Binding Points for Subject Identity

The case for standard Published Subject Indicators

Bernard Vatant
Mondeca
Les Aubergeries
05380
Chateauroux-les-Alpes, France
phone (33) 04 92 43 20 71
email bernard@universimmedia.com
web http://www.universimmedia.com

ABSTRACT

Exchanging information about a subject with semantic tools, and singularly topic maps, demands both human and computer readable ways of agreement about this subject identity. Since subjects are always addressed through representations, hence agreement about subject identity must be grounded on some sort of representation, be it a name, an addressable resource, an URL or an XML <topic> element in a <topicMap> document. An overview is made of how those different representations can achieve or not the task of indicating subject identity in a non-ambiguous way.

To address this problem, the XTM specification recommends that subject identity should be defined as far as possible by reference to Published Subject Indicators. But requirements for any standard definition, structure, management and validation of those PSIs remain to be defined, and in fact such requirements were deliberately considered out of the specification scope by its authors. Propositions are made for such requirements, grounded on the claim that a PSI should be a binding point for every possible representation of its subject.

Beyond this problem of definition lies another important one. What are the best semantic ways to use PSIs? Merging topics has been a widely addressed process in the Topic Map community, and is in fact the only one really tackled by the XTM specification. But using PSIs as binding points for collaborative Semantic Networks of independent topic maps is proposed as an alternative to merging Very Large Topic Maps. Such distributed process could lead to a new vision of Subject Identity, grounded on dynamic knowledge organized around such binding points.

KEYWORDS

Topic Maps; Subject Identity; Knowledge Representation; Knowledge Sharing; Published Subject Indicators; Networks

1 INTRODUCTION

Central to the Topic Map paradigm, and in fact to all Semantic Web technologies, is the question of what a subject is. Knowing pretty well it was a very elusive question, the authors of the Topic Map specification have given a completely elusive answer. A Topic is a system representation, or reification, of a subject. But what is a subject ? Anything you want, absolutely anything whatsoever you can think of, represent, talk or write about. Letting the users and authors deal with that metaphysical non-definition, they have nevertheless addressed the subject identity issue in its most pragmatic aspect: how will I know that the subject I am talking about is the same one you are talking about? This process of transferring the definition question from the high philosophy sphere down to the somehow more manageable social and technical levels, by making it an agreement problem, transforms it in a genuine technical issue. Setting standard efficient solutions for both humans and systems to know whether they agree or not on subject identity of a given topic, is really what is at stake.

2 FROM SUBJECT REPRESENTATIONS TO SUBJECT IDENTITY

Topic Map paradigm makes a fundamental distinction between addressable and non-addressable subjects. Non-addressable subjects are ideas, concepts, real world objects, anything that cannot be directly handled by a computer, and can be addressed only through representations or signs in any given language. Some questions arise in an identical way, whatever the type of representation is: name, image, icon, text, XML file, URI ... The first one is how to know if a given sign may represent more than a single subject, simultaneously or successively. It's the question of representation ambiguity. The second one is how to know that two or more different signs represent the same subject. It's the question of representation multiplicity. Let's see how those questions are addressed in Topic Maps through names, resources and subject indicators, all of which may be considered to begin with as ways to catch subjects in representation.

2.1 Names versus identity

Topic Maps specification claims an essential distinction between subject identity and topic name, the former being considered as an absolute definition of the subject, and the latter as a characteristic, likely to be valid only in a context, specified by a scope assignment. That distinction seems quite relevant, and is a core feature of the paradigm: semantics are in the "ontological" subject, whereas names deal with syntactic representation. But this smart principle tends to often be in conflict with common sense, which considers naming as the first and obvious way to define a subject identity, and this conflict is showing somehow in the XTM specification itself through the following sentence

"Base names are subject to the topic naming constraint that two topics with the same base name in the same scope implicitly reify the same subject and should therefore be merged"

Despite the in-principle above distinction, it seems that this constraint implicitly acknowledges that names support some weak form of identity, that other topic characteristics (roles and occurrences) don't support. And as a matter of fact, this topic naming rule, intended to be a safeguard against the risks of name ambiguity, is really a very restricting constraint, since whenever a new topic is added to a topic map, the author—be it human or computer program—has to somehow check if the name given to the new topic is not already used for another topic in the same scope, and if ever, cope with that and find another name, or disambiguate by assigning different scopes. In any case, the difference of identity between two topics with the same name must have been established by another way than these very identical names. What shall the teacher do if another student named Peter Martin enters the classroom where there is already one Peter Martin? He'll have to add some precision to one or both of the names on his list, like Peter Martin (Younger) and/or Peter Martin (Older), to avoid future mistakes due to ambiguity.

Moreover, it's unlikely that every topic map will be built by single hand-made human authoring, or even out of merging of smaller topic maps. One would like to add information from a non topic map data base into a topic map, and constantly update it, for example, or implement automatic attachment of topics and occurrences out of data mining, or run inference rules on existing topic maps to extend them through new associations ... and such problems of duplicate naming will occur of course, and the system will have to handle them. How it will cope with that sort of ambiguity situation is an open issue. It can transfer the name conflict up to the "human management" to settle them, or escape the difficulty by adding some random identity number to any next topic name in the TM, so that it will never get unexpected merging ... and indeed never any merging at all. But that would not be very conformant to intended Topic Map semantics and purpose.

On the other hand, the Topic Naming Constraint does not prevent from representation multiplicity, and from getting the same subject represented in the same or different Map(s) by different Topics with different identities and different names, or very subtle variants of the same name. If a Map is created by a single consistent author, that should not happen. But many Maps will be created through collaborative authoring tools, or out of automatic processes like merging or data mining. In such a situation, if Topics should be merged because they are indeed reifying the same subject, there is no way a process will find by looking only at their (different) names.

The conclusion to draw of all the above is quite clear. Even if topics are often searched and even first defined through their names (in a given scope), it's clear that names carry only a weak form of identity, and that inter-agreement on subject identity can't be grounded on names only. That's

why the Topic Map specification claims that identity must rely on other representations.

2.2 Resources and "addressable subjects"

To avoid addressing through a potentially ambiguous name, an addressable resource may be considered as a subject in itself. In this case, this resource is called an "addressable subject" by the XTM specification. It may look that such a definition of subject will prevent from any ambiguity in the subject identity, but it implicitly assumes some requirements, among which a certain stability of resource's both content and address.

Stability of content: Defining as subject "Document published at http://www.topicmaps.org/xtm/1.0" implicitly assumes that both this address and document have a certain character of permanence. If content changes, a thing which could happen even for such a standard specification (at least in the form of minor editorial changes, like fixing typos, changing fonts or images format ...), and is most likely to occur for any addressable resource on the Web, to what extent does the addressable subject change or not? Strictly speaking, every change in the document makes it a different version, and therefore should introduce a different addressable subject. But what is the amount of change beyond which one would really want to change the subject, hence reify it by another topic?

Stability of address: If the document has been moved to a new address, does the subject change? One may consider it does not, if the document is always "the same" at the new address. But it does not prevent the obsolete address in the Topic Map document from pointing to nowhere, which seems to indicate that if the addressable subject has not changed, at least it's moved.

Therefore it would be reasonable to consider that an addressable subject consists of a permanent address where one can find a permanent resource content. Unfortunately, permanence for both resource and address is wishful thinking for most on-line information. And given the dynamic state of the Web, permanence of an address is generally more likely to occur than permanence of content, and it would be dangerous to transfer here a way of thinking inherited from resource management in the library universe. Paper documents have stable content, even if they keep changing address. In a library, you can consider a book as an addressable subject, even if you have hard time finding out its address on the shelves because the library has been reorganized since your last visit. When you eventually find the book you're looking for, you know its content has not changed since it's been printed or the last time you handled it. On the Web, it's often the other way round. The book does not move, but the text inside is ever-changing ...

2.3 Subject Indicators

If now we consider a so-called non-addressable subject, the resource is considered as a subject indicator rather than the subject in itself, and the problematic is slightly different. At the above quoted address http://www.topicmaps.org/xtm/1.0 there is an addressable document which indicates quite well a non-ambiguous subject, of which a name is "XTM specification version 1.0". I can certainly download and print a document from there, and I could with good reasons pretend it to be a personal copy of the specification. This address is intended to be stable. The resource is a standard specification, so it should not change content. But maybe some minor modifications in comments or other parts of the document may occur with the time. Should I say that the subject has changed, and modify the subject identity? Certainly not, and at this point, we begin to have a different notion of what a subject indicator could be, and see that something is happening, at the boarder between addressable and non-addressable, at this address where the subject indeed is "living". A subject indicator may be understood as a person (the resource) living at a permanent address (the URL). Even if I know the person is ever changing with time, she's considered to have always the same identity. If I've not been visiting that address for a while, I hope anyway I will meet and identify there the "same ever changing person."

3 REQUIREMENTS FOR STANDARD PUBLISHED SUBJECTS INDICATORS

In principle just any resource and address, provided the above stability requirements are satisfied, could be used as Subject Indicator. But it should make sense that dedicated resources—Published Subject Indicators—should be defined and declared by their authors and publishers to be maintained for this specific and only use. Let's see what kind of requirements one would need for that.

3.1 Declared and dedicated resource and address

The first technical requirement is that authors and publishers declare in a standard non-ambiguous way that what they publish is intended to be a Subject Indicator, and that this publication has this unique dedicated purpose. This declaration could appear for example in the form of a dedicated standardized namespace, so that any user could recognize a standard PSI directly from its very address syntax. Another obvious way is to define a PSI as a <topic> in a <topicMap> document, and to declare it is intended to be a PSI through an <instanceOf> element, referring to some authoritative address like http://www.topicmaps.org/psi/psi.htm (a recursive PSI for the notion of PSI).

3.2 Bootstrapping authority and expertise

Any dedicated repository, following closely the previous rules of declaration and stability, would have right to pretend to be a reference PSI repository. But it's clear that only validation by a community will ratify this unilateral declaration. The authority and expertise of the authors and publishers will be of course a strong basis to begin with for this ratification. But,

as for any standard, the wide-scale ratification is likely to result from some bootstrapping process. A good standard is a standard which is widely used and acknowledged ... so it should be very interesting for any PSI repository to maintain an index of its most significant public users, and that every single PSI itself contains an index of occurrences, back linking to the <topic> elements in topic maps using it. Of course, if a PSI repository is itself organized as a topic map, it will be all the more easy to achieve that. The community of users implicitly or explicitly defined by the "back links" is what is called hereafter the "external content" of the PSI. It is the very basis of the bootstrapping process. New users will choose a reference PSI not only for the sake of its content's quality (definition, description or representation of the subject) but also for the sake of quality and width of its community of users. As time goes, this "external content" will be in fact the most important to build up authority.

3.3 Context of use and validation

The Topic Map specification does not allow <scope> to be attached to <subjectIndicator>. This seems to mean clearly that a subject identity should be defined in an absolute way, independent of any context. This viewpoint is highly questionable. An useful subject has to be considered as a "subject of conversation", and this conversation (otherwise called sometimes knowledge interchange) is taking place inside a given community of language, interest or knowledge, which can be called a context of use, and the PSI should somehow indicate its context of validity. But "context" seems more difficult to define, if possible, than "subject" itself, and context is itself a subject! Moreover, the authors and publishers of PSI are themselves parts of the context, and as such can't be aware of its fuzzy limits. All they should honestly declare is: "From where we are and around us as far as we can see, this is how we understand that subject." And the potential user will say : "My view of the world is slightly different, but all the same it makes sense to me and I'll use this PSI." That kind of conversation is expanding both context of validity and authority, in the same bootstrapping process, and eventually may lead to some auto-organized emerging answer to the difficult context question.

Therefore the context of validity of a PSI should better be defined in this recursive and bootstrapping way, by the very extent of the community actually using it, than declared to begin with by its first creators and users. Maybe we can identify the context defined this way with the above "external content" of back links to the community of users.

3.4 Content: towards a minimal standard topic map structure

All the above pleads clearly for a social and dynamic definition of subject, going far beyond the distinction between "authors" defining a subject inside a PSI and "users" pointing to this absolute definition. In the bootstrapping conversation process defined above, the PSI's address will be used as a binding point. But the question of content is still open. Should there be any resource content at all except the back links to users? If there is, is it possible to envision a standard form for this content?

An interesting approach is given by Seruba PSIs. Neither "definition" nor "description" is given for any subject (or concept), except, like in a thesaurus, the relationships with other subjects, given in terms of roles in standard associations. That kind of structure seems a most natural one for a consistent PSI repository, and the requirements for this kind of repository can be summed up as follows.

1. A standard PSI resource should be a <topic> element in a topic map where every other <topic> is itself a declared PSI.

2. The occurrences of this <topic> should link back to published topics using it as <subjectIndicator>, such back links being defined through <resourceRef> elements.

3. The association and role types linking those PSIs together inside the repository map should be defined by "meta-level" PSIs in some other topic map. A set of such "meta-level" PSIs could of course be subject of general recommendations inside a given community.

This last requirement is to be compared with efforts in various communities to leverage so-called "upper ontologies". Expressing the relationships of those upper ontologies in the form of general association and role types, and define them as such "meta-level" PSIs, would be a very interesting path of work to follow.

The above defined structure is of course questionable. Another sustainable viewpoint is to consider that a PSI should not even declare associations with other PSI, because this would constrain or create contradictions with associations in users' topic maps. This extreme case for "empty" PSIs, reduced to their external content, might make sense at the end of the bootstrapping process, the same way location names are losing along ages their original sense, but keep their locating capacity through road signs pointing at them. People keep on understanding efficiently where "Big Elm Cross" is, and follow roads to and from it, even ages after Big Elm is cut down, and even is they don't know any more how an elm looks like because elms have long ago disappeared in the surroundings.

4 USING PSIS AS NETWORK BINDING POINTS

One thing is to define PSIs, another one is to use them in a consistent way. In the XTM specification as well as in the ongoing debates about interoperability and knowledge interchange, the main if not only use of PSIs for interoperating topic maps seems to be merging of topics with the same subject. Though merging is a fundamental feature to insure inner consistency of any given topic map, or in specific applications like index, catalog or data base fusions, it may not be the killer approach for binding independently managed and continuously updated on-line topic maps. What could that

binding consist of, and in what ways is it different from merging.

Suppose we have two topic maps TM1 and TM2 developed in the same community context, or in different but somehow related or widely overlapping contexts. Some identical subjects are represented in both maps, this identity being discovered out of some query process, or human agreement between TM1 and TM2 authors, or more surely by reference to the same PSIs. The corresponding topics would be merged if someone was to build TM3 as a reunion of TM1 and TM2. But it happens that nobody wants or need that TM3 at all, because they do not want to support the cost of maintenance of another map, or because the contexts of TM1 and TM2 are different in many ways, the reunion of those contexts would not really make sense, and maybe it would take a hard time to avoid unwanted merging, and get rid of vocabulary ambiguities, or TM1 and TM2 may have all sorts of ontological, social, political, economical reasons to keep their independence. But they would like to somehow be bound through their common subjects, so that a query on any one may spider into the other one through the common subjects.

Let's say some subject S is represented in TM1 by topic T1 and in TM2 by topic T2. T1 would make for a good <topicRef> for T2 in TM2, and of course the other way round. There could be a bilateral agreement between TM1 and TM2 to do it that way. The two topic maps could then be linked through a bunch of well-chosen "strategic" common subjects, at a lower cost than full merging. If now we consider a network of dozens or hundreds of topic maps wanting to be bound that way, it would be clearly unmanageable.

But suppose each individual topic map in the network refers to a common PSI through a <subjectIndicatorRef> element, and this PSI links back to each of them through <resourceRef>. In this situation the PSI will act in fact as "Big Elm Cross", allowing spidering from one map to another through this common subject, without unnecessary merging. Such use of PSI will create kind of synapses linking in networks independent but interactive topic maps. Networks of relatively small but very ontology-consistent and human-manageable, not to mention human-browsable topic maps, that search engines could spider through PSI binding points is certainly a more exciting and sustainable perspective than building Very Large Topic Maps with all the related problems of management, updating and ontology leveraging. In such Networks, the Subject Identity could be thought in a radically new way: not "defined" at a central point, it will be distributed all over an attractor of resources pointing back and forth a same neutral point. Through collaborative authoring tools and Open Hyperdocument Systems, this Subject Identity could have a dynamic existence, maintained and updated through ongoing conversation around it.

5 ONGOING PROJECTS AND PERSPECTIVES

Such open topic maps networks remain to be developed and tested. But some ongoing projects are working in that direction. Already mentioned before, Seruba (http://www.seruba.com) has proposed a core of PSIs for association and role types used by its Lexicosaurus multilingual ontology (http://www.lex4.com), and a sample of PSIs linked by those associations (http://psi.seruba.com). This same Seruba Lexicosaurus is in the process to be used link conceptual subjects in the Semantopic Map project (http://www.universimmedia.com/semantopic.htm). Binding this last project to other "semantic websites" like XML.fr has also been considered. Another interesting PSI repository could be grounded on the XTM Cyc Upper ontology (http://www.doctypes.org/cyc/cyc-xtm-20010227.html

Of course, testing the scalability of the above model is linked to a widespread adoption of topic map technology for published ontologies, web directories and other on-line knowledge bases. But in any case, the debate around PSIs should lead to a better comprehension of the subject identity issues and a more widespread understanding that interoperability of systems is not a mere technical problem, but is not conceivable without parallel interagreement between knowledge communities. In the perspective of development of the Semantic Web as a tool for Global Knowledge Interchange, this fundamental social perspective should not be forgotten. Semantic Web will be a non-sense if people around it do not develop more efficient tools to agree on "what they are about". Beyond all names, representations, identities, processes, a subject has to remain above and before all a subject of conversation.

BIBLIOGRAPHY

[**JP 01**] XML Topic Maps: Creating and Using Topic Maps for the Web—Jack Park, Editor—Sam Hunting, Michel Biezunski, and Steven Newcomb, Technical Editors—Addison Wesley, September 2001.

[**BH 97**] Information Seeking and Subject Representation—An Activity-Theoretical Approach to Information Science—Birger Hjorland—Greenwood Press. Westport, Conn. 1997. LC 96-51136. ISBN 0-313-29893-9.

BIOGRAPHY

Bernard Vatant is a former high school mathematics teacher, graduate from ENSET (Cachan, France) in 1975, presently working as an independent consultant. He's been recently involved in the development of ontologies for Mondeca Topic Maps solutions, and is Participating Member in XTM Authoring Group.

He has a 20 years long background in astronomy and epistemology popularization, of which most recent episode is the management of a multidimensional website focused on solar astronomy, sustainable development and knowledge representation—www.universimmedia.com.

His main present research interests are collective knowledge representation and web indexing ; his experience as a former editor in Open Directory Project has led him to en-

vision further alternative non-hierarchical tools grounded in the Topic Maps paradigm.

In everyday life, Bernard and his wife are managing a (flexible) complex family of (up to) 5 children, in the French Southern Alps.