Enterprise Situational Awareness at MITRE: Defining the Next Generation Internet

Suzette Stoutenburg, Amy Kazura, Richard Panek, Jason Peterson, Karen Fox
The MITRE Corporation
1155 Academy Park Loop, Colorado Springs, Colorado 80922
{suzette, alk, rpanek, jasonp, kfox}@mitre.org

Abstract. Situational Awareness is a critical mission need for MITRE's Department of Defense (DoD) sponsors. Solutions for Situational Awareness need to be interoperable, dynamically adaptable, and extensible in a Network Centric environment. At the same time, there is a growing need within MITRE to better maneuver through the ever increasing amount of information on our intranet and the World Wide Web (www), necessitating solutions for enhanced integration of information sources across the Enterprise. We believe that the key to solving both challenges is to apply semantic technology, since these emerging solutions transcend language and specific implementations and offer integration at the semantic level. Further, this technology is inherently extensible, and builds the foundation for dynamic integration and a fully automated Machine-to-Machine (M2M) enabled enterprise. To demonstrate the potential, we initiated an effort to apply semantic technology to achieve Enterprise Situational Awareness at MITRE. We built an abstraction layer over the existing MITRE Information Infrastructure (MII) to exploit existing information sources and link this information together in a meaningful way. A number of use cases were identified for the effort, entitled MITRE Enterprise Situational Awareness (MESA). With our Center for Information and Technology (CI&T) partners, we selected a semantic "Dynamic Topic Watch" application as our demonstration. This application dynamically provides news, events and other information surrounding a user's specified topics of interest, including technical events, Community of Interest (CoI) news, research results, vendor announcements, publications and more. This was achieved by reusing and tailoring existing ontologies to represent enterprise objects such as people, events, and projects. We built an ontology based on the existing MITRE subject Taxonomy (MST) and automated the process of linking enterprise objects to the concepts that describe them. We linked existing data sources on the MII and Web with these ontologies to support query and subscription to the semantic Topic Watch service. The prototype capability has been deployed on the MII. This paper describes the solution we built, and summarizes our findings and recommendations.

1. Introduction

1.1. Motivation

The International Semantic Web Conference (ISWC)[1] is a growing international forum started in 2002, in which research in all aspects of the Semantic Web are exchanged between academia, industry and government. This conference has had significant sponsorship and participation of key leaders of the successful World Wide Web Consortium (W3C). Each year, the ISWC conducts a Semantic Web Challenge[2], the purpose of which is to encourage innovative applications of semantic technology, support deployment of useful semantic web applications and advance current research and standards. Former winners are trailblazers in technology innovation.

The lure of winning the Semantic Web Challenge was the inspiration for this project. We envisioned MITRE as an entrant in this prestigious competition by developing a semantic application on the MITRE Information Infrastructure (MII). As we explored the idea further, we found that the effort could address other key goals, including demonstrating how semantic technology could be used to integrate heterogeneous data sources, particularly for critical mission areas such as Situational Awareness. We also saw the opportunity to "eat our own dog food" by trying this technology on ourselves, the MII, and extrapolating our findings to mission applications. Finally, we saw this as an opportunity to offer a rich new capability to MITRE employees.

[1] http://iswc2005.semanticweb.org/
[2] http://challenge.semanticweb.org/

Our plan is to submit the MITRE Enterprise Situational Awareness (MESA) application to the 2006 Semantic Web challenge.

1.2. Purpose and Goals

Situational Awareness is a critical mission need for the Department of Defense (DoD). Solutions for enhanced situational awareness need to be extensible, dynamically adaptable, machine understandable and interoperable in a Network Centric environment. At the same time, there is a growing need to enhance the integration of disparate information across organizations, the enterprise and external information sources. An emerging solution to interoperability involves expressing information through semantic concepts and building software that can query at the concept level. This means that as new information sources are identified, software does not need to change; new sources are mapped to the concepts in the semantic knowledge base, thus offering inherent extensibility. Semantic solutions also transcend language and specific implementations, and enable a foundation for a fully automated (M2M enabled) enterprise.

To demonstrate the potential of semantic technology solutions, we initiated an effort to apply semantic technology to achieve Enterprise Situational Awareness at MITRE. Our specific goals for the project were as follows.

- Demonstrate how semantic technology could be applied to achieve Enterprise Situational Awareness at MITRE.
- Exploit existing MII information sources, including information about Enterprise Objects such as people, events, projects, etc. link these objects using ontologies, and automatically associate meta data to them.
- Provide users the capability to discover events of interest to enhance employee productivity and collaboration, through a query and subscription service.
- Determine how best to apply techniques and lessons learned to critical sponsor missions.
- Enter the 2006 ISWC Semantic Web Challenge.

2. Use Cases

As we developed the idea for Enterprise Situational Awareness at MITRE, we considered a number of use cases to demonstrate the concept. We identified quite a few applications in which integration of heterogeneous data on the MII would prove useful. The purpose of this section is to capture the complete set of use cases identified for future consideration, as well as describe the use case we selected and why.

2.1. Use Cases Identified

Below is a list of the primary use cases identified during the development of the MESA concept.

1. Dynamic Topic Watch – capability in which employees can query or subscribe to news and events of interest surrounding selected technical topics. Information sources across the MII and World Wide Web would be integrated semantically, and audience targeting and calendar integration could be applied. (Note that "audience" can refer to an employee, project, organization, community of interest or any entity.)
2. Semantic calendaring and event tracking – capability in which events across MITRE could be linked to employee, project and organization calendars. This could also be targeted to audience interest.
3. Automated Dynamic Knowledge Zones – this would involve automating the current MITRE Knowledge Zone capability. This capability provides topic related summaries of news, vendor announcements in advanced technology and more[3]. Traditional Knowledge Zones are currently manually assembled every 6 months; automation would result in less manual labor, more timely updates, reduced errors, reduced cost, and semantically rich links across topics and meta data.
4. Audience Targeting – apply semantic technology to audience targeting and social networking analysis; for example, a capability that could provide answers to "Who else is like me?"; that is,

[3] http://communityshare.mitre.org/sites/knowledgezones/default.aspx

"who else at MITRE is interested in the same things as me, who else has similar experience, responsibility, etc." ?

5. Semantic Event Correlation and Event Detection Management – This would be a capability in which we could link events to the interests of the audience, such as the following:
    - Conferences
    - MITRE Institute courses
    - Professional external courses
    - University courses and seminars
    - New book publishing
    - MITRE employee evaluation of classes and tools
    - W3C and other standards organization events

We also identified a number of other ideas during brainstorming sessions. These ideas are less well developed, but preserved here.

1. Identification of Enterprise Licenses across MITRE – what tools have we already purchased at MITRE? Are there licenses available for use?
2. Automatic generation of reports on technical stature (i.e., conferences attended, papers published, etc.)
3. Automated reports on research developments in particular areas of interest
4. Automated dashboard on new lab capabilities and connections across MITRE and sponsors
5. Summary of conference participation and evaluation by MITRE employees, past and future
6. Association of new contacts to events and conferences
7. Enhanced application of lessons learned and reuse across the enterprise
8. Identification of new tools used at MITRE
9. Vendor announcements of interest
10. Tracking of visits by VIPs and sponsors for better coordination
11. Reports on new prototypes and capabilities delivered across MITRE
12. Summary of key project findings
13. Identification of skill growth – how do we detect skills gaps?
14. Semantic enablement of mail lists
15. Indexing and annotation of email
16. Automatic generation of status reports and employee performance reviews
17. Support for new employees, people new to projects, new to a customer
18. Application to Know Your Customers, an effort to better our sponsors' critical mission challenges

2.2. Selected Use Case: Dynamic Topic Watch

We collaborated with CI&T to select the best use case to prototype. We considered the current strategic goals and work program within CI&T so that the application to prototype would contribute to that strategy. We also sought an application that would provide the most benefit to MITRE employees, with the least amount of obstacles (such as privacy with annotating email and calendars). Finally, we wanted to select an application that could make use of existing information on the MII to the extent possible.

Together, we selected Dynamic Topic Watch as the application that we would prototype. We envisioned a capability in which employees can query or subscribe to news and events of interest surrounding selected technical *concepts* of interest. We envisioned that concepts of interest could be based on a user profile, derived from existing information about the employee, such as project and organization membership, Sharepoint membership, email list membership, etc. We planned to integrate heterogeneous data sources across the MII and the World Wide Web by linking information based on those concepts of interest to each user.

2.3. How is this different from Google?

The MESA prototype employs semantic query and should not be considered to be a replacement for Google as the two are fundamentally different, though complementary. Google is a search function which works on large volumes of natural language text. (Text is still the most plentiful information type on the Web despite recent increases in multimedia content.) No special preparation, other than publishing the content in a Web-accessible location, is needed for content to be usable. Nor do users require any special knowledge or education in order to use Google (though they do need to guess what terms would appear in the target document). Google's ease of use for both publishers and consumers of information is one of the major reasons for its tremendous success and popularity.

In contrast, taxonomic query approaches are typically used for browsing purposes, as opposed to searches for a specific term. The ease with which users can find the desired content depends on their a priori knowledge of the classification scheme used. The producers (or third-party cataloguers) as well as the users of information have to guess and remember how the other one thinks and what vocabulary they use, and to use that vocabulary consistently. For example, when using the MITRE Subject Taxonomy, it might not be clear whether Semantic Web-related documents are found under "Computing Methodologies" or under “Information Management”. This means taxonomic solutions require more upfront effort to 1) develop a structured classification scheme, and 2) apply it consistently to the resources in the system. It also means users need *a priori* knowledge, namely knowledge of the classification scheme and how it is applied, in order to use the system. This is one of the reasons the Semantic Web has been mired in a "chicken and egg" problem for a while with respect to RDF/OWL content; information producers do not see the benefit of additional upfront effort to develop RDF and OWL since no one is using the Semantic Web yet, and information consumers do not seek to use the Semantic Web since there is little or no RDF/OWL content there.

If humans had a culturally-independent world-view and were good at applying classifications consistently, Google never would have happened and we'd all be very happy finding things in the Web via Yahoo categories. On the other hand, computers are not good at dealing with open vocabularies, so having a structured set of terms facilitates automated processing. The MESA prototype is investigating how to combine the best features of each of these approaches in order to provide better information services for the users. To this end we have created concept schemes that are organized as ontologies, and supplemented the elements of the ontologies with sets of natural language labels that may be used to refer to each element, rather than requiring universal knowledge of each element’s unique identifier.

So, in summary, the semantic query approach used in MESA avoids the pitfalls of either a purely text-based or purely taxonomic approach by searching on terms based on their *meaning*, not by keyword matching. This approach leads to more accurate, semantically rich results. Also, the foundation is in place for knowledge discovery through the power of standard semantic languages.

Another key point is that semantic technology is *inherently extensible.* New data sources can be linked in to existing ontologies by mapping concepts through properties. This approach therefore transcends dialects of organizations, joint forces, coalition partners, etc., and supports M2M integration. Therefore, Google indexes and search results would serve as another information source in a semantic query.

3. Prototype Design

The Dynamic Topic Watch application was designed to be n-tier, providing Web based Graphical User Interface (GUI), Service, Database and Semantic layers, as shown in Figure 1 below. Data sources are linked as instances through the Semantic Layer.

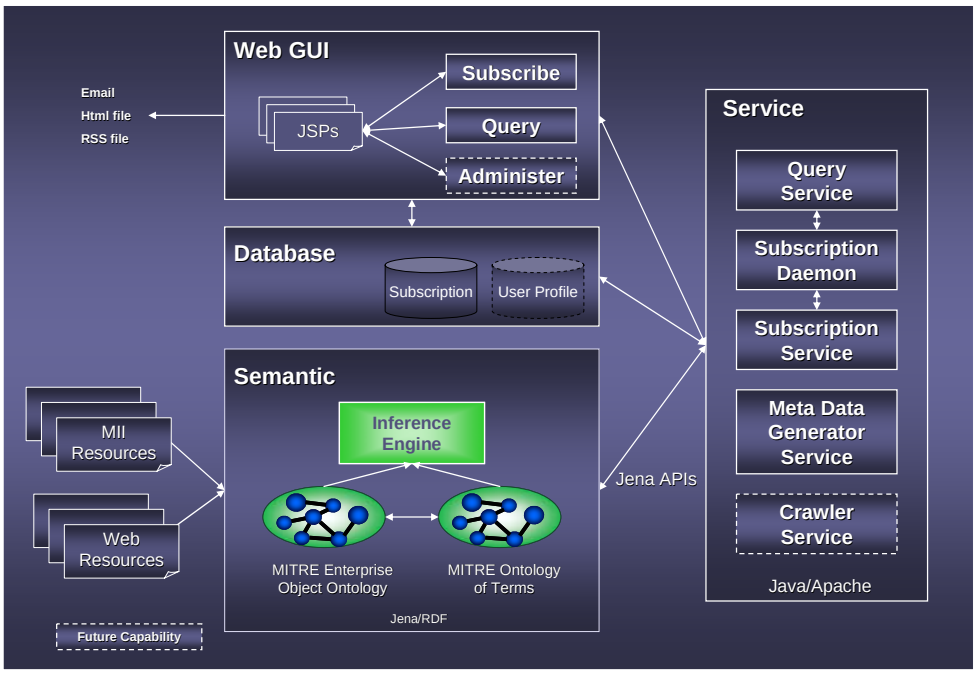

Figure 1. Design of Dynamic Topic Watch Prototype

The Web layer controls the flow of the application and the management of session objects. Users can query or subscribe to reports, and there is also an option to administer the application which is reserved for administrators. The Web layer does require MITRE employee identification, supporting retrieval of user information that is used throughout the application.

While the GUI layer invokes the services to access the data sources, the service layer is built to accommodate calls from any user, and in fact, some of the services invoke one another. There is a Query Service, Subscription Service, Subscription Daemon and Meta Data Generator service. The need for a Crawler Service was identified late in the project, and is currently under construction. (More details on this issue can be found in section 4.) The Database layer supports the Service layer, primarily by providing subscription information. In the future, we anticipate adding a user profile database.

The Semantic layer consists of the ontologies built to support Dynamic Topic Watch, as well as the inference engine that operates over those ontologies. This layer also provides the link to web resources on the MII and World Wide Web. Further details of the design are described below.

3.1 Data Sources

We had anticipated that information on the MII would be available in databases, streaming XML, or perhaps HTML with standard meta data tags. Instead, we quickly determined that no databases were available. Most of the sources were encoded with HTML and employed various approaches to embedding information. We also found that meta data is not applied consistently in all sources; in some cases, meta data is tagged manually. This is also an issue on most World Wide Web information sources.

In the absence of consistently structured data sources, we determined that construction of an instance layer would be necessary. The purpose of this layer would be to link information source instances to concepts in the MITRE Enterprise Ontology. To automate the creation of instances, we identified the need for a web crawler. This requirement was identified late in the project, and is currently in progress. Note that we considered using the MITRE Google engine as a source of instances, but found that the licensed APIs to Google are limited, and would not meet our needs. Late in the

project, we identified a set of alternative Google APIs, and are currently in the process of researching these further as a source of information. To keep the project on schedule, and to focus on the most difficult task of building the essential infrastructure, we decided to manually build a set of instances for demonstration purposes. We used a variety of MITRE internal sources as well as external Internet sources to create instance data, as described below.

Many of our internal sources closely followed those of the MITRE Knowledge Zones[4], including the following.

- Community Share sites
- Hot Topics
- MITRE Technology Program Project Sites
- MITRE Technical Reports
- Technical Exchange Meetings

We randomly selected a number of external data sources which are listed below. The crawler combined with the meta data generator will automate this task in the future.

- New Books Published
- Upcoming Conferences
- Conference Proceedings
- Journal Publishing
- Technical Papers
- E-zine Article
- Factiva News Items
- Government News Item
- MindSwap News Items (from the University of Maryland)
- W3C News Items and Resources

3.2 GUI Design

The basic concept of the GUI design was to keep it as simple and easy to understand as possible. Most pages include Java Server Page (JSP) code that made calls to the Java servlets operating behind the scene to do the querying, handling of subscriptions and semantic linking. The following sections describe the flow of the application, and include detailed descriptions of the Query, Subscription and filtering capabilities of the application.

3.2.1 Login

The main login page requires an employee number in order to identify the user using MITRE's Lightweight Directory Access Protocol (LDAP), enabling the tool to automatically fill in information for subscriptions. In future releases, user information will be used to tailor the results of queries.

Figure 2. Dynamic Topic Watch Login Page

[4] http://web1a.mitre.org/infolink/zones_cms/

3.2.2 Query

After login, the user is taken to the main query page. This page is pre-populated with a selection of all types of information (e.g., Hot Topics, Factiva). The user can select only those types of interest, or confirm a search on all types. One or more topics of interest are then entered by the user, and the form is submitted.

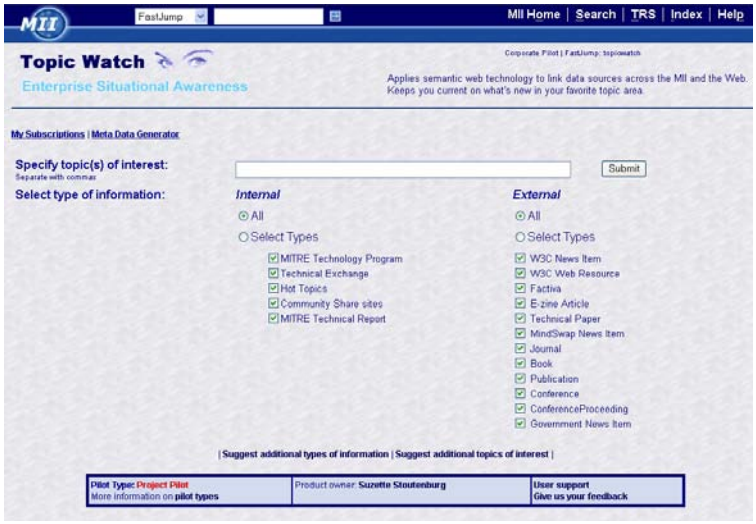

Figure 2. Query Form

Submission of the query form results in a query to the ontologies. Assume that the user selected Conferences and Hot Topics as the information sources, and "Semantic Web" as the concept to search. Essentially this query would be translated to, "Provide information on Conferences and Hot Topics related to *Semantic Web* or any concept related to *Semantic Web*." The query is served by the Jena RDF engine, and the application displays instances that meet the constraints of the query, as shown in Figure 3 below. Note that the query encompasses multiple ontologies; the MITRE Enterprise Ontology semantically describes Conferences and Hot Topics and the MITRE Ontology of Terms describes the "Semantic Web" and related concepts. The links across ontologies are described in section 3.3.

Figure 3. Query Results

In the example in Figure 3 above, the term *Semantic Web* was identified to be related to other concepts such as meta data, semantic network, folksonomy, etc. using the ontology, that is, the knowledge base. This is one of the biggest strengths of a semantic technology approach.

Note that the user can filter on the terms if too much information is returned. Once the user selects only those desired concepts, the form can be resubmitted.

Figure 4. Filter Terms

3.2.3 Subscription

The user can subscribe to information through the subscription option or save a query as a subscription by clicking on the Subscribe to Query link from the Query page. Each approach takes the user to the subscription page with fields pre-populated to match the query. The user must select the frequency of report and the method of delivery, then submit the form. The subscription page is shown in Figure 5 below.

Figure 5. Subscribe to Query

The user can view his current subscriptions at any time by clicking on the **My Subscriptions** link. This produces a list of subscriptions, which can then be edited or deleted. Users should create a new subscription if topics are to be modified.

Reports are automatically generated for each subscription at the selected timeframe with the latest information corresponding to the selected topics of interest. HTML and RSS subscriptions have links produced whereas email subscriptions provide reports directly through email.

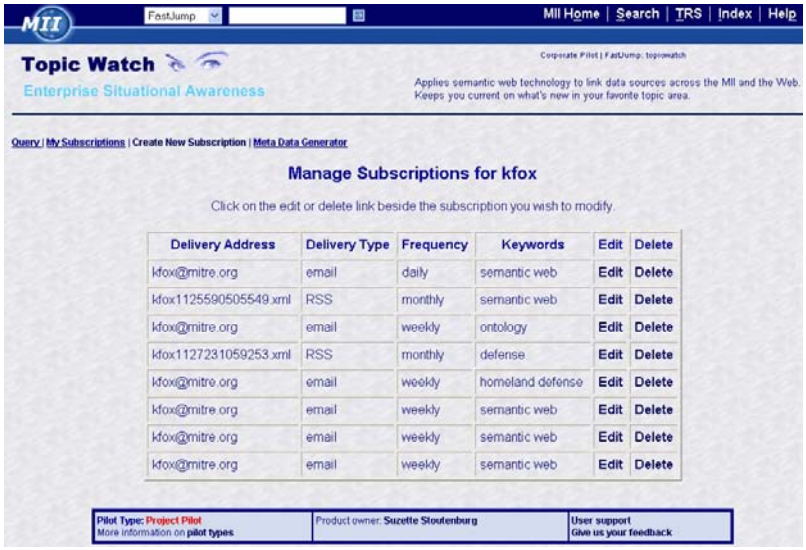

Figure 6. My Subscriptions

3.3 Ontology Design

Our primary objective with respect to the ontology design was to reuse existing ontologies since a primary goal of the Semantic Web is to facilitate reuse of modular, well-defined data. We identified several existing ontologies which describe some of the concepts that are part of our Enterprise Objects collection. We decided to reuse the friend-of-a-friend (FOAF) ontology for describing people and organizations, and the Simple Knowledge Ontology System (SKOS) to describe meta data. We developed the MITRE Enterprise Ontology (MEO) to describe Enterprise Objects and the MITRE Ontology of Terms (MOT) to describe meta data. These two main ontologies were linked with the skos:related property, as shown in Figure 7 below.

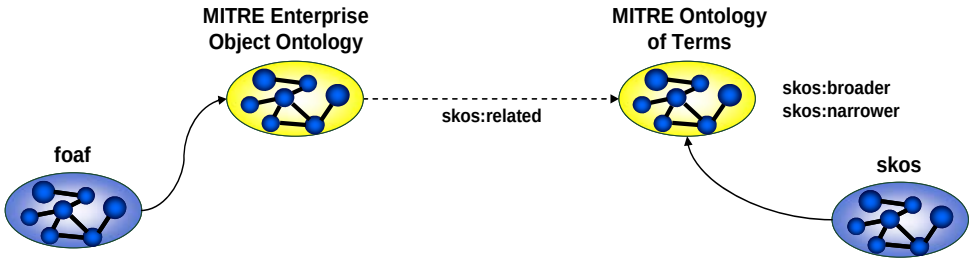

Figure 7. High Level Ontology Design

For example, suppose the user queries for "Conferences related to Homeland Security". The ontology design supports this query by first identifying all terms in the MOT that are related to "Homeland Security". Then, all instances of the class Conference (and all its subclasses) are then retrieved using the property skos:related. This example is depicted in Figure 8 below.

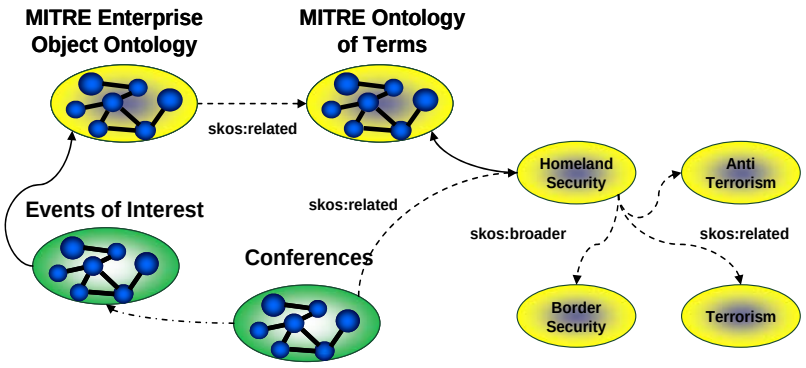

Figure 8. Example Query

The FOAF ontology[5] was chosen for describing persons, and associated projects and organizations. The MESA extensions to FOAF are shown in Figure 9. The Description of a Project (DOAP)[6] ontology was also identified for use in providing additional detail about projects, though we are not currently using this information.

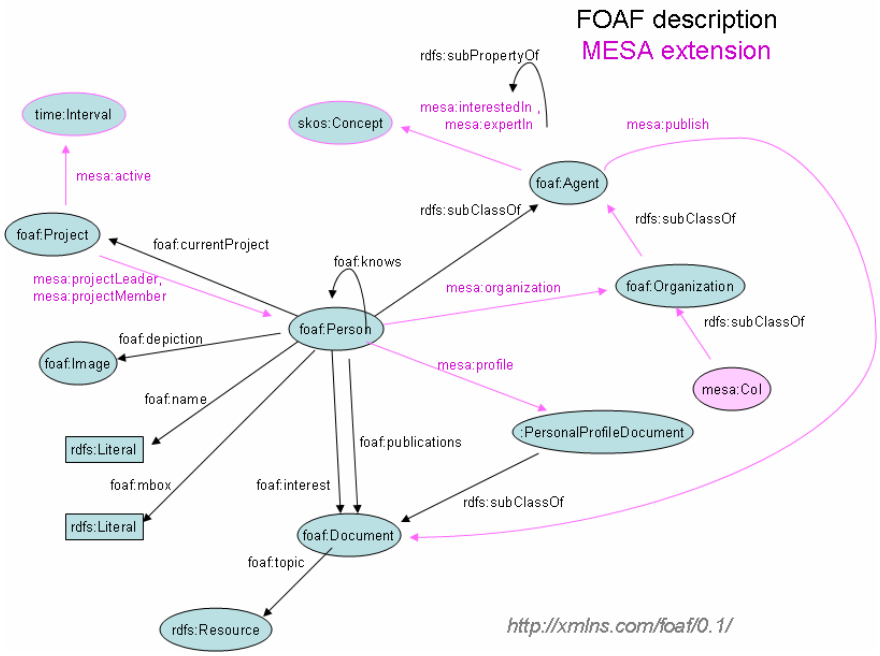

Figure 9. MESA Extends FOAF Ontology

We use the popular Dublin Core[7] document description vocabulary for providing information about the document and web resource instances the MESA prototype uses as information sources. For the data about the queries and subscriptions themselves, we were unable to find any existing ontologies, so we created this part of the MESA ontology. Figure 10 shows the MESA extensions to Dublin Core.

[5] http://xmlns.com/foaf/0.1/
[6] http://usefulinc.com/ns/doap#
[7] http://dublincore.org

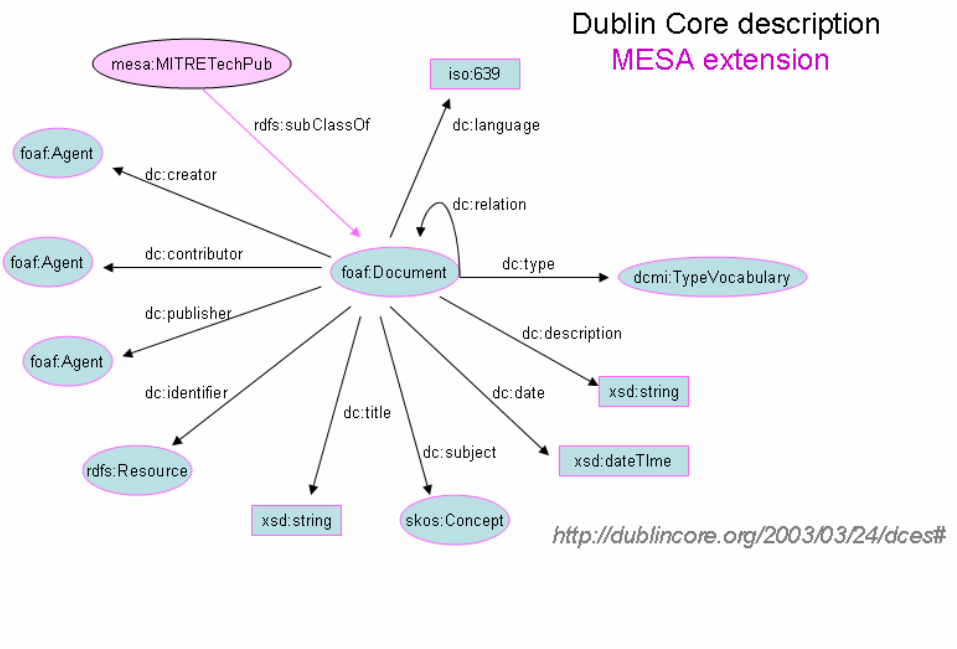

Figure 10. MESA Uses Dublin Core Concepts

In addition to creating a MITRE Enterprise Ontology which could capture information currently contained in the various databases backing the MII, the other main focus of our ontology effort was to create a concept scheme which would be used to tag the information sources in the MESA system. This concept scheme, the MITRE Ontology of Terms (MOT), models the subject (topic) concepts of information resources. For describing these abstract concepts and the terms used for them, we are using the SKOS framework[8], as shown in Figure 11. SKOS is a product of the W3C activity to develop standards for describing knowledge organization schemes such as thesauri, taxonomies, ontologies, etc.

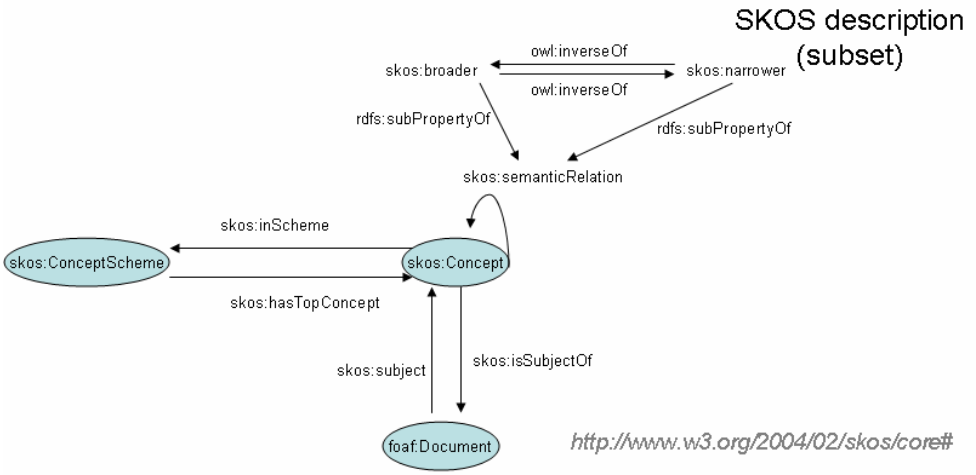

Figure 11. SKOS Ontology Subset

[8] http://www.w3.org/2004/02/skos

We first translated the existing MITRE Subject Taxonomy (MST)[9] to a SKOS-based representation[10] in order to enable more detailed concept schemes to be linked to the MST. Then we created a new concept scheme for Semantic Web-related concepts (our test subject area) and used the SKOS "broader", "narrower", and "related" properties to link our Semantic Web-related concepts to the MST. We used Wikipedia[11] as the source for definitions of concept since it is a world-wide community resource that does not reflect the view of just one cataloguer. An example concept entry is shown in Figure 12 below.

Additional concept schemes need to be defined for other areas requiring more detailed description than that provided by the MST. For MII products, this would include as a minimum concept schemes to be used for products related to Human Language Technology and Biotechnology TAT areas, since these areas are not currently represented in the MST. It is still an open question as to how ownership or stewardship of these concept schemes would be determined, but one possibility is to let the schemes develop dynamically through communities of interest or collaborative efforts such as social bookmarking or folksonomies.

```
mesa:c0001  a  skos:Concept ;
      skos:prefLabel  "Semantic Web"^^xsd:string ;
      skos:altLabel "SemWeb"^^xsd:string ;
      skos:altLabel "SWeb"^^xsd:string ;
      skos:broader mesa:c0007, mst:0601 ;
      skos:related mesa:c0005, mesa:c0006, mesa:c0011,
                   mesa:c0013, mesa:c0014, mesa:c0015,
                   mesa:c0017 ;
      dct:issued "2005-07-01"^^xsd:date ;
      skos:inScheme mesa:SemanticWebScheme ;
      skos:definition "From Wikipedia, the free encyclopedia.  The
      Semantic Web is a project that intends to create a universal
      medium for information exchange by giving meaning
      (semantics), in a manner understandable by machines, to the
      content of documents on the Web. Currently under the direction
      of the Web's creator, Tim Berners-Lee of the World Wide Web
      Consortium, the Semantic Web extends the ability of the World
      Wide Web through the use of standards, markup languages and
      related processing tools."^^xsd:string .
```

Figure 12. SKOS Example

3.4 Software Services Design

The original concept for the design of the Dynamic Topic Watch prototype application was that it would be a collection of loosely-coupled web services behind a web-based front end. It was hoped that by creating independent services there would be easy reuse with later prototypes or other projects requiring a semantic query capability. Apache Axis was the web service implementation available to us in the lab. We originally used Cerebra as the OWL engine, but switched to the Jena RDF engine when we discovered interoperability problems (see section 4 for more details.) Unfortunately, we found that Axis and Jena were incompatible because of the XML libraries on which they were based. Jena required Xerces while Axis was based on the Java Development Kit (JDK) XML library. The project schedule didn't allow time to resolve this issue, so web services as

[9] http://info.mitre.org/mii/infomgmt/taxonomy/v1_2/index.html

[10] We made a few modifications to the SKOS representation of the MST since the MST was lacking certain information: 1) We made dates be strings rather than xsd:date since the MST used a non-standard date format (e.g., unclear whether "05/07/2003" is supposed to designate "5 July 2003" or "7 May 2003"; 2) The MST only included "last modified" dates so we put that information directly on the concept even though it should be places on a change note (with one change note per change); 3) No descriptive information was included in the MST for the top elements (categories) so we assumed those elements are same elements as the sub-elements with the similar name plus the string "General"; 4) We split apart editorial notes from term definitions even though the MST did not since we did not want to display editorial information to end users.

[11] http://www.wikipedia.org/

a transport layer had to be discarded. However, the underlying components are still sufficiently decomposed to provide reuse by other systems.

The entire Dynamic Topic Watch prototype can be thought of as four different service calls: a call to check if a concept is in the MITRE Ontology of Terms, a call to query an ontology, a call to establish a subscription, and a call to retrieve a subscription. In addition, the Meta Data Generator can be called as a service, either by the MESA Crawler or by an individual to determine meta data for a particular source. Services are invoked by the GUI layer, a JSP application which controls the basic flow and object management needed to call the services. The actual calls are to regular Java methods, but these could be changed to SOAP calls if the incompatibility between Jena and Axis is resolved.

When the query service is invoked, the first service called is a concept checker, which checks if a search term is contained in the MITRE Ontology of Terms. We did not want to limit the user in their choice of terms, however, so users may enter concepts that are not in the MITRE Ontology of Terms. By capturing these terms we hope to be able to expand the MOT to include new topics of interest as they arise.

The call to the concept checker results in a list of instances that is passed to the query service. The query service also receives a list of information source types from the MITRE Enterprise Object Ontology that were chosen by the user. The query service calls the Jena RDF engine with a query constructed from all of the specified terms. The results are gathered into an XML file, which is returned to the caller. If the caller is one of the other Dynamic Topic Watch components (either the GUI or the Subscription daemon) an XSLT stylesheet is applied to provide a more graphically appealing output.

Subscriptions are managed by two simple services that store and retrieve a user's subscription preferences. These include the concepts and information source types of interest, a frequency at which reports should be provided, the delivery type for the subscription (HTML, Email or RSS), and the address to which the results should be sent or can be viewed. This information is stored in a simple relational database. This database is queried daily by the subscription daemon, a separate process that handles the actual performance and delivery of each subscription.

The subscription daemon wakes up and queries the database to gather all subscriptions that need to be fulfilled that day. If the day is the end of the week, month, or quarter, then those subscriptions are gathered too. For each subscription, a query is executed, and the results are stored or delivered depending on the delivery type. Email subscriptions have the results sent to the user's address, while RSS and HTML subscriptions are stored on the Dynamic Topic Watch server. RSS subscriptions also have their feeds updated with the addresses of the new results.

These services provide the core functionality of the Dynamic Topic Watch application. This functionality, especially the query functionality, is directed by the GUI but not tightly bound to it. Should additional use cases be found, these services can remain the foundation of any expanded capability.

3.5 Meta Data Generator

The ontologies describing the relationships between keywords (the MITRE Ontology of Terms) and attributes of data sources (the MITRE Enterprise Object Ontology) provide a useful model for discovering the semantic relationships between topics and publications. The full power of this model, however, can only be fully realized, by applying it to a set of actual sources of data. Preferably this model would be applied to as large a set as possible, but this can be tedious and time consuming if done by hand. In order to reach as large a space of sources as possible, the Meta Data Generator was developed.

The Meta Data Generator works by searching a resource specified by the user for concepts (and semantically related concepts) contained in the MOT. This source can be located anywhere that is

URL-addressable, including the World Wide Web and both local and networked file systems. Several types of sources can be searched, including PDF, XML, HTML and Microsoft Word and Powerpoint documents. If the document type has a designated metadata section, this is scanned in addition to the body text of the document. In addition, new source and document types can be supported by writing a Java extension using the Meta Data Generator's extensible API.

Once a match with a keyword is found, the results are written to an instance file, an RDF document that describes the relationships between actual sources and concepts. This is the document referenced by the Jena engine whenever a query is performed. The Meta Data Generator can be called manually, or by the Crawler as part of a larger search across several sources. In this way, an instance file describing a large number of sources can be generated without manual annotation or searching, giving the abstract model described in the ontologies a practical form usable by the query engine. The Meta Data Generator thus plays a key role in the overall Dynamic Topic Watch prototype.

4 Findings

First, we were successful in building a working prototype in which queries can be made over semantic concepts, integrating information sources on the MII and Web. We were successful in demonstrating that the technology is viable. However, in developing the MESA prototype, we obtained valuable experience with some of the drawbacks of Semantic Web-based solutions. We also identified some considerations for the way ahead for the MII. In this section we highlight these issues and provide suggestions for sponsors considering the application of the Semantic Web technology.

Even though attempts were made by the W3C Semantic Web Working Groups to layer OWL semantically and syntactically on RDF, we encountered interoperability problems when we attempted to query over RDF and OWL ontologies. Recall that we built the MEO and MOT in OWL, and both reused concepts from SKOS and FOAF, which are built in RDF. We found that it is currently not possible to query across RDF and OWL ontologies simultaneously. This is due to basic semantic incompatibilities between the languages. OWL makes a distinction between types of properties (i.e., whether a property relates a class instance to a class instance, or a class instance to a literal) that RDF does not make. Figure 13 illustrates this point. Also, RDF only recognizes named (or anonymous) classes whereas OWL permits other types of class descriptions, e.g., by defining restrictions on class members.

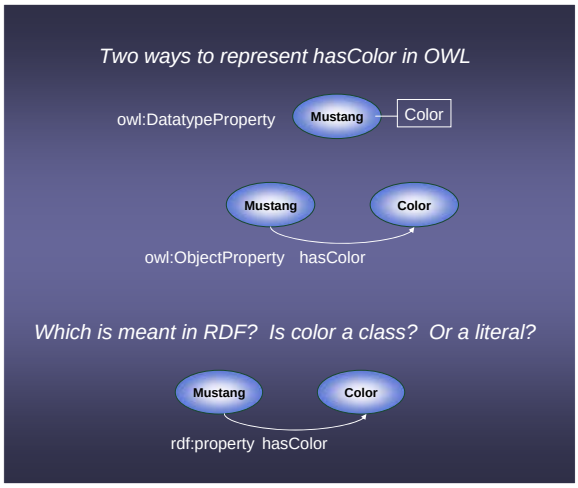

Figure 13. RDF and OWL Semantic Incompatibility Example

To keep the project moving, we resolved the problem by building the MEO and MOT in RDF. This delayed our effort significantly, since we had to rebuild ontologies and change inference engines (from Cerebra to Jena) in the middle of the construction. More importantly, this approach is not preferred because it runs contrary to the Semantic Web vision. The Semantic Web vision is one of interoperable resources exposed and available for use on the web, which is apparently not possible between RDF and OWL currently. In researching a more palatable solution, we shared our findings with the W3C and greater Semantic Web Community and found our that OWL Full addresses the semantic incompatibility we encountered. However, OWL Full is not decidable, and therefore no tools have yet implemented OWL Full. Also, it is our understanding that SPARQL[12], a proposed query language for the Semantic Web, is being extended to operate over OWL (in addition to its current ability to operate over RDF). These are two potential solutions to the problem. In the meantime, a choice must be made early in the design process regarding the level of expressivity (i.e., ontology language) is required for the application's ontologies. RDF-based ontologies are conceptually simpler; depending on the application, OWL ontologies may not be worth the extra complexity. OWL Full (like RDF) permits a resource being both a class and an instance, a feature useful in meta-modeling. This type of modeling (e.g. "F-15" as both a class of aircraft as well as an instance of the class of aircraft types) may be relatively common in mission applications; therefore, it would appear that RDF/OWL Full can sometimes be more desirable than OWL Lite/DL despite issues with computational guarantees for inference. And regardless of the ontology language chosen, it is far too easy to inadvertently move from one OWL species into another with the addition of even a single statement (e.g., when merging ontologies). While schema validation would catch those cases, resolving the issues may not be straightforward, particularly if multiple ontologies have been imported.

In addition to the semantic interoperability issues, we also encountered another technical issue. As discussed in section 3.4, we also found incompatibilities between the Apache Axis web service implementation package and the Jena RDF engine, since each were based on different XML libraries. To avoid this issue, we did not implement the services in SOAP, and hope to move to REST in the future or to a compatible implementation of SOAP.

Regarding MII data sources, we found minimal structured sources; most use HTML primarily, and employ various approaches to embedding information. Meta data is often applied manually and sometimes inconsistently. To extract relevant information for proper meta data application, ontology classes had to map to MII data sources based on data embedding methods. Because most MII sources are HTML based, we found that we needed an additional “data layer” to represent instances. We found that a web crawler capability is needed to “discover” information and populate instances. We did explore Google as a source of instances, but found that the APIs to MITRE’s Google were not sufficient to identify what we needed. We are currently exploring use of alternate Google APIs as a source of information.

Another consideration in the design process is whether instance data will be held in ontologies or in standard databases. There is a price to pay with using a database in that some simple relations in an ontology may need to be "reified" to fit into database tables. This extra mapping layer is probably worth it if all the data needed by the application already is located in relational databases. On the other hand, it's easier to derive hidden relations through inference if the data is in a graph (ontology). However, use of inference can also bog down a system’s performance (e.g., in the processing of a query). The system designer needs to decide when inference is worth the extra system resources and when it is not.

Finally, we found that semantic development tools are immature or nonexistent. Significant progress needs to be made in this area for wide spread use of semantic technology to occur.

[12] http://www.w3.org/TR/rdf-sparql-query/

5 Recommendations

As discussed in this paper, the MESA prototype effort has started to examine the question of how structured (formal taxonomies and ontologies) and unstructured (natural language, text search) approaches can be merged to provide some of the benefits of the former with the ease of use of the latter. Since this effort demonstrated that semantic technology is viable, we recommend that MITRE invest in additional work to expand the Dynamic Topic Watch prototype. The coincidence of the MESA effort with the Social Bookmarking pilot effort[13] could provide a unique opportunity to put the combination of these complementary approaches to the test. The Social Bookmarking application allows the MITRE staff to create and monitor a "folksonomy" for tagging web resources of interest. Such a set of user-defined tags is flat, though flexible, dynamic, and easily evolvable; these features make a folksonomy easy to use. This is in contrast with traditional tagging schemes, such as taxonomies, which are done by and for experts, require *a priori* definitions, and are difficult to evolve as the information environment changes. MESA's concept schemes, described using SKOS, allow concepts to be unambiguously identified with URIs and organized into strict hierarchies (taxonomies) or more semantically linked networks (ontologies) to enable machine processing, while also allowing the concepts to be associated with sets of natural language labels to facilitate user access. In the MESA prototype, these labels were defined and assigned by the prototype team based on knowledge of the subject area. But these labels could reflect the shared understanding of the user community even more by being defined by the community itself, i.e., by processing the tags resulting from Social Bookmarking to determine how the users perceive the concepts of interest to them and the terms they use to describe those concepts. There may be other integration opportunities, such as combining efforts with the RADAR project.[14] We are investigating these as part of the next phase of MESA development.

We also recommend investment in merging the MITRE Ontology of Terms with the MITRE Subject Taxonomy, thus semantically enabling MITRE Enterprise meta data. Further, we recommend that this ontology be implemented as a Wiki, a "MITRE Subject Wiki", so that we as MITRE engineers can own our meta data. This effort should be launched as an experiment, to see how the ontology grows and is used.

We recommend that MITRE work with the W3C and Semantic Web Community to resolve the interoperability issues with RDF and OWL. We should encourage vendors to participate in the resolution as well. We should closely review future proposed standards to ensure that additional interoperability issues are not introduced in the future. And, as discussed in section 4, we recommend that anyone considering use of semantic technology select the appropriate ontology standard up front, since the different languages are not fully compatible.

We recommend that MITRE evolve MII data sources to more robust technology, such as XML data streams and web services. We don't recommend replacement, but instead, a general migration in that direction. For example, as new MII applications are built, introduce these capabilities as web services or XML streams instead of HTML documents. Use of HTML should be reserved for the GUI tier of the architecture, and should be derived from the structured data sources. We also recommend that meta data be applied automatically and consistently, using automatic meta data generation tools in combination with the Social Bookmarking effort, as well as with a MITRE Subject Wiki.

An important point, and one that is overlooked far too often, is that the success of a system does not depend entirely on the technology used to implement it, but is highly dependent on *how* it is used. That is, social and cultural factors, as well as technical ones, can greatly affect a system. The background (or contextual) knowledge of a user base needs to be considered when designing any system, but is particularly important in the case of ontologically-based applications where the terms

[13] http://onomi.mitre.org/
[14] http://scm.developer.mitre.org/svn/radar

from a structured vocabulary are used in a very precise manner. Therefore, organizations considering employing semantic technology should develop a set of questions to assess whether this approach is appropriate for the application and the user base. Some questions that may be helpful in assessing whether more or less structured information is more suitable include the following.

- How homogeneous is the user base? Do they use the same concepts? The same terms? Concepts at the same level?
- Could the user base be motivated to tag (or make other efforts) for community good?
- What are the users' motivations? Users' expectations of benefits? Are they the same for all users?
- How often do user vocabulary and concepts change?
- Is precision or recall more important?
- Is it more important to be precise or dynamic with vocabulary?
- Is the users' concept scheme appropriate to a hierarchy in the first place?
- How much reclassification of previously classified (tagged) items is required as a concept scheme or vocabulary changes?
- What happens to the system if reclassification is not done (i.e., how much is staleness of tags a problem)?
- How many different tags may be reasonably applied to a single item? (e.g., many articles can be said to be about more than one topic)
- How much human effort required to tag and re-tag? Is manual tagging feasible or reasonable at all?

So, on the production side, the more we can automate the tagging of documents, the less the production effort will be, but the key to a successful system still lies in not requiring users to learn an artificial classification scheme in order to use the system.

Finally, we recommend that this technology be applied to a customer mission challenge. We recommend that a Mission Oriented Investigation and Experimentation (MOIE) proposal be developed to apply ontologies for heterogeneous data source integration. The data integration challenges faced by the Coalition Air Operations Center (CAOC) may provide a good starting point for such an experiment.