Analysis of general ontologies
(Technical report)

Blanca Vazquez, Samuel Vieyra, Hasdai Pacheco and Hugo Estrada
Fund of Information and Documentation for the Industry INFOTEC, San Fernando 37, Col. Toriello Guerra, Tlalpan, 14050 Mexico, D.F., Mexico
{blanca.vazquez, samuel.vieyra, ebenezer.sanchez, hugo.estrada} @infotec.com.mx

Summary
Ontology consists of a set of concepts, axioms, and relationships that describe a domain of interest. A general ontology is limited to concepts that are meta, generic, abstract and philosophical, and therefore are general enough to address (at a high level) a broad range of domain areas [1].

The objective of this report is to analysis several general ontologies to identify which ontology could be useful to carry out the enrichment of business models. In addition, a set of Semantic Annotation Rules (SARs)[1] was defined in order to select the best concept and instance candidates from the general ontology (resultant of this analysis) to enrich each BPMN 2.0 primitive of a BP model.

1. Introduction
According to Gruber [2], ontology is a formal, explicit specification of a shared conceptualization. Ontologies are considered as key elements for semantic interoperability and to share vocabularies for describing information relevant to a certain area of application [3]. Guarino in [4] classified the ontologies in three categories: *Upper level ontologies or general ontologies, domain ontologies and application ontologies*.

A *general ontology* describes very general concepts like space, time, matter, etc., which are independent of a particular problem or domain; a *domain ontology* describes the vocabulary related to a generic domain (like medicine or automobiles) and an *application ontology* describe concepts depending both on a particular domain and task, which are often specializations of both the related ontologies.

In this report, we have described five general ontologies along different criterias. The aim is to define an ontology that could help us in order to carry out the successfully the semantically annotation of business models. The aspects that considered to analyze the ontologies are described to below: dimensions, top level concepts, coding language, modularity, language,

[1] Definition of Semantic Annotation Rules for annotation process of Business Process Models (Technical report). http://www.semanticwebbuilder.org.mx/semanticAnnotation/definitionOfSARs

applications, semantic annotations (rdfs:label, rdfs:comment), domain concepts, specificity level, knowledge reuse, rules and axioms.

The general ontologies analyzed in this report were selected because have been applied in real projects, such as reasoning and language natural generation, the enrichment of lexical resources, word sense disambiguation, natural language processing and support modeling and simulation. The analysis of the general ontologies carried out is presented in Section 2, also a comparative table of the features most relevance of these ontologies is described in this section, and Section 3 present the conclusions of this report.

2. Analysis of general ontologies

This section describes the analysis of general ontologies carried out.

2.1 Sumo

The Suggested Upper Merged Ontology (SUMO) and its domain ontologies form the largest formal public ontology in existence today. They are being used for research and applications in search, linguistics and reasoning. SUMO is the only formal ontology that has been mapped to all the WordNet lexicon. The SUMO was initially developed at Teknowledge Corp.

Dimensions: ~25,000 terms, ~80,000 axioms.

Top level concepts: An overview of the top level concepts of SUMO is described to below.

- Physical
  - Object
    - SelfConnectedObject
    - ContinuousObject
    - CorpuscularObject
    - Collection
  - Process
- Abstract
  - SetClass
    - Relation
  - Proposition
  - Quantity
    - Number
    - PhysicalQuantity
  - Attribute

Specificity level: The domain concepts are usually generics.

Coding language: SUO-KIF (originally). SUMO is free and owned by the IEEE. The ontologies that extend SUMO are available under GNU General Public License.

Modularity: The SUMO is a modular ontology. That is, the ontology is divided into self-contained sub ontologies. Each sub ontology is indicated by a section header, and the dependencies between the sub ontologies are specified with statements of the form: INCLUDES '<SUBONTOLOGY>'". In SUMO exist three modularity levels: 1) SUMO (General Concepts), 2) Mid- Level Ontology, and 3) several modules related to specific domains.

Language: English.

Applications: SUMO has been applied to semantic annotation of image collections [5]; translating UNL expressions to logical expressions [6]; conceptual metaphors: ontology-based representation and corpora driven mapping principles [7].

Use of rdfs:label: SUMO applies the use to provide a human-readable version of a resource's name.

Use of rdfs:comment: SUMO does not apply this label.

Domain concepts: Several examples of domain concepts used by SUMO are: Communications, Countries and Regions, distributed computing, Economy, Finance, Engineering components, Geography, Government, Military, Transportation, Viruses, World Airports, etc.

Knowledge Reuse: SUMO was created merging a number of previously defined upper-level ontologies. In addition to this, SUMO has been aligned with WordNet Lexical Database to promote the use of this ontology in Natural Language Understanding applications.

Rules/Axioms: ~ 800 RULES

2.2 YAGO2

YAGO2s is a huge semantic knowledge base, derived from Wikipedia WordNet and GeoNames. Currently, YAGO2s has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities. YAGO2s is part of the YAGO-NAGA project at the Max Planck Institute for Informatics in Saarbrücken/Germany.

Dimensions: ~1,354,000 classes, ~10 million entities, ~120 million facts.

Top level concepts: An overview of the top level concepts of YAGO2 retrieved from yago simple taxonomy is described to below.

Abstraction
Artifact
Building
Organization
Person
Physical_entity
GeoEntity

Specificity level: Most of the domain concepts are specifically described in their domain context.

Coding language: YAGO model is expressed in a slight extension of RDFS.

Modularity: There are 39 modules. Each module receives a data source as input. Input sources are WordNet, Wikipedia, WordNet Domains, the Universal, WordNet, and Geonames. Others can be added [19].

Language: English, with multilingual annotations.

Applications: YAGO2 has been applied to information extraction [8], a spatially and temporally enhanced knowledge base from Wikipedia [9], and search for knowledge instead of Webpages [10], Jeopardy contestant WATSON [11].

Use of rdfs:label: YAGO2 applies the use of rdfs:label to provide a human-readable version of a resource's name. In YAGO2 the labels are usually descriptive, however the code language is omitted.

Use of rdfs:comment: YAGO2 applies the use of rdfs:comment to provide a human-readable version of a resource's name. In YAGO2 the comments are usually descriptive, however the code language is omitted.

Domain concepts: YAGO2 describes domain concepts in all WordNet domains (astrology, linguistics, literature, religion, theatre, sport, agriculture, alimentation, engineering, etc.).

Knowledge Reuse: YAGO2 is derived from Sumo, Wikipedia, WordNet and GeoNames.

Rules/Axioms: There are different types of rules: factual, implication, replacement and extraction rules [11]. Factual rules are simply additional facts for the YAGO2 knowledge base. They are declarative translations of all the manually defined exceptions and facts that the previous YAGO code contained. Implication rules serve to deduce new knowledge from the existing knowledge. Replacement rules imply that if a part of the source text matches a specified regular expression, a certain string should replace it. Extraction rules applies primarily

to patterns found in the Wikipedia infoboxes, but also to Wikipedia categories, article titles, and even other regular elements in the source such as headings, links, or references.

2.3 OntoSem

The OntoSem (Ontological Semantics) [12] ontology is a formal, language independent, unambiguous general ontology that provides a metalanguage for describing conceptual meaning. The root concepts of OntoSem are object, event and property. It contains around 8000 concepts and 350 properties with their respective meanings. OntoSem has been created by University of Maryland Baltimore County (USA).

Dimensions: ~9,000 concepts, ~350 properties, ~30,000 senses, Onomasticon ~350,000 entries.

Top level concepts: An overview of the top level concepts of OntoSem is described to below.

```
Event
        mental-event
        physical-event
        social-event
Object
        intangible-object
        metal-object
        physical-object
        social-object
Property
        attribute
        case-role
```

Specificity level: The domain concepts describe to depth each concept. For example:

```
Academic-role
        Academic-administrator
        Academic-specialist
Student
        School-student
                Elementary-school-student
                High-school-student
                University-student
Teacher
```

Coding language: Frame-based LISP notation.

Modularity: Each of the concepts in the ontology inherits either from the object, event or property.

Language: English, with multilingual annotation. But only the english version is available.

Applications: OntoSem has been applied to word sense disambiguation [13] and semantic analysis [14].

Use of rdfs:label: OntoSem applies the use of rdfs:label to provide a human-readable version of a resource's name.

Use of rdfs:comment: None identified.

Domain concepts: Several examples of domain concepts used by OntoSem are: Academic-event, work-activity, financial-event, geopolitical-entity, animal, social-role, and business roles.

Knowledge Reuse: OntoSem does not knowledge reuse.

Rules/Axioms: OntoSem contains rules and axioms.

2.4 Cyc Knowledge

The Cyc Knowledge Base (KB) is a formalized representation of a vast quantity of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. CYC KB is developed by Cycorp.

Dimensions: ~ 3.2 million assertions (facts and rules), ~ 280,000 concepts, ~ 12,000 concept-interrelating predicates.

Top level concepts: An overview of the top level concepts of Cyc Knowledge is described to below.

Thing
Intangible thing
Individual
Relations
Sets
Collections
Paths
Logic
Math

Specificity level: The domain concepts describe to depth each concept. For example:

Student
- School_children
- Secondary_school_student
- Grade_school_student
- Graduate_student
- Kindergarten_student
- Medical_residency_program_student

Coding language: Formal language CycL.

Modularity: The Cyc KB is divided into many (currently thousands of) "contexts" (or "microtheories"), each of which is essentially a collection of assertions that share a common set of assumptions; some microtheories are focused on a particular domain of knowledge, some a particular interval in time, some a particular level of detail, etc.

Language: English

Applications: CYC KB has been applied to semantic web infrastructure for clinical research and quality reporting [15], the develop of a service-oriented platform [16], and text mining [17].

Use of rdfs:label: CYC KB applies the use of rdfs:label to provide a human-readable version of a resource's name.

Use of rdfs:comment: None identified.

Domain concepts: Several examples of domain concepts used by CYC KB are: healthcare, computer security, command and control, mortgage banking, vehicles, buildings & weapons, social activities, military organizations

Knowledge Reuse: The knowledge reuse by CYC KB is based in SUS, FIPS 10-4, several large (300k-term) pharmaceutical thesauri, large portions of WordNet, MeSH/Snomed/UMLS, and the CIA World Factbook.

Rules/Axioms: CYC KB contains rules and axioms.

2.5 Cosmo

COSMO is an effort proposed and driven by PatCassidy. This Common Semantic Model (COSMO) is conceived as being made up of a lattice of ontologies which will serve as a set of basic logically-specified concepts (classes, relations, functions, instances) with which the

meanings of all terms and concepts in domain ontologies can be specified. The most important function of the COSMO is to serve as a Foundation Ontology that has a sufficient inventory of fundamental concept representations so that it can support utilities to translate assertions of fundamentally different ontologies into the terminology and format of each other. The COSMO can also be used as the starting ontology for creation of more specialized domain ontologies.

The COSMO ontology is intended to be a merged ontology, initially derived primarily from elements (Types [classes], relations, and Inference Rules) existing in the public ontologies OpenCyc (The version used was 0.78b OWL version) and SUMO (both SUMO and the MILO extension were used). Additional elements were adopted from the BFO and DOLCE ontologies.

Dimensions: 7339 types (OWL classes), 808 relations and 2039 restrictions.

Top level concepts: An overview of the top level concepts of Cyc Knowledge is described to below.

Object Attribute GenericLocation GenericSubstance GenericAgent Role TemporalLocation Group SituationProcessEventOrState.

Specificity level: The domain concepts describe to depth each concept. For example:

Human-role
- Student
  - CollageStudent
  - Pupil

Coding language: OWL

Modularity: None identified

Language: English

Applications: Cosmo ontology has been applied to the building of a hierarchical component framework to support component-based modeling and simulation [18]

Use of rdfs:label: None identified.

Use of rdfs:comment: Cosmo applies the use of rdfs:label to provide a human-readable version of a resource's name. Moreover, many elements are referenced to WordNet with the purpose of supporting Natural Language Understanding.

Domain concepts: Several examples of domain concepts used by CYC KB are: Person, organization, group (event, education, object group), role, object, artifact-generic, quantity, individual, synonym.

Knowledge Reuse: The COSMO ontology is intended to be a merged ontology, such as OpenCyc, SUMO, MILO, BFO, DOLCE.

Rules/Axioms: Non-Identified.

A comparative table of the features most relevance of the general ontologies is presented in Table 1.

Table 1: Comparative table of the general ontologies analyzed.

Ontology	Dimensions	Coding language	License	Modularity	Language	Application	Knowledge Reuse
Sumo	~25,000 terms, ~80,000 axioms.	SUO-KIF (originally)	GNU General Public License	1) SUMO (General Concepts), 2) Mid - Level Ontology, and 3) Several modules related to specific domains. .	English	Semantic annotation of image collection, Translating UNL expressions to logical expression, : ontology-based representation and corpora driven mapping principle	SUMO was created merging a number of previously defined upper-level ontologies. In addition to this, SUMO has been aligned with WordNet Lexical Database.
YAGO2	~1,354,000 classes, ~10 million entities, ~120 million fact	RDFS	Creative Commons Attribution 3.0 License	YAGO2 presents 39 modules. Each module receives a data source as input. Input sources are: WordNet, Wikipedia, WordNet Domains, Universal WordNet and Geonames.	English, with multilingual annotation	Information extraction, a spatially and temporally enhanced knowledge base from Wikipedia, search for knowledge instead of Webpages, Jeopardy contestant WATSON.	YAGO2 is derived from Sumo, Wikipedia, WordNet and GeoNames.
OntoSem	~9,000 concepts, ~350 properties, ~30,000 senses, Onomasticon ~350,000 entries.	Frame-based LISP notation	Hakia licenses	Each of the concepts in the ontology inherits either from the object, event or property.	English, with multilingual annotation. But only the english version is available.	Word sense disambiguation, semantic analysis.	OntoSem does not knowledge reuse.
Cyc Knowledge	~ 3.2 million assertions (facts and rules), ~ 280,000 concepts, ~ 12,000 concept-interrelating predicates.	Formal language CycL.	OpenCyc License.	The Cyc KB is divided into many (currently thousands of) "contexts" (or "microtheories"), each of which is essentially a collection of assertions that share a common set of assumptions.	English	Semantic web infrastructure for clinical research and quality reporting, the develop of a service-oriented platform, and text mining.	SUS, FIPS 10-4, several large (300k-term) pharmaceutical thesauri, large portions of WordNet, MeSH/Snomed/UMLS, and the CIA World Factbook.
Cosmo	7339 types (OWL classes),	OWL	The development	The COSMO ontology is intended to be a merged	English	Building of a hierarchical	OpenCyc, Sumo, MILO, BFO, DOLCE.

Ontology	Dimensions	Coding language	License	Modularity	Language	Application	Knowledge Reuse
	808 relations and 2039 restrictions.		of COSMO is fully open	ontology, taken modules of existing public ontologies, such as OpenCyc, Sumo, MILO, BFO, and DOLCE.		component framework to support component-based modeling and simulation.	

3. Conclusions

This technical report presented an analysis of five general ontologies. The objective were to define an ontology that could help us in order to carry out the successfully the semantically annotation of business models. In order to carried to the analysis, we defined a set of criterions: dimensions, top level concepts, coding language, modularity, language, applications, semantic annotations (rdfs:label, rdfs:comment), domain concepts, specificity level, knowledge reuse, rules and axioms. A comparative table to resume the features most relevance analyzed of the general ontologies also was presented.

From of the analyzed carried out, we concluded that several facts, such as Cyc knowledge is a commercial product that is not available for download. The OntoSem ontology is described in English, with multilingual annotation, but only the english version is available. YAGO2 is derived from Sumo, Wikipedia, WordNet and GeoNames, these merge additionally offer a wealth of axiomatic knowledge, e.g. that two people sharing the same parents must be siblings. We consider to select the YAGO2 mainly why its accuracy has been manually evaluated, proving a confirmed accuracy of 95%, moreover YAGO2 presents a Creative Commons Attribution 3.0 License and is available for download. The themes available of YAGO2 are:

- ✓ Taxonomy: All types of entities and the class structure of YAGO2s. Moreover, it has formal definitions of YAGO relations.
- ✓ Simpletax: An alternative, simpler taxonomy of YAGO.
- ✓ Core: Core facts of YAGO2s, such as the facts between entities, the facts containing literals, i.e., numbers, dates, strings, etc.
- ✓ Geonames: Geographical entities, classes taken from GeoNames.
- ✓ Meta: Temporally and spatially scoped facts together with statistics and extraction sources about the facts.
- ✓ Multilingual: The multilingual names for entities.
- ✓ Link: The connection of YAGO2s to WordNet, DBpedia, etc.
- ✓ Other: Miscellaneous features of YAGO2s, such as Wikipedia in-outlinks, GeoNames data etc.

4. References

[1] Niles, I., and Pease, A. 2001. Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001. Also see http://www.ontologyportal.org

[2] Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum.-Comput. Stud. 43(5-6) (December 1995) 907–928.

[3] Vazquez, B., Martinez, A., Perini, A., Estrada, H., Morandini, M.: Enriching Organizational Model through Semantic Annotation. In: Proceedings of the Iberoamerican Conference on Electronics Engineering and Computer Science, Procedia Technology (April 24-26 2013).

[4] Guarino, N.: Formal ontology and information systems, IOS Press (1998) 3–15

[5] Laura Hollink, Guus. Schreiber, Jan Wielemaker and Bob. Wielinga. Semantic Annotation of Image Collections. In S. Handschuh, M. Koivunen, R. Dieng and S. Staab (eds.): *Knowledge Capture 2003 -- Proceedings Knowledge Markup and Semantic Annotation Workshop*, October 2003.

[6] Suresh, Kumar.: Translating UNL expressions logical expressions. Master's thesis, Department of computer science and engineering Indian Institute of Technology, Bombay (2004).

[7] Sevcenko, M. (2003). Knowledge Support for Modeling and Simulation. Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, KES 2003 Oxford, UK, pp. 99 - 103.

[8] Sarawagi, Sunita (2008). Information Extraction. Journal Found, Trends databases. Vol. 1, number 3, ISSN 1931-7883. Now Publishers Inc.

[9] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Special issue of the Artificial Intelligence Journal, 2012.

[10] Fabian M. Suchanek and Gerhard Weikum. "YAGO - Search for Knowledge instead of Webpages".Article in the yearbook of the Max Planck Society 2007.

[11] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Max Planck Institute for Informatics, Germany. INRIA Saclay, France.

[12] Nirenburg, S., Raskin, V.: Ontological semantics. MIT Press (2004).

[13] McShane, M., Beale, S., Nirenburg, S.: Ontosem methods for processing semantic ellipsis. In: Proceedings of the Workshop on Computational Lexical Semantics.HLT-NAACL 2004, Boston, Massachusetts, USA, Association for Computational Linguistics (May 2 - May 7 2004) 1–8.

[14] Nirenburg, S., Beale, S., McShane, M.: Baseline evaluation of WSD and semantic dependency in OntoSem. In: Proceedings of the Conference on Semantics in Text Processing. STEP '08, Stroudsburg, PA, USA, Association for Computational Linguistics (2008) 315–326.

[15] Pierce CD, Booth D, Ogbuji C, Deaton C, Blackstone E, Lenat D. 2012. SemanticDB: A Semantic Web Infrastructure for Clinical Research and Quality Reporting. Current Bioinfomatics. 7(3).

[16] Cheptsov A, Assel M, Gallizo G, Celino I, Dell'Aglio D, Bradesko L, Witbrock M, Della Valle E. 2011. Large Knowledge Collider. A Service-oriented Platform for Large-scale Semantic Reasoning . International Conference on Web Intelligence, Mining and Semantics (WIMS'11), ACM International Conference Proceedings Series, Sogndal, Norway, May 2011.

[17] Grobelnik M, Mladenić D, Witbrock M, Sammut C, Webb GI. 2010. Text Mining for the Semantic Web. Encyclopedia of Machine Learning. :Part21,978-980.

[18] Yong Meng Teo; Nat. Univ. of Singapore, Singapore; Szabo, C.CODES: An Integrated Approach to Composable Modeling and Simulation

[19] Fabian M. Suchanek, Johannes Hoffart, Erdal Kuzey, Edwin Lewis-Kelham: YAGO2s: Modular High-Quality Information Extraction with an Application to Flight Planning. BTW 2013: 515-518.