel, Google Docs ...
  ○ DaaS - BigQuery, Redshift, ...
  ○ End-user tools - R, Tableau, ...
  ○ Code: Python, Ruby, Java, ...
Key Principles

1. Narrow Focus
   One (small) part of the data chain. One specific feature (packaging) or type of data (simple tabular)

2. Web Oriented
   Build for the web. Use formats that are web "native" (JSON) and work naturally with HTTP (e.g. CSV streaming, CSV plain text nature). Distributed not centralized.

3. Existing Tools
   Work as easily as possible with existing tools. E.g. everyone has tools to use CSV and it's supported by almost every language.

4. Minimal, Clean, Sufficient Structure
   Only most essential metadata (but allow for extensibility). Schema information. No blank lines at the top of the table, no footnotes inlined in rows at the bottom.

5. Simplicity
   Use most basic formats. Add the minimum, do the least needed. Reflected in all of the above.
“Data Package Standards”

Data Package

Data Package Tabular Profile (#tabular)
Data Package Geo Profile (#geo)
More ...

Community

Docs, Outreach, Events, ...

Tooling and Integration

Publish and Use [profile specific] Data Packages

Core Tools
- Command Line Tool
- Web Tool
- JS / Python / Ruby libs
- Validation
- Registry

Integration
- R / Excel / ...
- DBs, DaaS (Bigquery etc)
- OpenRefine, Dat etc
- Data API
- ...