Open Data on the Web, London, 2013

Position paper: Bill Roberts, Swirrl IT Limited
bill@swirrl.com
+44 7717 160378

Machine-readable metadata for open data discovery and quality assessment

Currently, a potential user of open data tends to encounter various catalogues or lists of datasets and usually wants answers to the following questions:

- What is this dataset about?
- Does it cover the area I am interested in?
- Does it cover the time period I am interested in?
- How detailed is it?
- Who created it?
- Am I allowed to use it?
- Is it accurate?
- Is it up to date?
- Will it be kept up to date in future?
- Will it still be available in the future?

Typically data publishers provide some form of metadata to describe their datasets. Currently there is a wide variety of forms and formats for such metadata, with varying degrees of machine readability.

To support easier discovery and use of open data as the number of datasets and number of open data catalogues explodes, interoperable machine readable metadata becomes steadily more important.

Interoperability tends to require some degree of standardization. Ideally this should be balanced with the need for simplicity and low barriers to entry for new publishers.

A useful topic to discuss at ODW would be how to agree guidelines or standards on:

- which aspects of a dataset should be described in metadata
- preferred vocabularies for representing metadata in machine readable form
- options for deployment methods that make life easy for automated discovery tools

In the case where the data itself is RDF, then it is clearly natural that the metadata should also be RDF. Also for open data in other formats, RDF is a natural approach for a machine readable representation of metadata.

There are a number of vocabularies in common use: Dublin Core, VoiD, DCAT. More recently schema.org has developed RDF-compatible vocabularies that provide a number of relevant terms and may be of particular benefit in assisting generic search engines to find and correctly categorise the data.

I am currently involved in the design of a metadata strategy for linked open data publications by the UK government Department for Communities and Local Government (DCLG). The outcomes of this could be applicable across the UK public sector and beyond and I would welcome the chance to broaden the discussion, look for international alignments and identify good examples to learn from. I would be happy to make a short presentation at ODW on the status of the DCLG work.

The Open Data Institute proposal for 'Open Data Certificates' (http://theodi.github.com/open-data-certificate/) is highly relevant to this discussion. The recommendation for which information to include in dataset metadata could be closely aligned to the requirements for ODI Open Data Certificates.

An issue that has come up in early discussions of the certificates is also relevant to the question of dataset metadata: that not all open data falls neatly into the model of a series of distinct datasets. Rapidly changing data in particular tends to be provided primarily via APIs and sometimes a static snapshot for download is difficult to create, or not particularly useful. Recommendations for dataset metadata should also be applicable to 'data service' metadata.