An RDF Prototype System for Federation and Semantic Alignment of Disparate Counter-Terrorism Information Sources

M. Personick, B. Bebee, B. Thompson, Y. Wang, M. Jaworowski
SAIC/Advanced Systems Concepts
4001 N. Fairfax Driver, Suite 350
Arlington, VA 22203, USA
{personickm, bebeeb, thompsonbry, wangyun, jaworowskim}@saic.com

B. Parsia, K. Clark
University of Maryland, College Park
Maryland Information and Network Dynamics Laboratory, Semantic Web Agents Group
8400 Baltimore Ave., Suite 200
College Park, MD 20742
clarkparsia@gmail.com

Curt Soechtig
ObjectSciences Corporation
6359 Walker Lane, Suite 100
Alexandria, VA 22310
curt.soechtig@objectsciences.com

Keywords: Info Sharing and Collaboration, Counter-Terrorism, Semantic Web, RDF, OWL

Abstract
Understanding terrorist actors and their relationships is critical to being able to discern and reveal terrorist group structure. Different organizations have different sources and data repositories. Each repository contains potentially new information on a particular terrorist group. The ability to perform analysis across these repositories is one of the significant promises of the Semantic Web. Through multiple efforts within the Intelligence Community (IC), these promises are beginning to be realized. However, questions remain as to whether the Semantic Web will enable, at scale, new capabilities in combining disparate information for analysis. Through work for a government customer, SAIC/ASC is developing a prototype using Semantic Web technologies (RDF & OWL). It enables analytical tools to find new information by operating on a federated view. The work is resulting in a cognizance of the balance of logical expressiveness and the performance of current reasoning technologies with the potential for new insights by combining information. The paper presents this work and its application on a live experimental network. It discusses the application of Semantic Web technologies and the new capabilities enabled. Finally, it provides an understanding of observed technical strengths and weaknesses as applied to counter-terrorism analysis.

1. Introduction
Early detection and pre-emption of terrorist threats is of paramount importance. Understanding terrorist entities and actors, along with their links and relationships, is critical to being able to discern terrorist group structure. The ability to perform analysis across disparate repositories of Counter-Terrorism (CT) data represents a significant step towards potential enhancements in early warning capabilities. Numerous research efforts are focused on new tools to aid in this analysis. These tools assist the analyst in discovering implicit links and relationships relevant to disrupting terrorist operations. However, CT data is often fragmented across numerous intelligence agencies. Each data source provides a different, but possibly incomplete, picture. Typically, tools only operate on one data source at a time. Technologies that create more views by federating fragmented data from different agencies create a significant advantage for analysts.

Through work for a government customer, SAIC/ASC is participating in an experimental network to enable research technologies to be tested by real users with real data. This environment enables the exploration of novel
solutions to the challenges of information sharing. Each data source brought onto the network has its own schema for storing data. To enable analysts to draw conclusions across data sources, the team has developed a prototype using the Semantic Web technologies of the Resource Description Framework (RDF) and Ontology Web Language (OWL). It enables analytical tools to find new information by operating on a federated view. Starting from the initial relational approach, this paper provides a background on the technologies and challenges the SAIC/ASC team has encountered as the prototyping work has progressed. It describes the initial approach in applying Semantic Web Languages and Technologies. Finally, it discusses ongoing work towards achieving scalable performance with current technologies by understanding the expressiveness required for the target CT application.

1.1 Problem Definition
A key challenge faced in the experimental environment is to federate a view of the collective understandings of multiple organizations. Many of the data sources brought onto the experimental network are within the CT domain. As such, they have a similar high-level structure consisting of entities, links, and attributes. However, each data source has its own definition of specific concepts related to the unique circumstances in which the information is gathered and maintained. Furthermore, each analytical tool has its own “tool ontology and syntax”. To move beyond existing capabilities, an approach to enabling analytical tools to operate across a federated or merged view of this information was required.

The initial efforts used a flexible, denormalized relational schema known as the Evidence Database (EDB) (Silk et al. 2003). On the data source side, the EDB was used as a common schema for new data sources on the experimental network: each repository brought onto the network was translated into EDB format. Likewise, each tool would be wrapped inside a proxy that could translate from EDB into the tool’s specific language.

The EDB schema was highly generic – generic enough to represent almost any entity-link-attribute data. However, the inability to efficiently created merged views and the mapping requirements ultimately lead the team to pursue other types of representations. The only way to combine two data sets in an EDB world is to create yet another replication by combining their table data. However this approach alone is not sufficient. If all table data is simply merged together, the resulting EDB instance can have multiple rows in the concept tables (i.e. Entity Type) for overlapping concepts. “Concept mapping” becomes important when considering two data sources simultaneously as in the EDB merging case. In Semantic Web circles this problem is known as “semantic alignment”.

A final problem lies in the propagation of updates from the data owners onto the experimental network. Propagating updates from the data owners onto the experimental network is difficult because once replicated into EDB, both concept data and instance data lost its connection to the authoritative source. This made any sort of incremental merge almost impossible; rather the EDB instance needed to be entirely re-created. In doing so, all updates made by the analysis tools, for example implicit links between explicit entities, were lost since the original entities could not be maintained without direct support from developers. A solution that enabled a direct mapping with the source schema was required.

2. Prototype System Architecture
In response to these challenges, the SAIC/ASC team created a prototype system to better understand how Semantic Web technologies could help to address these challenges. The goals were:
- Promote a separation of concerns between the application layer (analytical tools) and the data sources.
- Provide a federation of multiple data sources in a single, semantically aligned, fused view to the analytical tools.
- Allow for propagation of updates incrementally from agencies’ source data into the federation.
- Allow for creation of new data by the analysis tools within the federation without threatening data integrity.
- Maximize overall system scalability.

The creation of the solution architecture was decomposed into a set of unique technical problems:
- Bringing new data sources into the federation.
- Semantically aligning data sources within the federation.
- Defining a query language and data access protocol to allow clients to retrieve subsets of the federated data.
- Incrementally merging updates from authoritative sources into the federation.
- Allowing clients to create new data in the federation and possibly update authoritative source data.

2.2 Conclusions
SAIC/ASC and team members have had initial success in demonstrating the efficacy of applying Semantic Web Languages and Technologies. The explorations as part of the prototyping effort have lead to the empirical conclusion that federation and semantic alignment of multiple Counter-Terrorism data sources is achievable using the current generation of supporting tools and languages. However, it requires a careful approach to manage scalability challenges. Through the continued research into the limits of expressivity and reasoning performance, the team is plans to apply scaling approaches to successfully demonstrate this prototype on an experimental network with real users and data.