Understanding Subjects and Subject Proxies[1]

Patrick Durusau, Patrick@Durusau.net

1 The assistance of Steve and Vicky Newcomb, http://www.coolheads.com, in preparation of this paper is gratefully acknowledged. Any remaining errors are solely my responsibility.

This is a very "lite" treatment of subject proxies. For the full story, see "Understanding Subject Proxies," at: http://www.durusau.net/topicmaps. As of May 8, 2005, forthcoming.

Introduction

Concepts such as “subject” and “subject proxy” get tossed around a lot in discussions about topic maps. What's worse is that discussions of concepts tend to be really dull. The following tries to avoid dullness by being short, limiting itself to 7 pages of substantive content, and illustrating the use of subjects and subject proxies to: enable smarter searching, improve tracking of terrorists, avoiding or find photos of costume malfunctions, and finally, empowering users to say what they mean, unconstrained by systems that seek to enslave users for the benefit of certain software programs.[2]

What is a subject?

The starting question in most discussions of topic maps is ‘What is a subject?’ Fair question. The latest draft of the Topic Maps Reference Model (TMRM) says: “Any thing whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be indicated by any means whatsoever.” Stated in an equivalent but less formal way, a subject is anything about which one can converse.[3]

Conversations can be had about public figures, TV shows, products, characters from novels, religious texts or moral values, in short, anything.[4] But all statements of subjects, at least from the speaker's point of view, are transparent. That is to say that the speaker, presumably, knows what subject they are talking about, even if the listener doesn't.

And in a conversation, if there is uncertainty about what subject the speaker has in mind, a listener can always ask: “What do you mean?” Hopefully the speaker will respond with more information about the subject until the speaker and listener both come to understand the identity of the subject that is the object of the conversation.

The problem that topic maps address is how to enable the same result

2 There is a great deal of technical complexity that is being elided over in this paper. The point here is to make the payoff of studying that complexity readily apparent.

3 Granted that leaves out subjects that can't be talked about but then Wittgenstein advises those must be passed over in silence anyway. Tractatus Logico-Philosophicus, 7.

4 Relationships are also subjects but the means for indicating them are various and some of them are complex. Representing relationships is treated more fully in “Understanding Relationships as Subjects,” forthcoming.

as the ensuing conversation between speaker and listener in a computer environment. That is to say, how can software "know" when a subject has been spoken of and determine what subject was meant by the equivalent of a speaker (an electronic record of some sort) in the absence of any ensuing conversation. In other words, how does anyone, including a computer system, identify a subject?

What is subject Identity?

So how does one go about identifying a subject? Try the following experiment: take out your driver's license, a major credit card and your cellphone. On a piece of paper write down the driver's license number, the credit card number and your cellphone number.

All three of those numbers are identifiers for you as a subject. Note that none of those numbers are you, so in the topic maps paradigm, those values are said to 'indicate' a subject, in this case you.

Well, that seems simple enough, but there are problems with using unique numbers to indicate subjects. The most obvious one is that someone has to be responsible for generating the unique number for each subject. The credit card company obviously has an incentive to do so, at least at an interest rate of 18% or more per year. But there is no economic incentive for the credit card company to provide unique identifiers to indicate all other subjects. Not to mention that the number of potential subjects needing numbers is unbounded, which would make assigning unique numbers a practical impossibility.[5]

A less obvious problem is that if the unique number, name or URI is not recognized as indicating a particular subject, what good is it? When the system encounters a 9-digit string what does that mean?[6] The original provider of the data presumably knew what it meant, but they are not available for questions. What subject, if any, does it indicate?

How can a user know what subject it signifies? The answer is that they cannot. In other words, a unique number/name/URI is semantically opaque. That opaqueness arises in part because a unique identifier has significance only as part of something outside of itself, such as a catalog of

5 Although one might wish that the telemarketers for credit card companies would make the attempt. At least it would cut down on the number of interrupted dinners, TV shows and moments of domestic bliss.

6 All social security numbers, an identifier assigned by the United States government always has 9 digits. I am presuming the existence of other 9 digit numbers that are not social security numbers.

such identifiers. Or if way the identifier is encoded somehow identifies the subject. Let's take a closer look at that semantic opaqueness before examining a solution to that problem.

Unique Names, Numbers and URIs

To discuss the question of what subject is indicated by a unique name (number or URI[7]) and the issue of semantic opaqueness, let's consider a library based example. Libraries and librarians have spent centuries cataloging materials (originally clay tablets, then scrolls, more recently codexes/books and now audio-visual materials) so their experience will be useful here.

When a book is catalogued by a library a record is created that has in part the following information:[8]

- Library of Congress Number
- International Standard Book Number (ISBN)
- Author
- Title
- Publisher
- Date

There are several subject indicators in this example. The Library of Congress Number, the ISBN, as well as Author, Title, Publisher and Date, if taken together, indicate one book in particular.[9] A subject indicator is simply information that in a particular context, indicates a given subject.

All of the example seems transparent enough to most human users but is it transparent to a computer system? What if the computer was instructed to find entries where “dc:creator = Mark Twain”?[10] True enough a search of the electronic catalog finds the string “Mark Twain” but it has been labeled

7 URI is used herein to refer to current URI schemes and not URIs in general as defined in RFC 3986.
8 This listing is extremely incomplete from a library cataloging practice point of view. Further detail was omitted in order to advance the explanation and with apologies to all library catalogers.
9 A subject indicator may consist of several parts.
10 The “dc:” prefix is a namespace for Dublin Core metadata. See: http://www.dublincore.org.

“Author” instead of “creator” so the search merrily passes it by.

What is lacking in that scenario is any means for the computer to discover that it should be treating “Author” and “dc:creator” as indicating the same subject. Not the subject of the author of a particular book, but the subject of an author/creator of a book.

What is needed is a means to provide a mapping of indicators for the same subject and the context in which those indicators occur. But it will be quickly protested, there are several indicators for the book described by a record, such as the ISBN, Library of Congress Number and others. Yes, and the answer to that objection highlights the similarities (and differences) between a library record and a subject proxy.

What is a Subject Proxy?

A subject proxy is a surrogate for a subject. That is to say that in an electronic information system, a subject proxy represents a particular subject. That is quite similar to a library record for a particular book. And there are other similarities as well.

For example, the user of a library catalog can search for a particular book using its ISBN, its Library of Congress number, or if they happen to know it, its accession number.[11] In other words, a book in the library has multiple addresses that are held together in a single record. And any one of those addresses will allow the user to find a particular book.

Subject proxies also support multiple addresses (read subject indicators) for subjects but with one wrinkle that is not currently supported in non-topic map systems. The subject indicators themselves, such as “author” for example, are features of “subject proxies” and can be addressed in any number of ways – all the ways in which the subject may be addressed in in various contexts. A system that understands how subjects may be addressed in any context can determine when two or more subject proxies are in fact proxies for the same subject, even if they have different addresses for the same subject.

For example, assume that the earlier example of "author" was in fact a subject proxy that had the following subject indicators:

Subject Proxy
    Author
    dc:creator

Such that when the search encountered "author = Mark Twain" while searching for "dc:creator = Mark Twain", instead of passing over the entry, any systems that understands "author" and "creator" in their original contexts will return the same subject proxy.

It is important to note that neither of the information systems were required to change any legacy data nor present data entry practices in order to enable seamlessly searching across different way to indicate the same subject.

Well, that is interesting for libraries but what about other (read commercial) applications? Given the $billions that are being spent by the Department of Homeland Security, I suppose that will qualify as a "commercial" application area. Consider the following example drawn from the 9/11 Commission report.

Khalid al Mihdhar, one of the 9/11 terrorists traveled more than once to the US under his own name. In early 2001, it was discovered that the person who had directed the bombing of the USS Cole was an associate of Khalid al Mihdhar. A subject proxy for Khalid al Mihdhar would be the place to not only find indicators for him, but other information, such as his known associates. Topic maps can assist in extracting and providing useful access to this sort of information from the flood of data processed by intelligence agencies on a daily basis. Possessing information and having meaningful access are both required for action to be taken in a timely fashion.[12]

12 That is not to imply that topic maps would have made any difference in the events of 9/11. After reading "The 9/11 Commission Report" one may reasonably conclude that no technology, however clever, would have addressed the then current problems of US intelligence gathering efforts. Topic maps can quite easily support delivery of information to users with appropriate clearance or even alert users that additional information is available that requires someone with the proper clearance to view the information..

With a topic map engine, all the subject proxies for a particular subject can be merged, resulting in all the information about a particular subject being available at one place. Think of it as a subject, like the book in the library catalog, having multiple addresses that all lead to the same place.

The addresses of subjects need not be "semantically opaque." There are no prior constraints on the ways in which subjects can be addressed. A subject address can have multiple components, including components that are themselves subject addresses. And, the use of multiple addresses allows users to discover information entered by other users who have a completely different view of the same subject and its relationships to other subject.

Subject Proxies Here and Now

Topic map software is available now that uses the approach described above to reliably merge information about subjects.[13] But what if you are not quite ready to take the full plunge into topic maps? Can you still see the power of subject proxies for searching and filtering content?

Certainly, although not marketed as a topic map or subject proxy technology, the "tagging" used by Flickr, http://www.flickr.com/, is an example of a socially constructed set of subject proxies.

In a nutshell, the Flickr system allows users to "tag" their photographs with "annotations," ("subject indicators" in topic map lingo). There are the obvious problems such as someone saying "cat" and another saying "feline," but any interested person can add the additional subject indicator to an image just as would happen with a subject proxy in a topic map system. Such additions could also be automated on the basis of who posted the images or any other criteria.

Not only does that help with searching, but with filtering content as well. What if a student is doing a research paper on breast cancer. Does it make a lot of sense to block sites or stories that have the word "breast?" Using the Flickr system to create subject proxies, imagine that all the photos

13 See for example, http://www.versavant.com

of Janet Jackson at SuperBowl XXXVIII were tagged: breast, SuperBowl XXXVIII. Also imagine that all the images of breasts in Scientific American, American Cancer Society, etc., online publications are "tagged:" breast, SciAM or breast, ACS. It is not hard to imagine filtering software for a student's computer that looks at more than simply the term "breast" in order to determine whether to display the resource or not.

The real advantage to such a system is that it harnesses the interest of millions of users to supply "tags" or subject proxies if you like. Images with no "tags" or subject proxies could be automatically rejected or only those classified by trusted sources accepted. The resulting system would certainly be more flexible and useful than conventional filtering software.

Conclusion

Subject proxies enable users to describe their subject indicators for any subject they want to talk about. And subject proxies have the ability for others to indicate they are talking about the same subject, but using a different subject indicator. Rather than imposing on users a standard set of terms that may or may not meet their purposes, topic maps and subject proxies empower users. [14]

Subject proxies (with a topic map engine) enable the gathering of all the different indicators for a subject into one place, enabling you to follow indicators into different ways to describe relationships with the same subject. That is to say you can discover all the information about your subject, even if sometimes it is indicated completely differently than you would in your own universe.[15]

14 The idea of empowering users to exploit their own ways of communicating with themselves and others should be contrasted with the idea of requiring users to adapt to way of communicating that happen to be convenient for computers. We have the technology to make computers adapt to human beings; why why should the burden be placed on humans, instead of computers?

15 Topic maps do this without worm holes, black holes, translation nanites, bio-mimetic gels or other expensive, dangerous or difficult to obtain technologies.

Biography

Patrick Durusau is an independent consultant, advising clients on the design, implementation and integration of complex information systems.

He was the Director of Research and Development for the Society of Biblical Literature. (2000-2005)

He is the co-editor (with Steven R. Newcomb) of the Topic Maps Reference Model and Chair of the Published Subjects TC at OASIS.

Patrick is the Chair of the US National Body to JTC1/SC34, a member of the board of directors of the Text Encoding Initiative Consortium, technical lead for the OSIS Project, a joint markup project of the Society of Biblical Literature, the American Bible Society and the United Bible Societies.

To arrange speaking engagements, lectures, workshops, assistance with technology grant proposals, consulting on topic maps or other information systems, contact Patrick Durusau at: patrick@durusau.net or call 1-678-625-0995.