When Communities of Interest Collide:

Harmonizing Vocabularies Across Operational Areas

C. L. Connors, The MITRE Corporation

Three recent trends have had a profound impact on data standardization within the Department of Defense (DoD), and the ways DoD plans to create, implement, and manage standardized data assets used for information exchange. These trends are the DoD's Net-Centric vision, the adoption of the eXtended Markup Language (XML) family of specifications, and the creation of the Communities of Interest (COIs). In this article, we give an overview of current COI data integration processes, resultant issues and proposed solutions for determining and documenting shared vocabularies. We also discuss potential uses of a software tool called *Harmony* – freely available under a non-exclusive, royalty-free license to U.S. government customers – that provides automated assistance for harmonizing terms across COIs, and a proposed follow-on tool called *Unity* that is uniquely tailored to the needs of COI vocabulary development.

COIs Help Make Data Assets Visible

DoD Directive 8320.2, "Data Sharing in a Net-Centric DoD," states that "Data assets shall be made visible by creating and associating metadata ("tagging"), including discovery metadata, for each asset." To respond to the directive, DoD has stood up a number of COIs, whose activities include the development of vocabularies that support each COI's operational area.

COI vocabularies are represented in XML to provide the "tags" mentioned in the directive - and utilizing XML is one of the foundation steps required to support the Net-Centric Vision. For example, web pages rendered with HTML have tags like <font> embedded that explain to the browser software how the information should *look* on the screen. XML tags go beyond this by helping to define information content. This kind of tagging allows search engines to determine the *meaning* of data contained in tagged XML instance documents and vastly improves the engine's ability to find and return meaningful information to the consumer.

In general, DoD's plan is that each COI will create an initial standardized set of data, associated business rules, and data exchange representations (i.e., "messages" or "information objects"). Once its initial work has been reviewed and approved by a COI governance body, the COI will disband until it is needed again to respond to a new or urgent warfighter need. Resultant artifacts, including data and metadata, will be persisted in catalogs and registries, including the DoD's Metadata Registry (MDR).

Schemas Are Data Exchange Templates

COIs generally need to define specific information exchanges. An example of a critical exchange is the Air Tasking Order (ATO) message that describes the organizations and assets assigned to complete an air mission. The format of any information exchange data set, like the ATO, can be represented in a "schema" by using the W3C XML Schema language. A schema describes the structure and content of a data exchange in XML, including repetition and occurrence information. In general the schema acts as a template for the information exchange

data, and can also be used to validate the content of an XML document (akin to a message), as shown in Figure 1.

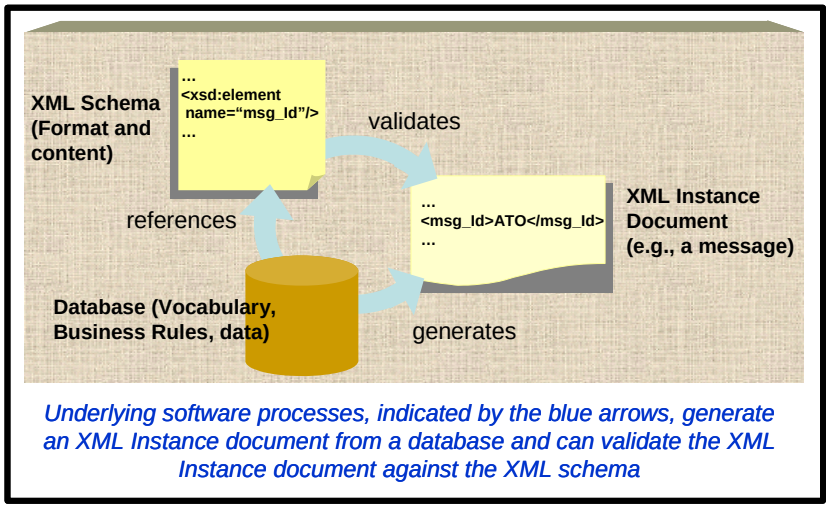
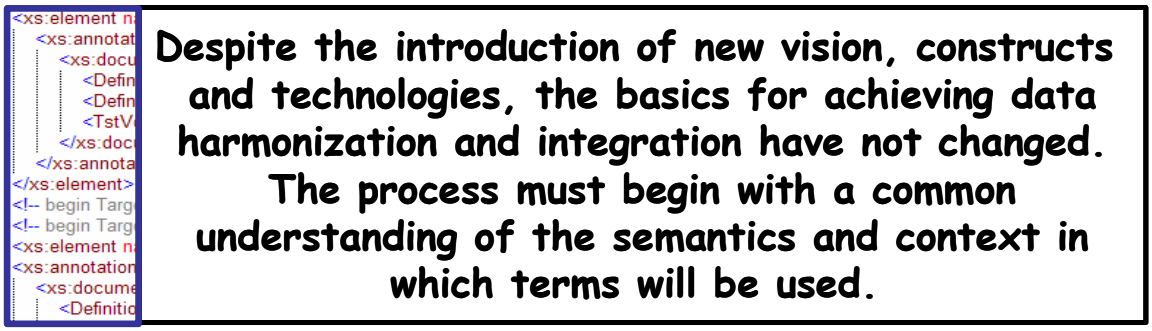


Figure 1. Notional Example of relationship between an XML Schema and XML Instance

Vocabulary as a COI Starting Point

While the ongoing migration to XML is desirable, the current methodology that DoD organizations have adopted for implementing XML has not removed the stovepipes already in place between heterogeneous systems. Despite the introduction of new vision, constructs and technologies, the basics for achieving data harmonization and integration have not changed – the process must begin with a common understanding of the semantics and context in which terms will be used.

With this realization in mind, the United States Air Force (USAF) advocates that any COI under its purview will develop a common vocabulary as its first data oriented activity. This standardized data set must then be used for data exchanges between COI participants and between any two COIs. The end result of this effort is the creation of a common terminology that is used to discover, utilize and integrate disparate data sources in order to successfully exchange information.



Schemas: Raw Material for Vocabulary Construction

For one USAF customer, we assisted the Time-Sensitive-Targeting (TST) COI to develop an initial common vocabulary. We'll narrate the investigative trail that let us to exploit XML schemas for this purpose.

First, we gleaned candidate terms and definitions relevant to TST from DoD publications, manuals and handbooks including "Joint Tactics, Techniques, and Procedures." We also wanted to reuse data from systems currently operating in the TST domain. While traditional products like data dictionaries (that can be very useful for vocabulary construction) are produced by some systems, we discovered the materials are not always available for review.

We discovered that several systems have XML schemas for use in web-based operations, and those aligned with COIs are making these schemas available in the MDR. Since we knew that XML schema embed a COIs terms to describe an exchange, we realized these schemas could be used as metadata source to cull out the COI's vocabulary.

A Time-Consuming Process

As members of the TST Data Panel, our team examined schemas from a number of communities working in the targeting domain, in order to accomplish this culling. It was not a pleasant activity. Because a manual review of the raw XML schemas was difficult, we decided to put the schema information into a database, but this wasn't easy either.

First, we had to first open the schemas using Microsoft Excel, which presented the schemas in a row and column format. In some instances, the schemas formats were so convoluted, we actually started over by first manually redesigning some of the schemas to make them easier to port into Excel.

Once the schemas were imported to Excel, we manually reformatted the data, and then imported it into a Microsoft Access database. With the data in the database, we were able to write queries in order to determine equivalencies between terms from different data sources. It was a lengthy, time-consuming process that we don't ever want to repeat!

However, creating a vocabulary for internal consumption by the participants in a COI was just the tip of the iceberg. We had to do the same work all over again in order to evaluate and then create a common vocabulary for exchanges *between COIs*. These "mappings," which indicate the degree of equivalence between two terms, were created using a one-by-one manual analysis process.

The development of an inter-COI vocabulary is not a trivial or short-term task. Even once an initial vocabulary has been developed, its still difficult to get the COI participants to approve it; impediments include competing needs, differing guidance materials, impact and needs of established systems, personal biases and different operational experiences. For example, arguments ensue over the determination of whether a term should be called "Location" or "Geographical Location" or what the correct type of measurement system that should be used for describing the "altitude" of an object. Another impediment is that engineers have limited software tools for developing a vocabulary and must do their work using time-intensive manual processes.

Harmony: Automated Support for Leveraging Schemas

MITRE's integral role in the development of these vocabularies for several COIs has allowed us to observe – and get hands-on experience with – the difficulty in establishing a COI's initial vocabulary, and the problems in then harmonizing vocabularies between COIs. Because establishing common vocabularies is so important in data integration efforts, MITRE has developed a software tool to simplify this effort, called *Harmony*.

Harmony is a developer's tool for creating a vocabulary by using existing XML or database schemas. The benefits are immediate – instead of the time-consuming process we described above, a user can instead quickly import two schemas and get automated support in determining equivalencies between terms in both schemas. Replacing the existing manual process with this tool will significantly reduce the time required for the analysis.

Harmony Proposes Schema Mappings

Harmony has a user-friendly process for creating the mappings. For the first step in the mapping process, a user imports the two schemas for analysis. *Harmony* represents the schemas on-screen in a side-by-side graphic format. Next, the *Harmony Engine* automatically proposes candidate correspondences between source and target schema elements and overlays these proposed mappings on-screen, as shown in Figure 2.

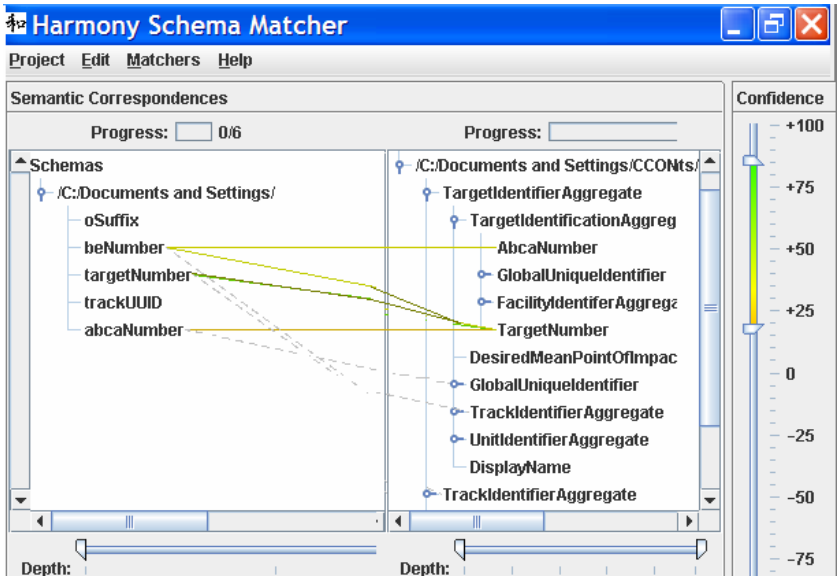


Figure 2. *Harmony* Proposes Schema Mappings

Harmony acts as a natural language processor to propose likely semantic correspondences, using definitions included in the XML schemas in the annotation tags, as well as the parsing the content of the XML tags. The *Harmony* GUI allows an integration engineer to view and edit the set of proposed mappings. A slide bar on the right side of the screen allows an engineer to change the degree of confidence by loosing or tighten the quality of the mappings. When complete, the resultant mappings can be produced as XML instance document, allowing this information to be used outside of the *Harmony* tool.

Harmony was built by The MITRE Corporation as a prototype effort, with an incremental, iterative design approach which provided an immediately useful tool for developers. MITRE provides a non-exclusive, royalty-free license for the Harmony schema matching software for U.S. government purposes only. Please see the "Resources" listed at the end of this article for contact information if you are interested in using *Harmony*.

How Harmony is Unique

While the original *Harmony* is similar to commercial products like BEA's *AquaLogic*, and MetaMatrix's *FOO*, these products do not provide the degree of integration support that meets the sponsor's needs; for example they can't support mapping at the enumeration level as noted; but *Unity* will provide this feature. Also, these tools create executable code in SQL or XQuery, from which a common vocabulary cannot easily be derived.

Unity: Beyond Harmony

COI developers have had the opportunity to evaluate *Harmony*. They provided feedback to the *Harmony* engineers which resulted in a set of recommended improvements. These recommendations include adding annotations and the ability to assign a name to each of the mappings between terms. This latter extension would allow these mappings to be taken outside of *Unity* as configuration managed items so that they can be persisted and updated. Another improvement would be the automatic generation of XSLT – an XML language for creating transformations between terms. For example, if the Air Operations COI uses the term "Altitude" and the TST COI uses the term "Vertical Distance," XSLT could be used to transform the "Altitude" into "Vertical Distance" and vice-versa when for inter-COI exchanges.

Another equally important recommendation is adding the capability to document mappings between information at the most granular level. For example, in addition to mapping between the terms "Type of Aircraft" and "Aircraft Type," data integrators need to map between the enumerated values that support them. But even here, different COIs use different terminology, so that we might also need to map between values like "F-15" and "F15 Eagle."

The developers wish to support these requirements by producing a newer version of *Harmony* called *Unity*, a tool that incorporates the recommendations described above and so is uniquely tailored to the needs of COI vocabulary development.

Summary

The DoD's strategic vision of Net-Centricity – including the COI-based approach for creating and managing data – and the adoption of the W3C XML specifications, has resulted in a need to use XML schemas as a data source for creating a common vocabulary, which acts as the foundation for information exchanges. The current manually-intensive processes that must be used to analyze XML schemas for use within and between COIs could be improved by using a software tool like *Harmony*, which provides automated support for harmonizing terms between any two WC3 XML schemas.

Acknowledgements

The author wishes to gratefully acknowledge the creators of the Harmony software: Peter Mork, Arnon Rosenthal, Len Seligman, Joel Korb, and Ken Samuel.

Resources

Peter Mork, Arnon Rosenthal, Len Seligman, Joel Korb, and Ken Samuel. "Integration Workbench: Integrating Schema Integration Tools," presented at *the Second International Workshop on Database Interoperability* (InterDB '06) at the *IEEE International Conference on Data Engineering*, Atlanta, GA, April 2006.

For additional information on licensing *Harmony*, please contact Mr. Peter Mork at pmork@mitre.org.