Promiscuous Semantic Federation: Semantic Desktops meet Web 2.0

Jack Park
SRI International
jack.park@sri.com

Abstract

What if there are ways to federate all of the many world views of actors everywhere. What would be the use cases under which such a federation makes sense? What would be the benefits to such a federation? What would be the drawbacks to such a federation? Indeed, what are the ways in which such a federation could be facilitated? Why promiscuous? By positing a rhetorical inquiry into the fabric of human cognition and its federation, we open Pandora's Box and unleash many questions, many avenues of inquiry. In this paper, we will limit the discussion to an overview of one approach to semantic federation of human thinking, the coupling of a Semantic Desktop platform, IRIS[1] and its associated *cognitive* framework CALO[2], to a Web 2.0 platform we call Tagomizer[3]. We will show how Tagomizer, a social bookmarking application running on top of a web-based subject map, can assist users in an aspect of Semantic Desktop platforms in their day-to-day cognitive tasks, that of finding information resources related to projects or tasks in which they are engaged. We highlight the fact that social bookmarking is one of many ways in which promiscuous semantic federation can be achieved. We will provide a framework with which answers to questions of the kind suggested above can be found.

1. Background

Ora Lassila once said [1]

> "Once the web has been sufficiently "populated" with rich metadata, what can we expect? First, searching on the web will become easier as search engines have more information available, and thus searching can be more focused. Doors will also be opened for automated software agents to roam the web, looking for information for us or transacting business on our behalf. The web of today, the vast unstructured mass of information, may in the future be transformed into something more manageable - and thus something far more useful."

Let us talk about transforming that unstructured mass of information through a brief introduction to aspects of CALO. SRI's CALO[4] project is one of two projects funded under DARPA's "Perceptive Assistant that Learns" (PAL) program[5]. The goal of the PAL program is to develop an enduring personal assistant that "learns in the wild" (LIW), evolving its capabilities as a personal assistant more and more through automated machine learning techniques rather than through code changes. We created Tagomizer to facilitate LIW for CALO and the program's users by installing a social bookmarking application designed specifically to interoperate with CALO through a web services interface. Tagomizer behaves in ways similar to the website known as "delicious"[6], where, as described below, users create bookmarks using reminder *tags*. A tag can be a word or word phrase. In the future, a tag could be, say, a Wikipedia URL, an image, or other symbol that serves the purpose. When coupled with a semantic desktop application, Tagomizer can serve the purpose of annotating emails, calendar events, chat sessions, files on the user's disk, and other local information resources. As a user evolves a *personal ontology* of tags, conceptual clusters form around those tags, aggregating a multitude of information resources according to the needs and habits of each user.

When a particular user tags some website and notices that others have also tagged the same website, one opportunity for LIW occurs when a user visits the bookmarks of other users. The opportunity for learning is

[1] IRIS: http://www.openiris.org/
[2] CALO: http://calosystem.org/
[3] Tagomizer: http://www.ai.sri.com/software/Tagomizer
[4] CALO is an acronym for "Cognitive Assistant that Learns and Organizes." CALO's name was also inspired by the Latin word *calonis*, which means "soldier's servant" and conjures an image of Radar O'Reilly from the M*A*S*H TV series.
[5] DARPA's PAL program: http://www.darpa.mil/ipto/programs/pal/
[6] delicious: http://del.icio.us/

predicated on similarities in thinking patterns among users and the fact that various users might bookmark websites unknown to others. In this paper, we describe Tagomizer, subject maps, knowledge representation, federation of knowledge through subject maps, and integration of all of those with semantic desktop applications. We will show that LIW can be facilitated by the emergence of what we call *wormholes* [2], links between different world views. We will show how the term *subject-centric federation* occurs when information resources that can be shown to describe the same *subject* are merged. We will show that *wormholes* can emerge when subject-centric federation of different world views brings together ideas expressed in different ways by different people. In some sense, subject-centric federation is a natural process inherent in social bookmarking websites; when integrated into a semantic desktop framework, social bookmarking tools become the foundation of subject-centric federation of personal knowledge assets with information resources found everywhere on the web.

2. Subject Mapping and Subject-centric Federation

The term "subject map" refers to the name given to implementations of the Topic Maps Reference Model (TMRM)[7]. This paper sets out, by means of an explanatory use case, to describe particular ways in which subject maps can add value to traditional knowledge representation (KR) methodology and applications. When comparing subject mapping to traditional KR methodologies for performing knowledge organization and representation, consider that there are at least two separate dimensions along which discussions occur, and from which the scenario presented here grows. At the most abstract level, we consider two primary dimensions:

- Discussions of *Problem-solving* systems (i.e., question answering)
- Discussions of *Understanding* systems (i.e., federations of disparate world views)

Problem-solving systems are well known among the important work going on in the AI and KR communities. Understanding systems might be thought to be a subclass of problem-solving systems, and they can very well be modeled as such.

An illustrative sketch: if one explores online resources with a personal medical problem in mind, one goal is to find a diagnosis of symptoms entered into an online query form. That would be a problem-solving scenario where one is most interested in ridding oneself of some visitation, really just looking for a diagnosis and prescription, perhaps a second opinion. Another mindset, one frequently the case for students, curious people, and others, is that of understanding the field. A description of symptoms might lead to links to a variety of resources, some of which include research reports, books on medical topics, and other resources aimed at deriving deep understandings of the nature of a visitation. In this scenario, one might be less interested in a single, thought to be accurate answer to a question, and more interested in a range of world views. There is ample reason to believe that both mindsets are important and worthy of continued inquiry[8]. Humans will always need answers and they will need understandings to go with those answers.

Aligned with the *understanding* dimension, our work seeks ways in which heterogeneous world views, as expressed in database schema, ontologies, stories of all kinds, can be *federated*. We label this process *subject-centric federation*. When we use the term *federation*, we distinguish that from *integration* and *fusion*, two well-used terms in the database and ontology fields. We think in terms of *lossless* merging. That is, federation, as we will describe the process here, involves merging information resources without loss of information from either of the merged resources. To anticipate, subject-centric merging only permits merging where the information resources are determined to be representations of the same subject. Thus, subject identity and the practices of subject identification are core considerations in this work.

This paper lays out an illustrative example that supports a particular set of claims. The claims are these:

1. Subject-centric federation of heterogeneous information resources is useful and appropriate for understanding (as compared to specific question answering).
2. Subject-centric federation requires close attention to the details of subject identity and to the rules or axioms appropriate to merging same-subject resources.

[7] TMRM: http://www.isotopicmaps.org/tmrm/

[8] Inquiry: more and more people with serious diagnoses are turning to the web for second opinions and deeper understandings.

3. Subject-centric federation leaves intact the messages or representations available in each federated resource.
4. Subject-centric federation is not a process of deriving semantic equivalence among federated resources; it is possible to federate resources that carry contradictory messages.
5. Subject maps, the paradigm, provide a working framework on which subject-centric federation can be facilitated.
6. Subject maps are an extension of any of a variety of traditional, well-developed KR methods.

Simply stated, subject-centric federation demands that, when two information resources, as, for instance, ontological entities or information resources, can be shown to describe or refer to the same subject, then they must be merged into a common representation container we call a *subject proxy*. A subject proxy serves as a container that marshals all descriptions, all references, and all stories that relate to a particular subject. The process requires that there can exist only one subject proxy in a given subject map for each subject.

3. Social Bookmarking as an Example of Federation

We describe Tagomizer as an illustration of the processes involved in subject-centric federation when social bookmarking occurs in a framework that supports subject-centric knowledge representation and organization.

3.1 Tagomizer

Tagomizer is built as an application of a subject map provider called TopicSpaces. Tagomizer exists as a test-bed for exercises in *learning in the wild* experiments with the CALO project at SRI. Figure 1 shows some of the author's bookmarks at Tagomizer. The program, together with TopicSpaces, is written in Java and will be available to the open source software community.

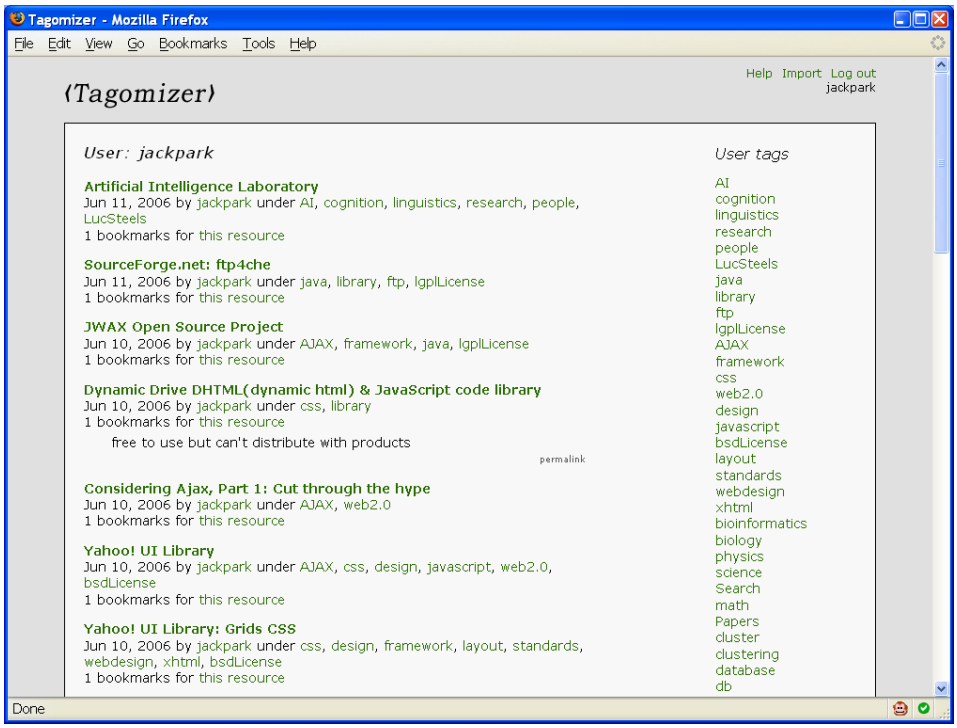

Figure 1. Tagomizer

Tagomizer is an application of TopicSpaces, a subject map provider described below. The application includes a servlet web interface, and an XML-RPC[9] web services interface for communications with CALO.

3.2 Social Bookmarking

Social bookmarking is facilitated through web-based portals that accept user-created *tags* in relation to specific web resources. Web resources are identified by the URL and by a brief descriptor of the resource (web page). Users link tags, typically in the form of single words or multiple-word phrases, to web resources through the medium of a *bookmark*. Thus, the ontology involved in social bookmarking includes the following classes:

- Users
- Web resources
- Tags
- Bookmarks

The relationships entailed by the bookmarking process include

- Users create bookmarks
- Bookmarks are created by users
- Users create tags
- Tags are created by users
- Users select web resources to bookmark
- Web resources are selected by users
- Bookmarks are associated with web resources
- Web resources are associated with bookmarks
- Tags are associated with bookmarks
- Bookmarks are associated with tags

Relationships exist between tags and resources. The classes and relationships (subjects) generated by bookmarking are illustrated in Figure 2 where one particular tag has satisfied the needs of two different users for two different resources.

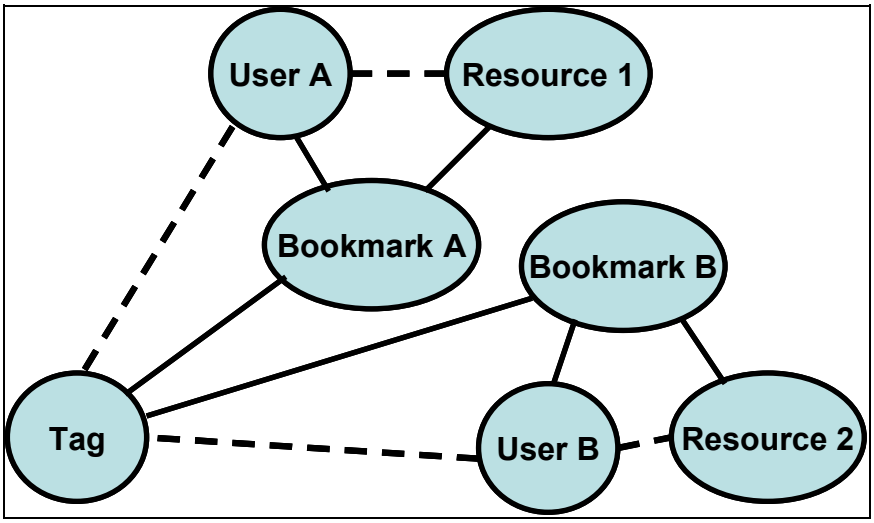

Figure 2. Generalized View of Social Bookmarking

The process of social bookmarking typically follows these sequential events:

1. User visits a webpage
2. User decides to “bookmark” the page

[9] XML-RPC: http://www.xmlrpc.com/

3. User applies a *bookmarklet*[10] to transport the chosen webpage and a brief description of it (the page's *title* and URL) to Tagomizer
4. At Tagomizer, user assigns some *tags* to the bookmark, where tags serve as reminders for later search and retrieval of the bookmarks, and for the formation of community links through those tags. The user can optionally include notes or comments, which can include information copied from the bookmarked webpage.

In this scenario, there are several subjects at play. Those subjects are associated with the following:

- The created bookmark itself, which is identified by the pair: `<user identity, bookmarked resource identity>`
- The bookmarked resource itself, which is identified by its `URL`
- The tags themselves
- The user as a particular individual

To that list of subjects, we must consider other subjects. For instance, the user's intentions, as expressed in the choice of tags assigned to a bookmark, can imply other semantics, other subjects at play. The list of subjects grows.

- The subject(s) *of* the bookmarked resource; the *intentions* of the bookmarked resource are found in the subjects presented there, and are not explicitly captured or represented in the bookmark.
- Implied semantics of the tags. Some tags might be simple random (promiscuous) choices of words or phrases to suit the whims of individual users. Other tags might serve some intended ontology. That is, a user might be working from a particular tag ontology, as might be the case with a community of CALO users.

Given that there are subjects covered in the bookmarked resources that, in some cases, are reflected in potential semantics of tag names used, working with subject-centric representation system offers the advantage of providing a uniform framework into which additional subjects can be harvested and represented. To fill out this picture, consider, for instance, Resource 1 in Figure 2. Let us substitute some "real" values. Resource 1 is actually Matthew West's publications page[11]. Of course, the lone tag in Figure 2 might be named, say "ISO18876". If that tag happens to carry a semantic meaning associated with those publications, then we might expect that Resource 2 just happens to be another resource that also deals with the same standard (Figure 3). Choice of tag names can have significant payoff as taggers evolve *languages*[12] of reminding through semantic ties with the subjects they tag.

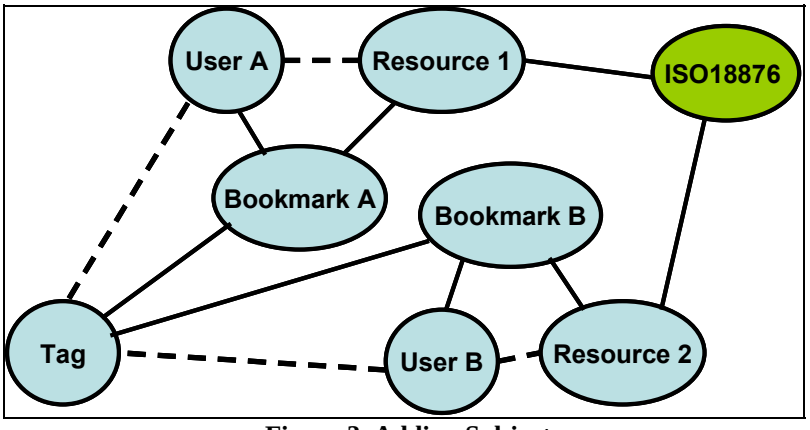

Figure 3. Adding Subjects

[10] bookmarklet: a tiny javascript snippet that behaves as a carrier of bookmark information. Bookmarklets are typically installed at the browser as a button. Click the button when visiting a webpage to start a new bookmark.

[11] Publications: http://www.matthew-west.org.uk/Publications.html

[12] Languages: http://weblog.infoworld.com/udell/gems/delicious.html

As Figure 3 illustrates, the specific subject of "ISO18876" appears in the graph. We will say, for illustration, that this subject was first harvested from Resource 1, then found to be one of possibly several subjects in Resource 2. We surmise, again for purposes of illustration, that User A made the first tie through the tag to the ISO standard, and User B later found a different resource that covers the same subject (the ISO standard) implied by the tag; the tag was reused by User B to remind of the same subject.

By any of several means, it is possible to harvest new subjects from bookmarked resources. Harvesting methods include

- Fully automated harvesting using combinations of indexing, clustering, and text analysis tools
- Human-entered subjects by means of New Subject forms, perhaps along the lines of Wikipedia creation of new chapters
- Combinations of automated harvesting of subjects coupled with human refinement of subject identity properties, that is, collaborative filtering of harvested subjects.

Each new subject added to a growing subject map becomes, itself, a binding point for further harvesting, linking to other bookmarks and to resources not yet bookmarked. In the Tagomizer example above, some level of federation of the world views of users is happening through the semantics of tags chosen. Consider the case where User A notices that User B used the same tag to bookmark a related resource. There is a hint that User B might be thinking along the same or closely related lines of inquiry, so User A decides that it might be useful to explore User B's other bookmarks. This is an instance of *learning in the wild*; in some sense, a *wormhole* has been opened between User A and the universe of thoughts of User B just through that single tag. In some sense, the personal *ontologies* of User A and User B were merged at that tag. User B, of course, enjoys the same opportunity to explore User A's bookmarks. We have more to say about merging and wormholes below

4. TopicSpaces—A Subject Map Provider

TopicSpaces is best described as a *framework* for representation of subjects. The representation scheme is that of properties and values, key-value pairs. As we will show below, those key-value pairs can be considered to form *frame-like* structures [8]. We interpret the TMRM to be a kind of recipe for the architecture, but not the fabrication of subject maps. That is, the TMRM specifies a means by which those ontological commitments necessary to fabricate a subject map are disclosed, but does not specify which commitments to make. TopicSpaces serves as a framework on which many different disclosures can be made. Disclosures are called *legends*. TopicSpaces provides a library of pre-defined subject property classes, and leaves room for the creation of new property classes where necessary.

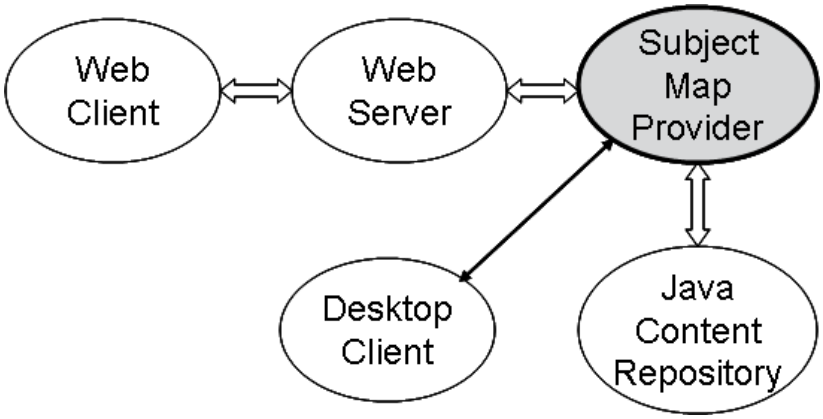

Figure 4. TopicSpaces Architecture

Figure 4 provides an architectural overview of TopicSpaces. TopicSpaces is a subject map provider (SMP) based on the TMRM, with a *core ontology* (see Figure 5) that permits the creation of a range of subject maps, including those that can federate ontologies. The SMP includes an extensible library of *agents*, each of which can be engaged in some specific task. Tasks include subject-centric merging within

the subject map, web harvesting to locate additional information related to specific subjects, and content management.

A core ontology is necessary for TopicSpaces to serve as a framework. The existence of that core ontology does not limit, in any way, the kinds of ontological commitments various legends can disclose; it merely serves as a library of potentially useful property classes, and convenient anchors for legends as they are created. Legends can come in many varieties. For instance, a particular legend can specify how people are to be identified, or how biomedical subjects are to be identified, and how subjects are to be merged using the disclosed properties. Tagomizer provides an instance of a particular legend.

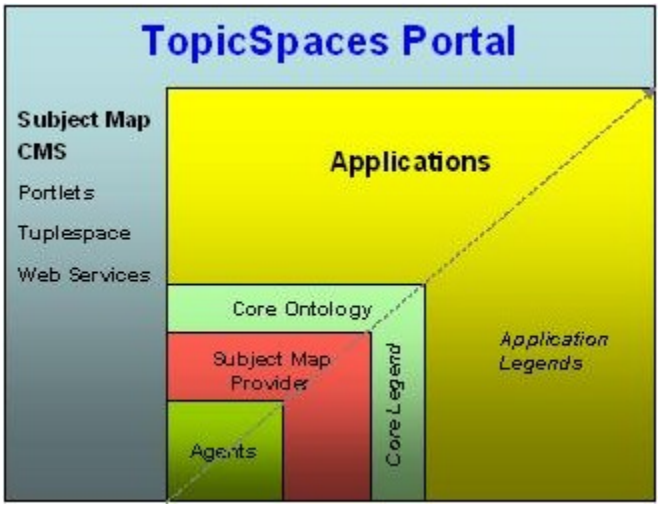

Figure 5. TopicSpaces Portal Architecture

TopicSpaces, itself, is very much a work in progress. The version of the system used for the Tagomizer work is implemented as a relational database with three tables: subjects, properties, and property values. The most recent version is implemented using the Apache Jackrabbit[13] version of the Java Content Repository (JCR)[14]. Jackrabbit provides a graph-based architecture consisting of nodes and properties, coupled with a variety of services including transactions, locking, version control, indexing, and a variety of harvesting tools. In the following section, we sketch how TopicSpaces represents knowledge using Jackrabbit's nodes and properties. We will show that the TMRM allows us to view KR in light of well-understood principles of *frame-based* methodologies. The TMRM dispenses with a predefined commitment to the ontology of XTM[15] or the TMDM[16], and is neutral in terms of the means by which any subject map is constructed, so long as certain *disclosure requirements* are met. XTM and the TMDM remain important to the topic mapping community; the TMRM offers alternative KR opportunities where they are deemed necessary.

4.1 Frame-based KR in TopicSpaces

We do not intend to survey the entire KR field here. Rather, we provide a brief sketch of frame-based KR in order to establish working visual and textual vocabularies with which to describe subject-centric KR and to explore federation of world views. Much of the *frame-based* KR discussion following is inspired by the Open Knowledge Base Connectivity (OKBC) specification [3]. This discussion applies directly to the representation system implemented in TopicSpaces. In this discussion, the terms *subject map* and *knowledge base* are considered interchangeable.

In the most general terms, the objects (entities, concepts,…) of any universe of discourse are represented by statements describing properties (also called *attributes*) of the object together with the values those properties take. For instance, a particular person will have a name, a birth date, and various other properties, each of which has one or more values, depending on the type of the property.

[13] Jackrabbit: http://jackrabbit.apache.org/
[14] JCR: http://jcp.org/en/jsr/detail?id=170
[15] XTM: http://topicmaps.org/xtm/
[16] TMDM: http://www.isotopicmaps.org/sam/

Most KR projects take place within the context of some means by which the artifacts of the project–the representations created by the project–are persisted in some database. For that reason, there is another property of each object in the universe that is the identity of that object *within the database*. We call that property a *locator* and the value of that property must be *unique* at least within the database. We now describe a fragment of a knowledge base that contains entries regarding a particular person, and we will do so in a stylized frame-like representation. We will also represent a fragment of a taxonomy as it might appear in a real knowledge base. We first define three frames, one to establish a descriptor of a class of entities, and one to describe a subclass, and one to describe a particular instance of that subclass. Each frame contains named properties and their values, also called *slots*. Slot values starting with "#" are local references to other frames within the database.

```
locator: 123455
      name:             Class
      psi:              http://www.topicspaces.org/psi#Class
locator: 123456
      name:             Person
      subClassOf:       #123455
locator: 58989859
      name:             Susan Sixpack
      instanceOf:       #123456
      birthdate:        01/21/1988
```

In each of these frames, the properties of the object, *slots*, are noted by being indented from the base frame locator. In some frame notations, it is possible to express meta-properties, or properties of the properties. Typical meta-properties include cardinality of the slot, data types, domain, and range. There have been two distinct ways in which to define these meta-properties. One way is to define each slot with its own frame, as, for instance:

```
locator: 124767
      name:             Slot
locator: 12469
      name:             date
locator: 124768
      name:             birthdate
      subClassOf:       #124767
      valueType:        #12469
```

Another approach has been to apply metadata directly to the slot, called *facets*. Let us illustrate facets by adding facets to the Susan Sixpack frame. First, we define the facets as frames:

```
locator: 8989
      name:             marriedName
locator: 8990
      name:             maidenName
locator: 58989859
      name:             Susan Sixpack
            nameType:   #8989
      name:             Susan Jones
            nameType:   #8990
      instanceOf:       #123456
      birthdate:        01/21/1988
```

In this illustration, we note that this individual has two different names, the name she was given at birth, and the name she chose to take when she married. To establish a working vocabulary for this discussion, we call this particular kind of metadata *scopes*. Thus, we have scoped her two names with the context in which they are valid.

TopicSpaces applies this same representation scheme to the construction of subject maps. What is missing from this example? Simply this: the definitions of some slots themselves, e.g. "name", "instanceOf", and "nameType" are not defined. The TMRM requires that all properties types used in a subject map must be treated as subjects, defined (e.g. with frames as above), and disclosed to those would make use of a subject map that includes those properties.

4.2 Subject-centric KR in TopicSpaces

Subject-centric KR posits that *all* entities existing within a particular universe of discourse are *subjects*. In the social bookmarking use case, the four enumerated classes are subjects, and each of the relationship kinds is also a subject. Each instance of each class is also a subject as is any individual relationship formed between each two class instances. Treating each entity as a *unique* subject requires that we provide for the unambiguous identification of each subject in the universe. Since users will continue to add instances to those subjects through bookmarking processes, subject identity processing takes a central role in maintaining the integrity of the growing knowledge base.

Establishment of subject identity properties is a requirement of the TMRM. In the trivial example above, a legend might specify that, for females, *maidenName* and *birthdate* are sufficient subject identifiers. A merging rule would compare those two values for detection of subject sameness. The same legend would specify that a *psi* (published subject indicator) is sufficient for identity of ontology entities such as `Class`, `psi`, and `birthdate`. In a subject map, all properties (slot types, in this case) are declared as subjects themselves. *Best practices* in subject identity appears to be a large field of inquiry, well beyond the scope of this paper.

5. Subject-centric Federation

Here, we consider federation of formal ontologies, but the discussion applies equally as well to the informal world of Folksonomies and less formal means of expressing world views. Implicit in this discussion is a contrast with *ontology integration* or *semantic integration* in the traditional sense where *semantic equivalence* is a goal. To anticipate, the achievement of semantic equivalence is not a goal of ontology federation as discussed here. Semantic equivalence, as used here, implies finding the same meanings. Consider a trivial example: two agents have analyzed the same scene. One agent determines that the scene describes a non-threatening situation. The other agent determines that the scene describes a serious threat. Subject-centric federation demands that the work product of both agents be merged as descriptions of the same subject, even though each analysis carries a completely different message. Merging for semantic equivalence would call for domain experts to make judgments and weed out the unacceptable analysis; subject-centric merging avails all world views, no matter how contradictory they might be.

Subject-centric federation of world views is defined as the process where heterogeneous world views are brought together "under the same umbrella". Federation involves this bringing together, and *subject-centric* federation demands that, where elements of different world views can be shown to be representations of the same subject, then those elements must be merged into a single knowledge base element that, alone, represents the particular subject, and that contains each of the merged statements in essentially its original *voice*. We use the term *voice* to mean that the re-representation process that occurs during merging does not alter the messages conveyed in the original representations.

In the simplest form of federation of two subjects, each representation (e.g., subject proxy, frame) is known to refer to the same subject, but each gave the subject a different name. Each representation is then merged into a single representation. If, for example, each representation is an ontology class from different ontologies, then each name in the merged representation is *scoped* to reference its source. In that manner, traceability to the original source is always preserved. Through that traceability, we say that *wormholes*[17] emerge. Wormholes are hypothesized to be topological features in space-time that constitute a *shortcut* between regions (Figure 5). We believe that, when different world views are federated on a subject-by-subject basis, different stories captured in same-subject representations offer shortcuts, through scoping, out to other world views.

[17] wormhole: aside from the enormous body of literature, both fictional and nonfiction, I credit first use of the term in relation to source traceability in merged ontologies to Patrick Durusau. In [3], Patrick credits me with first use. It doesn't matter. Neither of us was likely the first user of the term in this context.

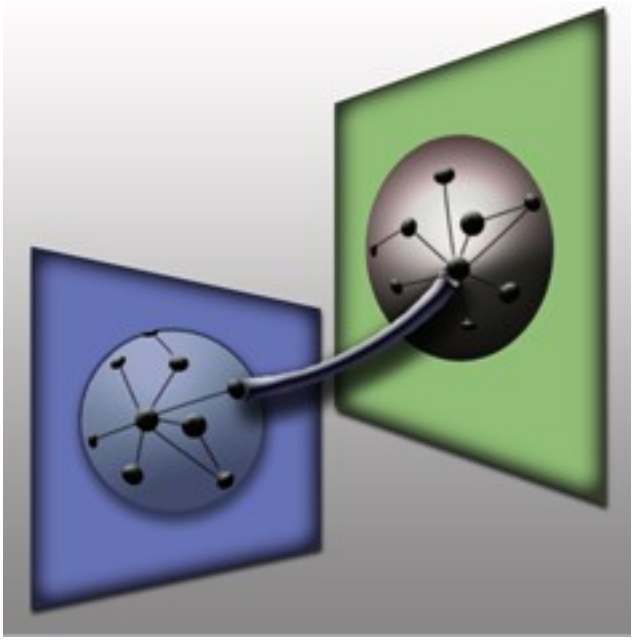

Figure 5. Wormhole Analogy[18]

Another way to look at federation is that of intersecting universes of discourse (Figure 6). The figure depicts scoping links into larger elements of each universe of discourse brought together in the subject proxy.

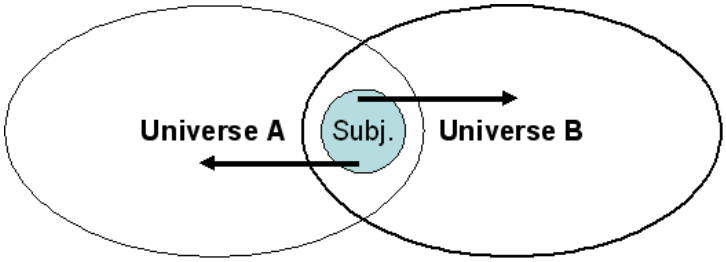

Figure 6. Intersecting Universes

As far as subject maps are concerned, all elements of each universe of discourse are represented with subject proxies (Figure 7). That is, each element in an ontology is, in fact, imported into a subject map, where merging is performed on those subject proxies that are found to represent the same subject. In Figure 7, we see that the nodes labeled "C" and "N" happen to be merged, and wormholes out to nodes "D" and "M" are created.

Following importing several ontologies, it is reasonable to expect that disagreements will be voiced both for lack of merging where there is belief that two subjects are the same but not detected during merging, and where subjects were merged and are believed, by some, to not be representative of the same subject. New information always enters a universe of discourse when disagreements lead to thoughtful discussions. Any web portal assigned to federation of heterogeneous world views should be expected to facilitate thoughtful discussion and even revisions to the subject map where consensus is achieved.

[18] Wormhole: image by Leslie Pound

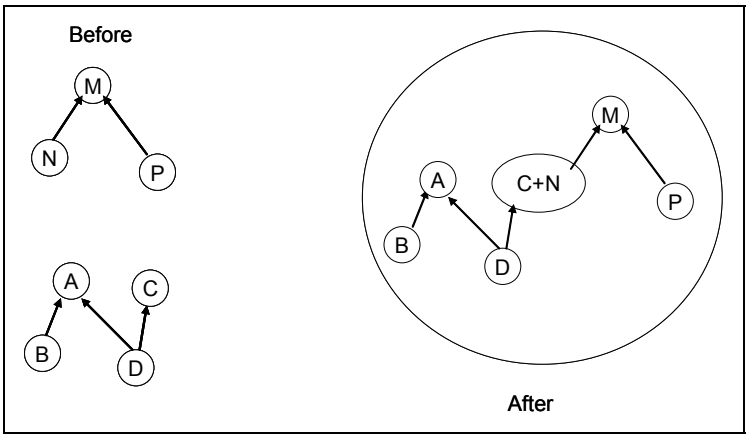

Figure 7. Ontology Federation into a Subject Map

6. Related Work

We limit this section to work related to semantic integration, federation, or fusion as the terms most often used in the literature. Perhaps the largest area of work in AI and KR has been that of the search for semantic equivalence, the quest to merge information resources that carry the same meanings. Subject-centric merging, as stated earlier, uses subject identity equivalence rather than semantic equivalence as the metric for merging decisions. An example of that is GLUE [9], a program that calculates similarity measures among resources in ontologies.

A kind of federation of information resources with topic maps on the web is an ongoing project [10], where multiple topic map portals are interconnected. That work describes a federation of portals, which contrasts to our federation of individual subjects into a single subject map.

Our claim 2 that subject-centric federation calls for close attention to the details of subject identity suggests examination of subject identity practices in traditional KR. We find useful support in [11], "Object Co-identification on the Semantic Web":

> "For any two parties to exchange data about an object, they need to have a **mutually comprehensible reference** for that object. Common names (or URIs) are one kind of mutually comprehensible reference. Extant work on the Semantic web ([]) is based on the assumption that standardized ontologies will provide these common names, so that everyone can use the same name for everything. We believe that this assumption is overly optimistic."

The paper [11] then goes on to describe a solution to the subject identity problem in the form of *Discriminant Descriptions*. A discriminant description is a formula constructed around keys and values. Opportunities for probabilistic matching are discussed.

In the commercial field, we find further interest in attention to subject identity. For instance, from [12]:

> "As we have seen, chunks of information that lack structure or have too rigid a structure tend to live in isolation, unable to discover and bond with others to create new information. What's required is a data-structure that provides unmistakable identity for every chunk of information but also facilitates flow and fusion."

We found evidence of interest in our claim 4 in the literature. From the abstract in [13], "Non-destructive Integration of Form-based Views":

> "The main idea of our approach is to keep the original views intact and to specify constraints between overlapping structures. For reasoning over constraints, we provide a set of inference rules that allows not only to infer implied constraints but also to detect conflicts."

The Harvard Law School has created *H2O Playlist*[19], a subject-centric tagging portal:

> "An H2O Playlist is a shared list of readings and other content about a topic of intellectual interest. It is a simple yet powerful way to group and exchange useful links to information–online and offline."

7. Discussion

> *Taxonomies: Not as boring as you think...*[20]
> *Promiscuity is the practice of making relatively casual and indiscriminate choices.*[21]

We first introduced the notion of federation to semantic desktops in [4], where we described a mismatch that comes into existence between users of semantic desktop workstations that are built on top of ontologies crafted by others. We labeled that situation *Just For Me*, with the explanation that workstations are crafted to support the productivity needs of individuals who already have deep-seated habits and ways of knowing. Those ways of knowing are often quite different from those of the ontologists and workstation designers. We showed how a subject map provider could serve as a kind of *interlingua* where the needs of the ontologists, maintenance of semantic interoperability between workstations, are federated with those of the users, creation of work products which include meeting plans, documents, content archival and retrieval, and more.

Just as the internet and web are founded on open source and creative entrepreneurship, Web 2.0 is bringing out even larger elements of creativity. With new user interface tools such as AJAX[22] to turn web browsers into powerful knowledge workstations, new ways are emerging to represent, manage, and present artifacts of human knowledge. Richard Dawkins first coined the word *meme* in his book *The Selfish Gene* [5] and later *used the term to refer to any cultural entity (such as a song, an idea or a religion) that an observer might consider a replicator*[23]. People continue to expand on application of that word and its extensions. For instance, *memography* is the subject of a wiki[24] where they say

> "Memography is our name for a simple three-step technique for tagging web pages that allows you to find them again with high precision and recall by leveraging the incredible full-text search databases of today's search engines - what Memography calls a memetic search."

At memography.org, as at many other online locations, people are talking about tagging as a means of forming associations with different subjects. Why tagging? Some might say tagging is messy. Here is what David Weinberger has to say about *messy* at his *Journal of the Hyperlinked Organization*[25] (JOHO) talking about his forthcoming book *Everything is Miscellaneous*:

> "As discussed in previous issues of Joho, the book pretends there are three orders of order. In the first, we organize the objects themselves. In the second, we separate the metadata from the data and organize the metadata (e.g., a card catalog). In the third order, the data and metadata are both digital, so we can come with new ways of organizing them free of the constraints of the physical. Chapter 8 begins by saying that messes in the first two orders are inefficient and make life worse, but in the third order, a messy pile with lots of implicit and potential relationships within it actually reverses entropy. The pile itself can stay messy as different people organize the metadata as they want. For example, if our family photos are in a messy pile, we can't find anything easily. If my wife wants to organize them by year and I want to organize them by person, one of us has to lose. But, we can each organize our digital pile of digital photos the way we want without actually

[19] H2O Playlist: http://h2obeta.law.harvard.edu/home.do
[20] Taxonomies quote: author unknown – http://finance.groups.yahoo.com/group/TaxoCoP/
[21] Promiscuity: restricted definition found at http://en.wikipedia.org/wiki/Promiscuity
[22] AJAX: http://en.wikipedia.org/wiki/AJAX
[23] Quoted from http://en.wikipedia.org/wiki/Meme
[24] Memography: http://www.memography.org/index.php/Main_Page
[25] JOHO quote: http://www.hyperorg.com/backissues/joho-jul23-06.html#book

rearranging the digital photos at all. The more metadata attached to the photos and the more relationships discerned among them, the more potential for fruitful ways of organizing it."

Perhaps, the more tags placed in a public digital universe the better. From recent discussions about promiscuity, some theories suggest that promiscuous sexual behaviors are good for the gene pool. Could we regard such theories as metaphors, mapping them into the world of memetics and conjecture that promiscuous tagging might prove valuable to the meme pool? Given the suggested emergence of wormholes between and among a diverse universe of taggers, it is reasonable to conclude that tags don't even have to be thoughtful or "correct" in any semantic sense. They simply need to be *seductive* in the sense that they attract a diverse audience to wonder what might be behind their existence, and follow the trails created by the tagging framework.

What, then, is the relationship between semantic desktop platforms and a messy universe of tagged information resources? There are many dimensions along which answers can be found for that question. Let us restrict this discussion to one particular dimension. That dimension is related to David Weinberger's comment: *But, we can each organize our digital pile of digital photos the way we want without actually rearranging the digital photos at all.* If we take a semantic desktop platform as a personal database qua workstation, we see immediately that the marriage of semantic desktops with Web 2.0-enabled subject map portals provides the necessary ingredients for individuals to manage federation of information resources in any way they choose without affecting the organizational behaviors of others.

Semantic desktop platforms built around subject map providers enable a kind of knowledge representation and knowledge organization suited to personal arrangement of views, including taxonomic views. Consider Tagomizer. Right now, as a web portal, Tagomizer, like "delicious" and some other social bookmarking websites, provide access to one particular kind of view. That view is known as a "pivot" view —also referred to as *pivot browsing* [6], where one is able to pivot around any of three dimensions, each centered on the bookmark object itself. Those dimensions are *users*, *tags* and *bookmarked (tagged) information resources*. Not available are other dimensions associated with each of those dimensions. That is to say, there are perhaps an infinite number of dimensions along which social bookmarking portals *could* provide views, but those views are not available. Consider users, those who do the tagging. What else can we know about them? At Tagomizer, we only know about their bookmarked resources and tags. Consider the bookmarked information resources, typically restricted to web pages. What else can we know about those resources? Indeed, what subjects are discussed at those web pages? What can we know about those subjects?

Since Tagomizer is built as an application using a subject map provider the program can and will eventually support more kinds of views. Bringing Tagomizer to that capability is the subject of present and planned work. Mirroring that capability in IRIS is also the subject of present and planned work. IRIS will become one of the first open source semantic desktop platforms to use a subject map provider as its core knowledge base. That knowledge base will remain coupled to the CALO ontology to maintain semantic interoperability among CALO platforms.

The value proposition behind CALO, largely available in CALO's semantic desktop workstation IRIS, is based on the notion that a framework is available on which individuals can organize their personal knowledge assets and bring them to bear on their day to day activities. A value proposition is only that, a proposition. It is up to users to utilize those facilities on which the value proposition is based. As we said here and elsewhere [2],[4], the many different ways in which knowledge is represented and organized can be federated under a *subject-centric* framework, one that facilitates individualization of a user's workspace without affecting the views of others.

Subject-centric federation, as a technology, is difficult to discern as different from integration processes already in practice in the KR communities. The process described here is different more by degree: while the few cases we find in the literature that acknowledge subject identity to be important to resource merging, none pay attention to the nature of subject identity to the degree that is supported by the TMRM. We do not claim that *subject-centric* requires the facilities of a subject map provider; indeed, many of the present and historical means of knowledge organization and representation can be adapted to a subject-centric framework through enriched representations of subject identity. We do claim that subject-centric representation and organization facilitates federation of heterogeneous information resources, the emergence of opportunities for chance discovery [7], and the potential for increased productivity.

8. Future Work

We are presently using Tagomizer as a stand-alone application on the web. Tagomizer is providing CALO developers with an opportunity to collaboratively tag websites related to various projects in which they are engaged. Some of those projects include the development of CALO itself. As a machine learning platform, CALO uses a web-services interface to query Tagomizer to collect and marshal new information for the user's benefit.

In future work, we plan to integrate TopicSpaces directly into the IRIS platform as a wrapper for the CALO ontology. As an ontology wrapper, TopicSpaces will serve our *Just For Me* objective by providing for federation of users' world views with those expressed in the ontology. At the same time, the Tagomizer application installed in IRIS will allow users to tag the various information assets IRIS supports.

In the case where IRIS interacts with web portals of various kinds, IRIS will be able to share some of the Tagomizer bookmarks created locally with public portals. Public websites bookmarked locally are a likely candidate, but other resources could be shared depending on community needs.

9. Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the Department of Interior-National Business Center (DOI-NBC).

This work has benefited greatly from interactions with Patrick Durusau, Steve Newcomb, Vinay Chaudhri, Adam Cheyer, Simon Buckingham Shum, Douglas Bowden, Glen Haydon, Richard Fikes, Joshua Levy, Eric Yeh, Annette Kramer, and Mark Szpakowski. Joshua Levy collaborated on the design and fabrication of Tagomizer. We thank the anonymous reviewers for valuable comments.

References

[1] Lassila, Ora, "Introduction to RDF Metadata", W3C NOTE 1997-11-13, http://www.w3.org/TR/NOTE-rdf-simple-intro-971113.html

[2] Park, Jack, and Patrick Durusau, "Avoiding Hobson's Choice in Choosing an Ontology", teleconference to the Ontolog Community, 27 April 2006; http://ontolog.cim3.net/cgi-bin/wiki.pl?ConferenceCall_2006_04_27

[3] Chaudhri, Vinay K., Adam Farquhar, Richard Fikes, Peter D. Karp, and James P. Rice, "Open Knowledge Base Connectivity 2.0.31", 9 April, 1998; http://www.ai.sri.com/~okbc/spec/okbc2/okbc2.html

[4] Park, Jack, and Adam Cheyer, "Just for Me: Topic Maps and Ontologies", in Lutz Maicher and Jack Park (Editors), *Charting the Topic Maps Research and Applications Landscape: First International Workshop on Topic Map Research and Applications, TMRA 2005*, Leipzig, Germany, 6-7 October 2005, Revised Selected Papers, Springer LNCS Volume 3873/2006 pp. 145-159.

[5] Dawkins, Richard, *The Selfish Gene*, Oxford University Press, 1976.

[6] Millen, David , Jonathan Feinberg, and Bernard Kerr, "Social Bookmarking in the Enterprise", Social Computing Vol. 3, No. 9 - November 2005; http://acmqueue.com/modules.php?name=Content&pa=showpage&pid=344&page=1

[7] Ohsawa, Yukio, and Peter McBurney (Eds), *Chance Discovery*, Springer-Verlag, 2002.

[8] Minsky, Marvin, "A Framework for Representing Knowledge", in P. Winston (Ed), *The Psychology of Computer Vision*, McGraw-Hill, 1975.

[9] Daon, AnHai, Jayant Madhavan, Pedro Domingos, and Alon Halevy, "Ontology matching: A Machine Learning Approach", in S. Staab and R. Studer [Eds}, *Handbook On Ontologies*, Springer-Verlag, 2004, pp. 385-403.

[10] Pepper, Steve, and Lars Marshal Garshol, "Seamless Knowledge: Spontaneous Knowledge Federation using TMRAP",, Extreme Markup Languages Conference, 2004, Montreal, Canada. Slides on the web at http://www.ontopia.net/topicmaps/materials/Seamless%20Knowledge%20with%20TMRAP.ppt

[11] Guha, R., "Object Co-identification on the Semantic Web", http://tap.stanford.edu/CoIdent.pdf

[12] Harbor Research Inc. "Designing the Future of Information: The Internet Beyond the Web", A Harbor Whitepaper, September 2005; available at http://www.maya.com/web/infocommons/infocommons.mtml

[13] Hidders, Jan, Jan Paredaens, Philippe Thiran, Geert-Jan Houben, and Kees van Hee, "Non-destructive Integration of Form-based Views" in 9th East European Conference on Advances in Databases and Information Systems (ADBIS 2005); nr. 3631, pages 74–86, Springer.