A Modeling Methodology to Harmonize Disparate Data Models

(A White Paper to aid intelligence gathering capabilities for the Office of Homeland Security)

Author:

Duane Nickull, duane@xmlglobal.com

Contributors:

Klaus Dieter-Naujok, knaujok@attglobal.net

Matt Mackenzie, matt@mac-kenzie.net

XML Global Technologies, Inc.

Version 1.0

January 2003

Abstract

This paper discusses a solution for the harmonization of disparate data. The logical application of this technology is for the United States Office of Homeland Security to aid with information interoperability, although other applications are foreseeable. The methodologies will facilitate mapping multiple instances of data into a unified view. The unified view of all relevant intelligence information will allow development of future rule and exception capturing techniques to aid the intelligence community in making decisions based on accurate and timely information. This harmonization is an important, yet neglected step, in many proposed architectures for the Office of Homeland Security.

The methodology discussed herein extends existing Object Oriented methodologies and modeling concepts. The main addition is the assertion of a “level of trust” placed on attribute values and trace-ability back to the System Actor who asserts the value. A “System Actor” is a term used to describe either a person or entity that interacts with the system. This attribute trust

can be mapped to specific requirements pertinent to the intelligence community stakeholders who may make strategic decisions based on their interpretation of aggregated intelligence information.

The primary audiences of this white paper are intended to be US Federal Level CIO, CTO and System Architects as well as software vendors providing technology and services for the Office of Homeland Security. Secondary audiences include anyone interested in integrating data from many disparate sources.

1.0 Introduction

Overview

Most initiatives undertaken by the Office of Homeland Security will focus on sharing information between disparate systems.[1] The bulk of current information contains large amounts of knowledge, however it is extremely difficult if not impossible to gain access to that information. Several factors affect this, as noted below.

By and large, most of the information is stored in databases on mainframe computers. Most of these systems are not accessible via network connections due to security concerns. This has hindered development of "real time" data aggregation efforts. Additionally, most of the data models for these systems are not harmonized since most have evolved separately from different sets of requirements. To further complicate matters, many different vendors have implemented most existing systems using different products and information models.

A real time, event-driven and harmonized view of all data is desired. To facilitate this development, a methodology and practice must be used to ensure any future work has a strong foundation of data harmonization to be built upon. The methodologies in this paper incorporate principles and techniques used for global electronic business initiatives and object-oriented programming and data models. This includes the United Nations CEFACT Modeling Methodology[2] (UMM), which relies heavily on the Universal Modeling Language[3] (UML) as the modeling syntax. [1]Unified Modeling Language is a standardized syntax for representing models, owned by the Object Management Group (OMG). UML is widely endorsed as an industry standard for models in software, data, analytical and logic extrapolation.

Use Case

The Office of Homeland Security will aggregate data from many different systems to create a large pool of intelligence, whether centralized or decentralized. The data aggregation will be

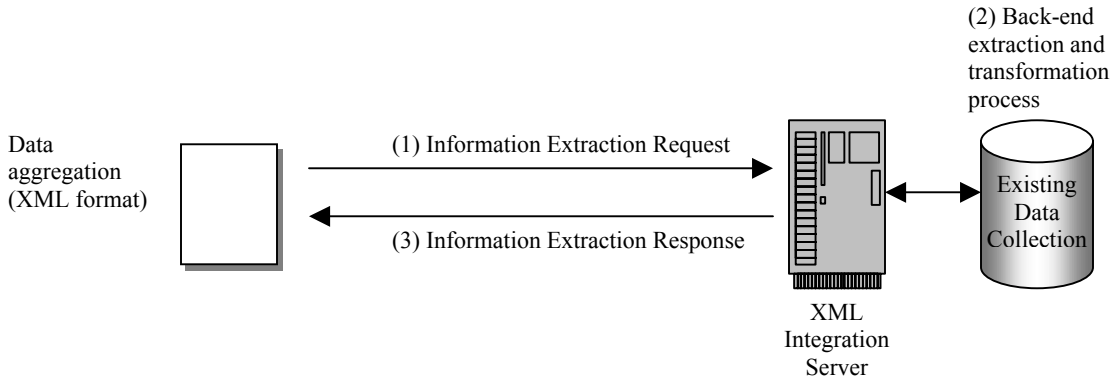
used, perhaps by searching or data mining algorithms, to aide federal agencies in on-going investigations of potential terrorist and other illegal activities. The integrity (accuracy) of the data is of utmost importance. This necessitates a well-planned data harmonization methodology.

Data Serialization and Syntax

Most existing instance data can be accessed via Relational Database Management Systems (RDBMS). A query is sent to the database that triggers result data to be returned. At some point, the result data is serialized into a stream and sent to the actor who requested it.

The eXtensible Mark-up Language[4] (XML) would be the most likely candidate format for data serialization. This offers several advantages. XML provides a common format for data serialization (avoids costly data parsing and interpreting errors), can represent complex, hierarchical data structures, is easily portable, is readable by both machines and humans and is supported by most software vendors.

The figure below is a simple view of an event-based architecture that extracts data out of databases, serializes it as XML and aggregates it into a collection of data.



An XML data integration server receives an Information Extraction Request (1). The request triggers data to be copied from existing data sources, serialized (transformed) into XML (2), then sent back to the data aggregation application that made the request. An alternative scenario may be to have a Listener that sits on the XML Integration Server and polls the data source looking for new transactions (data). There would likely be two types of transactions for that latter scenario – one would be a single batch mode copy or "dump" from an existing database to capture the entire contents followed by a secondary "transactional" extraction based on detecting new data. Transactional data flow may be triggered by a number of factors[5].

This simple architecture does not require much in terms of data modeling if there is a one to one relationship between the data aggregation and the existing data collection. However, if information is to be extracted from multiple data collections into a single unified data aggregation, harmonizing the data models must be done. A common information schema must be developed which represents a "superset" of all the existing data collections. Additionally, the superset schemata must be able to account for all possibilities in attributes to any objects of interest and facilitate a special set of needs that are possibly unique to stakeholders within the entire community.

Given that the number of disparate data structures is in the thousands, it is not realistic to expect that a single group could peruse all the different data structures and derive a super-set. Additionally, many of the existing data models have easily observable flaws in them. The United States Department of Justice Information Model has several differences that should preclude its' use as the data model for the Office of Homeland Security.

2.0 Critical Analysis of Existing Data Models

Current US Department of Justice Data Models

The existing data models were designed in an environment where data sharing was not a fundamental requirement. An entity, such as a person subject, was assigned attributes to identify characteristics of the person. As of late 2002, current data models of a person include the sex, age, weight, eye color, name, nationality, ethnic data, religion and a vast array of other information[6]. The attribute and values of a person's weight seems to be relatively trivial at first. It could be represented in a UML class diagram as follows:

The weight attribute is almost worthless unless you know the date associated with it and information about which *Agency* asserts the weight information. A person's weight, like their hair color, eye color, citizenship or sex, can change over their lifecycle. If there were two records of a single person, both with different values for certain fields, it would be difficult to know which information is the most current without the date association.

The agency relationship is one of the most important since knowing what level of trust can be placed on this information is one of the key stakeholder requirements. The sex attribute value can also be changed over a persons' lifecycle, as can name, nationality, religion, eye color along with most other values. An Office of Homeland Security user of this information will probably not place the same level of trust on a passport issued by an unfriendly government as it would on a passport issued by the United States government.

Another question of how you can positively identify a person must be raised. Most correctional institutes evolved using a person's full, current legal name as the primary identifier. A legal name may be changed (legally within several countries). It can also be easily spoofed and is also non-unique. Currently, biometric information is the favored method of positively identifying a person. Most Department of Justice data models still rely on a model where a primary name (current legal name) is used and there are a series of aliases that may be attributed to a certain individual. Fingerprinting has been successfully used as a biometric means of positive identification but is also not 100% foolproof.

3.0 Object Modeling Methodology

Top-down Object View

The logical places to start any modeling work are by collecting the requirements of stakeholders who will be interfacing with the data. This aids in the development of a data-object model and subsequent refinement of characteristics of each object to facilitate those specific requirements. The United Nations CEFACT Modeling Methodology (UMM N090 R10) would be a good methodology to follow since it allows for an audit trail back to the original set of stakeholder requirements. This paper focuses on the requirements of the intelligence community actors who will rely on this data to offer a wide variety of services, make strategic decisions and recommendations.

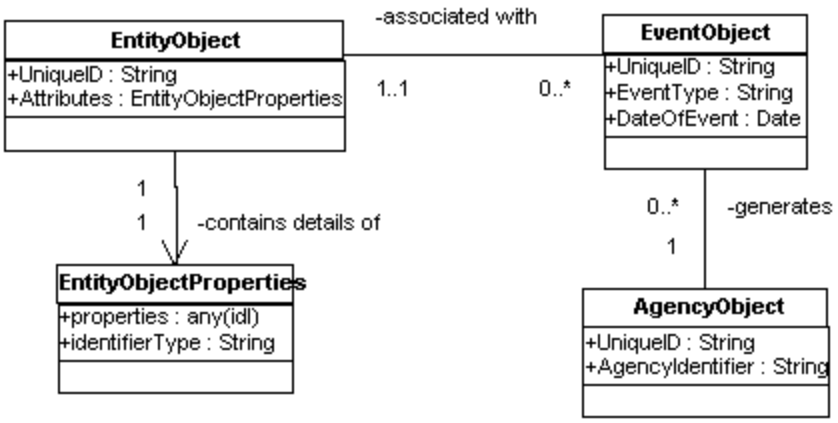
Object Oriented Model View

The highest-level object in most object hierarchies is the abstract *Object* Object. All other objects inherit from the top level abstract *Object* Object. In the intelligence community end-user domain, there are five distinct sub-types of Objects. They are described in the table below. Each of these five sub-objects have *type* attributes associated with them to further distinguish instances of these objects.

Object Type	Description	Examples
Entity Objects	Entities (nouns) as defined by	Types include Persons,

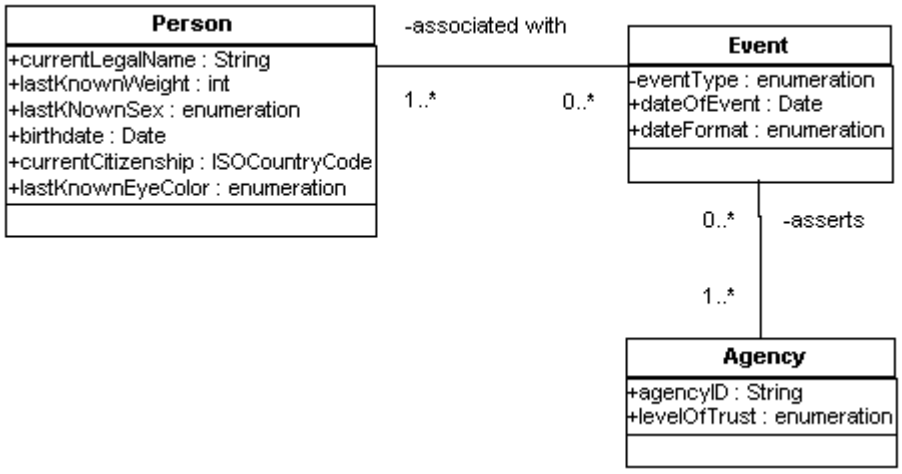
	stakeholders in the intelligence community.	Substances, Shipping Containers, Real Property, Information.
Event Objects	An Event (verb) is always associated with both an Entity Object Instance and an Agency Object (generated by).	Arrest: A Person (Entity Object) was arrested by an Agency Object instance.[7]
Agency Objects	Entities that generate Event Objects, generate or consume information	Local Law enforcement Agency, FBI, Canadian Border Patrol
Process Objects	Processes relevant to investigations, actions and other items that may trigger information flow about entity or Event objects to be shared between or within an Agency.	Arrest of an Entity Object of type “Person” may trigger a notification to a Federal Agency to check for prior arrests on other jurisdictions.
Condition Objects	Conditions are Rules that can trigger an Event or affect a Process.	If the subject of an arrest is a foreign national, an embassy MUST be contacted within 24 hours.

Event Objects are *always* associated with Agency and Entity Objects and must have a *date* attribute. This is to meet the requirements of the intelligence community actors who use the data. The association to an Agency Object instance is essential for determining the level of trust that may be given to information contained in an Event Object Instance. This simple relationship is expressed in the UML class diagram below.



The multiplicity is fairly simple – an Entity Object instance can have zero or more events associated with it. An Event Object Instance must be associated with at least one Entity Object. Entity, Event and Agency Objects all have a unique ID (UID) attribute to them. The unique ID must be of a type and granted by an agency that is intrinsically trustworthy and impossible to fake in order to maintain the integrity of the data.

Using this methodology, our previous model could be expressed as such:

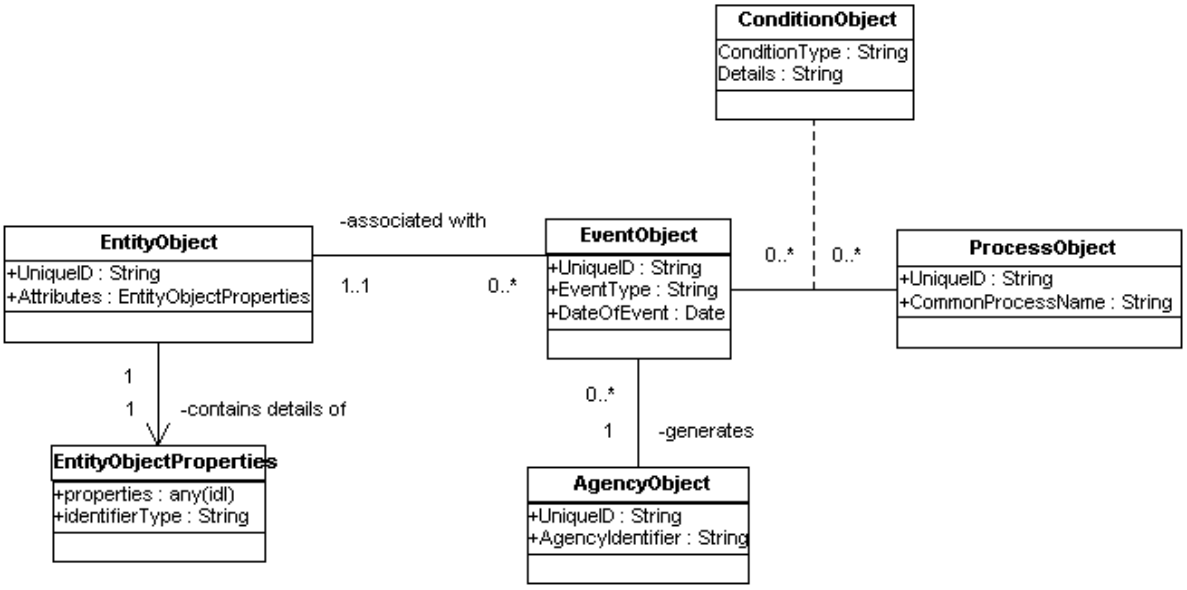


Note that the *Person Object* is now associated with an *Event Object* for each of the recorded events. An Event cannot exist without an Agency. The navigability allows the consumer of the information to ascertain which Agency is associated with the event, hence an implied level of trust can be asserted.

The association of an *Event* to a *Person* Object may be assigned or re-assigned by any agency. Other agencies may then weigh the integrity of that assignment base on a review of the facts or by assessing the credentials of the agency making the assertion.

In a real world instantiation, if an event of type “Crime” happens and the suspect is unknown, an “unidentified” Entity Object instance of type “person” (Person Object) may be created and properties may be associated with that instance based on known facts (possible eye witness descriptions, photographs, DNA from blood or semen samples). This anonymous Person Object instance may then be later identified as a known Person Object instance. At that time, the Object-to-Event association may be re-assigned to the identified object instance.

The next step is to add the two final object types to our model. Both *Process Objects* and *Condition Objects* affect how information may be used and may act as triggers for certain downstream processes such as sharing information, assigning resources to investigate *Event Objects* or making associations between events and people. While the *Process Object* class is a straightforward class that has an effect upon instances of other classes, the *Condition Objects* are really an association class. Our previous model may now be shown as follows:



The above model would need to be expanded to be implement-able. Such expansion would have to include details that would be of interest to the stakeholders who will mine the intelligence from

the data aggregation to base decisions upon. In the above figure, note how the association object class *Condition Object* constrains the processes that may be used based on rules. For an example, if a subject of arrest is under 19 years of age, a process of notifying the persons' legal guardians must be started within a certain number of hours after the arrest. The entire range of applicable conditions and rules must be present if someone were to automate law enforcement functionality.

XML schemas are available for the above data models upon written request.[8]

More on Entity Objects

Entity Objects will likely be the primary subjects of interest to those fighting terrorism. Entity Objects may be modeled to represent the entire range of objects of interest for all stakeholders of the information. The needs of the intelligence community are diverse and the list may be highly confidential. The following list represents a starter list of Entity Object types.

Object Type	Description	Examples
Person	Any person of interest to stakeholder agencies.	George Washington, Saddam Hussein, Osama Bin Laden
Substance	A substance that is of interest to stakeholder agencies	Radioactive substances, controlled or illicit pharmaceutical compounds
…		

Unique Keys for Entity Objects

Past Department of Justice Information models are not candidates for re-use since they often rely on a legal name as a unique key – a credential that can be easily faked or changed in the life-cycle of an *Entity Object* instance. Several investigations have been conducted into the use of non-intrusive, biometric authentication techniques for establishing a unique key for entity objects of type "person".[9] The leading candidates appear to be palm/fingerprint, facial recognition and retinal scan. Most other types have either been deemed too intrusive to civilians or have unacceptable False Positive Rates (FPR) or False Rejection Rates (FRR).

Summary

This paper briefly outlined a modeling methodology and object model for use within the Office of Homeland Security data aggregation architecture(s). In order to implement this methodology, modeling analysts should probably work on a pilot project to examine the full scope and breadth of issues and collect formal requirements. Most estimates have the number of agencies at around 22 and the number of data acquisition points at over 170,000. In any event, a carefully thought out modeling methodology must be used if the resultant data aggregation is going to be able to accommodate current and future requirements.

The author and many of the contributors of this white paper are actively involved in on-going work in the area of data harmonization, largely related to building global Business to Business (B2B) exchanges. Most of this work can be directly ported for use within any large integration projects.

Appendixes

Relationship to Existing Government Methodologies

The E-Government Act of 2002, enrolled as HR 2458, contains a variety of specific provisions relevant to information discovery and interoperability. The thrust of the law is clearly meant to accelerate the ongoing work that has been pioneered, and meshes nicely with recent developments in standards arenas. Although the U.S. Federal GILS law in Title 44 Section 3511 is not directly affected, it offers relationship opportunities to promote common interoperability solutions. The following specific examples are work related to (both directly and indirectly) the newly formed Office of Homeland Security architecture for sharing of vital information.

Section 207 calls for the establishment of an "Interagency Committee on Government Information". Within two years, the Committee is to recommend standards for the categorization of Government information in a way that is searchable electronically and interoperable across agencies. Within a year after Committee recommendations are made, the Office of Management and Budget (OMB) and the Archivist of the United States are to issue Federal policy requiring compliance and laying out. Due to the current political climate, a substantial budget has already been allocated for integration projects within the Office of Homeland Security.

Section 212 directs OMB, within three years and in consultation with agencies, the regulated community, public interest organizations, and the public, to study and report to Congress on progress toward integrating Federal information systems across agencies. The study is to address the integration of data elements used in the electronic collection of information and the

feasibility of software tools for assembling, documenting, and validating the information. It is also to address the feasibility of a distributed information system that provides access to information integrated across participating agencies. This data integration study effort is to be informed by a series of no more than 5 pilot projects.

Section 214 requires a research and implementation strategy on using information technology to enhance crisis preparedness, response, and consequence management of natural and manmade disasters. Harmonization of data elements is a critical component of that strategy.

Section 216 codifies under law the long-standing and on-going work of the Federal Geographic Data Committee in facilitating the development of common protocols for the development, acquisition, maintenance, distribution, and application of geographic information.

Resources and further reading:

(1) United Nation Modeling Methodology draft TMWG/N090, Sept 2002 – http://www.unece.org/cefact/drafts
(2) XML Global Technologies, Inc – http://www.xmlglobal.com
(3) Autonomy - http://www.autonomy.com/Content/Press/Archives/2002/1021

Footnotes:

[1] The President's Plan to Strengthen Our Homeland Security, Feb 4, 2002 - http://www.whitehouse.gov/news/releases/2002/02/20020204-2.html
[2] United Nation Modeling Methodology draft TMWG/N090, Sept 2002 – http://www.unece.org/cefact/drafts
[3] Object Management Groups' UML - http://www.omg.org/gettingstarted/what_is_uml.htm
[4] W3C XML Specification v 1.1 transitional - http://www.w3.org/XML/
[5] See more on "Transactional Data Flow Triggers" in the related Intelligence Data Aggregation Technical Architecture, Duane Nickull, 2002-3.
[6] United State Department of Justice Information Models – http://www.ojp.usdoj.gov.bjs
[7] This model is logically extensible to accommodate an expanded scope. It would be easy to add in events such as "was in XXXX country on XXXX date" or "Purchased XXXX substance".
[8] Send email with details of request to duane@xmlglobal.com
[9] http://ctl.ncsc.dni.us/biomet%20web/BMCompare.html