A Classification Vocabulary

Introduction

We start with a simple axiom: the less you ask people to agree upon, the more likely they are to agree. In other words, the simpler the model, the more chance there is that people will see that their own data fits your model.

With that in mind, the aim of the Classification Vocabulary is to provide the basic framework that can provide maximum interoperability between diverse classification systems. To achieve this we must separate out the vocabulary terms from the data itself.

Professional classification systems typically provide some combination of:

- a classification, a.k.a age rating, expressed as a string, that may or may not include numbers (for example, 'AL', '12', 'R18');
- a classification icon (an image that conveys the classification);
- short explanatory text ('contains moderate violence' etc.)
- long form text;
- icons that convey the reasons for the rating (such as the PEGI/Kijkwijzer icons);
- descriptors.

Few if any schemes use all of these elements but subsets are all in widespread use.

Professional classification is always carried out according to what in data modelling would be called a controlled vocabulary, that is, the categories are pre-defined and any piece of content is slotted in to one of those categories. Each category will be a summation of information about the work and that is provided through various means including icons, descriptors or text.

Finally, we must take account of practicality: the model must actually be usable and useful.

This gives us our basic model. We have a Class – a concept – of a classification that can have a number if different properties:

- age rating
- numeric rating
- classification icon
- short text
- long text
- descriptor icon
- descriptor

Each of these properties is discussed below. For the technically minded, details of how this would be modelled using RDF/Linked Data is also provided. The domain of

all properties listed here is `cfn:Classification` where `cfn` is the prefix for the Classification Vocabulary. Linked Data is the most flexible method of encoding the Classification Vocabulary but other technologies may be used, notably XML.

Non-technical participants may safely ignore all technical considerations. The key thing is the model, that is, the concept of a classification (for a given piece of content) and the properties of the classification.

Property age rating

This is the familiar AL, PG, M, 15 or whatever. Since these are always drawn from a well defined list, and each value is backed by carefully worded definition, these age ratings are themselves 'concepts.' The age ratings should have their own identifier that itself is machine readable (i.e. a URI) so that users can look up what 'AL' means, e.g. `http://example.org/class/AL`

Conceptually the value of the age rating property is drawn from a controlled list and is identified by a URI.

In RDF terms, `cfn:ageRating` has range `skos:Concept`.

Property numeric rating

This is where practicality takes over. One can imagine that somewhere within a computer program that is making use of the data there will be a line something like:

```
if the age of the user < the numeric rating, do X.
```

That only works if the age rating is a number. Conceptually therefore, the value of this property is a whole number.

In RDF terms, `cfn:numericRating` has range `xsd:integer`.

Property classification icon

Classification schemes typically have well known icons associated with them that are often seen on printed packaging and promotions as well as online. This property links a classification to its relevant icon. It is distinct from the descriptor icon below.

Conceptually, the value of this property is an image.

In RDF terms, `cfn:classificationIcon` has range `xsd:anyURI`

Property short text

This would be a short piece of text that sums up the reason for the classification. 'Contains moderate violence' etc. Where the text is available in multiple languages, the property should be used once each per language.

Conceptually, the value of this property is text for which the language may or may not be identified.

In RDF terms, `cfn:shortText` has range `rdfs:Literal`.

Property long text

This would be a longer piece of text that gives a detailed explanation of the content from a classifier's point of view (examples include BBFC's Extended Consumer Information). Where the text is available in multiple languages, the property should be used once each per language.

Conceptually, the value of this property is text for which the language may or may not be identified.

In RDF terms, `cfn:longText` has range `rdfs:Literal`.

Property descriptor icon

Some classification systems employ icons to convey the reasons for a given classification. The icon property can be used any number of times to point to the appropriate icon(s).

Conceptually value for this property is an image.

In RDF terms, `cfn:descriptorIcon` has range `xsd:anyURI`.

Property descriptor

Some classification systems employ descriptors to convey the reasons for a given classification. It is tempting to cast descriptors as individual properties but this is a mistake. The selection of, say, gambling, as a descriptor is common, however, it is a choice made the classification system and should not therefore be assumed to be universal. It is perfectly possible to have a classification system that takes no account of the presence or absence of gambling.

A comparison of App store descriptors illustrates this:

Google Play	Apple	RIM	ESRB / CTIA
Alcohol, tobacco, drugs	Alcohol, tobacco, drug use / refs	Alcohol, tobacco, drug use	Controlled substance
Gambling	Simulated gambling	Gambling	Gambling
Profanity / crude humor	Profanity / crude humor	Language	Language / Crude humor
Sexual / suggestive content	Sexual content / nudity	Sexual content	Sexuality
	Graphic sexual content / nudity		
Violence	Cartoon / fantasy violence	Violence	Violence
	Realistic Violence		
	Prolonged graphic / sadistic realistic violence		
User Generated Content & User to User Communication		Social networking, community or other user generated content	Shares personal information; Users create / exchange content
	Mature / suggestive themes		
	Horror / fear themes		
Hate			
Location			Displays user's location to others
		Minimum age in Terms of Use of Application	Minimum age requirement

Taken from Annie Mullins' presentation from the CEO Coalition meeting 14th September 2012.

Rather than try to encode all these different descriptors, and those not shown here but used by others as properties, it is simpler to cast these as terms in a controlled vocabulary. Google, Apple, RIM and ESRB/CTIA will have slightly different controlled vocabularies but those terms can each be used as a value for the descriptor property.

Publishers can use the Classification Vocabulary's descriptor property and give the value as terms taken from their choice of controlled vocabulary. As with the age ratings, each descriptor will have its own URI that can be looked up to provide more details, e.g. `http://example.org/descriptor/gambling`.

Conceptually, the value of the descriptor property is a URI that identifies a term from a list of potentially harmful content types.

In RDF terms, `cfn:descriptior` is `skos:Concept`.

Cardinality and Application Profiles

The preceding discussion has not said anything about specific properties being *required* or *required once* etc. Those restrictions, known as cardinality constraints, are likely to be important in specific contexts. For example, an app store may *require* classifications to include values for the age rating and short text properties; others may require numeric age rating and classification icon etc. These are known as *application profiles* of the Classification Vocabulary. Profiles may also stipulate *which* descriptor set to use. In other words, an application profile can be highly prescriptive in how the Classification Vocabulary is used in a specific context without restricting other users.

It's that flexibility around a common set of terms that maximises the task force's chances of achieving consensus.

Interoperability

If different applications use different descriptors and have different rules about which properties are required and which are optional, how does that help interoperability?

Interoperability comes about because developers know what to do with the data they can get. If an application requests details of a classification and gets back an age rating, a classification icon and some short text, it knows what to do with it, *irrespective* of the classification scheme. The end user will be presented with the information and they can then make an informed choice.

Summary

The Classification Vocabulary presented here is no more than an outline suggestion of how to achieve the following goals:

- a common system that can be used by any existing or future classification system;
- a system for making those classification schemes interoperable without any need to harmonise their content and age boundaries;
- a system that is conformant with the wider principles of Open Data.

Phil Archer
12 November 2012
Last modified 13 November
phil@philarcher.org