Open Data Overview

Author: Deirdre Lee, DERI, NUI Galway

Contents

What is Open Data

The goal of the Open Data initiative is to motivate governments and organisations to make information freely available and easily accessible online. The benefits of Open Data are economic, through the identification of new business opportunities, and social, through increased transparency, participation and accountability.

The Open Knowledge Foundation defines a work as 'open' if it satisfies the specified conditions of access, redistribution, reuse, absence of technological restriction, attribution, integrity, discrimination and license - in essence a work is open if it can be freely used, reused and redistributed by anyone [1]. Although any institution or organisation may produce Open Data, for example the World Bank[1], the Guardian[2], University of Southampton[3] and Enel[4], an emphasis is placed on publishing information from public authorities: Open Government Data. Open Government Data refers to data and information produced or commissioned by the government or government controlled entities [2].

Governments have traditionally been protective over the data they collect, citing national security and citizen privacy as reasons. However Open Government Data does not pertain to sensitive or individual information, but core public data on transport, infrastructure, education, health, crime, environment, etc. Many believe that public data should be open by default, as it is information pertaining to the public domain and has been collected using public finances, i.e. taxes.

Freedom of information legislation comprises laws that guarantee access to data held by the state. According to Wikipedia, over 85 countries around the world have implemented some form of such legislation, with Sweden's Freedom of the Press Act of 1766 is the oldest in the world[5]. In addition to national legislation, the European Union created the Directive 2003/98/EC of the European Parliament and the Council of 17 November 2003 on the re-use of public sector information[6],. This introduced a common legislative framework regulating how public sector bodies should make their information available for re-use in order to remove barriers such as discriminatory practices, monopoly markets and a lack of transparency.

The United States of America and the United Kingdom led the way in terms of Open Data. In January 2009, President Obama issued a memo on Transparency and Open Government as one of his first acts in office and data.gov[7] was launched in May 2009. Under the Gordon Brown government data.gov.uk[8] was launched in January 2010. London also launched a datastore for the capital[9]. Datacatalogs.org[10] claims that there are now 200 registered data catalogues available online, from

[1] http://data.worldbank.org/

[2] http://www.guardian.co.uk/data

[3] http://data.southampton.ac.uk/

[4] http://data.enel.com/

[5] http://en.wikipedia.org/wiki/Freedom_of_information_legislation

[6] http://ec.europa.eu/information_society/policy/psi

[7] http://data.gov

[8] http://data.gov.uk

[9] http://data.london.gov.uk/

[10] http://datacatalogs.org/

national and local public authorities, as well as some private enterprises. The European Commission are also taking the Open Data initiative seriously and in December 2011, Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda, presented the Open Data Package consisting of a Communication on Open Data, a proposal for a revision of the Directive and a proposal for a revision of the Commission's rules on re-use of the documents it holds [6].

How to Publish Open Data

Following on from Tim Berners-Lee's 5-star deployment scheme for Linked Open Data[11], Cyganiak devised the 5-shamrock scheme for publishing Open Data [3]

- Publish data on the Web
- Publish data in a machine-processable format
- Use an open standard format
- Publish under an open license
- List your data in a data catalogue.

Data catalogues are common registries that list multiple datasets, for example, the Spanish data catalogue[12], the Danish data catalogue[13] and the Dublin, Ireland data catalogue[14]. The data catalogue may host the datasets or it may link to datasets hosted elsewhere.

Open Data Technologies

CKAN

CKAN is open-source data portal software[15], initiated by the Open Knowledge Foundation[16]. CKAN makes it easy to publish, share and find data by providing a powerful database for cataloguing and storing datasets, with an intuitive web front-end and API. CKAN is used by the UK, Norwegian and Dutch governments, local government, and specialist data publishers. It has run ckan.net (now thedatahub.org) since 2007, and powers more than 40 data hubs globally.

INSPIRE

Although not pertaining to Open Data in particular, the INSPIRE Directive[17] is very relevant to the Open Data initiative, as much of the public information being made available is spatial data. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) was published in the official Journal on the 25th April 2007. The INSPIRE Directive entered into force on the 15th May 2007 and will be implemented in various stages, with full implementation required by 2019.

[11] http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/
[12] http://opengov.es/
[13] http://data.digitaliser.dk/
[14] http://www.dublinked.ie/datastore/
[15] http://ckan.org
[16] http://okfn.org/
[17] http://inspire.jrc.ec.europa.eu/

The INSPIRE directive aims to create a European Union (EU) spatial data infrastructure. This will enable the sharing of environmental spatial information among public sector organisations and better facilitate public access to spatial information across Europe. A European Spatial Data Infrastructure will assist in policy-making across boundaries. Therefore the spatial information considered under the directive is extensive and includes a great variety of topical and technical themes.

DCAT

To facilitate integration, discoverability and searchability of data catalogues, a common metadata schema is essential. The Data Catalog Vocabulary (DCAT)[18] is an RDF vocabulary to represent government data catalogues such as data.gov and data.gov.uk. It is a W3C note and a task force within the W3C Interest Group on eGovernment[19]. An overview of DCAT is shown in Figure 1. DCAT defines three main classes:

- dcat:Catalog represents the catalogue
- dcat:Dataset represents a dataset in a catalogue
- dcat:Distribution represents an accessible form of a dataset as for example a downloadable file, an RSS feed or a web service that provides the data.

More information on DCAT can be found in [4].

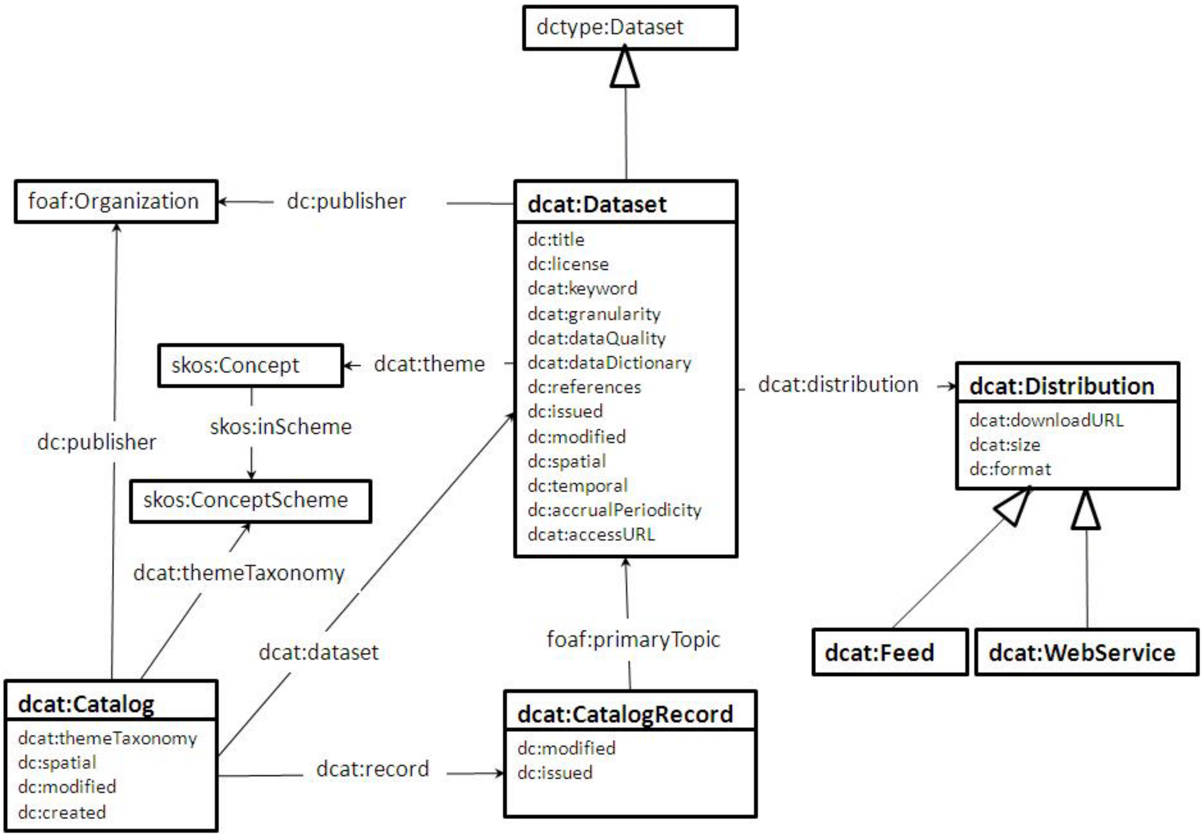

Figure 1: Data Catalog Vocabulary (DCAT)

[18] http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary/Vocabulary_Reference
[19] http://www.w3.org/egov/wiki/Main_Page

VOID

The Vocabulary of Interlinked Datasets (VOID)[20] is similar to DCAT, in that it is an RDF based metadata schema. However while DCAT can be used to describe any data catalogue, VOID is used to describe Linked datasets. With VoID the discovery and usage of linked datasets can be performed both effectively and efficiently. A Linked dataset is a collection of data, published and maintained by a single provider, available as RDF, and accessible, for example, through dereferenceable HTTP URIs or a SPARQL endpoint.

JoinUp Core Vocabularies & ADMS

In addition to common metadata schemas, common models of how to present and structure data are required to avoid creating islands of open data in Europe, which could lead to fragmented open data initiatives. Joinup[21] is a new collaborative platform created by the European Commission and funded by the European Union via the Interoperability Solutions for Public Administrations (ISA) Programme[22]. It offers a new set of services to help e-Government professionals share their experience with interoperability solutions and support them to find, choose, re-use, develop, and implement open source software and semantic interoperability assets.

Core Vocabularies[23], such as the Core Person Vocabulary and Core Business Vocabulary, are simplified, reusable and extensible data models that captures the fundamental characteristics of a particular asset or domain.

The Asset Description Metadata Schema (ADMS)[24] is a common way to describe semantic interoperability assets making it possible for everyone to search and discover them once shared through the forthcoming federation of asset repositories. The development of ADMS is being coordinated on More information on ADMS can be found on [5]

Linked Open Data Projects, Technologies & Tools

As discussed above, in order to ensure efficient reuse, Open Data should be published in a machine-processable format. To support advanced semantic interoperability, Linked Open Data (LOD) should be provided.

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. Key technologies that support Linked Data are URIs (a generic means to identify entities or concepts in the world), HTTP (a simple yet universal mechanism for retrieving resources, or descriptions of resources), and RDF (a generic graph-based data model with which to structure and link data that describes things in the world)[25]. RDF, the Resource Description Framework, is one of the key ingredients of Linked Data, and provides a generic graph-based data model for describing

[20] http://www.w3.org/TR/void/
[21] https://joinup.ec.europa.eu/
[22] http://ec.europa.eu/isa/
[23] https://joinup.ec.europa.eu/asset/all
[24] https://joinup.ec.europa.eu/asset/adms/home
[25] http://linkeddata.org/

things, including their relationships with other things. RDF data can be written down in a number of different ways, known as *serialisations*. Examples of RDF serialisations include RDF/XML, Notation-3 (N3), Turtle, N-Triples, RDFa, and RDF/JSON [6].

Projects, such as LOD2[26] and LOD-Around-The-Clock (LATC)[27] funded under the EC FP7 ICT Programme, support the publication and consumption of Linked Data through the provision of tools and best practices. Linked Data life cycles[28] identify fundamental phases and involved parties of Linked Data publication and consumption. Similarly, the LOD2 technology Stack comprises a number of tools for managing the life-cycle of Linked Data, in particular the stages

- Extraction of RDF from text, XML and SQL
- Querying and Exploration using SPARQL
- Authoring of Linked Data using a Semantic Wiki
- Semi-automatic link discovery between Linked Data sources
- Knowledge-base Enrichment and Repair

For example the D2R Server[29] is a tool for publishing relational databases on the Semantic Web. It enables RDF and HTML browsers to navigate the content of the database, and allows applications to query the database using the SPARQL query language.

The RDF Extension for Google Refine[30] adds a graphical user interface(GUI) for exporting data of Google Refine[31] projects as interlinked RDF data. Data can be reconciled against any SPARQL endpoint or RDF dump. The reconciled data can then be exported as RDF based on a template graph. These all are supported by a friendly graphical interface.

Bibliography

1. Open Knowledge Foundation: Open Definition. 2011 [cited 2011; Available from: http://opendefinition.org/okd/.
2. Open Knowledge Foundation: Open Government Data. 2011 [cited 2011; Available from: http://opengovernmentdata.org.
3. Cyganiak, R.: How to Publich Open Data. in Opening Up Government Data. DERI, NUI Galway. (2011).
4. Maali, F., Cyganiak, R., Peristeras, V.: Enabling Interoperability of Government Data Catalogues, in Electronic Government 10th International Conference, EGOV 2010 (2010).
5. Shukair, G., et al.: Towards a Federation of Government Metadata Repositories, in Share-PSI.eu workshop on Removing the roadblocks to a pan European market for Public Sector Information re-use (2010).
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS), (2009).

[26] http://lod2.eu/

[27] http://latc-project.eu/

[28] http://linked-data-life-cycles.info/

[29] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[30] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/

[31] http://code.google.com/p/google-refine/