## SUMO Overview

We will now provide an overview of SUMO. More detail is given later in the section "SUMO Details and Examples" on Page 120. A somewhat simplified version of the SUMO hierarchy is shown in Figure 56. The root node of the SUMO is, as in many ontologies, **Entity**, and this concept subsumes **Physical** and **Abstract**. The former category includes everything that has a position in space and time, and the latter category includes everything else.

```
Entity
  Physical
    Object
      SelfConnectedObject
        Substance
        CorpuscularObject
          OrganicObject
          Artifact
      Collection
      Agent
      Region
    Process
      DualObjectProcess
      IntentionalProcess
      NaturalProcess
      Motion
      InternalChange
  Abstract
    Quantity
      Number
      PhysicalQuantity
    Attribute
    SetOrClass
    Relation
    Proposition
    Graph
    GraphElement
```

*Figure 56: Top of the SUMO structure*

Under the concept of **Physical**, we have the disjoint concepts of **Object** and **Process**. The existence and nature of the distinction between these two notions was the subject of much debate. According to those who adopt a 3D orientation (or "endurantists", as they are sometimes called), there is a basic, categorical distinction between objects and processes. According to those who adopt a 4D orientation (the "perdurantists"), on the other hand, there is no such distinction. The 3D orientation posits that objects, unlike processes, are completely present at any moment of their existence, while a 4D orientation regards everything as a space-time worm (or a slice of such a worm). On the latter view, paradigmatic processes and objects are merely opposite ends of a continuum of spatio-temporal phenomena. The current version of the SUMO embodies a 3D orientation by making **Object** and **Process** disjoint siblings of the parent node **Physical**.

Immediately under the concept of **Object**,

there are two disjoint concepts: **SelfConnectedObject** and **Collection**. A **SelfConnectedObject** is any **Object** whose parts are connected with one another.

The concept of **SelfConnectedObject** is partitioned into two concepts: **Substance** and **CorpuscularObject**. A **Substance** is an **Object** in which every part is similar to every other in every relevant respect. More precisely, something is a **Substance** when all of its parts (down to an unspecified level of granularity) have the properties of the whole. Thus, things like water and clay would be subclasses of **Substance**, as would topographic locations like surfaces and geographic areas.

Disjoint from **SelfConnectedObject**s are **Collection**s. **Collection**s consist of disconnected parts, and the relation between these parts and their corresponding **Collection** is known as **member** in the SUMO.

Note that this **member** predicate is different from the **instance** and **element** predicates, which relate things to the **Class**es or **Set**s to which they belong. Unlike **Set**s, **Collection**s have a position in space-time, and **member**s can be added and subtracted without thereby changing the identity of the **Collection**. Some examples of **Collection**s are toolkits, football teams, and flocks of sheep.

Going back up the hierarchy of the ontology, we return to the concept of **Physical** and its other child **Process**. We wanted a complete typology of the various sorts of processes, but we found that there were very few resources to guide us here. SUMO's Process hierarchy begin with research into classifying over 3,000 English verbs into "semantically coherent verb classes" (Levin 1993). There are 48 main verb classes outlined in this work. However, not all of these are relevant to the task of developing a typology of processes. Some of the verb classes relate to static predicates in the ontology rather than to processes, and some classes seem to be syntactically motivated, for example the class of verbs that take predicative complements. We eliminated the verb classes that do not seem to refer to genuine processes, adjusted the hierarchy somewhat, and developed formal axioms for the remaining concepts.

Returning to the highest level distinction in the SUMO, that between

**Physical** and **Abstract**, let us now consider the **Abstract** branch. The class **Abstract** subsumes six disjoint concepts: **Set**, **Proposition**, **Quantity**, and **Attribute**, **Graph** and **GraphElement**. **Set** is the ordinary set-theoretic notion, and it subsumes **Class**, which, in turn, subsumes **Relation**. A **Class** is understood as a **Set** with a property or conjunction of properties that constitute the conditions for membership in the **Class**, and a **Relation** is a **Class** of ordered tuples. Note that **Relation** is immediately subsumed by **Class**, rather than **Set**, because we restrict **Relations** to those ordered tuples that express intensional content. For example, the predicate **mother** in the SUMO does not just denote the set of all ordered pairs whose first element is a mother and whose second element is an offspring of the first element. This predicate also expresses the property or conjunction of properties that comprise the notion of biological motherhood.

The concept of **Proposition** corresponds to the notion of semantic or informational content. However, the SUMO places no size restrictions on this content. Although some **Propositions** are expressed by single sentences, other **Propositions** are expressed by entire books or even libraries of books. This is a broader notion than is used in many ontologies, but it does not seem to be possible to make a principled distinction between the abstract content expressed by one sentence and the abstract content expressed by larger units of discourse. Some examples of **Propositions** would be the story line conveyed by a novel and the musical content denoted by a printed score.

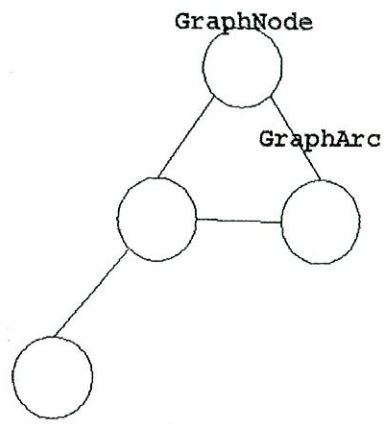

*Figure 57: A graph*

The class of **Attributes** includes all qualities, properties, etc. that are not reified as **Objects**. For example, rather than dividing the class of **Animals** under **Objects** into **FemaleAnimals** and **MaleAnimals**, we make **Female** and **Male** instances of **BiologicalAttribute**, which is a subclass of **Attribute**.

**Quantity** under **Abstract** is divided into **Number** and **PhysicalQuantity**. The former is understood as a count independent of an implied or explicit measurement system, and the latter is taken to be a complex consisting of a **Number** and a particular unit of measure. Thus, 1 meter and 39.37 inches would be two distinct instances of **PhysicalQuantity** in the SUMO, though they can be shown to be equivalent via two conversion axioms in the ontology.

A **Graph** (as shown in Figure 57) is an abstract representation of physically connected entities. It is an arrangement of **GraphElements**, which are **GraphNodes** and **GraphArcs**. The relation **abstractCounterpart** relates **Abstract** graphs to their **Physical** counterparts.

## SUMO's Eleven Modules

for axioms

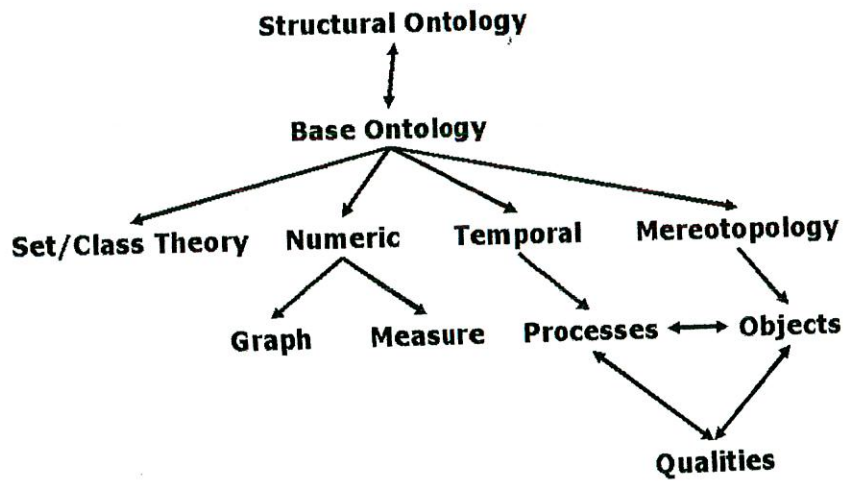


*Figure 58: SUMO's 11 modules*

This section goes through each of the modules of SUMO (Figure 58). There are many ways to group axioms in SUMO. These divisions were largely chosen for their separability, as a way to support modular and partial use of SUMO.

### Structural Ontology

This section of SUMO includes some of the most fundamental definitions, including the relations that are used to state that something in an instance of a class (**`instance`**), that one class is a subclass of another (**`subclass`**), that one relation is a subrelation of another (**`subrelation`**) and what the type restrictions are for a relation or function (**`domain`**, **`domainSubclass`**, **`range`**, **`rangeSubclass`**).

## Base Ontology

This section includes the top classes in SUMO: **Entity**, **Physical**, **Abstract**, **Object**, **Substance**, and several levels of the hierarchy below that. It also includes the relations **part**, **property**, **manner**, and some of the basic semiotic relations (detailed in the section Semiotics Content on page 147 below). Also included are the basic types of numbers, arithmetic functions, and mathematical inequalities. The fundamental different types of relations are defined, which include **TotalValuedRelation**, **PartialValuedRelation**, relations and functions of different arities, reflexivity, transitivity and other basic binary properties. The different basic kinds of **CaseRole** are defined, including **agent**, **patient**, **origin**, **destination** and others. Relations having to do with absolute and relative probability are defined. Various modals such as **needs**, **wants**, **desires**, **believes**, **capability**, **hasPurpose** are given. Relations of causality such as **causes**, **causesSubclass** and **causesProposition** are provided.

## Set/Class Theory

This section defines basic set-theoretic notions, such as union, difference, instersection etc. Class relations like mutual disjointness and cardinality defined. One very powerful function is **KappaFn**, although practical reasoning with such a construct is very difficult. **KappaFn** allows one to define a class "dynamically" by stating the intensional axioms for a new but unnamed class.

## Numeric Functions

This section is relatively straightforward, defining functions for addition (**AdditionFn**), subtraction, square root etc.

## Temporal

Most of the original definitions and axioms were derived from James Allen's work. This part of the section was extensively revised on the basis of comments from Pat Hayes. Another part of this section was an attempt to incorporate the

Simple-Time ontology from the Ontolingua server into the SUMO. Included are relations between events, as well as metric times, like **January** and **Thursday**.

## Mereotopology

Most of this content was originally taken from Barry Smith's and Nicola Guarino's papers on the subject. Content that describes "negative" space such as **Holes**, is taken mostly from (Casati&Varzi 1995) . We define various abstract parts of objects such as **bottom**, and relationship between objects, such as that they are **connected**.

## Graph

This section defines many basic graph-theoretic notions. Many real-world entities exhibit graph-theoretic structures, includes computer networks, transportation systems, and electrical systems. We define nodes and arcs, paths, weights, shortest path and many other useful theoretical notions.

## Measure

Various abstract classes of measures are defined such as **AreaMeasure** and **VolumeMeasure**, as well as specific measures such as **Acre** and **Liter**. Many functions such as **MiliFn** are defined to aid in unit conversion. Functions that construct ratios, such as **PerFn**, are provided. Relations that relate physical objects to measurements such as **length**, **width** and **height** are defined.

## Processes

The process ontology is one of the largest and most significant parts of SUMO. The top levels of the hierarchy under process are **DualObjectProcess**, **NaturalProcess**, **IntentionalProcess**, **Motion** and **InternalChange**. There are however three "orphans" that have resisted a more principled high level classification. They are **PowerGeneration**, **CausingHappiness** and **CausingUnhappiness**.

## Lower Level Ontologies

There are a number of ontologies that extend SUMO. Together, they number some 20,000 terms and 70,000 axioms (not counting the automatically added content noted briefly above and shown in Figure 55), making the SUMO family the largest publicly available formal ontology. These additional ontologies include a MId-Level Ontology (MILO), comprising some 2000 terms with associated formal definitions, for concepts that are still quite general, but are arguably too low-level for SUMO itself. Domain ontologies include theories of geography, terrorism, political systems, economic systems, as well as the GOLD linguistic ontology described in (Farrar 2004).

MILO is the most significant of the lower level ontologies. Rather than being driven by a specific application, it was developed largely as a by-product of the WordNet mapping effort. As mentioned in the section above on SUMO History, we guided the MILO development by using a version of the Brown corpus that was manually annotated with fully disambiguated WordNet word senses. We started with the most frequently appearing senses and progressed through to all senses that appeared at least three times in the corpus. If there was an existing SUMO concept that was equivalent to each word, we simply recorded the mapping equivalence. If there was not an equivalent concept however, we attempted to create a new formal concept. While this occasionally resulted in some new concept in SUMO proper, and some very domain specific concepts that clearly belonged in a domain ontology, there were many others that equally clearly applied to many domains and therefore were placed in MILO. We have arbitrarily capped SUMO at roughly 1000 terms. Subsequent domain ontology development efforts have also resulted in the "discovery" of mid-level concepts that are too general to belong in one particular domain ontology, yet are no more general than existing concepts in SUMO. They were therefore placed in MILO.

Several domain ontologies were created in attempt to ontologize the CIA World Factbook (Brasseys 1997). While there have been other such attempts (Frank et al 1999), to our knowledge, none have captured more than simple binary relations, leaving alone the bulk of the content. The ontologies that resulted were Communications, Government, Geography, People,

TransnationalIssues, Transportation, Economy, and Finance. However, in the subsequent years to that project, those files have been considerably expanded. The CountriesAndRegions ontology is largely a list of facts one would find in an atlas, such as that France is a **Nation**. The Distributed Computing ontology was developed in the course of a large project that was focused on distributed operating systems and other system-level computing issues. In order to perform automated system configuration it was beneficial to have a formal ontology that could support a decision support system. Most of this ontology is highly specific. The Engineering Components ontology was developed by Michal Sevcenko (Sevcenko 2002). The Military ontologies grew out of several projects for the US military. The North American Industrial Classification system ontology was an initial and rather shallow attempt to convert the NAIC taxonomy standard into an ontology. Only the first level of the taxonomy has been properly integrated and defined. This effort did however motivate a number of additions to MILO for occupations and retail products. The Physical Elements ontology is a simple collection of facts about the fundamental physical elements, such as that lead has an atomic weight of 207. The Viruses ontology was developed by Mike Pool of IET while working on a project to represent information about biological warfare. The World Airports ontology is also a simple collection of facts about the airports of the world.