

Open Data



Open Data Publishing Pipeline (ODPP)

Richard Cyganiak
DERI, NUI Galway
SEMIC Conference 2013

The Problem

Open Data



Data wranglers
deliver new value
from existing data.

The Problem

Open Data



Data wranglers
deliver new value
from existing data.

Stats package

Open Data

Visualisation/
Infographics

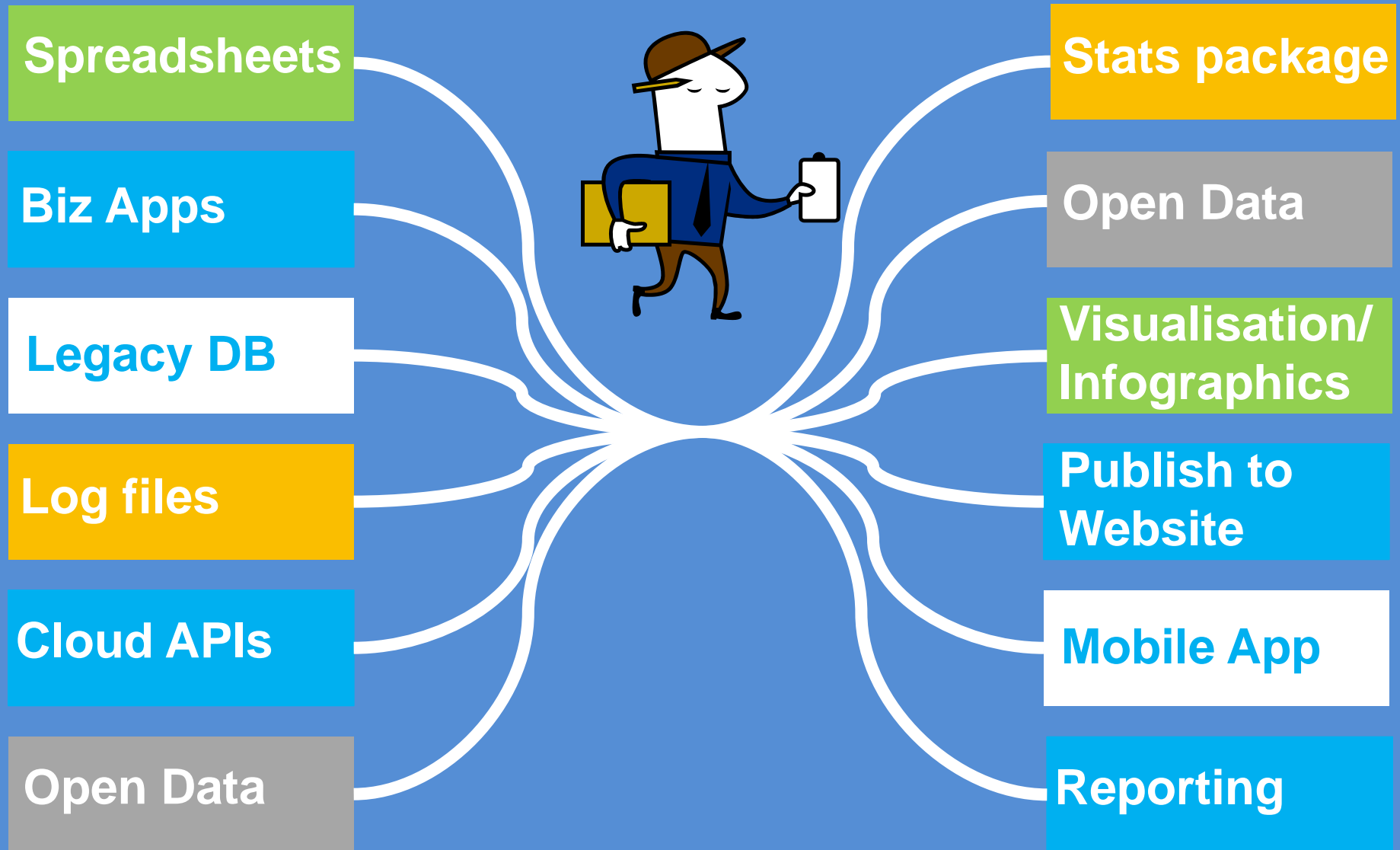
Publish to
Website

Mobile App

Reporting

The Problem

Open Data



The Problem

Open Data

Spreadsheets

Stats package

Biz Apps

Open Data

Legacy DB

Visualisation/
Infographics

Log files

Publish to
Website

Cloud APIs

Mobile App

Open Data

Reporting



The V's of Big Data

Open Data

1. Volume

2. Velocity

3. Variety

Catalogs

Standards

Tools

Community



Welcome to the Data Hub!

Find data

the Data Hub contains **3324 datasets** that you can browse, learn about and download.

Share data

Add your own datasets to share them with others and to find other people interested in your data.

[Sign up »](#)

Collaborate

Find out more about working with open data by exploring these resources:

- [GetTheData.org](#)
- [DataPatterns.org](#)
- [Open Data Manual](#)

Who else is here?

Canada

Datasets for <http://www.datadotgc.ca/>. DataDotGC, which launched, in February 2010, is a Canadian, citizen-led effort to promote open data and help share data that has already been...

Canada has 523 datasets.

Bibliographic Data

This group comprises open bibliographic datasets according to the Principles on Open Bibliographic Data and a few not yet really open bibliographic datasets. It

Linking Open Data Cloud

This group catalogs data sets that are available on the Web as Linked Data and contain data links pointing at other Linked Data sets. The descriptions of the data sets in this group are...

Linking Open Data Cloud has 326 datasets.

Climate Data

Weather, temperature, carbon, water, soil and all other kinds of climate related open data. Including (but not limited to) data about: the atmosphere the weather...

Linking Open Data

A group for Linking Open Data datasets. The initial import of data for this group was done in October 2009 from the list of RDF datasets dumps provided by the W3C Linking Open Data...

Linking Open Data has 81 datasets.

OpenSpending

Datasets to be imported to the [OpenSpending.org](#) site. Packages listed here will automatically be available for selection in the OpenSpending web importer.

prefix.cc

namespace lookup for RDF developers

examples: [foaf](#) [foaf:knows](#) [dc,foaf](#) [rdfs,dc,foaf,geo.sparql](#) [http://xmlns.com/foaf/0.1/name](#)

[popular](#) [latest](#) [about](#) | [prefix.cc](#)

- Namespace registry for RDF/SPARQL users
- Side effect: Create database of vocabularies



NUI Galway
OÉ Gaillimh



DERI Galway

DCAT — Data Catalog Vocabulary

VoID — Vocabulary of Interlinked Datasets

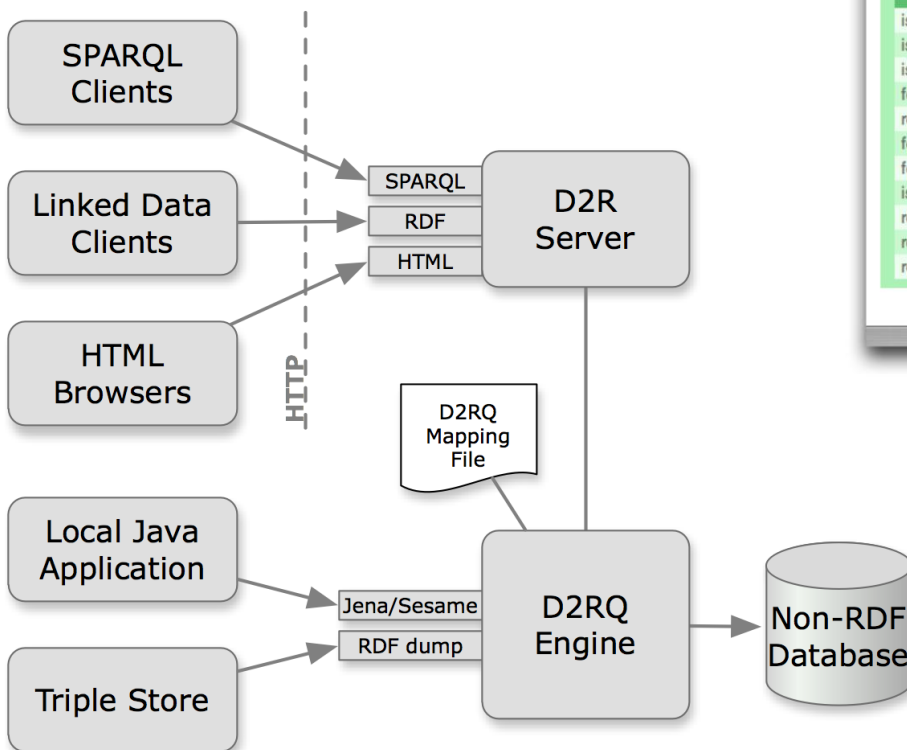
DDI Discovery Vocabulary

Data Cube Vocabulary

D2RQ

Open Data

Database-to-RDF mapping



Two screenshots of the D2RQ interface are shown. The top screenshot is a web browser displaying the D2R Server for Andy Seaborne, showing a table of properties and values for a resource. The bottom screenshot is a SPARQL Explorer showing a query and its results.

Andy Seaborne | D2R Server
Resource URI: <http://www3.wiwiw.fu-berlin.de/2020/resource/persons/6>

Home | All Persons

Property	Value
iswc:address	Hewlett-Packard Laboratories, Bristol, BS34 8QZ, UK
is do:creator of	http://www3.wiwiw.fu-berlin.de/2020/resource/persons/6
iswc:has_affiliation	http://www3.wiwiw.fu-berlin.de/2020/resource/persons/6
foaf:homepage	http://www3.wiwiw.fu-berlin.de/2020/resource/persons/6
rdfs:label	Andy Seaborne
foaf:mbox	andy@wiwiw.fu-berlin.de
foaf:name	Andy Seaborne
iswc:research_interests	http://www3.wiwiw.fu-berlin.de/2020/resource/persons/6
rdfs:seeAlso	http://www3.wiwiw.fu-berlin.de/2020/resource/persons/6
rdfs:type	http://www3.wiwiw.fu-berlin.de/2020/resource/persons/6
rdfs:type	http://www3.wiwiw.fu-berlin.de/2020/resource/persons/6

SPARQL Explorer for <http://localhost:2020/sparql>

SPARQL:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX map: <file:///Users/richard/d2r/imdb.n3#>
PREFIX do: <http://purl.org/dc/elements/1.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX movie: <http://localhost:2020/resource/>
PREFIX imdb: <http://www.csd.abdn.ac.uk/~ggrinnes/dev/imdb/IMDB#>

SELECT ?title WHERE {
  ?movie imdb:title ?title .
  ?movie imdb:year "1977" .
  ?movie imdb:genre ?genre .
  ?genre rdfs:label "Sci-Fi" .
}
```

Results:

SPARQL results:

title
"Empire of the Ants"
"Cinderella 2000"
"Wizards"
"The Brain Leeches"
"Star Wars"
"Buck Rogers"
"The Magnificent Major"

Facet / Filter Undo / Redo 5

Refresh Reset All Remove All

Type of Contract [change](#)

783 choices Sort by: **name** count [Cluster](#)

Firm Fixed Price. Sun Server Maintenance. 1

Firm Fixed Price. Sun Server/Hardware Maintenance. 1

Firm Fixed Price. Technical Service Support (mod for ING06HQFS0056) 1 [edit](#) [include](#)

Firm Fixed Price.Sun Hardware Maintenance 1

Firm Fixed Price/For support of multiple investments within Infrastructure i.e. BCT LAN EAD ESN 1

Firm Fixed Price/For support of multiple investments within Infrastructure i.e.

5200 rows

Show as: rows [records](#) Show: 5 10 25 50 rows

	All	Contract ID	Contractor Name	Type of Contract	Date
☆	1.	1939	ASAP SOFTWARE EXPRESS INC DELL MARKETING L.P.	Microsoft Enterprise Agreement	04/01/200
☆	2.	1940	BMC SOFTWARE DISTRIBUTION INCORPORATED	Remedy Service Desk Maintenance	04/01/200
☆	3.	1941	GOVCONNECTION INCORPORATED	Cisco SmartNet	05/01/200
☆	4.	1942	ITS CORPORATION	Time & Materials	12/31/200
☆	5.	7490	SENET INTERNATIONAL CORPORATIO	Firm Fixed Price C&A	05/04/200
☆	6.				200
☆	7.				200
☆	8.				200
☆	9.	1948		firm fixed price	11/05/200

RDF Export for Google Refine

Base URI

Properties

RDF nodes populated from spreadsheet columns

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by clicking which column to substitute into which node.

Base URI: <http://lab.linkeddata.deri.ie/test/> [edit](#)

RDF Skeleton

[RDF Preview](#)

Available Prefixes:

[rdfs](#) [dcat](#) [foaf](#) [dct](#) [myns](#) [rdf](#) [skos](#)

[+ add prefix](#)

[manage prefixes](#)

Organisation URI



[X](#) [foaf:name](#) →

[X](#) [foaf:Organization](#)

[add rdf:type](#)

Organisation cell



[X](#) [foaf:member](#) →

(row index) URI



[X](#) [foaf:Person](#)

[add rdf:type](#)

[X](#) [foaf:name](#) →



[X](#) [foaf:title](#) →



[X](#) [:annualPayRate](#) →



[add property](#)

Name cell

Job Title cell

Annual pay rate - including taxable benefits

Classes

[add property](#)

Define multiple trees

[Add another root node](#)

OK

Cancel

Neologism

RDF Schema Vocabulary Editor and Publishing System

DERI Vocabularies

Home > dady

Dataset Dynamics (dady) vocabulary

[View](#) [Edit](#)

Author:
Michael Hausenblas (LIDRC) (Email)

Last update:
27 January 2010

Namespace URI:
<http://purl.org/NET/dady#>

Abstract
A vocabulary that allows for coarse-grained descriptions of the data-source dynamics as well as the discovery of change notification mechanism(s).

All terms at a glance

Classes: [AtomUpdateSource](#) | [ChangeSetUpdateSource](#) | [HighFrequentUpdates](#) | [IrregularUpdates](#) | [LowFrequentUpdates](#) | [MidFrequentUpdates](#) | [NoUpdates](#) | [RegularUpdates](#) | [TriplifyUpdateSource](#) | [UpdateDynamics](#) | [UpdateFrequency](#) | [UpdateRegularity](#) | [UpdateSource](#)

Properties: [dynamics](#) | [notification](#) | [update](#)

[Create new class](#) | [Create new property](#)

Classes

- [dady:UpdateDynamics](#)
 - [dady:UpdateFrequency](#)
 - [dady:HighFrequentUpdates](#)
 - [dady:LowFrequentUpdates](#)
 - [dady:MidFrequentUpdates](#)
 - [dady:NoUpdates](#)
 - [dady:UpdateRegularity](#)
 - [dady:IrregularUpdates](#)
 - [dady:RegularUpdates](#)
- [dady:UpdateSource](#)
 - [dady:AtomUpdateSource](#)
 - [dady:ChangeSetUpdateSource](#)
 - [dady:TriplifyUpdateSource](#)

Properties

- [dady:dynamics](#)
- [dady:notification](#)
- [dady:update](#)

Overview diagram

Zoom: [Config](#) [Save Layout](#) [About](#)

```
graph TD
    A[an update source offering updates as Atom feeds] --> D[dady:UpdateSource]
    B[an update source offering updates as Talis' Changesets] --> D
    C[an update source offering updates as Triplify's Updates] --> D
    D -- offers update source --> E[dataset update dynamics]
    F[void:Dataset] -- dataset has update dynamics --> E
    E --> G[dady:UpdateFrequency]
    G --> H[high-frequent]
```

Great stuff, but...

Open Data

Great stuff, but...

Open Data

This is still a full-time job!

- **Many separate tools** that require setup and have their own learning curve
- **Command-line interfaces, complex languages**
- **Web server configurations** for publishing data and vocabularies according to best practices
- *Difficult to transfer this knowledge to our partners*

ODPP: Open Data Publishing Pipeline

Automated discovery

Catalogue

Reformat

Standardise

Publish

- Web-based data publishing platform
- Map input data (relational, CSV, Excel) to semantic standards
- Visual mapping editor and vocabulary search, based on UML

Open Data Publishing Pipeline

- **Already knows all (?) semantic standards that have been published as RDF Schemas**
- **Special support for generating SKOS, Data Cube and Schema.org**
- **Under the hood: proven open-source technologies**

ODPP Development

Open Data

Under development since early 2013

Trials with LGMA, Galway and Fingal CCs

Public beta in Q3

How would *you* use ODPP?

Open Data



richard.cyganiak@deri.org
+353 91 49 5730
@cygri

Thank You!