Chapter 4

A Multi-INT Semantic Reasoning Framework for Intelligence Analysis Support

Terry JANSSEN*, Herbert BASIK†, Mike DEAN‡, Barry SMITH§

*Lockheed Martin Corporation, †BBN Technologies Inc., ‡University of Buffalo

Abstract: Lockheed Martin Corp. has funded research to generate a framework and methodology for developing semantic reasoning applications to support the discipline of Intelligence Analysis. This chapter outlines that framework, discusses how it may be used to advance the information sharing and integrated analytic needs of the Intelligence Community, and suggests a system/software architecture for such applications.

Keywords: intelligence analysis, semantic technology, reasoning, common logic

Introduction

The volume of data available to intelligence agencies and the complexity of the national security environment are increasing so rapidly as to overwhelm a finite workforce of analysts. Machines, running knowledge-based applications, are needed to augment human cognitive capacity in order to achieve the levels of situational awareness desired by decision-makers and commanders. We describe the state of the art in semantic approaches (i.e., ontology-based solutions) to this problem, in both the intelligence domain and in other domains with similar information and knowledge requirements. We also outline the results of Lockheed Martin Corp. research to address some of the specific challenges confronting the integration and fusion of data generated by multiple intelligence agencies.

1. State of the Ontological Art

Over time, each intelligence agency has developed its own mechanisms for representing data, information, and knowledge. The divergence of resulting representations and standards poses severe obstacles to the automated integration and management of data, etc. This gave rise to a US Government Executive Order [1] of August 27, 2004, which expressed a mandate to the effect that intelligence agencies must strengthen their mechanisms for the sharing of terrorist information, for example through more widespread and systematic use of XML and similar markup standards.
Experience in other domains (especially in bioinformatics) suggests that ontologies can play a crucial role in supplementing such standards and in guiding their development and use. The following paragraphs discuss the Gene Ontology, from which the reader can observe parallels with similar issues faced by intelligence agencies. The Gene Ontology (http://www.geneontology.org) is the most successful ontology initiative thus far when measured by numbers of users and of supporting software tools, and by the diversity of problems to which it has been applied. The Gene Ontology is a controlled vocabulary for describing the attributes of gene products. It has been used to annotate data derived from biomedical experiments. It makes these data, which would otherwise exist in multiple separate silos maintained by multiple separate research communities (i.e., akin to the current structure of the various intelligence agencies), more easily integratable and comparable. Lessons from experiments on the effects of given drugs or toxins or pathogens on mice, for example, can be drawn to help in the understanding of what the comparable effects might be in human beings.

This success has led to the creation of other ontologies covering other domains of biology, including the cell ontology or the phenotype ontology, which now form part of a higher-level ontology-based integrating framework that is being realized already within the context of the Open Biomedical Ontologies (OBO) Foundry initiative [2, 3]. A plurality of ontology modules is being created by different community groups using both Web Ontology Language (OWL) and OBO-specific ontology formats against a background of common development principles designed to ensure their interoperability.

The OBO Foundry ontologies contain terms designed to capture in an algorithmically reasonable form the qualitative aspects of biomedical phenomena. We focus on qualitative data here, similarly, in the security area, pertaining for example to intention or threat, to religion and family relationships, or to relative spatial location, as expressed for example in observation reports [4]. In this we go beyond traditional methods of what is called “information fusion” having been developed primarily for integration of quantitative data.

Experience with OBO Foundry has shown that a combination of semantic technologies is appropriate for capturing such qualitative data. Our goal here, more specifically, is to advance the needs of intelligence agencies in integrating and interpreting very large bodies of qualitative data through the application of semantic technology. Little et al. describe elsewhere in this volume those aspects of our project which pertain to the use of ontologies to support multi-INT data fusion when enhanced through the consideration of probabilities [5]. Here we confine ourselves to describing the general approach and to giving some outline of the results.

2. Premise of Current Research

The premise of this research is that a system can be built using ontological models which enables multi-INT data to be semantically fused; that is, to be integrated in such a manner as to allow automated reasoning to be performed over the combined dataset. This is by virtue of the fact that the data is associated with terms for which reasoning rules are defined in ontologies. The semantic benefits of ontologies result from the fact that they capture certain aspects of the meanings of those terms, in addition to their taxonomic and other structural relationships.

For intelligence data to be semantically fused, it must exist as “stand-alone” facts or assertions. That is, extracted from, or tagged within, intelligence products, such as images, reports, cables, etc. They must also be able to be uniquely accessed from relational and other types of databases.

Part of this research investigated the integration of data extraction tools in a semantic architecture. Considerable effort has been expended on the entity extraction problems from free text, and other unstructured data formats. In this work, we focused on entity extraction from imagery. For this, we used an image semantic mark-up tool, GARCON™ [6] which will be discussed in greater detail in a later section.

In particular, our research focused on using COMINT (Communications Intelligence) data, such as telephone transactions, to illustrate and demonstrate our conclusions. The problem addressed was the automated generation and analysis of social networks. The work of associating persons-of-interest with telephone transactions, and of investigating degree-of-separation-type problems, had been performed previously. In this effort, we attempted to demonstrate the integration and use of multi-INT data (i.e., data provided by two or more intelligence agencies or disciplines) to improve the quality and utility of the social networks generated. Specifically, the integration of IMINT (Imagery Intelligence) information such as geo-location and geo-temporal relationships, with social relationships.

For example, an ontological model may allow the system to conclude (on the analyst’s behalf) that if a cell phone has been used in a communications transaction and that that cell phone is geo-located to a particular place and time, then the owner of that cell phone may be similarly located. In this case, simple rules allow the performance of inferential steps linking communication events to specific telephone numbers and thus to people; thereby adding intelligence value to the social networks involved.

The ultimate goal of this research is to extend such reasoning to include different data types, in order to advance our total situational knowledge. For example, to data types associated with intelligence questions dealing with activity, behavior, and intent. Our assumption at this point is that entity extraction and markup will have previously occurred (e.g., from natural language text), and that the “facts” extracted have been entered into a knowledgebase, perhaps virtual.

2.1. Differing Agency Missions and Vocabularies

There are many efforts in the DNI/IC/DoD/DHS to develop common vocabularies for particular agencies and their missions. Some examples are DISA’s Directive to establish a common vocabulary and define a set of services and interfaces common to DoD information systems [7], and DoD’s Directive including Unique Identification (UID) Standards for a Net-Centric Department of Defense, with common vocabulary and definitions [8].

An example is the information and vocabulary associated with the domain of “terrorism.” This is defined as all information, whether collected, produced, or distributed by intelligence, law enforcement, military, homeland security, or other activities relating to [9]:

a) The existence, organization, capabilities, plans, intentions, vulnerabilities, means of finance or material support, or activities of foreign or international terrorist groups or individuals, or of domestic groups or individuals involved in transnational terrorism.
b) Threats posed by such groups or individuals to the United States, United
States persons, or United States interests, or to those of other nations,
c) Communications of or by such groups or individuals, or
d) Groups or individuals reasonably believed to be assisting or associated
with such groups or individuals.

Each agency and mission area dealing with a particular subset or aspect of
terrorism and may use differing terminology and interpretations. A key conclusion of
our research is that these differences must be respected (from a practical perspective),
and we do not advocate attempting to define and enforce a single vocabulary. We
intend the concept of semantic fusion to overcome these obstacles.

2.2. Overview of the Concept of Semantic Fusion

The solution proposed by this research is to integrate the conceptual knowledge models
of the various intelligence agencies by constructing a formal comprehensive conceptual
model (CCM) which spans the Intelligence Community and DoD. Our proposed
concept is illustrated in Figure 1. The purpose of the CCM is twofold. First, it models
those aspects of the real world which are not contained in INT-specific models. Second,
its provides an integrating framework in which INT-specific concepts can be understood.

![Diagram of Integration Via Comprehensive Ontological Framework]

Figure 1. Integration Via Comprehensive Ontological Framework

The first conclusion reached by this project is the recognition that conceptual
models (i.e., ontologies) written in the same system of logic cannot be effectively
integrated. This is true because small differences in “common” concepts cannot be
detected (or resolved) if both versions are represented in languages with the same
“expressivity.” That is, the difference itself may exceed the language’s ability to
represent it.

The relevance of this conclusion to the problem at hand is that the Intelligence
Agencies and DoD need to develop (or already have developed) INT-specific
conceptual models, which define terms and relationships in similar, but not exact, ways
as noted above. They have behaved as stove-pipes. In general, the most advanced of
these models are represented in the Web Ontology Language / Description Logic
(OWL-DL). Description Logic (DL) refers to a system of logic intended to provide a
basis for “describing” information, entities, and relationships, and for supporting
inference within these descriptive models.

The full CCM proposed by this project will need to be written in a more expressive
language than OWL-DL. It will be written in the dialect of Common Logic (CL).
Being a first order logic, CL allows much greater resolution in modeling the real world.
It is therefore capable of representing differences between the OWL-DL models
developed by the individual intelligence agencies. An analogy to this line of reasoning
is the use of real mathematics to understand integer arithmetic. The CCM will also
need to address the concept of uncertainty [10].

Pragmatically, this approach does not impose a common vocabulary across the
Intelligence Community, or force substantial rework to harmonize agency-specific
ontologies. Rationalization across ontologies is achieved via mapping to a higher-order
logical standard.

2.3. Rationale for Approach

Rationale for the viability of this approach is derived from a prototype ontology-based
knowledge framework, for which we have been able to demonstrate both fusion and
reasoning capabilities of the sort required. Our framework is designed to provide
services useful to intelligence analysts by allowing them to draw on the functionality of
ontologies without imposing the use of a common vocabulary.

At the same time however our approach addresses need being increasingly felt
across the intelligence community to force substantial harmonization of agency-
specific approaches to knowledge representation. As in the biological domain of the
OBO Foundry, it exploits the benefits of modularity in building a common upper-level
framework to which agency-specific representations can be mapped according to
specific needs, in ways which guarantee the interoperability of the modules used.

As this research continues, we will continue to extend the range of incorporated
modules, again drawing on the example of the OBO Foundry family of ontologies, to
achieve broader coverage across the intelligence / security domains.

3. Discussion of Research and Conclusions

3.1. Semantic Image Markup Tool

One component in our framework is the GARCON™ tool referenced above, which is
receiving attention especially in the GEOINT (Geospatial Intelligence) community as a
tool for annotating images. GARCON uses an ontology, written in OWL-DL, as the
basis for an analyst’s semantic mark-up of an image. Semantic mark-up implies that the entities and relationships extracted from an image are able to be placed in the correct context (as specified by the ontology).

The value of semantic markup of imagery is illustrated in the following example. If an analyst is viewing an image of an airfield, and detects the presence of aircraft and other entities, semantic mark-up (as opposed to non-semantic markup) enables the following types of inferences to be made automatically, in addition to the mere augmentation of the metadata associated with that image.

a) The relative position of the aircraft with respect to other entities in the image (e.g., a fuel truck) may indicate the operational status of the aircraft,
b) The type of aircraft itself indicates the range of possible activities that it may be engaged in, and possible insight into the intent of cognitive entity (e.g., the commander) which caused that aircraft to be in that location,
c) The possible situation/context which “best explains” the pattern of all the entities in the image (i.e., abductive logic).

The role that semantic image mark-up played in this research is that it demonstrated how one type of intelligence data (e.g., from GEOINT) can be represented in a manner which enabled it to be automatically fused with other types of intelligence data, to allow deeper situational awareness.

3.2. Semantic Multi-INT Data Integration

Our envisioned unified, but modular, framework for intelligence-related information will contain three levels of ontological models: general, intel-related, and INT-specific. These levels correspond to the industry terminology of upper, mid-level, and domain ontologies. Upper ontologies describe concepts such as time, location, part-ness, process, event, etc. Mid-level ontologies introduce concepts of interest across the intelligence domain (as well as other large domains), such as person and organizational relationships, threat, capability, etc. These models enable the integration of information described in domain-specific ontologies. For our purposes, domain-specific ontologies contain the concepts unique to a particular intelligence agency. For example, electron transactions. The framework provides a coherent methodology for capturing, expanding, and reusing sets of models, at all levels, developed by master analysts and subject matter experts.

The various types of intelligence (e.g., Geospatial (GEOINT), Signals (SIGINT), Human (HUMINT), Open Source (OSINT), etc.) will be integrated as shown in Figure 1. Our hypothesis is that even though different bodies of information are described using different ontologies based on different logical approaches, they can be unified and reasoned over by automated tools given the right sort of representational and computational framework. Crucial to this framework is the use of a more powerful logic - the CL (Common Logic) standard - to integrate information annotated using ontologies formulated using semantically weaker languages, particularly OWL-DL (Description Logic).

Following the lesson of the OBO-OWL conversion project referred to above, we have created similar facilities to convert OWL-DL ontologies to the CL format within the framework of our IRAD project. We have converted information from different sources into a common format, and been able to reason over the aggregated knowledgebase.

Among the benefits of this approach is that first-order logic allows a level of model sophistication that is not easily achieved (or may not be possible) with DL. That is, CL provides significantly enhanced expressivity, as befits a first-order logic (FOL). We believe that pulling OWL-DL content into an FOL reasoner is the only viable approach to providing what is called “all-source” analysis.

As far as we know, this project is the only application of this approach. The Vampire project by the OWL working group Manchester translated OWL into FOL [11], as did Stanford University, but neither focused on end-user application. Trade-offs are of course involved, since with the increased expressivity of CL means that the resultant framework is no longer marked by the feature of decidability. We believe, however, that decidability is of lesser concern, when compared with expressivity. This assertion has been demonstrated in projects such as the Large Knowledge Collider [12], trading off completeness for scalability. This trade-off is justified as seen in an effort supported by IKRIS [13] as well as the results described below.

In biology and other sciences, gaps or inconsistencies in data can be addressed via increased experimentation. In the intelligence domain, gaps and inconsistencies, along with a variety of related problems, are not typically resolvable in this way. Specifically, our information comes from multiple uncoordinated sources, many of which will produce either:

a) Similar information about the same instances,
b) Similar information about distinct instances (which may accordingly be confused),
c) Differing information about the same instance (this can produce conflicting information, e.g., concerning the spatial or temporal location of an event),
d) Differing information about different instances (there may be two separate but related items being tracked in different ways).

There is inevitable uncertainty whenever we are attempting to reason about how instance-level data fit together to form a common operating picture. Intelligence reports are noisy and information is incomplete. In addition, there is a conscious attempt by adversaries to both deny access to information and to increase the uncertainty/ambiguity of that information which is observable. This means that all of the mentioned alternatives will generate knowledge problems, for example because we sometimes believe that two instances are identical when they are in fact distinct. In practical terms it means that combining probabilistic reasoning with semantic technology is an important enabling capability for multi-INT fusion. Facilities for probabilistic reasoning are accordingly an essential component of our project.

3.3. The Need for High-Expressivity Ontology Languages

Intelligence agencies have developed INT-specific terminologies for describing qualitative data which define terms and relationships in semantically similar but not necessarily identical ways. Many of the most advanced of these models have been encoded in OWL-DL. OWL-DL is a W3C standard with many attractive algorithmic properties. However, OWL-DL does not provide the degree of expressivity needed for
practical models in the intelligence domain. That is, it cannot easily express complex qualitative information, especially in areas where time and change are involved [14].

For this reason our project draws on the resources not only of OWL-DL but also of the more expressive power of Common Logic (CL), an ISO standard language for first-order logic and related logics. A code fragment written in a Common Logic is shown below. This construct states that a "meeting" is an event (i.e., an instance of type event), in which two people (i.e., agents) are inferred to have met if they are both related to the same event which occurred at a particular time.

:Prop Meeting
:Inst Event
:Sup Event
:Name "Meeting"
:Lex "?I is a meeting"

(! (and (Meeting ?e)
 (agent ?one ?e)
 (agent ?two ?e)
 (occursAt ?e !))

(holdsIn !t (meetWith ?one ?two))

We draw specifically on the resources of CL with Well-Founded Semantics [15], which has desirable computational properties when used to reason over large bodies of data. CL also has the high expressivity we need to represent complex real-world situations, particularly the expressivity needed to describe things that are changing / evolving over time. For example, dynamic entities such as organizations which gain and lose members, change locations, establish differing relationships and associations, etc.[16]. Well-Founded Semantics [17] provides fast and efficient query-answering capabilities even when addressing large data collections containing 10s of millions of assertions. Like OWL-DL, Common Logic is XML compliant. At the same time, CL is marked by a high degree of syntactic flexibility, so that individual CL systems may use a variety of non-XML syntactic frameworks which are always mapable to a fully XML-compliant syntax.

Our system will be useful only if it is responsive to the query needs of intelligence analysts in both accuracy and timeliness. It must therefore be computationally tractable. The solutions provided by HighFleet, Inc. (formerly, Ontology Works) addresses this tractability by including proprietary heuristic algorithms in the reasoner which detect and resolve conditions of non-decidability or runaway.

Many organizations in the intelligence and security domains have adopted XML for interoperability between systems and organizations; as is evidenced for example by the XML Directive from the Dept of Navy [7]. XML is, or will become, the de facto standard for the DoD and IC. RDF and OWL-DL, and eventually CL, are logical steps in the progression of standardization, in part because they have been developed on top of XML.

3.4. System / Software Architecture

Figure 2 shows our proposed system architecture for a semantically-enabled system to support intelligence analysis. The primary interface with analysts is through a query engine operating on a combined, multi-INT knowledgebase. This virtual knowledgebase, has access to all of the information extracted from the intelligence products produced by each of the intelligence agencies, subject to security constraints. We assume that the extraction and markup is based on each agency's native ontological models and representation methods. The transformation of this information to our proposed comprehensive ontological framework will occur on the knowledgebase side of the interface. Therefore, these transformations (if needed) will be transparent to the contributing agencies, after the semantic compatibility issues have been resolved.

The ontological model (or set of ontological models) is the proposed comprehensive model for the intelligence domain. This is what we propose to write in Common Logic. It will drive the knowledgebase, as well as the transformations of agency-unique data into the common format. The set of automated tools (e.g., reasoners, etc.) will operate on the knowledgebase itself, and will support the understanding and response to queries received via the query engine.

It must be noted that although the intelligence analyst appears to be peripheral to this architecture, the entire system is intended to be a support tool to him. The analysts and other subject matter experts are also integrally involved in the development of the ontologies themselves.

3.5. Experiment

This project itself was designed to validate the concepts discussed above by demonstrating how the utility of a social network model could be improved via the integration of dynamic spatial and temporal relationships. This would in effect illustrate how data from various intelligence agencies could be fused. For example, social relationships generated by one agency could be combined with location information generated by another agency.
A fragment of code is shown below. It provides the logic to determine if there is an entity at a certain location during a particular time interval.

\[
\text{\( \rightarrow (\text{and} \ (\text{find.entity} \ ?x12 \ ?x11)) \)} \\
\text{\( (\text{find.temporalIndex} \ ?x12 \ ?x109) \)} \\
\text{\( (\text{time.intervalContains} \ ?x109 \ ?x106) \)} \\
\text{\( (\text{time.intervalStartedBy} \ ?x106 \ ?x107) \)} \\
\text{\( (\text{garcon.time.xsdDateTime} \ ?x107 \ ?\text{start}) \)} \\
\text{\( (\text{or} \ (\text{georss.where} \ ?x12 \ ?x39)) \)} \\
\text{\( (\text{find.spatialRegion} \ ?x12 \ ?x39)) \)} \\
\text{\( (\text{or} \ (\text{geo.representativePoint} \ ?x39 \ ?x38)) \)} \\
\text{\( (\text{rcc.invPart} \ ?x39 \ ?x38)) \)} \\
\text{\( (\text{or} \ (\text{gml.pos} \ ?x38 \ ?\text{coords}) \)} \\
\text{\( (\text{garcon_space.pos} \ ?x38 \ ?\text{corrdcs})) \)} \\
\text{\( (\text{holdsIn} \ ?\text{start} \ (\text{locatedIn} \ ?x11 \ ?\text{coords})\)) \)
\]

\[
\text{\( \rightarrow (\text{and} \ (\text{find.entity} \ ?x12 \ ?x11)) \)} \\
\text{\( (\text{find.temporalIndex} \ ?x12 \ ?x109) \)} \\
\text{\( (\text{time.intervalContains} \ ?x109 \ ?x106) \)} \\
\text{\( (\text{time.intervalFinishedBy} \ ?x106 \ ?x107) \)} \\
\text{\( (\text{garcon.time.xsdDateTime} \ ?x107 \ ?\text{start}) \)} \\
\text{\( (\text{or} \ (\text{georss.where} \ ?x12 \ ?x39)) \)} \\
\text{\( (\text{find.spatialRegion} \ ?x12 \ ?x39)) \)} \\
\text{\( (\text{or} \ (\text{geo.representativePoint} \ ?x39 \ ?x38)) \)} \\
\text{\( (\text{rcc.invPart} \ ?x39 \ ?x38)) \)} \\
\text{\( (\text{or} \ (\text{gml.pos} \ ?x38 \ ?\text{coords}) \)} \\
\text{\( (\text{garcon_space.pos} \ ?x38 \ ?\text{corrdcs})) \)} \\
\text{\( (\text{holdsIn} \ ?\text{finish} \ (\text{locatedIn} \ ?x11 \ ?\text{coords})\)) \)
\]

The resultant Lockheed Martin multi-INT ontology (organized in the framework structure) was created to accommodate data gathered from a variety of intelligence sources. All data is stored in a knowledge server, which is retrieved / processed via a Common Logic query engine. The knowledge server used in this project was provided by HighFleet, Inc. (formerly known as Ontology Works). The ontology defines both a data model and a conceptual model, effectively tying together the data from multiple sources into a single understanding of the world.

To test the viability of our approach, we conducted the following experiment. Three datasets from the National Counterterrorism Center (NCTC), the Mondial project (using the 1996 CIA World Factbook), and the Garcon-F project were incorporated into our pilot framework together with fictional (randomly created) HUMINT and SIGINT data. The ontology was created in Common Logic to integrate the semantics of these datasets.

Data ingest programs were written to pull the data into a knowledge server, which was then used to answer queries that spanned multiple domains. In our chosen scenario, we asked for details about a kidnapping (i.e., an event recognized by NCTC), the membership and leadership of groups blamed for the action, communications between group members mentioning particular keywords, and imagery that positioned those communications in a geospatial, geotemporal framework.

Figure 3 illustrates how an analyst may write a query, associated with the above problem, in a graphical language provided by HighFleet, Inc. The relationships needed are defined in the ontology, and the instance data is contained in the knowledgebase. Once the query is defined, the knowledge server will respond with whatever data matches the defined patterns. Figure 4 contains the response to this query, also in graphical format, based on representative data developed by this project.

One of these rules of inference is used to conclude that two people met at a particular time given an event description of a meeting. These rules of inference provide clarity both for the ontologist and for the query writer, who can concentrate on the relatively simple relationship between two people rather than a complex event.

The rules of inference were also used to integrate data from different sources. In one case, the Garcon-F geospatial/imagery ontology was incorporated into the Multi-Int ontology. Rules were written that combined facts based on this OWL-DL ontology to conclude the location of objects at particular times.
4. Conclusions

The execution of this experiment did in fact demonstrate the ability to integrate multi-INT data, natively expressed in lesser semantic formats, into a cohesive whole. That is, to provide the capability to reason over a consolidated knowledgebase. The models in the framework provided the information needed to make the implicit semantics of the native models more explicit.

The experiment also confirmed, albeit in a non-exhaustive manner, that systems based on Common Logic (with Well-Founded Semantics) could overcome some of the computational difficulties associated with first order logics. Even if not “provable,” such systems could greatly enhance the analytical capabilities provided to intelligence analysts.

In the foregoing we have described only the basic outlines of the project. In addition we are realizing a number of additional components, including image annotation, data import and results visualization. Our major focus is to construct the engineering required to take our approach to multi-INT data integration into production, and to bring our pilot testing on artificial data to the level where the approach can be thoroughly tested by information analysts on large bodies of real-world data.

Acknowledgments

We are grateful to the Lockheed Martin Corporation, and also to the following for crucial contributions to this work: William Andersen, Ryan Kohl, Mitch Kokar, Kathryn Blackmond Laskey, Eric Little, Sarah Taylor, Jim Pridgen, Tony Chen, Kirsten Gheen and Jessica Hartnett.

References