Published Subjects: Introduction and Basic Requirements

OASIS Published Subjects Technical Committee Recommendation, 2003-06-24

Document identifier:

pubsubj-pt1-1.01-cs

Location:

http://www.oasis-open.org/committees/documents.php?wg_abbrev=tm-pubsubj

Editor:

Steve Pepper <pepper@ontopia.net>

Contributors:

Bernard Vatant (TC Chair), Suellen Stringer-Hye (TC Secretary), James David Mason (TC Liaison with ISO/IEC JTC1/SC34), Thomas Bandholtz, Vivian Bliss, Patrick Durusau, Peter Flynn, Eric Freese, Lars Marius Garshol, Kim Sung Hyuk, Motomu Naito, Eamonn Neylon, Mary Nishikawa, Michael Priestley, H. Holger Rath, Don Smith

Abstract:

This document provides an introduction to Published Subjects and basic requirements and recommendations for publishers of Published Subjects.

Status:

Committee Specification

Table of Contents

1. Introduction

The goal of the OASIS Topic Maps Published Subjects Technical Committee is to promote Topic Maps interoperability through the use of Published Subjects. A further goal is to promote interoperability between Topic Maps and other technologies that make explicit use of abstract representations of subjects, such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL).

Published Subjects as defined in this Specification provide an open, scaleable, URI-based method of identifying subjects of discourse. They cater for the needs of both humans and applications, and they provide mechanisms for ensuring confidence and trust on the part of users. Published Subjects are therefore expected to be of particular interest to publishers and users of ontologies, taxonomies, classifications, thesauri, registries, catalogues, and directories, and for applications (including agents) that capture, collate or aggregate information and knowledge.

This document provides an introduction to Published Subjects and defines requirements and recommendations for publishers.

> NOTE: Terms shown in **boldface** in this Specification are defined in the Topic Maps specifications. At the time of writing the ISO standard is undergoing a revision that will lead to the replacement of [ISO 13250:2003] and [XTM 1.0] by a new multipart standard. Definitions are therefore best sought in the [SAM] Committee Draft.

2. A Gentle Introduction to Published Subjects

2.1 Subjects and Topics

A **subject** can be an individual, like Isaac Newton, the apple that fell on his head, or a document (such as this one). It may be a class of individuals, like scientists, fruits, or OASIS specifications. It may also be a more abstract concept like gravity or inevitability. In short, a subject can be any *subject of discourse* that an author wishes to identify, name, represent, or otherwise make assertions about.

Applications deal with subjects through formal representations using symbols as proxies. The term **topic** is used for such symbols in this document, in order to conform to Topic Maps terminology (other formalisms, for example, RDF, use other names for what amounts to the same thing). A topic is thus the representation, inside some application, of a unique, clearly identified, and non-ambiguous subject.

> NOTE: The term 'subject' is defined in Topic Maps as "anything whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever." It corresponds exactly to the term 'resource' in RDF (defined in [RFC 2396] as "anything that has identity").

> In Topic Maps, 'resource' is used in the more restricted sense of "network retrievable information resource." In order to avoid confusion, this document will never use the term 'resource' in the RFC 2396 sense. It will only be used in the sense "network retrievable information resource" and will always be qualified, usually as "information resource," unless the meaning is absolutely clear from the context.

2.2 The Identification of Subjects

The goal of any application which aggregates information, be it a simple back-of-book index, a library classification system, a topic map or some other kind of application, is to achieve the "collocation objective;" that is, to provide *binding points* from which everything that is known about a given subject can be reached. In topic maps, binding points take the form of topics; for a topic map application to fully achieve the collocation objective there must be an exact one-to-one correspondence between subjects and topics: Every topic must represent exactly one subject and every subject must be represented by exactly one topic.

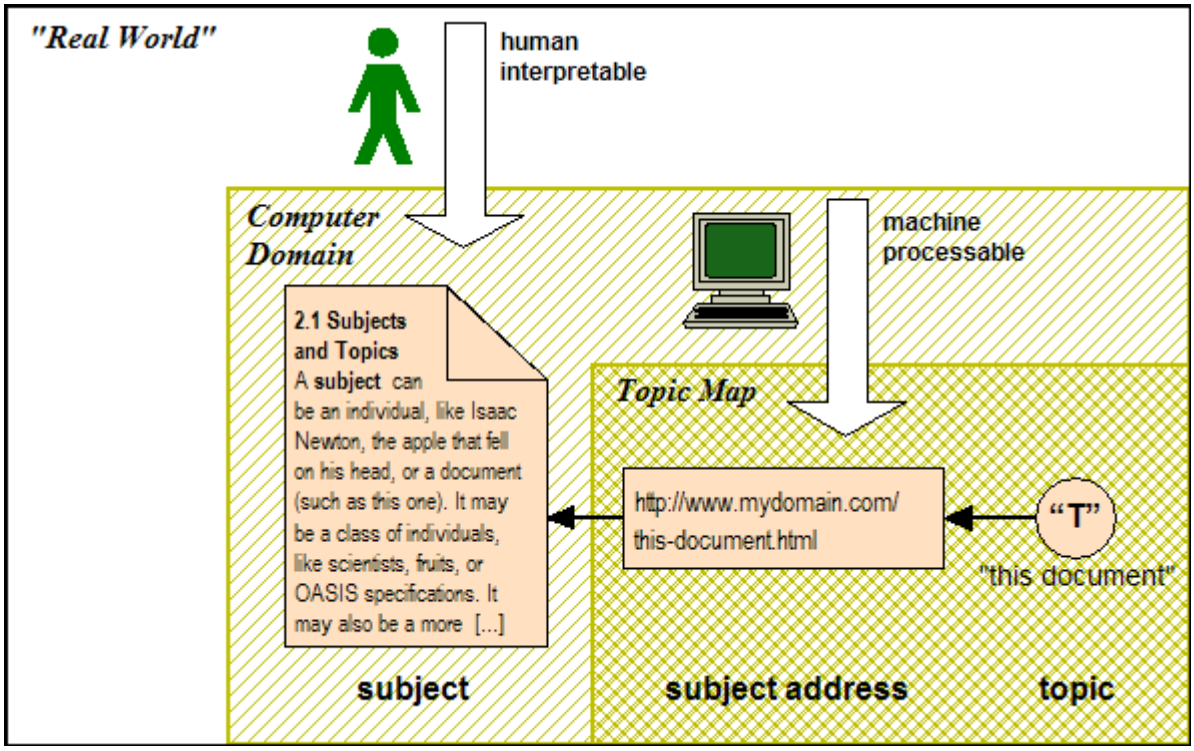
When aggregating information (for example, when merging topic maps), comparing ontologies, or matching vocabularies, it is crucially important to know when two topics represent the same subject, in order to be able to combine them into a single topic. To achieve this, the *correspondence* between a topic and the subject that it represents needs to be made clear. This in turn requires subjects to be identified in a non-ambiguous manner.

The identification of subjects is not only critical to individual topic map applications and to interoperability *between* topic map applications; it is also critical to interoperability between topic map applications and *other applications* that make explicit use of abstract representations of subjects, such as RDF.

2.3 The Addressability of Subjects

The most widespread method of identifying subjects in today's networked environments is through the use of *addressing,* and, more particularly, through the use of URIs. This method works fine when a subject has an unambiguous address, which is the case for "this document" (in our examples, above) and for (network-retrievable) information resources in general.

Any information resource can (at least in theory) be retrieved and displayed for human interpretation. In addition, its address can be used as a unique identifier by computers for the purpose of automated merging (since if two such resources have the same address they are, by definition, the same thing). The address of a subject which happens to be an information resource is called, quite simply, a **subject address**.



A scenario that uses a subject address

In the scenario shown above, the subject about which assertions are being made is a information resource ("this document"). It therefore has a network address which can be used as a unique identifier by topics that represent it. When merging topic maps an application compares subject addresses: If two topics have the same subject address then, by definition, they represent the same subject and will be merged.

> NOTE: The syntax for establishing the identity of such a subject in XTM (XML Topic Maps) is as follows:

```
<topic id="this-document">
  <subjectIdentity>
    <resourceRef
      xlink:href="http://www.mydomain.com/this-
document.html"/>
  </subjectIdentity>
  <!-- names and occurrences -->
</topic>
```

2.4 Subject Indicators and Subject Identifiers

However, most subjects are not information resources and thus do not have unambiguous and resolvable addresses. This is the case for all of the examples given above (with the exception of "this document"). The identity of such "non-addressable" subjects can only be established *indirectly*. Topic Maps provides a approach for doing this which may also be used in other paradigms, such as RDF. That approach is based around the notion of *subject indicators*.

2.4.1 Subject Identification for Humans: Subject *Indicators*

A **subject indicator** is an information resource which provides some kind of compelling and unambiguous *indication* of the identity of a subject to humans. It may be a textual definition, description or name; it may be a visual, audio or other representation of the subject; or it may be some combination of these. A subject indicator is distinct from the subject that it indicates.

Equipped with a subject indicator, human users should be able to know exactly what subject is being referred to. Whenever applications are considered media for human transactions, subject indicators provide a common reference to human users connected through the application, and agreement on the subject indicator can be used as the external expression of agreement as to the identity of a subject.

2.4.2 Subject Identification for Computers: Subject *Identifiers*

Since subject indicators, by definition, are information resources, they can always be given unambiguous addresses. Those addresses can then be used by computers as a basis for establishing identity (i.e., as unique identifiers for purposes such as automated merging) in *almost* the same way as computers use the addresses of addressable subjects: If two topics have the same subject indicator, then by definition they represent the same subject and should be treated accordingly. From the perspective of an application, if the identity of two topics is established by subject indicators that have the *same address*, they should be regarded as representing the same subject. The address of a subject indicator is called a **subject identifier**.

Subject indicator and subject identifier are two faces of the *same identification mechanism*, or two sides of the same coin; the former being for humans and the latter for applications. This identification mechanism is the basis for agreement on the identity of subjects throughout the network: between applications, between users, and between applications and users.

2.4.3 Distinguishing between Subject Addresses and Subject Identifiers

The word "almost" was emphasised in the preceding section in order to underscore the fact that applications need to know whether a URI is being used as a subject address or as a subject identifier, that is, as the address of an (addressable) subject or as the addresses of a subject indicator. An information resource may both *indicate* a subject and be a subject in its own right.

Topic Maps allow this distinction to be made in both model and syntax (using `<resourceRef>` and `<subjectIndicatorRef>` elements, respectively), as the syntax examples show. Not all paradigms that make explicit use of abstract representations of subjects may have the same capability.

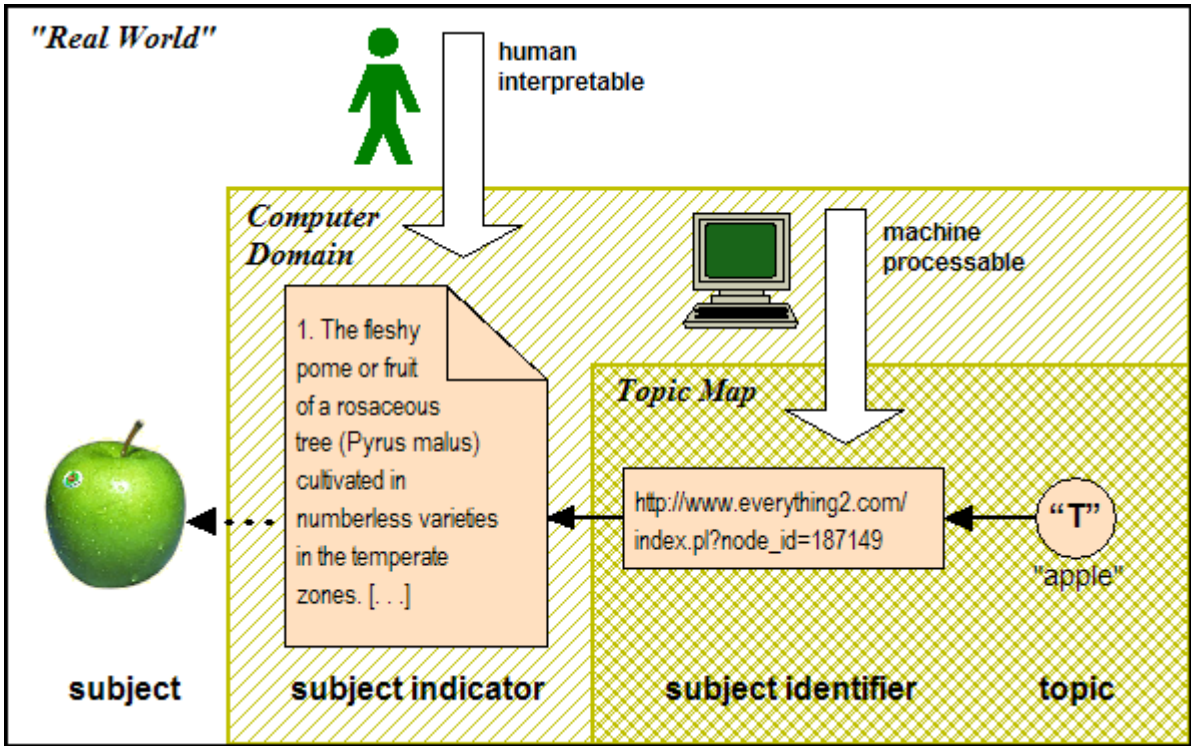
2.4.4 Example: Identifying the Subject "Apple"

To see how subject indicators work, let us imagine that you wish to make a statement about apples - perhaps that they are used to make cider - and that you wish to distinguish apples as fruit from the company Apple and the version of the APL programming language called Apple that was developed for the Illiac IV.

Your "apple" is obviously not an information resource; using a subject address is therefore not an option because it doesn't have one! You choose instead to establish the identity of your subject using a subject indicator, which means you need an information resource that provides some compelling indication of the identity of "apple". You could create your own, or you could use something that already exists. One possible choice would be the classic definition from the 1913 edition of Websters Unabridged Dictionary:

> 1. The fleshy pome or fruit of a rosaceous tree *(Pyrus malus)* cultivated in numberless varieties in the temperate zones.
> Note: The European crab apple is supposed to be the original kind, from which all others have sprung.
> 2. *bot.* Any tree genus *Pyrus* which has the stalk sunken into the base of the fruit; an apple tree.
> 3. Any fruit or other vegetable production resembling, or supposed to resemble, the apple; as, apple of love, or love apple (a tomato), balsam apple, egg apple, oak apple.
> 4. Anything round like an apple; as, an apple of gold.

Let us assume you decide that this definition is sufficient for your purposes. You find it online at *Everything2* on a page with the URL http://www.everything2.com/index.pl?node_id=187149. When this resource is used as a subject indicator the resulting scenario is as shown in the figure below:



A scenario that uses a subject indicator

In this scenario the identity of topic "T" cannot be established directly since it represents a non-addressable subject (apple, the fruit). Instead, subject identification is done *indirectly* using a subject identifier, which resolves to a subject indicator, which in turn provides a human-interpretable *indication* of the identity of the subject. The subject indicator is only used by humans. For applications it is sufficient to compare subject identifiers to know when two topics represent the same subject.

NOTE: The XTM syntax for the example above is as follows:

```
<topic id="apple">
  <subjectIdentity>
    <subjectIndicatorRef

xlink:href="http://www.everything2.com/index.pl?node_
id=187149"/>
  </subjectIdentity>
  <!-- names and occurrences -->
</topic>
```

2.5 Published Subjects

2.5.1 Shortcomings of the above scenario

The scenario described above works well within a controlled environment, but it is too simple to be sustainable in a broader context. Any information resource can be considered a subject indicator simply by being referred to as such by an application, *whether or not that resource was intended by its publisher to be a subject indicator, and whether or not the publisher is aware of (or even cares about) its use as a subject indicator.* Hence subject indicators and subject identifiers defined in such a way are not trustworthy, and are likely to be either ambiguous, or unstable, or both.

Regarding ambiguity, how are we to know exactly what subject is indicated by the subject indicator in the example above? The Webster dictionary entry actually defines at least three other subjects in addition to apple (the fruit) and might conceivably be used by someone else to indicate a different subject, for example the apple tree *(Malus domestica)*. Because it was not the specific intent of the publisher of this resource to provide a subject indicator for a single subject, the degree of precision necessary to avoid ambiguity cannot be expected.

In addition, since the resource was not published with the express intent of being used as a subject indicator, one cannot expect any commitment to stability of the URI and the resource to which it resolves. (This becomes very clear when one realises that every URI in use at the Everything site changed when the system was upgraded to Everything2 in 2001. Anyone relying on those resources as subject indicators before that date would have been confronted with a major maintenance effort.)

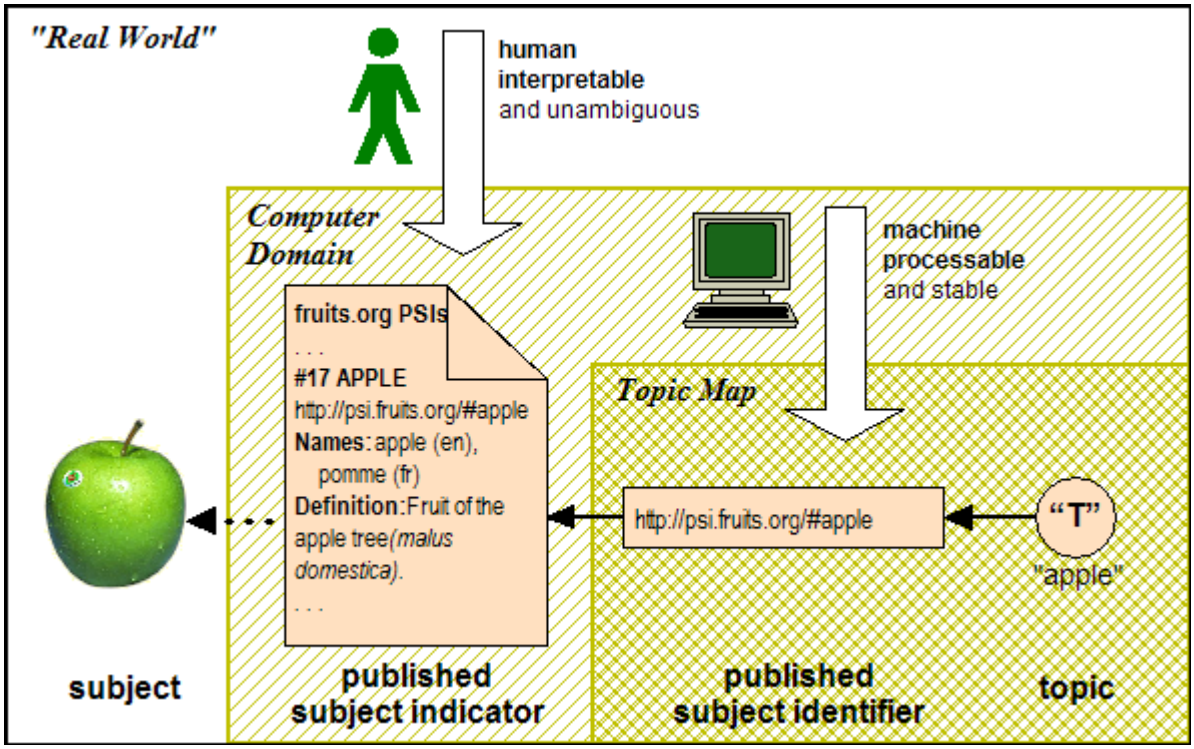
2.5.2 Publishers in the loop

If publishers are aware of the shortcomings described above and want to provide applications and users with non-ambiguous, stable, reliable, and authoritative subject indicators and identifiers, the situation is far better. The publishers can provide sets of subject indicators and subject identifiers published in a standard way, and declare their intention to maintain their stability, reliablility and availability on the network. At that point, applications and users will be provided with *published* subjects, *published* subject indicators (PSIs) and *published* subject identifiers (PSIDs).

The publication space within which such published subjects will be used is a network of interconnected applications and of users allowed to access those applications. It can be as wide and open as the Web, but it can be also a more or less closed network like an enterprise intranet or community portal. For that reason, "published" does not necessarily mean "public".

2.5.3 Example: A Published Subject for "Apple"

In the figure below, the subject identified for the computer by the (fictitious) URL "http://psi.fruits.org/#apple" is indicated to humans by a dedicated information resource, a *published subject indicator* created by Fruits.Org, in order to provide a non-ambiguous and stable definition. The Publisher (Fruits.Org) has declared this resource stable and intended to be used as a PSI. A human user can trust the URI resolution to provide him with a stable on-line resource as long as he has access to the network.



A revised scenario that uses a published subject indicator

NOTE: The XTM syntax for the example above is as follows:

```
<topic id="apple">
  <subjectIdentity>
    <subjectIndicatorRef
      xlink:href="http://psi.fruits.org/#apple"/>
  </subjectIdentity>
  <!-- names and occurrences -->
</topic>
```

2.6 The Adoption of PSIs

Anyone can publish PSIs, from the largest international organizations to communities of interest, enterprises and even individuals. There is no approval process and no registration authority. The adoption of PSIs can therefore be an open, bottom-up, and distributed process.

PSIs are usually published in sets that correspond to a particular controlled vocabulary or ontology (such as those published by OASIS based on the ISO 3166 and 639 country and language codes), or a group of subjects that are of special interest to a certain organization, application, or individual.

Whether a particular set of PSIs will actually be adopted by others will naturally depend on a number of factors, not least the publisher's visibility and the degree of trust it enjoys. The more authoritative a publisher is regarded as being, the greater will be the likelihood that its PSIs will enjoy widespread adoption. (Of course, the definition of "authoritativeness" will vary according to the circumstances: different criteria will apply if the goal is adoption within a relatively narrow community, such as a single company, than when the goal is adoption across the whole of the World Wide Web.)

The adoption of PSIs is therefore likely to be an evolutionary process in which stable, authoritative sets of PSIs gradually emerge and achieve widespread adoption. However, the mechanism is also sufficiently flexible and open that there is no need to await the emergence of "definitive" sets of PSIs. Any user that needs a PSI for a particular purpose should first consider adopting one that already exists, and then, if nothing suitable is found, create his or her own. If something more definitive becomes available later, it is a simple matter to map from the one to the other. Such mappings can be created once and leveraged repeatedly. Over time, as de facto standards emerge, they will become less and less necessary.

Thus there is no reason not to start using published subjects immediately. The sooner one starts, the greater the short- and long-term benefits in terms of interoperability within and between applications.

3. Requirements and Recommendations for PSIs

3.1 Requirements for PSIs

This section contains basic requirements for the publication of PSIs.

3.1.1 Requirement 1

A Published Subject Identifier must be a URI.

3.1.2 Requirement 2

A Published Subject Identifier must resolve to an human-interpretable Published Subject Indicator.

> NOTE: It has been widely discussed whether URNs should be used as PSIDs, or only URLs. Although general best practice will certainly use URLs, URNs are not completely ruled out as PSIDs, provided the publisher defines some easily accessible resolution mechanism (in order to conform to Requirement 2). URNs can be used by Topic Map authors as valid interoperable, machine-processable Subject Identifiers, but in the absence of easy resolution to a human-interpretable Subject Indicator, they will not be considered as conforming Published Subject Identifiers with respect to the present requirements.

3.1.3 Requirement 3

A Published Subject Indicator must explicitly state the unique URI that is to be used as its Published Subject Identifier.

> NOTE: PSIDs should be used exactly as published since processors cannot be expected to perform URI normalization.

3.2 Recommendations for PSIs

This section contains general recommendations for publishers of PSIs on the use of metadata. Deliverable 2 of the OASIS Published Subjects TC will provide more specific recommendations and examples of both human-readable and machine-readable metadata.

3.2.1 Recommendation 1

A Published Subject Indicator should provide human-readable metadata about itself.

3.2.2 Recommendation 2

A Published Subject Indicator may provide machine-processable metadata about itself.

> NOTE: Machine-processable metadata is recommended so that applications can help users discover and evaluate the suitability of PSIs. Human-readable as well as machine-processable metadata can be included in the Subject Indicator itself (e.g., as RDF metadata) or in a separate information resource referenced from the Subject Indicator (e.g., as XTM metadata).

3.2.3 Recommendation 3

Metadata defined in Recommendations 1 and 2 should be consistent, but need not necessarily be equivalent.

> NOTE: Consistency between human-readable and machine-processable metadata guarantees consistent "interpretation" by applications and humans. This can be achieved, for example, by human-readable metadata being a rendition of machine-processable metadata. This issue will be addressed by Deliverable 2.

3.2.4 Recommendation 4

A Published Subject Indicator should indicate that it is intended to be used as a PSI.

> NOTE: This Statement of Purpose should be clearly endorsed by the publisher (see below).

3.2.5 Recommendation 5

A Published Subject Indicator should identify its publisher.

> NOTE 1: Publisher is to be understood here according to the Dublin Core definition: "An entity responsible for making the resource available."
>
> NOTE 2: The Statement of Purpose and Publisher Identification are the principle warrants of trust, which is fundamental to the effective functioning of the PSI mechanism.

4. References

ISO 13250:2003

ISO/IEC 13250:2003 Topic Maps, International Organization for Standardization, http://www.y12.doe.gov/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf

RFC 2396

Uniform Resource Identifiers (URI): Generic Syntax, IETF, August 1998, http://www.ietf.org/rfc/rfc2396.txt

SAM

Garshol, Lars Marius and Graham Moore (eds):*The Standard Application Model for Topic Maps*, April 3 2003, http://www.isotopicmaps.org/sam/sam-model/

XTM 1.0

Pepper, Steve and Graham Moore (eds): *XML Topic Maps (XTM) 1.0 Specification*, TopicMaps.Org, March 2001, http://www.topicmaps.org/xtm/1.0/