Analytic Technology™ Industry Roundtable

Upper Ontology for the Analytic Technology Industry Roundtable's Analysis Exchange Model 1.0

'Approved for Public Release; Distribution Unlimited. Case Number 17-4108'.

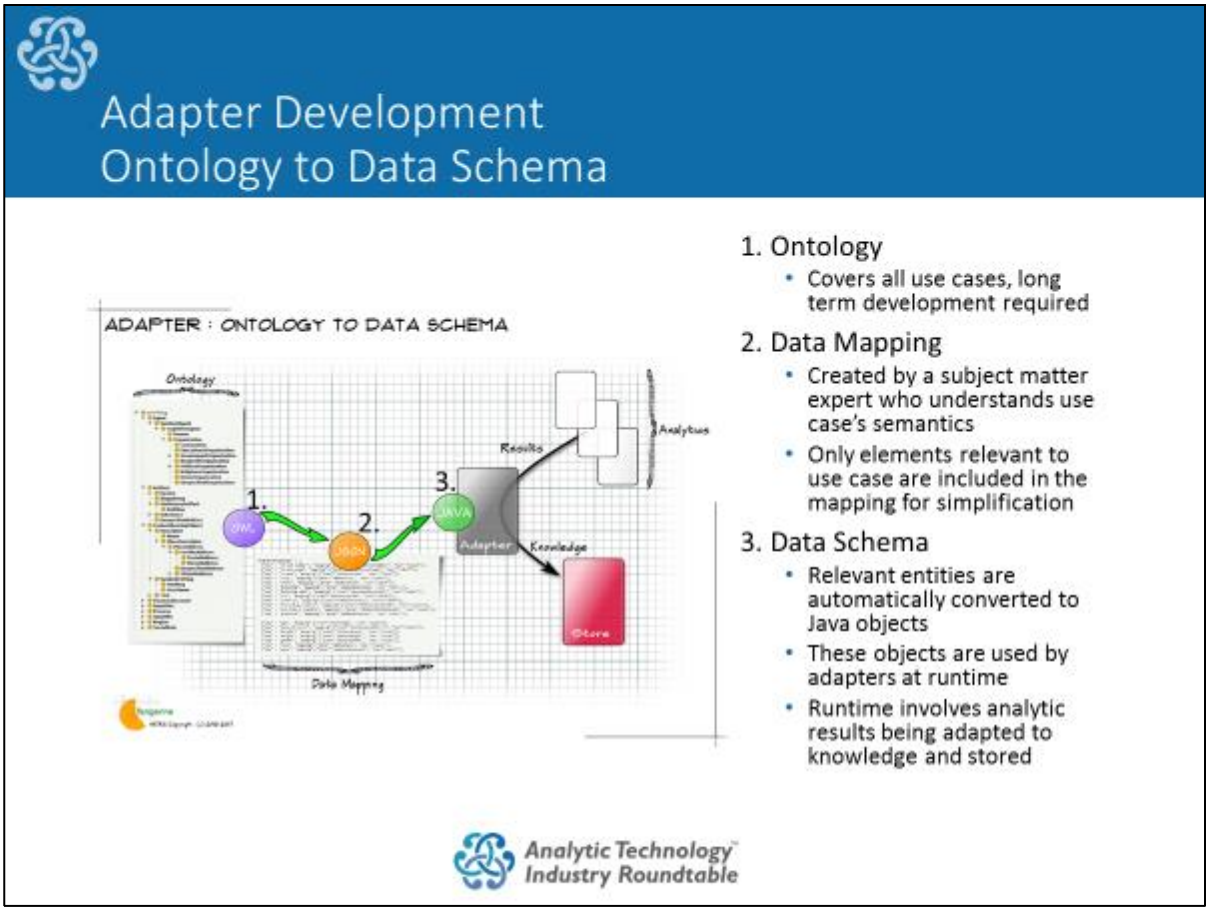
1. Introduction

The Roundtable's Analytic Exchange (AE) Model 1.0 is chartered with the goal of facilitating the exchange of information between different analytic and analysis vendors. The Analysis Exchange Model 1.0 was made available as an open source technology via GitHub in November 2017.[1] MITRE and the Roundtable Architecture Working Group are studying the feasibility of using an ontology as the basis for unifying information, to accomplish this technical goal and explore other technical topics of interest. This document describes the upper portion of the Analysis Exchange Model 1.0's ontology and how it is used to support integrating information from different vendors.

Error! Reference source not found. depicts the role of the Analysis Exchange Ontology. The ontology will provide the classes and properties that model the semantics of concepts in select vendor analytics. The ontology will be used as a guide for creating mappings for individual vendor models such that a consistent knowledge base can be established with content from different vendors. In addition to containing classes and properties, the ontology will contain rules that model semantics which cannot be modeled via the ontology modeling language and axioms that are selected for expressing the ontology. With the classes, properties, and rules, the consistent

[1] https://analytic-roundtable.github.io/

knowledge base can then be queried to retrieve information from any analytic using a consistent querying mechanism.



Adapter Development
Ontology to Data Schema

1. Ontology
   - Covers all use cases, long term development required
2. Data Mapping
   - Created by a subject matter expert who understands use case's semantics
   - Only elements relevant to use case are included in the mapping for simplification
3. Data Schema
   - Relevant entities are automatically converted to Java objects
   - These objects are used by adapters at runtime
   - Runtime involves analytic results being adapted to knowledge and stored

Figure 1: Role of the Analysis Exchange Ontology

It is often recommended that existing ontologies should be reused when developing a new ontology. Well-developed preexisting ontologies can provide valuable content and reduce the burden of developing a new ontology. The Analysis Exchange Ontology is based on the Suggested Upper Merged Ontology (SUMO). Whenever possible, concepts from SUMO are used to model vendor content. SUMO was selected because it was created by a team of accomplished ontologists.

This document is a snapshot of the current state of the Analysis Exchange Ontology as it supports the Analysis Exchange Model 1.0. Because the project is evolving and expanding into next year as work moves toward achieving Analysis Exchange Model 2.0, there may be changes to the upper ontology, its content, and our approach to support the new requirements emerge. As these changes are introduced and solidified, new iterations of this document detailing the changes and their rationale will be released.

2. Requirements

This section presents the various motivations for the ontology. The Analysis Exchange Ontology is intended to serve as a consistent set of data semantics for analytic results. It is a central ontology where the semantics of external analytic stacks' results are aligned such that ambiguities between different representations are mitigated. Because it is impractical to develop a comprehensive ontology for all use cases at the outset, the Analysis Exchange Ontology is designed to be extensible to new use cases, while it supports several from its first release.

2.1. Use Cases

Three of the initial use cases that motivate the creation of the Analysis Exchange Ontology are listed below with brief descriptions. The Analysis Exchange Ontology has to contain classes and properties that model entities that participate in the use cases.

Fraud Use Case: This use case is based on the submissions of fraudulent vouchers. In particular, we focus on a travel voucher scenario in which patients may request reimbursement for travel to facilities for medical treatment. In this scenario, the medical institutions need to recognize vouchers that are legitimate and avoid funding those that are not. Fraudulent vouchers may be submitted by travel providers that offer transportation services for patients that are unable to commute independently. The fraud may exist as excessive fees, incorrect distances traveled, incorrect dates of travel, and non-existent travel events among others.

Cyber Forensics Use Case: This use case is based on assessing cyber events. In particular, we focus on "reactive" analysis and "proactive" analysis. A cyber event may target a computer network using malware or a denial of service attack. Cyber analysts may want to identify the perpetrators, determine the extent of the attack, assess data loss, and identify remediation measures. Additionally, given the increase in cyber criminals and cyber events, cyber analysts are tasked with defending against future cyber events by, among others, reviewing available resources and actors who may be targeting organizations.

Threat Analysis Use Case: This use case is based on gauging the threat posed by terrorist organizations. In particular, we focus on supporting capabilities that reduce the costs of doing complex data exploration and increase the effectiveness in identifying trends and patterns of targeted violent events over time and space. In this scenario, terrorist organizations engage in attacks and government organizations want to identity and assess threats as soon as possible and neutralize specific threat actors.

2.2. Competency Questions

Competency questions are often used in the ontology development process. The questions can be used to assess how well the ontology supports its original goals. A few competency questions are listed below.

Q1: How many vouchers did PersonX submit between January 1, 2017 & January 31, 2017 that included travel distances greater than 50 miles?

- Concepts: Voucher, Event, Human, Address, travel distance

Q2: Has PersonX submitted requests for toll reimbursement despite the fact that the shorter routes between his home and the destination facility does not contain toll routes?

- Concepts: Voucher, Address, Fee, Facility, Route, travel distance

Q3: Who perpetrated the cyber attack?

- Concepts: Human, Event, CyberAttack, agent

Q4: What is the extent of the cyber attack and which systems have been affected?

- Concepts: CyberAttack, Computer, Damage

Q5: How many people died as a result of the terrorist attack?

- Concepts: Event, TerroristAttack, Organization, Human, perpetrator, victim, death quantity, causality

Q6: What is the distribution of weapons used in terrorist attacks between June 1, 2017 and August 1, 2017?

- Concepts: Event, TerroristAttack, Weapon, event date

3. Ontology Curation

The Suggested Upper Merged Ontology (SUMO) [1] was selected as the basis for the Analysis Exchange Ontology, whose upper levels are shown in **Error! Reference source not found.**. SUMO was selected for a variety of reasons. (1) It is an extensive ontology covering domains such Finance, Transportation, IT, and Military. The Basic Formal Ontology (BFO) [2] was influenced by SUMO. As such, the AE has a direct path to integrating with the variety of ontologies that are based on BFO since SUMO and BFO are compatible. This decision reduces the effort to align the AE with other ontologies based on SUMO or BFO in the future. (2) Every concept in SUMO is documented. The documentation is useful when searching for a SUMO concept to use for a vendor concept. (3) WordNet (WN) synsets are aligned with SUMO [1]. The benefit of the WN-SUMO alignment is that a vast array of words in WN can be used when searching for SUMO concepts.

SUMO is a large upper model ontology. To minimize the size of the Analysis Exchange Ontology, the team decided to not "import" SUMO via owl:import directives but rather to selectively extract subtrees in SUMO to support the AE conceptual needs. The advantage of this approach is that the

Analysis Exchange Ontology is as compact as it can be while supporting the needs of the AE. The disadvantage of this approach is that the AE could become out of sync with SUMO in the future.

We believe that the severity of the synchronization issue will not be great since SUMO is an established ontology and the core is not likely to change significantly.

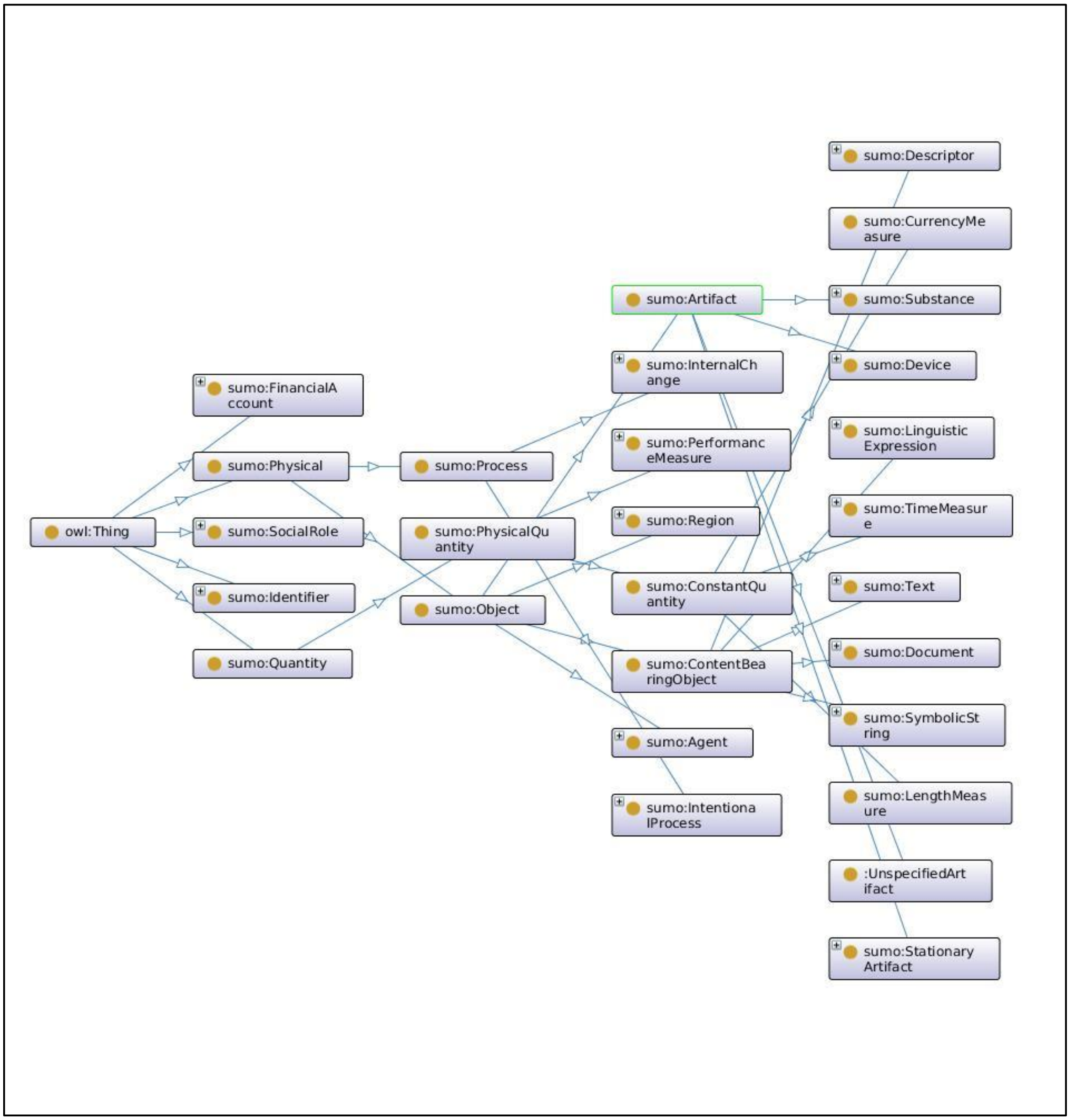


Figure 2: Top layer of the Analysis Exchange Ontology

Since the Analysis Exchange Ontology is based on SUMO, the Analysis Exchange Ontology follows the SUMO naming conventions as closely as possible. Concept URIs are expressed in camel case. Classes begin with capital letters while properties begin with lower case letters. Every concept is documented. The documentation includes the source of the documentation. The concepts that were extracted from SUMO are annotated with the original SUMO documentation. Concepts that were not extracted from SUMO include documentations from authoritative sources whenever possible or are based on the team's interpretation when authoritative sources could not be identified. Whenever possible properties are expressed as has{X} where X refers to a class in the Analysis Exchange Ontology or a synonym of a class in the Analysis Exchange Ontology.

Concepts in the Analysis Exchange Ontology exist in one of two different namespaces. Concepts that were extracted from the upper model are prefixed with the prefix for the upper model. Concepts that were added because the upper model did not contain viable option are prefixed with the 'AE' prefix. This naming convention facilitates tracking the origin of concepts in the Analysis Exchange Ontology.

At this phase of development, the Analysis Exchange Ontology is driven by the vendor models. It is far too difficult to predict in advance the concepts that will be needed to facilitate vendor fusion. The Analysis Exchange Ontology development process is thus evidence driven. The process begins with reviewing vendor documentation and sample data to gauge the semantics of the vendor data. The vendor product is often chosen to support a specific use case. As such, the use case layer indirectly influences modelling decisions. For each item in the vendor model, the upper model ontology is reviewed to determine if there exists a concept in the upper model that sufficiently characterizes the semantics of the vendor item under review. If such a concept exists it is added to the Analysis Exchange Ontology. The advantage of preferring concepts from the upper model is that we maintain consistency with the upper model whenever possible. On the other hand, if it is determined that the upper model does not contain a sufficient concept a new concept is added to the Analysis Exchange Ontology. Prior to adding a new concept to the ontology, an anchor for the new concept must be identified. The anchor is selected by searching the upper model for the most specific concept that is a super class or super property of the new concept to add. The search is based on documentation of the vendor item, if available, and the assumptions that are made about the example data compared to the documentation of the upper model concepts. The search for an upper model anchor concept includes searching the WN-SUMO alignment for potential SUMO concepts. The vendor concept is searched in WN, an appropriate synset is selected, and the SUMO mapping for the synset is considered. If an appropriate WN matched is found, the corresponding SUMO concept is selected as an anchor. A new concept is created and placed “under” the SUMO or AE anchor that was previously identified. The new concept is added using the naming conventions that have been agreed upon. The new concept is documented based on the concept documentation conventions that have been agreed upon.

An example of a vendor item that maps directly to SUMO is “entity:address” in the NetOWL model. sumo:PlaceAddress was selected as the class for “entity:address”. The ancestors of sumo:PlaceAddress are also added to the Analysis Exchange Ontology.

An example of a vendor item that required a new concept is “entity:artifact:weapon:explosive” in the NetOWL model. AE:ExplosiveWeapon was created for “entity:artifact:weapon:explosive”. AE:ExplosiveWeapon is anchored by sumo:Weapon. The ancestors of sumo:Weapon as added to the Analysis Exchange Ontology.

A major challenge in creating the Analysis Exchange Ontology is maintaining conceptual consistency. Ideally, and ontology is created to facilitate machine interpretation of information. Consistent modelling and interpretation facilitate consistent retrieval of information. To achieve consistency, we adopt a policy of *modelling the meaning of data and not the structure of data*. The structure refers to formats in which the data is presented and the labeling policies. Upon inspecting the vendor models every attempt is made to understand the semantics of the content. The vendor content is then integrated with the existing ontology such that if two or more vendors provide the same “type” of information, the information will be modeled similarly in the Analysis Exchange Ontology. This decision reduces the complexity of extracting information and facilitates consistent machine interpretation of information.

An example of maintaining consistency can be found in the mapping of NetOWL "link:person:person_transient_in". This NetOWL item models a travel event as a link between a person and a destination. The Analysis Exchange Ontology adopts the Neo-Davidsonian [4] model of events. Events are modeled as Event instances with thematic roles. The NetOWL "event" items are not modeled according to the Neo-Davidsonian model. To maintain consistency across all events the NetOWL "link:person:person_transient_in" is modeled as an event with *agent* and *destination* roles. The origin of the travel event is unknown. This decision ensures that all events are consistently represented and knowledge bases derived from NetOWL can be consistently queried to retrieve events associate with a particular entity. If the NetOWL content was modeled as it is structured, there would exist two different event models in the Analysis Exchange Ontology.

The Analysis Exchange Ontology is constructed with the goal of supporting rule based inferencing in the future. As the ontology develops every attempt is made to ensure that the inferencing rules will be consistent with new content when appropriate.

4. Core Concepts

This section contains brief descriptions of core concepts. Core concepts are those that are not specific to any particular use case but are needed to support use cases or are needed as anchors for concepts that were added for specific use cases.

4.1. Classes

4.1.1. Core Classes

A partial listing of some of the core classes are presented. The classes include descriptions attributed SUMO or to other sources if a SUMO description was available. In some cases, example usage patterns are provided as a guide for the intended uses of the concepts.

```
sumo:CognitiveAgent
```

- @SUMO "A CognitiveAgent is an Agent that has the ability to reason, deliberate, make plans, and experience emotions. Although Human is a subclass of CognitiveAgent, there may be instances of CognitiveAgent which are not also instances of Human. For example, Primates, dolphins, whales, and some extraterrestrials (if they exist) might be considered CognitiveAgents." [1]

`sumo:Device`

- @SUMO "A Device is an Artifact whose purpose is to serve as an instrument in a specific subclass of Process." [1]

`sumo:TransportationDevice`

- @SUMO "A TransportationDevice is a Device which serves as the instrument in a Transportation Process which carries the patient of the Process from one point to another." [1]

`sumo:Weapon`

- @SUMO "The Class of Devices that are designed primarily to damage or destroy Humans/Animals, StationaryArtifacts or the places inhabited by Humans/Animals." [1]

`sumo:Descriptor`

“sumo:Name” is a subclass of sumo:Descriptor. Names are typically designators for Human and Organizations but are used with other entity types. A Name is intended to function as a container for a name grammar. For example, a Human identified by Human_1 can have the name “John Leroy Smith”. This example will result in the following assertions:

- `Human_1 rdfs:type sumo:Human`
- `Name_1 rdfs:type sumo:Name`
- `Human_1 rdfs:label “John Leroy Smith”`
- `Human_1 AE:hasName Name_1`
- `Name_1 rdfs:label “John Leroy Smith”`
- `Name_1 AE:hasFirstName “John”`
- `Name_1 AE:hasMiddleName “Leroy”`
- `Name_1 AE:hasLastName “Smith”`

In this example, the instances Human_1 and Name_1 both have labels. In the AE every instance has a label that contains a human interpretable representation of the instance. The sumo:Name instance (Name_1) allows us to model the individual components of name. This modeling decision supports searching for entities by any name component.

`sumo:Document`

- @SUMO "Instances of Document are ContentBearingObjects that are intended to convey propositional content via Text (LinguisticExpressions, seen or heard), Images, or some combination of these (e.g., an audio clip included in an electronic document consisting mostly of VisualText and some Images). Formally, a Document constitutes any ContentBearingObject that is an Artifact conventionally typically intended to be transmitted and assimilated as a meaningful whole. An Article or a Book would be a Document, but a Word or Paragraph typically would not." [1]

`sumo:LinguisticExpression`

- @SUMO "This is the subclass of ContentBearingPhysical which are language-related. Note that this Class encompasses both Language and the elements of Languages, e.g. Words." [1]

`sumo:Text`

- @SUMO "A LinguisticExpression or set of LinguisticExpressions that perform a specific function related to Communication, e.g. express a discourse about a particular topic, and that are inscribed in a CorpuscularObject by Humans." [1]

`sumo:IntentionalProcess`

- @SUMO "A Process that has a specific purpose for the Agent who performs it." [1]

4.1.2. Events

Events in the Analysis Exchange Ontology are influenced by the Davidsonian model of events [4]. In the AE, an event is an occurrence in space and time. The event instance is a container for the event details. The event details are specified using thematic roles. A listing of common thematic roles extracted from Jurafsky and Martin, Chapter 22,[2] is provided below:

- agent → “The volitional causer of an event”
- theme → “The participant most directly affected by an event”
- patient → “The participant undergoing change of state, causally affected by another participant”
- experiencer → “The experiencer of an event”
- instrument → “An instrument used in an event”
- source → “The origin of the object of a transfer event”
- goal → “The destination of an object of a transfer event”

Ideally, every event has a location and a time.

Given the example, “John and Mary met in Washington DC on January 1, 2017 at 1:00 PM to discuss their projects” and event instance would be created and represented as:

- `Event_1 rdfs:type sumo:Meeting`
- `Event_1 rdfs:label “John and Mary met in Washington DC...”`
- `Event_1 AE:hasAgent Person_1 (John)`
- `Event_1 AE:hasAgent Person_2 (Mary)`
- `Event_1 AE:hasLocation Location_1 (Washington DC)`
- `Event_1 AE:hasTime “2017-01-01 13:00:00”`

Generic thematic roles are used with events for consistency. Specific role types can be inferred with rules. In the example “The pirates attacked the crew” the following event instance will be created:

- `Event_2 rdfs:type sumo:Attack`
- `Event_2 rdfs:label “The pirates attacked the crew”`
- `Event_2 AE:hasAgent Person_4 (Pirates)`

[2] https://web.stanford.edu/~jurafsky/slp3/22.pdf

  - Event_2 AE:hasPatient Person_5 (Crew)

In this example, the pirates are the attackers and the crew is the victim. Rules can be used to model these specific event roles. The following rules are expressed in F-Logic.

- ?Event[hasAttacker→?Attacker] :- ?Event : sumo#Attack, ?Event[AE:hasAgent→?Attacker].
  - The attacker role of an event is filled by the agent of an Attack event.
- ?Event[hasVictim → ?Victim] :- ?Event : sumo#Attack, ?Event[AE:hasPatient → ?Victim].
  - The victim role of an event is filled by the patient of an Attack event.

sumo:Identifier

- @SUMO "Identifier is the Class of ContentBearingObjects that identify some entity, such as a bank account, a person, or a location (e.g., as identified by a specific street address or GPS coordinates), perhaps uniquely under some circumstances." [1]

sumo:TimeMeasure

- @SUMO "The class of temporal durations (instances of TimeDuration) and positions of TimePoints and TimeIntervals along the universal timeline (instances of TimePosition)." [1]

sumo:LengthMeasure

- @SUMO "A subclass of ConstantQuantity, instances of which are measures of length." [1]

sumo:CurrencyMeasure

- @SUMO "Instances of this subclass of ConstantQuantity are measures of monetaryValue stated in terms of some UnitOfCurrency such as UnitedStatesDollar, UnitedStatesCent, RMB, Lira, Yen, etc." [1]

Instances of sumo:CurrencyMeasure should have two properties. The numericValue predicate is used to express the values of the instances. hasUnitOfMeasure is used to express the type of currency. A predefine list of measures will be created. The numericValue property and the hasUnitOfMeasure property should also be used with instances of sumo:LenthMeasure.

`sumo:SocialRole`

- @SUMO "The Class of all Attributes that specify the position or status of a CognitiveAgent within an Organization or other Group." [1]

4.1.3. Other Classes for Multiple Use Cases

The following list of classes are significant examples included in the core because they are needed to model content extracted from NetOWL. The NetOWL model is considered use case agnostic. Many of the entity types in NetOWL are used in several cases. For this reason, it was decided that concepts that are needed for NetOWL will be included in the AE core.

`sumo:InternetAddress`

- @Wikipedia “An Internet Protocol address (IP address) is a numerical label assigned to each device connected to a computer network that uses the Internet Protocol for communication. An IP address serves two principal functions: host or network interface identification and location addressing.”[3]

`sumo:Artifact`

- @SUMO "An Object that is the product of a Making." [1]

`sumo:FinancialAccount`

- @SUMO "A formal banking, brokerage, or business relationship established to provide for regular services, dealings, and other financial transactions." [1]

`sumo:Organization`

[3] https://en.wikipedia.org/wiki/IP_address

- @SUMO "An Organization is a corporate or similar institution. The members of an Organization typically have a common purpose or function. Note that this class also covers divisions, departments, etc. of organizations. For example, both the Shell Corporation and the accounting department at Shell would both be instances of Organization. Note too that the existence of an Organization is dependent on the existence of at least one member (since Organization is a subclass of Collection). Accordingly, in cases of purely legal organizations, a fictitious member should be assumed." [1]

`sumo:Region`

- @SUMO "A topographic location. Regions encompass surfaces of Objects, imaginary places, and GeographicAreas. Note that a Region is the only kind of Object which can be located at itself. Note too that Region is not a subclass of SelfConnectedObject, because some Regions, e.g. archipelagos, have parts which are not connected with one another." [1]

4.2. Constants

Constants are instances in an ontology that provide representative instances of classes. Ideally, the constants should be used when creating instances of those classes for which the constants exist. For example, when mapping analytic results to the AE, instances in the analytic results that map to constants in the AE should use the URI's for the constants as opposed to generating new instances. Examples of the good candidates for constants include the continents of the world, well known artifacts such as monuments, and historical events. These types of instances are statics and unambiguous. Whenever necessary, the AE will contain constants of classes.

5. Ontology Language

The Analysis Exchange Ontology is expressed in Web Ontology Language (OWL). The ontology uses the following OWL axioms:

- `owl:subClassOf`

- `rdfs:type`
- `owl:ObjectProperty`
- `owl:DatatypeProperty`
- `rdfs:comment`

The ontology is translated to F-Logic to be used in an F-Logic inspired architecture. The translation to F-Logic retains the semantics of the classes and properties in the ontology because the OWL axioms that are used are limited to those that can be implemented in F-Logic. Examples of OWL to F-Logic translations are presented below:

```
SubClassOf(sumo:Human sumo:CognitiveAgent) → sumo#Human ::
   sumo#CognitiveAgent.
```

Note that in OWL, the namespace prefix is denoted with “:”. However, in F-Logic “:” is the equivalent of rdfs:type. In the Analysis Exchange Ontology, the “#” symbol is used to denote the namespace prefix.

```
ClassAssertion(sumo:Human Person_1) → Person_1 : sumo#Human.
```

6. References

1. Suggested Upper Merged Ontology (SUMO) http://www.adampease.org/OP/
2. Basic Formal Ontology (BFO) http://ontology.buffalo.edu/bfo/
3. Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) http://www.loa.istc.cnr.it/old/DOLCE.html
4. Briscoe, T. Introduction to Computational Semantics for Natural Language https://www.cl.cam.ac.uk/teaching/1213/L107/semantics2.pdf