Towards the Semantic Superhighway: A Manifesto for Published Subjects

Steve Pepper
Ontopia
Waldemar Thranes gt. 98
N-0175 Oslo, NORWAY
+47 23233080
pepper@ontopia.net

ABSTRACT

This paper describes the need for a simple mechanism for defining and assigning unique global identifiers for arbitrary subjects on the World Wide Web in order to solve the problem of information overload.

It presents the case for Published Subjects and published subject indicators (PSIs) being the best solution to this problem, and briefly characterizes the strengths and weaknesses of alternative approaches. It ends with a call to action.

It might look like a scientific paper, but it is not. It does not represent scholarly work that is being published for the first time, and it ought to be understandable by anyone into whose hands it is likely to fall. Nor is it a standards document (although parts of it may read like one) because the ideas and proposals it contains are so simple and obvious as to hardly seem worth standardizing. Rather, it is a call for action, aimed at absolutely anyone who aids and/or abets in the publication of information, or dissemination of knowledge, on the World Wide Web, especially those concerned with semantic interoperability.

(Yes, that does indeed mean you.)

Categories and Subject Descriptors

H.1.2 [Models and Principles]: User/Machine Systems – human information processing.
H.3.7 [Information Storage and Retrieval]: Digital Libraries – Standards.
H.3.7 [Information Storage and Retrieval]: Systems and Software – Distributed systems.

General Terms: Design, Human Factors, Management, Standardization, Theory.

Keywords: Identifiers, Metadata, Taxonomies, Thesauri, Ontologies, Semantic Integration, Semantic Interoperability, Semantic Web, Topic Maps.

1. INTRODUCTION

1.1 The Problem of Disconnected Information

A spectre is haunting the World Wide Web (and not just the Web): the spectre of infoglut.

None of us need telling that the world is drowning in information, or that one of the biggest problems faced by most organizations in today's "knowledge economy" is how to get the right information to the right person at the right time, in order to enable the correct and efficient execution of some task. The nature of this problem is not new (indexers and librarians have been helping us solve it for centuries), but the scale of it is. The advent, over the last few decades, of word processors, personal computers, and, most recently, the Internet has resulted in an almost unimaginable increase in the amount of information that surrounds us; and that same technology has simultaneously increased our dependency on information, which has becomes *the* core asset in many sectors of today's economy. In other words, the more we need the stuff, the harder it becomes to find what we are looking for.

Fortunately, the very technology that (1) enables us to create this glut of information, and (2) induces our ever-increasing dependence upon it, is also responsible for bringing about a gradual awareness of the underlying cause of the problem, because the more we are able to connect our computers, the more obvious it becomes how horribly *disconnected* our information is.

And disconnected information costs money: spread across multiple systems, it becomes hard to find; duplicated across different departments, it breeds redundancy and becomes unreliable; isolated in information silos that don't communicate with one another, it fails to achieve its full value.

(Does any of this sound familiar?)

Disconnected information means disconnected knowledge: knowledge that cannot be shared, insights that cannot be drawn, and decisions that cannot be taken because the right information was *not* available to the right people at the right time. Somehow we need to connect our information – and knowledge – such that it attains its full value. The only question is *how*.

There are those that put their faith in the miracle cures offered by search engine vendors: they may have some of their pain alleviated for a while, but they usually end up being disappointed; perhaps they switch to another vendor, in which case they probably end up being disappointed once again.

Others start to feel their way towards a different solution, one in which "metadata", "taxonomies", "thesauri", and perhaps even "ontologies" play a role. These are all knowledge organization techniques that allow us to organize information *by subject*. The underlying intuition is thus that the key to solving the "findability problem" is *subject-based organization* of information.

1.2 The Centrality of Subjects

And why not, because isn't that how humans work? Our starting point is invariably some *topic* or *subject of discourse* that interests us and about which we need more information. The subject in

question could be anything at all, from an abstract concept to a physical object; it could be something that exists or something that does not; it could be animal, vegetable or mineral – or none of the above; but it is always a subject – by definition. Even when our starting point is an author or a date (as opposed to the kind of "subject" that might be found in a *keywords* or *dc:subject* field), the person or date in question can also be construed as a subject about which we want more information.

Those indexers and librarians who have been helping us find information (in books and libraries, respectively) for the longest time have known about the centrality of subjects for centuries: that is why books are usually organized by subject on library shelves and why indexes consist primarily of lists of subjects (or topics), along with the locators (page numbers) of information *about* those subjects. Even the lowly folder mechanism on file systems is a recognition of the fact that humans like to – need to – order their information *by subject*.

If subject-based organization of information is the key to making it easy for users to find information, then it must also be the key to making it possible to *connect* information – at least when that information is to be consumed by humans.

The problem of distributed information is just that – that it is distributed, spread all over the shop, in different physical locations and in different information systems (not to mention in different data formats, under the control of different applications, and requiring different logons and passwords). The physical connectivity brought about through networking (and, in particular, the Internet and the World Wide Web) provides the *potential* to connect all our information and also helps us understand the enormous benefits that connected information would bring, but it once again raises the question: *how*, or, more precisely, *by what criteria* should two pieces of information be connected?

1.3 The Importance of Collocation

It seems altogether too obvious that the only possible answer to this question, at least in a scenario that involves humans, is that two pieces of information should be connected *if they are about the same subject*.[1] Once again, the reason is because humans think in terms of subjects and they seek out information on the basis of it being about a certain subject. The most useful information retrieval application is one which allows humans to find "everything there is to know" about a particular subject from a single point of access, irrespective of the physical and logical locations of the individual pieces of information. This goal of providing users with a single point of access to all relevant information is what librarians call the *collocating* (or *collocation*) *objective* [15].

In libraries, the most obvious form of collocation is physical; books are relatively small physical objects that are usually organized as discrete entities (rather than, say, on a chapter by chapter basis), and can thus be physically collocated on the same shelf. A back-of-book index, on the other hand, employs a more "conceptual" form of collocation; rather than collect together all the separate pieces of information within a book that are about the same subject, an index uses locators (page numbers) to point to the information in question.

These two kinds of collocation – physical and conceptual – each have advantages and disadvantages. From the point of view of the user looking for information, seeing it collected together physically provides for a better overview and more ease of comparison. From the point of view of someone managing the information, physical collocation may not be an option, for a variety of technical, political, and economic reasons.

Fortunately, in the era of digital information, we can get the best of both worlds: information can (in theory) be organized, indexed, mapped, and managed in a decentralized, distributed fashion, and then brought together – physically collocated (again, in theory) – at runtime when being presented to end users.

However, this only works if we have a reliable, machine-processable way of knowing when two (or more) pieces of information are "about" the same thing, or, more generally, if humans and computers can somehow know when they are talking about the *same subject*.

1.4 The Need for Identifiers

The primary mechanism for denoting subjects in human discourse is names. Names are essentially conventionally agreed upon labels for subjects. Despite being fraught with ambiguity (because of synonymy, homonymy, polysemy, and the like), names generally suffice in human communication as a means of establishing when people are talking about the same subject. This is because of our ability as humans to utilize *context* as a disambiguator and to engage in *negotiation based dialogue*.

Computers, of course, are not as smart as us. Their ability to utilize context and engage in negotiation is practically non-existent: given the string "paris", a machine cannot distinguish between the capital of France, the hero of Troy, and the character in *Romeo and Juliet* (to name just a few of the many possibilities). That is why data processing systems traditionally rely on *identifiers* rather than names. An identifier can be thought of as a name that is unique within some carefully defined domain of discourse and hence more suitable for computational purposes.[2]

Identifiers are all around us, from file names, to primary keys in databases, to language codes (like "nor" for Norwegian) in HTML documents and elsewhere. As long as you stay within the carefully delineated domain of discourse (or "namespace") in which they are defined, such identifiers are unique and therefore absolutely unambiguous. They can thus be used as the basis for all kinds of automated processing, including the collation of information about the subjects that they identify (such as the language Norwegian in the example above).

---

[1] They might also be connected for other reasons, e.g. because they are by the same author, or published by the same organization, but this should be *in addition to* rather than instead of connection based on shared "aboutness".

[2] The term *name* is sometimes used indiscriminately to denote both labels used by humans and labels used by machines. However, the distinction between labels used in fuzzy, negotiated contexts, and labels that are purposely defined to be unique within some carefully delineated domain of discourse, is central to the issue at hand and this warrants the use of two separate terms, *name* and *identifier*.

The great benefit of common identifiers, of course, is that they allow us to use different names and still know when we are talking about the same thing.

This is all well and good, but it only works within the narrow confines of a single application or family of applications; if the goal is to solve the problem of information overload described above on a larger scale (which could easily be corporate-, industry-, or even world-wide), that is not enough. As soon as we move outside the bounds of the local context, these kinds of identifiers break down and we no longer know whether the identifier "nor" stands for the language Norwegian, the country Norway, or the Icelandic airport Nordfjörður.

What we need therefore is a mechanism for defining and applying unique identifiers on a global scale. Once we have that, computers will have the potential to automatically collate information and knowledge by subject across any conceivable physical or logical boundary. We will have begun the construction of a "spine of concepts" – or *semantic superhighway* – to complement the physical superhighway of the Internet and the World Wide Web, and we will be well on the way to getting a handle on infoglut.

2. REQUIREMENTS ON A GLOBAL IDENTIFIER MECHANISM

A number of different proposals have been put forward over the years for global identifier mechanisms, but before looking at those it is important to list some of the requirements that an acceptable solution should meet. I propose the following:

1. The mechanism as a whole should be:
    a. democratic
    b. scaleable
    c. easy to adopt
2. The identifiers themselves should be:
    a. easy for humans to use
    b. easy for computers to use

We will look at each of these requirements in turn.

2.1 Democratic

The mechanism by which identifiers are defined should be open and democratic. The reason for this is not simply ideological (although that would be reason enough on its own); the reason is rather that we cannot expect identifiers that have been imposed from above to achieve the degree of widespread adoption required to solve the findability problem. The days of monolithic, "one size fits all" solutions are gone; open, collaborative systems are the order of the day and users will not accept identifiers that have been imposed from above (although they will probably accept identifiers that are *proposed* from above by authorities that are regarded as being sufficiently objective).

2.2 Scaleable

The mechanism must be able to scale to billions of identifiers.

In an ideal world there would be a unique identifier for every conceivable subject under the sun. In reality this is never going to be the case: partly because humans invent new concepts faster than it would be possible to mint identifiers for them; and partly because the effort (however small) of creating a new identifier will sometimes be greater than the value it brings (if only because for some obscure and/or transitory concepts the amount of relevant information available to be connected would be too small).

Nevertheless, identifiers are needed for a vast number of subjects. Wikipedia, which at the time of writing (April 2006) contains slightly over a million articles in its English version, merely scratches the surface: for every article on Wikipedia, there tens, hundreds, or even thousands of subjects for which there are no articles (and may never be) but which are still worthy of being assigned global identifiers. These range from people to places, from species to substances, and from objects to processes, and they include (for example) the number 28 bus that runs past my house in Oslo. For all of these, or at least all those for which there exists more than a couple of pieces of information spread around the globe, having an identifier would aid findability.

2.3 Easy to adopt

In order to succeed, a mechanism for assigning and using global identifiers must be easy to adopt: the threshold must be as low as possible, and the advantages both obvious and immediate. Any solution that requires users to change their habits in a fundamental way, or that requires a new tool set, is unlikely to succeed. What we need, therefore, is a mechanism that is a natural extension of existing mechanisms and that offers a smooth migration path based on existing practices.

2.4 Easy for humans

Humans use names rather than identifiers in their daily discourse with one another, but they are also very much "in the loop" when it comes to machine-based information processing. While the purpose of identifiers is to make it possible for computers to know when two pieces of information are about the same subject, it is humans who are mostly responsible for defining and assigning those identifiers.

This means that identifiers must be easy for humans to 1) locate, 2) mint, 3) interpret, and 4) apply, i.e.

1) given a subject, it should be easy to find an identifier (if one already exists);
2) given a subject, it should be easy to mint an identifier (if nothing suitable already exists);
3) given an identifier, it should be easy to find out what subject it identifies; and
4) given an identifier, it should be easy to insert it into a document, index, database or other source of information.

2.5 Easy for computers

While identifiers are defined and assigned by humans, they are mostly used by computers in order to ascertain when two subjects are the same, or, more precisely, when two pieces of information (or two assertions) are about the same subject. The simpler the processing involved, the more scaleable and robust the solution will be.

The simplest way to use identifiers to determine whether two things are equal or equivalent, is to simply compare the identifiers as strings: if two subjects, named A and B, have identifiers A' and B', which are lexically identical, then A and B can be assumed to be different names for the same subject.

Any processing that requires more than simple string comparison (whether it be string normalization, network retrieval of a resource, or some form of computation), will invariably lead to more heavyweight and less reliable applications.

3. PUBLISHED SUBJECTS

This section introduces the Published Subjects mechanism and evaluates it on the basis of the foregoing requirements.

3.1 Overview of Published Subjects

The concept of Published Subjects originated within the Topic Maps community in late 1999 as the ISO standard [8] was nearing finalization.[3] It was refined during the development of the XML Topic Maps (XTM) specification in 2000-2001 [12], and in work performed by an OASIS Technical Committee in 2002-2003 [11]. The basic ideas are extremely simple and can be summarized as follows:

1) A *published subject identifier* is a IRI[4] that by definition is deemed to identify a single subject and that was expressly created in order to serve as the identifier of that subject.

2) A published subject identifier resolves to a human-interpretable document, known as a *published subject indicator*, that is intended to convey a compelling and unambiguous "indication" of the identity of its subject and explicitly declares itself to be a subject indicator.

An attempt was made in [11] to reserve the acronym PSI for "published subject *indicator*", but in practice it is also used for "published subject *identifier*", and, even more appropriately, for the composite of an identifier and its corresponding indicator.

Figure 1 provides a schematic overview of the PSI model. A few explanatory comments are in order:

- The subject identified by a PSI can be anything whatsoever, "regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever."[5]

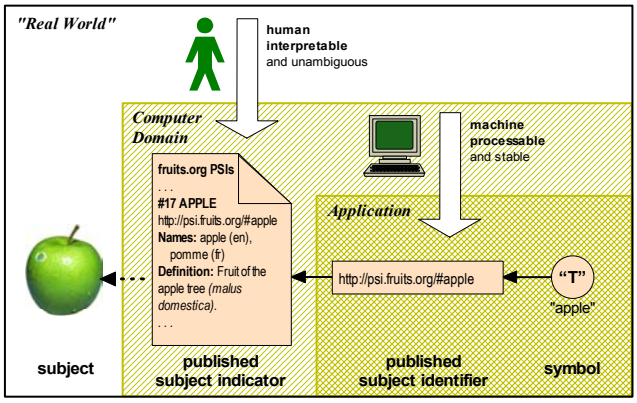


Figure 1: The roles of published subject identifier and published subject indicator in identifying a subject

- Anyone may define and publish a PSI: there is no requirement that authorization be sought or given.
- Although a PSI can be any kind of IRI, in practice the most common form of PSI is an http IRI (or URL). This is because of the requirement that the IRI be resolvable. The resolution mechanism for URLs is very simple and well-understood, and it is supported by every web browser and many other applications as well. Most other URI schemes and URN namespaces do not have this advantage.
- In order to be regarded as a *published* subject identifier, a subject identifier (and its corresponding subject indicator) must be made available to members of some wider community for use outside a single application.[6]
- The duality of the PSI, consisting of an indicator and an identifier, reflects that of the human/computer dichotomy:
- *Indicators* exist for the benefit of humans whose job it is to assign identifers to information resources; by examining and interpreting the contents of the subject indicator, a human can gain a notion of the identity of the subject that is sufficient for deciding whether or not to use that PSI.
- *Identifiers* exist for the benefit of computers that need to be able to ascertain whether two pieces of information (or assertions) are about the same subject: if the identifiers are identical, this can be assumed to be the case; if they are not, no such assumption can be made.

3.2 Advantages of Published Subjects

The basic principles of PSIs are thus extremely simple: A PSI is simply a URL that is defined (and published) for the express purpose of serving as an identifier and that resolves to a document that in some way describes or "indicates" the subject that it identifies. The first major advantage of PSIs is that the concept is easy to explain.

In addition, PSIs extend current practice rather than promote an entirely new paradigm: URLs are already well understood as identifiers for *documents*, and they have been used as identifiers for arbitrary subjects in a number of communities (including the RDF/OWL and Topic Maps communities) for several years. This means that there already exists a very large number of identifiers that are URLs. All that needs to be done in order to turn these into PSIs is to create subject indicators to which they resolve. In many cases this can be done automatically, using existing information resources, such as definitions.

Furthermore, existing systems of (non-global) identifiers can easily (often authomatically) be turned into PSIs by prefixing an existing alpha or numeric code with a namespace URI based on an internet domain and an (optional) subject domain component. Thus PSIs were created by the OASIS Geolang TC for all of the ISO 639 language codes by the simple expedient of prepending the namespace URI `http://psi.oasis-open.org/iso/639/#` to the 3-letter alpha codes (such as `nor` for Norwegian) [10].

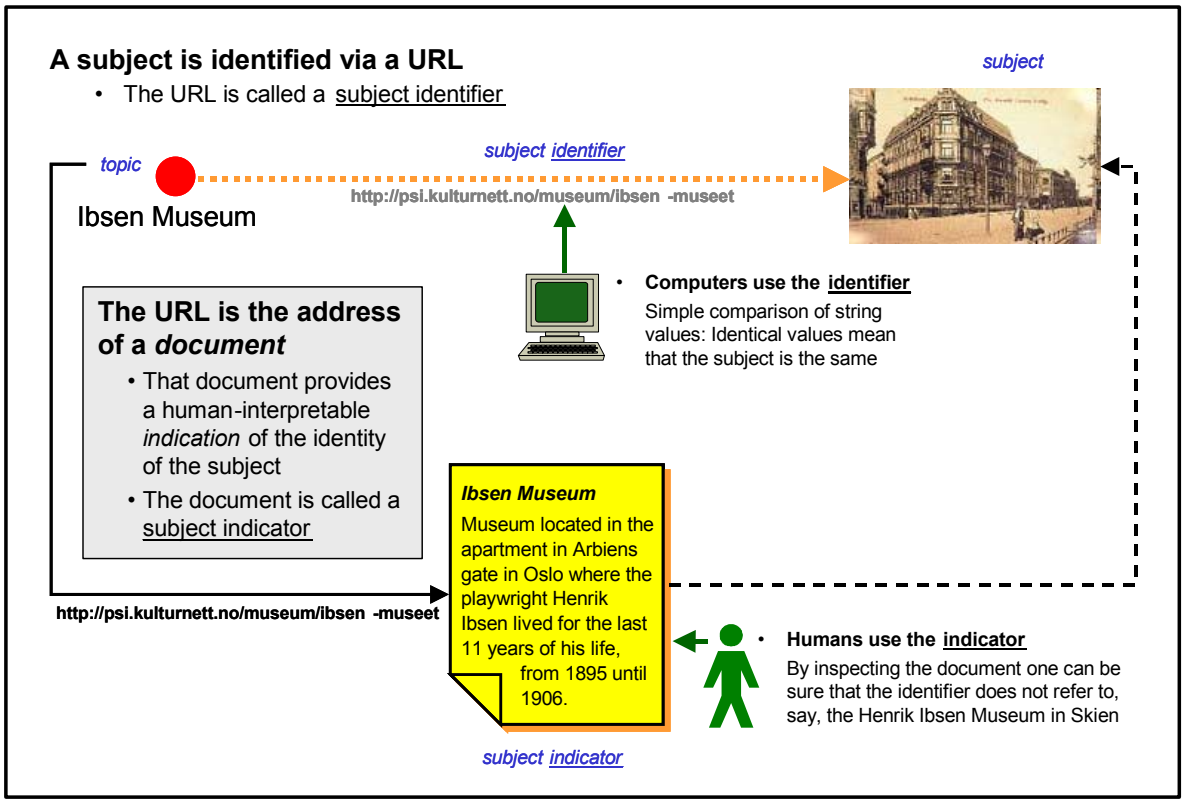
---

[3] The term originally used was "public subject", but this was changed in 2001 to "published subject" in order to avoid giving the impression that the concept could not be applied within a closed community.

[4] The original work done on Published Subjects talked in terms of URIs, but there is, of course, absolutely no reason why IRIs should not be used instead.

[5] This is the definition of "subject" in [8]. Intuitively it is the same as that of "resource" as used on the Web.

[6] Naturally this does not prevent URLs from being used as identifiers within a single application, with or without the creation of corresponding subject indicators, and without being made available to others; these can be regarded as subject identifiers, but not as *published* subject identifiers.



The Published Subjects paradigm meets the requirement of openness and democracy by allowing PSIs to be defined by anyone at all, from the loftiest international organizations, such as governments, NGOs (e.g., the United Nations), and multinational corporations, to individuals and micro-communities. Provided compelling ways of utilizing PSIs become available, this openness will also make it possible to mobilize people to create and use identifiers on the kind of scale necessary to solve the infoglut problem. Recent developments in social bookmarking and community-based tagging and content creation (ref. del.icio.us [3], Flickr [4], and Wikipedia [17]) indicate that this ought to be a real possibility.

PSIs also meet the requirement of being easy to use for humans. They are easy to mint – provided the publisher has access to an internet domain and the ability to create a simple web page – and they are also easy to interpret: simply paste the identifier into a web browser and retrieve the indicator, which should be readily interpretable. (If it is not, complain to the publisher and consider using a different PSI.) Most PSIs will also be easy to insert into documents, indexes, databases and other source of information; doing so will at least be no more difficult than inserting a URL.

Finally, comparing PSIs when collating information is a simple matter of performing a string comparison: it is not necessary for computers to retrieve the subject indicator since that is intended only for humans. No additional network traffic is involved, so there is no performance penalty.

3.3 Some Objections

Certain objections have been raised to the idea of Published Subjects as described above. The first, most important, and in our opinion only really telling objection, is that there is currently no mechanism for aiding in the discovery of PSIs: it is all very well saying that minters of PSIs should publish them, but how are other users supposed to locate them when needed?

Fortunately this is an objection that can be easily dealt with: for example, by agreeing on some distributed system of PSI registries or by establishing conventions for the contents of published subject indicators that make them discoverable using web search techniques. That no such mechanism as yet exists is simply due to the fact that the whole idea of PSIs is relatively new and people are still experimenting with alternative approaches to both PSI metadata, PSI management, and PSI discovery.

A second objection, which at first sight appears more substantial, is that the "bottom up" nature of the PSI minting mechanism, whereby anyone can define PSIs, is bound to lead to the existence of multiple, "competing" PSIs for the same subject and thereby confound the achievement of collocation.

It is true that multiple PSIs will inevitably be defined for the same subject, but this need not be a problem. If the process of defining PSIs is as random as, say, genetic mutation (which it almost is), we can look forward to an evolutionary process of natural selection through which certain PSIs (or sets of PSIs) will emerge as *de facto* standards on the basis of criteria such as stability, trust, authoritativeness, and first-mover advantage. Users have a vested interest in ensuring that this happens, otherwise the whole purpose of using PSIs will be subverted.

This process will, of course, be somewhat messy (like the Web itself) but over time it will lead to stable, authoritative, and trusted sets of PSIs. In the meantime, as "market consolidation" of PSIs takes place, identity services can be offered for those needing to map one set of identifiers to another.

For example, while the rest of the world uses ISO 3166 codes to identify countries, the CIA (ever a law unto itself) has its own

identifiers, referring to Germany as GM instead of DE, the Republic of Korea as KS instead of KR, and so on. To achieve collocation across information that uses both sets of identifiers, all that is required is a simple mapping service; actually defining the mappings only needs to be done once.

A third objection might be that it is wrong in principle to use URLs for two distinct purposes: identifying information resources (directly) and identifying arbitrary subjects (indirectly, via documents). Certainly there is a potential problem for systems that ignore the distinction between these two uses, as pointed out in [13]. But it remains to be shown that this is sufficient reason to completely abandon the use of URLs as identifiers, especially when the PSI paradigm offers a perfectly good mechanism (namely, subject indicators) for ascertaining when a URL has been minted in order to be used as an identifier.

In conclusion, none of these objections outweigh the advantages of PSIs described above.

4. ALTERNATIVE PROPOSALS

This section looks briefly at some alternative proposals that explicitly or implicitly address the need for global identifiers for arbitrary subjects, and characterizes them on the basis of the requirements defined in section 2. Readers who do not need to be convinced that PSIs are the best option can skip this section and go straight to the Call for Action.

4.1 Uniform Resource Names (URNs)

It was generally assumed in the early days of the Web [7] that identifiers would fall into one of two (or possibly more) classes: locators and names, represented by URLs and URNs respectively.

While URLs have been a phenomenal success, URNs, which had the potential to become the *de facto* standard for assigning global identifiers to arbitrary subjects, are hardly used at all. As of this writing, twelve years after the URN mechanism was approved, only 25 formal URN namespace IDs have been registered, most of them related to the Internet and many of them unused.

There seem to be two reasons for this:

1) URNs do not have a single, well-defined resolution mechanism. Each URN scheme defines its own and most of these are not supported by web browsers. Given a URN, it is therefore non-trivial to discover what subject it identifies. (An example, taken from UBL, is given later.)
2) URNs are based on a "top-down" approach whereby URN schemes must be registered with an authority (IANA) in order to become official. Whether or not IANA's approval policy has been restrictive, this represents a level of bureaucracy that could easily be intimidating.

PSIs suffer from neither of these problems, since they have a well-defined resolution mechanism and can be minted by anyone.

4.2 The Hash/Slash Distinction

The advent of the Semantic Web led to an enormous increase in the need for identifiers for subjects that are not information resources, including properties, classes, and instances of classes other than InformationResource. Instead of using URNs, the Semantic Web community used URLs for this purpose and this led to the Great Debate, referred to as the "httpRange-14 issue" but also known as the "Identity Crisis of the Web", which has gone on for years and continues to this day [13], [1].

At the heart of this debate is the question: How can we know what a URL identifies (or "means")? Does it identify the information resource that it dereferences to, or does it identify the subject that is denoted, described, or (in PSI parlance) indicated by that resource?

One position that was put forward was that the answer depends on the *form* of the URL. A URL that contains a hash character is deemed to be fundamentally different from one that does not: a URL with a hash identifies an arbitrary subject, whereas one without always identifies an information resource.[7]

The philosophical basis for this distinction was never entirely clear and the position has not been accepted by the Web community as a whole, for a number of reasons. Firstly, it ignores the fact that a hash URL in fact identifies a fragment of a document, which itself is an information resource. Secondly, the inclusion of a fragment identifier in a URL has a number of negative consequences in terms of processing (not least because the fragment identifier portion of the URL is not seen by the Web server). And thirdly, there already exists a considerable legacy of slash URIs that identify non-information resources: in other words, the horse has already bolted.

4.3 The TAG's httpRange-14 Resolution

A "final solution" to the httpRange-14 issue was proposed (and accepted) by the Technical Architecture Group (TAG) of the W3C in June 2005 [16]. It reads as follows:

> The TAG provides advice to the community that they may mint "http" URIs for any resource provided that they follow this simple rule for the sake of removing ambiguity:
>
> – If an "http" resource responds to a GET request with a 2xx response, then the resource identified by that URI is an information resource;
>
> – If an "http" resource responds to a GET request with a 303 (See Other) response, then the resource identified by that URI could be any resource;
>
> – If an "http" resource responds to a GET request with a 4xx (error) response, then the nature of the resource is unknown.

While this seems to permit the use of either hash or slash URIs for any kind of subject, it imposes a burden on minters and users that they are unlikely to accept. It also leaves unanswered whether the same URI can identify both a location within an HTML document and (say) a concept in an ontology. For this and other reasons the resolution is still causing controversy in the Semantic Web community. [1]

4.4 The tdb URN namespace

Larry Masinter's proposal [6] for a URN namespace called "tdb" (thing described by) is intended to be "useful as a way of creating URNs that refer to physical objects or even abstractions that are not themselves networked resources." The idea is to use the URL of a web page describing an arbitrary subject as the principal component of a URN identifying that subject. A tbd URN has the following form:

```
urn:tdb:<date>:<encoded-URI>
```

[7] As Tim Berners-Lee once put it: "You jump into a whole new world when you add the "#"." [14]

For example, `urn:tdb:2001:http://www.ietf.org` "can be used to designate the Internet Engineering Task Force organization, at least as it was described by or referenced by its home page at the first instant of 2001."

The idea of defining a URN namespace and then opening it up for anyone to use gets around the problem of the lack of openness of the URN mechanism in general (provided, of course, that IANA approves such an "irresponsible" idea). However, it only provides a partial solution to the problem caused by the lack of a general and widely supported resolution mechanism for URNs, since in order to be dereferenced, the URN must be unpacked and the URL both extracted and decoded. There are no widely available tools that do this today.

The idea of using a web page that "describes" a subject as (part of) an identifier for that subject is similar to that of using a published subject indicator that "indicates" the subject, but there is one very important difference: whereas a PSI must have been expressly created in order to serve as an identifier and must declare itself as such, a tbd URN allows any arbitrary web page to be used for this purpose. A description that was *not* created expressly in order to provide a "compelling and unambiguous indication" of the identity of a subject will, in general, be:

- less precise (because real precision is usually only attained as the result of a conscious effort);
- less objective (because the page is likely to include assertions about the subject – some of which are likely to be subjective – over and above those necessary to convey its identity); and
- less stable (because the publisher of the web page has made no commitment to providing a stable "indicator").

One can also question the wisdom of including the date in the URN. While doing so is an understandable response to the lack of stability of arbitrary web pages, it complicates matters (especially resolution) and does nothing to encourage stability. PSIs, on the other hand, promote stability by requiring publishers of subject indicators to make a commitment, and by actively espousing a policy of survival of the fittest (i.e., most stable and trusted).

4.5 thing-described-by.org

The tdb URN proposal appears not to have been pursued since it lapsed as an IETF Internet Draft in October 2004 and may have been superseded by *thing-described-by.org*. This domain (and its abbreviated sister, *t-b-d.org*) hosts a 303-redirect service that provides "a convenient mechanism for minting dereferenceable http URIs for things that are not necessarily Web resources." [2]

t-d-b.org essentially provides a way of "annotating" a URL to flag that it is intended to be used as an identifier for an arbitrary subject, without modifying the URL itself and without having to resort to a new URI or URN scheme that is unsupported by browsers. The idea is to construct a ("compound") URL by using the URL of a web page that describes a subject of interest as a *query parameter*. For example, the (compound) URL

```
http://t-d-b.org?http://dbooth.org/2005/dbooth/
```

essentially states (in PSI terminology) that the information found at `http://dbooth.org/2005/dbooth/` is intended to function as a "subject indicator" and that the (compound) URL itself can be used as an identifier for whatever the subject indicator describes.

Although some of the problems associated with the tdb URN proposal are avoided, there are still issues with this solution:

1) It depends on a service that is owned and operated by a single individual and thus has no guarantee of stability. Even if ownership and maintenance were to be transferred to another, more authoritative organization, the dependency on a third party would remain and would likely discourage many users.
2) All identifiers are 17 characters longer than they need to be.[8]

Once again, Published Subjects do not suffer from either of these drawbacks.

4.6 Web Proper Names

Web Proper Names is an initiative described in [5] whose goal is to provide "a distributed approach to creating and sharing Web names for things." It is unique in proposing the creation of a special URI scheme (called "wpn") for this purpose, and even more unusual in that it is based on the use of search technology. The idea is to construct a URI consisting primarily of a search expression that returns a set of documents that describe (or depict) a particular subject.

A WPN is actually a ten-tuple that consists, in addition to the namespace identifier (wpn), of the following components: owner, short name, engine, date, terms, language, result sequence size, checked sequence size, and percent correct. The last three of these specify, respectively, the number of hits produced by the search; how many of these were checked by the minter of the WPN; and the percentage of the latter that were actually about the intended referent (a measure of precision).

[5] provides the following example:

```
wpn://www.ltg.ed.ac.uk/~ht/WPN/EiffelTower?
terms=eiffel+tower+paris+-hotel+-webcam&
ln=en&se=www.google.com&dt=2004-05-21&
rs=17&cs=5&pc=84
```

This 140 character string, it is suggested, could serve as a globally unique identifier for the Eiffel Tower.

An Expanded Web Proper Name (EWPN) is a WPN that is stored as a web page with the following additional information: correct checked sequence (a list of the URIs in the result set that really were "about" the intended referent), incorrect checked sequence (a list of URIs that were *not* about the intended referent), and "further optional data about the referent that could be useful."

An EWPN may be stored anywhere. However, [5] encourages people to store them "so they are addressed by an http URI formed by adding the shortname to their owner identification." Thus it is recommended that the EWPN for the Eiffel Tower be stored at the following location (this URL differs slightly from that given in [5], which appears to be inconsistent with the WPN example and the explanatory text):

```
http://www.ltg.ed.ac.uk/~ht/WPN/EiffelTower
```

From the Published Subjects perspective, such a document could easily be regarded as a published subject indicator. If, in addition, its URL were to be used as an identifier, this proposal would be almost completely compatible with Published Subjects. However, that is not what is being proposed. The URL is not intended to be

[8] We will continue to maintain that these 17 characters are unnecessary until such time as a convincing argument has been adduced explaining why it should be possible to distinguish direct identifiers from indirect identifiers merely by inspecting the identifier itself. See below for more on this.

used as an identifier; only the 140 character wpn URI is to be so used.

This leads to our first objection to the proposal: the identifiers it results in are extremely unwieldy; they consume a lot of space and they are hardly suitable for hand-typing. On the other hand, there is no denying that WPNs can be minted very quickly and yield high precision – at least for a topic like the Eiffel Tower. A Google search on April 17 2006, using the same terms as the example above, yielded approximately 1,76 million hits; a cursory check of the first 30 indicated 100% precision (in the sense that every page talked about or mentioned the Eiffel Tower we all know and love, rather than some other Eiffel Tower).

This exercise took about three minutes, so there is no denying that it can be done efficiently (at least for some subjects). One wonders, though, whether the process of minting a WPN is not a little *too* simple. Three minutes is rather less than it would take to create and post a single subject indicator, and less than it would take (at least today) to locate a pre-existing PSI. We are all in favour of convenience, but too much of it can lead to problems: given the ease of creating my own, why should I bother to look for ~ht's WPN? Here is mine:

```
wpn://www.ontopia.net/~pepper/WPN/EiffelTower?
terms=eiffel+tower+paris+-hotel+-webcam&
ln=en&se=www.google.com&dt=2006-04-17&
rs=20&cs=10&pc=100
```

A simple string comparison with ~ht's WPN will not result in a match, so achieving collocation between ~ht and ~pepper is not going to be trivial.

This, however, is where the search terms come in: the reasoning behind the WPN format is that the list of URLs resulting from a search will allow machines to map automatically (with degrees of certainty) between identifiers. Beyond this, basing identifiers on search parameters does not seem to offer any benefits compared to a simple definition or description of the type that might be found in a one volume encyclopedia – that is, the contents of a typical subject indicator.

On the other hand, it does mean that WPNs depend on proprietary black boxes (search engines) over which users have no control. While it is hard to see what practical consequences this might have, it does feel distinctly uncomfortable.

But perhaps the most serious objection to the WPN proposal is its focus, as the name suggests, on proper names. Although [5] states that "Web Proper Names do not restrict referents to only those things that have proper names", one wonders just how easy it would be to construct a WPN for the more general concept of, say "tower". In another cursory attempt with Google, the search terms *tower+-records+-hobbies+-insurance+-comic* were required just to get the first five hits to be somewhat relevant, after which the results went completely haywire. On the (probably optimistic) assumption that a reasonably precise set of hits could be generated using ten search terms, it is hard to see how this would be preferable to a simple definition culled from, say, WordNet [18]

> **tower** (a structure taller than its diameter; can stand alone or be attached to a larger building)

or from Wikipedia [17]:

> A tower is a tall man-made structure, always taller than it is wide. Towers are often built as landmarks to be impressive or beautiful, however the main concept of towers is to save surface area. Skyscrapers are often not classified as towers, although most have the same design and structure of towers.

Such a definition or description, located at, say

```
http://psi.ltg.ed.ac.uk/~ht/Tower     /1/
```

or even

```
wpn://ltg.ed.ac.uk/~ht/Tower          /2/
```

would surely be infinitely more useful?

This brings us to our final objection, which is that WPNs require the definition of a new URI scheme which is unsupported by current tools. Getting such a scheme approved is likely to take a long time; getting it implemented in browsers and other tools will take even longer. In the meantime, Rome burns. With PSIs we can start extinguishing the fire today.

The rationale for introducing a new URI scheme, as in /2/, rather than simply using the existing, widely supported http URI scheme, as in /1/, is the conviction that "it is necessary for [a WPN's] primary role as a *name* that it be intrinsically (i.e. notationally) distinguishable [from] normal http: URIs." Why this should be the case is nowhere stated. However it is a conviction that seems to be widely held, and that also lies behind the design decisions of the several of the other proposals discussed here.

However, experience to date in both the RDF/OWL and the Topic Maps communities suggests that it is perfectly possible to use URLs as identifiers even when they are not *notationally* distinguishable from URLs used as addresses. Of course, if they are to be applied correctly it is important to know (1) that they are intended to be used as identifiers, and (2) *what* they are intended to identify, but this is precisely what subject indicators are meant to convey.

The real problem, perhaps, is that certain knowledge organization models do not cater for the fact that the same string may be used, in different contexts, as either a locator or an identifier, but this is a problem with those models, not with the notion of using one string for multiple purposes. After all, telephones and telefaxes have been able to share the same addressing mechanism, even though their purposes are quite different. While this may have been the cause of some irritation once in a while, it is unlikely that the telefax would ever have taken off if fax numbers had been required to be "notationally distinguishable" from phone numbers, since that would have required a major change to the existing telephone infrastructure.

5. CALL TO ACTION

The preceding sections have described the need for unique global identifiers for arbitrary subjects in order to address the problem of infoglut; we have presented a simple mechanism for achieving this; and we have reviewed the most important alternative proposals.

However, the point is to change the world, not interpret it, so this section suggests how those that support the PSI proposal might proceed in order to get it adopted on a broad scale.

There seem to be three main tasks:

1) plugging the most important gap in the proposal by agreeing upon a discovery mechanism for PSIs;
2) encouraging the creation of PSIs;
3) and getting people to use PSIs.

A discovery mechanism could be a PSI registry (or registries), conventions for the content of subject indicators that would make them searchable, or something else. Discussions on the best approach could take place in a number of fora, including OASIS and the Semantic Web and Topic Maps mailing lists. Implementing a prototype would be a very suitable subject for a Masters thesis. Any initiatives along these lines will receive the moral support (at least) of the current author and practical support from many of his friends and relations.

But to get the ball rolling, we need a critical mass of PSIs. This can be achieved by winning the argument for Published Subjects with communities such as the following:

- Creators of topic maps, RDF schemas, and OWL ontologies, who all use URLs as identifiers anyway: These people should be persuaded to create human-interpretable subject indicators to which their identifiers resolve.
- Creators of XML schemas that define concepts based on arbitrary subjects, for example UBL, which among much else defines the concept of "Floor" *and* gives it an identifier (albeit a URN).[9]
- Creators of subject heading lists, library classification schemes, and thesauri, many of whom are already intuitively defining identifiers in the form of URLs and who simply need to ensure that these resolve to the definitions they are also creating.
- Creators of metadata sets who define unique codes for use within specific domains (e.g., the ISO country and language codes mentioned above). These should be persuaded to define a http URI namespace for their domains that can be prefixed to their codes, and ensure that the resulting URL resolves to the appropriate descriptions.
- People with time on their hands, access to a suitable internet domain, and the ability to take a set of codes and turn it into a set of Published Subjects, either on behalf of the publisher of the original codes, or on their own behalf. Follow the example set by the OASIS GeoLang TC [9, 10].

A few million PSIs and a good discovery mechanism should be enough for a rudimentary semantic superhighway.

Finally, we need to encourage the use of PSIs and devise applications that demonstrate their power. There is something of a chicken and egg problem here, but once we have a discovery mechanism, it ought to be possible to dream up compelling Web 2.0 applications that blow people's minds, ensure the adoption of Published Subjects, and mark the beginning of the end of infoglut.

[9] `urn:oasis:names:specification:ubl:schema:xsd:CommonBasicComponents-1.0:Floor`, a URN that requires the abilities of a Sherlock Holmes to track down and dereference. It can be found in http://docs.oasis-open.org/ubl/cd-UBL-1.0/xsd/common/UBL-CommonAggregateComponents-1.0.xsd, along with the following definition "Identification by name or number of the floor in a building, as part of an address." How much easier if the identifier had been `http://psi.oasis-open.org/UBL/floor` and the definition located at that same address...

6. ACKNOWLEDGEMENTS

Thanks to Suellen Stringer-Hye, Patrick Durusau, Motomu Naito, Gabriel Hopmans, Alexander Sigel, and Sylvia Schwab for their comments.

7. REFERENCES

[1] Booth, David, *Confusion on httpRange-14 decision*, 2006-02-20, http://lists.w3.org/Archives/Public/www-tag/2006Feb/0067.html.

[2] Booth, David, *thing-described-by.org*, http://thing-described-by.org/.

[3] *del.icio.us*, http://del.icio.us/.

[4] *Flickr*, http://www.flickr.org/.

[5] Halpin, Harry; Thompson, Henry S., *Web Proper Names: Naming Referents on the Web*, http://www.webpropernames.org/paper/.

[6] Masinter, Larry, *"duri" and "tdb" URN namespaces based on dated URIs*, 2004, http://larry.masinter.net/duri.html.

[7] Mealling, M.; Denenberg, R. (eds.), *Report from the Joint W3C/IETF URI Planning Interest Group: Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): Clarifications and Recommendations*, IETF, August 2002, http://www.ietf.org/rfc/rfc3305.txt.

[8] Newcomb, Steve; Biezunski, Michel; Bryan, Martin, *ISO/IEC 13250:2000 Topic Maps*, ISO, Geneva 2000.

[9] OASIS Published Subjects TC, *Published subjects for countries in ISO 3166*, OASIS 2005, http://psi.oasis-open.org/iso/3166/.

[10] OASIS Published Subjects TC, *Published subjects for languages in ISO 639*, OASIS 2005, http://psi.oasis-open.org/iso/639/.

[11] Pepper, Steve (ed.), *Published Subjects: Introduction and Basic Requirements*, OASIS TC Recommendation, 2003-06-24, http://www.oasis-open.org/committees/download.php/3050/pubsubj-pt1-1.02-cs.pdf.

[12] Pepper, Steve; Moore, Graham (eds.), *XML Topic Maps 1.0*, TopicMaps.Org, 2001, http://www.topicmaps.org/xtm/1.0/.

[13] Pepper, Steve; Schwab, Sylvia, *Curing the Web's Identity Crisis*, Ontopia, 2004, http://www.ontopia.net/topicmaps/materials/identitycrisis.html.

[14] *Semantic Web Interest Group IRC Chat Logs for 2003-04-10* http://chatlogs.planetrdf.com/rdfig/2003-04-10.html#T20-59-04.

[15] Svenonius, Elaine, *The Intellectual Foundation of Information Organization*, MIT Press: Cambridge, MA, 2000.

[16] W3C Technical Architecture Group, *What is the range of the HTTP dereference function*, W3C, 2006, http://www.w3.org/2001/tag/issues.html#httpRange-14.

[17] *Wikipedia*, http:wikipedia.org.

[18] *WordNet,* http://wordnet.princeton.edu.