Semantics for Information Sharing and Discovery in the Intelligence Community

Martin Thurn
Northrop Grumman Corp., 4805 Stonecroft Blvd., Chantilly VA 20151, 703-449-3803,
martin.thurn@ngc.com

***Abstract*—One of the distinguishing characteristics of the intelligence community is the strict security framework that is used to control classified information. A counterproductive side-effect of this strict security is that intelligence analysts are often not aware of information that is relevant to their analysis. Semantic technology and ontologies can help analysts discover relevant information even if that information is under the strictest controls and even if the analysts are not cleared to access the data. These techniques can be applied immediately within the current security framework of the intelligence community.**

***Index Terms*—Discovery, Information Sharing, Metadata, Redaction, Semantics**

I. Introduction

THE many agencies of the United States intelligence community – and the corresponding organizations of her friend and partner countries around the world – employ a strict security framework to protect and control classified information. The basis of this framework is that a person is granted access to a sensitive document only if they need to know those data to perform their duties.

This basis creates two immediate impediments to information sharing and discovery across the boundaries of security levels and compartments. When sensitivity classifications are assigned to an entire document, it prevents an unapproved analyst from seeing any portion of the document, even when the document may actually contain a mixture of sensitive and unclassified information. To make matters worse, it is often the case that an unapproved analyst is prevented from knowing even the existence of that document. In the former case, the analyst can at least ask for permission to read the document and fulfill her duties; in the latter case, there is virtually no hope for the analyst ever to see the data.

II. Physically-Separate Standard Semantic Metadata

We have developed an approach to discovering and sharing information that is particularly well-suited to the intelligence community, an approach based on physically-separate standard semantic metadata. "Metadata" is a general term that refers to data that describes other data. Metadata for a document may explicitly identify the title of the document, provide a table of all the geographic locations mentioned in the document, or include any other information about the properties or content of the document. "Physically separate" means that the metadata is stored in a separate file rather than being embedded within the data file itself – an important contrast to the dominant practice of embedding all metadata within documents. "Semantic" means that the metadata represents the meaning of the data, as opposed to just syntactic sugar. In particular, our approach focuses on expressing the semantics of the *content* of the document, i.e. the actual body text, rather than facts *about* the document which are typically found in the header. "Standard" means that the metadata is represented using semantics standards such as the Resource Description Framework (RDF) and Web Ontology Language (OWL). In addition, "Standard Semantic" means that the metadata strictly corresponds to an ontology so that the meaning is explicit and can be processed by automated tools.

Using physically separate semantic metadata for discovery is not a new idea – this is a technique that has been used successfully by libraries for centuries. A card (whether paper or electronic) in a card catalog is metadata for a book in a library's holdings. The card for a rare and delicate book is itself neither rare nor delicate, and therefore does not have to be subject to the same protections as the book itself. Whereas the book may be held in a special collection accessible only to approved scholars, the card describing the book can be publicly accessible, updated frequently, and copies can be distributed to other libraries. In contrast, metadata that is not physically separate from the data – metadata that appears in the front matter of a book, for instance – cannot be updated and can only be accessed by those who already have access to the book itself.

Within the intelligence community, working with physically separate metadata has all the advantages of working with catalog cards, and also solves fundamental security problems that stand in the way of discovery and sharing of information. There are two keys to this aspect of the solution. First, the physically separate metadata can be at a lower level of classification than the data itself. It is entirely possible that the very nature of the metadata makes it lower level; or the system can be specifically designed so that the metadata is of a lower classification, if necessary. Second, the physically separate metadata can be stored on a different network (or several different networks) than the original data. The bottom line is,

while an organization may not be able to share much of its data for security reasons, it may be able to share a great deal of metadata. That metadata will allow intelligence analysts to discover the existence of information that is important to them even if they have not been cleared to access the data itself.

It should also be noted that since electronic metadata files can be much larger than physical index cards in a traditional card catalog, the metadata may easily contain a wealth of valuable content information that can be exploited independently of the actual data file. Of course, the metadata might not have the same authority as the actual data (see the sample scenario below), but it certainly can be used to suggest hypotheses.

III. Ontologies for Discovery

Rich ontologies are essential to the success of the approach to discovery described here. Ontologies allow semantic searches to match even if the query concept is more specific or more general than the concept in the metadata. Semantic metadata is data about the meaning of the data. Meaning has the property that it can be abstracted, which is important for both discovery and security reasons. An aircraft ontology, for instance, may indicate that the B-2 is a stealth bomber, a stealth bomber is a type of bomber, and that bombers are a type of airplane. This will allow a semantic search for the concept "airplane" to discover documents that mention specific types of aircraft such as the B-2 (even when the documents do not contain the query word "airplane"). And if the fact that a B-2 was used for a particular mission makes a document classified, unclassified metadata can be generated by referring to the more abstract concepts of "stealth bomber" or, if necessary, "bomber" or just "airplane". By abstracting as little as possible to meet security requirements, the semantic metadata can make the maximum amount of information available for discovery and exploitation. Rich standard ontologies facilitate this type of searching and abstraction. In the ideal case, the ontologies themselves will be standards used across the intelligence community – a central topic of this conference.

Discovery based on physically-separate metadata is often viewed as a last resort – a technique to be used only when security restrictions prevent access to the data itself. Indeed, one could argue that it should be a last resort when only very basic document metadata (e.g. Title, Author, Date) is available. However, semantic metadata can be arbitrarily rich, containing a detailed, unambiguous, machine-interpretable version of the information contained in a document. Since rich metadata provides an unambiguous and direct representation of the meaning of a document, metadata can serve as a better basis for discovery and automatic exploitation than even the document itself. As rich semantic metadata becomes available for more and more documents in a repository, search recall should increase, because exact matches are not necessary; and as the metadata becomes richer, the precision should increase as well, since fine-grained concepts from an ontology are less ambiguous than English words. Once sufficiently rich semantic metadata is available, metadata-based discovery can exceed both the recall and the precision of keyword searching against full text documents.

IV. Sample Scenario of Semantic Discovery

An intelligence analyst is creating a map of the locations of certain objects of interest. In the past, creating such maps required reading intelligence cables that describe, in ordinary English, the locations of the objects at various times. The analyst would then have to type all the coordinates into a geographical information system (GIS) to create the map – a tedious and error-prone task.

In our approach, as each cable arrives, a metadata file is created that contains RDF descriptions of what objects were at what locations at what times based on standard ontologies. This RDF can be automatically generated using existing information extraction technology such as NetOwl from SRA International, TextTrainer from Northrop Grumman, or AeroText from Rocket Software. A semantic metadata search – either a live search initiated by the analyst, or an automated "batch" query that runs overnight – is then used to discover all the metadata files that describe locations of objects of interest. Having standard ontologies greatly facilitates the indexing and retrieval required for this type of search. Since RDF is completely structured, the resulting locations can automatically be loaded into the GIS application. As a result, maps that previously took weeks to create manually are now automatically generated in seconds more accurately from a more comprehensive set of sources.

After automatically generating a new map, the analyst sees an alarming pattern and decides to write a report. Of course, she can't use metadata as source information for a formal intelligence report, so she logs on to the data repository (to which she has access) to verify the pattern against the original reporting. However, she is denied access to several of the cables because they are stored in a restricted collection. Through official channels (referenced in the metadata) she requests access to the restricted collection, receives access, confirms the accuracy of the map, and produces an important report. In the past, she never would have seen the pattern in the first place because she wasn't aware of the reports in the restricted collection.

V. Ontologies for Information Sharing

The approach and claims described above for using semantic metadata to improve discovery hold true equally well for information sharing – one can simply view the sharing as a "push" of metadata across security boundaries whereas discovery is like a "pull". However, the use of ontologies and rich semantic metadata can enhance information sharing in a radical way.

Recall that our semantic metadata is represented in a standard language (RDF) that is well-defined and machine-interpretable, and that we can create rich ontologies in OWL

that are also machine-interpretable. For discovery, these ontologies enable semantic searching by abstracting the query concepts; to aid information sharing, ontologies can be used to automatically abstract or redact the semantic metadata itself.

Another feature of OWL is that it can encode inferences and other logical constructs which can then be automatically processed in software. Classification guides rules and policies can be represented in OWL, and the computer can automatically apply those rules and policies to semantic metadata. This allows the automatic redaction or abstraction classified metadata so that it conforms to the lower classification level. Semantic technologies that exist today enable us to automatically redact metadata for information sharing.

We can actually take this one step further. We can write a classification guide in OWL in such a way that a theorem prover can be used to *mathematically prove* that the redacted data does not violate any classification rules. Pellet is one example of a widely-used and well-respected open source theorem prover.

VI. Sample Scenario of Semantic Sharing

Local law enforcement has a need-to-know whenever FBI identifies an individual in the local community with terrorist connections. However, local law enforcement does not have the need-to-know (nor do they even care) the source or methods FBI used to obtain such information. In the past, whenever a new terrorist connection was established and documented, the entire data record was classified because it described how FBI obtained the information to create the connection. The only way local law enforcement came to know about the connection would be if an FBI agent read the entire report, distilled it down to an unclassified version, obtained the relevant approvals, and finally sent the information to local law enforcement.

In our approach, as each suspect interview summary report is generated, an RDF metadata file is generated containing names and known-terrorist connections. Again, this can be automatically generated using existing information extraction technology. This RDF metadata is automatically routed to local law enforcement via a fully accredited hardware/software guard device at the FBI network boundary. This guard reads the RDF, compares it to classification guides and policies encoded in OWL, and performs a logical redaction of the simple metadata facts. The redacted RDF metadata is then allowed to pass outside the FBI network and travels on to local law enforcement, where it can automatically be added to a database or reformatted into a textual message. Through official channels (referenced in the RDF), local law enforcement can request confirmation of the information at any later date.

VII. Conclusion

Discovering information in an environment with strict security constraints is a critical problem for the intelligence community. Physically-separate metadata can be used to overcome some of these problems. Metadata can have a lower level of classification than the data itself, and can reside on a different network than the data itself. In this way, more accessible metadata indexes can be created and exploited while fully maintaining the security of the source data. This means that even the most sensitive documents can be discoverable, and much of the information they contain can be exploited – even by analysts that have absolutely no access to the source documents themselves. Effective discovery and exploitation, however, depends on the availability of rich content metadata that is based on extensive ontologies.

There is an inherent conflict in the intelligence community between the responsibility to share information and the responsibility to protect it. This dilemma can be finessed by protecting *data* and sharing rich *metadata*. This approach can be implemented within the current strict security framework and will benefit significantly from the type of ontology work discussed at this conference.

Semantic technologies that exist today enable us to automatically convert documents to metadata, automatically redact that metadata to any security level, and automatically prove that the redaction is sound and complete.

Acknowledgment

Martin Thurn thanks Dr. Terry Patten for 20 years of friendship and mentoring, and for his pioneering work in computational linguistics, natural language processing, information extraction, and most recently, application of semantic technologies to the problem of secure information sharing.

References

[1] D. Nardi and R.J. Brachman, “An Introduction to Description Logics”, *The Description Logic Handbook*, Jan. 2003.

[2] F. Baader and W. Nutt, “Basic Description Logics”, *The Description Logic Handbook*, Jan. 2003.

[3] A. Uszok, J. Bradshaw, R. Jeffers, N. Suri, P. Hayes, M. Breedy, L. Bunch, M. Johnson, S. Kulkarni, and J. Lott, “KAoS Policy and Domain Services: Toward a Description-Logic Approach to Policy Representation, Decomfliction, and Enforcement”, *Proceedings*, pp 93-96 [IEEE 4th International Workshop on Policies for Distributed Systems and Networks, June 4-6 2003].

[4] J. Bradshaw, A. Uszok, R. Jeffers, N. Suri, P. Hayes, M. Burstein, A. Acquisti, B. Benyo, M. Breedy, M. Carvalho, D. Diller, M. Johnson, S. Kulkarni, J. Lott, M. Sierhuis, and R. Van Hoof, “Representation and Reasoning for DAML-Based Policy and Domain Services in KAoS and Nomads” [AAMAS ’03, July 14-18 2003, Melbourne Australia].

[5] N. Suri, J. Bradshaw, M. Burstein, A. Uszok, B. Benyo, M. Breedy, M. Carvalho, D. Diller, P. Groth, R. Jeffers, M. Johnson, S. Kulkarni, and J. Lott, “DAML-Based Policy Enforcement for Semantic Data Transformation and Filtering in Multi-agent Systems” [AAMAS ’03, July 14-18 2003, Melbourne Australia].