The Use of Ontologies to Support Intelligence Analysis

Richard Lee
Booz Allen Hamilton
8283 Greensboro Drive
McLean, VA, 22102, USA
lee_richard@bah.com

Keywords: Information Extraction, Metadata, Ontologies

1. Overview

In this paper we describe the Metadata Extraction and Tagging Service (METS) system in use at DIA. We briefly describe the purpose and function of the system. We explain why we chose to use OWL and ontologies rather simple XML for the representation of the data it produces. We discuss an experiment we conducted on using ontologies for multi-int data fusion. We describe the OWL ontologies we've developed. We conclude with a list of the ontology and data coordination we hope to do in the future.

2. Background

A few years ago, we were tasked with evaluating the accuracy and usability of commercial Information Extraction (IE) tools and with determining the benefits of using them to "tag" many years of message traffic.

IE tools process free-text documents and extract from them items of interest. These items can cover a wide range of types of entities (persons, organizations, locations, equipment, dates, etc), and events. It is important to note that the tools do far more than simply identify the presence of such an item in the document – they extract information *about* an item. For a person, this information could include name(s), title, profession, age, hair color, etc. It could also include information about relationships between the person and other entities and events – associates and relations, membership in a group, ownership of things, instigation of or participation in an event, etc.

We considered the traditional mechanism for XML "tagging" of documents. This consists of placing XML tags around the references to an item in the document, creating XML elements. For example, the Intelligence Community Metadata Standard for Publication (IC-MSP) defines a set of "in-line" tags for this purpose. In the latest version (4.0), it allows for a set of 18 such tags, including a catch-all.

Although the IC-MSP standard does allow for a modest number of attributes, including the xlink set, it was apparent that it – or indeed any representation based on such in-line tags – would be hard-pressed to capture all the useful information produced by IE. Consider the following sentence from a sample document:

"South of Baghdad near the town of Hillah, a suicide bomber blew up his car outside the house of Police chief Maj. Ahmed Suleiman, killing himself and wounding seven, officials said."

While the text indicating specific entities and events can be tagged, all their properties and relationships are another matter:

- owner of car
- owner of house
- occupation, name, title of the intended victim
- agent, location, instrument, victims, etc of the bombing event

– spatial relations amongst the locations

Many of these concerns could be addressed by abandoning the inline-tag representation in favor of a more item-centric representation. This allows for a cleaner and more complete representation of the information, which facilitates discovery and linking of information across data sources. We have therefore gone that route.

However, the lack of a semantic underpinning for XML made us reluctant to use it as the representation for METS data. We wanted to see the data used throughout DoDIIS, across COIs, and we wanted to ensure it could be used to support automated inferencing.

Accordingly, we elected to use an RDF-based semantic representation. Initially, we used DAML (DARPA Agent Markup Language), and then the W3C standard OWL (Web Ontology Language).

3. Ontologies

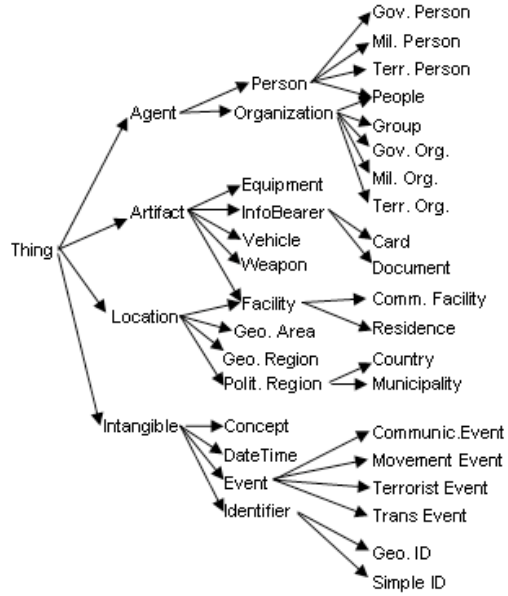
At this time, METS uses a set of three inter-related OWL ontologies which were developed on the program.

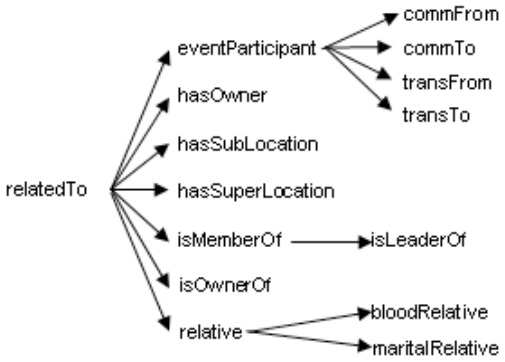
The *core* ontology was designed to arrange a broad set of domain-independent concepts into a class hierarchy. A large set of properties, both for simple text values (name, color, etc) and for relations (memberOf, uses, eventParticipant, etc) was also arranged into a hierarchy. The properties are also identified where appropriate as *transitive*, *inverses* of each other, etc, to further facilitate inferencing.

The *ct* ontology was designed, in like fashion, to cover classes and related properties that were deemed to be specific to the Counter-Terrorism (CT) domain; these were tied into the core hierarchies (via *subClassOf* and *subPropertyOf* declarations).

The *icmsp* ontology was designed to mirror the IC-MSP PublicationMetadata specification, following its XML structure as closely as possible within the added constraints of RDF. It deviates a bit from that specification to use key items (*Person*, *Organization*, *Date*, etc) out of the core ontology.

Fragments of the class and property (relationship) hierarchies are shown below.





4. METS Description

METS is a system for processing text documents. It is fronted by 4 web services:

- *Persistence* service ties to a feed of messages and newswire articles and processes them into a set of data stores
- *On-demand* service accepts arbitrary documents and processes them back to the submitter

- *Query* service retrieves processing results from the data stores matching the query
- *Bulk-transfer* service retrieves all results produced and stored in the specified time interval

METS incorporates a normalization component to convert an input document (text, HTML, XML, word, PDF) into standard XML and OWL forms (see below), and to identify the metadata. It applies a commercial categorization tool and multiple commercial extraction tools, translating the results into the standard forms. It applies heuristics and commercial tools to merge (de-conflict) and clean up the extraction results.

The result of the processing is represented as an OWL/RDF document. All the document metadata (security, date, source, etc information), including the categorization results, is represented in the OWL, using the *icmsp* ontology. All the results of the extraction -- entities, events, and relationships -- are also represented, in conformance with the *core* and *ct* OWL ontologies

Each input document is normalized into XML compliant with the IC-MSP specification. The metadata about the document is represented as called for by the PublicationMetadata portion of the specification. The categories identified by the categorization are included in the IC-MSP metadata as well. The entities and events identified by the extraction are flagged via in-line tags (the set of tags used is actually much larger than the set allowed by the specification, indicating the larger set of entity and event types extracted).

METS is operational at DIA on JWICS, processing live WISE message traffic. Multiple projects are developing interfaces to submit documents and data requests to the METS web services.

5. A Multi-INT Experiment

The data processed by METS for storage is message traffic (largely HUMINT) and newswire articles from WISE. As an experiment, we supplemented the system with a new component which produced OWL from IMINT data, and one which attempted to correlate the data from the two INTs based on location. We enhanced the core ontology with more geographic and geometric concepts to support this; this is of course a prime candidate for carving out and replacing with standard ontologies. The results were encouraging, but suffered from the inability of METS' extractors to disambiguate (and therefore provide coordinates for) location references in many cases.

6. Future Ontology and Data Coordination

We will continue to work on improving the coverage and accuracy of the IE in METS.

While the current ontologies were developed in-house, in consultation with CT analysts and their data schemas, we will continue to track and participate in efforts toward standardization such as this conference, work on Catalyst, TWPDES, Universal Core, etc, with the goal of helping devise ontologies that are used and interconnected across the community.

We also hope to be coordinating with other projects to:

- identify coreferential items across the METS-processed documents and other data sources
- discover more knowledge by using the ontology-based inferencing capabilities

7. References

Information about METS is at http://mets.d2lab.net (internet) and http://mets.dodiis.ic.gov (JWICS). The three ontologies are at http://mets.d2lab.net/onts (internet) and http://mets.dodiis.ic.gov/onts (JWICS).