Metadata for the open data portals

Discussion Paper No. 6, December 2016, Joined-up Data Standards Project

Beata Lisowska

Data Scientist, Development Initiatives

Contents

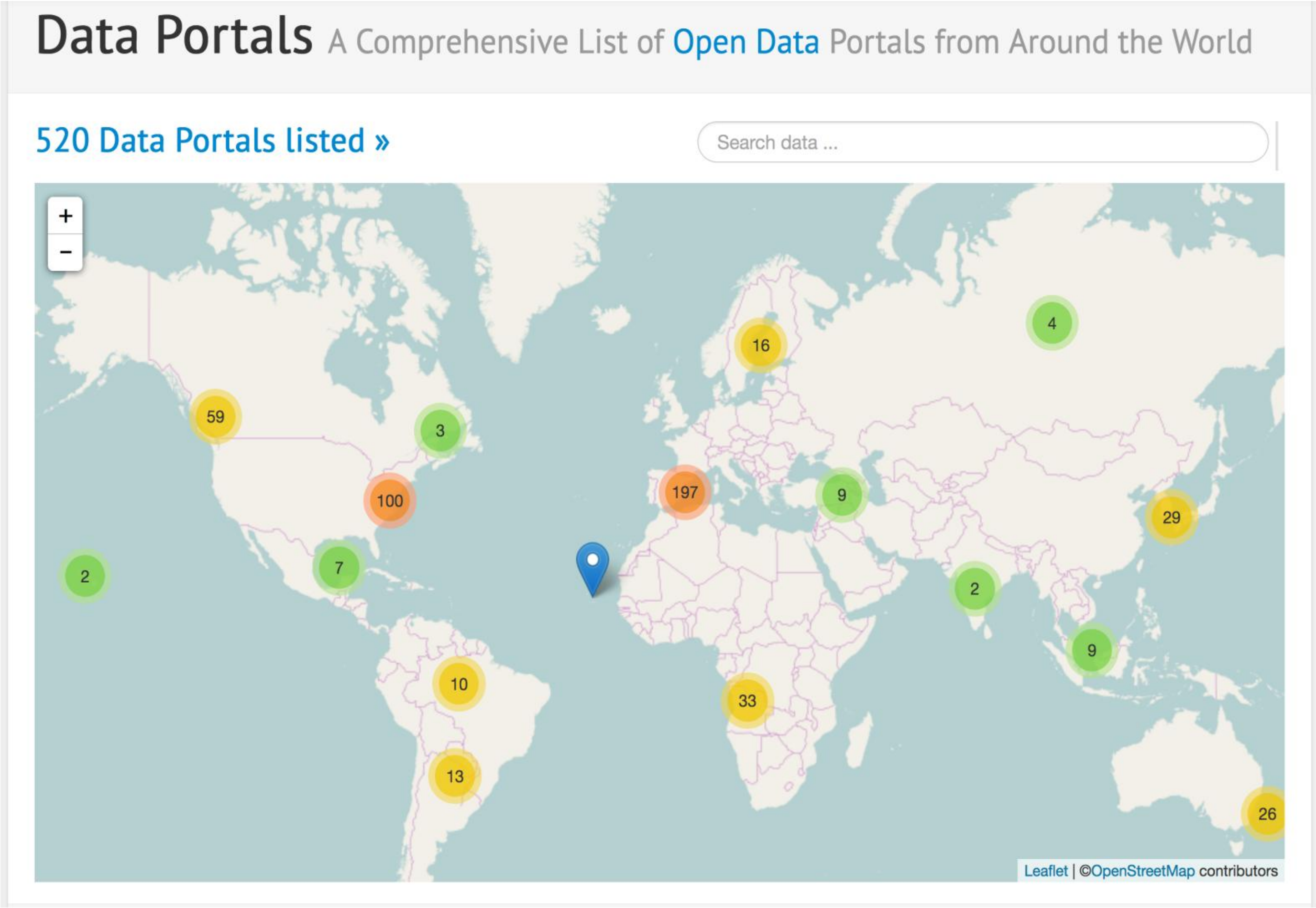
Introduction

A recent Open Data Institute (ODI) summit in London featured a number of talks where a range of stakeholders discussed open data: how important it is, how it unleashes the true potential of data, what it means, what possibilities if offers, and where the future of the open data lies. Open data, should be accessible to all, usable and sharable by all, and as such is a key tool in seeking to advance sustainable development and be used for good governance.

However, despite more data being published in open formats, data scientists, journalists and analysts are often left with a daunting and time-consuming task of not only finding relevant data and discovering new datasets, but most importantly understanding it before any analysis can be done. That information should be found in the metadata that should couple the data published.

Metadata is, in essence, structured information that makes it easier to retrieve, use or manage an information resource. In practice, metadata describes a dataset and its structure, and helps users discover it. The information usually includes such basic elements as: title, who published the dataset, when it was published, how often it is updated and what license is associated with the dataset. These are classed as 'descriptive metadata' as opposed to 'structural metadata', which describes for example information on page layout or an object's component and their relationships (such as chapters or tables in a book).

Just as the number of open datasets has exponentially increased, so too has the number of open data portals and associated standards. There are currently 521 open data portals listed on Data Portals, a list curated by a group of experts from around the world. Staggeringly 197 are associated with Europe, 100 registered in the USA, and only 33 in Africa. Simple analysis of the resources reveal that 118 of the portals are classed as Government (or 'government'), 12 as Community, 5 as Institutional and 6 as Research. In total 141 open data portals are assigned a publisher, the remaining 380 are not assigned on a publisher classification basis. Five of the portals are listed as 'inactive'.



Data Portals A Comprehensive List of Open Data Portals from Around the World

520 Data Portals listed »

Search data ...

Figure 1: Map of open data portals in the world according to OpenDataPortals.org

This information was pooled from the website's metadata; however, it seems that although the metadata is present, only a quarter has annotated fields. This in itself can stem from a variety of issues, but one stands out in particular: this data is incomplete because the data portals listed do not provide comprehensive metadata that describes their own platforms. 385 portals have metadata associated with them. Only 8 provide links to full metadata downloads and only 12 provide a working API (application programming interface) point.

The above problem combined with an ever-increasing number of open data portals begs the question: which metadata standards are used, if any, and which platforms are most prevalently used for these portals?

This paper investigates how open data portals share their metadata and explores the most prevalent underlying metadata standards used. It seeks to understand to what extent the metadata standards used by the predominant open data platforms are interoperable. Interoperble metadata across open data portals enables datasets to be discoverable, re-useable and searchable across portals rather than 'siloed' within them (this is called a federated search).

Platforms used for open data portals

When looking at the software governments and organisations use to publish their open data and generate metadata, CKAN, Socrata and Junar were the most commonly listed by opendata.org (Figure 2). However, it must be noted that this website did not provide a comprehensive picture of the distribution of how the metadata is generated since the sample size was only 55 and this information was not available for 330 remaining open data portals.

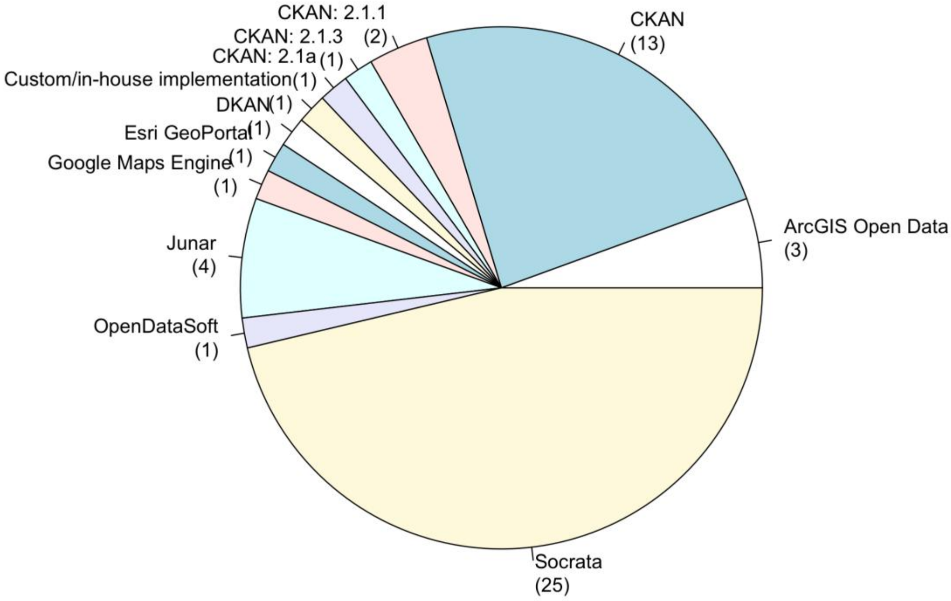


Figure 2. Distribution of open data 55 portals generators listed on opendata.org. This chart excludes 330 open data portals that did not disclose the metadata generator type. The metadata was accessed on the 27th October 2016.

Metadata standards

The way in which metadata is generated by the open data portals is dictated by the data they publish, such as statistical, geographical, or financial, as these are linked to the metadata standards that they employ. These standards are a direct response to the need for descriptive, structured information on the data published on any given subject. For example, the INSPIRE Infrastructure for Spatial Information in Europe standard (INSPIRE metadata schema) stemmed from the need to address the specific metadata requirements for the geo-spatial information that is in itself based on the ISO 19115 standard. There are a number of metadata vocabularies, ontologies and standards weaved into the architecture of the open data portals and Table 1 shows the ones most predominantly used by the most popular open data platforms.

Table 1. The underlying metadata standards behind commonly used metadata generators used by open data portals

Open data platform	Metadata model
CKAN	CKAN
OpenDataSoft	DCAT
SOCRATA	Socrata
DKAN	DCAT, INSPIRE
ArcGis Open Data	INSPIRE
Esri Geoportal Server	Open Geospatial Consortium (OGC) compliant CS-W 2.0.2 service
Junar	DCAT, INSPIRE

However, these metadata models applied by individual portals are based on established core metadata standards. For example the Dublin Core standard is a basic layer for nearly all metadata standards in use today. It is the specific layer of a metadata standard (Figure 3) that determines the subject-specific use of the standard, while the distribution layer is the one which makes the metadata model for each open data portal.

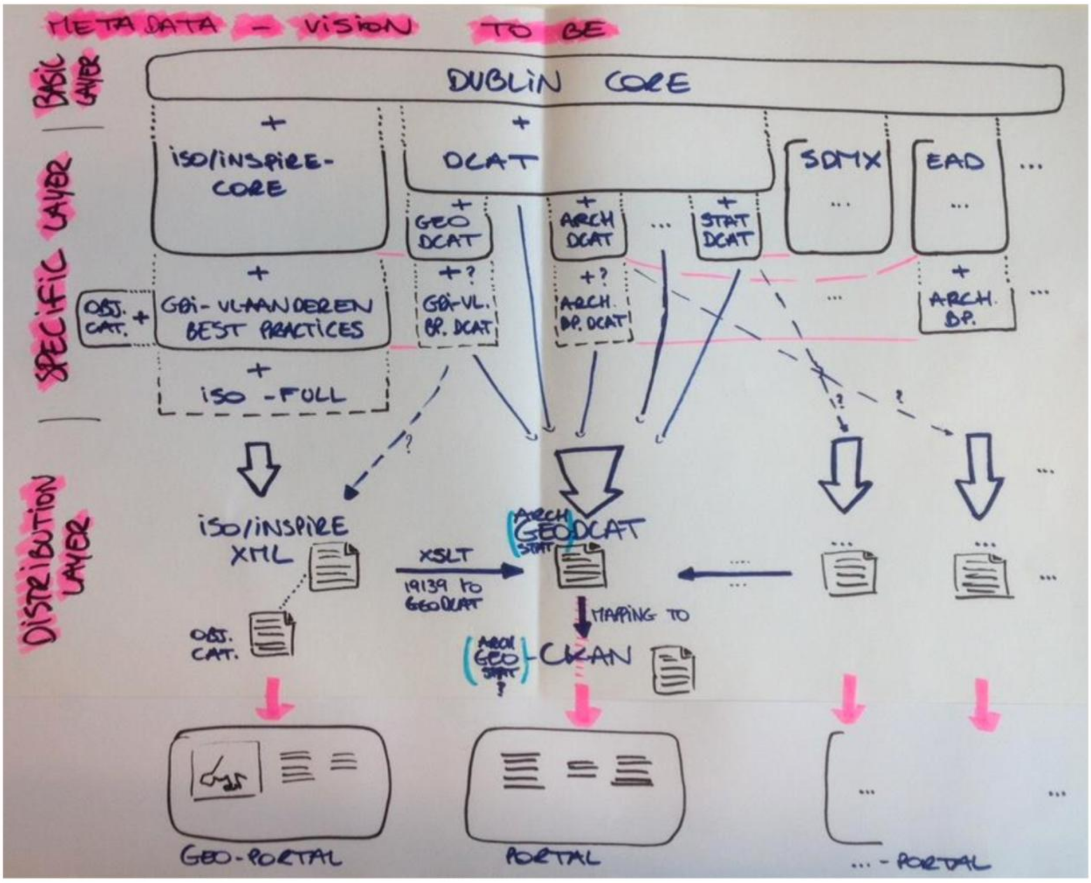


Figure 3. Metadata standards for geospatial, open government, statistical and archival information (Geraldine Nolf, Informatie Vlaanderen). The picture shows the different layers of the metadata standards

In the following sections we describe the core metadata standards, focusing on metadata standards that are domain-specific: geographic and spatial-temporal information.

Core metadata standards

RDF Data Cube vocabulary

This vocabulary provides the means to publish multidimensional data, such as statistics, on the web in such a way that it can be linked to related datasets and concepts using the W3C RDF (Resource Description Framework) standard. The model underpinning the Data Cube vocabulary is compatible with the cube model that underlies SDMX (Statistical Data and Metadata eXchange), an ISO standard for exchanging and sharing statistical data and metadata among organisations.

Dublin Core

A basic, domain-agnostic standard that can be easily understood and implemented, and as such is one of the best known and most widely used metadata standards. Sponsored by the Dublin Core Metadata Initiative, Dublin Core was published as ISO Standard 15836 in February 2009.

Data Catalog Vocabulary (DCAT)

The most commonly used metadata standards are in some way a version of the basic DCAT standard. DCAT in itself has gained popularity due to its flexibility and elegant design. The main goal of DCAT is to improve the data catalogues' interoperability so applications can easily consume metadata from multiple catalogues. An example of a basic DCAT is DCAT-AP, which was designed to meet the metadata publishing needs in the context of the European Commission's Interoperability Solutions for European Public Administrations (ISA) programme: 'Improving semantic interoperability in European eGovernment systems'. This standard uses the main classes defined by the DCAT standard, and builds on properties to expand the fundamental version of the standard to fit the needs defined by the European Commission.

Other profiles of DCAT include: Common Core Metadata Schema (CCMS), which provides vocabulary that other schema can map to Asset Description Metadata Schema (ADMS) for describing Semantic Assets within a catalog, or most recently, the metadata standard that enables the sharing of metadata across different data catalogs called Data Catalog Interoperability Protocol (DCIP).

DCAT profiles and structure have been described in detail in our discussion paper on data catalog vocabulary, where the principles of DCAT application to databases are also explored.

Other metadata standards

Table 2: Other metadata standards

Metadata standard	Description
Project Open Data Metadata Schema v1.1	The metadata schema is based on DCAT. This specification defines three types of metadata elements: Required, Required-if (conditionally required), and Expanded fields. These elements were selected to represent information that is most often looked for on the web. To assist users of other metadata standards, field mappings to equivalent elements in other standards are provided.
VoID	VoID (Vocabulary of Interlinked Datasets) is an RDF vocabulary, and a set of instructions, that enables the discovery and usage of linked datasets.
Schema.org	Schema.org provides a set of extensible schemas to mark-up HTML pages that enables webmasters to embed structured data on their web pages for use by search engines.
PREMIS	The PREMIS (Preservation Metadata: Implementation Strategies) Data Dictionary defines a set of metadata that most repositories of digital objects need to record and preserve those objects over the long term. It has its roots in the Open Archival Information System Reference.

Geographic metadata standards

ISO 19115

This is an internationally-adopted schema for describing geographic information and services. It provides information about the identification, extent, quality, spatial and temporal schema, spatial reference, and distribution of digital geographic data.

Sponsored by the ISO (International Organization for Standardization), the first edition of ISO 19115 was published in 2003. It has since been split into parts: ISO 19115-1:2014 contains the fundamentals of the standard; ISO 19115-2:2009 contains extensions for imagery and gridded data; and ISO/TS 19115-3:2016 provides an XML schema implementation for the fundamental concepts compatible with ISO/TS 19138:2007 (Geographic information Metadata XML, or GMD).

Table 3. Table of metadata schemas, standards, vocabularies and ontologies implemented by open data portals

Metadata standard	Description
INSPIRE Metadata Schema	This style allows users to view and edit metadata following the FGDC Content Standard for Digital Geospatial Metadata (CSDGM) guidelines, export metadata in this standard's XML format, and validate it using the CSDGM XML DTD.
Google DSPL (Dataset Publishing Languages)	DSPL is a data and metadata format designed from the ground up to support powerful, interactive visualisations like those in the Google Public Data Explorer.
FGDC CSDGM Metadata	This style allows users to view and edit metadata following the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) guidelines, export metadata in this standard's XML format, and validate it using the CSDGM eXtensible Markup Language (XML) Document Type Declaration (DTD).
ISO 19139 Metadata Implementation Specification GML3.2	This style is identical to the one above, except the exported files use the GML 3.2 namespace, and therefore can be validated with versions of the ISO 19139 XML schemas that reference the GML 3.2 namespace. For example, use this style if you plan to validate the exported metadata files using the NOAA NCDDC XML schemas.
ISO 19139 Metadata Implementation Specification	This style allows users to view and edit a complete metadata document that complies with ISO standard 19139, Geographic information – Metadata – XML schema implementation, export metadata in this format, and validate it using the standard's XML Schemas. Use this style to create metadata that complies with ISO standard 19115, Geographic information – Metadata
North American Profile of ISO 19115 2003	This style allows users to view and edit a complete metadata document that complies with the North American Profile of ISO 19115:2003 – Geographic information – Metadata, export metadata in this format, and validate it using the ISO 19139 XML schemas.

Open data portals and metadata

W3C 'Data on the Web Best Practices' encourages publishers to couple data with metadata at the time of publication. The majority of open data publishers respect this rule. However, the format in which metadata is published depends highly on the open data portal that publishes it.

Open data portal software frameworks are either built on their own standards or use an already existing standard. The two predominant platforms – CKAN and Socrata – have each been developed on their own respective frameworks that are rooted in major standards such as Dublin Core and RDF vocabulary. The platforms tend to use either one standard, as described in the previous sections, to generate metadata or a combination of a few. The following section describes major platforms in relation to the metadata fields and standards that they use.

Socrata and CKAN

Socrata is based on the RDF metadata (Dublin Core and DCAT) with enrichment from custom metadata fields. CKAN stores the datasets as a folder that hosts datasets or resources. The metadata is served as RDF and the platform supports DCAT, Dublin Core and INSPIRE format. The generated metadata fields are shown in Table 4.

Table 4. Comparison between metadata fields defined by Socrata and CKAN

	Fields	Description
CKAN	**Title**	Field used to label datasets. This attribute is intended to allow search, sharing and linking of datasets
	Unique identifier	This attribute assigns a unique URL to a dataset. This is one of the Dublin Core recommendations
	Groups	A customisable group that the dataset belongs to
	Description	Human readable description of the dataset
	Data preview	Quick preview in the comma separated value (CSV) format of the dataset
	Revision history	Provides revision history
	Licence	Allows user to check what licence a given dataset is
	Tags	Allocating tags to datasets makes them more discoverable through tag search and faceting by tags
	Formats	Provides information on the format datasets is available for download in
	API key	Allows for a developer access to the metadata fields
	Customizable extra fields	Such as location data or extra information relevant to the publisher or the dataset
SOCRATA	**Name**	Title of the dataset
	ID	Unique identifier for the dataset
	Description	The human-readable description of the asset
	Attribution	The attribution of the dataset
	Type	What sort of asset is described
	Updated at	Timestamp
	Page views	Set to provide statistics on page view of a dataset per day/week/month or all time
	Columns name	An array of column names in the dataset
	Columns description	An array of the descriptions matching the column name
	Columns field name	This serves as an identifier for columns and describes the field names of columns
	Categories	Categories are assigned using statistically derived models
	Tags	Tags are also assigned based on statistically derived models
	Domain category	Given by the owning domain
	Domain tags	Array of tags assigned to the dataset by the owning domain
	Domain metadata	‘Key’ and ‘value’ of any custom metadata given to this asset by the owning domain

DKAN

DKAN metadata fields are compatible with CKAN. The compatibility with CKAN is translated into identical API. Metadata is presented at a dataset level using such standards as: Dublin Core, DCAT and INSPIRE geospatial format. Indeed this data portal uses the ISO array of specific protocols for various types of data. Other standards setting organizations include the US FGDC and the European INSPIRE Metadata Directive.

Junar

Junar delivers cloud-based open data platform for businesses, governments, non-governmental organisations (NGOs) and academia. Junar manages its content based on the SaaS (Software-as-a-service) Open Data Platform. Junar uses the RDF metadata standard as presented in Dublin Core and DCAT. This platform is favoured for its ease of deployment, however, Junar does not support structural metadata.

OpenDataSoft

OpenDataSoft natively uses a subset of DCAT to describe datasets and INSPIRE for geospatial data. The following metadata fields are available in its standard form: title, description, language, theme, keyword, license, publisher and references. It is possible to activate the full DCAT template, thus adding the following additional metadata: created, issued, creator, contributor, accrual periodicity, spatial, temporal, granularity and data quality.

JKAN

Using Jekyll, JKAN allows for a quick deployment of static pages from underlying files. This data portal is based on CKAN and it is aimed at data publishers in the government that would like to deploy their data quickly.

ArcGis Open Data

The following metadata styles are provided to support ArcGis Open data portal: FGDC CSDGM Metadata, INSPIRE Metadata Directive, ISO 19139 Metadata Implementation Specification GML3.2, ISO 19139, Metadata Implementation Specification, and ISO 19139 Metadata Implementation Specification.

Each of these open data portals exposes their metadata using their own fields and in essence standards (Table 4). Although some, such as CKAN, provide handy extensions plugins to expose and consume from other catalogs such as DCAT, it is still not common practice and as a result the federated search across different platforms is hindered.

The lack of interoperability of metadata in the open data portals

A question has been posed on the Open Data Forum, on dealing with the problem of choosing one open data platform above another – in this case, Socrata or CKAN. One of the users, Joe Pringle, answered “I’m seeing more and more cases where Socrata and CKAN are both part of a federated ecosystem of data publishing activities rather than one monolithic catalog that must serve everyone.” Although this answer refers to choosing one platform over another, it in essence argues that as long as data is interoperable, who or what publishes it is not important. The idea of a federated ecosystem of open data applies as well to metadata. If open data portals expose their metadata in different standards then searching and discovering datasets across platforms is impossible.

Vlaamse Overheid tackles the same question of lack of interoperability across metadata standards. Figure 4 – which was presented by Geraldine Nolf at the recent SDSVoc workshop in Amsterdam – shows how the core metadata standards used by open data portals relate to one another. That relationship and the common building blocks of metadata standards allow, for example, linked data to map across different standards.

Nolf argues the metadata profiles should be linked, in the short term, in an uni-directional way from the more specific niche metadata profiles into the more basic/core metadata standard (in this case Dublin Core). She argues that a better solution would be to reconsider, possibly even reshape, one or more standards, so that the differences that data providers encounter converge. This is an argument that we have made before in our consultation paper: if you are talking about the same thing, you should ideally speak the same or in a short term use an interpreter (map across data standards).

Nolf and this paper are not the only ones pointing out these issues. In a 2015 paper Assaf et al. explore how similar the models of open data portals are and performed various steps to establish what the mappings between them should be. Discussing CKAN, DKAN, POD, DCAT, VoID, Schema.org and Socrata, the authors recognise the need for “a harmonized dataset metadata model containing sufficient information so that consumers can easily understand and process datasets".

There are a number of methods that can be used to ensure interoperability between metadata standards in the short term. A 2006 study of methodology behind metadata interoperability and standarisation by Zeng and Chan suggests there are a variety of approaches to deal with the problem of interoperability at a schema level, namely:

1. **Derivation**: when a new schema is created from an existing one, in which case the basic structure and common elements are conserved.
2. **Application profiles (AP)**: this approach ensures similar structure and common elements. APs usually emerge as a response to accommodating specific needs.
3. **Crosswalks**: entails mapping of the elements between metadata standards.
4. **Switching-across**: using switching schema as a switching mechanism among multiple schema. This is an alternative to crosswalks.
5. **Metadata framework**: a skeleton of a metadata that can be used to concentrate efforts of a variety of actors towards a common standarisation in a given field.
6. **Metadata registry**: collection of metadata schemas.

Yet, to the authors' knowledge, the interoperability is integrated in a limited scope into the fibre of open data platforms.

Conclusions

Our July 2016 consultation paper uses a simple linguistic analogy to deal with the lack of interoperability between data standards; you should either speak the same language or use a translator. The same sentiment becomes clear from the literature review of the metadata standards and the approaches to dealing with the lack of interoperability as listed above. What is needed is a standarised way of publishing metadata by the open data portals:

1. to allow the users to find the context of the data in a simple and accessible way
2. to allow a federated search across platforms so that the true power of metadata can be unleashed: the power of machine-readable discoverability of data.

Metadata can range from basic to advanced, from merely stating that the data exist to defining relationships between datasets. Making metadata machine-readable unleashes its huge use potential; making metadata interoperable allows for datasets to be discoverable, re-usable, and searchable in a federated[*] way. However, to achieve that, metadata must not only be defined but also its fields encoded in a common standarised way. If that cannot be achieved in the short term, there must be a machine-readable map between data standards. Otherwise the true potential of open data with its vast resources will be truly limited – and so it is open data portals' responsibility to standardise the way their metadata is exposed.

As more efforts are being directed to crosswalks of metadata schema, we would like to invite the Open Data community and open data portals to work with us on mapping open data metadata standards using our Online Thesaurus, where the mapping can be done using SKOS: Simple Knowledge Organization System.

[*] Federated search allows user to search not only within one portal but across platforms.

Bibliography

Assaf A et al., 2015 ‘HDL – Towards a Harmonized Dataset Model for Open Data Portals’.

Hillmann, DI, 2008, ‘Metadata Standards & Applications.’ Trainee Manual.

National Information Standards Organization, 2001, ‘Understanding Metadata.’

Nolf G, 2016, ‘Interoperability between metadata standards: a reference implementation for metadata catalogues’. Vlaamse Overheid.

World Bank, 2014, ‘Technology Options for Open Government Data Platforms.’

World Bank, 2014, ‘Technical Assessment of Open Data Platforms for National Statistical Organisations.’

Zeng ML and Chan LM, 2006, ‘Metadata Interoperability and Standarization – A Study of Methodology Part I’, volume 12, number 6. D-Lib Magazine.