Introduction to Data Packages

PRESENTED BY

Rufus Pollock
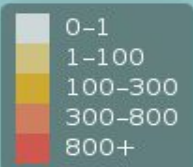
OPEN KNOWLEDGE

HDX Context

HDX Beta

Q ebola

DATA | LOCATIONS | ORGANISATIONS | CRISES ▼

+
-

Cumulative Cases of Ebola
Hover over a country/district

	0–1
	1–100
	100–300
	300–800
	800+

West Africa: Ebola Outbreak

Cumulative Cases of Ebola

27,987

WHO
Data - Explore - Aug 17, 2015

Cumulative Deaths from Ebola

11,299

WHO
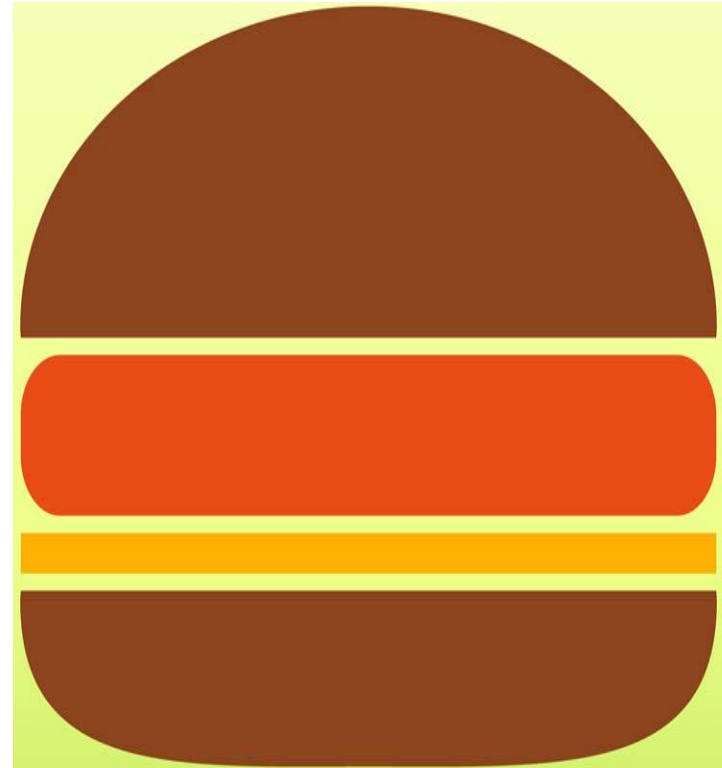Data - Explore - Aug 17, 2015

Response Plan Coverage

68.7%
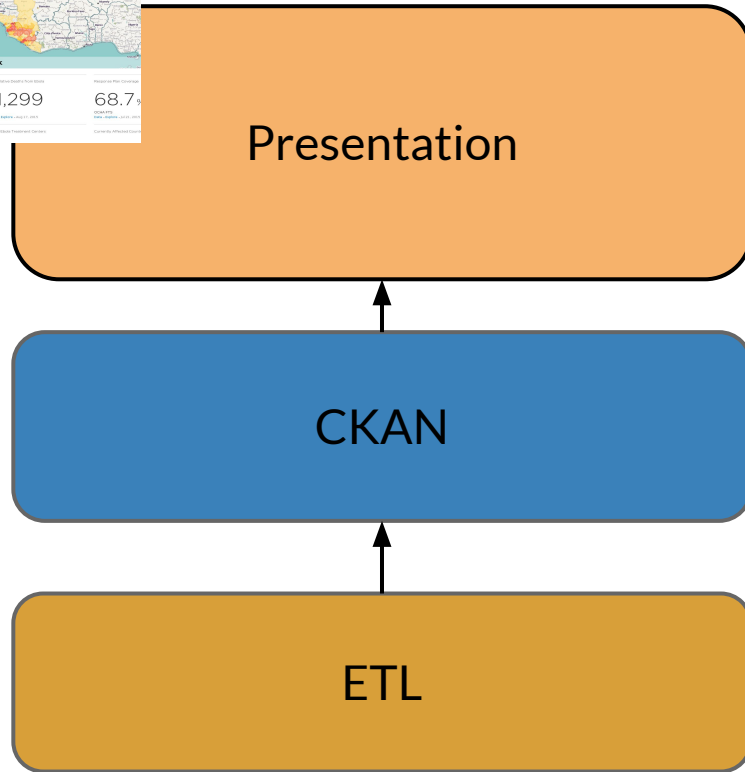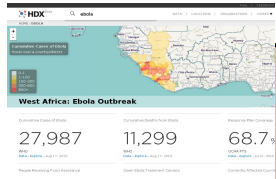
OCHA FTS
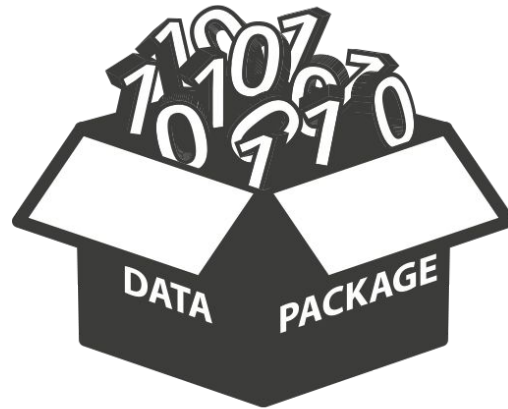Data - Explore - Jul 21, 2015

People Receiving Food Assistance

Open Ebola Treatment Centers

Currently Affected Countr

OPEN KNOWLEDGE

Hamburger Analogy

Presentation
↑
CKAN
↑
ETL

OPEN KNOWLEDGE

	year	admin1	admin2	admin3	admin4	admin5	admin6	func1	func2	econ1	econ2	fin_source	exp_type	transfer	approved	adjusted	executed
0	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	111 Remuneration	111.00 Remuneration of	Base	Personnel	Excluding	30269300	30269300	27849186
1	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	112 Mandatory state	112.00 Mandatory state	Base	Personnel	Excluding	5564000	5564000	5401021
2	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.03 Office supplies,	Base	Non-	Excluding	1557700	1262700	1115740
3	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.03 Office supplies,	Special	Non-	Excluding	15000	15000	5282.72
4	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.06 Books and	Base	Non-	Excluding	150800	120800	92372.12
5	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.10 Drugs and	Base	Non-	Excluding	4400	400	0
6	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.11 Telecommunication	Base	Non-	Excluding	1257600	952600	763776.13
7	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.13 Transportation	Base	Non-	Excluding	130000	0	0
8	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.17 Current	Base	Non-	Excluding	650000	35000	21689
9	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.18 Current repair of	Base	Non-	Excluding	239000	163000	86035.64
10	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.20 Use of government	Base	Non-	Excluding	35000	65000	56812
11	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.21 Trainings	Base	Non-	Excluding	300000	0	0
12	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.22 Printing services	Base	Non-	Excluding	189300	114300	114204
13	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.23 Representation costs	Base	Non-	Excluding	1610000	810000	335058.53
14	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.30 IT and PC works	Base	Non-	Excluding	120000	20000	14373.88
15	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.38 Compensation for	Base	Non-	Excluding	430500	430500	193208
16	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.39 Compensation for	Base	Non-	Excluding	672000	272000	236280
17	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.45 Goods and services	Base	Non-	Excluding	50000	20000	15729.51
18	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	114 Duty travels	114.01 Travels within the	Base	Non-	Excluding	1000	0	0
19	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	114 Duty travels	114.02 Travels abroad	Base	Non-	Excluding	2124800	933200	429432.75
20	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	116 Employer's	116.01 Employer's share of	Base	Personnel	Excluding	846700	846700	813966.06
21	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	135 Transfers to	135.25 Other transfers made	Base	Non-	Excluding	6199300	5899300	5601770
22	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	242 Purchase of	242.00 Purchase of fixed	Base	Capital	Excluding	1810100	1140100	244455.59
23	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	243 Capital repair	243.04 Capital refurbishment	Base	Capital	Excluding	600000	0	0
24	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	243 Capital repair	243.06 Capital refurbishment	Base	Capital	Excluding	1000000	28300	28228
25	2009	Central	101	1001 Directorate			218 Building	01	01.01	111 Remuneration	111.00 Remuneration of	Base	Personnel	Excluding	2466000	2693500	2693496.5
26	2009	Central	101	1001 Directorate			218 Building	01	01.01	111 Remuneration	111.00 Remuneration of	Special	Personnel	Excluding	10000	10000	10000
27	2009	Central	101	1001 Directorate			218 Building	01	01.01	112 Mandatory state	112.00 Mandatory state	Base	Personnel	Excluding	486600	521100	521021.16
28	2009	Central	101	1001 Directorate			218 Building	01	01.01	112 Mandatory state	112.00 Mandatory state	Special	Personnel	Excluding	2300	2300	2300

Tabular Data Package

	year	admin1	admin2	admin3	admin4	admin5	admin6	func1	func2	econ1	econ2	fin_source	exp_type	transfer	approved	adjusted	executed
0	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	111 Remuneration	111.00 Remuneration of	Base	Personnel	Excluding	30269300	30269300	27849186
1	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	112 Mandatory state	112.00 Mandatory state	Base	Personnel	Excluding	5564000	5564000	5401021
2	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.03 Office supplies,	Base	Non-	Excluding	1557700	1262700	1115740
3	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.03 Office supplies,	Special	Non-	Excluding	15000	15000	5282.72
4	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.06 Books and	Base	Non-	Excluding	150800	120800	92372.12
5	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.10 Drugs and	Base	Non-	Excluding	4400	400	0
6	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.11 Telecommunication	Base	Non-	Excluding	1257600	952600	763776.13
7	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.13 Transportation	Base	Non-	Excluding	130000	0	0
8	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.17 Current	Base	Non-	Excluding	650000	35000	21689
9	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.18 Current repair of	Base	Non-	Excluding	239000	163000	86035.64
10	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.20 Use of government	Base	Non-	Excluding	35000	65000	56812
11	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.21 Trainings	Base	Non-	Excluding	300000	0	0
12	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.22 Printing services	Base	Non-	Excluding	189300	114300	114204
13	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.23 Representation costs	Base	Non-	Excluding	1610000	810000	335058.53
14	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.30 IT and PC works	Base	Non-	Excluding	120000	20000	14373.88
15	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.38 Compensation for	Base	Non-	Excluding	430500	430500	193208
16	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.39 Compensation for	Base	Non-	Excluding	672000	272000	236280
17	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	113 Payment for	113.45 Goods and services	Base	Non-	Excluding	50000	20000	15729.51
18	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	114 Duty travels	114.01 Travels within the	Base	Non-	Excluding	1000	0	0
19	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	114 Duty travels	114.02 Travels abroad	Base	Non-	Excluding	2124800	933200	429432.75
20	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	116 Employer's	116.01 Employer's share of	Base	Personnel	Excluding	846700	846700	813966.06
21	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	135 Transfers to	135.25 Other transfers made	Base	Non-	Excluding	6199300	5899300	5601770
22	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	242 Purchase of	242.00 Purchase of fixed	Base	Capital	Excluding	1810100	1140100	244455.59
23	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	243 Capital repair	243.04 Capital refurbishment	Base	Capital	Excluding	600000	0	0
24	2009	Central	101	0101 Parliament			010 Central apparatus	01	01.01	243 Capital repair	243.06 Capital refurbishment	Base	Capital	Excluding	1000000	28300	28228
25	2009	Central	101	1001 Directorate			218 Building	01	01.01	111 Remuneration	111.00 Remuneration of	Base	Personnel	Excluding	2466000	2693500	2693496.5
26	2009	Central	101	1001 Directorate			218 Building	01	01.01	111 Remuneration	111.00 Remuneration of	Special	Personnel	Excluding	10000	10000	10000
27	2009	Central	101	1001 Directorate			218 Building	01	01.01	112 Mandatory state	112.00 Mandatory state	Base	Personnel	Excluding	486600	521100	521021.16
28	2009	Central	101	1001 Directorate			218 Building	01	01.01	112 Mandatory state	112.00 Mandatory state	Special	Personnel	Excluding	2300	2300	2300

datapackage.json

- metadata
- csv schema

CSV

The best slides may have little or no text at all

OPEN KNOWLEDGE

Data Packagist

Create, edit, and validate Data Packages.

Select Profile   Data Package ⬍

Download   Uploa

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.
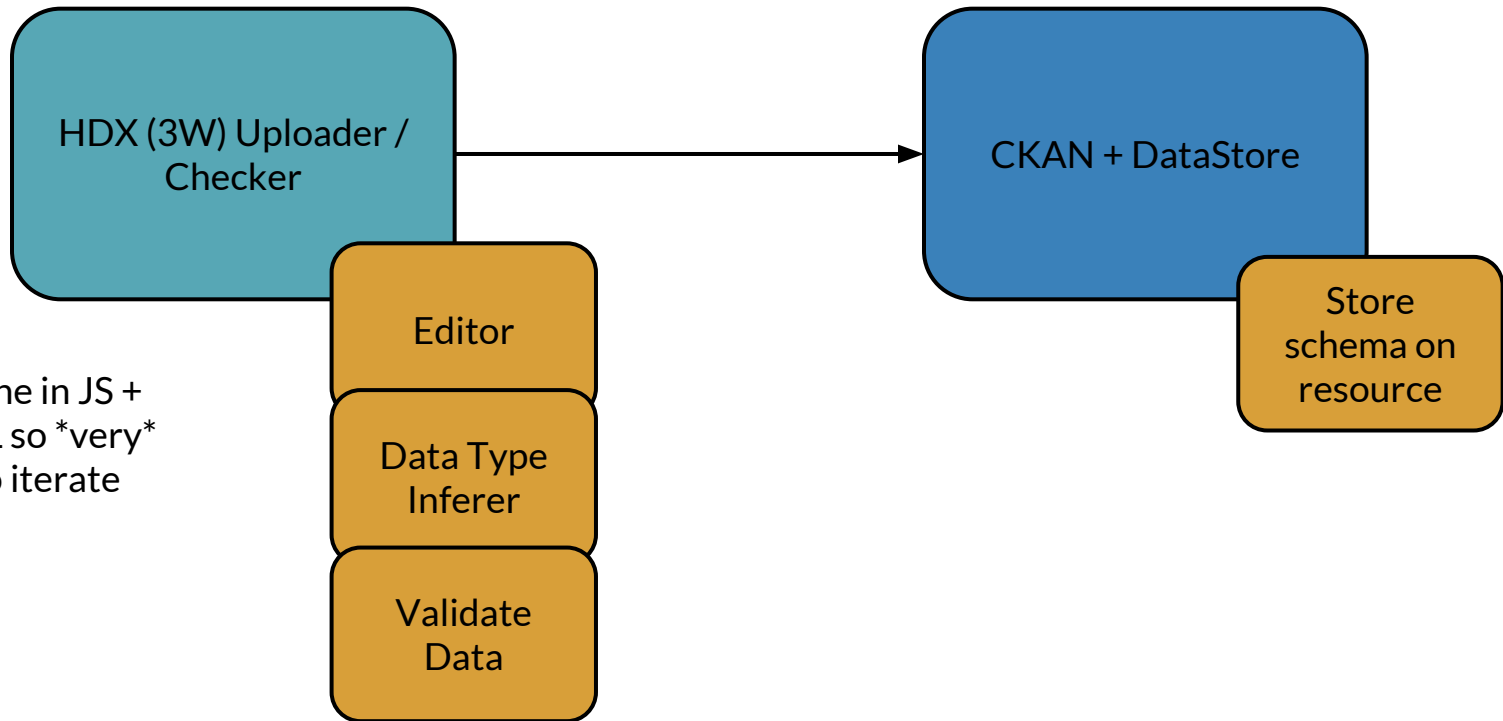
Title

My Data Package

A human-readable title for this package.

Description

```
{
  "name": "my-data-package",
  "title": "My Data Package"
}
```

http://datapackagist.okfnlabs.org/

OPEN KNOWLEDGE

HDX (3W) Uploader / Checker

CKAN + DataStore

Editor

Data Type Inferer

Validate Data

Store schema on resource

All done in JS + HTML so *very* fast to iterate

OPEN KNOWLEDGE

Appendix

Data Containerization

Cargo Transport ~ 1955

Manual, Slow, Costly
(and Dangerous)

Containerization

Data is Shipping Pre-Containerization

OPEN KNOWLEDGE

"Standards"

Data Containerization

Connect different tools with a **lightweight** common interchange format
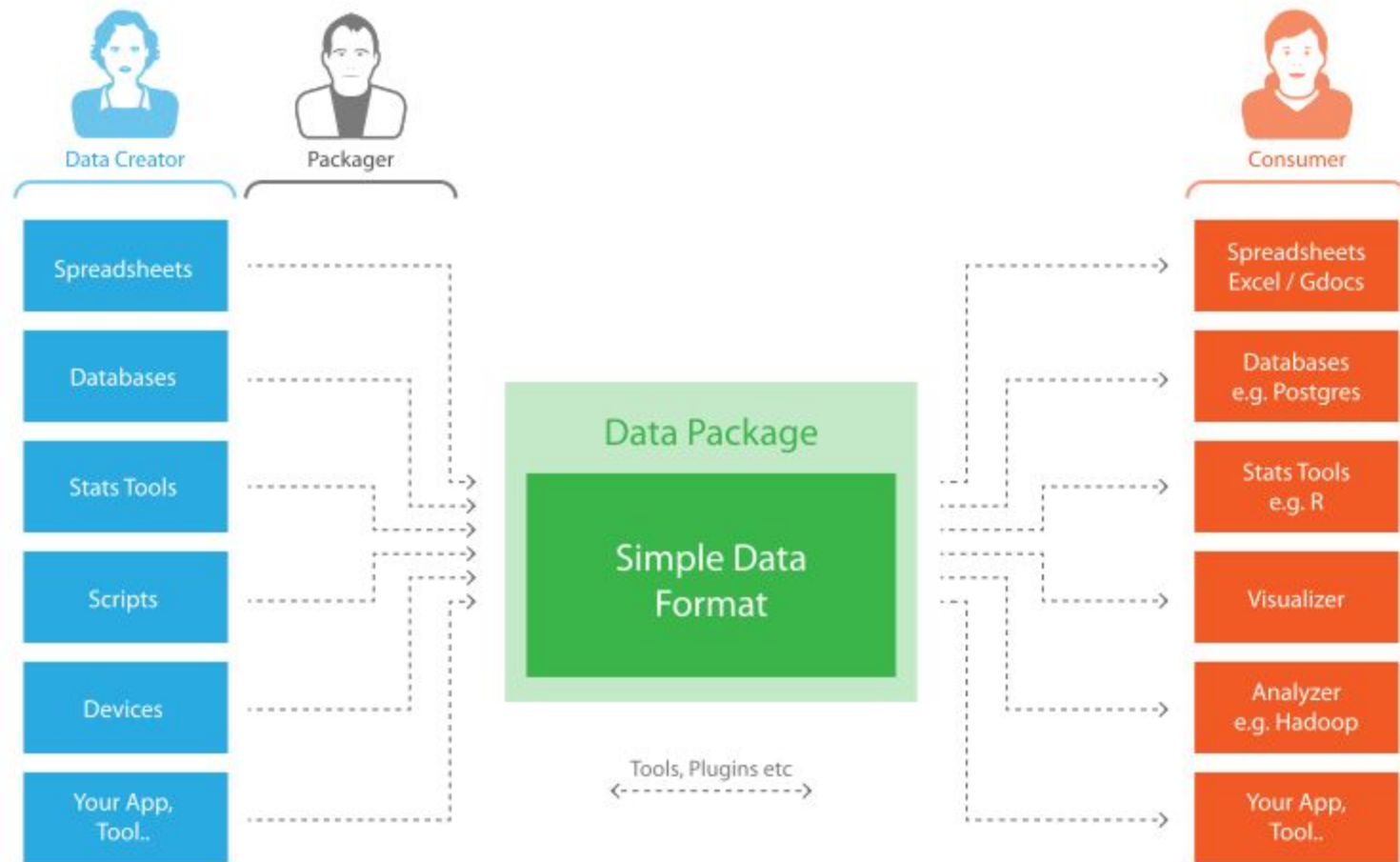
3 minor "innovations"

- 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 "Standards"
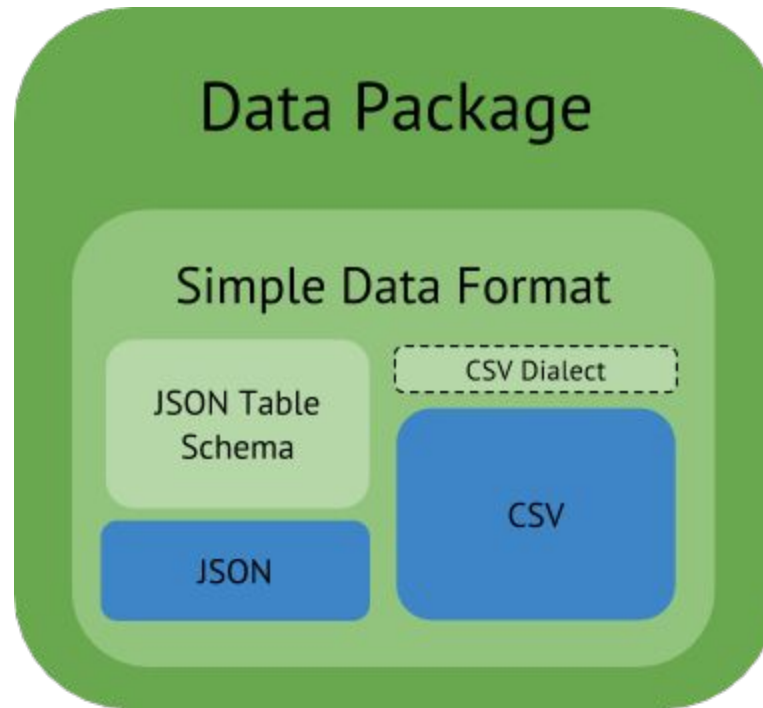
Data package: minimal wrapping, agnostic about the data its "packaging", designed for extension. Very flexible: can be used as a transport for any kind of data BUT limits integration and tooling. http://data.okfn.org/standards/data-package/

Tabular Data Package: tabular data only. Extends data package but requires data to be "good" CSVs & provides schema in simple JSON structure.
http://data.okfn.org/doc/tabular-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. The Web

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 its 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 ...

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
- ...

Community

Docs, Outreach, Events, ...