Who's on First?

By

Patrick Durusau

Patrick@snowfallsoftware.com

February 21, 2007
About the author:

Patrick Durusau is the standards lead for Snowfall Software. He has served on the Text Encoding Initiative Board of Directors and is the technical lead for the OSIS project (a standard for encoding bibles in XML). Today, most of his energies are focused on topic maps and related technologies both in ISO and OASIS. He is currently the chair of INCITS V1, the US National Body representative to ISO/IEC JTC1/SC34, which is the committee responsible for SGML, HyTime, Document Schema Definition Languages and Topic Maps. He is the ISO Project Editor for OpenDocument Format (ISO 26300), co-editor OpenDocument Format at OASIS, a member of the OpenDocument Format TC and Chair of the OpenDocument Format Metadata SC at OASIS. He also serves as a co-editor of the ISO/IEC Topic Maps Reference Model (ISO/IEC 13250-5) and is acting Convenor for WG 3 of ISO/IEC JTC 1/SC 34.
Illustration 1: A promising modeling career cut short by an author.
Introduction:

I want to begin by thanking Professor Kathryn La Barre for this opportunity to talk to you about topic maps. Although I have spent the last six years working on various aspects of topic maps, you can see from the photo that was made for a promotion for the local library that topic maps aren't my only area of interest.

The purpose of this presentation is two-fold: First, to give you an overview of topic maps and their uses. Second, I am hopeful that this presentation will interest some of you in a topic maps course that will be offered this summer at GSLIS.

I spent ten years as a trial lawyer years ago and so speak with the use of few if any notes. My idea of a good time as a speaker or listener doesn't include reading from a set manuscript. But, I do remember what it was like to try to take notes, understand new material and follow the discussion that occurs in a graduate class. (In both law and theology graduate programs.) These lecture notes are not a word for word transcript of what I will be covering in my presentation but are meant to be helpful both in terms of preparing for discussion as well as reviewing what was discussed in class.

As is the case with most authors who have spent perhaps too much time doing anything, I will occasionally lapse into jargon that is transparent to only those within a very small circle. Should that happen, please don't hesitate to ask for more explanation. Obscurity isn't a sign of profoundness, just poor explanation.
The Problem:

Some years ago, the comedy duo of Abbot and Costello performed a routine known as “Who's on First?” By way of background, Abbott has been named as the coach of a baseball team so long as Costello is on the team. Costello asks Abbot to name the players so he will know their names. A brief setup discussion of the odd names of players starts the conversation and the following ensues in part:

Abbott: ..... Well, let's see, we have on the bags, Who's on first, What's on second, I Don't Know is on third...

Costello: That's what I want to find out.

Abbott: I say Who's on first, What's on second, I Don't Know's on third.

Costello: Are you the manager?

Abbott: Yes.

Costello: You gonna be the coach too?

Abbott: Yes.

Costello: And you don't know the fellows' names.

Abbott: Well I should.

Costello: Well then who's on first?

Abbott: Yes.

Costello: I mean the fellow's name.

Abbott: Who.

Costello: The guy on first.

Abbott: Who.

Costello: The first baseman.

Abbott: Who.

Costello: The guy playing...

Abbott: Who is on first!

Costello: I'm asking you who's on first.
Abbott: That's the man's name.

Costello: That's who's name?

Abbott: Yes.

Costello: Well go ahead and tell me.

Abbott: That's it.

Costello: That's who?

Abbott: Yes. PAUSE

Apologies to all Abbot and Costello fans as I don't do the voices very well! An important thing to note is how the conversation ends:

Costello: Same as you! Same as YOU!!! I throw the ball to who. Whoever it is drops the ball and the guy runs to second. Who picks up the ball and throws it to What. What throws it to I Don't Know. I Don't Know throws it back to Tomorrow, Triple play. Another guy gets up and hits a long fly ball to Because. Why? I don't know! He's on third and I don't give a darn!

Abbott: What?

Costello: I said I don't give a darn!

Abbott: Oh, that's our shortstop.

Costello: (makes screaming sound)

(You can see the entire transcript and listen to an MP3 audio recording of “Who's on First” at: http://www.phoenix5.org/humor/WhoOnFirst.htm)

Unfortunately for many users they reach the same conclusion as Costello and simply give up, even when they know the information must exist but they simply cannot find it.

While the comedy routine between Abbott and Costello is amusing, users don't find the same experience as amusing. What is more, not finding information can have serious consequences.

Medical libraries store the results of clinical trials using MESH (Medical Subject Headings) from the National Library of Medicine. As we all know, medical terminology changes daily and subject headings change in response to those changes.

The medical school at Tulane University participated in a clinical trial and duly stored the results according to the then current MESH categories. Several years later, and after the categories had changed, they had a patient who was treated with a drug tested in the clinical trial. Unfortunately for that patient, a search using the now current MESH headings failed to find the results of the earlier
clinical trial. Results that would have indicated that the patient should not be given the drug in question. As a consequence of not finding that prior clinical trial, the patient died.

In the ensuing investigation, which involved lots of lawyers, it was discovered that Tulane “knew” that the patient should not have been given the drug in question. They had, after all, performed the very clinical trial that demonstrated this patient should not have been given the drug in question. The information had been properly recorded and with enough effort, it could have been discovered. But, to find it a researcher would have to know the significance of the older terminology.

If the consequences of failing to find information can range from frustration to death, depending upon the situation, simply because we are using different terminology, it looks like a problem that merits our attention.

What are our options?

We could all agree to use the same terminology. That seems plausible, at least for the serious cases like medical research.

But is it really plausible? Medical researchers are making new discoveries every day and they have to talk about them to describe their discoveries. Not to mention that medical research is going on all over the world in different languages.

On the other hand, Abbott and Costello were using the same terminology, but they meant different things by it. So, maybe using the same terminology isn't as good of a choice as it might seem.

The problem simply gets worse as more information becomes available. Compare a search of your local website with that of the WWW, both using a search engine like Google. The larger the data set in question, the less useful the results. The larger the data set the more variance there is in the terminology used for the same subjects and the more the same terminology is used to mean different subjects.

And the problem isn't limited to situations where users are simply free to throw up whatever they like onto websites. The DOD Performance & Accountability Report, FY 2005, has the following statement:

    To date, the Department-wide financial statements have received a disclaimer of opinion from the auditors, which means that the financial information displayed in the statements is in such poor condition that the auditors are unable to express an opinion.

Or as noted by the GAO:

    DOD’s stovepiped, duplicative, and nonintegrated systems environment contributes to these operational problems and costs the American taxpayers billions of dollars each year. For fiscal year 2005, the department requested approximately $13 billion to operate, maintain, and modernize its reported 4,150 business systems—an increase of about 1,900 in the number of reported systems since last year. ¹

So the problem isn't limited to people who simply generate uncontrolled content but is also found

where systems are intentionally constructed by people who are trying to solve specific problems. But, those systems are “stovepiped” to serve a specific need. Each of those systems has its own methods to identify subjects in those systems.²

Side note: It isn't recommended that you tell any of your future directors that your part of the library is “un-auditable.”

**Why should librarians care?**

Well, all that sounds like a real mess, but then we all know that the world outside of libraries is a mess. So what's new?

**Why you should fall to your knees and worship a librarian³**

Ok, sure. We’ve all got our little preconceived notions about who librarians are and what they do. Many people think of librarians as diminutive civil servants, scuttling about "Sssh-ing" people and stamping things. Well, think again buster.

Librarians have degrees. They go to graduate school for Information Science and become masters of data systems and human/computer interaction. Librarians can catalog anything from an onion to a dog’s ear. They could catalog you.

Librarians wield unfathomable power. With a flip of the wrist they can hide your dissertation behind piles of old Field and Stream magazines. They can find data for your term paper that you never knew existed. They may even point you toward new and appropriate subject headings.

People become librarians because they know too much. Their knowledge extends beyond mere categories. They cannot be confined to disciplines. Librarians are all-knowing and all-seeing. They bring order to chaos. They bring wisdom and culture to the masses. They preserve every aspect of human knowledge. Librarians rule. And they will kick the crap out of anyone who says otherwise.

I must confess this is my favorite librarian image. I bought one of the t-shirts for my wife, who is a children’s librarian at the local library. By way of additional background, one of my first part-time jobs, while I was in middle school was at a library. I worked at a law library while in law school and have spent most of my life in libraries of one sort or another.

So, why should librarians care if the world outside the walls of libraries can be charitably described as

---

² Note that the number of systems increased from 2,250 one year to 4,150 systems the next year. Not that many new systems, simply overlooked 1,900 systems in the prior report.
³ http://librarianavengers.org/?page_id=3
info-glut?
Well, there is the case I mentioned above about the medical school at Tulane where changing subject headings lead to the death of a patient. But I think there are other, equally compelling reasons.

As the amount of information outside of libraries continues to increase, the question that libraries face is how to effectively integrate that information with their collections? Rather than being islands of sanity libraries can become useful portals to the largely untamed information outside their walls.

Being trained to think outside of categories, that there is more than one way to find information on any given subject, librarians are in a unique position to contribute to solving the problem of effective access to ever increasing amounts of information. Hopefully one of the techniques that you will choose to use in that quest will be topic maps.

**Topic Maps: A Short History**

Topic maps, ISO/IEC 13250, began, as all standards do, in response to a particular problem and that problem shaped the resulting response. (I am in the process of obtaining interviews with those present at the first conversation that can be recognized as leading up to the topic maps proposal for the topic maps course.)

Briefly summarized, O'Reilly, a leading publisher of computer books, had finished a multi-volume series on X-Windows (think about MS Windows if properly written). A number of vendors were selling Unix software that all used X-Windows and they wanted to license the content of the X-Windows series to offer it as documentation with their software products.

That sounds innocent enough doesn't it? The publisher gets to reuse their content, vendors can avoid having to write their own documentation and hopefully, users will get better documentation than they would otherwise. Recall this was in the heady early days of markup and since the text was in SGML, how hard could it be?

A number of experts were assembled and given the task that I just outlined. Several years later, the project split into what is now known as DocBook and the topic maps effort. Oh, I forgot to mention one fact: all the vendors had their own terminology that did not match that used in the X-Windows series of books. Whoops!

The question that was faced by the group was how to map between different terminologies, where neither terminology could be changed to match the other? Sometimes the same words meant different things and sometimes different words were used to mean the same things. Eventually the effort simply crashed and burned.

The topic maps part of the group decided to continue to work on the integration of information issue that arose in the context of mapping the O'Reilly materials to vendor documentation.

Given the markup background of the participants it should not be surprising that the issue was seen as one that could be addressed with markup technology.

The result of that effort was the first topic maps standard, ISO/IEC 13250:1999, which was based on the HyTime standard.

HyTime, well, what to say? HyTime was an effort to enable the development of addressing systems that could address any content that was based upon SGML syntax. It is in use today but only in a limited number of systems and is not widely supported. We will be covering HyTime briefly in the summer course because it remains one of the best explorations of the issues of addressing to date.
The topic maps community fairly quickly realized that using a little implemented standard might not have been the best choice to promote the adoption of topic maps. At the same time there was some dissatisfaction with the time required to produce standards in ISO.

As a result of the perceived problems with continuing in ISO, TopicMaps.org was formed by Steven Newcomb and Michel Biezunski, the editors of the ISO standard for topic maps. Steve Newcomb was the editor of the Standard Music Description Language and a co-editor of the second edition of the HyTime standard. (Noting that Steve DeRose and David Durand were the editors of the first edition of the HyTime standard.) The goal of that effort was to create an XML version of the topic maps standard. The result of that effort was the production of XTM (XML Topic Maps) in a little less than two years. (see, http://www.topicmaps.org/xtm/index.html) Note that version of XTM is purely of historical interest.

The XML version of topic maps was added to the ISO topic maps standard and work began on a comprehensive revision of the topic maps standard.

As it stands today, the topic maps standard has seven (7) parts:

- **13250-1** - Information technology - Topic Maps - Overview and basic concepts
- **13250-2** - Information technology - Topic Maps - Data model
- **13250-3** - Information Technology - Topic Maps - XML Syntax
- **13250-4** - Information technology - Topic Maps - Canonicalization
- **13250-5** - Information technology - Topic Maps - Reference model
- **13250-6** - Information technology - Topic Maps - Compact syntax
- **13250-7** - Information technology - Topic Maps - Graphical notation

Two separate standards detail a constraint and query language for topic maps:

- **18048** - Information technology - Topic Maps - Query Language (TMQL)
- **19756** - Information technology - Topic Maps - Constraint language (TMCL)

**Subject Identity:**

As we saw with the Abbott and Costello and other examples, the question of how subjects are identified plays a critical role in the integration of information about any given subject. If we are limited to only one way to identify a subject we are going to miss information that was recorded that identified the subject differently.

The topic maps paradigm has defined a subject as: “anything whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever.” (TMDM, 3.14)

One obvious solution to the problem of subjects being identified differently is to simply have a single identification for every subject. But, who would enforce such a system? Moreover, what do we do about new subjects that arise in the course of research or even every day life?
Well, perhaps we could simply create a mapping from diverse systems to the system that we want to use. That leaves others to use whatever identification they want but at least we can construct a system that works for us. That works, but only to a degree. The most obvious problem is that the mapping will have to be updated on a continuing basis as new subjects emerge.

But, there is a more subtle problem that is lurking in the background that makes both the uniform identification and mapping solutions extremely problematic.

Let's think about how we are going to identify a subject. Let's start with something rather simple, let's say we are going to identify this class. We know the title by which the university knows the course, its professor, time and location of meeting, etc. That doesn't sound hard does it?

But, what we often miss is that the identification of a subject includes, of necessity, other subjects. That is to say that “university,” “professor,” “time,” and “location,” for example are all subjects as well. Not to mention that the values we associate with those subjects may also be subjects in their own right.

While that may seem like a trivial observation, the impact of it on an information system for integration of information is profound. If we adopt the mapping method to integrate information, we have of necessity abandoned the connections that the subject we are mapping to its original domain. That is the subjects that were originally used to define it.

Which means, of course, that users from that domain, can no longer find their subject other than by adopting a foreign means of describing that subject. Note that this isn't a foreign language issue (although it occurs there as well) but an issue of forcing a user out of their 'universe of discourse” and into one that is not familiar.

Recall the Tulane medical library example? There is little doubt had the material been findable in the “universe of discourse” familiar to the researchers that it would have been found. But because the researcher had a different “universe of discourse,” the information went unnoticed, with tragic consequences.

Another consequence to consider is the investment that users have (by training) as well as their institutions in particular ways of organizing their information. To ask them to adopt even a clearly better system of organization requires not only proof that the proposed system is better but that it is also worth the cost of transition.

If it were possible to preserve the various universes of discourse such that users did not have to change how they access information and yet at the same time have access to information recorded in other universes of discourse, that would be a net win wouldn't it? Rather than viewing semantic diversity a a problem to be solved, why not embrace it to the benefit of all users? That task is the one that topic maps has been working on and we will be turning to the specifics of those efforts.

**Specifics on Topic Maps:**

I am going to cover the mechanisms of the original HyTime version of topic maps rather lightly today so if you want the full detail on that you need to sign up for the summer course.

One of the odder features of the HyTime version was something we called the “topic naming constraint.” Basically that was a doctrine that said that if two topics (we'll get to that in a moment) had the same name in the same scope, that the two topics represented the same subject and should be merged.

After much wailing and gnashing of teeth over the topic naming constraint (TNC), it was abandoned in the current revision. Quite honestly it turned out that the TNC was an effort to enforce a particular
method of information “hygiene.” Yes, standards writers are not immune from thinking they know the “best” way other people should handle their data. I suppose that is part and parcel of writing standards but that one seemed a bit over the top. Particularly in light of the expanded notion of merging that emerged in the revision to ISO 13250.

In a nutshell, the revised version of 13250 recognizes that merging of representatives for subjects is actually a consequence of allowing subjects to have defined identifications. That is to say that if the creators of topic maps can define how to identify their subjects, then merging is simply a natural consequence of comparing those identifications in order to gather all the information about a given subject to a single location. Bearing in mind that any given subject may have multiple, independent identifications.

User specify their subject identity rules via a “legend.” Just as you use a legend on a road map to learn how to use it, a legend specifies for a topic map the rules that govern the representatives of subjects in the topic map. The Topic Maps Data Model, which specifies the rules for the XML syntax in the topic maps standard is one example of a legend.

**Topic Maps Data Model (TMDM):**

There are three basic parts to a topic map as defined by the TMDM:

- Topics
- Associations
- Occurrences

**Topics:**

In the TMDM, subjects are represented by topics, relationships between subjects by associations and occurrences are information about a subject. There are some nuances to those statements but that is enough to get us started in talking about topic maps.

One of the more useful distinctions that topic maps introduced is the difference between an “addressable” versus a “non-addressable” subject. An “addressable” subject is one that can be addressed over a network as a series of bytes. A “non-addressable” subject is one that cannot be addressed over a network. An example of the latter would be the Statue of Liberty.

That distinction avoids the various proposals that plague RDF in order to distinguish what is meant by a URI. (Note that if you are working on international standards, the current preference is for IRIs, which don't disadvantage CJKV languages.)

A term you will encounter in topic maps standards and papers is “reification.” It is one of the more unfortunate holdovers from the original topic maps standard. Basically it means to be represented by a topic. But, when you read the TMDM, you will notice that names (of topics), associations, and occurrences, for example, all merge with each other even if they have not be “reified.” But, reification is required, by the TMDM, for an association to be a role player in another association or to provide an occurrence with additional properties.

**Associations:**

Associations represent relationships between subjects that are represented by topics. In the context of
an association, topics are referred to as role players. The roles that they play are also topics. An example of an association would be the marriage of George and Mary where George is said to play the husband role and Mary plays the wife role. Associations are assigned a static type and can have scope (scope in just a minute).

Finding good associations that can be used consistently is one of the more difficult parts of topic map authoring.

**Occurrences:**

Occurrences originally began as a means of indicating where a subject (represented by a topic) could be said to “occur” in a text. Remember that the original use case was indexing so that is really not surprising.

As the standard developed, a need to record other information about the subject represented by a topic was recognized. So occurrences expanded beyond their original role to include simply information about the subject. In the current version this can also include XML encoded data.

It is worth noting that an occurrence is actually a light weight form of association with fixed roles for the topic and the occurrence. Occurrences can have a type (which is a topic) as well as scope.

**Topic Name:**

The topic name is used to attach a name to a topic. Actually a form of occurrence. Topic names can have a type as well as scope.

**Scope:**

Scope is used to indicate that information is only useful or valid in a particular context. Scope can be used with names, associations and occurrences.

**The Future of Topic Maps:**

Obviously this brief overview of topic maps is insufficient to enable anyone to go forth and begin conquering incompatible information systems using topic maps.

One missing piece that will be covered in the summer course is topic maps syntax. There is the “official” XML syntax but I would not recommend using it unless you really like typing. As I mentioned earlier there is a pending item to construct a “compact” syntax that should aid in hand authoring of topic maps. Personally I think that actually writing topic maps by hand slows down the authoring process so that one has to consider the choices that have been made.

On the other hand, we will need to touch on the automatic creation of topic maps as well as “virtual” topic maps, a term used by Professor Barta at Bonds University in Australia. Prof. Barta has created topic maps that query the DNS and merge those topics with others representing equipment on their local network. Given the size and dynamic nature of the DNS service, the creation of a static topic map really isn't a possibility.

There are any number of instances of real world topic maps that would prove worthwhile to study. The topic maps in use at the Office of Naval Intelligence and at the weapons facility at Oakridge are obviously out of bounds because of security reasons. However, the IRS TaxMap, the topic map at the Danish Royal Library and others, should provide more than enough fertile ground for investigation that we won't have to become spies.

I am hopeful that the Topic Maps Query Language and Topic Maps Constraint Language will both be stable enough by late Spring to include material on both of those standards in the course. As all of you
know as librarians it isn't sufficient to simply record information. Recorded information without some means to access it is fairly useless. The constraint language will add an important dimension to topic maps generally as it will enable users to exchange (and enforce) constraints on their topic maps.

We will also be covering the Topic Maps Reference Model (TMRM) with a focus on the issues as they relate to subject identity and not with much focus on the formal parts of the TMRM. It took several years to produce the agreement that is expressed in about 12 pages of formalism and I don't think covering that in detail would be practical in an introductory course.

In terms of software, I don't anticipate that students will have to buy any software. I know most of the topic map software firm founders and am working on obtaining free copies of current software for use in connection with the class.

If I had to summarize why librarians should be interested in topic maps (and you as students in a topic maps course) I think I would put it this way: Libraries have historically been the bastions of semantic and cultural diversity. Shouldn't they take the lead in gathering resources together and providing access in ways that preserve that semantic and cultural diversity

Added post-lecture:

The Gene naming problem I mentioned is discussed in: The success (or not) of HUGO nomenclature, Javier Tamames, and Alfonso Valencia, Genome Biology 2006, 7:402
http://genomebiology.com/2006/7/5/402