Open Data in Practice

http://training.theodi.org/InPractice

David Tarrant · @davetaz

Session 3

Data Publication Platforms

Specialist Solution	Integrated Solution
+ Easy to get setup and maintain. + Open Data focused + Clear workflows for publishing open data + Visualisation tools + Data mashing tools + Best for transactional data	+ No new platform to learn + Data is provided in parallel to web pages + No separation from authoritative data + Easy discovery of data + Best for reference data + Best for Linked Open Data

Key characteristics of specialist solution

1. Separate from your main org website

2. Designed to publish open data, not to fulfill other org goals

Key characteristics of integrated solution

1. It is your main website

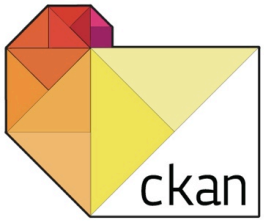
2. Publishes data alongside everything else that the org does

Merging specialist and integrated

Method 1: *Build the functionality of your current website into a new open data platform.*

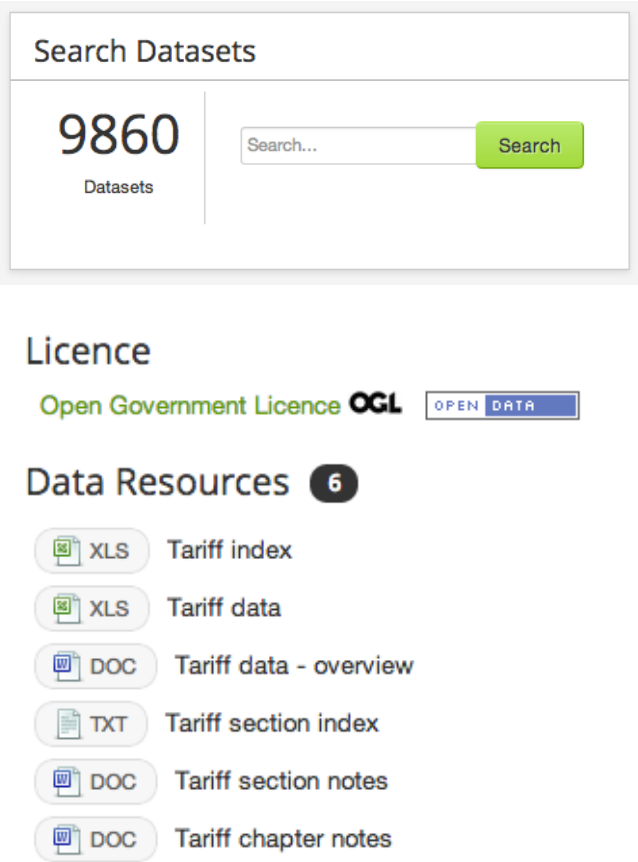
Method 2: *Hide the specialist solution behind your main website and use it as a loosely coupled CMS.*

Specialist Solutions



comprehensive knowledge
archive network

http://www.flickr.com/photos/okfn



ckan

Open Knowledge Foundation Supported

Data Catalogue

Open Source

Feels like a record manager

Simple API and search

http://demo.ckan.org/

Evolution of CKAN

Updated July 2014

Early) Dataset catalogue (data.gov.uk)

no data hosted or searched

Mid) Data and dataset catalogue

no data hosted but it is searchable

Now) Integrated data driven web site

data platform is integrated with data, search and content

Features…

Publish, Store and Manage Data and Metadata

Visual and Geospatial

Social

Full Stored History

Federate Your Data With Other Organizations

Rich RESTful JSON API for Developers

ckan

Socrata

Search

View Types

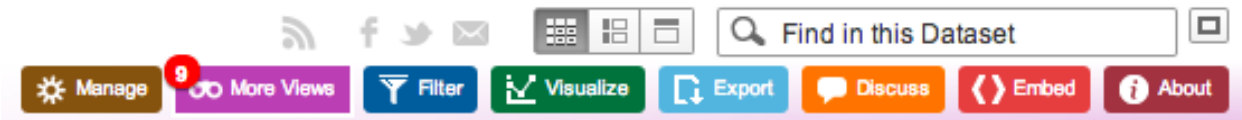
- Datasets
- Charts
- Maps
- Calendars
- Filtered Views
- External Datasets
- Files and Documents
- Forms
- APIs

Data as a Service (DaaS)

Hosted enterprise solution

Rich Interface

Powerful Dataset API (SODA)



https://opendata.socrata.com/

Data **Publishing**, Optimized for Business Users

Flexible **Metadata** Management

Federate Your Data With Other Organizations

Metrics of the Success of Your Initiative in Real-time

Anyone Can Create **Maps** and **Charts**

Data Becomes **Social**

Developers Are Supported Every Step of the Way

Socrata

Socrata

Search

View Types

- Datasets
- Charts
- Maps
- Calendars
- Filtered Views
- External Datasets
- Files and Documents
- Forms
- APIs

Data as a Service (DaaS)

Closed Source (main product)

Hosted Solution

US Based

Clean Interface

Powerful API (SODA)

Proprietary API

Find in this Dataset

Manage 9 More Views Filter Visualize Export Discuss Embed About

ckan

Comparison

History

- Publish
- Store
- Manage
- Metadata
- Visual

- Geospatial
- Social
- Federate
- Developers

Metrics

Socrata

Comparison

ckan

- Manages Metadata
- Points to the Authoritative Data
- Loosely Coupled

Socrata — The Data Experience Company™

- Manages the Data Itself
- Issues Identifiers
- Manages Schemas

CKAN you can install on one machine and have it catalogue large files. You will need more power for the data search. CKAN can store billions of records but not process them.

Socrata will index all the data, allowing it to be visualised and processed in the platform quickly. A visulisation of 6 Million rows can be done in seconds. This is because Socrata uses upwards of 15 different distributed services to run the entire platform, more that 40 bits of software tied together.

Integrated solutions

Integrated solutions expose data using the current infrastructure (web pages).

Data driven web site

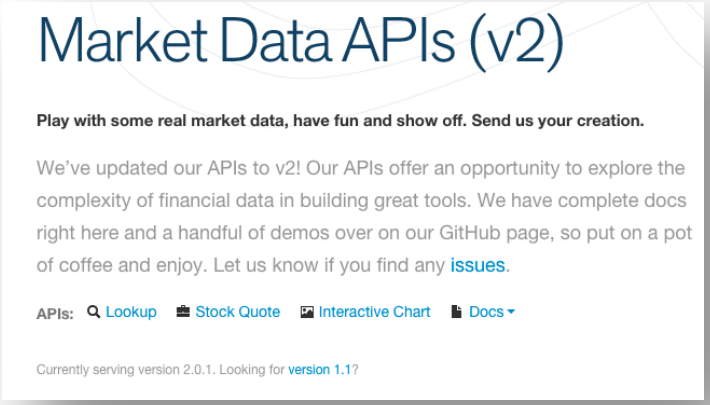
Best for reference and live data

Live data

Most commonly exposed via an API

transportAPI

powering change and innovation in transport



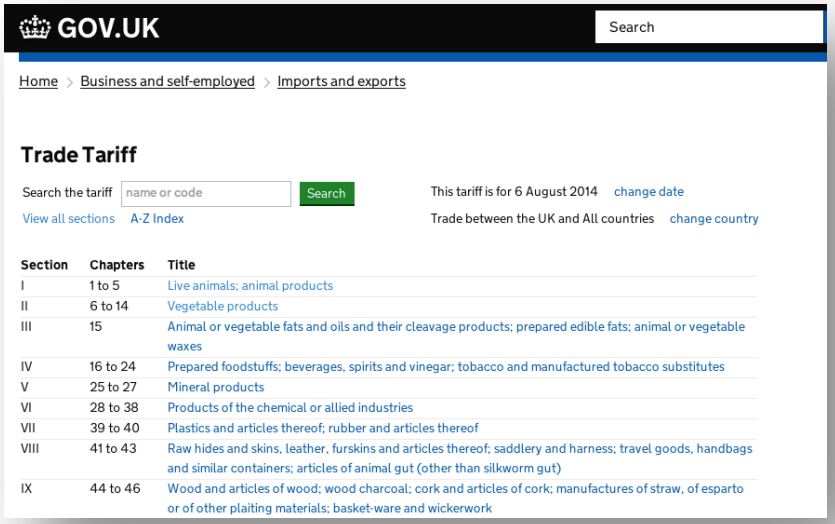
Reference data

Two methods:

- Make it a download (like transaction data).
- Embed the data in the page.

Integrated data example (download)

R 0-2 Tree Enabling



UK Trade Tariff

BBC Music and Programmes

Try using the following: .csv .json .xml .rss .rdf

Integrated data example (embedding)

Open Graph

DCAT

Hybrid approaches?

OPEN KNOWLEDGE

Tabular data package

The data hidden in the web

Guides

Marking up your dataset with DCAT

The Data Catalog Vocabulary (DCAT) defines a standard way to publish machine-readable metadata about a dataset.

The simplest way to publish a description of your dataset is to publish DCAT metadata using RDFa. RDFa allows machine-readable metadata to be embedded in a webpage. This means that publishing your dataset metadata can be easily achieved by updating the HTML for your dataset homepage.

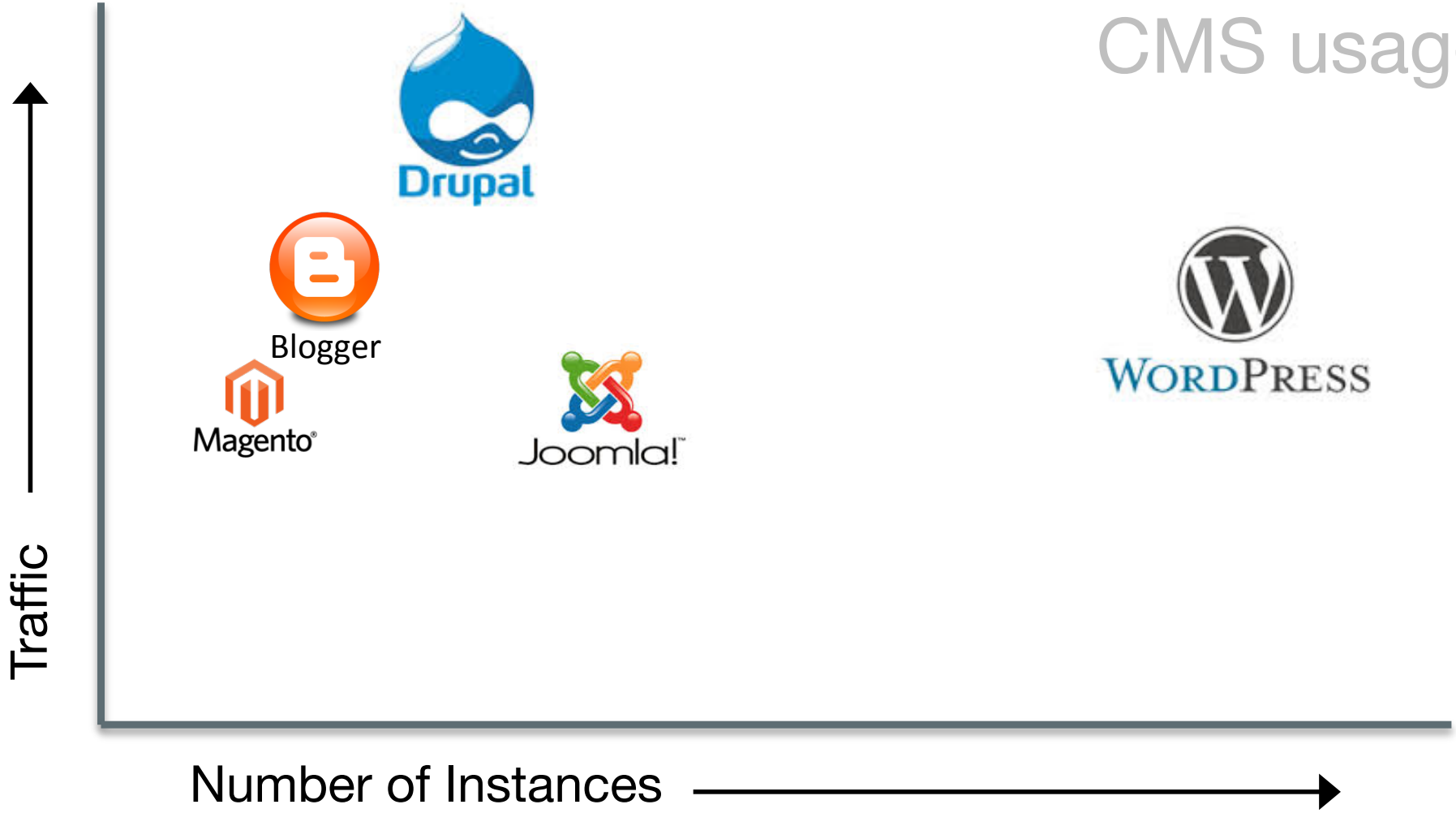
This guide provides a short introduction to publishing DCAT metadata using RDFa. For more advanced use cases, including publishing data in other formats, take a look at the official W3C documentation for DCAT. The RDFa primer may also be useful background reading.

The Open Data Certificates application supports reading DCAT published as RDFa. So as well as providing machine-readable metadata for data consumers, using DCAT will simplify the process of certifying your dataset as the application will be able to automatically populate some of the answers for you.

Getting started

The first thing to do is to let applications know that your web page is describing a dataset. To do this we need to declare the metadata schemas we will be using to describe the dataset and then indicate the type of thing being described.

CMS usage



Integrated data example
R/T 3 Tree/Net Enabling

London Waterloo: passenger departures

Due	Destination	Expected	Platform
12:27	London Waterloo	Starts here	4
12:28	Windsor & Eton Riverside	Starts here	16
12:30	Portsmouth Harbour	Starts here	14
12:33	Guildford	Starts here	3
12:33	London Waterloo	Starts here	17
12:35	Weymouth	Starts here	8
12:36	Hampton Court	Starts here	2
12:37	London Waterloo	Starts here	18
12:39	Guildford	Starts here	1
12:39	Poole	Starts here	11
12:42	Basingstoke	Starts here	10
12:42	Shepperton	Starts here	4
12:45	London Waterloo	Starts here	16
12:45	Portsmouth & Southsea	Starts here	12
12:46	Chessington South	Starts here	3
12:50	Reading	Starts here	19

UK Transport Data

Live traffic information from the Highways Agency

Published by Highways Agency. Licensed under OGL Open Government Licence.
Openness rating: ★★★☆☆

NII

Transport

Live traffic information data showing traffic information on the strategic road network in England, maintained by the Highways Agency.

Update: 12th August 2013 Following a change of supplier, the NTIS system is being re-developed and will eventually replace the legacy system. Please consult the updated document which describes the NTIS Legacy DATEX II v1.0 Publisher.

New services will be delivered in DATEX II v2 format using web services to push data to subscribers. Full details of the new services can be found in document NIS P TIH 008 available from the TIH website. Potential subscribers', Project Managers and engineers seeking to develop a new interface are...

Read More

DATA RESOURCES (11)

XML Current planned events

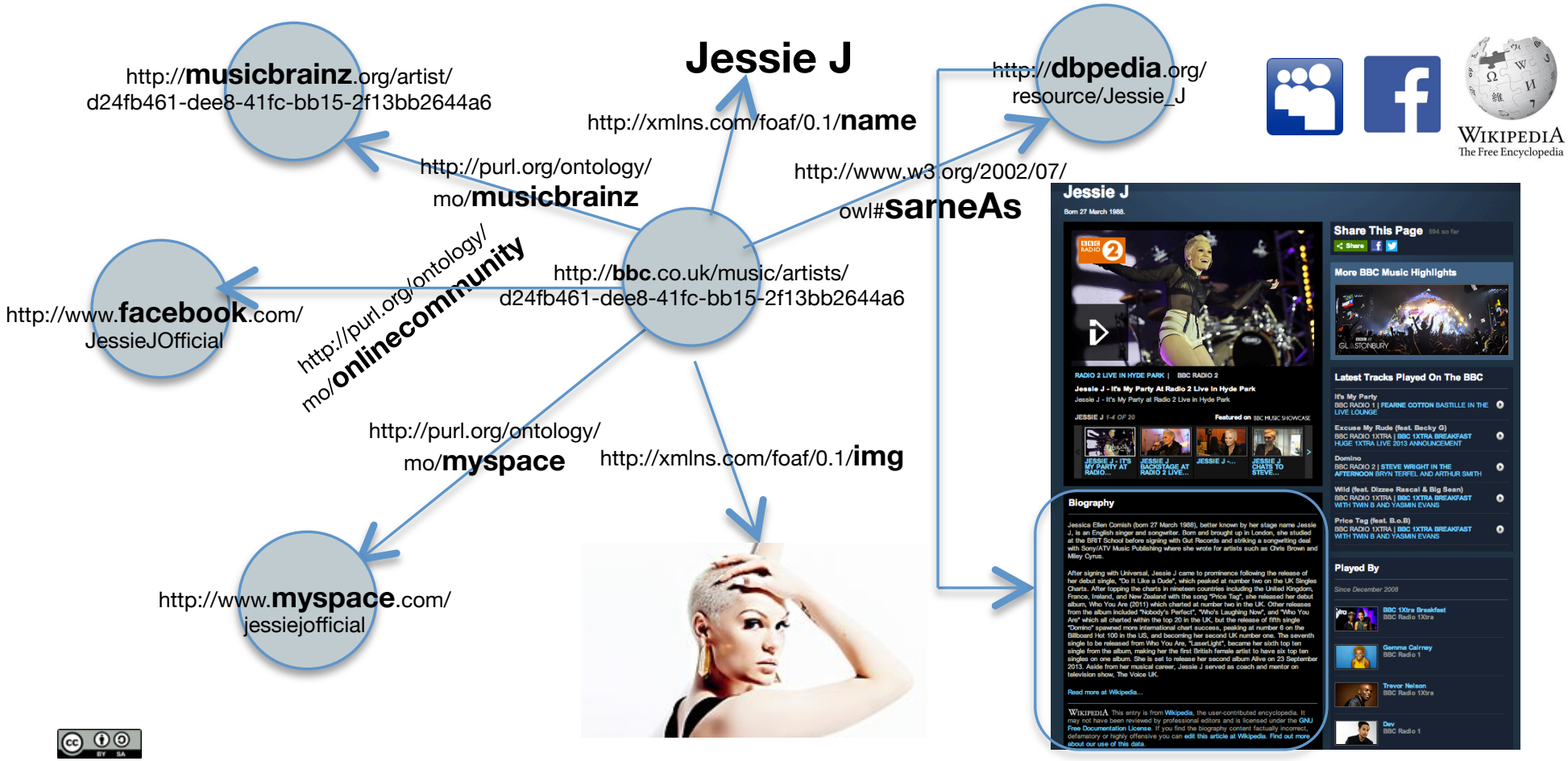
XML Future planned events

Highways agency data

Example

Doctor who series and episodes publishing in Github.

Linked data


Jessie J
http://musicbrainz.org/artist/
d24fb461-dee8-41fc-bb15-2f13bb2644a6
http://xmlns.com/foaf/0.1/name
http://dbpedia.org/
resource/Jessie_J
http://purl.org/ontology/
mo/musicbrainz
http://www.w3.org/2002/07/
owl#sameAs
http://www.facebook.com/
JessieJOfficial
http://purl.org/ontology/
mo/onlinecommunity
http://bbc.co.uk/music/artists/
d24fb461-dee8-41fc-bb15-2f13bb2644a6
http://purl.org/ontology/
mo/myspace
http://xmlns.com/foaf/0.1/img
http://www.myspace.com/
jessiejofficial
WIKIPEDIA
The Free Encyclopedia
Jessie J
Share This Page
More BBC Music Highlights
Latest Tracks Played On The BBC
Biography
Played By

Specialist Solution	Integrated Solution
+ Easy to get setup and maintain.	+ No new platform to learn
+ Open Data focused	+ Data is provided in parallel to web pages
+ Clear workflows for publishing open data	+ No separation from authoritative data
+ Visualisation tools	+ Easy discovery of data
+ Data mashing tools	+ Best for reference data
+ Best for transactional data	+ Best for Linked Open Data

Both great for open data

Integrated solutions more suited for building a web of linked data

Formats, Structures and Files

Data formats are complex and have ~~suffered~~ benefited from years of development in many different domains.

One domain is bringing them all together…

Session 4

Hands on: Publication

Task

Each team is to publish a high quality usable dataset.

1 each in:

- CKAN (demo.ckan.org)
- Github (http://theodi.github.io/data-publishing-template/)
- Socrata (https://demo.socrata.com/)

Sub tasks

- Create the entry for the dataset.
- Upload the dataset.
- Assess the usability using 5-star rating
- Create an Open Data Certificate and publish this for the dataset.
- *Improve the quality of the dataset.*

The dataset

http://data.gov.uk/dataset/financial-transactions-data-dfe

JULY 20xx

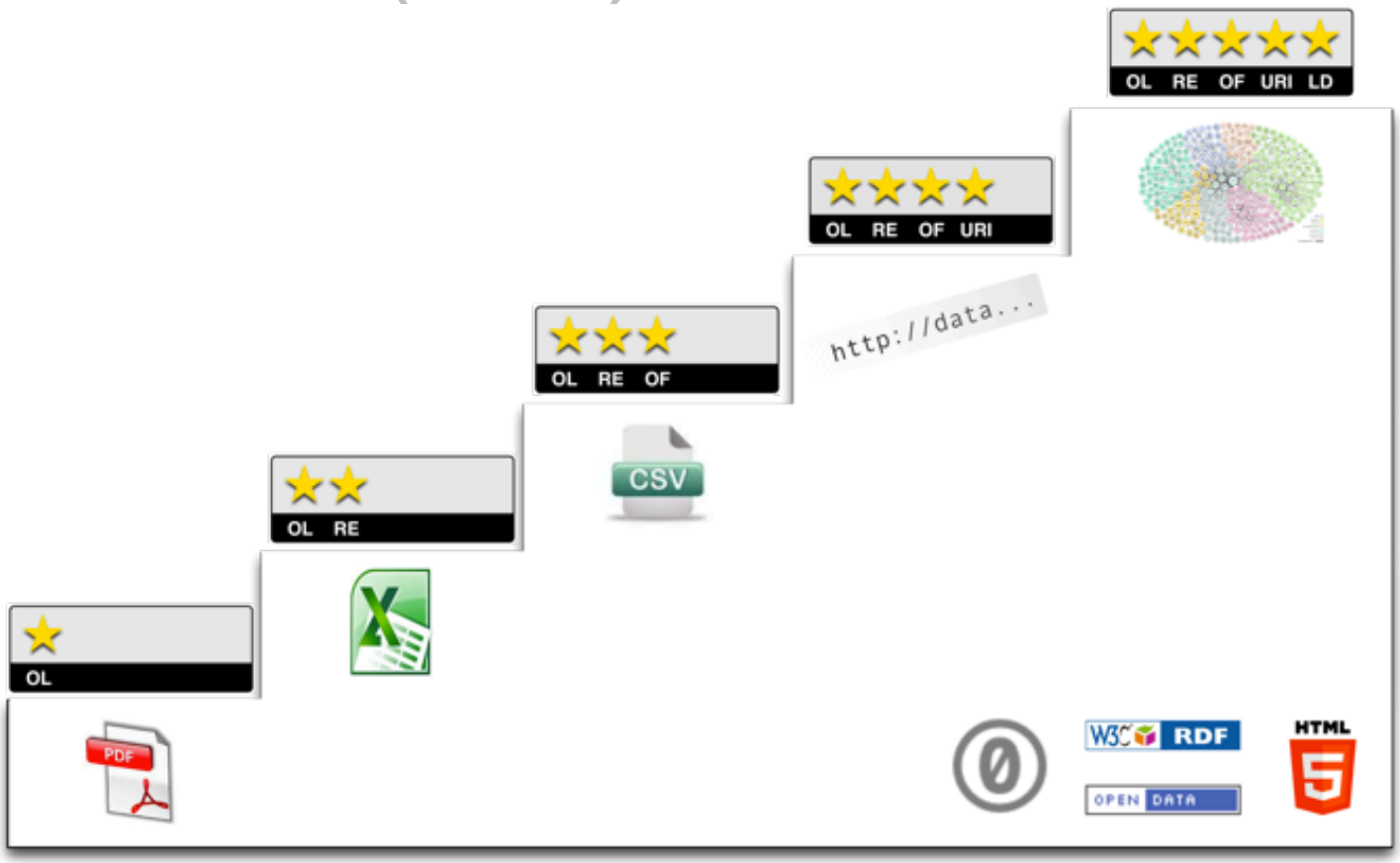
data only

By using this dataset we will need to conform to the terms of the license ☺

Session 4

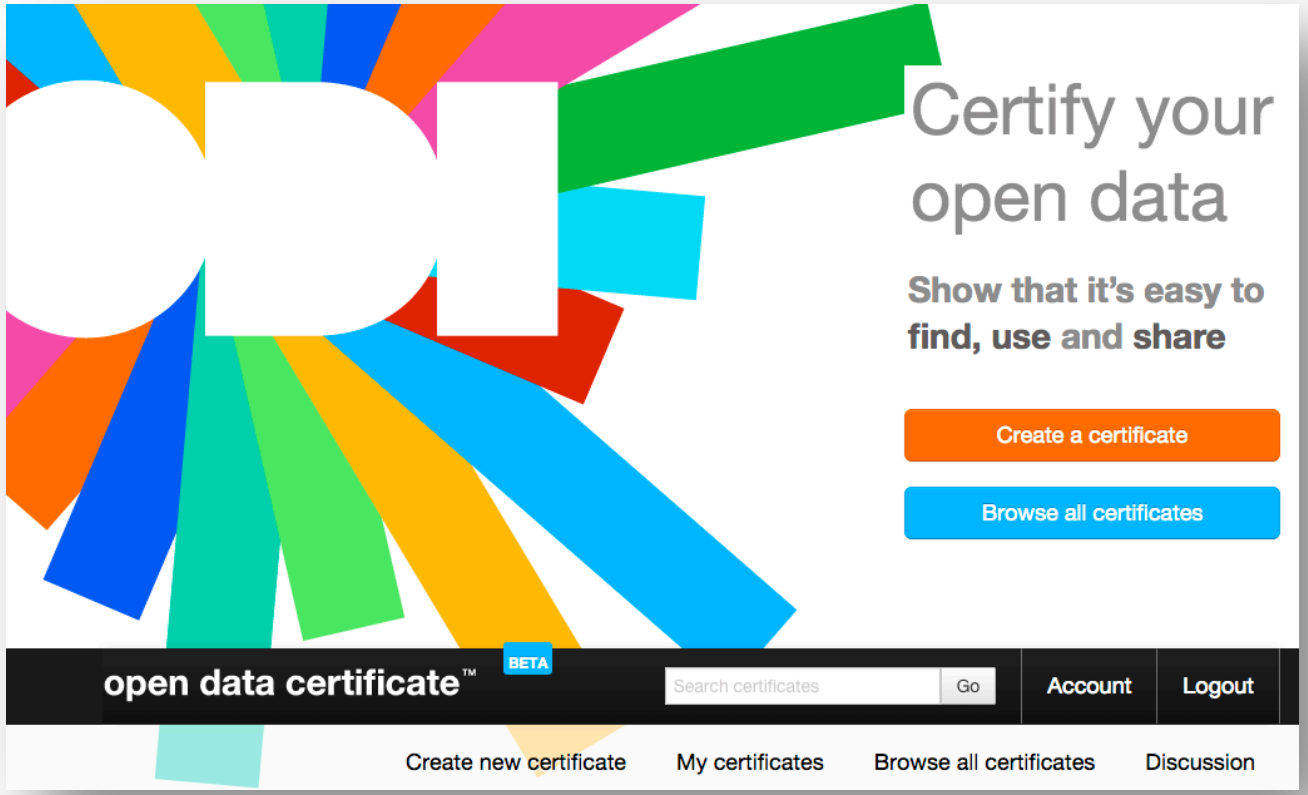
Tools and guides

5-Star Data (.info)



Open Data Certificates

certificates.theodi.org



CSVLint.io

CSV Lint

About | Recent validations | Recent schemas

Check your CSV files with CSVLint

CSV looks easy, but it can be hard to make a CSV file that other people can read easily.

CSVLint helps you to check that your CSV file is readable. And you can use it to check whether it contains the columns and types of values that it should.

Just enter the location of the file you want to check, or upload it. If you have a schema which describes the contents of the CSV file, you can also give its URL or upload it. Read more...

Enter a link to your CSV:

Enter URL +

Or upload a file:

Choose a file

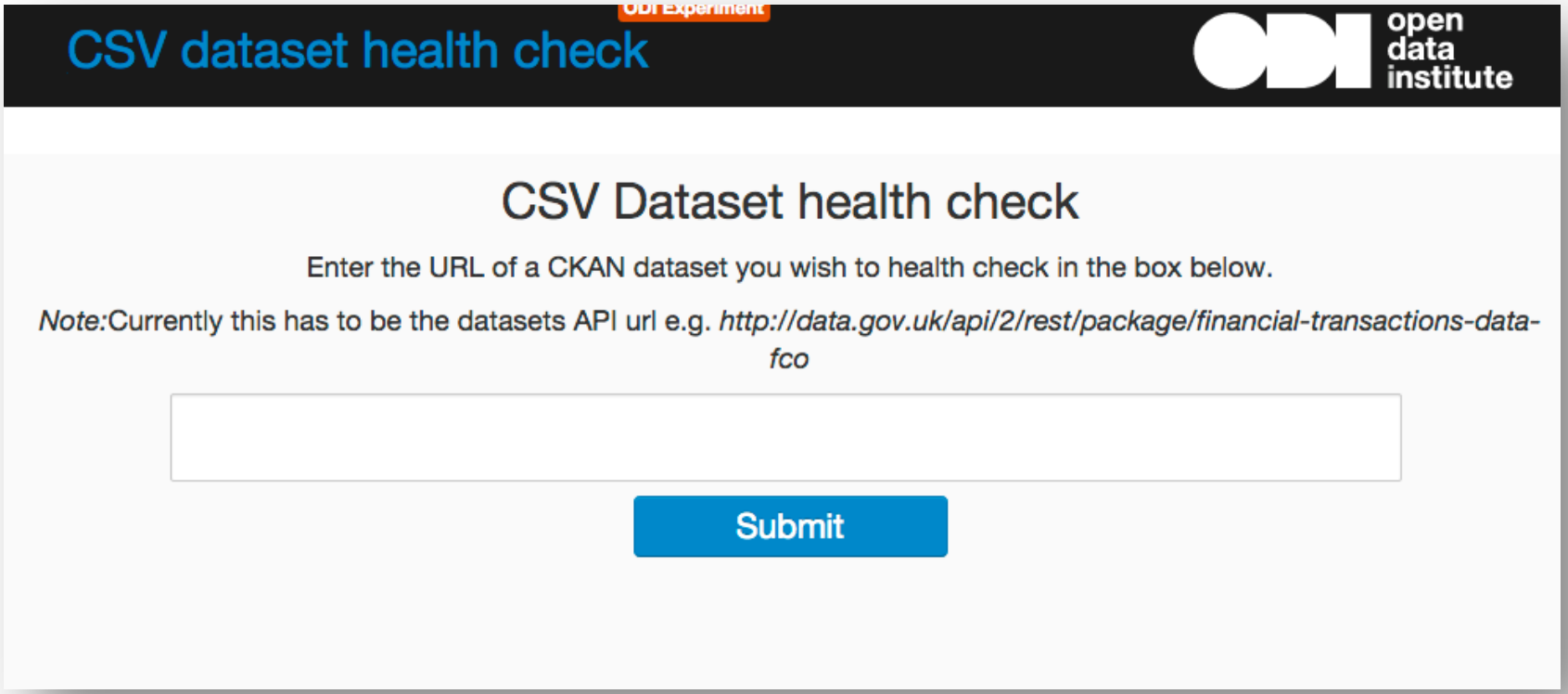
Add optional schema

Validate

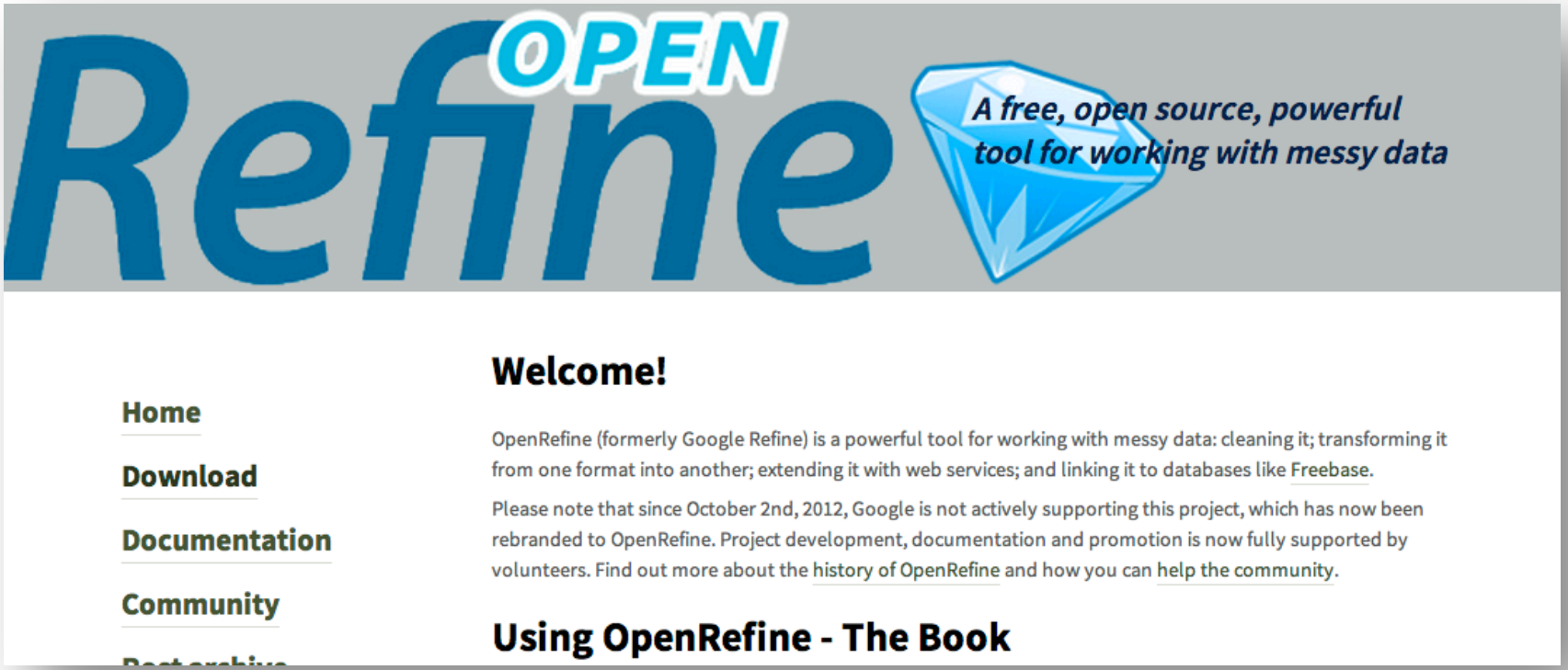
Submitted urls are recorded in a public list of validation reports. If you want to validate private data then upload a file from your computer, using the Browse button below.

CSV Dataset health check (CKAN only)

http://theodi.github.io/csv-dataset-validator/



Open Refine (.org)



Thank-you