Ontology Archeology: Mining a Decade of Effort on the Suggested Upper Ontology

Adam Pease and Christoph Benzmüller*

Articulate Software, Angwin, CA, USA

ARCOE-10, August 16-17, 2010 Lisbon, Portugal

*supported by DFG grant BE 2501/6-1
Ontology: An (Over-)Stretched Notion?

Ontology Archeology: Mining a Decade of Effort on SUMO
Ontology versus Taxonomy
Notion of Taxonomy

Characteristics

- simple hierarchical categorization and classification of entities in a domain (simple ontology in form of a hierarchy)
- classes, super-/subclasses, related classes, maybe some extras such as cardinality constraints, etc.
- computational properties $\approx$ expressivity

Analogy?

- taxonomy development $\approx$ GUI builder (restr. lang. features)
Notion of Taxonomy

Characteristics

- simple hierarchical categorization and classification of entities in a domain (simple ontology in form of a hierarchy)
- classes, super-/subclasses, related classes, maybe some extras such as cardinality constraints, etc.
- computational properties $\geq$ expressivity

Analogy?

- taxonomy development $\approx$ GUI builder (restr. lang. features)
Notion of Ontology

Characteristics

- rich semantic specification of a conceptualization; describes terms & entities and their relationships in a domain (often includes a taxonomy)
- expressive rules, e.g. first-order, sometimes even higher-order
- expressivity > computational properties

Analogy?

- ontology development \(\approx\) unrestricted, large-scale software development
Notion of Ontology

Characteristics

- rich semantic specification of a conceptualization; describes terms & entities and their relationships in a domain (often includes a taxonomy)
- expressive rules, e.g. first-order, sometimes even higher-order
- expressivity > computational properties

Analogy?

- ontology development ≈ unrestricted, large-scale software development
Ontology: An (Over-)Stretched Notion?

Taxonomies

Simple Ontologies

- OBO
- DublinCore

Ontologies

- OpenCyc
- Cyc
- BFO

SUMO
- proprietary
- open source
- just 113 terms

DOLCE

- very small

Adam Pease and Christoph Benzmüller®
Ontology: An (Over-)Stretched Notion?

sufficiently expressive
Simple Taxonomy (excerpt from SUMO's base taxonomy)

(subclass Physical Entity)
(subclass Abstract Entity)
(partition Entity Physical Abstract)

(subclass Object Physical)
(subclass Process Physical)
(partition Physical Object Process)

...
The SUMO Ontology is more than just a Taxonomy

(subclass Physical Entity)
(subclass Abstract Entity)
(partition Entity Physical Abstract)

(subclass Object Physical)
(subclass Process Physical)
(partition Physical Object Process)

(<=>
    (instance ?PHYS Physical)
    (exists (?LOC ?TIME)
        (and
            (located ?PHYS ?LOC)
            (time ?PHYS ?TIME))))
The SUMO Ontology is more than just a Taxonomy

(subclass Physical Entity)
(subclass Abstract Entity)
(partition Entity Physical Abstract)

(subclass Object Physical)
(subclass Process Physical)
(partition Physical Object Process)

(<=>
  (instance ?PHYS Physical)
  (exists (?LOC ?TIME)
    (and
      (located ?PHYS ?LOC)
      (time ?PHYS ?TIME)))
  (subclass ObjectAttitude IntentionalRelation)
  (=>
    (and
      (instance ?REL ObjectAttitude)
      (?REL ?AGENT ?THING))
    (instance ?THING Physical))
More About SUMO
The SUMO Ontology: Some Facts

- developed since: > 10 years
- original motivation: education support; did not really narrow the research focus though; observation in different specialist domains: need for common upper level ontology
- open source: www.ontologyportal.org
- SUMO versus SUMO:
  - SUMO: Suggested Upper-level Ontology
  - MILO: Mid-level Ontology
  - Specific Domain-level Ontologies
  \{SUMO\}
- representation language: SUO-KIF (adaptation of the Knowledge Interchange Format KIF)
- logic: mainly first-order; some higher-order extensions

Adam Pease and Christoph Benzmüller

Ontology Archeology: Mining a Decade of Effort on SUMO
The SUMO Ontology: Some Facts

- Developed since: > 10 years
- Original motivation: Education support; meta data for government training applications; did not really narrow the research focus though; observation in different specialist domains: need for common upper level ontology
- Open source: www.ontologyportal.org
- SUMO versus SUMO:
  - SUMO: Suggested Upper-level Ontology
  - MILO: Mid-level Ontology
  - Specific Domain-level Ontologies
- Representation language: SUO-KIF (adaptation of the Knowledge Interchange Format KIF)
- Logic: mainly first-order; some higher-order extensions
The SUMO Ontology: Some Facts

- developed since: > 10 years
- original motivation: education support
  meta data for government training applications; did not really
  narrow the research focus though; observation in different
  specialist domains: need for common upper level ontology
- open source: www.ontologyportal.org

- SUMO versus SUMO:
  
  SUMO: Suggested Upper-level Ontology
  MILO: Mid-level Ontology
  Specific Domain-level Ontologies

- representation language: SUO-KIF
  (adaptation of the Knowledge Interchange Format KIF)
- logic: mainly first-order; some higher-order extensions
The SUMO Ontology: Some Facts

- developed since: > 10 years
- original motivation: education support
  meta data for government training applications; did not really narrow the research focus though; observation in different specialist domains: need for common upper level ontology
- open source: www.ontologyportal.org
- SUMO versus SUMO:
  - SUMO: Suggested Upper-level Ontology
  - MILO: Mid-level Ontology
  - Specific Domain-level Ontologies
  SUMO

- representation language: SUO-KIF
  (adaptation of the Knowledge Interchange Format KIF)
- logic: mainly first-order; some higher-order extensions
The **SUMO Ontology:** Some Facts

- developed since: $>$ 10 years
- original motivation: education support meta data for government training applications; did not really narrow the research focus though; observation in different specialist domains: need for common upper level ontology
- open source: www.ontologyportal.org
- **SUMO versus SUMO:**
  
  **SUMO**: Suggested Upper-level Ontology  
  **MILO**: Mid-level Ontology  
  Specific Domain-level Ontologies

  \[
  \begin{align*}
  \text{SUMO} &\quad \text{SUO-KIF} \\
  \text{SUMO} &\quad (\text{adaptation of the Knowledge Interchange Format KIF})
  \end{align*}
  \]

- representation language: SUO-KIF  
- logic: mainly first-order; some higher-order extensions
<table>
<thead>
<tr>
<th>SUMO</th>
<th>Total Terms</th>
<th>relations</th>
<th>Total Axioms</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1173</td>
<td>353</td>
<td>4741</td>
<td>932</td>
<td></td>
</tr>
<tr>
<td>MILO</td>
<td>Total Terms</td>
<td>relations</td>
<td>Total Axioms</td>
<td>Rules</td>
</tr>
<tr>
<td>1662</td>
<td>159</td>
<td>5116</td>
<td>1183</td>
<td></td>
</tr>
<tr>
<td>Domain ontologies</td>
<td>Total Terms</td>
<td>relations</td>
<td>Total Axioms</td>
<td>Rules</td>
</tr>
<tr>
<td>17312</td>
<td>708</td>
<td>77974</td>
<td>2041</td>
<td></td>
</tr>
<tr>
<td>Total for all ontologies</td>
<td>Total Terms</td>
<td>relations</td>
<td>Total Axioms</td>
<td>Rules</td>
</tr>
<tr>
<td>20147</td>
<td>1220</td>
<td>87831</td>
<td>4156</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: SUMO term and axiom statistics
SUMO Example Entries

**SUMO — ’part’**

(domain part 1 Object)
(domain part 2 Object)
(instance part PartialOrderingRelation)
(instance part SpatialRelation)
...

**MILO — ’typical-part’**

(domainSubclass typicalPart 1 Object)
(domainSubclass typicalPart 2 Object)
(instance typicalPart BinaryRelation)
(instance typicalPart SpatialRelation)
(relatedInternalConcept typicalPart part)
...

**Government Domain Ontology — ’capitalCity’**

(domain capitalCity 1 City)
(domain capitalCity 2 GeopoliticalArea)
(instance capitalCity BinaryPredicate)
(subrelation capitalCity administrativeCenter)
...
How Developed?

**SUMO**
- initially by mining theories of common sense knowledge (e.g. James Allen’s theory of temporal relations)
- by reflection and inspection of world; most of SUMO now original

**MILO**
- WordNet used to check coverage of SUMO/MILO
- criterion: every WordNet synset that occurred ≥ 3× in Brown corpus considered worthwhile for inclusion
- inclusion not always possible: ’better’ hardly formalizable without context (should be handled at NL level: e.g. CELT)

**Domain Ontologies**
- deliveries for researchers and projects: e.g. Geography Ontology (for government), Media Ontology (for contractor in London)
What goes in **SUMO** and what in **MILO**?

- no 'deep' answer
- simply keep **SUMO** around 1000 terms
- hand-select most specific terms in **SUMO** and move to **MILO**
The Growth of SUMO
The Growth of SUMO

100,000 lines

80,000 lines

60,000 lines

40,000 lines

20,000 lines

2000 2001 2002 2003 2004 2005 2006 2007 2009 2010

NAICS
engineering
elements
WMD
Virus
Transport
Transnat
Sports
QoS
People
MProcess
MPeople
MDevices
Military
Media
Justice
Language
Government
Geography
Financial
Economy
Comm
Arabic
MILO
SUMO
The Growth of SUMO

With YAGO, OBO, DBPedia, mondial (semi-automatic, partial inclusions)
Tool Support (not only) for SUMO Development
Tool Support (not only) for SUMO Development

Sigma Knowledge Engineering Tool

- browsing and displaying tool
- supports inspection and debugging
- open source: sigmakee.sourceforge.net

FOL & HOL Theorem Provers — Integrated with Sigma

- KIFVampire (FOL), SystemOnTPTP (FOL), LEO-II (HOL)
- support for ontology debugging and question answering

CELT

- Deep NL processing with respect to own ontology
  (cf. LogAnswer which does not have its own ontology)
Tool Support (not only) for **SUMO** Development

Adam Pease and Christoph Benzmüller

---

**Walking**

appearance as argument number 1

*(documentation* *Walking* *EnglishLanguage* "ambulating relatively slowly, i.e. moving in such a way that at least one foot is always in contact with the ground.")

appearance as argument number 2

*(partition* *Ambulating* *Walking* *Running*)

*(subclass* *Waiting* *Walking*)

*(termFormat* *EnglishLanguage* *Walking* "walking")

antecedent

*If a process is an instance of walking and process is an instance of running and an agent is an agent of process and agent is an agent of process and the measure of agent is a length measure per a time duration holds during the time of existence of process and the measure of agent is length measure per time duration holds during the time of existence of process, then length measure is greater than length measure*
Tool Support (not only) for SUMO Development

Adam Pease and Christoph Benzmüller

Ontology Archeology: Mining a Decade of Effort on SUMO
Tool Support (not only) for SUMO Development

(SZS Status Theorem)

1. (exists (?X6)
   (instance ?X6 PrimaryColor))
2. (instance Red PrimaryColor)
3. (instance Red PrimaryColor)
4. (not
   (instance ?X6 PrimaryColor))
5. (not
   (instance Red PrimaryColor))
6. True

[Query]

(SZS Status Theorem)
Answer 1. [definite] ?X = Blue

1. (instance Blue PrimaryColor)
2. (exists (?X1)
   (instance Blue PrimaryColor))
3. (not
   (exists (?X1)
    (instance Blue PrimaryColor)))
4. (not
   (instance Blue PrimaryColor))
5. (instance Blue PrimaryColor)
6. (not
   (instance Blue PrimaryColor))
7. True
8. True
9. True

[KB]

[Instantiated Query]

2
3
4
3 5
Insights, Challenges, Future Work
Insights, Challenges, Future Work

Overall

- biggest surprise so far: no big surprise so far
  (although problem expectation was high for some contractors)
- local problems are continuously being detected and revised
- periodic minor revisions as required/appropriate
- no fundamental issues detected so far (which does not mean there are no fundamentally different theories about the world)
- interesting recent problem: semantics of embedded formulas and modal operators → solution proposed in next talk
- interesting and demanding nature of field
- SUMO interesting and relevant to many communities:
  Computer Science, Linguistics, Philosophy (hard though for students to get work accepted as MSc, PhD)
- example use in Linguistics: checking analogies in NL
Insights, Challenges, Future Work

Overall

- biggest surprise so far: no big surprise so far (although problem expectation was high for some contractors)
- local problems are continuously being detected and revised
- periodic minor revisions as required/appropriate
- no *fundamental* issues detected so far (which does not mean there are no fundamentally different theories about the world)
- interesting recent problem: semantics of embedded formulas and modal operators → solution proposed in next talk
- interesting and demanding nature of field
- SUMO interesting and relevant to many communities: Computer Science, Linguistics, Philosophy (hard though for students to get work accepted as MSc, PhD)
- example use in Linguistics: checking analogies in NL
Insights, Challenges, Future Work

Overall

- biggest surprise so far: no big surprise so far (although problem expectation was high for some contractors)
- local problems are continuously being detected and revised
- periodic minor revisions as required/appropriate
- no fundamental issues detected so far (which does not mean there are no fundamentally different theories about the world)
- interesting recent problem: semantics of embedded formulas and modal operators  \( \quad \rightarrow \) solution proposed in next talk
- interesting and demanding nature of field
- SUMO interesting and relevant to many communities: Computer Science, Linguistics, Philosophy (hard though for students to get work accepted as MSc, PhD)
- example use in Linguistics: checking analogies in NL
Insights, Challenges, Future Work

Overall

- biggest surprise so far: no big surprise so far (although problem expectation was high for some contractors)
- local problems are continuously being detected and revised
- periodic minor revisions as required/appropriate
- no fundamental issues detected so far (which does not mean there are no fundamentally different theories about the world)
- interesting recent problem: semantics of embedded formulas and modal operators → solution proposed in next talk
- interesting and demanding nature of field
- SUMO interesting and relevant to many communities: Computer Science, Linguistics, Philosophy (hard though for students to get work accepted as MSc, PhD)
- example use in Linguistics: checking analogies in NL
Insights, Challenges, Future Work

Overall

- biggest surprise so far: no big surprise so far (although problem expectation was high for some contractors)
- local problems are continuously being detected and revised
- periodic minor revisions as required/appropriate
- no *fundamental* issues detected so far (which does not mean there are no fundamentally different theories about the world)
- interesting recent problem: semantics of embedded formulas and modal operators → solution proposed in next talk
- interesting and demanding nature of field
- SUMO interesting and relevant to many communities: Computer Science, Linguistics, Philosophy (hard though for students to get work accepted as MSc, PhD)
- example use in Linguistics: checking analogies in NL
Open Source

- SUMO is largest open source formal ontology
- motivation: make SUMO ubiquitous, hope for real impact
- disappointment: not the big growth as originally anticipated
- some contractors do not want their additions to become open source → no need to publish new domain ontologies
- concrete license: BSD like license for SUMO, GPL for MILO (creative commons license would be choice today)
Open Source

- SUMO is largest open source formal ontology
- motivation: make SUMO ubiquitous, hope for real impact
- disappointment: not the big growth as originally anticipated
- some contractors do not want their additions to become open source $\rightarrow$ no need to publish new domain ontologies
- concrete license: BSD like license for SUMO, GPL for MILO (creative commons license would be choice today)
Ongoing and Future Work

- better archeology (cvs logs available for purpose):
  - detailed analysis/inspection of modifications to
    - comments
    - hierarchy
    - axioms
    - ...
  What can be learned?
- proper treatment of modal operators → possible world semantics for SUMO?
- appropriate treatment of (other) higher-order aspects
- modeling of interesting problem scenarios (e.g. Smullyan’s puzzles) and testing of SUMO and Sigma for them
- more domain ontologies
- use of CELT & SUMO in NL applications
Ongoing and Future Work

- better archeology (cvs logs available for purpose):
  detailed analysis/inspection of modifications to
  - comments
  - hierarchy
  - axioms
  - ...
  What can be learned?

- proper treatment of modal operators → possible world semantics for SUMO?

- appropriate treatment of (other) higher-order aspects

- modeling of interesting problem scenarios (e.g. Smullyan’s puzzles) and testing of SUMO and Sigma for them

- more domain ontologies

- use of CELT & SUMO in NL applications
Ongoing and Future Work

- better archeology (cvs logs available for purpose): detailed analysis/inspection of modifications to
  - comments
  - hierarchy
  - axioms
  - ...
  What can be learned?

- proper treatment of modal operators \(\rightarrow\) possible world semantics for SUMO?

- appropriate treatment of (other) higher-order aspects

- modeling of interesting problem scenarios (e.g. Smullyan’s puzzles) and testing of SUMO and Sigma for them

- more domain ontologies

- use of CELT & SUMO in NL applications
Ongoing and Future Work

- better archeology (cvs logs available for purpose):
  - detailed analysis/inspection of modifications to
    - comments
    - hierarchy
    - axioms
    - ...
  What can be learned?

- proper treatment of modal operators $\rightarrow$ possible world semantics for SUMO?

- appropriate treatment of (other) higher-order aspects

- modeling of interesting problem scenarios (e.g. Smullyan’s puzzles) and testing of SUMO and Sigma for them

- more domain ontologies

- use of CELT & SUMO in NL applications
Ongoing and Future Work

- better archeology (cvs logs available for purpose):
  detailed analysis/inspection of modifications to
  - comments
  - hierarchy
  - axioms
  - ...

- What can be learned?

- proper treatment of modal operators $\rightarrow$ possible world semantics for SUMO?

- appropriate treatment of (other) higher-order aspects

- modeling of interesting problem scenarios (e.g. Smullyan’s puzzles) and testing of SUMO and Sigma for them

- more domain ontologies

- use of CELT & SUMO in NL applications
Insights, Challenges, Future Work

Ongoing and Future Work

- Better archeology (cvs logs available for purpose): detailed analysis/inspection of modifications to:
  - comments
  - hierarchy
  - axioms
  - ...

  What can be learned?

- Proper treatment of modal operators $\rightarrow$ possible world semantics for SUMO?

- Appropriate treatment of (other) higher-order aspects

- Modeling of interesting problem scenarios (e.g. Smullyan’s puzzles) and testing of SUMO and Sigma for them

- More domain ontologies

- Use of CELT & SUMO in NL applications
Challenge

expressive, general purpose knowledge representation

robust and effective integration

heterogeneous reasoning systems (general purpose & specialist)
Thank You!