Core Vocabularies and Grammar for Public Sector Information Interoperability

Chris Harding, The Open Group

The European Directive on the re-use of public sector information focuses on economic aspects, with the aim of stimulating re-use by commercial enterprises. Enterprises cannot discover information that could be valuable to their customers, and integrate it into their products and services, unless the information is clearly described in a way that they can understand. Core vocabularies provide a way of doing this, but they are not by themselves enough. A basic grammar is needed also.

The Open Group is developing the Open Data Element Framework (O-DEF), which has a core vocabulary and a basic grammar for describing atomic units of data. This paper describes the main lessons learned from this development, and their application to public sector information. It is illustrated by a simple example of information use in smart cities.

Commercial Re-Use of Public Sector Information

Commercial re-use is a primary concern of the SHARE-PSI project, which focuses on the practical and technical challenges arising from the Directive. Its workshops have given valuable insights.

A particular conclusion that emerges from several of these workshops is that re-use requires effort on the part of both publishers and re-users. The following points captured in the workshop reports make this clear.

- Even though valuable data is released, additional effort is required to energize external stakeholders to create something with it (Samos).
- Open data, even when freely available, is not free to use since so much time has to be spent cleaning it up, converting it, integrating and maintaining it (Lisbon).
- There is a need to describe the quality of data in a consistent manner if potential consumers are to make informed choices (Timişoara).
- Raw data is of almost no value except to a small number of people with the skills and motivation to work with it. Commercial re-users add value to the raw data by analysis, transformation, and enhancement (Krems).

Applications and Data Models

Use of data generally involves an application program. Such a program will be based on a data model, and this model is often radically different from the data model of the information provider.

For example, the administration of a “smart city” might collect information about temperature, humidity, air quality, etc. at many locations (possibly by distributing measurement kits to citizens, as in Amsterdam). Such information could be stored and made available using a single data model with a single entity – *Reading* – illustrated in the figure.

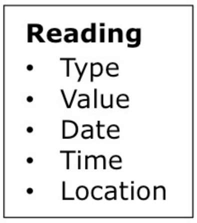

Figure 1: Smart City Information Provider Data Model

A shop or restaurant in the city might have an application that analyses how various aspects of the customers' experience, such as store layout and decoration, featured promotions, and so on, affect what they buy. The outside temperature could be a factor – for example in whether a restaurant customer buys soup or ice cream – and the analysis could take account of it if data is available. A simplified version of the application data model is shown below.

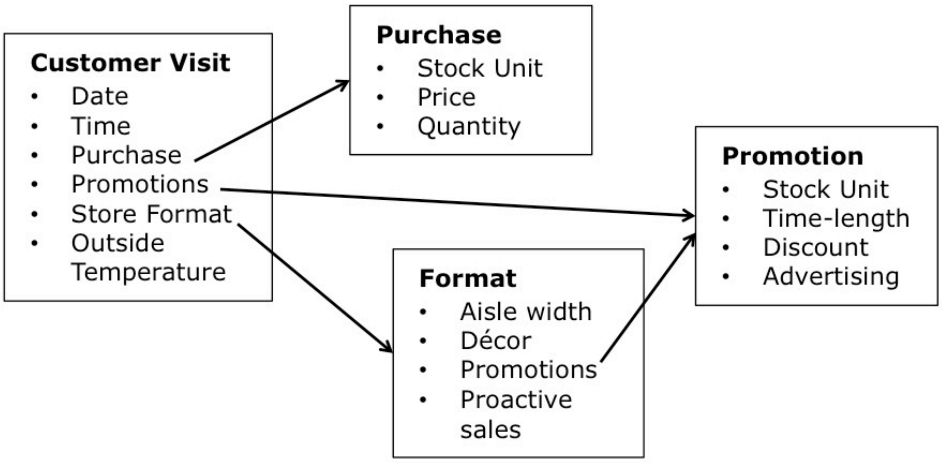

Figure 2: Customer Experience Application Data Model

The city provides the information that is needed, but finding it may not be straightforward. It may not be obvious that a "reading" is a measurement, and that its "value" could be a temperature. The data must be described in such a way that it is clear that the reading is a measurement, the type indicates whether it is of temperature, humidity, or air quality, the value is its value, the date and time are the date and time when it was taken, and the location is where it was taken. The store can then search for temperature readings taken nearby during a customer visit in order to populate its "Outside Temperature" field.

The Berlin meeting of the SHARE-PSI project is about maximizing interoperability. The simple example above illustrates one of the main interoperability problems for re-use of public sector information: understanding the information available, and integrating it into applications, without spending a large amount of effort.

The question of how much effort is needed is a crucial one. It is always possible to understand and integrate information if the provider is prepared to spend sufficient time explaining it, and the user is prepared to spend sufficient time learning about it and writing transformation programs. Infinite time is not available, either in the commercial sector or the public sector. It must be possible to create data definitions without much effort, and those definitions must enable the data to be used without much effort. Increasingly, this means that the definitions should be able to be interpreted by software tools for discovery and integration.

Core Vocabularies

Core vocabularies are a key topic of the Berlin workshop. They address part of the problem. If the term “Reading” is in a core vocabulary, or is identified as being equivalent to a term in a core vocabulary, then it will be understood.

There are a number of established vocabularies that can be used as core vocabularies for interoperability. They include vocabularies with machine-interpretable definitions such as that of the Dublin Core® Metadata Initiative and the core vocabularies of the Semantic Interoperability Community (SEMIC). They also include vocabularies whose definitions are human-interpretable, such as the United Nations Standard Products and Services Code® (UNSPSC®) and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT).

As well as core vocabularies containing commonly-used terms, application-specific vocabularies are needed to describe some items of data. For example, the customer experience application has a *proactive sales* flag to indicate whether a salesperson greets every customer and asks what they want, or whether the sales staff stay behind the counters and wait for customers to approach them. A piece of data with precisely the same meaning is unlikely to be used by other applications, and *proactive sales* belongs in an application-specific vocabulary rather than a core vocabulary.

The O-DEF is derived from vocabularies previously published by The Open Group that contained terms commonly used by business enterprises, with parallel versions in Dutch, English, and French. Experience with these vocabularies leads to conclusions that are interesting, not so much as regards the terms themselves, but more as regards the basic grammar that is used to make data descriptions from them.

Basic Data Description Grammar

Why is a basic grammar needed? It is needed so that combinations of vocabulary terms can be interpreted consistently, by people and, more importantly, by software programs. How do you interpret “reading value”, for example? It might mean the benefit a child gains from learning to read. But, if you know that “reading” is an object class and “value” is a property then you can interpret it as intended in the example given earlier.

A basic grammar is particularly important for machine interpretation; people are still much better than software at dealing with ambiguity and uncertainty.

Don’t we have one already? We have natural languages such as English, we have conceptual frameworks for data management, particularly that of relational databases, and we have the Resource Description Framework (RDF) of the World-Wide Web Consortium.

We need a grammar for data description. Natural language grammars enable us to express thoughts and emotions, and to describe things, including data, but they are not specialized for data description. Data description requires grammatical constructs such as *object class* and *property*, rather than *noun*, *verb*, etc.

The grammar must be consistent with use of relational databases. The relational database approach to data storage and management has been established for many years, and is in common use by the majority of enterprises today. It includes the discipline of data modeling, with a grammar containing constructs such as *entity*, *relationship*, and *attribute*. It has the ISO 11179 standards for metadata registries that define constructs including *object class* and *property*.

The grammar must also be usable for data expressed in other commonly-used languages, such as XML and JSON.

RDF is an excellent and established semantic standard, with defined machine-interpretable representations, which can be processed by a growing body of sophisticated software. It is highly appropriate as a technical basis for data descriptions, but it is more oriented to describing the real world than to describing data. It is easy to use RDF and RDF Schema to express ideas such as *person* or *name*, not so easy to express ideas such as *piece of data giving the name of a person*.

We have many ways of expressing data, and we have conceptual frameworks for understanding data, but we do not have a commonly-accepted basic grammar for describing data.

Conclusions

Commercial re-use of public sector information requires effort by the commercial companies concerned. Discovery of relevant information and integration of that information into applications must be easy, and able to be supported by software tools, for commercial re-use to be viable.

Discovery and integration of information is made easier if that information is clearly described. In many cases it is impossible if there are no descriptions. To enable software support, the descriptions must be machine-processable.

Descriptions should use core vocabularies containing commonly-used terms to enable interoperability with and between applications. They will also need to use application-specific terms.

It should be possible to use the existing vocabularies that have been developed by industry bodies and standards organizations.

A basic grammar is needed for the data descriptions. This should be consistent with relational database usage and able to accommodate other data representation approaches.

Vocabularies and grammar should be able to be expressed using RDF and RDF Schema to facilitate processing by semantic software.

References

The Directive. DIRECTIVE 2003/98/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 November 2003. Refer to http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:02003L0098-20130717

The SHARE-PSI Project. Refer to http://www.w3.org/2013/share-psi/

Open Public Sector Data Business Scenario. The Open Group publication K140, December 2014. Refer to https://www2.opengroup.org/ogsys/catalog/K140

Amsterdam Smart Citizen Kit. Refer to http://amsterdamsmartcity.com/projects/detail/id/69/slug/smart-citizen-kit

Dublin Core Metadata Initiative. Refer to http://www.dublincore.org/

Semantic Interoperability Community (SEMIC). Refer to https://joinup.ec.europa.eu/community/semic/description

Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). Refer to http://systems.hscic.gov.uk/data/uktc/snomed

Universal Data Element Framework (UDEF). Refer to http://www.opengroup.org/udef

Resource Description Framework (RDF). Refer to http://www.w3.org/standards/techs/rdf

ISO/IEC 1179 Parts 1-6: Metadata Registries. Refer to http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=45342