Ref: Story1

THE USE OF STANDARDS AND CORE REFERENCE DATA FOR LINKING OPEN DATA

INTRODUCTION

Local open data is most useful when like datasets from different localities can be identified, compared and combined. This is complicated where power is devolved down to regional and local government, as consistency of open datasets cannot be mandated across a country or the EU as a whole. However mechanisms can be implemented to aid local governments in aligning where they see public and mutual advantage in doing so.

Open standards make it possible for citizens and entrepreneurs to access data consistently in a multitude of apps or APIs, compliant with the architecture of an open platform.

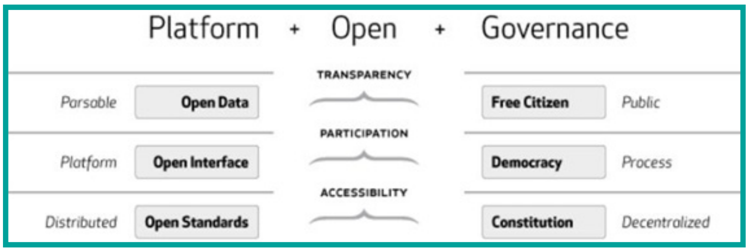


Figure 1: Open standards

An example of open standards is the open311.org standard. Open311 is a form of technology providing open channels of communication for issues concerning public space and public services. Primarily, Open311 refers to a standardised protocol for location-based collaborative issue-tracking.

For some time now, central government in many countries has made data more available for consumption by other organisations and individuals, in both public and private sectors. As technology and understanding of data publishing become better understood, a demand has emerged for the same approach by local government.

However, local government data presents extra challenges and opportunities that are not found in national/departmental data:

- municipalities provide hundreds of types of service, so can publish a hugely diverse range of data;
- there are typically hundreds of municipalities, each publishing a subset of the same type of data, potentially in different formats;
- given the tiered structure of local government in many countries, each community might be served by several authorities of various types;
- the value of local data greatly increases if it can be combined and compared across similar data from multiple other sources;
- the context for each community and local environment is crucial in making meaningful conclusions.

Most of these challenges can be overcome by defining standards for use by all municipalities to publish their data in a common way.

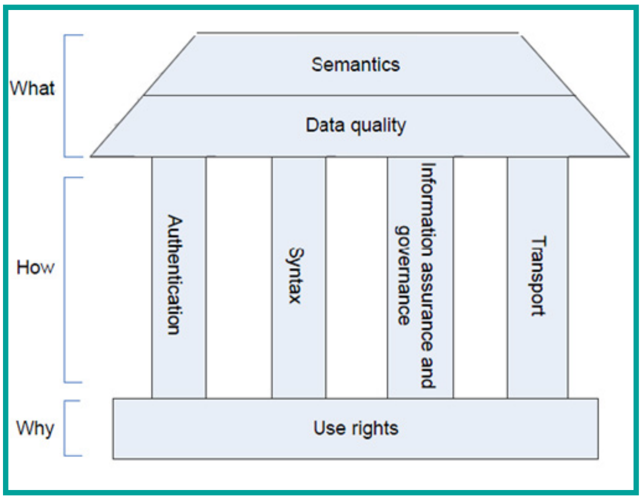


Figure 2 – Seven layers of standards

Figure 2 above shows the different aspects which need to be considered when looking at standards. Some areas in the diagram deal with the more

technical aspects of applying standards – such as authentication, use rights or transport – rather than the definition of the standards themselves.
When considering standards for publishing and linking open data, two of the 'layers' in the diagram are key. These are:

Semantics, or what things are called. Standards used to categorise datasets and enable searching to find data in the first place are known as core reference data. Typically this entails defining names for a concept, then assigning a Uniform Resource Identifier (or URI) which uniquely identifies each 'thing'. A URI is an identifier of something (eg a municipality, a service, an area) that has precise meaning. URIs can normally be entered in the address bar of a web browser to look up their meaning. URIs provide machine readable data and allow computer programmes to link to associated data.

Syntax, or standards defining the way the data is published, enabling aggregation of multiple datasets from different sources.

As a notable example of syntax standard, let's look at a machine readable standard format for publishing inventories of local open datasets. An inventory is a catalogue of datasets held by an organisation. It can include both published and unpublished datasets. It contains references to:

- **datasets**, sets of records however structured – not just datasets according to standard database terminology. Datasets contain:
  - **data resources**, files or feeds/streams of data records available in zero, one or more: **renditions**, different formats of the data, such as: CSV, XML and PDF. Renditions may represent physical files or API calls.
- **document resources**, files or web pages describing datasets. They may include an API page for a dataset and documents describing a dataset (eg rules for data inclusion, how data was redacted). Document resources are expressed as one or more:
  - renditions, different formats of the data, such as: ODF, PDF and HTML.

An inventory lets other people know what data has been published by municipalities. It encourages consistent referencing of datasets across councils to make it easier to find datasets on the same topic or in the same format.
Below are some examples of how and why standards are being developed for sharing open data in partner countries and municipalities.

MAKING OPEN FINANCE (BUDGET AND ACTUAL) DATA COMPARABLE

Financial information, such as details of budgets and spending, is frequently the information about local government most sought after by local newspapers and community, but this also tends to be the data the administration would most like to be selective about publishing. This is for several reasons:

Budgets: The process of budgeting may be spread over 8-9 months before any political decisions are made. 1-2 months of detailed budgeting follows, with the detail drilling down the business hierarchy to the lowest delegated result units.
Typically, budgets at corporate, departmental and at service level are passed in December. These can be and are published, but there are still a lot of details to be decided by the sector's administration and budgeting unit. The work also includes allocation of time periods to budget items, which is very important for later reporting of results. A complete budget is usually not publishable before mid/late February.

The budget should be a complete set of traceable details giving the assumed spending for the coming year. The accounts may have even more details, since some parts of the organisation leave out one or more dimensions of the account-coding from the budget. The basic information provided should be with the following coding:

1. **Organisation**
   usually a sector/service-provider
2. **Subdivision of organisation**
   department, unit, geographical location
3. **Type of expenditure**
   wages, travel, office equipment etc.
4. **Project**
   cross organisational information
5. **Object**
   usually a physical object like a building

The challenge in accounting is to provide transparent and comparable data. To give an example: in accounting each organisational unit is typically identified by a code. If you want the data to be comparable across many municipalities, these codes need to be defined as a set of core reference data (ideally with a URI for each code). So a code would need to be defined for each organisational unit and each function the unit performs, including internal functions such as HR, finance and so on. Additionally, services may be provided by external organisations, which again may have internal services that need to be accounted for.

Ideally, this should be built up as a matrix, assigning codes to different organisations or professional units in a consistent way, perhaps with the sector or organisational unit identified by the first number of the code, management function by the second number, internal functions by second and third, and so on.

This is seldom the case, which makes comparison between administrative services and their cost across organisations, for example, impossible without adjusting the budget and costs to ensure the comparison is accurate. The accounts often require a significant amount of work to make changes to the totals of values in budgets and accounts so figures are directly comparable. To outsiders this may look like figure-fixing, but are simple adjustments to ensure financial information can always be linked to a comparable structure. (This is not the case for all organisations and sectors, but still for so many that it should be taken into account.)

Hopefully, this is less a problem elsewhere. However, in Norway, there is a lot of reporting from local levels to central points (KOSTRA-reporting, public comparable data distributed to services). Before this can be done, numbers must be adjusted and accounts rearranged due to differing organisational structures, accounting-codes, and more.

CATEGORISING DATASETS AND INDICATORS

Combining and comparing datasets is much easier where they take the same format. Formats are defined by individual 'schemas' for datasets. Schemas themselves take different formats. The main ones are:

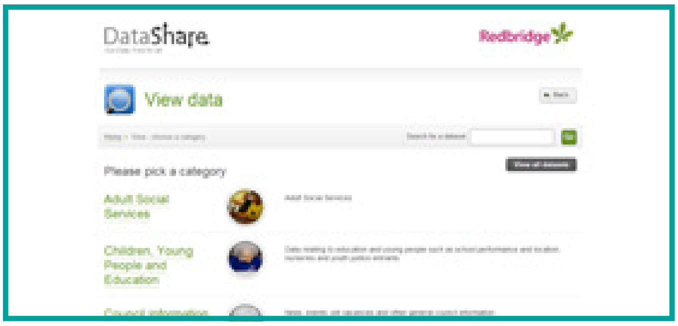
- for tabular data, often expressed via the JSON table schema or the DataShare definition format (see example)
- for richer XML data, expressed via an XML schema
- linked data, which might be expressed via a linked data profile or an ontology

In the UK, a council officer can publish the schemas used to define the datasets they publish on the esd-standards site, cross-referenced against appropriate functions and services. The schemas and the references can be viewed and searched by officers from other municipalities, who can choose to use these same schemas to publish their own datasets. All datasets using the same schema can automatically be aggregated.

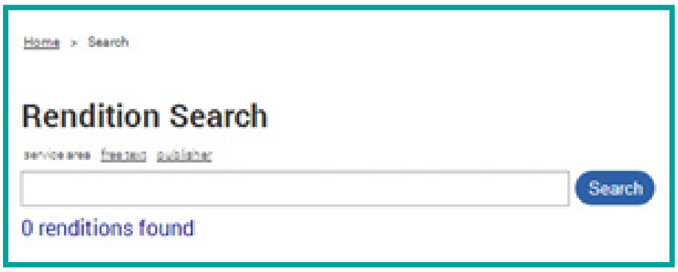
Categorising the datasets according to the functions of a local municipality and the services they may provide make searching and finding schemas for a dataset easier.

Peterborough City Council has worked with data.gov.uk, the London Borough of Redbridge and esd-toolkit under the guidance of the Local eGovernment Standards Body to develop standard formats for popular datasets, and a format for listing published datasets as an inventory. Councils are invited to use these formats as they publish open data so data about the same things can be easily discovered and compared.

The inventory includes properties which identify the council (or other organisation) to which the data belongs, the area described and the subject matter of the data. Where appropriate, datasets are indexed according to esd-toolkit's standard Function List and Service List. This means that you can easily find every council's datasets for a specific service or function. Hence all data on the same topic can be found and combined to get a national picture.



The search tab lets you search through inventories previously uploaded to find datasets applicable to a specified service, topic or location.


Home > Search

Rendition Search

service area free text publisher

0 renditions found

STANDARDISING SEMANTICS

Standardising semantics makes searching for, grouping and analysing data, not to mention comparing results, possible across the borders of countries in the NSR.

Standards are built in most areas and countries (at least in the NSR), but each country seems to build standards for information without consideration of sharing outside their own borders. A European standard has been set for some data, but often, the country standards are still the ones followed.

The Opening Up project has shown that building comparable and consistent data for transnational use is difficult, if not impossible, without a great deal of work and a thorough knowledge of local standards, culture, politics, semantics and regulations.

It is known from earlier projects and experience that services around the North-sea are defined differently. With that as reference, there is also evidence that structure and reporting is different from country to country.

Transferring this information to open data is relevant, and helps us conclude that there is a need for standards within the EU to give public organisations, as well as politicians and the public, the opportunity to do comparative studies between countries.

In Norway, a system for comparing municipalities has been introduced, with compulsory deliveries of data from the municipalities to Statistics Norway, which has overall responsibility for official statistics in Norway. However, no compulsory standards for defining organisations, accounting codes, services or goods have been implemented. The result is that comparisons are doubted and a lot of control activity is introduced for securing the quality of data.

The issues discussed above shows there are several vital standards to be defined from a semantic perspective:

1. Service definitions. Potentially, these can be handled by the EU-service list.
2. Service name. The same service often has several different names, even within regions and countries, suggesting a need for some form of synonym search.
3. Occurrences/incidents that can be of value for planning, well-being, safety and more are labelled differently, often based on countries' and regions' cultural issues. When considering comparability, this can lead to serious misunderstandings.
4. Government systems. There are some deviations in the way central, regional and local government is organised in the different countries, which proves a hindrance when searching for data. All data should be accessible in a common searchable database.
5. In most countries, business, enterprises and industries information and data concerning employment and tax is published and available in digital format, which is very good. However this can leave the companies free to label themselves, update information and so on. This works very well for most companies, but may be a source of misinformation. There are examples (from Norway) of companies no longer in operation but still appearing in searchable data, of merged companies which show as two separate organisations and so on.
6. Accounting standards. Should be able to be handled, but doubtful.
7. Structural standards. Based on services, what sort of data should be included in reporting data and how to report.
8. Tangibles, definition and use should be understood.
9. Demographic data differ and make comparison difficult, such as age-groups, legal status and more.
10. Standards for geographic issues seems to be international and working, even the terms and metrics seems to be in standards that are followed all over the world.
11. And there are more.

In short, with so much work left to do, a series of new EU-projects on semantics and standards has the potential be the greatest added value project in the history of the EU.

What is discussed above is the problem of lack of standards and the following difficulties of semantics. There is no question that these issues are costing the EU and every country, region and municipality a lot of money in additional work (man-hours), lack of capacity for services, not to mention the demands on politicians, civil servants and the public. For the sake of effectiveness and efficiency, the matters of standards and semantics urgently need to be addressed in a serious manner.

STANDARDISING FORMATS AND REFERENCES

A prerequisite to realise the benefits of open and free data is standardisation - both when storing data and in the services you use to display it. This is certainly one of the hypotheses or assumptions worked with throughout the geographic data area in Denmark for several years.

Simon Bent Holm says: "You have got the idea that it's enough to make data freely available - there's probably someone who finds a way to use them. But it's no guarantee. Large amounts of messy data is not different from smaller amounts of messy data."

When creating solutions based on local data for use immediately in several municipalities, it makes sense to harmonise and standardise the underlying data models. This standardisation was one of the drivers in a large inter-municipal project on "Cost effective management on a geographical basis" (OGF). This is a continuation of INSPIRE and other national data projects in Denmark.

It is easier and far more efficient where there is only a single way of identifying something, rather than 98 different ways to recognise the same thing. When the data is open, it is certainly much easier to access for all parties.

Following the OGF decision in 2011 to create a common municipal data model for themes created and maintained in the municipalities and not already part of another joint public data model, there was a resulting list of approximately 80 themes.

The purpose of establishing the logical FKG data model is to identify common and unique identifiers of relevant core concepts of municipal geodata. The principles for which data is covered by the FKG data model are as follows:

- Data generated by the municipality
- Data that supports the workflows in the municipality
- Data used by more than 20% of the municipalities
- Data which is not part of another joint public data model
- Data which is not the result of an analysis of other data
- Data which is not single case data. So no data only relevant for one particular file. In its present form, the FKG data model consists of 74 themes grouped in 14 broad thematic groups.

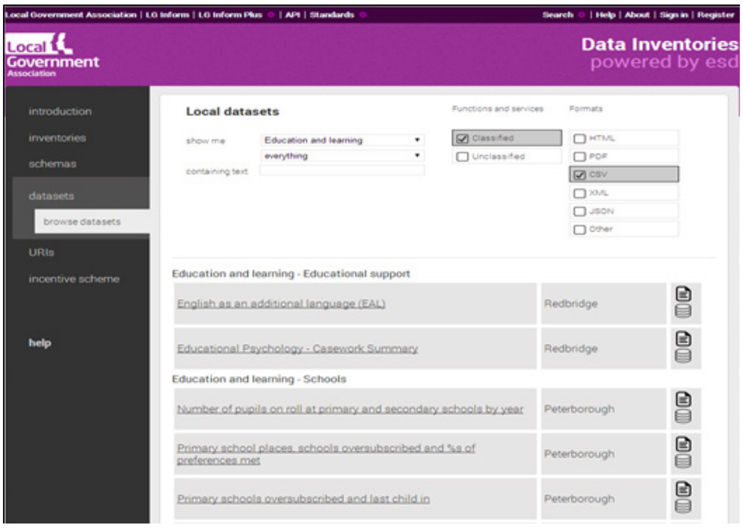
In the UK, the benefits of the emerging infrastructure for councils to share details of their published open data are just becoming visible.

Take a look at the screen shot overleaf from esd's open data page for browsing datasets. It lists council datasets published in specified formats for selected service areas.

The page shows a filtered selection of datasets read from different councils' inventories of open data. Inventories are generated automatically and available for anyone to read from councils using the latest version of DataShare to publish their open data.

The inventory includes both datasets and the schemas that define their structures, so it's possible to see schemas used by different councils for each service area. One council can apply another council's schema if it chooses. Using DataShare, it's a simple process of pointing to another council's schema to create a local dataset of the same format. The Open

Data User Group's local open data incentive scheme provides a DataShare compatible schema for each of its three themes.


Local Government Association | LG Inform | LG Inform Plus | API | Standards
Search | Help | About | Sign in | Register
Local Government Association
Data Inventories
powered by esd
introduction
inventories
schemas
datasets
browse datasets
URIs
incentive scheme
help

Local datasets

show me: Education and learning ▾
everything ▾
containing text:

Functions and services: ☑ Classified ☐ Unclassified

Formats: ☐ HTML ☐ PDF ☑ CSV ☐ XML ☐ JSON ☐ Other

Education and learning - Educational support

English as an additional language (EAL)	Redbridge	
Educational Psychology - Casework Summary	Redbridge	

Education and learning - Schools

Number of pupils on roll at primary and secondary schools by year	Peterborough	
Primary school places, schools oversubscribed and %s of preferences met	Peterborough	
Primary schools oversubscribed and last child in	Peterborough	

The inventory format was defined in work led by Peterborough council. Other councils now aim to implement the same schema in open source portals, so more councils' dataset inventories are published in the same way.

data.gov.uk has written a harvester, which is now in beta testing, for the inventory. As soon as the harvester is released, councils which use the inventory can register for automated harvesting so they no longer need to manually add each of their datasets to data.gov.uk.

Using the above infrastructure, a scheme is being implemented in England to encourage municipalities to publish datasets according to a single national standard for three themes:

- Planning applications
- Public toilets
- Premises licences

The schemas for each of these three datasets have been agreed with experts for publication on the esd Local Government Data Schemas pages.

All councils in England are encouraged with a financial incentive to publish their data for the three datasets.

To comply with the scheme, municipalities are required to:

- publish open data in comma separated variables (CSV) format compliant with a schema which defines the columns and rules for their content
- use URIs from recognised public URI sets for common values like geographical area and service category
- publishing metadata for each dataset on data.gov.uk
- publishing data under open government license
- self-register for an Open Data Institute Open Data Certificate

Municipalities are also encouraged to include their datasets in inventories that comply with the inventory standard.

METHODOLOGIES FOR STANDARDISING ACROSS MUNICIPALITIES

Data model building

In Denmark, data modelling work is coordinated as far as possible between the data model for Denmark Area Information (DAI), the data model for PlansystemDK (PlanDK3) and to a lesser extent with the data model for the FOT. The FOT data model is the basic structure of a more generic model, with joint geometry and attribute tables for all object types. This means that it cannot really be coordinated with the FKG data model.

The data model consists conceptually of the following parts:

- A general data model (data dictionary)
- Standardised fields in the theme-specific data models
- Theme-specific data models
- Any individual fields

The general data model contains a number of fields that are common and compulsory for all themes. These fields consist of both system-generated values and some fields that require user input.

Generic standards are defined independent of any theme and can be used in any of the definitions to link across themes, whilst theme specific fields are unique to the theme for which they have been defined.

There has been no data model work for any individual fields or for specific information regarding individual issues that are only relevant to a single municipality. The individual fields can be well handled, as all FKG data model themes are identified by a feature code that has a value in the range 5000 to 7999. This code range is selected because the codes do not coincide with feature codes in other common public data collections. Basically, the FKG data model follows a general section, one theme specific / common share and possibly an individual component.

The FKG data model, as mentioned, is a logical data model. It does not take a position on how the physical data model is to be built in a database. However, it is important for setting up the physical data model that data can be converted between different database systems without loss of information. This requires that you stick to the agreed field definitions, and ANSI SQL standards for table naming and data types.

Table names and field names in the database must meet the ANSI SQL standard for the use of signs. This ensures that the database can be deployed on all major database platforms, and that data can be moved to another database or the most common file-based desktop GIS tools.

Organising

The work of the Joint Municipal Geodatamodel is organised with a board consisting of representatives from the Association of Municipal Engineering and Local Government Denmark.

Under the Board, two working groups are established. One group handles the development and maintenance of the logical data model and the other physical implementation of various GIS systems. The goal is to establish a joint municipal portal from which data is exhibited through standardised services. The majority of this data will be freely available
Each group works towards the goal of issuing a maximum of two new versions of the data model per year. In the beginning there were many adjustments to the data model as it was implemented in the municipalities. There is always an ongoing dialogue with the professionals in charge of the themes in the municipalities. It may therefore take some time before all the data sets coming through working groups use the standards.

Some municipalities already have data in other or local data models. This data must be converted so that it fits with the common data model. This is a work that requires great enthusiasm and generosity, but the board has an ambitious goal for the work to be completed by the end of 2014!

The Interreg IVB North Sea Region Programme
Investing in the future by working together for a sustainable and competitive region