Preferred nature of vocabularies

8 June 2017

Approved by the TSC as FHISO technical policy on 8 June 2017.

This is a FHISO policy developed after the November 2015 Board meeting resolved that "the TSC will draft and release … a general policy on vocabularies". It was released by the TSC for wider comment on the tsc-public mailing list and has been updated to reflect the consensus reached on that list.

Although neither absolutely binding nor completely immutable, future FHISO standards and other technical documents should follow this FHISO policy insofar as it applies, unless there are exceptional circumstances not to do so.

The key words MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, MAY, and OPTIONAL in this document are to be interpreted as described in RFC 2119.

The examples given in this documented are intended as illustration only. This document does not define any actual *vocabularies*. As actual *vocablaries* are defined, the examples herein may be updated to reflect them.

Vocabularies and terms

The World Wide Web Consortium (W3C) answers the question "What is a vocabulary?" as follows:

> Vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms. In practice, vocabularies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only).

A **vocabulary** is defined to be a set of related *terms*, where each *term* identifies some specific concept, idea or relationship. How the *terms* in a *vocabulary* are related to each other is not specified, and in particular there is no assumption that they can be used in the same context, for example as possible values of some field. A *vocabulary* is simply a collection of *terms* that might usefully be discussed together, for example, all the *terms* defined in a particular standard or section of a standard. Given this rather nebulous definition, the word SHOULD be avoided in FHISO standards and other technical documents if a more precisely-defined word or phrase is available.

A **term** consists of a unique, machine-readable identifier, known as the *term name*, paired with a clearly-defined meaning for the concept or idea that it represents. This meaning SHALL be written in a natural language (such as English), though aspects of the definition MAY also be available in machine-parsable form.

Term names

A **term name** SHALL canonically be an IRI matching the `IRI` production in §2.2 of RFC 3987. IRIs have been chosen in preference to URIs because it is recognised that certain culture-specific genealogical concepts may not have English names, and in such cases the human-legibility of IRIs is advantageous.

FHISO standards MAY define new *terms*, and MAY use *terms* defined in third-party standards. Where possible all *terms* used in FHISO standards SHOULD:

— be defined in an open standard;
— have a scheme of `http` or `https`;
— be defined on a domain under the control of the standard's authors; and
— be a IRI which, when visited, provides a human- or machine-readable description of the meaning of the vocabulary *term* in a common, open format.

In additional, *terms* defined in FHISO standards SHALL use the following IRI patterns:

— A scheme of `http`.
— An authority with just the host `terms.fhiso.org`.
— A path constructed from the following slash-separated path segments, each of which SHOULD match the `NCName` production in the XML Namespaces specification:
    1. zero or more short names used to name of the *vocabulary* (or the literal string `type` — see below in the section on *classes*);
    2. a short name for the *term*.
— No query or fragment.

Although this permits an arbitrary number of path segments between the literal `term` and the name of the *term*, exactly one path segment SHOULD normally be used. For example, a *term* named "birth" in some future FHISO "events" vocabulary might have the *term name* `http://terms.fhiso.org/events/birth`.

The use of additional segments to partition a *vocabulary* into parts is permitted but NOT RECOMMENDED in normal circumstances. It is also permitted but NOT RECOMMENDED to have zero path segments between the literal `term` and the name of the *term*; this should only be done if a term has sufficiently wide applicability that it is not naturally part of one specific vocabulary.

Term names are case-sensitive, but FHISO SHOULD NOT define multiple *terms* that differ only in their capitalisation. Implementors are warned that some current third-party standard do contain *terms* differing only in their capitalisation, and FHISO standards MAY use such *terms*.

A **namespace** is an abstract container for *term names* that all share a common prefix, where the common prefix is itself a valid IRI and is known as the **namespace name**. *Namespace names* SHOULD normally end with a delimiter character such as `/` or `#`. For example, the *term* `http://terms.fhiso.org/events/birth` is within the *namespace* with a *namespace name* of `http://terms.fhiso.org/events/`. The content of a *namespace* is a *vocabulary*, but not all *vocabularies* equate with a whole *namespace* — it can also be convenient to talk about *vocabularies* that span several *namespaces*, and also *vocabularies* that are a subset of a namespace.

Comparison of *term names* uses the "simple string comparison" algorithm given in §5.3.1 of RFC 3987. This is how XML Namespaces compares namespace IRIs and it involves no normalisation of *term names* before comparing.

Applications MUST NOT make inferences about the meaning or usage of a *term* based solely on its *term name*.

Compact term names

FHISO standards or other technical document MAY allow *terms* to be written in ways other than as IRIs, and if so SHALL provide an unambiguous definition of how such *terms* are converted to and from their canonical form as an IRI. This conversion process SHALL NOT depend on anything external to the document or data stream.

FHISO's preferred way of shortening *term names* is to use some form of **compact term name**. A standard doing so SHOULD define a mechanism like that in XML Namespaces to bind a *namespace name* to shorter, more convenient identifier called the **namespace prefix**. A *compact term name* comprises a *namespace prefix*, followed by a separator (which will typically but not necessarily a colon, depending on the host language), followed by a **local part**. The *term name* in IRI form is found by concatenating the *namespace name* corresponding to the *namespace prefix* with the *local part*. For example, if the IRI `http://terms.fhiso.org/events/` is bound to the prefix `ev`, then `ev:birth` could be the compact representation of the term `http://terms.fhiso.org/events/birth`.

Compact term names might take the syntactic form of a `QName` in XML Namespaces, allowing *terms* to be used as element or attribute names in XML formats. In an XML 1.0 document, *namespace names* are URIs, not IRIs; algorithms for converting IRIs to and from URIs can be found in §3 of RFC 3987. More generally, *compact term names* might take the form of a CURIE; and CFPS 37 gives an example of how they might be used backwards-compatibly in GEDCOM.

IRI resolution

An HTTP 1.1 GET request made without an `Accept` header to a *term name* IRI (once coverted to a URI per §3.1 of RFC 3987) SHOULD result in a 303 "See Other" redirect to a document containing a human-readable definition of the *term*. This document SHOULD have a `text/plain` or `text/html` content-type, and SHOULD use either an ASCII or UTF-8 encoding which SHOULD be explicitly specified in the content-type.

A 303 redirect is considered best practice for Linked Data to avoid confusing the concept represented by the *term* with the definition of that *term*, which can be found at the post-redirect URL. Following W3C policy on vocabulary IRIs, *term* IRIs defined by FHISO MUST NOT result in a 200 response unless the *term* actually denotes the document being retrieved rather than a concept defined in it.

The webserver serving the *term* IRI SHOULD support content negotiation (per §3.4.1 of RFC 7231) which MAY allow *term* definitions to be fetched in additional human-readable formats. Future FHISO policy is expected to define a **discovery** mechanism by which *vocabulary* authors can provide machine-readable information to applications on the properties and expected usage of otherwise-unknown *terms*. This will involve applications making a GET request to the *term name* IRI with an appropriate `Accept` header, and receiving a 303 "See Other" redirect to a machine-readable resource in the expected format. Support for *discovery* will be OPTIONAL for clients and RECOMMENDED for servers.

Webservers that do not support content negotiation SHOULD (and other webservers MAY) provide a `Link` header, as defined in RFC 5988, to locate the machine-readable description of the *term*. Clients that support *discovery* SHOULD support this mechanism too.

Classes

A **class** is a *term* used to denote the set of values or entities that may be used in some particular context. There might, for example, be an "individual" *class* to denote people of genealogical interest; the nature of the such entities is beyond the scope of this policy, but this policy permits a *class* to be defined representing individuals, howsoever they may be represented.

Other *classes* might represent various types of literal, such as strings, dates, integers, and booleans. Such general *classes should not* be defined in domain-specific *vocabularies* as they are likely to required by many FHISO standards. It is anticipated that a future FHISO policy will provide for common definitions of these basic *classes*, possibly by reference to a third-party standard such as XML Schema Datatypes. Where domain-specific *classes* are required, for example to represent strings in some microformat, these MAY be defined included in the appropriate *vocabulary*.

Finally, *classes* can be used to represent a collection of *terms*. Such *classes* of *terms* are examples of *vocabularies* and are known as **vocabulary classes**. Not all *classes* are *vocabularies*, as the individual and integer examples demonstrate.

An example of a *vocabulary class* might be an "event type" *class* consisting of terms for "birth", "burial" and such like. A FHISO standard that defines a *vocabulary class* SHALL state whether or not it is **extensible**: that is, whether or not third parties are be permitted to define additional *terms* of that *class*. Due to the widely variable nature of genealogical data, *vocabulary classes* SHOULD be extensible unless there are compelling reasons to the contrary.

(Some authorities use different nomenclature such as "enumerations" or "controlled vocabulary" to describe these concepts. FHISO does *not recommend* the use of this terminology as it is used ambiguously in the literature: some authorities use them for any *vocabulary class* while others reserve them for just those that are not *extensibile*.)

As for any *term*, the *term name* of a *class* is an IRI, called its **class name**. These MAY be put in a *vocabulary*-specific *namespace* alongside other *terms*, for example `http://terms.fhiso.org/events/Type`; alternatively, they MAY be placed in the `http://terms.fhiso.org/type/` *namespace*, known as the **type namespace**. *Classes* SHOULD be put in *type namespace* if they are too general to belong in a

vocabulary-specific *namespace*, or if the natural choice of name would conflict with another *term* in the *vocabulary*-specific *namespace*.

FHISO standards SHOULD respect the convention that *class* names have a upper-case first letter, for example `http://terms.fhiso.org/type/Sex`.

Properties

A **property** is a *term* used to identify a particular attribute of an entity, where the attribute has an associated value, which may simply be a boolean. A link between two entities can also be expressed as a *property* where the value is another entity (or a reference to another entity; the distinction is beyond the scope of this policy).

Any FHISO standard that defines a *property* SHALL also define:

— its **domain** — the context in which it is to be used;
— its **cardinality** — whether it can appear multiple times on the same entity, and if so whether any meaning is ascribed to their order; and
— its **range** — the expected form of the *property*'s value, such as whether it is free-form text, an integer or date, another *term* (of a particular *class*), or another entity of some form.

The *domain* and *range* of a *property* SHOULD normally be specified in terms of a *class*. For the *domain* this is the *class* of entity on which the *property* may be used; for the *range* it is the *class* of the *property*'s value.

In some cases when the expected value of a *property* is another *term* (i.e. when its *range* is a *vocabulary class*), the natural choice of name for the *class* of values might be the same as the obvious choice of *property* name. For example, the *property* denoting an individual's sex and the *class* of possible sexes might both naturally be called `http://terms.fhiso.org/indi/sex` (or differ only in capitalisation). This is an example of when the *term namespace* should be used.

FHISO standards SHOULD respect the convention that *property* names have a lower-case first letter; some example *properties* include `http://terms.fhiso.org/indi/occupation` and `http://terms.fhiso.org/indi/religion`.

Properties shall not have default values, and no information shall be assumed from the absence of a *property*.

Compatibility

Once standardised, the definitions of *terms* MUST only be change in a backwards-compatible way. As *term names* have no version number embedded in them, this means their meaning MUST essentially stay unchanged, and neither expand nor contract in scope in a future standard.