

# **ONTOLOGY EXTENSION USING TEXT MINING FOR NEWS ANALYSIS**

Inna Novalija

**Doctoral Dissertation**  
**Jožef Stefan International Postgraduate School**  
**Ljubljana, Slovenia, September 2011**

Supervisor: Prof. Dr. Dunja Mladenec, Jožef Stefan Institute and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

**Evaluation Board:**

Prof. Dr. Marko Bohanec, Jožef Stefan Institute, Ljubljana, Slovenia

Assist. Prof. Dr. Irena Nančovska Šerbec, Faculty of Education, University of Ljubljana, Slovenia

Dr. Michael Witbrock, Cypcorp, Austin, USA

**MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA**  
**JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL**



Inna Novalija

# **ONTOLOGY EXTENSION USING TEXT MINING FOR NEWS ANALYSIS**

**Doctoral Dissertation**

## **RAZŠIRITEV ONTOLOGIJE Z UPORABO METOD ANALIZE PODATKOV ZA ANALIZO NOVIC**

**Doktorska disertacija**

*Supervisor:* Prof. Dr. Dunja Mladenić

Ljubljana, Slovenia, September 2011



**To my family and friends**



# Index

<b>1 Introduction.....</b>	<b>1</b>
1.1 Background.....	1
1.2 Terminology Specification .....	3
1.3 Motivation, Aims and Hypothesis .....	4
1.4 Scientific Contributions .....	5
1.5 Thesis Structure .....	6
<b>2 Related Work .....</b>	<b>7</b>
2.1 Existing Approaches of Ontology Development.....	7
2.1.1 Natural Language Processing Based Approach .....	9
2.1.2 Pattern Based Approach .....	9
2.1.3 Networks/Graphs Based Approach .....	10
2.1.4 User Dialogue Based Approach .....	10
2.1.5 Approaches of Ontology Population .....	10
2.1.6 Other Approaches.....	10
2.2 Cyc Extension and Population.....	11
2.3 Existing Techniques of News Analysis .....	12
2.3.1 Artificial Intelligence in Business World.....	12
2.3.2 News Analysis Systems .....	12
2.3.3 Question Answering.....	13
<b>3 Materials and Methods.....</b>	<b>15</b>
3.1 Problem Definition .....	15
3.1.1 Ontology Extension Problem .....	16
3.1.2 Ontology Population Problem .....	17
3.2 Methodology for Ontology Extension .....	18
3.2.1 Phases of Methodology for Ontology Extension.....	20
3.2.2 Text Mining Usage.....	24
3.2.3 Concept-similarity Identification .....	24
3.2.4 Relation Extraction.....	24
3.2.5 User Interaction .....	24
3.3 Methodology Adaption for Cyc Ontology .....	25
3.3.1 Structure of Cyc Knowledge Base .....	25
3.3.2 Specific Aspects of OntoPlus Application to Cyc Extension .....	25
3.4 Methodology for News Analysis .....	27
3.4.1 Phases of Pipeline for Business News Analysis .....	28
3.4.2 Methodology Support .....	30

<b>4 Experiments .....</b>	<b>35</b>
4.1 Evaluation Methods .....	35
4.1.1 Tagging Experiments .....	35
4.1.2 Ranking Experiments.....	36
4.1.3 Question Answering Experiments .....	38
4.2 Domains of Interest .....	38
4.2.1 Financial Domain.....	38
4.2.2 Fisheries and Aquaculture Domain.....	39
4.3 Data Description.....	40
4.4 Experimental Settings .....	41
4.4.1 Tagging Experiments .....	41
4.4.2 Ranking Experiments.....	42
4.4.3 Question Answering Experiments .....	42
<b>5 Results.....</b>	<b>45</b>
5.1 Tagging Experiments .....	45
5.2 Ranking Experiments .....	45
5.2.1 Concept Ranking.....	46
5.2.2 Relation Ranking .....	50
5.2.3 Examples of Cyc KB Extension .....	54
5.3 Question Answering Experiments.....	58
5.3.1 Illustrative Question Answering.....	58
5.3.2 News Based Question Answering.....	62
5.4 Software Demonstration.....	67
<b>6 Discussion .....</b>	<b>71</b>
6.1 Methodology for Ontology Extension .....	71
6.2 Pipeline for Business News Analysis.....	73
<b>7 Conclusions .....</b>	<b>75</b>
7.1 Summary .....	75
7.1.1 Methodology for Ontology Extension .....	76
7.1.2 Pipeline for Business News Analysis .....	77
7.2 Scientific Contributions .....	78
7.3 Future Work .....	80
<b>8 Acknowledgements .....</b>	<b>81</b>
<b>9 References.....</b>	<b>83</b>
<b>Index of Figures .....</b>	<b>89</b>
<b>Index of Tables .....</b>	<b>91</b>
<b>Index of Algorithms .....</b>	<b>93</b>



<b>Appendix 1: OntoPlus Applications in Financial Domain .....</b>	<b>95</b>
<b>Appendix 2: OntoPlus Applications in Fisheries &amp; Aquaculture Domain.....</b>	<b>99</b>
<b>Appendix 3: OpenCalais to Cyc Mappings .....</b>	<b>101</b>
<b>Appendix 4: Example of Financial News Analysis .....</b>	<b>107</b>
<b>Appendix 5: Example of News Annotation with Cyc Tagger .....</b>	<b>111</b>
<b>Appendix 6: Publications .....</b>	<b>115</b>
<b>Appendix 7: Biography .....</b>	<b>117</b>



## Abstract

In computer science, ontologies enable formalized knowledge representation. The goal of ontology extension is to correctly augment the existing ontology with new formalized knowledge (e.g., concepts, relationships etc.).

This thesis addresses the ontology extension process based on text mining methods. News analysis is the application of the extended ontology. A novel **OntoPlus** methodology introducing usage of the ontology content, structure and the co-occurrence information is proposed for semi-automatic ontology extension. The **OntoPlus** methodology allows transforming textual information into a structured conceptualized form. The **OntoPlus** methodology is able to perform within different domains and different information sources. The methodology enables extension of very large multi-domain ontologies.

The proposed **OntoPlus** methodology is evaluated using a well known Cyc ontology and textual material from two domains – financial domain and fisheries & aquaculture domain. We have found that the best results are achieved by combining content, structure and co-occurrence information, where the combination of weights depends on the domain. In our case, the ontology content and structure are more important than co-occurrence for data in financial domain. At the same time, the ontology content and the co-occurrence have higher importance for data in fisheries & aquaculture domain.

The thesis also addresses the process of business news analysis by (1) the ontology extension with relevant concepts, (2) ontology population with entities, facts and events extracted from text and (3) reasoning based on the obtained ontology. We introduce a **pipeline for business news analysis**, which utilizes entity, event and fact extraction service, the **OntoPlus** methodology and the Cyc ontology. Furthermore, we populate the Cyc ontology with a set of entities, events and facts extracted from a collection of financial news. We use ontology structural and lexical features for obtaining matches between existing ontology instances and new instances extracted from the Web. The **pipeline for business news analysis** constitutes a whole strategy of business news analysis and question answering based on the ontology reasoning and information from the news.

The experimental results demonstrate that using the proposed **OntoPlus** methodology, based on the combination of the ontology content, structure and the co-occurrence information and using the proposed **pipeline for business news analysis** provide a potential to aggregate new knowledge into the existing ontology. The user obtains a support in analysis of financial texts and business information.



## Povzetek

Ontologije v računalništvu omogočajo formalno predstavitev znanja. Cilj razširitve ontologije je, da pravilno poveča obstoječo ontologijo z novim formaliziranim znanjem (npr. s pojmi, odnosi, itd.).

Disertacija obravnava procese razširjanja ontologije na osnovi metod analize besedil in uporabo tako razširjene ontologije pri analizi novic. Za polavtomatsko razširitev ontologij predlagamo novo metodologijo **OntoPlus**, ki uvaja uporabo vsebine in strukture ontologije ter informacijo o sopojavitvah pojmov v besedilih. Metodologija **OntoPlus** omogoča preoblikovanje besedila v konceptualizirano obliko, lahko jo uporabimo na različnih področjih in z različnimi viri informacij. Metodologija omogoča razširitev ontologije tudi v primerih, ko le-ta pokriva več domen.

Predlagano metodologijo **OntoPlus** smo ocenili z uporabo znane ontologije Cyc in besedil iz dveh domen – finančne domene in domene ribištva in ribogojstva. Ugotovili smo, da se najboljši rezultati dosežejo s kombiniranjem vsebine ontologije, strukture ontologije in sopojavitve pojmov, pri čemer je delež prispevka vsakega od teh treh vidikov odvisen od domene in podanih virov podatkov. V našem primeru sta za podatke v finančni domeni vsebina in struktura ontologije bolj pomembni kot sopojavitve. Po drugi strani sta v domeni ribištva in ribogojstva bolj pomembni vsebina ontologije in sopojavitve.

Disertacija se ukvarja tudi s procesom analize poslovnih novic s pomočjo razširitve ontologije z ustreznimi pojmi in primeri, izločenimi iz besedila. Predlagamo **cevovod za analizo poslovnih novic**, ki uporablja izločanje entitet, dogodkov in dejstev, metodologijo **OntoPlus** in ontologijo Cyc. Poleg tega smo ontologijo Cyc razširili z množico entitet, dogodkov in dejstev, izločenih iz zbirke finančnih novic. Pri tem smo uporabili strukturo ontologije in leksikalne značilnosti vsebine ontologije, da bi našli ustreznice med primeri obstoječe ontologije in novimi primeri, izločenimi iz besedil s svetovnega spleta. **Cevovod za analizo poslovnih novic** predstavlja celotno strategijo analize poslovnih novic in odgovorov na vprašanja, ki temeljijo na obrazložitvi s pomočjo ontologije in na informaciji iz novic.

Izsledki poskusov kažejo, da uporaba predlagane metodologije **OntoPlus**, temelječa na kombinaciji vsebine in strukture ontologije ter sopojavitve informacij, kakor tudi uporaba predlaganega **cevovoda za analizo poslovnih novic**, omogoča odkrivanje novega znanja v obstoječih podatkih. S tem uporabnika podpremo pri analizi finančnih besedil in poslovnih podatkov.



## Abbreviations

AI	=	Artificial Intelligence
API	=	Application Programming Interface
ASFA	=	Aquatic Sciences and Fisheries Abstracts
AURA	=	Automated User-Centered Reasoning and Acquisition System
Cyc KB	=	Cyc Knowledge Base
DIM	=	Domain Information Module
DSEM	=	Domain Subset Extraction Module
FCA	=	Formal Concept Analysis
FOPC	=	First-Order Predicate Calculus
HR	=	Hit Rate
IPO	=	Initial Public Offering
IT	=	Information Technology
KE	=	Knowledge Entry
LA	=	Learning Accuracy
Mt	=	Microtheory
NC	=	Number of Concepts
OEM	=	Ontology Extension Module
QA	=	Question Answering
SRA	=	Semantic Research Assistant
TFIDF	=	Term Frequency - Inverse Document Frequency





# 1 Introduction

This chapter introduces the terminology used in this thesis, presents the motivation, the hypothesis and the goals of the performed work, and provides a list of specific scientific contributions of the thesis.

## 1.1 Background

Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries (W3C, 2011). Semantic technology is a general term referring to approaches and software that involves some kind and level of understanding the meaning of the information it deals with (Semantic Technology and Linked Data Annotation, 2011). Ontologies are considered one of the pillars of Semantic Web and Semantic Technologies (Semantic Web. Ontology, 2011).

Gruber (1993) defined Ontology as an explicit specification of a conceptualization consisting of the following main components: concepts, relations, functions, axioms and instances. Furthermore, ontologies enable effective domain knowledge representation, knowledge sharing and knowledge reuse (Chandrasekaran et al., 1999).

This thesis explores the process of ontology extension motivated by potential usage of the extended ontology for the analysis of textual information. For instance, in question answering based on the news articles from the financial domain using semantic information enables providing better answers, assuming that the semantic information matches content of the articles. Usage of ontologies allows to search not only within the terms occurring in the query, but also within their semantically related concepts.

According to Jarrar (2005), the following challenges should be dealt with while building any kind of ontology: Ontology reusability, Ontology application/task-independence and Ontology evolution. Ontology reusability implies the maximization of ontology's use for different purposes, across different applications and tasks. Ontology application/task-independence suggests that ontologies capture semantics at the domain level and are independent of application requirements. Ontology evolution deals with continuous growth and intensive maintenance of the emerging ontologies.

News reports are considered to be one of the largest sources of information about society. News has several characteristics that differentiate it from other domains, such as containing many named entities (i.e., names of people, companies, locations, etc.). Following that, applying semantic technologies has been found beneficial for analyzing news (Grobelnik and Mladenić, 2004). Therefore, large common sense ontologies can be an effective instrument for news analysis.

However, there exist several challenges while using text mining and semantic technologies in news analysis, including the fact that news agencies produce huge amounts of content. Additionally, news sources are dynamic, interactive and socially biased. We can also find news on the same event reported by different agencies of diverse

opinion, in different natural languages.

In our research we select the Cyc ontology (Cycorp, 2011) as a foundation for ontology extension, ontology population and news analysis. The Cyc Knowledge Base is a common sense ontology, which has been being developed for more than 20 years (more than 900 human years of effort) and is used as a knowledge source in the Cyc Artificial Intelligence system. It already aggregates more than 15.000 predicates, 300.000 concepts and 3.500.000 assertions (Cycorp, what's in Cyc, 2011), but the knowledge is still very sparse in various domains. For example, the annotation experiments, conducted on a randomly selected subset of business news, have shown the insufficient representation of financial domain in Cyc and the ways to effectively improve it by the means of the Cyc Knowledge Base extension (Novalija and Mladenović, 2008). Cyc was characterized as relatively sparse and very tangled hierarchy by Noy and Hafner (1997). Manual building of large ontologies, such as the Cyc Knowledge Base, demands a substantial amount of human efforts, which is the reason that all domains are not covered yet in all details. Further extension of such a large ontology is challenging as well because of its complexity and interconnectivity. The **OntoPlus** methodology proposed in this research is meant to speed up the process of building an extensive ontology and to make ontology development more efficient.

In this thesis, the news sources from business and financial domain are considered. The extension of Cyc was needed to cover business and financial domain and, as expected, it is a demanding task. According to Zhang et al. (2000), in the financial environment the tasks are dynamic, distributed, global, and heterogeneous in nature. They are characterized by the large amount of continually changing, and generally unorganized, information available, the variety of all kinds of information (like market data, financial report data, breaking news, etc.) and many sources of uncertainty in the environment. Martínez Montes et al. (2005) as well mention several reasons explaining why the creation of the ontologies in business and financial domain is difficult. Slow standardization efforts and high complexity of the financial standards, high competition and dynamics of the financial sector influence the implementation of the new technologies. Consequently, there exist a very few ontologies connected to the financial sphere of life. At the same time, there is a high necessity in the creation of the extensive financial ontologies which could be effectively used and reused by the financial institutions.

In order to analyze news, we apply ontology based text annotation and ontology based reasoning. Ontologies are commonly used for annotating textual data mainly based on human language technologies (Bontcheva et al., 2006). The reasoning systems operate based on logically formulated knowledge. As Panton et al. (2006) state, in order to mimic human reasoning, Cyc uses background knowledge regarding science, society and culture, climate and weather, money and financial systems, health care, history, politics, and many other domains of human experience.

What is obvious or natural for a human mind, should be “explained” to the machine using formalized knowledge representation and logical rules. For instance, depending on the context, *palm* can have different meaning:

*The coconut **palm** is typically found along sandy shorelines → PALM is a plant*

*The doctor was holding pills on his **palm** → PALM is part of hand*

Cyc's representation language is known as CycL. It is essentially an augmented version of first-order predicate calculus (FOPC). Using CycL it is possible to correctly encode the above statements, so that the machine explicitly understands them and uses them for question answering.

In this thesis we show that ontology extension and population with formalized information extracted from business news and, reasoning based on the extended and populated ontology support the user in analysis of business and financial news.

## 1.2 Terminology Specification

In this section we define the terminology used in our research. In the process of ontology extension, the **OntoPlus** methodology, proposed in this thesis, uses the ontology content, the ontology structure information and the co-occurrence data between the existing and candidate ontology concepts.

**Ontology content** of a particular concept is defined as the available textual representation of the referred concept. The ontology content includes a natural language concept denotation (lexical entries for a particular concept) and textual comments about the concept. **Ontology structure** of a particular concept is defined as the neighbourhood concepts involved in the hierarchical and non-hierarchical relations with the referred concept. For instance, an example Cyc concept *CommonStock* has the associated content "*Share; Ordinary share; The collection of Stock whose instances represent owners (stockholders) who have only a residual claim on an Organization's assets after all debts and claims generated by PreferredStocks have been met..*" and the associated structure "*Equities; shares; stocks; issues; class B stocks; class A stocks...*" etc. **Co-occurrence** information is represented by the occurrence of two or more concepts within a defined textual block. The available textual information is used to find the co-occurrences between existing ontology concepts and new domain concepts suggested for ontology extension.

**Ontology extension** in this thesis stands for adding new concepts to the existing ontology or, augmentation of the existing textual representation of the relevant concepts present in the ontology with new available textual information – extension of the concept comments, changing or adding concept denotation. By **Ontology population** we consider adding new instances of a concept (e.g., *LehmanBrothers* as *Business*, *MingchunSun* as *Person*) or relation instances (e.g., *positionOfPersonInOrganization MingchunSun LehmanBrothers Economist*) into the ontology.

When representing a new term inside ontology, the information is described at two layers – **Specification layer** and **Instantiation layer**. While structural information is defined by finding existing related ontology concepts and their relationships at the specification layer, factual information is obtained during ontology population at instantiation layer.

For example, for Cyc concept *CommercialOrganization*, which represents a subclass of *Organization*, whose primary goal is to generate a profit for its owners, usually through the buying and selling of goods or services, a specification layer includes a number of related Cyc concepts, such as: *BusinessRelatedThing*, *BankingOrFinanceCompany*, *ManufacturingOrganization* etc., a number of Cyc relations, such as: *companyIsInIndustry executiveVicePresident companyHasGeneralCounsel* etc.

The instantiation layer for *CommercialOrganization* includes instances, such as: *FileMaker-CommercialOrganization*, *Symantec-CommercialOrganization*, *ChevroletTheCompany* etc.

The relation instances at the instantiation layer of the *CommercialOrganization* are the following: (*companyIsInIndustry* *MicrosoftInc* (*IndustryOfRegionFn* *SoftwareIndustry* *UnitedStatesOfAmerica*)), (*mainBusinessActivityOfOrgOccursAt* *CVSCorp* *UnitedStatesOfAmerica*) etc.

In our research we deal both with specification and instantiation layers. Specification layer is mainly utilized in ontology extension (e.g., adding new concepts) and instantiation layer is important for ontology population (e.g., adding new instances).

### 1.3 Motivation, Aims and Hypothesis

In this thesis a novel **OntoPlus** methodology is proposed for semi-automatic ontology extension based on text mining methods. Furthermore, we suggest a **pipeline for business news analysis**, which utilizes entity, event and fact extraction services, the **OntoPlus** methodology and exploits Cyc for reasoning and question answering.

The motivation for the performed work comes from the idea that business news aggregates a large amount of interesting business, financial, organizational and personal information, which can be discovered, combined and used in reasoning and question answering by the reasoning tool-kits. Using ontologies allows reasoning within semantically related concepts and instances.

The importance of ontology extension is interconnected with the dynamic nature of the ontologies. When extending large ontologies with new concepts and populating them with new instances, it is necessary to identify the equivalent concepts and instances already present in the ontology. It is also important to find the correct location and context for new concepts and instances we insert into the ontology.

The hypothesis in this thesis is that applying the suggested methodology for ontology extension and the **pipeline for business news analysis** on a set of business news allows for the efficient ontology based reasoning and news related question answering. Moreover, the user obtains a possibility of knowledge extraction from business news, ontology extension and population with extracted concepts, entities, facts and events. Usage of the obtained extended and populated ontology for reasoning and question answering provides the user with feasibility to automatically analyze textual financial and business information, to detect important information and to save time.

The main goal of this research is to contribute to the analysis of the financial news by the means of semantic technologies - in particular by extending and populating the business and financial parts of the Cyc ontology, which is known to have one of the largest knowledge bases in the world (OpenCyc, 2011).

Furthermore, we set a goal to show that by extending the Cyc ontology with new concepts, populating Cyc with new instances extracted from business news and applying Cyc reasoning technology, we can efficiently obtain the important information “hidden” in the financial textual data, answer various questions and combine the pieces of puzzle together.

Utilizing the ontology lexical and structural features, we aim to make the process of ontology extension more productive.

In addition, in this research we set a goal to:

1. Specify the terminology used in the research;
2. Propose the methodology for ontology extension and the **pipeline for business news analysis**;
3. Adapt the proposed methodology to the Cyc Knowledge Base extension;
4. Experimentally evaluate the proposed methodology for ontology extension in several different domains and different knowledge representations;
5. Develop a user interface of the Cyc Knowledge Base extension for a selected domain;
6. Experimentally evaluate the proposed **pipeline for business news analysis**. Explore a set of financial news and determine the relevant financial concepts and instances;
7. Obtain a number of financial events and facts from a set of business news;
8. Identify a set of interesting, non-trivial queries for business news analysis.

## 1.4 Scientific Contributions

There are several scientific contributions of this doctoral dissertation.

First, in this thesis we propose a novel **OntoPlus** methodology introducing usage of the ontology content, the ontology structure information and the co-occurrence data between existing and candidate ontology concepts for ontology extension. There are several methodologies for ontology learning, which use the available lexical information from the ontology. Our methodology uses not only lexical, but also structural ontological information, which allows it to achieve better results in ontology extension and ontology population processes.

Second, we adapt the **OntoPlus** methodology for one particular scenario – the Cyc ontology extension. The suggested methodology contributes to large multi-domain ontology extension, which is rarely handled by the methodologies for ontology learning.

Third, we conduct experiments using the **OntoPlus** methodology on real world data in two different domains having two knowledge representation levels – financial domain represented by a glossary of financial terms (Harvey, 2003) and fisheries & aquaculture domain represented by Aquatic Sciences and Fisheries Abstracts (ASFA) thesaurus (ASFA thesaurus, 2010). We find that using a combination of the ontology content, structure and the co-occurrence information is more beneficial for the extension of large multi-domain ontologies, than using only content, only co-occurrence or only concept denotation information.

Forth, we propose a **pipeline for business news analysis** and explore the process of business news analysis by the ontology extension and ontology population with entities, facts and events extracted from text and reasoning based on the obtained ontology.

Fifth, we extend and populate Cyc with business and financial terms. In news analysis experiments we use a collection of the financial news from the Yahoo! Finance website (Yahoo! Finance, 2010). We crawl (automatically collect from Web) a set of news stories, extract entities, facts and events with fact extraction service (OpenCalais, 2011) and apply the **OntoPlus** methodology to map the extracted knowledge into the Cyc KB. Following that, a number of queries are analyzed through the Cyc reasoning interface.

Finally, applying the **OntoPlus** methodology and the **pipeline for business news analysis** we are able to competently deal with a specific and non-trivial task of business

news analysis, to discover, combine and use in reasoning and question answering the important information extracted from business news.

## 1.5 Thesis Structure

The thesis is structured as follows: Chapter 2 presents the related work; In Chapter 3 we discuss the materials and methods. This chapter presents the new **OntoPlus** methodology for ontology extension and contains the adaptation of the proposed methodology for one concrete scenario – the Cyc Knowledge Base extension. Furthermore, in Chapter 3 we suggest the **pipeline for business news analysis**; Chapter 4 describes the evaluation strategies, data and experiments; The results of the experiments are provided in Chapter 5; the discussion is covered in Chapter 6; we provide the results summary, conclusions and future work in Chapter 7. In addition, the thesis contains the following appendices: Appendix 1: **OntoPlus** Applications in Financial Domain, Appendix 2: **OntoPlus** Applications in Fisheries & Aquaculture Domain, Appendix 3: OpenCalais to Cyc Mappings, Appendix 4: Example of Financial News Analysis, Appendix 5: Example of News Annotation with the Cyc Tagger, Appendix 6: Publications, Appendix 7: Biography.

## 2 Related Work

In this chapter we describe the related work in the fields of ontology development and news analysis.

### 2.1 Existing Approaches of Ontology Development

The automatic and semi-automatic ontology extension processes are usually composed of several phases. Most approaches include defining the set of the relevant ontology extension sources, preprocessing the input material, ontology augmentation according to the chosen methodology, ontology evaluation and revision phases.

While developing ontologies, it is important to follow a number of ontology design criteria. While methodologies define the particular phases of ontology development (e.g., first, defining the sources, then, preprocessing material etc.), the design criteria display the principles of building the ontology (e.g., clarity, extendibility etc.).

Gruber (1995) defines the following design criteria for ontology development: Clarity, Coherence, Extendibility, Minimal encoding bias and Minimal ontological commitment. Jarrar (2005) states two additional methodological principles: the Ontology double articulation principle and the Ontology modularization principle. The ontology double articulation principle implies the idea of separate domain and application axiomatizations, where the domain axiomatization characterizes the vocabulary at the domain level and the application axiomatization focuses on the usability of this vocabulary for certain applications. The idea of the ontology modularization is that an ontology is built as a set of small modules and later composed and used as one modular ontology. The mentioned principles of ontology engineering allow for creation of highly reusable, easily built and maintained ontologies.

The classic methods and methodologies of ontology creation are: Uschold and King's method (1995), Grüninger and Fox's methodology (1994), METHONTOLOGY (Corcho et al., 2005), On-To-Knowledge (Sure and Studer, 2002).

Uschold and King's methodology for developing ontologies includes the following stages:

- Identify Purpose;
- Building the Ontology;
- Ontology capture;
- Ontology coding;
- Integrating Existing Ontologies;
- Evaluation;
- Documentation.

The methodology by Grüninger and Fox can be described by the next steps:

- Capture of motivating scenarios;
- Formulation of informal competency questions;
- Specification of the terminology of the ontology within a formal language;
- Getting informal terminology;
- Specification of formal terminology;
- Formulation of formal competency questions using the terminology of the ontology;
- Specification of axioms and definitions for the terms in the ontology within the formal language;
- Establishing conditions for characterizing the completeness of the ontology.

One the most famous and frequently used methodologies are METHONTOLOGY and On-To-Knowledge methodology. According to the developers of METHONTOLOGY, the METHONTOLOGY framework includes:

- The identification of the ontology development process;
- A life cycle based on evolving prototypes and
- the methodology itself, which specifies the steps for performing each activity, the techniques used, the products to be output, and how the ontologies are to be evaluated.

On-To-Knowledge methodology distinguishes such phases of the ontology development:

- Feasibility Study;
- Kickoff;
- Refinement;
- Evaluation and
- Application & Evolution.

Buitelaar et al. (2005) state that the process of ontology development from text can be organized in a layer cake of increasingly complex subtasks: term extraction at the bottom, synonym extraction, concept definition, establishment of concept hierarchies, relation identification and rule definition on the top. As Reinberger and Spyns (2005) state, the following steps can be found in the majority of methods for ontology learning from text: collecting, selecting and preprocessing of an appropriate corpus, discovering sets of equivalent words and expressions, establishing concepts with the help of the domain experts, discovering sets of semantic relations and extending the sets of equivalent words and expressions, validating the relations and extended concept definition with help of the domain experts and creating a formal representation. As suggested in (Grobelnik and Mladenić, 2006), ontology learning from text is just one phase in the methodology for semi-automatic ontology construction preceded by domain understanding, data understanding and task definition and followed by ontology evaluation and ontology refinement. In our research, we focus on ontology extension assuming that the main challenge is in finding the relevant concepts and relations in the existing ontology.

Prieto-Diaz (2002) utilizes top-down and bottom-up processes for ontology development. A more general top-down process embodies domain experts identifying the key concepts in order to capture the high level ontology. The instruments for the text analysis are used in the bottom-up process for keywords extraction. In a similar way, the proposed in this thesis the **OntoPlus** methodology incorporates top-down and bottom-up process, where the user is providing relevant keywords or glossary while the system uses



the data to identify relevant parts of the existing ontology.

Fortuna et al. (2007) developed an approach to semi-automatic data-driven ontology construction focused on topic ontology. The approach combines machine learning and text mining techniques with an efficient user interface. The domain of interest is described by keywords or a document collection and used to guide the ontology construction. OntoGen (Fortuna et al., 2007) uses the vector-space model for document representation. The tool operates based on applying unsupervised, semi-supervised and supervised learning methods.

In the following sections we proceed with discussion on different approaches of the ontology development.

### **2.1.1 Natural Language Processing Based Approach**

Natural language processing is notably used for learning or extending ontologies (Burkhardt et al., 2008; Sabrina et al., 2001). Unsupervised text mining for ontology learning was elaborated by Reinberger and Spyns (2005). Cimiano et al. (2005) suggest an approach for learning concept hierarchies from text based on Formal Concept Analysis (FCA), a method mainly used for the data analysis. The Web is considered a source of text suitable for ontology extension in (Agirre et al., 2000), where the English lexical ontology WordNet (WordNet, 2011) is extended based on clustering word senses. However, our approach is more general by enabling extension of any ontology that has some lexical description of the concepts.

### **2.1.2 Pattern Based Approach**

Lexico-syntactic pattern-based ontology learning is handled by Text2Onto (Cimiano and Völker, 2005), a framework for ontology learning and data-driven change discovery. The main aspects of the Text2Onto framework include using so called Probabilistic Ontology Model, user interaction and operation strategies for data-driven change discovery. Text2Onto allows learning ontological structures from text in form of modeling primitives, such as concepts, subclasses, instances etc. without connection to a certain representation language. SPRAT (Maynard et al., 2009) is a tool for automatic semantic pattern-based ontology population. SPRAT system combines named entity recognition, ontology-based information extraction and relation extraction in order to define patterns for the identification of a variety of entity types and relations between them.

SOFIE (Suchanek et al., 2009) is a system for automated ontology extension, which can parse natural language documents, extract ontological facts from them and link the facts into ontology. SOFIE uses logical reasoning on the existing knowledge and on the new knowledge in order to disambiguate words to their most probable meaning, to reason on the meaning of text patterns and to take into account world knowledge axioms. The work by Suchanek et al. (2009) resembles our approach of ontology extension applied to the Cyc ontology augmentation. Cyc extension also involves interaction with logical constraints from the knowledge base.

In our work, patterns are not utilized, as we are assuming availability of keywords or glossary terms that already represent new concepts used for ontology extension.

### 2.1.3 Networks/Graphs Based Approach

The networks/graphs methods of ontology extension include the work of McDonald et al. (1990) that apply the pathfinder networks approach, allowing for representing proximity data between pairs of items. Furthermore, the spreading activation technique was also shown to be suitable for semi-automatic ontology extension (Liu et al., 2005).

### 2.1.4 User Dialogue Based Approach

The semi-automatic approach for ontology extension presented in (Witbrock et al., 2003) is based on the user-interactive dialogue system for knowledge acquisition, where, the user is engaged in a natural-language mixed-initiative dialogue. The system contains a natural language generation module, parsing module, post-processing module, dictionary assistant, user interaction agenda and salient descriptor.

In our approach the user plays an important role validating the proposed new formalized knowledge.

### 2.1.5 Approaches of Ontology Population

A number of approaches for automatic ontology population have proven themselves as effective tools of information extraction. Very often pattern based approaches for ontology learning from text are used for ontology population.

Described first by Hearst (1992), the pattern based approach for instance and hyponym extraction uses a defined set of patterns while analyzing textual sources. Etzioni et al. (2000) developed a KnowItAll system for named entity classification. The approach performs pattern learning and can iteratively obtain new rules and new seeds.

Carlson et al. (2010) present a method of coupling the semi supervised learning of category and relation instance extractors for ontology population with category and relation instances. In (Carlson et al., 2010) a number of categories (e.g., academic fields, athletes) and relations (e.g., *PlaysSport(athlete, sport)*) are extracted from Web pages, starting with a handful of labeled training examples of each category or relation, plus hundreds of millions of unlabeled Web documents. Carlson et al. (2010) state that much greater accuracy can be achieved by further constraining the learning task, by coupling the semi-supervised training of many extractors for different categories and relations.

Several methods discussed below operate with enlarging of the Cyc Knowledge Base (Cyc KB) (e.g., Sarjant et al., 2009; Shah et al., 2006).

### 2.1.6 Other Approaches

Extension of the existing ontology by automatically extending its relations was addressed by several researchers. The approaches include learning taxonomic (Cimiano et al., 2004)/non-taxonomic relations (Maedche and Staab, 2000) and extracting semantic relations from text based on collocations (Heyer et al., 2001). However, in our work we first of all, assume suggesting relevant existing relation instances.

Turney (2001) has used a co-occurrence analysis technique for mining synonyms from Web. Besides, the ontology structure has been adequately used in the collective entity resolution (Štajner and Mladenić, 2009). Usage of the co-occurrence in the **OntoPlus**

methodology presented in our research was inspired by the work on collective entity resolution and synonym extraction.

## 2.2 Cyc Extension and Population

Several methods of automatic ontology extension operate with enlarging of the Cyc Knowledge Base. As it was stated by Lenat (1995), one can think of Cyc as an expert system with a domain that spans all everyday objects and actions. For example:

- *You have to be awake to eat.*
- *You can usually see people's noses, but not their hearts.*
- *Given two professions, either one is a specialization of the other or else they are likely to be independent of one another.*
- *You cannot remember events that have not happened yet.*
- *If you cut a lump of peanut butter in half, each half is also a lump of peanut butter; but if you cut a table in half, neither half is a table.*

Building of the Cyc ontology was initiated over 20 years ago. According to Cyc method (Lenat and Guha, 1990), the phases to build the Cyc ontology are following:

- Manual encoding of the explicit and implicit knowledge appearing in the knowledge sources;
- Knowledge codification that is aided by tools using knowledge already stored in the Cyc KB;
- Delegating to the tools the majority of the work.

In each phase two tasks are performed: 1. Development of a knowledge representation and top level ontology containing the most abstracts concepts. 2. Representation of the rest of the knowledge using these primitives.

The automated population of Cyc with named entities involves the Web and a framework for validating candidate facts (Shah et al., 2006). The user-based dialogue system for the Cyc KB extension was presented by Witbrock et al. (2003). Medelyan and Legg (2008) describe the methodology for integrating Cyc and Wikipedia, where the concepts from Cyc are mapped onto Wikipedia articles describing correspondent concepts. Sarjant et al. (2009) use Medelyan and Legg (2008) method to augment the Cyc ontology using pattern matching and link analysis.

Taylor et al. (2007) have conducted research on Cyc microtheories. In their work, they considered the problem of how to automatically determine where to place new knowledge into an existing ontology.

An interesting approach of extending and using Cyc for answering clinical researchers' ad hoc queries is described in (Lenat et al., 2010). Even long and complex queries are parsed into CycL fragments, which often can be united together only in a single way after applying various constraints. The developed Semantic Research Assistant (SRA) performs a set of database calls and then combines their results into answers to a specified query.

In our research Cyc extension and population plays a substantial role. Selected for our research task of news analysis, Cyc aggregates large common-sense ontology, suitable for news formalization, and an inference system, which allows performing reasoning based on the formalized knowledge.

## 2.3 Existing Techniques of News Analysis

The analysis of news sources represents an important research challenge of our times. News not only reflects the different processes happening in the world, but also influences the economic, political and social situation. Moreover, news sources contain an enormous amount of information, which can be compiled together and analyzed through reasoning and question answering.

The study of business news is interdisciplinary combining artificial intelligence techniques and financial data analysis.

### 2.3.1 Artificial Intelligence in Business World

The discoveries and developments in the 21<sup>st</sup> century Artificial Intelligence (AI) have changed the ambitions of scientists in different research areas. As Duong (2008) stated in her article about the industrial application of artificial intelligence, AI has transformed our way of thinking and solving problems, has changed the consumer behaviors and improved quality of life. Duong suggested that higher forms of AI would be able to increase the productivity of the economy.

On the other side, the information technology community have developed and implemented a number of systems, applying various AI approaches for economic purposes and business news analysis. The information technology artifacts for business and financial tasks include, for instance, such tool as ATRANS (Lytinen and Gershman, 1986). Developed in 1986, ATRANS was created to operate in the domain of international banking telexes. This historically important system automatically extracted the information required to complete the transfer (the various banks mentioned in the telex, their roles in the money transfer, payment amounts, dates, security keys, etc.) and formatted it for entry into the bank's automated transaction processing system. For text analysis, the ATRANS developers utilized case-frame analysis and conceptual dependency formalism.

AI techniques, methods and tools constitute a central part of our research. The use of the Cyc ontology and the Cyc reasoning interface allows us to effectively analyze queries based on the data obtained from the news.

### 2.3.2 News Analysis Systems

Starting from 1990s, a number of other systems dealing with news analysis have been developed (Andersen et al., 1992; Losch and Nikitina, 2009; Iacobelli et al., 2010). JASPER (Andersen et al., 1992) is a fact extraction system developed and deployed by Carnegie Group for Reuters Ltd., which uses a template-driven approach, partial understanding techniques, and heuristic procedures to extract certain key pieces of information from text. JASPER combines frame based knowledge representation with object-oriented processing, pattern matching, and heuristics which allows it to analyze textual sources efficiently and quickly.

The newsEvents Ontology developed by Losch and Nikitina (2009) allows modeling of business events, the affected entities and relations between them. Losch and Nikitina use a pattern-based approach with defined and specified EventRole patterns for ontology

design.

Iacobelli et al. (2010) have presented a system called Tell Me More, which mines Web for news stories based on the seed news story and selects snippets of text from those stories which offer new information beyond the seed story. The obtained new content is classified as supplying: additional quotes, additional actors, additional figures and additional information depending on the criteria used to select it.

Several interesting approaches, which can be applied to news analysis, deal with data mining and knowledge extraction from Web.

In (Chang et al., 2006) the authors compare the existing Web data extraction approaches. Ghani et al. (2000) demonstrated the possibility to discover interesting regularities about companies by extracting, and then mining information on the Web. Etzioni et al. (2008) presented an Open Information Extraction from Web wherein the identities of the relations to be extracted are unknown and the billions of documents found on the Web necessitate highly scalable processing.

Soderland et al. (2010) performed research on the adaptation of the Open information extraction to domain-specific relations. The key ideas of this approach include domain specific class recognition with minimum manual effort, learning rules for relation extraction based on limited training data and active learning over learned rules to increase precision and recall.

### 2.3.3 Question Answering

Since part of our research includes utilization of large common-sense ontologies for reasoning and question answering, in this subsection we present a number of different question answering systems. The discussed question answering systems are mainly based on large knowledge bases or ontologies (Tunstall-Pedoe, 2010; Gunning et al., 2010; Bradeško et al. 2010).

In Tunstall-Pedoe (2010) a system called True Knowledge is described. True Knowledge is a commercial, open-domain question-answering platform. True Knowledge aggregates large knowledge base of common sense, factual and lexical knowledge, natural language translation system and inference system.

An interesting approach was taken by the IBM team (Ferrucci et al., 2010), which created a system to play in a quiz show Jeopardy. In order to develop Watson (Ferrucci et al., 2010), researchers have used a number of machine learning techniques and developed DeepQA architecture. DeepQA is a massively parallel probabilistic evidence-based architecture. The overarching principles in DeepQA are massive parallelism, many experts, pervasive confidence estimation and integration of shallow and deep knowledge.

The Halo project (Gunning et al., 2010) was updated to include design and evaluation of a tool called AURA, which enables domain experts in physics, chemistry and biology to author a knowledge base and then allows a different set of users to ask novel questions against that knowledge base.

Bradeško et al. (2010) have presented a system which enables contextualized question answering and provides document overview functionalities. Based on ontologies and domain specific document collections, the system is able to obtain a high number of relevant answers. The system employs AnswerArt (Dali et al., 2009) technology for question answering and the Cyc ontology for providing semantic context to the document collection from a particular domain of interest. In order to contextualize question

answering, Bradeško et al. (2010) have selected ASFA abstracts for a document collection and extended Cyc (Cycorp, 2011) by using WordNet (WordNet, 2011) and ASFA ontology (ASFA thesaurus, 2010). For their experiments, Bradeško et al. (2010) have used the subject-predicate-object triplets extracted from ASFA documents. The technology provided by Bradeško et al. (2010) successfully illustrates how the extended ontology can contribute to question answering. In our research the same resources (the Cyc ontology and the ASFA thesaurus) are used in different type of question answering.

### 3 Materials and Methods

This chapter describes the methodologies proposed in our research. We provide a problem definition, a detailed specification of the **OntoPlus** methodology applied in the processes of ontology extension and ontology population, a methodology adaption for the Cyc ontology extension and a **pipeline for business news analysis**.

#### 3.1 Problem Definition

In this section we describe the formal background of the proposed methodology. According to Maedche and Staab (2001), Ontology ( $O$ ) is a tuple:

$$O := \{L, C, H_C, R, H_r, F, G, A\} \quad (1)$$

where

- $L$  represents lexical entries for concepts and relations;
- $C$  is a set of concepts;
- $H_C$  is a taxonomy of concepts, where taxonomy represents a collection of terms organized hierarchically;
- $R$  is a set of non-taxonomic relations;
- $H_r$  is a set of taxonomic relations;
- $F$  and  $G$  are the relations connecting concepts and relations with lexical entries from  $L$ ;
- $A$  is a set of axioms.

In order to formalize the ontology extension and population in more detail, we adapt the Maedche and Staab ontology definition in the following way:

$$O := \{L, C, H_C, R, H_r, I, F, G, K, A\} \quad (2)$$

Where, additionally

- $I$  is a set of instances (individuals like *GeorgeWBush*, *LehmanBrothers* etc.);
- $K$  are the relations connecting instances with lexical entries from  $L$ ;

### 3.1.1 Ontology Extension Problem

Following the ontology definition and the ontology extension problem can be defined in the following way (Fortuna et al., 2007):

$$f^e : (O, T^E) \rightarrow O^E \quad (3)$$

- $f^e$  is a transformation from existing ontology and textual sources to extended ontology;
- $O$  represents an existing ontology, which we are extending;
- $T^E$  is a domain glossary - textual source of information we use for ontology extension;
- $O^E$  is an extended ontology.

The proposed **OntoPlus** methodology enables extending the existing ontology by (a) adding a new hierarchically related concept, by (b) augmenting textual representation of the existing concept or, by (c) adding new axioms.

(a) The following formula corresponds to adding a new hierarchically related concept to the existing ontology  $O$  (2):

$$f_{HRC} : (L, C, H_c, R, H_r, I, F, G, K, A) \rightarrow (L \cup \{l_c\}, C \cup \{c\}, H_c^e, R, H_r, I, F^e, G, K, A) \quad (4)$$

- $c$  is a new hierarchically related concept;
- $l_c$  is a lexical entry for a new hierarchically related concept;
- $H_c^e$  is an extended taxonomy of concepts (by extending taxonomy we consider e.g., adding new concept to the existing taxonomy);
- $F^e$  is an extended set of relations connecting concepts with lexical entries from  $L$  (by extending a set of relations connecting concepts with lexical entries we consider e.g., adding new lexical entry for a specific ontology concept).

For instance, in the Appendix 1 we provide the following example of the Cyc ontology extension. The financial glossary term *DEALER\_LOAN* is being added to the Cyc ontology. Hence, the extended taxonomy of concepts includes the new ontology concept *DealerLoan*, which is a SUBCLASS of *LoanAgreement*.

(b) Augmentation of the existing textual representation of the relevant concepts of the existing ontology  $O$  (2) with new lexical entries is displayed as:

$$f_{LE} : (L, C, H_c, R, H_r, I, F, G, K, A) \rightarrow (L \cup \{l_c\}, C, H_c, R, H_r, I, F^e, G, K, A) \quad (5)$$

For instance, for the Cyc ontology concept *OrganizationMerger* the existing textual



representation can be augmented with new lexical entry from the financial glossary: “*Business combination*”.

(c) Adding a new axiom to the existing ontology  $O$  (2) is presented in the following way:

$$\begin{aligned} f_A : (L, C, H_C, R, H_R, I, F, G, K, A) \rightarrow \\ (L, C, H_C, R, H_R, I, F, G, K, A^e) \end{aligned} \quad (6)$$

- $A^e$  is an extended axiom set (by extending a set of axioms we consider e.g., adding an associative relationships between two existing ontology concepts).

In Appendix 2, for instance, a set of Cyc axioms is extended with associative relationship *Chloroplasts* is CONCEPTUALLY RELATED to *Pigment*.

### 3.1.2 Ontology Population Problem

We define the ontology population task analogically to the ontology extension task:

$$f^P : (O, T^P) \rightarrow O^P \quad (7)$$

- $f^P$  is a transformation from existing ontology and textual sources to populated ontology;
- $T^P$  is textual source of information we use for ontology population;
- $O^P$  is a populated ontology.

The **pipeline for business news analysis** enables population of the existing ontology by adding (a) a new instance, by (b) adding a new fact. The following formula corresponds to adding a new instance to the existing ontology  $O$  (a):

$$\begin{aligned} f_{HRI} : (L, C, H_C, R, H_R, I, F, G, K, A) \rightarrow \\ (L \cup \{l_i\}, C, H_C, R, H_R, I \cup \{i\}, F, G, K^e, A^e) \end{aligned} \quad (8)$$

- $i$  is a new possible instance;
- $l_i$  is a lexical entry for a new instance;
- $K^e$  is an extended set of relations connecting instances with lexical entries from  $L$ ;

For example, in Appendix 4, from business news a new Company instance *BacheCommoditiesLimited* with lexical entry “*Bache Commodities Limited*” is extracted and added to the Cyc ontology. The Cyc ontology is also populated with Person instance *ChristopherBellew* (lexical entry: “*Christopher Bellew*”).

(b) Adding a new fact or event to the existing ontology  $O$  (2) is presented in the following way:

$$f_{FE} : (L, C, H_c, R, H_r, I, F, G, K, A) \rightarrow (L \cup \{l_{i1}, l_{i2} \dots l_{in}\}, C, H_c, R, H_r, I \cup \{i_1, i_2 \dots i_n\}, F, G, K^e, A^e) \quad (9)$$

- $i_1, i_2 \dots i_n$  are new possible instances;
- $l_{i1}, l_{i2} \dots l_{in}$  are new possible lexical entries for new instances;

Also, Appendix 4 illustrates how a new fact (*positionOfPersonInOrganization ChristopherBellew BacheCommoditiesLimited SeniorVicePresident-CorporateOfficer*) is added to the Cyc ontology.

### 3.2 Methodology for Ontology Extension

As a cardinal part of our research, we propose a new **OntoPlus** methodology for text-driven ontology extension, which combines text mining methods with user-oriented approach and supports the extension of multi-domain ontologies.

Since the **OntoPlus** methodology is user based, users play a substantial role in the **OntoPlus** application. The target user group for the **OntoPlus** methodology is a group of ontology engineers, who build and maintain large common sense ontologies.

The creation of the **OntoPlus** methodology is motivated by the fact that large common sense ontologies with thousands of different concepts, instances and relationships are hard to develop. Hence, the goal of the **OntoPlus** methodology is in providing the ontology engineers with a simple way to transform the unstructured textual information into a structured ontological form. We have developed an application for the Cyc ontology augmentation, which allows the user to extend the Cyc ontology with information obtained from the specified domain glossary. Taking the unstructured textual information (domain glossary and domain keywords) and ontology as an input, the system is required to provide the user with formalized ontological knowledge. The software application is discussed in chapter 5.

The proposed methodology for ontology extension includes the following phases: Domain information identification, Extraction of the relevant domain ontology subset from a multi-domain ontology, Domain relevant information preprocessing, Composing the list of potential concepts and relationships for ontology extension, User validation, Ontology extension, Ontology reuse. The detailed description of seven methodology phases and application of the methodology for extension of the Cyc Knowledge Base are given below.

The proposed methodology embodies three main modules: the Domain Information Module (DIM), the Domain Subset Extraction Module (DSEM) and the Ontology Extension Module (OEM). The graphical illustration of methodology modules is provided below in Figure 1. Each of the modules pursues its own goal – i.e., the goal of the Domain Information Module is in user identification of the domain relevant information; the goal of the Domain Subset Extraction Module is in automatically acquiring the existing ontology concepts and relationships for a specified domain; the goal of the

Ontology Extension Module is in transforming textual resources specified by user to valid ontological form.

The main task of the Domain Information Module is accumulating the relevant domain information, needed for the ontology extension. The Domain Information Module contains the domain keywords, determined by the user and a domain relevant glossary of terms with descriptions.

In the Domain Subset Extraction Module initially the multi-domain ontology is limited to the particular domains of interest. For instance, the Upper-Level Domain Extractor extracts the domains of interest, in the way they are represented, from multi-domain ontology. Subsequently, the domain related knowledge is extracted from the ontology by the Domain Knowledge Extractor and the Relevant Ontology Subset is obtained.

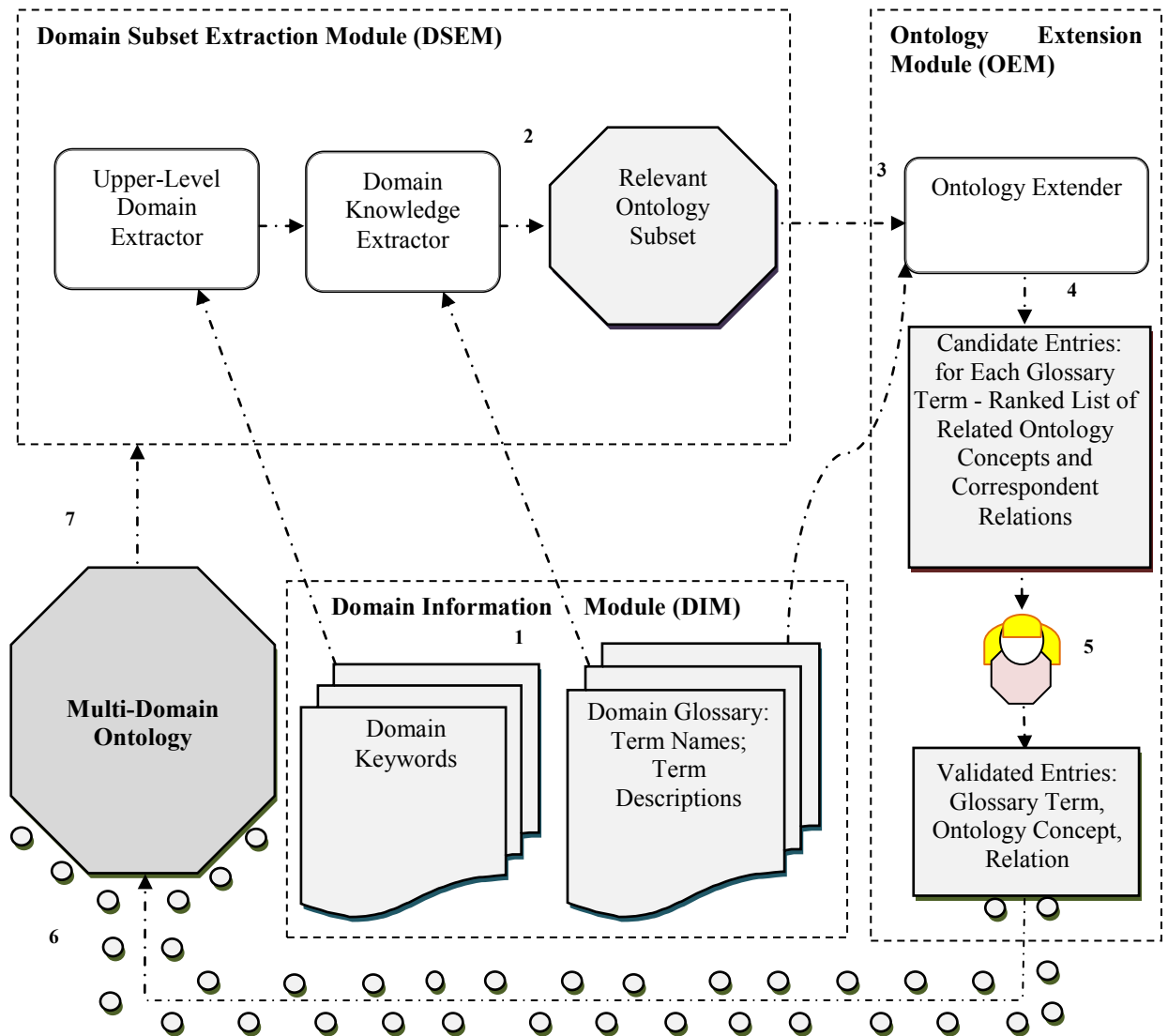


Figure 1: *Text-Driven Ontology Extension (OntoPlus Methodology).*

The Ontology Extension Module is the most important methodology module where the actual procedure of ontology extension takes place. For each glossary term the Ontology Extender from OEM produces the ranked list of textually related existing ontology concepts. In addition, the relationship for every term-concept pair is suggested. Based on the suggested list of related concepts with relationships, the user makes a decision of which terms from the domain relevant glossary should be added to the ontology and how the glossary terms should be connected with the existing ontology concepts. Ontology is extended after the user validation (the new formalized knowledge displayed with little circles in Figure 1 is added to the ontology). The technical aspects of the proposed methodology are described in the remaining of this section.

### 3.2.1 Phases of Methodology for Ontology Extension

In detail, the proposed methodology for text-driven ontology extension accounts for the following phases (illustrated with number in Figure 1):

1. *Domain information identification.* This takes place in the Domain Information Module. The user (e.g., ontology engineer) identifies the appropriate domain keywords. As well, in this module a domain relevant glossary, containing terms with descriptions is determined. We assume that the glossary terms are the candidate entry concepts for the existing ontology. Consequently, the glossary terms might be in the following relationships with the existing ontology concepts:

- Equivalence relationship: the candidate concept represented by a glossary term is equivalent to the existing ontology concept;
- Hierarchical relationship: the candidate concept represented by a glossary term is in the superclass-subclass relationship with an existing ontology concept;
- Non-hierarchical relationship: the candidate concept represented by a glossary term is in the associative relationship with an existing ontology term. The nature of the relationship is not hierarchical;
- No relationship: the candidate concept represented by a glossary term is not related to the existing ontology concept.

2. *Extraction of the relevant domain ontology subset from multi-domain ontology.* Extraction of the relevant domain ontology subset from multi-domain ontology based on the specified domain information takes place in the Domain Subset Extraction Module. In case of large common-sense ontologies, such as the Cyc Knowledge Base, the user entering new knowledge very often needs a particular ontology subset of his domain interest.

Therefore, the domain keywords are mapped to the natural language representation of the ontology domain information and a set of the relevant domains of interest is identified.

Further, ontology concepts defined in these domains are extracted. By concept extraction we mean obtaining the content and structure of the ontology concept. Correspondently, we find the textual representation (natural language denotation and comments) as content for the particular ontology concept. The ontology structure of the particular concept is represented by the natural language denotations of the hierarchically and non-hierarchically connected ontology concepts. Besides that, the names of the glossary terms are mapped to the natural language denotations of the concepts from other domains and the corresponding concepts are also extracted.

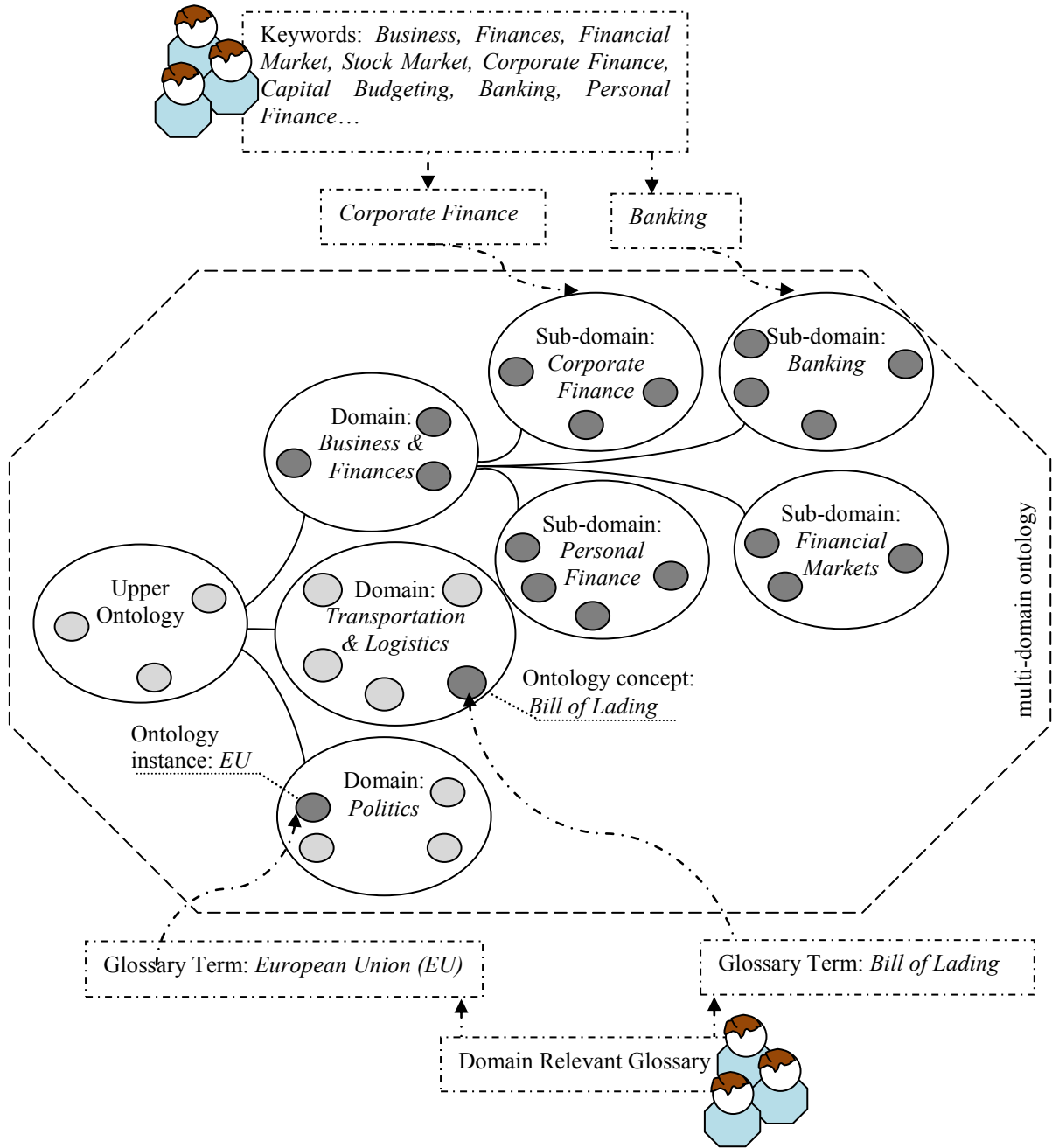


Figure 2: *Illustrative Extraction of Business & Finances Domain Subset from a Multi-Domain Ontology. Dark circles represent the extracted relevant concepts.*

Figure 2 demonstrates the extraction of Business & Finance knowledge subset from the multi-domain ontology. The ontology contains three domains - Business & Finance, Transportation & Logistics and Politics.

The domain relevant glossary is composed of the financial terms with descriptions. The correspondent financial concepts are extracted from the Business & Finances domain defined by the user-specified keywords. Moreover, concepts from other domains with concept denotations equivalent to the glossary term names are extracted.

3. *Domain relevant information preprocessing.* The information from the domain relevant glossary and the extracted relevant ontology subset are linguistically preprocessed in the Ontology Extension Module. The preprocessing phase includes

tokenization, stop-word removal and stemming. Textual information is represented using a bag-of-words representation with TFIDF weighting and similarity between two text segments is calculated using cosine similarity between their bag-of-words representations, as commonly used in text mining (Grobelnik and Mladenić, 2006). For each term from the domain relevant glossary we compose a bag-of-words aggregating preprocessed textual information from: (1) the glossary term name and (2) the term comment. For each concept from the extracted relevant ontology subset the following information is considered: (1) the ontology concept content consisting of the preprocessed natural language concept denotation and concept comment; (2) the ontology concept structure consisting of the preprocessed natural language concept denotation and natural language denotations of hierarchically and non-hierarchically related concepts. In addition, for relation identification, for each ontology concept we compose two additional bags-of-words: one with natural language denotation of the concept and natural language denotations of superclasses of this concept, another with natural language denotation of the concept and natural language denotations of subclasses of this concept.

4. *Composing the list of potential concepts and relationships for ontology extension.* The ranked list of the relevant concepts and possible relationships suitable for ontology extension is composed in this phase.

The combined content, structure and co-occurrence similarity,  $similarity(t,c)$ , is used to rank ontology concepts for each glossary term:

$$\begin{aligned}
 similarity(t,c) := & \delta_1 * similarity_{content}(t,c) \\
 & + \delta_2 * similarity_{structure}(t,c) \\
 & + \delta_3 * similarity_{co-occur}(t,c)
 \end{aligned} \tag{10}$$

$$\sum_{i=1}^3 \delta_i = 1$$

The cosine similarity  $similarity_{content}(t,c)$  between glossary term  $t$  and ontology concept  $c$  content is calculated and weighted with weight  $\delta_1$  defined by the user. The cosine similarity  $similarity_{structure}(t,c)$  between glossary term  $t$  and ontology concept  $c$  structure is calculated and weighted with weight  $\delta_2$ . We use Jaccard similarity to measure the co-occurrence of glossary term  $t$  and ontology concept  $c$ :

$$similarity_{co-occur}(t,c) = \frac{N(t,c)}{N(t) + N(c) - N(t,c)} \tag{11}$$

$N(t,c)$  is the number of textual documents where glossary term  $t$  and ontology concept  $c$  occur together.  $N(t)$  is the number of documents where glossary term  $t$  occurs and  $N(c)$  corresponds to the number of documents which contain ontology concept  $c$ . Co-occurrence similarity is calculated based on the names of glossary terms and ontology concepts denotations. Each textual document is composed either of the content of an ontology concept or of the textual information about a particular glossary term (name and description).

Ontology concepts with similarity  $similarity(t,c)$  larger than  $similarity_{max}(t, c^{max}) * (1 - \beta)$  are suggested to the user, where  $similarity_{max}(t, c^{max})$

represents the highest similarity value between ontology concepts and a glossary term  $t$  and  $\beta$  is a user defined parameter.

$$\begin{aligned} \text{similarity}(t,c) &\geq \text{similarity}_{\max}(t, c^{\max}) * (1 - \beta) \\ 0 &\leq \beta \leq 1 \end{aligned} \tag{12}$$

To propose the relationship of equivalence we use string-edit distance between glossary term names and the related concept names. In the case of equivalence, the user can extend textual representation of the related ontology concept. Hierarchical relationship: “glossary term  $t$  is a subclass for concept  $c$ ” is proposed, when the similarity  $\text{similarity}_{\text{sub}}(t,c)$  between the glossary term  $t$  and subclasses of the related concept  $c$  is higher than the similarity  $\text{similarity}_{\text{sup}}(t,c)$  between the glossary term  $t$  and superclasses of the related concept  $c$ ;  $\gamma$  is a user defined parameter:

$$\begin{aligned} \frac{\text{similarity}_{\text{sub}}(t,c)}{\text{similarity}_{\text{sup}}(t,c)} &\geq \gamma + 1 \\ 0 &\leq \gamma \leq 1 \end{aligned} \tag{13}$$

We rank the hierarchical (subclass) relations using the quotient in the left part of formula (13).

Currently we do not propose hierarchical relationship “glossary term  $t$  is a superclass of concept  $c$ ” since we assume that the existing ontology concepts, which are already embedded into the hierarchy, contain a valid superclass information. If we do not find equivalence or hierarchical relationships between glossary term  $t$  and the related concept  $c$  or if the nature of the relationship is not clear (for instance, when the related ontology concept has no subclasses), we propose non-hierarchical associative relationship: “glossary term  $t$  is conceptually related to ontology concept  $c$ ”.

5. *User validation.* Furthermore, in OEM the user validates the candidate entries results consisting of the glossary terms, existing ontology concepts and glossary term-ontology concept relationships. In case of the equivalence relationship the user can extend the textual representation of the existing ontology concept by adding a comment, or adding or changing the natural language denotation. In case of the hierarchical relationships the user can add subclasses to the existing ontology concepts. If the nature of the relationship is not clear, the user can create an associative relationship or choose any other relationship between a glossary term and existing ontology concept. Moreover, the list with validated entries in the relevant format is created.

6. *Ontology extension.* The ontology extension takes place in the Ontology Extension Module. It represents adding new concepts and relationships between concepts into the ontology. Since, in phase 5, the user validates the candidate entry results, in phase 6 we have a list of final entry results. For instance, from Appendix 1 we have the following user validated relationships connected to the new ontology concept *DealerLoan*:

*Suggested relationship: DealerLoan is SUBCLASS to LoanAgreement*

*Suggested relationship: DealerLoan is CONCEPTUALLY RELATED to FinancingByBorrowing*

7. *Ontology reuse.* The ontology reuse phase serves as the connection link between separate ontology extension processes. As a part of the new extension process, we reuse

the previously extended ontology in the Domain Subset Extraction Module and in the Ontology Extension Module.

### 3.2.2 Text Mining Usage

Text mining plays a central part in our ontology extension methodology. Since users extend the ontology with information from their domain of interest, as domain information for ontology extension we use domain keywords and domain relevant glossary. With text mining techniques we are able to transform unstructured textual information into formalized knowledge.

As it is described above, in section 2.2, the textual information in our methodology is represented using a bag-of-words representation with normalized TFIDF weighting, and similarity between two text segments is calculated using cosine similarity between their bag-of-words representations. In addition, we use a chain of linguistic components, such as tokenization, stop-word removal and stemming, which allows for normalizing the textual representation of ontology concepts and a domain relevant glossary of terms with their descriptions.

### 3.2.3 Concept-similarity Identification

For concept-similarity identification in the **OntoPlus** methodology we use a combination of the ontology concept content, the ontology concept structure and the co-occurrence data between existing and candidate ontology concepts.

The ontology concept denotation and the ontology concept comment carry the same semantic weight. As well, we attribute the same semantic weight to the glossary term name and the glossary term comment.

Experiments on different domains and knowledge representations allow defining the  $\delta$  coefficients from formula (10) – in particular, how much weight should be attributed to content, structure or co-occurrence. In addition,  $\beta$  coefficient allows varying the number of results presented to the user in a concrete usage scenario.

### 3.2.4 Relation Extraction

The initial observations of the domain relevant glossaries show that the name and comments of the glossary terms often contain references to the superclasses and subclasses of the related ontological concepts. Following this idea, we have included into the **OntoPlus** methodology a relation identification part (formula 13). Eventually, this step allows automatic classification of the unstructured terms from the domain glossary into the ontology.

### 3.2.5 User Interaction

The **OntoPlus** methodology is mainly targeted at the ontology engineers – people, who develop and maintain large ontologies, such as the Cyc KB.

User interaction comes in two forms. The users define the domain relevant



information – domain keywords and domain relevant glossary.

Afterwards, user interaction plays a filtering role in the **OntoPlus** methodology. At the end, the users decide, which concepts and relationships they want to have in the ontology.

With user validation we are able to avoid the insertion of the non-relevant ontology concepts and relationships.

### 3.3 Methodology Adaption for Cyc Ontology

We have adapted the proposed methodology in order to obtain an exhaustive specific methodology for the Cyc Knowledge Base extension and demonstrated its suitability for ontology extension using real-world data. The motivation for the **OntoPlus** adaptation to the Cyc ontology extension lies in the actual usage of extended the Cyc ontology for business news analysis. The **OntoPlus** methodology adapted to Cyc extension is required to provide a user with relevant new Cyc concepts and their relationships to the existing Cyc concepts. The methodology adaptation for the Cyc ontology is intended for usage by Cyclists – agents, who inspect and modify the Cyc Knowledge Base.

#### 3.3.1 Structure of Cyc Knowledge Base

Currently, Cyc operates on one of the largest knowledge bases in the contemporary IT world.

It is stated as “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” (Cycorp, what's in Cyc, 2011) and divided into the large number of “microtheories”, each of which represents the set of assumption for a particular knowledge domain.

At the present time, the Cyc KB contains nearly two hundred thousand terms and an average of several dozen hand-entered assertions about/involving each term. New assertions are continually added to the KB by human knowledge enterers. Additionally, term-denoting functions allow for the automatic creation of millions of non-atomic terms, such as (*LiquidFn Nitrogen*); and Cyc adds a vast number of assertions to the KB by itself as a product of the inference process (Cycorp, what's in Cyc, 2011).

There exist several reasons for using Cyc in our news analysis task:

- The large number of common-sense assertions currently existing in the Cyc KB;
- Context sensitivity of the Cyc ontology;
- Existence of several version of the system available under different licenses (OpenCyc, ResearchCyc);
- Cyc API, which allows one to programmatically connect to the Cyc ontology.

#### 3.3.2 Specific Aspects of OntoPlus Application to Cyc Extension

Figure 3 displays the Cyc adaptation of the proposed methodology for semi-automatic ontology extension. The graphical representation of the ontology extension process, shown in Figure 3, demonstrates adding new concepts to the Cyc Knowledge Base. The

methodology phases are illustrated with numbers. The specific aspects of the **OntoPlus** adaptation to Cyc extension are based on the microtheories (Mt) (Contexts in Cyc, 2011) that Cyc uses to represent thematic subsets of the ontology. Microtheories are collections of facts, which appear to be true for a particular topic. Namely, the knowledge base in Cyc is divided into various microtheories, which contain a set of facts valid in a particular context. In such way, while extending the Cyc ontology, we can select the microtheories relevant for our domains of interest, and obtain the concepts and relationships from the selected microtheories. Because the Cyc KB is an extremely large common sense knowledge base, restriction to the relevant domain subset allows for more efficient methodology application.

In the *Domain information identification phase* in the Cyc adaptation of the methodology for ontology extension we identify the Domain Keywords and the Domain Glossary for the domains of interest. For research purposes we use a financial glossary, composed by Harvey (2003) and the ASFA thesaurus (ASFA thesaurus, 2010).

The Relevant Ontology (Cyc KB) Subset is extracted in the *Domain Subset Extraction Module*. The Upper-Level Domain Extractor uses Domain Keywords to obtain a number of domain relevant Cyc microtheories by mapping microtheory names to domain keywords from the list. Furthermore, the Knowledge Extractor provides a set of concepts defined in the domain relevant microtheories. Additionally, the concepts that are defined in other microtheories, but contain the natural language denotations correspondent to glossary term names, are extracted into the Relevant Ontology (the Cyc KB) Subset.

The *Domain relevant information preprocessing phase* and *Composing the list of potential concepts and relationships for ontology extension* follow subsequently. The Ontology Extender takes the Domain Glossary and extracted Cyc concepts as an input. The bag-of-words containing term name and term description is composed for each glossary term. A set of bag-of-words is composed for every extracted Cyc concept: concept denotation and concept comment; concept denotation, denotations of concepts in the hierarchical and non-hierarchical relationships with extracted Cyc concept; concept denotation and denotations of superclasses of the extracted concept; concept denotation and denotations of subclasses of the extracted concept. In order to find the related Cyc concepts for each glossary term we use TF-IDF weight and cosine similarity for content and structure similarities and Jaccard similarity for co-occurrence similarity.

Our experiments show that the best results are obtained giving more weight to content and structure for the financial domain and more weight to content and co-occurrence for fisheries & aquaculture domain.

Cyc concepts with combined similarity larger than  $similarity_{max}(t, c^{max}) * (1 - \beta)$ , where  $similarity_{max}(t, c^{max})$  represents the maximum combined similarity value between Cyc concepts and a glossary term  $t$  for a particular glossary term  $t$ , are suggested to the user.

We use string-edit distance between glossary term names and related concept denotations to propose the relationship of equivalence. In this case the user can extend Cyc textual representation of the related concept – add information to Cyc comment, add or change the Cyc concept denotation.

In case the similarity between the glossary term  $t$  and subclasses of the related Cyc concept  $c$  is higher than the similarity between the glossary term  $t$  and superclasses of the related Cyc concept  $c$  by  $\gamma$ , we propose hierarchical relationship: “*glossary term  $t$  is a subclass for Cyc concept  $c$* ”. If we do not find equivalence or hierarchical relationships between glossary term  $t$  and the related Cyc concept  $c$ , we propose a non-hierarchical associative relationship: “*glossary term  $t$  is conceptually related to Cyc concept  $c$* ”.

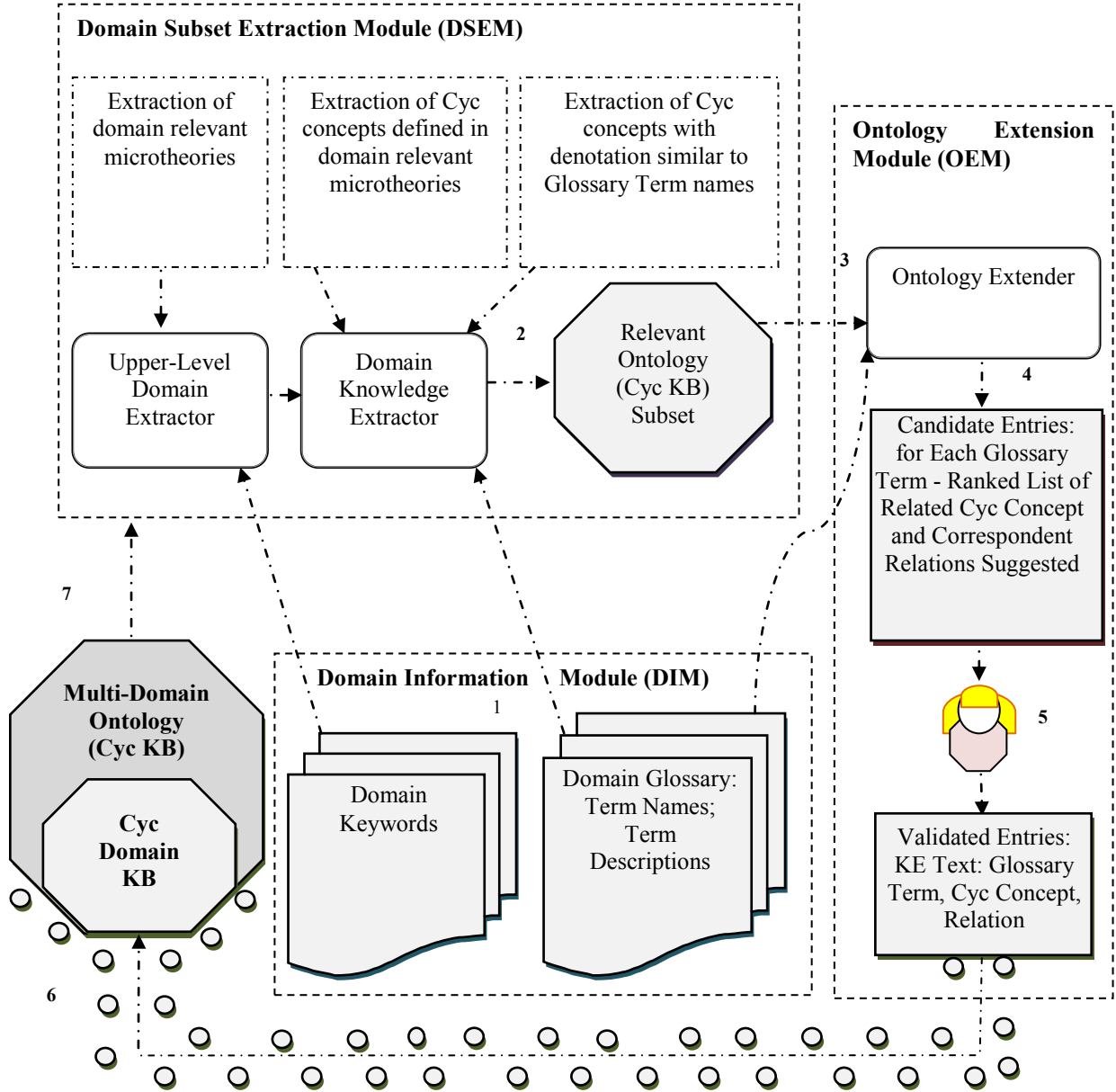


Figure 3: *Text-Driven Ontology Extension (Cyc KB Adaptation)*.

In the *User validation phase* the user has the opportunity to dynamically construct new assertions in the Knowledge Entry (KE) format which can be then automatically integrated into the Cyc KB.

Afterwards, the Cyc KB extension takes place in the *Ontology extension phase*. The *Ontology reuse phase* occurs when the new set of knowledge is added to the Cyc KB.

### 3.4 Methodology for News Analysis

In this thesis we explore how a large common-sense ontology can be used for analysis of textual information, in particular, for business news analysis.

In order to analyze news, we should characterize the nature of the information present. Concretely, business news usually contains specific domain information (such as business, economic, financial terms), common-sense information, different named entities (such as names of people, locations, organizations etc) and information from other domains.

With the **OntoPlus** methodology presented in the previous sections we can effectively handle the extension of a large ontology with specific domain information. At the same time, the ontology used for news analysis, already contains common-sense knowledge, which allows operating with common-sense information extracted from news.

The Cyc ontology population with named entities extracted from news can be done with assistance of different entity, event and fact extraction services.

For our research task we have utilized OpenCalais fact extraction service (OpenCalais, 2011). OpenCalais (Iacobelli et al., 2010) is a Thomson-Reuters free Web service that performs named entity recognition and extracts relationships and events from text. OpenCalais uses natural language processing techniques and machine learning to recognize instances of named entities. OpenCalais uses not only manually created databases of entities, but also textual features, such as capitalization, for new entities identification.

OpenCalais supports a rich set of semantic metadata, including entities (39 types), events and facts (76) (OpenCalais - English Semantic Metadata, 2011). The examples of OpenCalais entities include Company, Person, Country, Product, ProgrammingLanguage etc. Typical OpenCalais events and facts are CompanyFounded, CompanyLocation, Merger, Arrest, MovieRelease etc. In addition, OpenCalais also provides a *GenericRelation* type, which basically contains a triple subject-predicate-object.

The Cyc Knowledge Base and OpenCalais tool provide a user with a unique base for news analysis. The knowledge base in Cyc contains knowledge represented in different business and finance related contexts - microtheories, such as *FinancialTransactionMt*, *BusinessGMt*, *BusinessMoneyMt*, *AccountingMt*, *ProductGMt* etc. These microtheories define and describe money, general (capitalist) business practices, accounting concepts and principles, the major business-related organizations and activities, formal products etc. In addition, a number of business and financial concepts are defined in *UniversalVocabularyMt* - the microtheory which contains the definitional assertions about everything in Cyc's universe of discourse.

In this section we propose a **pipeline for business news analysis**, which allows analyzing large amounts of business and financial textual information. The **pipeline for business news analysis** is targeted at people who are interested in analyzing large amount of textual information from media, connected to business. The purpose of the proposed pipeline is in providing a line of steps for transforming large numbers of texts with business and financial knowledge into formalized ontological form and performing analysis based on the obtained ontologies. The pipeline requires a set of business news as an input. It is based on the fact extraction service and ontology with reasoning tools.

### 3.4.1 Phases of Pipeline for Business News Analysis

In Figure 4 we have defined a **pipeline for business news analysis** using OpenCalais for entity, event and fact extraction, the **OntoPlus** for the Cyc ontology population and reasoning based on the extended and populated ontology.

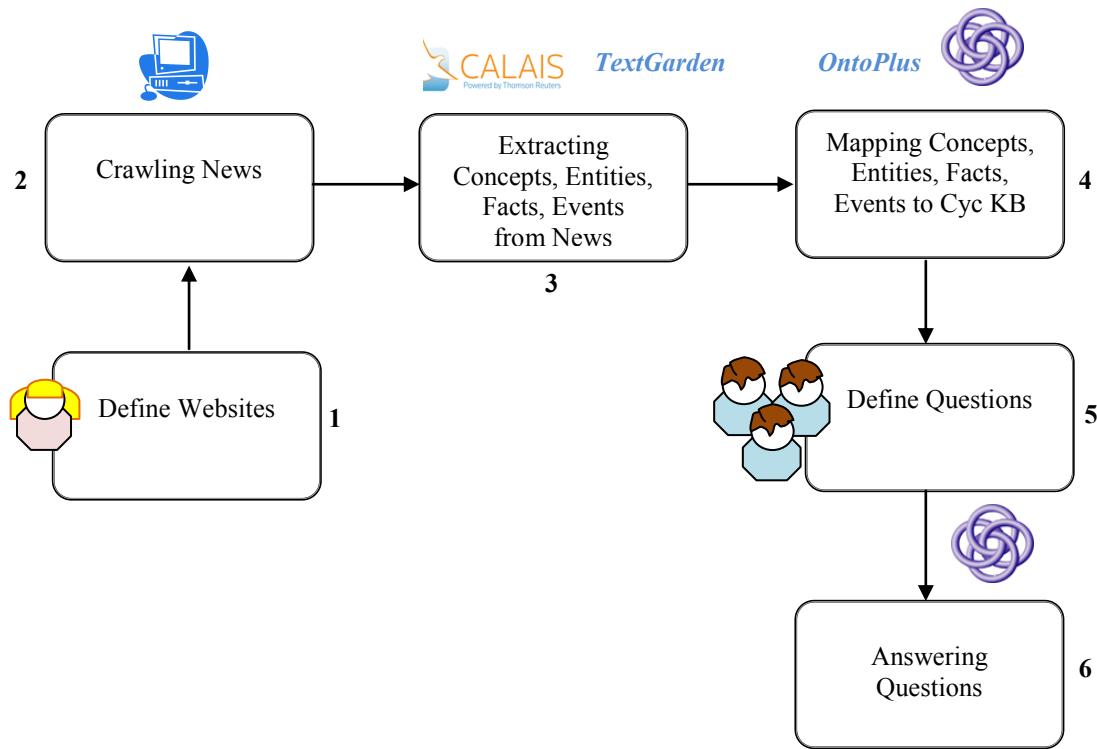


Figure 4: *System Pipeline (Analyzing Business News Using OpenCalais, OntoPlus and Cyc)*

In detail, the proposed **pipeline for business news analysis** accounts for the following phases: News website definition, News crawling, Concept, entity, event, fact extraction from news, Concept, entity, event, fact mapping to the Cyc KB, Questions definition, Questions answering.

1. *News website definition.* In the first phase of the **pipeline for business news analysis** a list of websites, which contain business news, is defined by the user (e.g., business or financial analyst).

2. *News crawling.* The news articles are crawled from the RSS feeds of the provided websites and afterwards, news cleaning is performed. Every news item represents a separate textual file.

3. *Concept, entity, event, fact extraction from news.* In this phase a set of financial concepts is extracted from business news. Using N-grams extractor from TextGarden tools (Text-Garden, 2011), it is possible to get all N-grams from the textual news collection and map them to the terms in the Harvey financial glossary (Harvey, 2003).

With a fact extraction service, such as OpenCalais tool, we are able to extract the information about entities, events and facts present in our news collection.

4. *Concept, entity, event, fact mapping to the Cyc KB.* In this phase ontology extension and ontology population are performed. With the **OntoPlus** methodology we are able to extend the Cyc KB with terms from the financial glossary, which occurred in our news collection. For ontology population we have created a set of mappings between OpenCalais entities, events and facts types and Cyc concepts – collections and predicates.

We also apply the **OntoPlus** methodology for concept disambiguation in the ontology population process.

5. *Question definition.* For question definition a set of questions, involving reasoning aspects is composed. For the business news analysis we have composed business related questions.

6. *Question answering.* The questions are asked using the Cyc reasoning interface and Cyc proofs are analyzed.

### 3.4.2 Methodology Support

Extraction of entities, events and facts with fact extraction service and mapping them to Cyc using the **OntoPlus** methodology provides a simple and effective way of the Cyc KB population with information from the news. Figure 4 demonstrates that the **pipeline for business news analysis** in phase 4 uses the **OntoPlus** methodology to map entities, events and fact extracted from news to the Cyc KB. The **OntoPlus** methodology is used in mapping for named entity disambiguation between new entities extracted from the news and existing instances in the Cyc ontology.

In order to perform the ontology population, we have created a set of mappings between OpenCalais types and Cyc concepts and relations.

Table 1: *Mapping Support.*

Mapping Support in Knowledge Entry Format	Mapping Support in Knowledge Entry Format
<i>Constant: OpenCalaisFactExtractionMt.</i>	<i>Constant: OpenCalaisAttributeFn.</i>
<i>In Mt: UniversalVocabularyMt.</i>	<i>In Mt: OpenCalaisFactExtractionMt.</i>
<i>isa: Microtheory.</i>	<i>Truth Value: :default.</i>
	<i>isa: BinaryFunction.</i>
<i>Constant: OpenCalaisFactEvent.</i>	<i>resultIsa: OpenCalaisAttribute.</i>
<i>In Mt: OpenCalaisFactExtractionMt.</i>	<i>argIsa: (1 CharacterString).</i>
<i>Truth Value: :default.</i>	<i>argIsa: (2 Thing).</i>
<i>isa: Collection.</i>	
<i>Constant: OpenCalaisAttribute.</i>	<i>Constant: extractedOpenCalaisFactEvent.</i>
<i>In Mt: OpenCalaisFactExtractionMt.</i>	<i>In Mt: OpenCalaisFactExtractionMt.</i>
<i>Truth Value: :default.</i>	<i>isa: BinaryPredicate.</i>
<i>isa: Collection.</i>	<i>arg1Isa: Event.</i>
	<i>arg2Isa: OpenCalaisFactEvent.</i>
<i>Constant: OpenCalaisFactEventFn.</i>	
<i>In Mt: OpenCalaisFactExtractionMt.</i>	<i>Constant: extractedOpenCalaisAttribute.</i>
<i>Truth Value: :default.</i>	<i>In Mt: OpenCalaisFactExtractionMt.</i>
<i>isa: UnaryFunction.</i>	<i>isa: BinaryPredicate.</i>
<i>resultIsa: OpenCalaisFactEvent.</i>	<i>arg1Isa: OpenCalaisAttribute.</i>
<i>argIsa: (1 CharacterString).</i>	<i>arg2Isa: Event.</i>

The examples of mapping support operations are given in Table 1. The support operations include creating Cyc microtheory *OpenCalaisFactExtractionMt* - the contextual space, where all information about the entities, events and facts extracted with OpenCalais are added. In addition, we create Cyc functions (*OpenCalaisFactEventFn*,

*OpenCalaisAttributeFn*), Cyc collections (*OpenCalaisFactEvent*, *OpenCalaisAttribute*) and Cyc predicates (*extractedOpenCalaisAttribute*, *extractedOpenCalaisFactEvent*), which are used in the mapping process.

Table 2 presents an example of mapping OpenCalais Fact/Event type of Merger into the Cyc KB.

Table 2: *Mapping OpenCalais Fact/Event Merger → Cyc Fact/Event Merger.*

Mappings
<p><i>In Mt: OpenCalaisFactExtractionMt.</i></p> <p><i>Direction: :forward.</i></p> <p><i>f: (implies</i></p> <p><i>(and</i></p> <p><i>(extractedOpenCalaisFactEvent ?FACTEVENT (OpenCalaisFactEventFn "Merger"))</i></p> <p><i>(extractedOpenCalaisAttribute (OpenCalaisAttributeFn "company" ?C) ?FACTEVENT)</i></p> <p><i>(extractedOpenCalaisAttribute (OpenCalaisAttributeFn "company" ?CI) ?FACTEVENT)</i></p> <p><i>(not (equals ?C ?CI))</i></p> <p><i>)</i></p> <p><i>(and</i></p> <p><i>(isa ?FACTEVENT OrganizationMerger)</i></p> <p><i>(isa ?C Business)</i></p> <p><i>(isa ?C LegalCorporation)</i></p> <p><i>(isa ?CI Business)</i></p> <p><i>(isa ?CI LegalCorporation)</i></p> <p><i>(mergees ?FACTEVENT ?C)</i></p> <p><i>(mergees ?FACTEVENT ?CI)</i></p> <p><i>)).</i></p>

From the Table 2 it is possible to see, that if the extracted event or fact (*?FACTEVENT*) is a Merger, then the correspondent different company attributes (*?C*, *?CI*) are instantiated as Cyc *Business* and *LegalCorporation*, and the relation instances about companies as *mergees* in event or fact (*?FACTEVENT*) are added.

Appendix 3 provides more examples of OpenCalais to Cyc mappings.

Algorithm 1 presents the procedure for mapping OpenCalais Generic Relation type to the Cyc Knowledge Base.

Generic relations in this research represent a relation, for which no predefined type was found. The generic relations aggregate subject, predicate and object triplets obtained from text.

Procedure *mapGenericRelationToOntology* as an input takes the ontology *O*, and the generic relation *GR*, consisting of subject, predicate and object. As an output, we retrieve the modified ontology, populated with named entities from the generic relation and the information connected to the event present in text.

The function *findNamedEntities* provides a list of all named entities present in text, e.g., in the subject or object of the generic relation.

---

Algorithm 1: *Mapping OpenCalais Generic Relations to Cyc KB.*

---

O: ontology used in mapping (see formula 1)  
 GR< SUBJECT; PREDICATE; OBJECT>: generic relation  
 SUBJECT: subject from the text, part of the generic relation  
 PREDICATE: predicate from the text, part of the generic relation  
 OBJECT: object from the text, part of the generic relation  
 E: set of events in the ontology O  
 NES: set of textual representations of named entities with correspondent types  
 NE: set of named entities. Each named entity is an actual or potential instance in O

```

procedure  mapGenericRelationToOntology  (O,  GR<SUBJECT,  PREDICATE,
OBJECT>)

  NESSUBJECT = findNamedEntities (SUBJECT)
  for NESi ∈ NESSUBJECT
    NEi = mapNamedEntityToOntology (O, NESi)
    if (not (isInOntology (NEi)))
      O.POPULATE (NEi)
      NEACTORS.ADD(NEi)

  NESOBJECT = findNamedEntities (OBJECT)
  for NESj ∈ NESOBJECT
    NEj = mapNamedEntityToOntology (O, NESj)
    if (not (isInOntology (NEj)))
      O.POPULATE (NEj)
      NEACTORS.ADD(NEj)

  E = findOntologyEvents(O)
  NEp = mapPredicateToOntologyEvent (O, E, PREDICATE)
  if ( not (isInOntology (NEp)))
    O.POPULATE (NEp)

  for NEa ∈ NEACTORS
    O.ASSERT (ACTORS, NEa, NEp)

return (O)
  
```

---

For each discovered named entity we find a relevant instance from the ontology using function *mapNamedEntityToOntology*. This similarity based function obtains from the ontology all possibly relevant instances. By applying the **OntoPlus** methodology we are able to identify the most similar instance for the defined named entity. If no relevant instances are obtained, the ontology is populated with correspondent named entity.

With the function *findOntologyEvents* we are able to find a set of events in the ontology.

With the function *mapPredicateToOntologyEvent*, which also uses the **OntoPlus** methodology, we obtain the most relevant event type for the predicate from the generic



relation. Finally, we assert the information about named entities as “actors” in the corresponding event.

Following the mapping of extracted entities, facts and events into the Cyc KB, Cyc is used for reasoning based on the obtained new knowledge and existing ontological rules.



## 4 Experiments

The evaluation methodology is provided in details in this chapter. In our research we specify ranking, tagging and question answering experiments in order to determine how successfully our methods help to extend ontology and to analyze business news.

### 4.1 Evaluation Methods

In order to evaluate the proposed **OntoPlus** methodology for ontology extension and the **pipeline for business news analysis**, we have conducted a series of experiments on the data sources, addressing different aspects of the proposed methodologies. Three types of conducted experiments included:

- Tagging experiments;
- Ranking experiments;
- Question answering experiments.

#### 4.1.1 Tagging Experiments

Because the available news collection included only business and financial news (and no fisheries & aquaculture news), tagging experiments have been executed in the financial domain. Tagging experiments show how the business news tagging with ontology components improves after ontology extension with the domain relevant glossary. For tagging experiments we have calculated the precision ( $Precision_{TAG}$ ) and recall ( $Recall_{TAG}$ ) of news tagging before and after adding terms to the ontology:

$$Precision_{TAG} = \frac{TP_{tag}}{TP_{tag} + FP_{tag}} \quad (14)$$

where

- $TP_{tag}$  represents the financial terms tagged correctly in the business news;
- $FP_{tag}$  represents the non-financial terms tagged as financial terms in the business news.

$$Recall_{TAG} = \frac{TP_{tag}}{TP_{tag} + FN_{tag}} \quad (15)$$

where

- $FN_{tag}$  represents the non-tagged financial terms in the business news.

Appendix 5 provides an example of news annotation with the Cyc tagger – a tool, which allows tagging text with formalized knowledge from the Cyc ontology.

### 4.1.2 Ranking Experiments

We have attributed special attention to our ranking experiments and to the evaluation of the **OntoPlus** methodology for ontology extension. The ranking experiments are conducted in two domains: financial domain and fisheries & aquaculture domain. Ranking experiments demonstrate how using the **OntoPlus** methodology we can semi-automatically extend the large lexical ontology with new concepts and identify the corresponding relationships between existent ontology concepts and domain glossary terms.

For the proposed **OntoPlus** methodology evaluation we have used two evaluation techniques - manual evaluation by human experts and gold standard based approach (Dellschaft and Staab, 2008).

The evaluation of the **OntoPlus** methodology is performed at the lexical, taxonomic (concept hierarchy) and non-taxonomic levels. For the lexical evaluation the mapping of the glossary terms to the existent ontology concepts is performed. At the taxonomic level the evaluation of the suggested hierarchically related concepts and suggested superclass-subclass relationships is implemented. Finally, at the non taxonomic relations level the evaluation of the suggested associatively related concepts and associative relationships is done. While the gold standard based approach is used to perform lexical and taxonomic evaluation, the manual evaluation is used at the non-taxonomic level.

Maedche and Staab (2002) have used the normalized string edit distance to identify how similar two ontologies are. Normalized string edit distance between ontology concept denotations and glossary term names is used as a baseline measure in the evaluation of the proposed **OntoPlus** methodology.

In order to define how successful the proposed methodology is in practice, we have used a number of evaluation measures commonly used for ontology learning evaluation, as follows.

Precision of the top suggested concept ( $Precision_{RANK}$ ) defines the percentage of the glossary terms for which the equivalent and hierarchical, associative or any related ontology concepts have obtained the highest position in the suggested ranked related concept list:

$$Precision_{RANK} := \frac{TP_{rank}}{TP_{rank} + FP_{rank}} \quad (16)$$

where

- $TP_{rank}$  represents the correct related concepts identified;
- $FP_{rank}$  represents the false related concepts identified.

Learning Accuracy (Hahn and Schnattinger, 1998) shows the degree to which the proposed methodology correctly predicts the superclass for the candidate ontology concept (represented by a glossary term) to be learned:

$$LA := \sum_{i=1}^n \frac{LA_i}{n} \quad (17)$$

$$LA_i := \begin{cases} \frac{CP_i}{SP_i} & \text{if } FP_i = 0 \\ \frac{CP_i}{FP_i + DP_i} & \text{if } FP_i \neq 0 \end{cases} \quad (18)$$

where

- $n$  represents the number of concept hypotheses for the target;
- $SP_i$  is the length of the shortest path from the top node of the concept hierarchy to the maximally specific concept subsuming the instance to be learned in hypothesis  $i$ ;
- $CP_i$  is the length of the path from the top node to that concept node in hypothesis  $i$  which is common both to the shortest path (as defined above) and the actual path to the predicted concept (whether correct or not);
- $FP_i$  is the length of the path from the top node to the predicted false;
- $DP_i$  is the node distance between the predicted false node and the most specific common concept still correctly subsuming the target in hypothesis  $i$ .

In addition, we have used a hit rate measure used in the evaluation of recommendation systems. The hit rate displays the number of hits and their position within top  $N$  suggestions (Deshpande and Karypis, 2004). We specify the hit rate measure as following:

$$HR := \frac{\sum_{t \in T} HR_t}{|T|} \quad (19)$$

$$HR_t := \frac{\sum_{u \in U} HIT(t,u)}{|U|} \quad (20)$$

where

- $t$  is a candidate concept for ontology extension;
- $u$  represents a user;
- $HIT(t,u)$  is a binary function. For the candidate concept  $t$  and user  $u$  it returns 1 if the correspondent related ontology concepts have been found among the top  $N$  suggestions and 0 otherwise;
- $U$  is a set of users;
- $T$  represents the set of candidate ontology concepts (glossary terms).

### 4.1.3 Question Answering Experiments

The **pipeline for business news analysis** is evaluated within question answering experiments. Question answering experiments reveal the capacity of Cyc to answer business news related questions before and after the extension and population of the Cyc Knowledge Base. As well as in the tagging experiments, because the available news collection included only business and financial news (and no fisheries & aquaculture news), question answering has been conducted only in business and financial domain. For question answering experiments we have calculated precision of ontology population ( $Precision_{NEWS POPUL}$ ) and the precision of news related question answering ( $Precision_{QA}$ ):

$$Precision_{NEWS POPUL} = \frac{TP_{news popul}}{TP_{news popul} + FP_{news popul}} \quad (21)$$

where

- $TP_{news popul}$  represents the correct instances inserted to the ontology;
- $FP_{news popul}$  represents the false instances inserted to the ontology.

$$Precision_{QA} = \frac{TP_{qa}}{TP_{qa} + FP_{qa}} \quad (22)$$

where

- $TP_{qa}$  represents the correct query answers;
- $FP_{qa}$  represents the false query answers.

## 4.2 Domains of Interest

According to the first phase of our methodology, domain knowledge identification should be made in the initial phase. As described above, the financial (business and financial) domain has been selected as a primary domain of interest. In addition, we have performed ranking experiments in fisheries & aquaculture domain. Two domains are described below in the following subsections.

### 4.2.1 Financial Domain

The financial (business and financial) domain is characterized by the dynamic and concentrated information flow.

Business and financial domain represents the notion of business organizations - organizations, which provide goods and services, and available information about them. The area of finance includes knowledge about money, investments, markets, risks.

The following are example concepts from the business and financial domain:

*ACCOUNT-* In the context of bookkeeping, refers to the ledger pages upon which various assets, liabilities, income, and expenses are represented. In the context of investment banking, refers to the status

*of securities sold and owned or the relationship between parties to an underwriting syndicate. In the context of securities, the relationship between a client and a broker/dealer firm allowing the firm's employee to be the client's buying and selling agent. See: Account executive; account statement (Harvey, 2003).*

*BROKER - An individual who is paid a commission for executing customer orders. Either a floor broker who executes orders on the floor of the exchange, or an upstairs broker who handles retail customers and their orders. Also, person who acts as an intermediary between a buyer and seller, usually charging a commission. A "broker" who specializes in stocks, bonds, commodities, or options acts as an agent and must be registered with the exchange where the securities are traded. Antithesis of dealer (Harvey, 2003).*

*COMPANY - A proprietorship, partnership, corporation, or other form of enterprise that engages in business (Harvey, 2003).*

*MONEY - Currency and coin that are guaranteed as legal tender by the government (Harvey, 2003).*

*COMMERCIAL PAPER - Short-term unsecured promissory notes issued by a corporation. The maturity of commercial paper is typically less than 270 days; the most common maturity range is 30 to 50 days or less (Harvey, 2003).*

*MORAL HAZARD - The risk that the existence of a contract will change the behavior of one or both parties to the contract, e.g., an insured firm will take fewer fire precautions (Harvey, 2003).*

Although in our experiments we have used a stable collection of business news, the data obtained from business and financial information sources in general is non-structured and continuously changing. There exist a large number of documents displaying different types of business and financial data – they often contain numerical values, date and time stamps, and references to various named entities, such as people, organizations, and locations.

#### 4.2.2 Fisheries and Aquaculture Domain

The fisheries & aquaculture domain in our research is represented by the data from areas of science, technology, management of marine, brackish water, and freshwater resources and environments.

The typical concepts in this domain are the following:

*ALLOZYMES - Variant forms of an enzyme that are coded by different alleles at the same locus are called allozymes (Wikipedia, 2011).*

*ANHYDRITE - A mineral - anhydrous calcium sulfate,  $\text{CaSO}_4$ . It is in the orthorhombic crystal system, with three directions of perfect cleavage parallel to the three planes of symmetry (Wikipedia, 2011).*

*TRANSPIRATION - A process similar to evaporation. It is a part of the water cycle, and it is the loss of water vapor from parts of plants (similar to sweating), especially in leaves but also in stems, flowers and roots (Wikipedia, 2011).*

*CHOLESTEROL - A waxy steroid of fat that is manufactured in the liver or intestines. It is used to produce hormones and cell membranes and is transported in the blood plasma of all mammals. It is an essential structural component of mammalian cell membranes (Wikipedia, 2011).*

*FOVEA - A part of the eye, located in the center of the macula region of the retina (Wikipedia, 2011).*

Since for our experiments in fisheries & aquaculture domain we are using a thesaurus, which contains a set of terms with specified relations described below, our fisheries & aquaculture data is characterized by the defined structure. Moreover, the nature of data is more academically grounded, since selected data sources cover mainly research papers.

### 4.3 Data Description

For our news tagging experiments we have used the RSS feeds data Yahoo! Finance (Yahoo! Finance, 2010) website.

The news collection used in the current experiment accounts for around 3400 Yahoo! Finance news.

In ranking experiments, for the financial domain, we have selected the Harvey (2003) financial glossary which can be found at the Yahoo! Finance website (Yahoo! Finance, 2010). The Harvey financial glossary (2003) contains around 6000 hyperlinked financial terms. The typical financial glossary entries are demonstrated on the Figure 5.

*TERM: Endowment*

*COMMENT: Gift of money or property to a specified institution for a specified purpose.*

*TERM: Recession*

*COMMENT: A temporary downturn in economic activity, usually indicated by two consecutive quarters of a falling GDP.*

Figure 5: Example Financial Glossary Entries.

Fisheries & aquaculture domain, represented by the Aquatic Sciences and Fisheries Abstracts (ASFA) thesaurus (ASFA thesaurus, 2010), has been selected as a source for our second domain of interest. The ASFA thesaurus contains around 9900 terms involving several types of relationships: equivalence relationships (USE, Use For UF), hierarchical relationships (Broader Term BT, Narrower Term NT), associative relationships (Related Term RT) and notes (SN).

Figure 6 shows how thesaurus relationships are used to compose a COMMENT for a particular term in the fisheries & aquatic thesaurus, in order to get the equivalent content as provided by a glossary (term and its comment/description). In detail, we take the names of the equivalent, hierarchical, associative terms for the specified ASFA term, notes about the specified ASFA term and merge them together into textual comment about the specified term. For instance, for the ASFA term *Collagen* displayed in Figure 6 the comment is composed out of the hierarchically related (BT) ASFA term *Proteins* and the associatively related (RT) ASFA term *Connective tissues*.



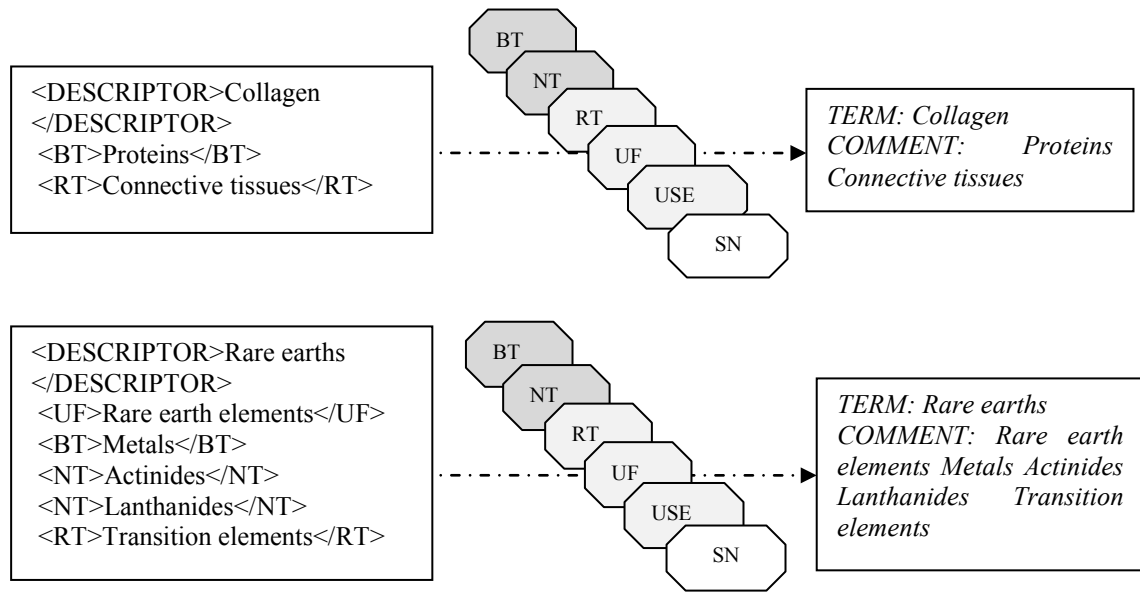


Figure 6: ASFA Thesaurus Transformation.

In our question answering experiments we have used the collection of around 3400 Yahoo business news items.

Cyc tools have been selected as a base for news tagging, ontology extension and question answering in our research experiments.

All datasets used for the experiments can be obtained by the request from the author of the thesis.

## 4.4 Experimental Settings

This section describes the settings for the tagging, ranking and question answering experiments conducted in this research. We have attributed different specifications to each type of experiments according to its purposes.

### 4.4.1 Tagging Experiments

In spite of the fact that Cyc contains a very extensive knowledge base, the representation of the financial and economical information in it is far from complete.

Tagging experiments show how the business news tagging improves after ontology extension with the domain relevant glossary.

The tagging/annotation experiments provide testing on a random subset of 100 Yahoo! Finance news items. We have identified the financial terms, occurring most frequently in the selected news, tagged the terms with the Cyc Tagger and checked the precision and recall of news tagging. Furthermore, we have added the simplest assertions about the missing financial terms into Cyc and again found the precision and recall of news tagging.

### 4.4.2 Ranking Experiments

Ranking experiments demonstrate how using the **OntoPlus** methodology we can semi-automatically extend the large lexical ontology with new concepts and identify the correspondent relationships between existent ontology concepts and domain glossary terms.

In order to evaluate the suggested methodology, we have conducted a number of experiments on a subset of 100 randomly selected terms from each domain resource (Harvey financial glossary (Harvey, 2003), the ASFA thesaurus (ASFA thesaurus, 2010)).

The selected terms might be in the equivalent, hierarchical and associative relationships with the existing ontology concepts. The human experts annotated the selected terms from the financial glossary and fisheries & aquaculture thesaurus with the correspondent equivalent and hierarchically related terms from the Cyc Knowledge Base. The extensive size of the Cyc Knowledge Base does not allow the experts to annotate the selected glossary term with all associatively related concepts from the ontology.

We have performed the domain information preprocessing and extraction of the relevant domain ontology subset from the Cyc Knowledge Base according to methodology phases described in chapter 3. Using formulas (10), (12) and (13) from the **OntoPlus** methodology, we are able to define a list of related Cyc concepts and a list of possible relationships for each glossary term. The Cyc Knowledge Base is then extended with the concepts corresponding to the chosen terms based on the ranking proposed by the methodology.

We have used precision and hit rate measures to identify the importance of the ontology concept content, ontology concept structure and co-occurrence for establishing relatedness between glossary terms and ontology concepts. Subsequently, we have evaluated each measure by estimating the quality of concept ranking.

In addition, within the gold standard based approach, we have used learning accuracy to measure the quality of the hierarchical (subclass) relation identification.

### 4.4.3 Question Answering Experiments

Question answering experiments demonstrate the capacity of Cyc to answer business news related questions before and after the extension of the Cyc Knowledge Base. Question answering experiments contain the evaluation of the **pipeline for business news analysis**.

We have performed the question answering evaluation according to two scenarios. Firstly, we assumed that we have a simple question and we want to get an answer using an unextended and extended Cyc Knowledge Base.

Secondly, we have extracted all correspondent financial concepts using TextGarden tools (Text-Garden, 2011) and all entities, events and facts, using OpenCalais service from the pool of around 3400 business news from year 2008, and inserted them into the Cyc KB using the **OntoPlus** methodology.

Following that, we have composed and tested with Cyc reasoning tools a number of queries, given in Table 3. We have evaluated obtained results with precision of question answering.

In addition, in order to evaluate how efficiently the ontology population is performed, we have selected 50 business news articles and calculated the precision of the suggested new ontology instances population.

Table 3: *Experimental Queries.*

№	Query
1.	Get companies with more than 100 employees
2.	List companies, which participated in mergers (or acquisitions) and the bankruptcy of which was reported in the news
3.	List people in high positions in companies/organizations with residence in District Of Columbia (US)
4.	Are there any people, accused of something or convicted in something, who work in large companies or organizations (>100 employees)?
5.	Which corporations were reported in the news as issuing securities?
6.	Get business partners from IT sector for particular company (expl: Enterra Solutions)
7.	List companies, which produce cars and their affiliates
8.	Get companies-competitors and products, which they produce
9.	Where are company customers located? – Get locations of customers for companies
10.	Was any company founded before year 1990? What does it produce?
11.	Get companies, which were involved into layoff activity, and their residence. List layoffs in Michigan-State and in Georgia-State.
12.	Get companies, against which there were reported lawsuits
13.	List company meetings of a particular company (Atsc) and conference calls of a particular company (ParTechnologyCorporation)
14.	Were there reported any companies, which have participated in reorganization or restatement?
15.	Get stock ticker symbols of companies in difficulties - bankruptcy of which was reported in the news or which had any labor issues or lawsuits reported
16.	Get person profile - contact details of a person, person attributes (age, birthdate, birthplace, gender etc). Examples: AlexisMcgee, JenniferPaige
17.	Were there any family relation found between employer/employee?
18.	Who works where? - Get all analysts
19.	Get people reported working as economists in companies from financial sector
20.	Which US companies were involved in IPO activity?



## 5 Results

The results of the experiments provided in this chapter confirm the applicability of the suggested methodology for ontology extension to the Cyc Knowledge Base augmentation. Extending and populating the Cyc ontology, we can better analyze textual data, in particular business news and effectively perform reasoning on the news related information.

We have organized results of the experiments into three groups: tagging experiments results, ranking experiments results consisting of concept ranking evaluation, relation ranking evaluation, the Cyc KB extension and question answering results. In addition, in this chapter we demonstrate the software prototype for the Cyc ontology extension.

### 5.1 Tagging Experiments

As we already mention in section 4.1, the tagging experiments have been conducted only for the financial domain, since no news was available for fisheries & aquaculture domain.

In the tagging experiment we have found 231 financial terms in the random sample of 100 Yahoo! Finance news articles. We have manually performed the evaluation of business news tagging. The precision of business news tagging increases from 61 % to 87 % and the corresponding recall - from 46 % to 81 % after adding the simplest assertions (lexical labels and instance/type information) about the missing terms into Cyc. This is confirming selection of the Cyc ontology as a base for ontology extension experiment. The Cyc ontology has still space for extension as for financial domain with terms relevant for financial news analysis.

### 5.2 Ranking Experiments

According to the **OntoPlus** methodology, for each glossary term the user wants to add to the ontology, a ranked list of the related ontology concepts is created.

As discussed in chapter 3, we assume the following relationships between the glossary terms and existing ontology concepts:

- Equivalence relationship;
- Hierarchical relationship;
- Non-hierarchical (associative relationship);
- No relationship.

### 5.2.1 Concept Ranking

In the present experiment we have evaluated the quality of concept ranking depending on the different proportions of ontology concept textual content, ontology concept structure and co-occurrences of glossary terms and ontology concepts. Additionally, we have taken into the account the importance of the established relationship between the top suggested related ontology concept and candidate concept for the ontology extension. We have grouped equivalent and hierarchical relations in one group assuming that these relations are the most important in the ontology extension process. Besides, we have also considered associative relations and a union of all the three considered relations (equivalent, hierarchical and associative relations) referred to as any relations.

Figure 7 and Figure 8 show the precision of the top suggested concept depending on content, structure and co-occurrence information for the financial glossary and the AFSA thesaurus respectively. The vertical grey lines are positioned in order to mark the different content weights. For instance, Figure 7 provides the performance in the financial domain when only structure is used (the content weight and the co-occurrence weight are set to 0).

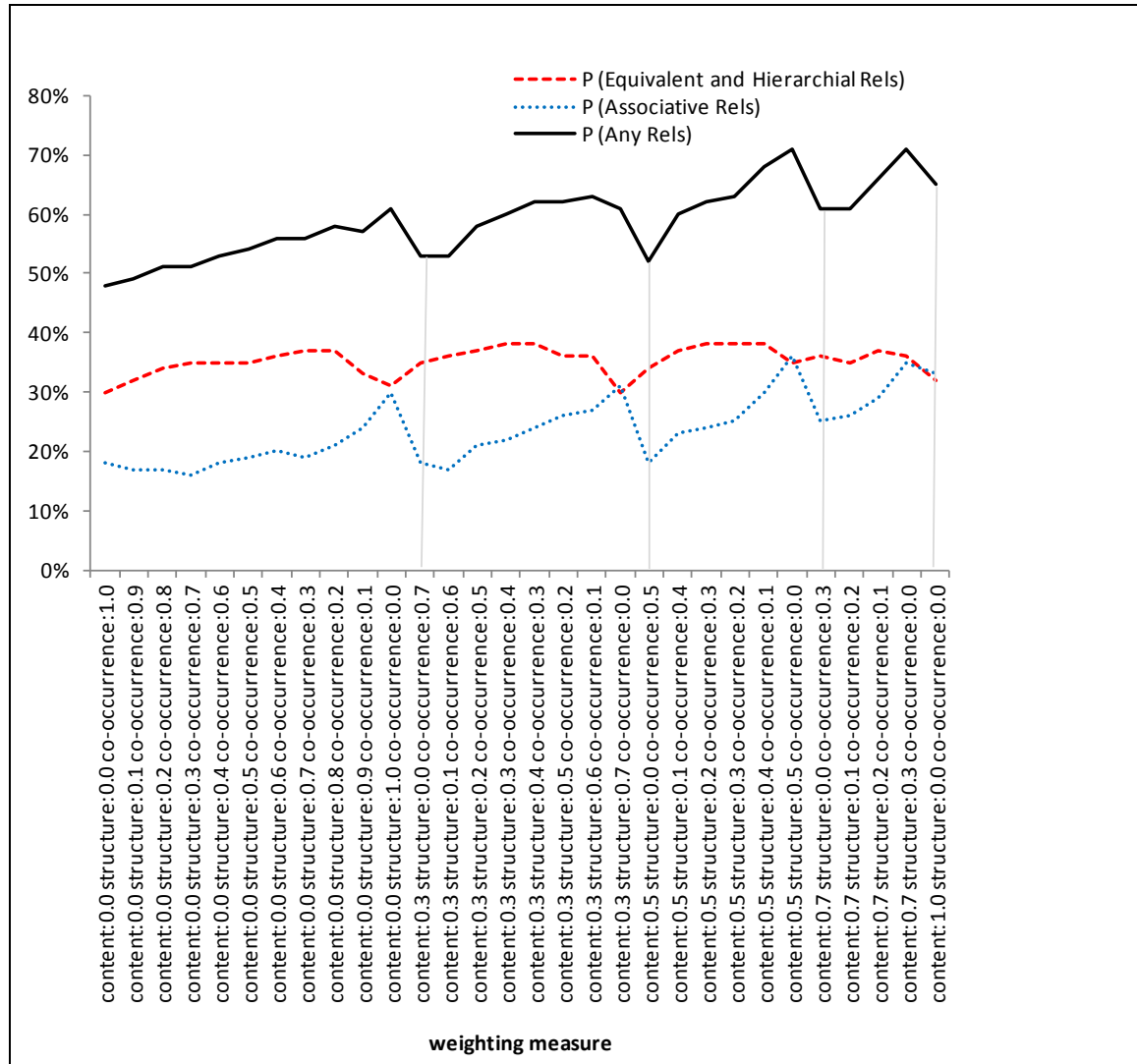


Figure 7: Performance of the content, structure and co-occurrence weighting measures (Financial glossary). Precision (P) of concept ranking (Top 1).

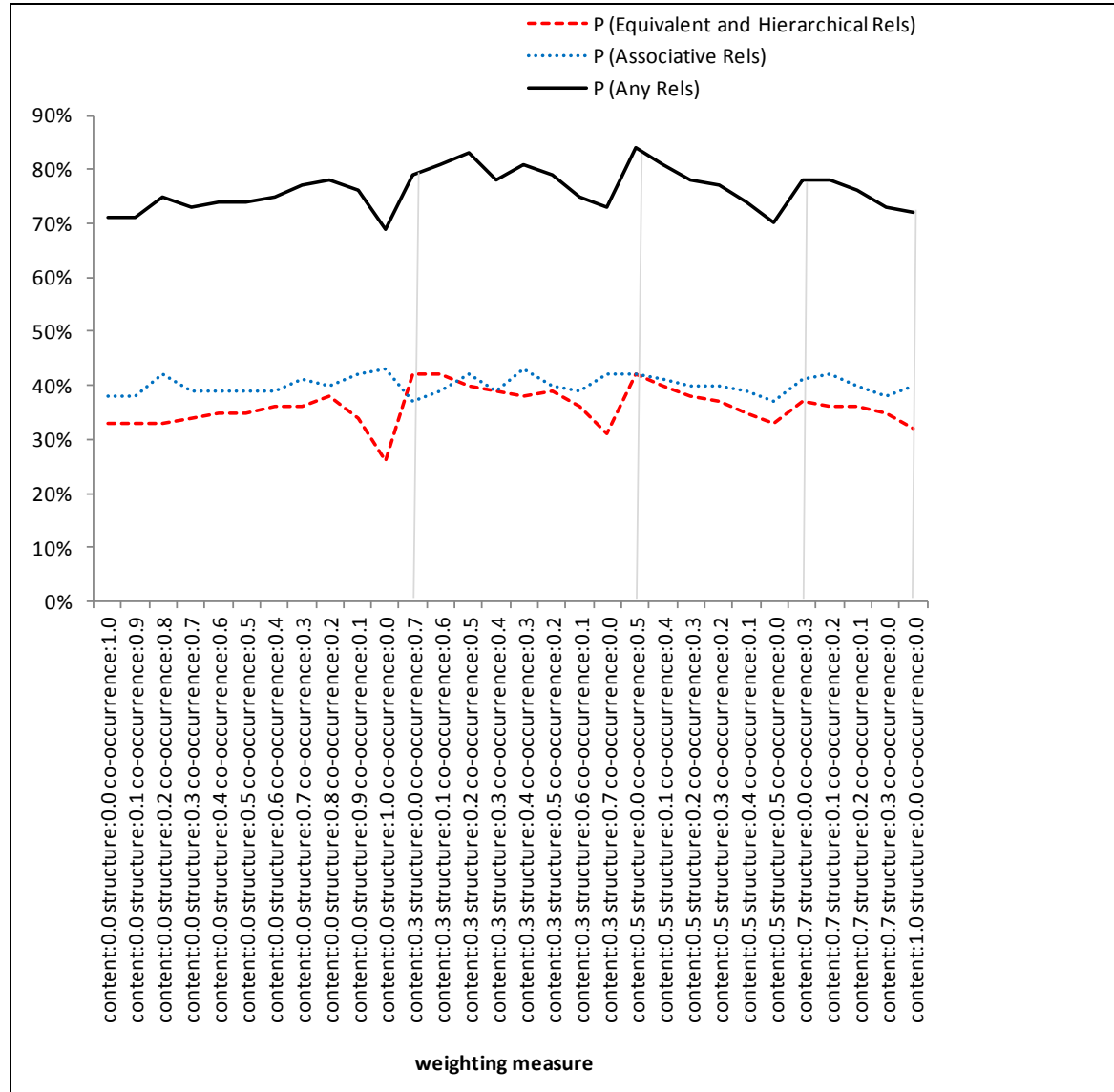


Figure 8: *Performance of the content, structure and co-occurrence weighting measures (ASFA thesaurus). Precision (P) of concept ranking (Top 1).*

Precision of the top ranked concept is 61 % when analyzing any of the three considered relations (Any Rels), 30 % when considering associative relations (Assoc Rels) and 31 % when the top ranked concept is in the equivalent or hierarchical relations (Eqv and Hier Rels) with a correspondent glossary term.

For the financial glossary the best results are obtained by combining the ontology content, the ontology structure and the co-occurrence information with giving more weight to content and structure and less weight co-occurrence. For the ASFA thesaurus, co-occurrence plays a substantial role. In this case the best results are obtained by giving more weight to content and co-occurrence and less weight to structure. The potential explanation of such performance can refer to the different domain and structural nature of the financial glossary and fisheries & aquaculture thesaurus.

In addition, Table 4 and Table 5 provide more detailed evaluation of the quality of ranking for the best performing weighting measures. The following weighting measures have been used for the financial domain: content weight  $\delta_1=0.5$ , structure weight  $\delta_2=0.4$

and co-occurrence weight  $\delta_3=0.1$ .

Table 4: *Evaluation of the top suggested candidate concepts for ontology extension (Financial glossary).*

Weighting Measure		100 Random Terms					
		HR (Top 1)		HR (Top 5)		HR (Top 10)	
		Eqv or Hier Rels	Any Rels	Eqv or Hier Rels	Any Rels	Eqv or Hier Rels	Any Rels
Baseline - Name:	[1.0]	18	28	24	36	25	40
Content (cosine similarity):	[1.0]	32	65	60	92	68	95
Co-occurrence (Jaccard similarity):	[1.0]	30	48	48	62	52	73
Content:	[0.5]						
Structure:	[0.4]	38	68	66	95	76	98
Co-occur:	[0.1]						

Table 5: *Evaluation of the top suggested candidate concepts for ontology extension (ASFA thesaurus).*

Weighting Measure		100 Random Terms					
		HR (Top 1)		HR (Top 5)		HR (Top 10)	
		Eqv or Hier Rels	Any Rels	Eqv or Hier Rels	Any Rels	Eqv or Hier Rels	Any Rels
Baseline - Name:	[1.0]	24	37	25	38	27	40
Content (cosine similarity):	[1.0]	32	72	52	88	56	91
Co-occurrence (Jaccard similarity):	[1.0]	33	71	49	89	51	90
Content:	[0.5]						
Structure:	[0.0]	42	84	63	96	66	96
Co-occur:	[0.5]						

For fisheries & aquaculture domain we have set content weight  $\delta_1=0.5$ , structure weight  $\delta_2=0.0$ , co-occurrence weight  $\delta_3=0.5$ .

The efficiency of using the **OntoPlus** methodology in a certain domain is validated through the specified baseline. As a baseline measure we use mapping glossary term



names to Cyc concept denotations, using normalized string-edit distance to rank the relations (equivalent or hierarchically related Cyc concepts and any related concepts) for each glossary term. Furthermore, we have compared the best performing weighting measures with other methods, which use only cosine similarity between textual content of the documents (Fortuna et al., 2007) or only co-occurrence analysis (Turney, 2001).

Table 4 and Table 5 contain the information on the hit rates (HR) for top 1, top 5 and top 10 suggested candidate concepts for ontology extension. The results of baseline are given under Baseline - Name [1.0] and show that by using baseline measure on both datasets for 40 % of terms, the related Cyc concepts have been found among the top 10 suggested concepts (when considering any of the three relations – the last column in the tables marked as Any Rels). Comparing it to the results of the best performing combination of content, structure and co-occurrence (98 % for financial domain and 96 % for fisheries & aquatic domain respectively), we demonstrate that by combining textual the ontology content, the ontology structure and the co-occurrence information we can provide the user with more than double number of concepts suitable for the ontology extension than using the baseline.

In addition, we have performed a sensitivity analysis for  $\beta$  parameter (12). Figure 9 displays the hit rate (HR) for the equivalent and hierarchically related concepts depending on  $\beta$  coefficient. It is possible to notice that higher hit rates both in financial and fisheries & aquaculture domains are obtained with higher  $\beta$  values. Figure 10 shows how  $\beta$  coefficient influences the number of concepts presented to the user. The lower  $\beta$  values imply the fewer number of the displayed concepts.

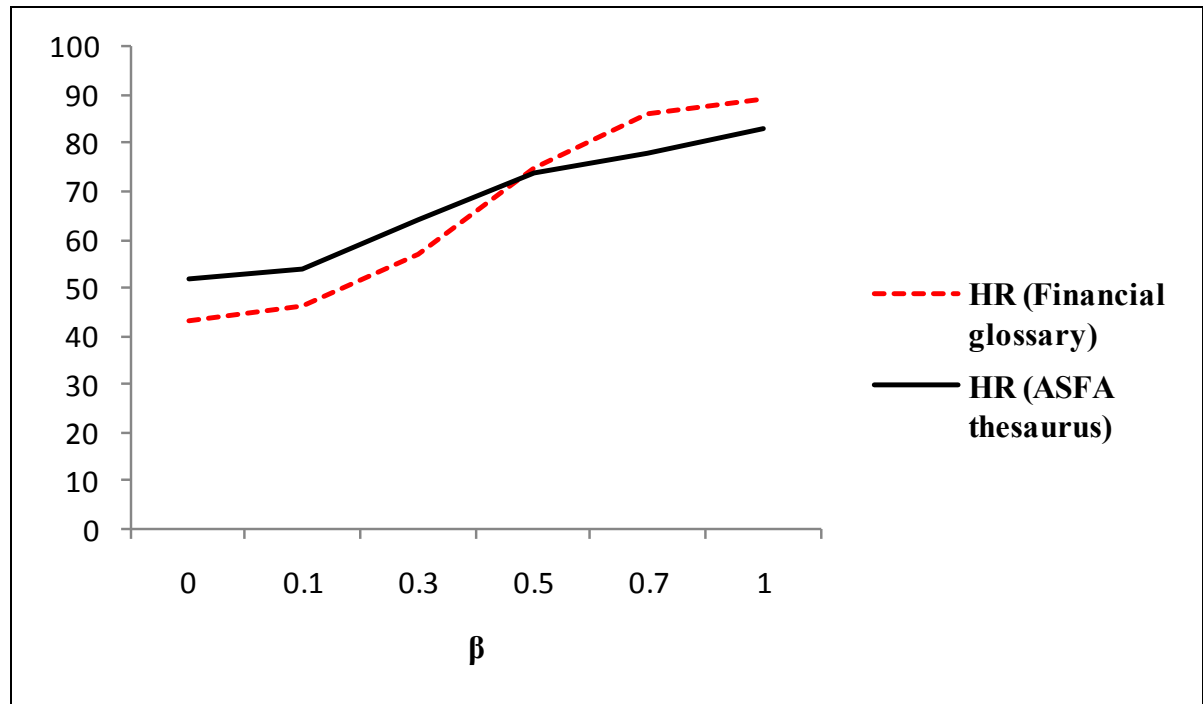


Figure 9: Hit Rate (HR) depending on  $\beta$  (Equivalent and hierarchically related concepts).

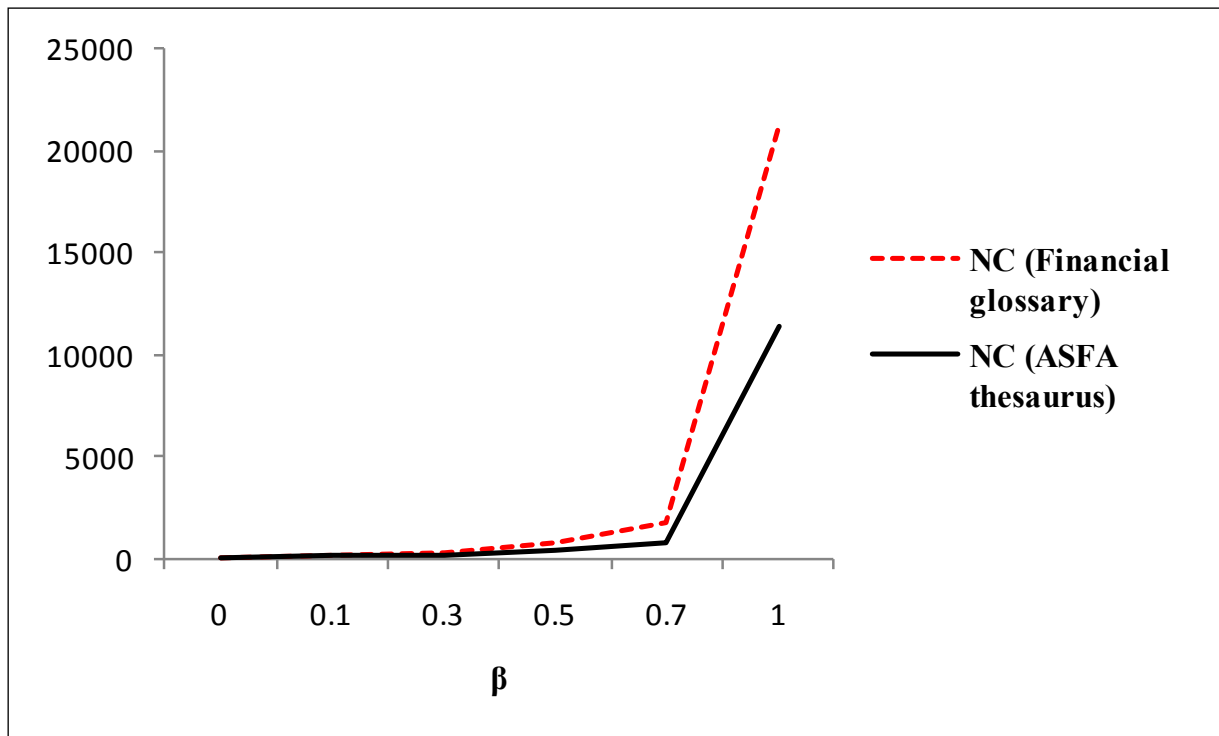


Figure 10: *Number of Concepts (NC) depending on  $\beta$ .*

### 5.2.2 Relation Ranking

In our methodology we can automatically suggest not only the related existing ontology concepts, but also a relation: the equivalent, hierarchical (subclass) and associative relations between existing ontology concepts and candidate concepts for ontology extension.

For the relation identification experiment we have evaluated the precision of the suggested equivalent relations and a hierarchical (subclass) relation which obtained the highest position according to our methodology.

Table 6 displays the evaluation of the equivalent relations and top 1 subclass relations suggested for each candidate ontology concept.

The results in Table 6 show the precision (P) of 29.5 % for subclass relation identification in financial domain and the precision of 20.3 % for subclass relation identification in fisheries & aquaculture domain. The precision for equivalent relation identification is 60.0 % for the financial domain and 94.7 % for fisheries & aquaculture domain.

Although it is not possible to statistically compare the results of the **OntoPlus** methodology with the results of other methods and tools, in this thesis we report on the achievements of Text2Onto framework (Cimiano and Völker, 2005) and SPRAT tool (Maynard et al., 2009). The authors of Text2Onto framework report a precision of 17.38 % for subclass-of relation identification on the subset of tourism-related texts. The

evaluation in SPRAT performed on 25 randomly selected Wikipedia articles about animals shows the precision of 48.5 % for subclass identification and 48.0 % for synonym recognition.

Table 6: *Evaluation of the equivalent, hierarchical (subclass) relation identification.*

Glossary/ Weighting Measure	100 Random Terms	
	P (Eqv Rels, %)	P (Top 1 Subclass Rels, %)
Financial glossary:		
Content: [0.5]		
Structure: [0.4]	60.0	29.5
Co-occur: [0.1]		
ASFA thesaurus:		
Content: [0.5]	94.7	20.3
Structure: [0.0]		
Co-occur: [0.5]		

The evaluation of the top suggested equivalent or hierarchical (subclass) relations is given in Table 7. The first column displays the number of concepts for which either equivalent or subclass relations, or both of them have been suggested automatically. The other three columns show the number of concepts for which the correct automatically suggested either equivalent or subclass relations, or both of them have been found among top 1, top 5 and top 10 suggested relations.

Table 7: *Evaluation of the top suggested equivalent & hierarchical (subclass) relations.*

Glossary/ Weighting Measure	Number of Concepts with Eqv or Subclass Rels Found	100 Random Terms		
		HR (Top 1)	HR (Top 5)	HR (Top 10)
Financial glossary:				
Content: [0.5]				
Structure: [0.4]	97	36	56	60
Co-occur: [0.1]				
ASFA thesaurus:				
Content: [0.5]	80	31	40	41
Structure: [0.0]				
Co-occur: [0.5]				

Evaluation of the suggested equivalent or hierarchical (subclass) relations shows that for 60 terms from the financial domain and for 41 concepts from the fisheries & aquaculture domain the correct automatically suggested relations have been found among top 10 suggested relations.

The sensitivity analysis for  $\beta$  and  $\gamma$  parameters is given in Figure 11 and Figure 12.

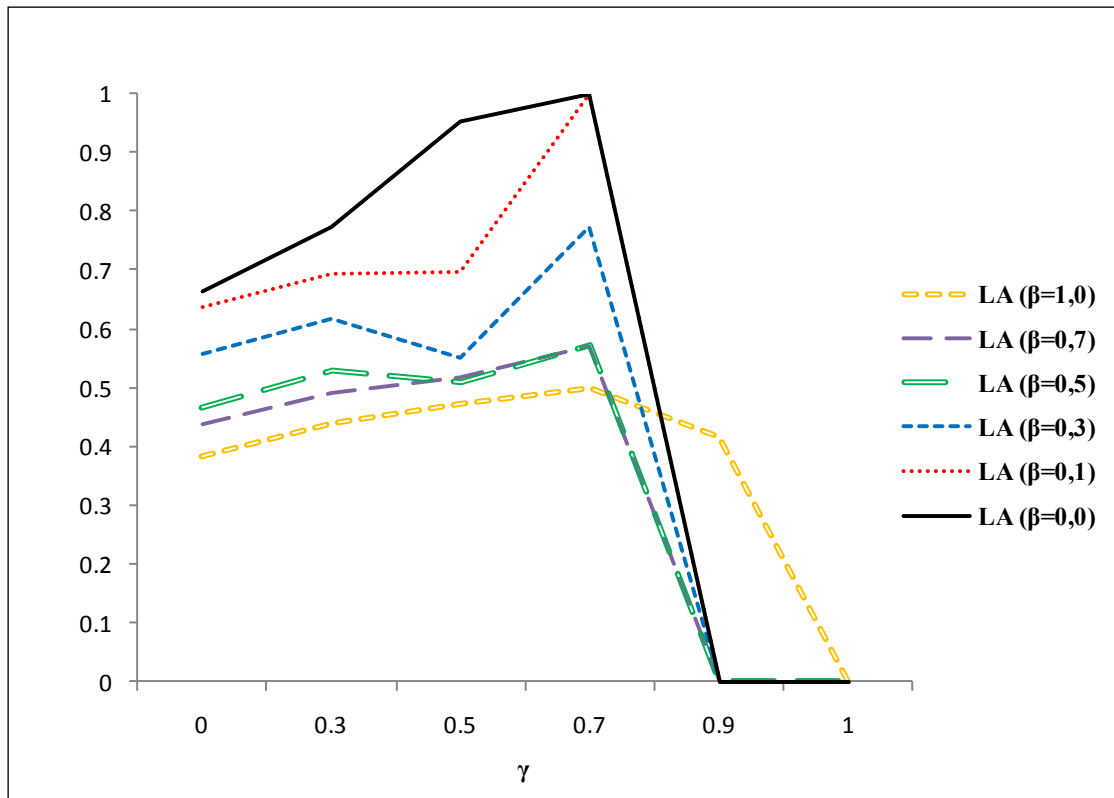


Figure 11: Learning Accuracy (LA) depending on  $\beta$  and  $\gamma$  (Financial glossary).

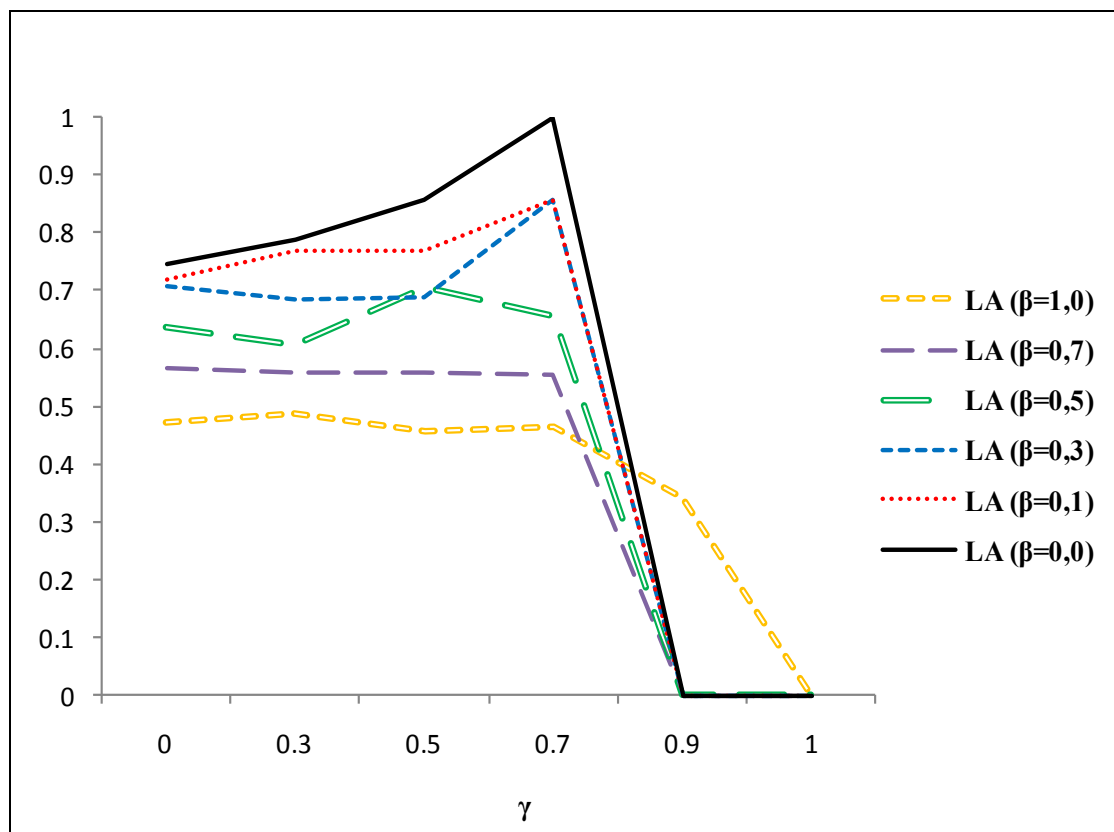


Figure 12: Learning Accuracy (LA) depending on  $\beta$  and  $\gamma$  (ASFA thesaurus).

Figure 11 and Figure 12 display the learning accuracy depending on  $\beta$  (12) and  $\gamma$  (13) for the candidate ontology concepts from the financial glossary and ASFA thesaurus.

It is possible to notice that the higher learning accuracy (LA) is obtained with  $\gamma = 0.7$  in the financial domain and  $\gamma = 0.7$  ( $0.0 \leq \beta \leq 0.3$ ) in fisheries & aquaculture domain.

Figure 13 and Figure 14 show the number of proposed relationships depending on  $\beta$  (12) and  $\gamma$  (13) for the candidate ontology concepts from the financial glossary and ASFA thesaurus.

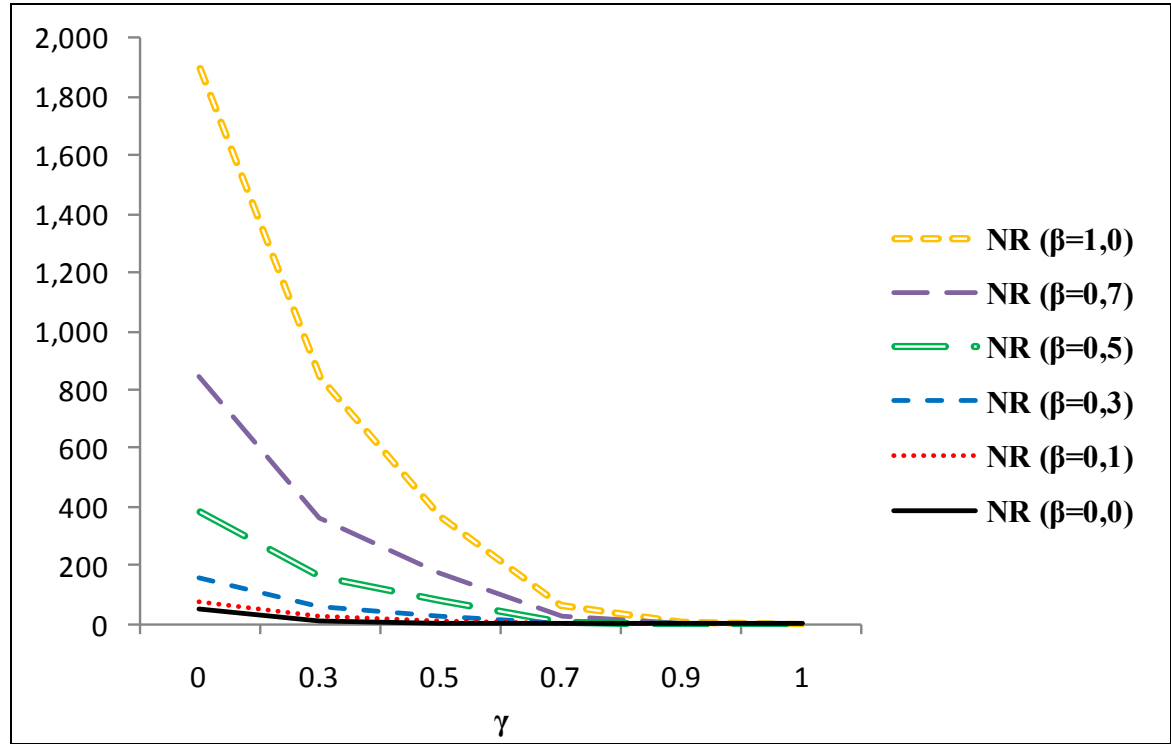


Figure 13: *Number of Proposed Relationships (NR) depending on  $\beta$  and  $\gamma$  (Financial glossary).*

Both in the financial and fisheries & aquaculture domains higher  $\gamma$  values lead to the fewer number of relationships proposed to the user.

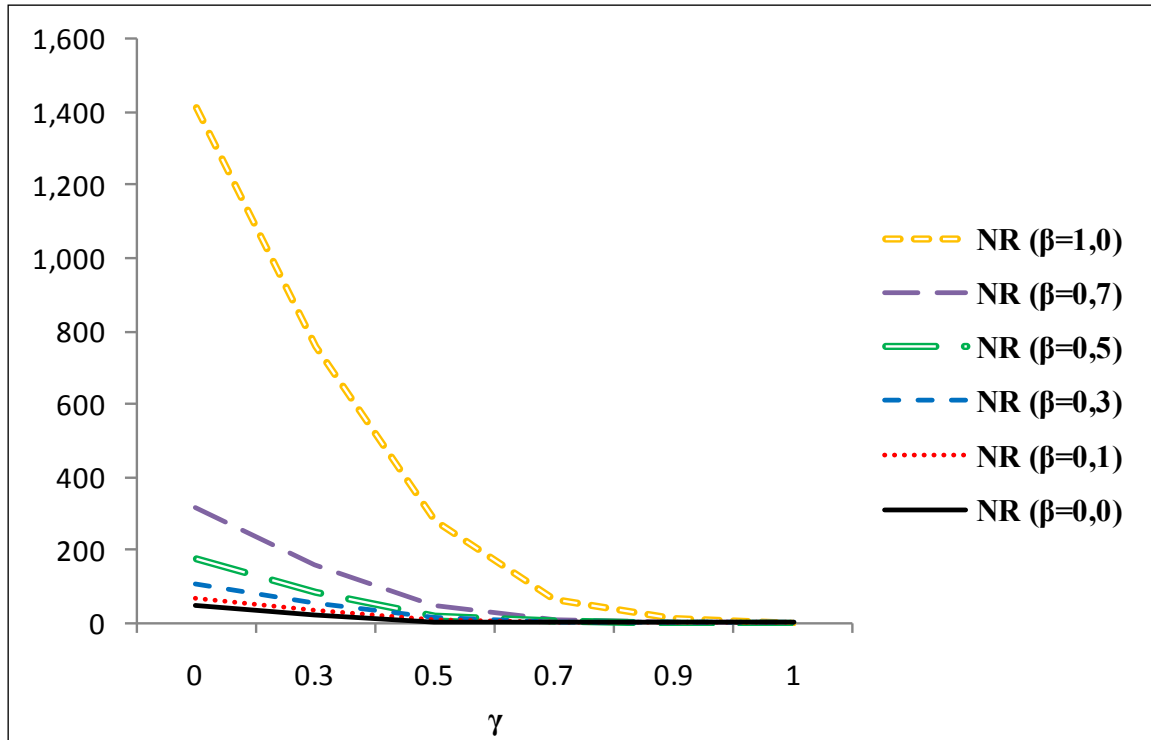


Figure 14: Number of Proposed Relationships (NR) depending on  $\beta$  and  $\gamma$  (ASFA thesaurus).

### 5.2.3 Examples of Cyc KB Extension

Table 8 and Table 9 provide concrete examples of Cyc extension according to the proposed methodology. The related Cyc concepts, which obtained the top position among the suggested related Cyc concepts and in the suggested equivalent, hierarchical (subclass) and associative relations, are highlighted in bold. An example from Table 8 shows that using the proposed methodology and assuming that we would like to extend the Cyc KB with a term *Life insurance* from the financial glossary, we get the composed list of the ranked related Cyc concepts:

- *LifeInsurance*;
- *Endowment-LifeInsurance*;
- *InsurancePlan*;
- *FHAMortgageInsurance*;
- *VAMortgageInsurance*;
- *InsuranceClaimForm*;
- *MedicalInsuranceClaimForm*.

After automatic relation identification the equivalent relation with Cyc concept *LifeInsurance*, hierarchical (subclass) relation with Cyc concept *InsurancePlan* and associative relations with Cyc concepts *Endowment-LifeInsurance*, *FHAMortgageInsurance*, *VAMortgageInsurance*, *InsuranceClaimForm*, *MedicalInsuranceClaimForm* are suggested.

Table 8: *Examples of Cyc KB extension (Financial glossary).*

Concept	Suggested Related Cyc Concepts	Suggested Eqv Rels	Suggested Hier-Subclass Rels	Suggested Associative Rels
FOREX	<b>Foreign</b>			<b>Foreign</b>
	MoneyModeExchange		MoneyModeExchange	
	FinancialExchange		<b>FinancialExchange</b>	
	ExchangeOfUserRights			ExchangeOfUserRights
	BondExchange			BondExchange
	MSExchangeServer			MSExchangeServer
	objectTendered			objectTendered
	NewYorkMercantile Exchange			NewYorkMercantile Exchange
	MonetaryExchangeOf UserRights			MonetaryExchangeOf UserRights
LIFE_ INSURANCE	<b>LifeInsurance</b>	<b>Life Insurance</b>		
	Endowment- LifeInsurance			<b>Endowment- LifeInsurance</b>
	InsurancePlan		<b>InsurancePlan</b>	
	FHAMortgage Insurance			FHAMortgage Insurance
	VAMortgageInsurance			VAMortgageInsurance
	InsuranceClaimForm			InsuranceClaimForm
	MedicalInsurance ClaimForm			MedicalInsurance ClaimForm
ROTATION [Strategy]	<b>Movement-Rotation</b>		Movement-Rotation	
	IndustryOrEconomic SectorType		IndustryOrEconomic SectorType	
	Strategy Assets		<b>Strategy</b>	<b>Assets</b>

Table 9: *Examples of Cyc KB extension (ASFA thesaurus).*

Concept	Suggested Related Cyc Concepts	Suggested Eqv Rels	Suggested Hier-Subclass Rels	Suggested Associative Rels
AQUATIC_ ORGANISMS	<b>AquaticOrganism</b>	<b>Aquatic Organism</b>		
	OrganismType ByHabitat		<b>OrganismType ByHabitat</b>	
SEDIMENT_ MIXING	<b>Sediment</b>			<b>Sediment</b>
	Mixing		<b>Mixing</b>	
SECRETION	<b>SecretionEvent</b>			<b>SecretionEvent</b>
	Secretion-Bodily	<b>Secretion -Bodily</b>		
	Gland			Gland
	Excreting		<b>Excreting</b>	

From Table 9 it is possible to notice that for the term *Aquatic organism* from ASFA thesaurus, we get the following list of the ranked related Cyc concepts:

- *AquaticOrganism*;
- *OrganismTypeByHabitat*.

Furthermore, we get a suggested equivalently related Cyc concept *AquaticOrganism* and hierarchically related Cyc concept *OrganismTypeByHabitat*.

Figure 15 displays two illustrative examples of the Cyc KB extension with user interaction, one in fisheries & aquatic domain (with term *Rare earths* from ASFA thesaurus) and the other in financial domain (term *Recession* from financial glossary).

As proposed in our methodology, the user gets a ranked list of relevant Cyc concepts for each glossary term and confirms the relationships between the glossary term and the proposed concepts. ASFA thesaurus defines *Rare earths* as a narrow term for *Metal*. Using our methodology, for *Rare earth* the user obtains a related concept *Metal* and suggested relationship: “*Rare earth is a subclass of Metal*”. For financial glossary term *Recession* the user obtains two related Cyc concepts – top ranked *Recession-Economic* with suggested hierarchical relationship: “*Recession equals to Recession-Economic*” and Cyc concept *Downturn* with associative relationship: “*Recession conceptually related to Downturn*”.



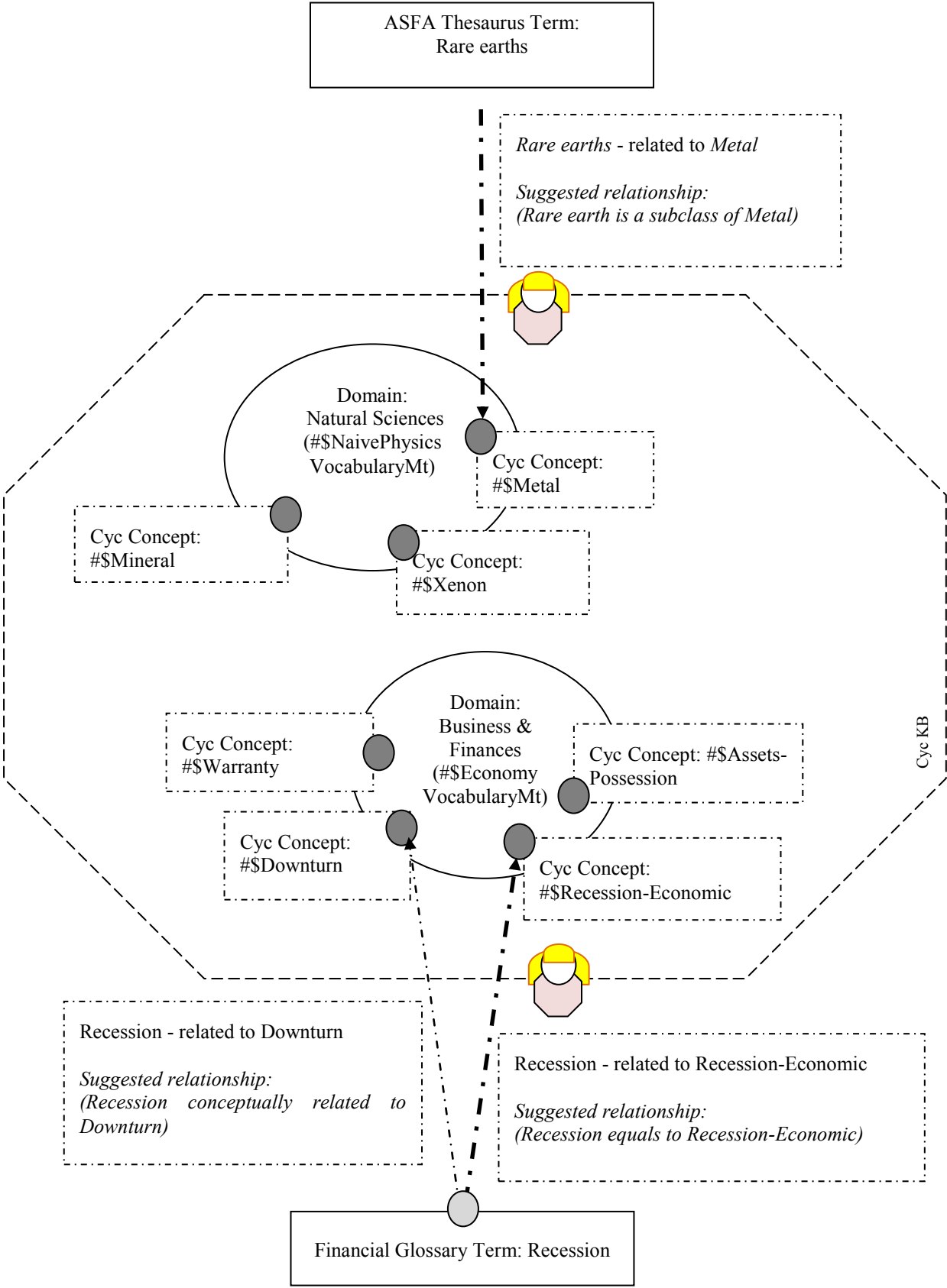


Figure 15: Cyc KB Extension - User Validation.

The combination of user-interaction approach with automatic concept suggestions for ontology extension prevents automatic method from establishing wrong relationships between ontology concepts, at the same time making the extension process faster and more effective than purely manual. It means that using the proposed methodology the user is able to compare large numbers of Cyc concepts with glossary terms and establish relationships in more effective way than just using the manual search for the relevant concepts in Cyc.

### 5.3 Question Answering Experiments

In this section we demonstrate the results of the evaluation of the proposed **pipeline for business news analysis**. In particular, we illustrate the relevance of the proposed Cyc ontology extension for news analysis and question answering in business and financial domain.

As it is stated in the experimental settings, we have performed the question answering evaluation using two scenarios. Firstly, we assumed that we have a simple question and we want to get an answer using the original (prior to the extension) and the extended Cyc Knowledge Base. The first scenario illustrates how specific users in a bottom-up approach can semi-automatically extend an ontology based on their questions of interest applying the **OntoPlus** methodology.

#### 5.3.1 Illustrative Question Answering

The following example illustrates the relevance of the proposed Cyc ontology extension for question answering in the financial domain.

For the research purposes we have selected the following questions about business cycle phases:

***What phase of the business cycle was Egypt in in 2008?***  
***Was Indonesia in contraction in 2008?***

Using the existing original Cyc KB we get no appropriate answers because of the insufficient representation of business cycles in Cyc.

Figure 16 presents the textual definition of business cycle and its phases which we use to implement the notion of business cycles in Cyc.

Using the proposed methodology for semi-automatic ontology extension, we obtain a ranked list of related Cyc concepts for the correspondent glossary term (Table 10).

To enter new assertions into the Cyc KB we use the KE text format which facilitates the knowledge entry process. We select the Cyc concept *Cycle-Situation* as a superclass for glossary term *Business Cycle*:

**KE text:**

```
Constant: BusinessCycle.
In Mt: UniversalVocabularyMt.
isa: TemporalObjectType.
genls: Cycle-Situation.
```

*comment: "Repetitive cycles of economic expansion and recession. The official peaks and troughs of the U.S. cycle are determined by the National Bureau of Economic Research in Cambridge, MA."*

*TERM: BUSINESS CYCLE*  
*COMMENT: Repetitive cycles of economic expansion and recession. The official peaks and troughs of the U.S. cycle are determined by the National Bureau of Economic Research in Cambridge, MA.*

*Phases of Business Cycle:*

*TERM: CONTRACTION*  
*COMMENT: A slowdown in the pace of economic activity.*

*TERM: TROUGH*  
*COMMENT: The lower turning point of a business cycle, where a contraction turns into an expansion.*

*TERM: EXPANSION*  
*COMMENT: A speedup in the pace of economic activity.*

*TERM: PEAK*  
*COMMENT: The upper turning of a business cycle.*

Figure 16: Business Cycle Definition.

Table 10: Related Cyc Concepts for Glossary Term “Business Cycle”.

Glossary Term	Ranked Related Cyc Concepts
BUSINESS CYCLE	Cycle-Situation Recession-Economic MacroeconomicEvent Trough (a type of FluidReservoir)

Furthermore, we create a set of business cycle phases (*Contraction*, *Expansion*, *Peak* and *Trough*) as subclasses for Cyc concept *MacroeconomicEvent*. The following code displays the example of the *Contraction* phase definition:

```
KE text:
Constant: ContractionBusinessCyclePhase.
In Mt: UniversalVocabularyMt.
isa: TemporalObjectType.
genls: MacroeconomicEvent.
comment: "A slowdown in the pace of economic activity".

In Mt: UniversalVocabularyMt.
f:(relationAllExists properSubSituations BusinessCycle
ContractionBusinessCyclePhase).
```

In addition, we create a predicate used for answering questions connected to business cycle phases of the specific countries.

**KE text:**

```
Constant: economyInBusinessCyclePhase.
In Mt: UniversalVocabularyMt.
isa: TernaryPredicate.
arity: 3.
arg1Isa: GeopoliticalEntity.
arg2Isa: TemporalThing.
arg3Isa: MacroeconomicEvent.
```

For the illustrative question answering example we estimate the business cycle phases by using the GDP growth rate - the percentage increase or decrease of Gross Domestic Product (GDP) from the previous measurement cycle. We identify that a term *GDP* is already implemented in the Cyc KB as *grossDomesticProduct*.

The following rule defines the conditions of being in the contraction business cycle phase for the particular country in the specified year. We assume that the contraction phase occurs when the real growth rate of GDP in the referred year  $GR(GDP)_{Y_n}$  decreases comparatively to the real growth rate of GDP in the previous year  $GR(GDP)_{Y_{n-1}}$  but is still higher than the real growth rate of GDP in the following year  $GR(GDP)_{Y_{n+1}}$ :

$$GR(GDP)_{Y_{n-1}} > GR(GDP)_{Y_n} > GR(GDP)_{Y_{n+1}} \quad (23)$$

**KE text:**

```
In Mt: UniversalVocabularyMt.
f:
(implies
  (and
    (evaluate ?SUCCESSOR1 (PlusFn ?Y 1))
    (evaluate ?PREDECESSOR1 (DifferenceFn ?Y 1))
    (evaluate ?PREDECESSOR2 (DifferenceFn ?PREDECESSOR1 1))
    (grossDomesticProduct ?X (YearFn ?SUCCESSOR1) ?S1GDP)
    (grossDomesticProduct ?X (YearFn ?PREDECESSOR1) ?P1GDP)
    (grossDomesticProduct ?X (YearFn ?PREDECESSOR2) ?P2GDP)
    (grossDomesticProduct ?X (YearFn ?Y) ?YGDP)
    (evaluate ?S1GR (QuotientFn ?S1GDP ?YGDP))
    (evaluate ?YGR (QuotientFn ?YGDP ?P1GDP))
    (evaluate ?P1GR (QuotientFn ?P1GDP ?P2GDP))
    (greaterThan ?P1GR ?YGR)
    (greaterThan ?YGR ?S1GR)
    (isa ?PHASE ContractionBusinessCyclePhase)
    (dateOfEvent ?PHASE (YearFn ?Y)))
  (economyInBusinessCyclePhase ?X (YearFn ?Y) ?PHASE)).
```

The expansion, peak and trough phases occur under the following conditions:

*Expansion:*

$$GR(GDP)_{Y_{n-1}} < GR(GDP)_{Y_n} < GR(GDP)_{Y_{n+1}} \quad (24)$$

*Peak:*

$$GR(GDP)_{Y_{n-1}} < GR(GDP)_{Y_n} > GR(GDP)_{Y_{n+1}} \quad (25)$$

*Trough:*

$$GR(GDP)_{Y_{n-1}} > GR(GDP)_{Y_n} < GR(GDP)_{Y_{n+1}} \quad (26)$$

For question answering the information from the Cyc KB about the GDP levels of Egypt and Indonesia in 2006-2009 is used:

**Cyc KB assertions:**

```
(grossDomesticProduct Egypt (YearFn 2009)
  (BillionDollars 470.4))
(grossDomesticProduct Egypt (YearFn 2008)
  (BillionDollars 450.1))
(grossDomesticProduct Egypt (YearFn 2007)
  (BillionDollars 419.9))
(grossDomesticProduct Egypt (YearFn 2006)
  (BillionDollars 392.1))

(grossDomesticProduct Indonesia-TheNation
  (YearFn 2009) (BillionDollars 968.5))
(grossDomesticProduct Indonesia-TheNation
  (YearFn 2008) (BillionDollars 927.7))
(grossDomesticProduct Indonesia-TheNation
  (YearFn 2007) (BillionDollars 874.4))
(grossDomesticProduct Indonesia-TheNation
  (YearFn 2006) (BillionDollars 822.6))
```

After extending the Cyc KB with notion of business cycle and business cycle phases, using the information about GDP from the Cyc KB, it is possible to get answers for the previously asked questions:

**Query:**

```
(economyInBusinessCyclePhase Egypt (YearFn 2008) ?PHASE)
```

**Query result:**

```
*[Explain]PeakBCPhase2008
```

**Query:**

```
(economyInBusinessCyclePhase
  Indonesia-TheNation (YearFn 2008) ContractionBCPhase2008)
```

**Query result:**

```
Query was proven True *[Explain]
```

According to the rules introduced into the Cyc KB, Egypt was in the peak business cycle phase and Indonesia was in the contraction phase of the business cycle in 2008. PeakBCPhase2008 and ContractionBCPhase2008 are the correspondent

instances of PeakBusinessCyclePhase and ContractionBusinessCyclePhase Cyc collections.

The results obtained in the illustrative question answering experiment are comparable with GDP growth rates in Egypt and Indonesia in 2007-2009 (Central Intelligence Agency, The World Factbook, 2010) displayed in Table 11.

Table 11: *GDP Growth Rates in Egypt and Indonesia.*

Country	GDP Growth Rate	Year est.
Egypt	7.2 %	2008
Egypt	4.5 %	2009
Indonesia	6.3 %	2007
Indonesia	6.1 %	2008
Indonesia	4.4 %	2009

### 5.3.2 News Based Question Answering

In our second question answering experiment we have performed news analysis extracting financial concepts with a help of TextGarden (Text-Garden, 2011) tools, OpenCalais service on all entities, events and facts from the news collection and inserting them into the Cyc KB using the **OntoPlus** methodology.

Table 12: *Extracted Entities by OpenCalais.*

Selected Calais Entity Types	Number of Extracted Entities
Company	8322
Continent	587
Country	3950
Currency	2409
EmailAddress	49
Facility	1143
FaxNumber	1
IndustryTerm	13537
MarketIndex	1702
Organization	8207
Person	8300
PhoneNumber	227
Position	8218
Product	349
ProvinceOrState	2973
Region	278
Technology	845
URL	985

We have used N-grams extractor from TextGarden tools for analysis of the collection

of around 3400 business news and for mapping the extracted N-grams to all relevant concepts from the financial glossary (Harvey, 2003).

Applying the **OntoPlus** methodology, we have managed to put the knowledge about 981 financial concepts into the Cyc KB.

Following that, a number of queries (Table 3) were tested using Cyc reasoning tools.

From the collection of around 3400 news stories a fact extraction service managed to extract 55607 entities, and 33294 facts & events, out of which 16335 generic relations.

The number of extracted entities by OpenCalais type is provided in Table 12. It is possible to notice that the most popular entity types, which occur in business news, are Industry terms (13537), Companies (8322), People (8300), Positions (8218) and Organizations (8207).

The number extracted facts and events by OpenCalais type is provided in Table 13.

Table 13: *Extracted Facts/Events by OpenCalais.*

Selected Calais Relation Types	Number of Extracted Relations	Selected Calais Relation Types	Number of Extracted Relations
Alliance	41	ContactDetails	74
AnalystEarningsEstimate	5	CreditRating	47
AnalystRecommendation	56	DebtFinancing	13
Bankruptcy	43	DelayedFiling	5
BonusSharesIssuance	3	Dividend	45
BusinessRelation	91	CompanyTicker	757
Buybacks	28	CompanyUsingProduct	15
CompanyAccountingChange	0	EmploymentChange	28
CompanyAffiliates	167	EmploymentRelation	45
CompanyCompetitor	107	EquityFinancing	10
CompanyCustomer	82	FDAPhase	14
CompanyEarningsAnnouncement	247	IndicesChanges	19
CompanyEarningsGuidance	52	IPO	7
CompanyEmployeesNumber	83	JointVenture	19
CompanyExpansion	32	Merger	38
CompanyForceMajeure	12	PatentFiling	0
CompanyFounded	110	PatentIssuance	0
CompanyInvestment	87	PersonAttributes	167
CompanyLaborIssues	58	PersonCareer	5870
CompanyLayoffs	56	PersonCommunication	254
CompanyLegalIssues	15	PersonEducation	42
CompanyListingChange	4	PersonEmailAddress	27
CompanyLocation	795	PersonRelation	27
CompanyMeeting	11	PersonTravel	44
CompanyNameChange	2	ProductIssues	56
CompanyProduct	128	ProductRecall	19
CompanyReorganization	134	ProductRelease	18
CompanyRestatement	0	SecondaryIssuance	6
CompanyTechnology	231	StockSplit	0

According to the extracted events and facts, business news contain a substantial amount of information about Personal careers (5870), Communication between business related people (254), company tickers (757), location of companies (795) etc.

What rarely occurred in our business news collection or did not occur at all, is the information about patents (Patent filing (0), Patent issuance (0)), information about changes of some company properties (Company accounting change (0), Company listing change (4), Company name change (2), Company restatement (0) etc.).

Using the **OntoPlus** methodology and OpenCalais to Cyc mappings we have populated the Cyc Knowledge Base with entities, events and facts extracted from business news.

Table 14 illustrates the precision of ontology population for selected entity, event and fact types from selected 50 business news. The precision of ontology population displays both the quality of the entity, event and fact extraction with fact extraction service and the quality of entity, event and fact mapping to Cyc. For instance, if a continent entity has been extracted by OpenCalais service and the same continent instance is already modeled in Cyc, then the precision illustrates how correctly OpenCalais continent entity is mapped to Cyc continent instance.

Table 14: *Precision of Ontology Population.*

Selected Calais Entity, Fact, Event Types	Number of Extracted Entities	Precision of Ontology Population
Continent	6	100 %
Country	17	100 %
Currency	3	100 %
Facility	31	100 %
IndustryTerm	123	82 %
MarketIndex	13	100 %
Organization	100	85 %
Person	78	92 %
Position	64	88 %
Product	3	100 %
ProvinceOrState	29	93 %
Technology	8	100 %
URL	23	100 %
All:		~91 %

It is possible to notice that the precision is high for companies, geographical entities, such as countries, continents, market indices, products, URLs etc.

Following that, we have conducted a question answering experiment using queries presented in experimental settings and Cyc reasoning tools for ontology based reasoning. The precision of question answering is provided in Table 15.

The result values provided in Table 14 state that ontology population with information from business news with a combination of the ontology based reasoning can contribute to the effective query processing.

Different types of queries and different queries give different resulting precision, but for most of the selected queries the precision of question answering exceeds 90 %.



Table 15: *Evaluation of News Based Question Answering.*

№	Experimental Queries (Questions)	Precision
1	Get companies with more than 100 employees	95 %
2	List companies, which participated in mergers (or acquisitions) and the bankruptcy of which was reported in the news	100 %
3	List people in high positions in companies/organizations with residence in District Of Columbia (US)	100 %
4	Are there any people, accused of something or convicted in something, who work in large companies or organizations (>100 employees)?	100 %
5	Which corporations were reported in the news as issuing securities?	67 %
6	Get business partners from IT sector for particular company (expl: Enterra Solutions)	100 %
7	List companies, which produce cars and their affiliates	100 %
8	Get companies-competitors and products, which they produce	100 %
9	Where are company customers located? – Get locations of customers for companies	100 %
10	Was any company founded before year 1990? What does it produce?	100 %
11	Get companies, which were involved into layoff activity, and their residence. List layoffs in Michigan-State and in Georgia-State.	100 %
12	Get companies, against which there were reported lawsuits	100 %
13	List company meetings of a particular company (Atsc) and conference calls of a particular company (ParTechnologyCorporation)	100 %
14	Were there reported any companies, which have participated in reorganization or restatement?	90 %
15	Get stock ticker symbols of companies in difficulties - bankruptcy of which was reported in the news or which had any labor issues or lawsuits reported	100 %
16	Get person profile - contact details of a person, person attributes (age, birthdate, birthplace, gender etc). Examples: AlexisMcgee, JenniferPaige	98 %
17	Were there any family relation found between employer/employee?	100 %
18	Who works where? - Get all analysts	97 %
19	Get people reported working as economists in companies from financial sector	100 %
20	Which US companies were involved in IPO activity?	100 %
All:		~ 97 %

Illustrative results of ontology based reasoning and question answering are given in Figure 17 and Figure 18. Figure 17 shows the Cyc reasoning tools proof for query:

*Are there any economists working in companies from the financial sector?*

Proof:  
 (companySector-SP *LehmanBrothers Financials-SectorSP*) in *StandardAndPoorsDataMt*  
 (positionOfPersonInOrganization *MingchunSun* *LehmanBrothers* *Economist*) in  
*OpenCalaisFactExtractionMt*  
**:ISA** (*isa MingchunSun Person*) in *EverythingPSC*

Figure 17: *Economists in companies from financial sector: **Mingchun Sun**.*

The proof in Figure 17 says that Lehman Brothers (*LehmanBrothers*) is a company from the financial sector (*Financials-SectorSP*) and *MingchunSun* is a person, who works at a position (*positionOfPersonInOrganization*) of economist (*Economist*) in Lehman Brothers (*LehmanBrothers*) company.

Figure 18 demonstrates how Cyc reasoning tools find an answer for the following query:

***Are there any companies from Georgia-State in USA involved into layoff activities?***

Proof:  
 (capitalCityOfState *Georgia-State* *CityOfAtlantaGA*) in *UnitedStatesGeographyMt*  
 (actors *CompanylayoffsAnnouncedDeltaAirLines-DT-2011-03-14-16-19-59* *DeltaAirLines*) in  
*OpenCalaisFactExtractionMt*  
 (residenceOfOrganization *DeltaAirLines* *CityOfAtlantaGA*) in *OpenCalaisFactExtractionMt*  
**:GENLPREDS** (*genlPreds capitalCityOfState geographicallySubsumes*) in *EverythingPSC*  
**:ISA** (*isa DeltaAirLines Business*) in *EverythingPSC*  
**:ISA** (*isa CompanylayoffsAnnouncedDeltaAirLines-DT-2011-03-14-16-19-59 EmployeeLayoff*) in  
*EverythingPSC*

Figure 18: *Get companies, which were involved into layoff activity from Georgia-State (USA): **Delta AirLines**.*

From the proof at Figure 18 we can notice that Atlanta (*CityOfAtlantaGA*) is a capital city (*capitalCityOfState*) of Georgia state, USA (*Georgia-State*). Delta Air Lines (*DeltaAirLines*) is a company, which has residence (*residenceOfOrganization*) in Atlanta (*CityOfAtlantaGA*).

It is possible to see that state geographically incorporates (*geographicallySubsumes*) capital city. Also, the proof says that Delta Air Lines (*DeltaAirLines*) participated in an event, connected to the layoffs (*CompanylayoffsAnnouncedDeltaAirLines-DT-2011-03-14-16-19-59*).

Consequently, the extension of the Cyc Knowledge Base according to the proposed **OntoPlus** methodology, population of the Cyc Knowledge Base with entities, events and facts extracted from business news allows users to perform a question answering based on the extended and populated ontology.

## 5.4 Software Demonstration

In order to demonstrate the **OntoPlus** methodology in practice, as a part of this research, we have developed a software prototype.

The software (size: 4.85 MB) is implemented in Java programming language. The software architecture includes several components, such as user interface component, preprocessing component, ontology subset extraction component and similarity calculation component. The goal of the user interface component is the communication with the user – specification of the domain relevant information (domain glossary; domain keywords; content, structure and co-occurrence weights etc.) by the user and displaying the results (a list of glossary terms with related ontology concepts and potential relationships) to the user. The preprocessing component has a number of functions for stemming, tokenization, stop words removal etc. The goal of ontology subset extraction is to obtain the relevant domain subset of the Cyc ontology based on domain glossary and domain keywords – relevant Cyc concepts with comments. Finally, the similarity calculation component contains functions for similarity calculation between existing and candidate ontology concepts.

The software works with the Cyc Knowledge Base and glossary files in textual formats.

According to the **OntoPlus** methodology, the user should initially select the relevant domain information – domain glossary and domain keywords.

Figure 19 and Figure 20 demonstrate the process of entering information about domain into the software - choosing the domain relevant glossary and domain keywords.

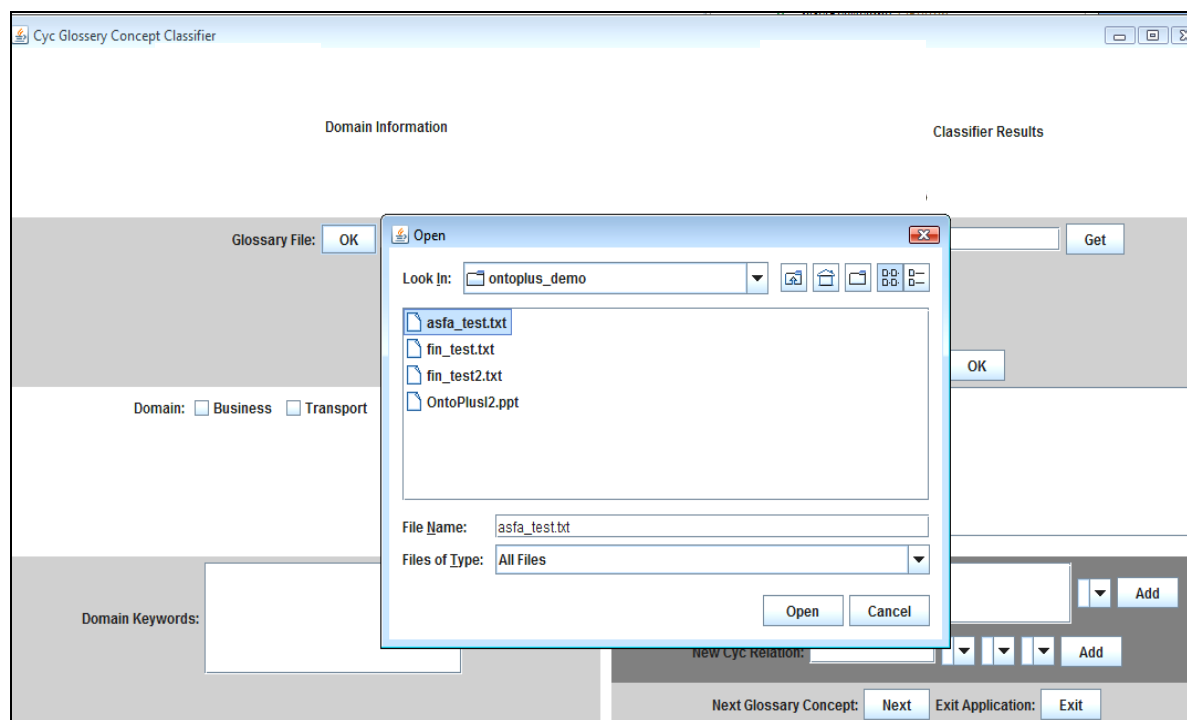


Figure 19: **OntoPlus** software demonstration. Choosing glossary file.

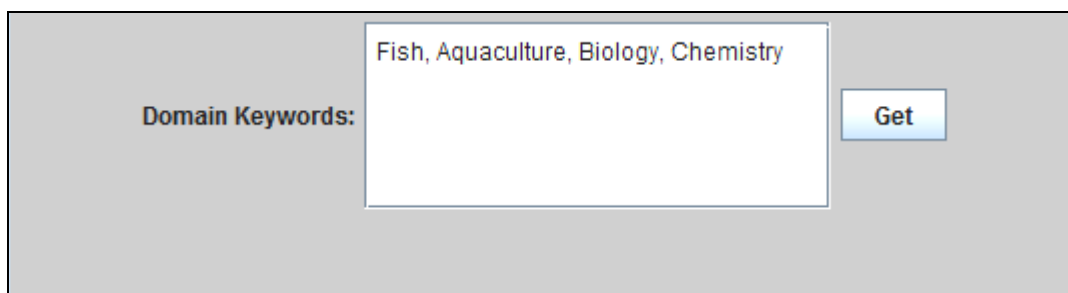


Figure 20: *OntoPlus* software demonstration. Domain keywords.

The user chooses the similarity weights – weights, which are put on content, structure and co-occurrence information (formula 10). The weights are assumed to be between 0 and 1. The content is represented by the string value “*coms*”, the structure is represented by the string value “*genls+specs*”, where “*genls*” is referred to the upper-level structural terms and “*specs*” is referred to the lower level structural terms, and the co-occurrence is represented by the string value “*co-occurrence*”. In addition, the user can regulate the deviation measure “*deviation*”, which is also assumed to take values between 0 and 1 (formula 12).

Figure 21 shows the process of the weight selection (where the user put half weight on content and the other half on structure).

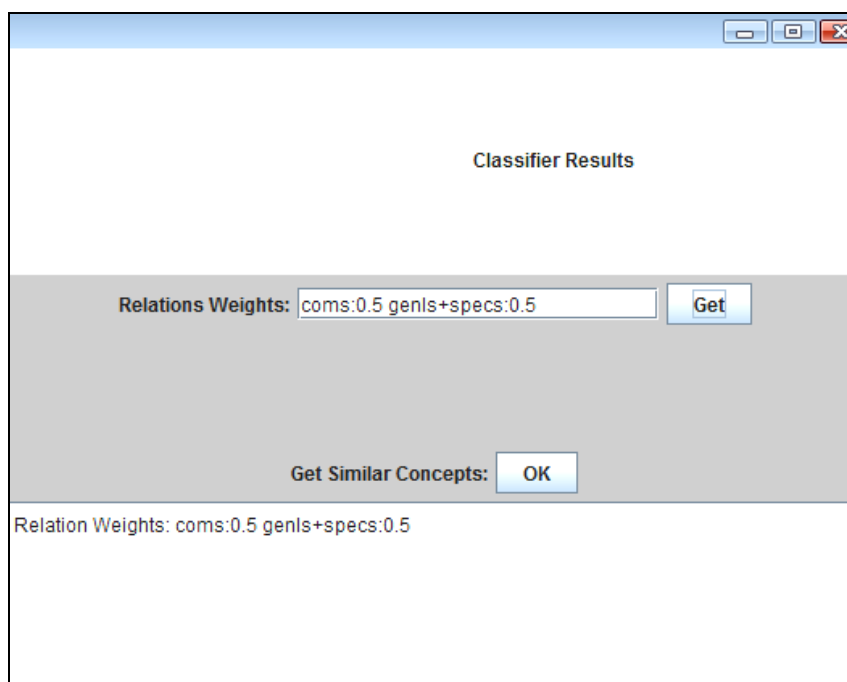


Figure 21: *OntoPlus* software demonstration. Content, structure, co-occurrence weights.

After entering the available domain information and selecting similarity weights, the user can obtain the relevant related concepts and relationships and extend the ontology with correspondent new knowledge.

Figures 22 and 23 demonstrate obtaining related concepts and suggested relationships



Classifier Results

---

Relations Weights:

Get Similar Concepts:

---

Concept: ALLOZYMES - Enzymes  
 Similarity: 0.11431397786957938 EnzymeMolecule - This is the collection of (individual) enzyme molecules, which are globular protein r  
 -----  
 Suggested relationship: Allozymes might be SPECS with difference in similarities 0.5109967623076722 to EnzymeMolecule

---

◀  ▶

---

New Cyc Concept:

New Cyc Relation:

---

Next Glossary Concept:  Exit Application:

Figure 23: *OntoPlus* software demonstration. Results – fisheries & aquaculture domain example.

The created software allows working with the Cyc Knowledge Base in the interactive regime. Moreover, the user can also obtain a separate textual file with the pool of suggested concepts and relations for the whole domain glossary.

## 6 Discussion

In this thesis we address the process of semi-automatic text-driven ontology extension using the ontology content, structure and the co-occurrence information with application of the extended ontology for news analysis. In this chapter we compare the proposed **OntoPlus** methodology with different ontology learning methodologies. The proposed **pipeline for business news analysis** in comparison to various news analysis and question answering approaches is as well discussed in this chapter.

### 6.1 Methodology for Ontology Extension

In this thesis we proposed a novel **OntoPlus** methodology for text-driven ontology extension. In contrast with many other methodologies for ontology extension, our methodology deals with ontologies and knowledge bases, potentially covering more than one domain. However, it allows restricting the area of ontology extension to a specific domain and users deal only with their sphere of interest.

The **OntoPlus** methodology is based on considering advantages and shortcomings of a number of ontology learning approaches. The **OntoPlus** methodology aggregates several ontology extension phases discussed in details in chapter 3. While developing the **OntoPlus** methodology, we took to the account the classical stages of ontology creation (Uschold and King, 1995; Grüninger and Fox, 1994; Corcho et al., 2005; Sure and Studer, 2002), such as purpose and motivating scenarios identification, building the ontology, evaluation etc.

In general, the majority of methods for ontology learning from text include the following steps (Reinberger and Spyns, 2005): collecting, selecting and preprocessing of an appropriate corpus, discovering sets of equivalent words and expressions, establishing concepts with the help of the domain experts, discovering sets of semantic relations and extending the sets of equivalent words and expressions, validating the relations and extended concept definition with help of the domain experts and creating a formal representation. Automating the search of related concepts and correspondent relationships for candidate ontology concepts, the **OntoPlus** allows for more efficient solutions of ontology extension.

In a similar way to Prieto-Diaz (2002), the **OntoPlus** methodology incorporates top-down and bottom-up processes, where the user is providing relevant keywords or glossary while the system uses the data to identify relevant parts of the existing ontology.

Analogically to the approach presented by (Witbrock et al., 2003), the **OntoPlus** methodology relies upon user, in particular, upon user interaction and user validation of suggested relationships between existing and candidate ontology concepts.

Along with many other researchers (Lenat and Guha, 1990; Lenat, 1995; Shah et al., 2006; Witbrock et al., 2003; Medelyan and Legg, 2008; Sarjant et al., 2009; Taylor et al.,

2007; Lenat et al., 2010), we have selected Cyc for our ontology extension and population tasks and ontology based news analysis. Cyc aggregates a large common-sense ontology, suitable for news formalization, and an inference system, which allows one to perform reasoning based on the formalized knowledge.

Text2Onto framework (Cimiano and Völker, 2005) for ontology learning and SPRAT tool (Maynard et al., 2009) for ontology population can be compared to our methodology in a number of ways. In Text2Onto, the user specifies a corpus - text collection used in ontology learning. In our case, the user defines a set of domain keywords and determines the domain relevant glossary. We expect that the efforts of keywords and glossary specification do not exceed the efforts of text corpus identification. As in Text2Onto, user interaction in the proposed methodology helps in avoiding adding to the ontology irrelevant concepts and relationships. Unlike the authors of Text2Onto and SPRAT tools, we do not use linguistic patterns for concept and relation identification. Instead, we use statistically driven approaches what makes our methodology more language independent.

The experimental results show that by exploiting ontology structure information, the **OntoPlus** methodology achieves a precision of 29.5 % for subclass relation identification in the financial domain and a precision of 20.3 % for subclass relation identification in the fisheries & aquaculture domain. The precision for equivalent relation identification is 60.0 % for the financial domain and 94.7 % for fisheries & aquaculture domain. Although Text2Onto and SPRAT results cannot be directly compared to our results, in order to provide a different perspective on ontology extension problem, we report them in this thesis. The authors of Text2Onto framework obtained a precision of 17.38 % for subclass-of relation identification on the subset of tourism-related texts (Cimiano and Volker, 2005). The creators of the SPRAT tool report the precision of 48.5 % for subclass identification and 48.0 % for synonym recognition on a subset of Wikipedia articles about animal (Maynard et al., 2009). Although there is no specified bottom line for ontology extension problem, we can say that quantitative results of the **OntoPlus** evaluation look competitive if we compare them with other methodologies for ontology learning – for instance, with methodologies, which use only content or only co-occurrence information.

If we compare the proposed methodology to the approaches used in OntoGen (Fortuna et al., 2007), we can notice resemblance in using text mining technology for handling textual data and measuring similarity. While in our case, we deal with extension of general multi-domain ontology, OntoGen focuses on topic ontology construction and applies several machine learning and data visualization methods that are not used in our approach. In such way, the **OntoPlus** methodology is able to perform within different domains and different information sources. For this reason, we can say that the proposed methodology goes beyond the topic ontology construction (Fortuna et al., 2007).

The **OntoPlus** methodology allows for the effective extension of the very large ontologies, such as the Cyc Knowledge Base. The methodology provides the user with required concepts and relationships in the form of the ranked list. The evaluation of the **OntoPlus** methodology confirms that combining textual the ontology content, the ontology structure and the co-occurrence information we can provide the user with higher number of concepts suitable for the ontology extension than using only concept denotations, only the ontology content or only the co-occurrence analysis.

The **OntoPlus** methodology allows transforming textual information organized at different knowledge representation levels into a structured conceptualized form. Unlike the approach in (Sarjant et al., 2009), the proposed methodology works even if no taxonomically structured data is available as an input. Manual building of large ontologies, such as the Cyc Knowledge Base, demands a substantial amount of human



effort. Further extension of such a large ontology is challenging as well. From the hundreds thousands of concepts, the proposed methodology is able to find the concepts and relationships the user needs and present them in a ranked list based on their relevance. The utilization of the lexical and structural information of the extensive knowledge bases and ontologies contributes to their infinite extension and reuse.

Consequently, the applicability to very large multi-domain ontologies, the possibility of diverse textual sources utilization and the usage of the language independent approaches represent the strengths of the proposed methodology. However, the proposed methodology as presented in this thesis is mainly applicable for extension of the ontology, which has a sufficient lexical representation of its components.

In this research the evaluation of the **OntoPlus** methodology was performed with Cyc running on the server and the **OntoPlus** implementing software running on the client machine. More extensive evaluation of the **OntoPlus** methodology can follow from the deployment of the proposed methodology in public – e.g., on the OpenCyc (OpenCyc, 2011) website, with crowdsourcing tools (e.g., Mechanical Turk, 2011) for review. With help of Mechanical Turk, domain specialists can identify the correct proposed related Cyc concepts and relationships.

## 6.2 Pipeline for Business News Analysis

The thesis addresses the process of business news analysis by ontology extension, ontology population with entities, facts and events extracted from text and reasoning based on the obtained ontology.

The **pipeline for business news analysis** presented in this thesis is based on the wide spectrum of business entities, events and facts, ontologically represented. We do not perform Open information extraction, as do (Etzioni et al., 2008; Soderland et al. 2010). However, at the same time, while ontologically representing subject-predicate-object textual triplets, we automatically connect the predicate to the most related event from the ontology.

As in (Soderland et al., 2010), we populate the ontology with new information. Analogically to (Tunstall-Pedoe, 2010) and (Lenat et al., 2010), we use a large common sense ontology for reasoning and question answering. We use news as the primary source of data as approaches (Andersen et al., 1992; Losch and Nikitina, 2009; Iacobelli et al., 2010) do.

As in Bradeško et al. (2010), we use the Cyc ontology and ASFA in our experiments – the **OntoPlus** methodology tests are performed on the Cyc KB with ASFA thesaurus as domain relevant resource. However, our question answering experiments are based on the extension and population of the Cyc ontology and reasoning based on the extended and populated ontology. In their question answering experiments Bradeško et al. (2010) use only synonymic and hierarchically related terms. At the same time, Cyc has an implemented reasoning system, which goes far beyond “hierarchical” and “synonymic” question answering.

If we compare our techniques to the existing methods of news analysis, we can notice that similarly to (Losch and Nikitina, 2009), we utilize semantic information from the ontology. As in (Iacobelli et al., 2010), we extract facts from a set of news stories. However, our approach for news analysis is further semantically driven, since we also use ontology to reason based on the facts obtained from business news and common-sense

facts, existing in the ontology.

The **pipeline for business news analysis** presented in this thesis constitutes a whole strategy of business news analysis and question answering based on the ontology reasoning and information from the news. In our research we combine different tools – fact extraction system and common-sense knowledge base with inference system, which allows to automatically extract facts from textual sources, populate the ontology and reason based on the populated ontology.

Comparing our techniques of ontology based news analysis with other existing approaches, we can say that using extended large common-sense ontologies, such as the Cyc Knowledge Base and reasoning based on the information existing in these ontologies can provide users with a wider spectrum of results than using specific separate ontologies for news events or different pattern-based systems. The results of the experiments justify the suitability of the **OntoPlus** methodology for ontology extension and the applicability of large lexical ontologies, such as the Cyc Knowledge Base, to analysis of textual sources, in particular business news.

## 7 Conclusions

### 7.1 Summary

In this research we have set a goal of contributing to the analysis of the financial news by means of semantic technologies - in particular by extending and populating the business and financial ontology in Cyc, which is known to have one of the largest knowledge bases in the world. Utilizing the ontology lexical and structural feature, we aimed to make the process of ontology extension more productive.

The **OntoPlus** methodology for text-driven ontology extension, combining text mining methods and user-interaction approach, has been suggested and exposed to evaluation. The evaluation of our methodology has been accomplished in two rather different domains; for the financial domain a glossary was available while for the fisheries & aquaculture domain a thesaurus has been used as a source of terms to be added to the existing ontology. In this research we demonstrated that the proposed methodology works for textual data structured at different knowledge representation levels.

Manual building of large ontologies, such as the Cyc Knowledge Base, demands a substantial amount of human effort. Further extension of such a large ontology is challenging as well because of its complexity and interconnectivity. The **OntoPlus** methodology presented in this research is meant to speed up the process of building an extensive ontology and lower the price of doing it.

With the **OntoPlus** methodology for each glossary term the user is provided with a ranked list of related ontology concepts and a ranked list of potential relations. We have found that the importance of the ontology content, structure and the co-occurrence information can vary for different domains and knowledge representations used in the process of ontology extension. The best results are achieved by combining content, structure and co-occurrence information for our data in the financial domain. At the same time, the ontology content and the co-occurrence are more important for our fisheries & aquaculture data.

Furthermore, the thesis addresses the process of business news analysis by ontology extension and ontology population with formalized knowledge extracted from text and reasoning based on the obtained ontology. In our research we proposed the **pipeline for business news analysis**. We showed that ontology extension can provide better annotation possibilities for textual data, such as business news. Analyzing news with means of semantic technologies, such as large common-sense ontologies and reasoning systems, we were able to answer a number of business queries.

We have found that using the **OntoPlus** methodology for the ontology extension with relevant concepts, using the **pipeline for business news analysis** for ontology population with entities, events and facts extracted from news and utilization of the obtained extended and populated ontology for reasoning and question answering provide a user with a possibility to automatically analyze textual financial and business data, to detect important information and to save time.

### 7.1.1 Methodology for Ontology Extension

In our research we have proposed a novel **OntoPlus** methodology for ontology extension. The proposed methodology operates with lexical and structural ontology information. The **OntoPlus** methodology uses the combination of the ontology content, the ontology structure information and the co-occurrence data between existing and candidate ontology concepts and consists of several methodology phases:

1. *Domain information identification.* The user identifies the appropriate domain keywords. As well, in this module a domain relevant glossary, containing terms with descriptions is determined.

2. *Extraction of the relevant domain ontology subset from multi-domain ontology.* The relevant domain ontology subset is obtained based on the specified domain information. The domain keywords are mapped to the natural language representation of the ontology domain information and a set of the relevant domains of interest is identified. Further, ontology concepts defined in these domains are extracted.

3. *Domain relevant information preprocessing.* The information from the domain-relevant glossary and the extracted relevant ontology subset are linguistically preprocessed. The preprocessing phase includes tokenization, stop-word removal and stemming. Textual information is represented using bag-of-words representation with TFIDF weighting and similarity between two text segments is calculated using cosine similarity between their bag-of-words representations. For each term from the domain relevant glossary we compose a bag-of-words aggregating preprocessed textual information from: (1) the glossary term name and (2) the term comment. For each concept from the extracted relevant ontology subset the following information is considered: (1) the ontology concept content consisting of the preprocessed natural language concept denotation and concept comment; (2) the ontology concept structure consisting of the preprocessed natural language concept denotation and natural language denotations of hierarchically and non-hierarchically related concepts. In addition, for relation identification, for each ontology concept we compose two additional bags-of-words: one with natural language denotation of the concept and natural language denotations of superclasses of this concept, another with natural language denotation of the concept and natural language denotations of subclasses of this concept.

4. *Composing the list of potential concepts and relationships for ontology extension.* The ranked list of the relevant concepts and possible relationships suitable for ontology extension is composed in this phase. Cosine similarity between glossary term and ontology concept content is calculated and weighted with weight defined by the user. Cosine similarity between glossary term and ontology concept structure is calculated and weighted with weight defined by the user. We use Jaccard similarity to measure the co-occurrence of glossary term and ontology concept.

Ontology concepts with similarity larger than a defined similarity threshold are suggested to the user.

To propose the relationship of equivalence we use string-edit distance between glossary term names and the related concept names. In the case of equivalence, the user can extend textual representation of the related ontology concept. Hierarchical subclass relationship is proposed, when the similarity between the glossary term and subclasses of the related concept is higher than the similarity between the glossary term and superclasses of the related concept.

5. *User validation.* Furthermore, the user validates the candidate entries results consisting of the glossary terms, existing ontology concepts and glossary term-ontology concept relationships. In case of the equivalence relationship the user can extend the

textual representation of the existing ontology concept by adding comment, adding or changing the natural language denotation. In case of the hierarchical relationships the user can add subclasses to the existing ontology concepts. If the nature of the relationship is not clear, the user can create an associative relationship or choose any other relationship between a glossary term and existing ontology concept. Moreover, the list with validated entries in the relevant format is created.

6. *Ontology extension*. It represents adding the new concepts and relationships between concepts into the ontology.

7. *Ontology reuse*. The ontology reuse phase serves as the connection link between separate ontology extension processes. As a part of the new extension process, we reuse the previously extended ontology.

The novel features of the **OntoPlus** methodology include (a) the combination of the ontology content, structure and the co-occurrence information in the ontology extension process, (b) the usage of ontology structure in textual format, (c) the possibility to obtain the relevant domain ontology subset from multi-domain ontology, (d) the feasibility of application to the Cyc ontology extension.

Text mining plays a central part in our ontology extension methodology. As domain information we use domain keywords and domain relevant glossary. With text mining techniques we are able to transform the unstructured textual information into formalized knowledge.

The **OntoPlus** methodology is mainly targeted at the ontology engineers – people, who develop and maintain large ontologies, such as the Cyc KB. The user interaction comes in two ways. The users define the domain relevant information – domain keywords and domain relevant glossary and afterwards, user interaction plays a filtering role in the **OntoPlus** methodology. With user validation we are able to avoid the insertion of the non-relevant ontology concepts and relationships.

Furthermore, we have adapted the proposed **OntoPlus** methodology in order to obtain a comprehensive specific methodology for the Cyc Knowledge Base extension. The main adaptations compared to the methodology described above are based on microtheories (Contexts in Cyc, 2011) that Cyc is using to represent thematic subsets of the ontology.

In experiments the suggested methodology for text-driven ontology extension, aggregating the elements of text mining and user interaction approach for ontology extension, was used for inserting the new knowledge into the Cyc Knowledge Base.

The experiments have been performed in two different domains having two knowledge representation levels – financial domain represented by the glossary of financial terms (Harvey, 2003) and fisheries & aquaculture domain represented by Aquatic Sciences and Fisheries Abstracts (ASFA) thesaurus (ASFA thesaurus, 2010).

### 7.1.2 Pipeline for Business News Analysis

We have handled news analysis problem with a **pipeline for business news analysis**. The **pipeline for business news analysis** aggregates obtaining entities, events and facts with fact extraction service, the **OntoPlus** methodology for the Cyc ontology population and reasoning based on the extended and populated ontology.

The following phases have been defined in the **pipeline for business news analysis**:

1. *News website definition*. A list of websites, which contain business news, is defined.

2. *News crawling.* The news is crawled from the RSS feeds of the provided websites and afterwards, news cleaning is performed. Every news article represents a separate textual file.

3. *Concept, entity, event, fact extraction from news.* A set of financial concepts is extracted from business news. Using N-grams extractor from TextGarden tools (Text-Garden, 2011), it is possible to get all N-grams from the textual news collection and map them to the terms in the Harvey financial glossary (Harvey, 2003). With fact extraction service, such as OpenCalais tool, we are able to extract the information about entities, events and facts present in our news collection.

4. *Concept, entity, event, fact mapping to the Cyc KB.* In this phase ontology extension and ontology population are performed. With the **OntoPlus** methodology we are able to extend the Cyc KB with terms from the financial glossary, which occurred in our news collection. For ontology population we have created a set of mappings between OpenCalais entities, events and facts types and Cyc concepts – collections and predicates. We also apply the **OntoPlus** methodology for concept disambiguation in the ontology population process.

5. *Question definition.* For question definition a set of questions, involving reasoning aspects is composed. For the business news analysis we have composed business related questions.

6. *Question answering.* The questions are asked using the Cyc reasoning interface and Cyc proofs are analyzed.

In order to perform the Cyc ontology population, we have created a set of mappings between fact extraction service types and Cyc concepts and relations. Additionally, we have proposed an algorithm for subject-predicate-object textual triplet mapping to the Cyc Knowledge Base.

In our news analysis experiments we have used a collection of the financial news from the yahoo website (Yahoo! Finance, 2010). We have crawled set of news stories, extracted entities, facts and events with fact extraction service (OpenCalais, 2010) and applied the **OntoPlus** methodology to map the extracted knowledge into the Cyc KB. Following that, a number of queries have been analyzed through the Cyc reasoning interface. The experiments confirmed utilizing a **pipeline for business news analysis** it is possible to effectively obtain and analyze information from the news and perform ontology based reasoning.

## 7.2 Scientific Contributions

The major contribution of this thesis is in proposing a novel **OntoPlus** methodology for text-driven semi-automatic ontology extension using the ontology content, the ontology structure information and the co-occurrence data between existing and candidate ontology concepts.

Second contribution of our work lies in the **OntoPlus** methodology adaptation to one particular scenario – the Cyc ontology extension.

Third contribution of this thesis is the evaluation of the proposed **OntoPlus** methodology for ontology extension on real world data in two different domains having two representation levels – financial domain represented by the glossary of terms and fisheries & aquaculture domain represented by the thesaurus. The experimental results

have demonstrated that using a combination of the ontology content, structure and the co-occurrence information is more beneficial for the extension of large multi-domain ontologies, than using only content, only co-occurrence or only concept denotation information. Moreover, our contribution to the ontology extension based on textual information is in defining the best combination of content, structure and co-occurrence measures for financial domain and fisheries & aquaculture domain. We have found that combining content, structure and co-occurrence information for our data leads to the best results in the financial domain. However, combining the ontology content and the co-occurrence information is more effective for fisheries & aquaculture data.

Then, we have explored the process of business news analysis by the ontology extension with relevant financial concepts, ontology population with entities, facts and events extracted from text and reasoning based on the obtained ontology. The forth contribution of our research lies in proposing a **pipeline for business news analysis**.

As a fifth contribution, we have extended and populated the Cyc ontology with business and financial terms, entities, events and facts obtained from news.

Finally, applying the **OntoPlus** methodology and a **pipeline for business news analysis** we were able to competently deal with a specific and non trivial task of business news analysis, to discover, combine and use in reasoning and question answering the important information extracted from business news.

In order to achieve the described contributions we were guided by the aims stated in Introduction section. Here we summarize the work done for each goal:

1. We have specified the terminology used in the research (section 1.2);
2. We have proposed the methodology for ontology extension (section 3.2) and the **pipeline for business news analysis** (section 3.4);
3. We have adapted the proposed methodology to the Cyc Knowledge Base extension (section 3.3);
4. We have experimentally evaluated the proposed methodology for ontology extension in different domains and different knowledge representations (section 5.2);
5. We have created software with user interface for the Cyc Knowledge Base extension for a selected domain (section 5.4);
6. We have evaluated the proposed **pipeline for business news analysis** (section 5.3). We have explored a set of financial news and determined the relevant financial concepts and instances (we have extended the Cyc ontology with 987 financial concepts and 55607 entities);
7. We have obtained a number of financial events and facts from a set of business news (by analyzing business news, we have found 33294 facts & events; section 5.3);
8. We have identified a set of interesting non trivial queries for business news analysis and answered them using the Cyc ontology and the Cyc reasoning interface (section 5.3).

### 7.3 Future Work

Different paths for future work arise from the presented research ideas and methods. We expect our **OntoPlus** methodology to be exploited for various ontology learning and ontology alignment purposes.

The future work in the direction of ontology extension should include further augmentation of the Cyc Knowledge Base and testing the textual data analysis in different domains.

For number of other domains (besides financial domain and fisheries & aquaculture domain) the best combination of content, structure and co-occurrence for ontology extension should be identified. As well, more exhaustive experiments across several domains involving a number of ontologies from the same domain would be needed to investigate how properties of the ontology relate to the usefulness of content, structure and co-occurrence information.

Particular attention should be devoted to better automatic identification of hierarchical relations and extraction of different types of non-hierarchical relations (part-whole relations, relation instances of the existing ontology predicates for the particular domain).

In case of ontology population, the named entity disambiguation problem should be resolved at a higher level and several approaches for concept disambiguation should be tested.

Also, an interesting future work approach would be better fact identification and extraction from text with subsequent ontology population with facts.

Finally, the **pipeline for business news analysis** should be modified in order to analyze various textual information, for examples, news from other domains.



## 8 Acknowledgements

I would like to express my gratitude to all people, who were with me on my PhD research path.

First, I would like to thank to my supervisor Dunja Mladenić, who supported me and provided enormous help and advice.

I would also like to thank to Marko Grobelnik and Mitja Jermol, whose positive energy and enthusiasm surrounded me for four years and who gave me the opportunity to participate in compelling European FP7 projects.

My gratitude goes to the members of my doctoral committee, Dr. Marko Bohanec, Dr. Irena Nančovska Šerbec and Dr. Michael Witbrock, for their valuable comments and remarks.

Special thanks go to my research colleagues at Artificial Intelligence Laboratory at Jožef Stefan Institute. To Luka Bradeško, who always had time to answer my research questions and to check my research papers. To Blaž Novak, who provided me with a crawled set of business news – a required material for my thesis. I am thankful to Delia Rusu, Lorand Dali, Alexandra Moraru and Mario Karlovčec for their life and scientific advice and for the great office time spent together.

I would also like to thank to all other people surrounding me at Jožef Stefan Institute.

I am thankful to all my friends in different parts of the world for being with me even thousands kilometers apart.

Finally, I would like to express my deepest appreciation to my family, to the most important people in my life - my parents, who always supported and loved me wherever I was and whatever I did, and my husband, who inspired me every day and loved me “no matter what” all these years.



## 9 References

- Agirre, E.; Ansa, O.; Hovy, E.; Martínez, D. Enriching very large ontologies using the WWW. In: *Proceedings of ECAI 2000. Workshop on Ontology Learning* (2000).
- Andersen, P. M.; Hayes, P. J.; Weinstein, S. P.; Huettner, A.K.; Schmandt, L. M., and Nirenburg, I.B. Automatic Extraction of Facts from Press Releases to Generate News Stories. In: *Proceedings of ANLP*, 170-177 (1992).
- ASFA thesaurus, <http://www4.fao.org/asfa/asfa.htm> (accessed June 2010).
- Bontcheva, K.; Cunningham, H.; Kiryakov, A. and Tablan, V. Semantic Annotation and Human Language Technology. In: Davies, J.; Studer, R.; Warren, P. (eds.) *Semantic Web Technology: Trends and Research* (John Wiley and Sons Ltd , 2006).
- Bradeško, L.; Dali, L.; Fortuna, B.; Grobelnik, M.; Mladenić, D.; Novalija, I.; Pajntar, B. Contextualized question answering. In: *Proceedings of ITI 2010* (2010).
- Buitelaar, P.; Cimiano, P.; Magnini, B. (eds.) *Ontology Learning from Text: Methods, Evaluation and Applications* (IOS Press, 2005).
- Burkhardt, F.; Gulla, J. A.; Liu, J.; Weiss, C.; Zhou, J. Semi Automatic Ontology Engineering in Business Applications. In: *Proceedings of Workshop Applications of Semantic Technologies, INFORMATIK 2008* (2008).
- Carlson, A.; Betteridge, J.; Wang, R. C.; Hruschka, Jr. E. R.; Mitchell, T. M. Coupled Semi-Supervised Learning for Information Extraction. In: *Proceedings of the third ACM international conference on Web search and data mining*, 101-110 (2010).
- Central Intelligence Agency, The World Factbook: <https://www.cia.gov/library/publications/the-world-factbook> (accessed December 2010).
- Chandrasekaran, B.; Josephson, J. R.; Benjamins, R. V. What are Ontologies and why do we need them? In: *IEEE Intelligent Systems and Their Applications* **14**, 20-26 (1999).
- Chang, C. H.; Kaye, M.; Girgis, M. R.; Shaalan, K. A Survey of Web Information Extraction Systems. *Journal of IEEE Transaction on Knowledge and Data Engineering* **18/10**, 1411-1428 (2006).
- Cimiano, P.; Hotho, A.; Staab, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research (JAIR)* **24**, 305-339 (2005).
- Cimiano, P.; Pivk, A.; Schmidt-Thieme, L.; Staab, S. Learning Taxonomic Relations from Heterogeneous Evidence. In: *Proceedings of ECAI 2004, Workshop on Ontology Learning and Population* (2004).

- Cimiano, P.; Völker, J. Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery. In: *Proceedings of NLDB 2005*, 227-238 (2005).
- Contexts in Cyc, <http://www.cyc.com/cycdoc/course/contexts-basic-module.html> (accessed June 2011).
- Corcho, O.; Fernández-López, M.; Gómez-Pérez, A.; López-Cima, A. Building legal ontologies with METHONTOLOGY and WebODE (Springer-Verlag, LNAI, 2005).
- Cycorp, Inc., <http://www.cyc.com> (accessed July 2011).
- Cycorp, Inc., what's in Cyc, [http://www.cyc.com/cyc/technology/whatisincyc\\_dir/whatsincyc](http://www.cyc.com/cyc/technology/whatisincyc_dir/whatsincyc) (accessed July 2011).
- Dali, L.; Rusu, D.; Fortuna, B.; Mladenić, D.; Grobelnik, M. Question Answering Based on Semantic Graphs. In: *Proceedings of the WWW-2009 Workshop on Semantic Search (SemSearch2009)* (2009).
- Dellschaft, K.; Staab, S. Strategies for the Evaluation of Ontology Learning. In: Cimiano P. and Buitelaar P. (eds.) *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text* (IOS Press, 2008).
- Deshpande, M.; Karypis, G. Item-Based Top-N Recommendation Algorithms. *ACM Trans. Inf. Syst.* **22/1**, 143-177 (2004).
- Duong, W. N. Ghetto'ing Third World Workers With Hi-Tech : Industrial Application Of Artificial Intelligence And Its Effect On Foreign Direct Investment in the Third World -- Exploring Regulatory Solutions Through An Emblematic Case for the New Economy. *Temple International and Comparative Law Journal* **22/1** (2008).
- Etzioni, O.; Banko, M.; Soderland, S.; Weld, D. S. Open Information Extraction from the Web. *Communications of the ACM* **51/12** (2008).
- Financial glossary, <http://biz.yahoo.com/f/g/> (accessed June 2010).
- Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; Schlaefter, N.; Welty, C. Building Watson: an Overview of the DeepQA Project. *AI Magazine* **31/3** (2010).
- Fortuna, B.; Grobelnik, M.; Mladenić, D. OntoGen: Semi-automatic Ontology Editor. *HCI* **9**, 309-318 (2007).
- Ghani, R.; Jones, R.; Mladenić, D.; Nigam, K.; Slattery, S. Data Mining on Symbolic Knowledge Extracted from the Web. In: *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining, KDD-2000* (2000).
- Grobelnik, M.; Mladenić, D. Visualisation of News Articles. *Informatica journal* **28/4** (2004).
- Grobelnik, M.; Mladenić, D. Knowledge Discovery for Ontology Construction. In: Davies, J.; Studer, R.; Warren, P. (eds.) *Semantic Web Technologies: Trends and Research in Ontology-Based Systems* (John Wiley & Sons, 2006).

- Gruber, T. R. A translation approach to portable ontologies. *Knowledge Acquisition* **5/2**, 199-220 (1993).
- Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies* **43/5-6** (1995).
- Gruninger, M.; Fox, M. The role of competency in enterprise engineering. In: *Proceedings of the IFIP WG5.7 Workshop on Benchmarking - Theory and Practice*, IFIP (1994).
- Gunning, D.; Chaudhri, V.K.; Clark, P.; Barker, K.; Chaw, S.Y.; Greaves, M.; Grosz, B.; Leung, A.; McDonald, D.; Mishra, S.; Pacheco, J.; Porter, B.; Spaulding, A.; Tecuci, D.; Tien, J.. Project Halo Update - Progress toward Digital Aristotle. *AI Magazine* **31/3** (2010).
- Hahn, U.; Schnattinger, K. Towards Text Knowledge Engineering. In: *Proceedings of the AAAI'98* (1998).
- Harvey, C. R. Yahoo Financial Glossary (Fuqua School of Business, Duke University, 2003).
- Hearst, M. A. Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of COLING* (1992).
- Heyer, G.; Läuter, M.; Quasthoff, U.; Wittig, T.; Wolff, C. Learning Relations using Collocations. In: *Proceedings of IJCAI-2001, Workshop on Ontology Learning* (2001).
- Iacobelli, F.; Birnbaum, L.; Hammond, K. J. Tell me more, not just "more of the same". In: *Proceedings of the IUI 2010*, 81-90 (2010).
- Jarrar, M. Towards Methodological Principles for Ontology Engineering, *PhD thesis* (Vrije Universiteit Brussel, 2005).
- Lenat, D. B. Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* **38/11** (1995).
- Lenat, D. B.; Guha, R.V. Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project (Addison-Wesley, Boston, 1990).
- Lenat, D.; Witbrock, M.; Baxter, D.; Blackstone, E.; Deaton, C.; Schneider, D.; Scott, J.; Shepard, B. Harnessing CYC to Answer Clinical Researchers' Ad Hoc Queries. *AI Magazine* **31/ 3** (2010).
- Liu, W.; Weichselbraun, A.; Scharl, A.; Chang, E. Semi-Automatic Ontology Extension Using Spreading Activation. *Journal of Universal Knowledge Management* **1**, 50 – 58 (2005).
- Losch, U.; Nikitina, N. The newsEvents Ontology? An Ontology for Describing Business Events. In: *Proceedings of Workshop on Ontology Design Patterns, ISWC* (2009).
- Lytinen, S. L.; Gershman, A. ATRANS: Automatic Processing of Money Transfer Messages. In: *Proceedings of AAAI*, 1089-1095 (1986).

- Maedche, A.; Staab, S. Discovering conceptual relations from text. In: Horn W. (ed.) *ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence* (Berlin, August 21-25, 2000; IOS Press, Amsterdam, 2000).
- Maedche, A.; Staab, S. Ontology learning for the Semantic Web. *IEEE Intelligent Systems* **16/2**, 72–79 (2001).
- Maedche, A.; Staab, S. Measuring similarity between ontologies. In: *Proceedings Of the European Conference on Knowledge Acquisition and Management - EKAW-2002* (Madrid, Spain, 2002).
- Maynard, D.; Funk, A.; Peters, W. SPRAT: a tool for automatic semantic pattern-based ontology population. In: *Proceedings of International Conference for Digital Libraries and the Semantic Web* (Trento, Italy, 2009).
- Martínez Montes, M.; Bas, J.; Bellido, S.; Corcho, O.; Losada, S.; Benjamins, R.; Contreras, J. WP10: Case study eBanking D10.3 Financial Ontology. *DIP - Data, Information and Process Integration with Semantic Web Services* (2005).
- McDonald, J.; Plate, T.; Schvaneveldt, R. Using pathfinder to extract semantic information from text. In: *Schvaneveldt*, 149–164 (1990).
- Mechanical Turk, <https://www.mturk.com/mturk/welcome> (accessed September, 2011).
- Medelyan, O.; Legg, C. Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. In: *Proceedings of Wiki-AI Workshop at the AAAI'08 Conference* (Chicago, US, 2008).
- Novalija, I.; Mladenović, D. Extending Ontologies for Annotating Business News. In: *Proceedings of siKDD 2008* (Ljubljana, Slovenia, 2008).
- Noy, N. F.; Hafner, C. The state of the art in ontology design: a survey and comparative review. *Artificial Intelligence Magazine*, **18**, 53-74 (1997).
- OpenCalais, <http://www.opencalais.com> (accessed June 2011).
- OpenCalais - English Semantic Metadata: Entity/Fact/Event Definitions and Descriptions, <http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-index-and-definitions> (accessed June 2011).
- OpenCyc, <http://www.opencyc.org/> (accessed September 2011).
- Panton, K.; Matuszek, C.; Lenat, D.; Schneider, D.; Witbrock, M.; Siegel, N.; Shepard, B. Common Sense Reasoning – From Cyc to Intelligent Assistant. In: Cai, Y. and Abascal, J. (eds.) *Ambient Intelligence in Everyday Life*, LNAI 3864, 1-31 (Springer, 2006).
- Prieto-Díaz, R. A faceted approach to building ontologies. In: *Proceedings of the 21st International Conference on Conceptual Modeling* (2002).
- Reinberger, M.-L.; Spyns, P. Unsupervised Text Mining for the Learning of DOGMA-

- Inspired Ontologies. In: Buitelaar, P.; Handschuh, S.; Magnini, B. (eds.) *Ontology Learning from Text: Methods, Evaluation and Applications* (IOS Press, 2005).
- Sabrina, T.; Rosni, A.; Enyakong, T. Extending Ontology Tree Using NLP Technique. In: *Proceedings of National Conference on Research & Development in Computer Science REDECS 2001* (2001).
- Sarjant, S.; Legg, C.; Robinson, M.; Medelyan, O. "All You Can Eat" Ontology-Building: Feeding Wikipedia to Cyc. In: *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI'09* (Milan, Italy, 2009).
- Semantic Technology and Linked Data Annotation,  
<http://gate.ac.uk/wiki/TrainingCourseMay2011/4th-gate-training.pdf> (accessed August 2011).
- Semantic Web. Ontology, <http://semanticweb.org/wiki/Ontology> (accessed August 2011).
- Shah, P.; Schneider, D.; Matuszek, C.; Kahlert, R. C.; Aldag, B.; Baxter, D.; Cabral, J.; Witbrock, M.; Curtis, J. Automated population of Cyc: Extracting information about named-entities from the web. In: *Proceedings of the Nineteenth International FLAIRS Conference*, 153-158 (2006).
- Soderland, S.; Roof, B.; Qin, B.; Xu, S.; Mausam; Etzioni, O. Adaptation Information Extraction to Domain-Specific Relations. *AI Magazine* (Fall 2010).
- Štajner, T.; Mladenić, D. Entity Resolution in Texts Using Statistical Learning and Ontologies. In: *Proceedings of Asian Semantic Web Conference* (2009).
- Suchanek, F. M.; Sozio, M.; Weikum, G. SOFIE: A Self-Organizing Framework for Information Extraction. In: *Proceedings of the 18th World Wide Web Conference, www2009* (Madrid, Spain, 2009).
- Sure, Y.; Studer, R. On-To-Knowledge Methodology - Final Version. *On-To-Knowledge deliverable D-18* (Institute AIFB, University of Karlsruhe, 2002).
- The Syntax of CycL, <http://www.cyc.com/cycdoc/ref/cycl-syntax.html> (accessed June 2011).
- Taylor, M. E.; Matuszek, C.; Klimt, B.; Witbrock, M. Autonomous Classification of Knowledge into an Ontology. In: *Proceedings of the Twentieth International FLAIRS Conference* (Key West, FL, 2007).
- Text-Garden -- Text-Mining Software Tools, <http://ailab.ijs.si/dunja/textgarden> (accessed June 2011).
- Tunstall-Pedoe, W. True Knowledge: Open-Domain Question Answering Using Structured Knowledge and Inference. *AI Magazine* **31/3**, 80-92 (2010).
- Turney, P. D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: *Proceedings of the Twelfth European Conference on Machine Learning* (2001).
- Uschold, M.; King, M. Towards a methodology for building ontologies. In: *Proceedings of Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95* (Canada, 1995).

W3C, <http://www.w3.org/2001/sw> (accessed August 2011).

Wikipedia, <http://en.wikipedia.org/wiki> (accessed June 2011).

Witbrock, M.; Baxter, D.; Curtis, J.; Schneider, D.; Kahlert, R.; Miraglia, P.; Wagner, P.; Panton, K.; Matthews, G.; Vizedom, A. An Interactive Dialogue System for Knowledge Acquisition in Cyc. In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* (2003).

WordNet – Princeton University Cognitive Science Laboratory, <http://wordnet.princeton.edu> (accessed June 2011).

Yahoo! Finance, <http://finance.yahoo.com> (accessed June 2010).

Yahoo! Finance news, <http://biz.yahoo.com/top.html> (accessed June 2010).

Zhang, Z.; Zhang, C. and San Ong, S. Building an Ontology for Financial Investment. In: *Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents*. Lecture Notes In Computer Science **198** (2000).



## Index of Figures

Figure 1: <i>Text-Driven Ontology Extension (<b>OntoPlus</b> Methodology).</i> .....	19
Figure 2: <i>Illustrative Extraction of Business &amp; Finances Domain Subset from a Multi-Domain Ontology. Dark circles represent the extracted relevant concepts.</i> .....	21
Figure 3: <i>Text-Driven Ontology Extension (Cyc KB Adaptation).</i> .....	27
Figure 4: <i>System Pipeline (Analyzing Business News Using OpenCalais, OntoPlus and Cyc)</i> .....	29
Figure 5: <i>Example Financial Glossary Entries.</i> .....	40
Figure 6: <i>ASFA Thesaurus Transformation.</i> .....	41
Figure 7: <i>Performance of the content, structure and co-occurrence weighting measures (Financial glossary). Precision (P) of concept ranking (Top 1).</i> .....	46
Figure 8: <i>Performance of the content, structure and co-occurrence weighting measures (ASFA thesaurus). Precision (P) of concept ranking (Top 1).</i> .....	47
Figure 9: <i>Hit Rate (HR) depending on <math>\beta</math> (Equivalent and hierarchically related concepts).</i> .....	49
Figure 10: <i>Number of Concepts (NC) depending on <math>\beta</math>.</i> .....	50
Figure 11: <i>Learning Accuracy (LA) depending on <math>\beta</math> and <math>\gamma</math> (Financial glossary).</i> .....	52
Figure 12: <i>Learning Accuracy (LA) depending on <math>\beta</math> and <math>\gamma</math> (ASFA thesaurus).</i> .....	52
Figure 13: <i>Number of Proposed Relationships (NR) depending on <math>\beta</math> and <math>\gamma</math> (Financial glossary).</i> .....	53
Figure 14: <i>Number of Proposed Relationships (NR) depending on <math>\beta</math> and <math>\gamma</math> (ASFA thesaurus).</i> .....	54
Figure 15: <i>Cyc KB Extension - User Validation.</i> .....	57
Figure 16: <i>Business Cycle Definition.</i> .....	59
Figure 17: <i>Economists in companies from financial sector: <b>Mingchun Sun</b>.</i> .....	66
Figure 18: <i>Get companies, which were involved into layoff activity from Georgia-State (USA): <b>Delta AirLines</b>.</i> .....	66
Figure 19: <i><b>OntoPlus</b> software demonstration. Choosing glossary file.</i> .....	67
Figure 20: <i><b>OntoPlus</b> software demonstration. Domain keywords.</i> .....	68
Figure 21: <i><b>OntoPlus</b> software demonstration. Content, structure, co-occurrence weights.</i> .....	68
Figure 22: <i><b>OntoPlus</b> software demonstration. Results – financial domain example.</i> .....	69
Figure 23: <i><b>OntoPlus</b> software demonstration. Results – fisheries &amp; aquaculture domain example.</i> .....	70



## Index of Tables

Table 1: <i>Mapping Support.</i> .....	30
Table 2: <i>Mapping OpenCalais Fact/Event Merger → Cyc Fact/Event Merger.</i> .....	31
Table 3: <i>Experimental Queries.</i> .....	43
Table 4: <i>Evaluation of the top suggested candidate concepts for ontology extension (Financial glossary).</i> .....	48
Table 5: <i>Evaluation of the top suggested candidate concepts for ontology extension (ASFA thesaurus).</i> .....	48
Table 6: <i>Evaluation of the equivalent, hierarchical (subclass) relations identification.</i> .....	51
Table 7: <i>Evaluation of the top suggested equivalent &amp; hierarchical (subclass) relations.</i> .....	51
Table 8: <i>Examples of Cyc KB extension (Financial glossary).</i> .....	55
Table 9: <i>Examples of Cyc KB extension (ASFA thesaurus).</i> .....	56
Table 10: <i>Related Cyc Concepts for Glossary Term “Business Cycle”.</i> .....	59
Table 11: <i>GDP Growth Rates in Egypt and Indonesia.</i> .....	62
Table 12: <i>Extracted Entities by OpenCalais.</i> .....	62
Table 13: <i>Extracted Facts/Events by OpenCalais.</i> .....	63
Table 14: <i>Precision of Ontology Population.</i> .....	64
Table 15: <i>Evaluation of News Based Question Answering.</i> .....	65



## Index of Algorithms

Algorithm 1: <i>Mapping OpenCalais Generic Relations to Cyc KB.</i> .....	32
---	----



## Appendix 1: OntoPlus Applications in Financial Domain

Appendix 1 shows the application of the **OntoPlus** methodology in the financial domain. For each financial glossary term (glossary term name – glossary term description), the most similar related concepts from the Cyc ontology (similarity; the Cyc concept denotation – the Cyc concept comment) are provided. In addition, for each financial term, a list of potential equivalent, hierarchical and associative relationships with the Cyc ontology concepts is shown below.

**Glossary Term: CREDIT\_UNION - A not-for-profit institution that is operated as a cooperative and offers financial services such as low-interest loans, to its members.**

Similarity: 0.41030743790527824 CreditUnion - Each instance of #CreditUnion is a financial cooperative #Organization of individuals with a common affiliation (such as employment, labor union membership, or place of residence). Credit unions accept deposits of members, pay interest (or dividends) on them out of earnings, and primarily provide consumer installment credit to members.

Similarity: 0.235638122037419 SaleByCreditCardWithLowAPR - A specialization of #SaleByCreditCard. Each instance of #SaleByCreditCardWithLowAPR is a purchase in which the buyer (see the predicate #buyer) pays by using some instance of #CreditCardWithLowAPR (q.v.).

Similarity: 0.21863625313983168 CooperativeOrganization - A cooperative is an organization in which each of its members contribute to the organization's capital. In return the members get dividends and the right to vote in the organization's elections (for board of directors).

Similarity: 0.2030559187186898 TradeUnion - An organization of workers formed to mutually benefit all of its members. If an employer establishes itself as a union shop, then all of its employees must be members of that particular trade union.

Similarity: 0.15709656044375866 CooperationEvent - A collection of all instances of #SocialOccurrence that can be called 'cooperations'. For each #CooperationEvent COOP the following hold: (i) there are at least two #IntelligentAgents AGT1 and AGT2 such that (#partnersInCooperation AGT1 AGT2 COOP), (ii) AGT1 and AGT2 share a GOAL, (iii) there are subevents ACT1 and ACT2 of COOP, both of which are #PurposefulActions, such that ACT1 is performed by AGT1 with the intent of furthering GOAL and ACT2 is performed by AGT2 with the intent of furthering GOAL, (iv) AGT1 believes that ACT2 furthers GOAL and AGT2 believes that ACT1 furthers GOAL, and (v) the fact ACT1 and ACT2 both further GOAL is not an unexpected coincidence: it was a part of AGT1's expectation that an act like (i.e. of the same kind as) ACT2, performed by AGT2, would further GOAL; and it was a part of AGT2's expectation that an act like ACT1, performed by AGT1, would further GOAL.

Similarity: 0.14808573359998406 FinancialOrganization -  
 #FinancialOrganization is a specialization of #Organization. Each  
 instance of #FinancialOrganization is primarily or significantly  
 engaged in the #FinancialIndustry or whose activities focus on that  
 industry. Instances of both #CommercialServiceOrganizations (e.g.,  
 banks and brokerage houses) and #NonProfitOrganizations (e.g.,  
 #InternationalMonetaryFund) may be instances of  
 #FinancialOrganization. Specializations of #FinancialOrganization  
 include #BankOrganization, #FinancialExchange, and  
 #InvestmentOrganization.

Similarity: 0.14156338878911331 Institution - A sub-collection of  
 #Organization. Each instance is an organization founded and united for  
 a specific purpose.

Similarity: 0.13429055303235912 FederalUnion - The collection of all  
 federal unions. A type of #GovernmentRelatedEntity and  
 #GeopoliticalEntity. The collection #FederalUnion is an  
 #ArtifactualFeatureType and a #FacetInstanceCollection.

Similarity: 0.13078561183986503 Operation - A specialization of  
 #MedicalCareEvent. Each instance of #Operation is a medical care  
 event in which a medical professional cuts a part of a living body,  
 either to examine what's inside (a diagnostic, exploratory surgery) or  
 to treat an ailment (an instance of #MedicalTreatmentEvent).  
 Generally, an operation will have one or more proper sub-events that is  
 an instance of #Surgery (or, specifically, #SurgicalProcedure (q.v.)).

Similarity: 0.12617666645287612 OfferingForSale - The collection of  
 OfferingForSale events includes events in which an agent offers one or  
 more things for sale to one or more other agents.

-----

Suggested relationship: CreditUnion might be EQUIVALENT to: CreditUnion

Suggested relationship: CreditUnion might be SUBCLASS with difference in  
 similarities 0.7906575760833375 to FinancialOrganization

Suggested relationship: CreditUnion might be SUBCLASS with difference in  
 similarities 0.6414717954484694 to FederalUnion

Suggested relationship: CreditUnion might be SUBCLASS with difference in  
 similarities 0.5099274027331723 to Operation

Suggested relationship: CreditUnion might be SUBCLASS with difference in  
 similarities 0.498593683386054 to TradeUnion

Suggested relationship: CreditUnion might be SUBCLASS with difference in  
 similarities 0.4822113409875642 to OfferingForSale

Suggested relationship: CreditUnion might be SUBCLASS with difference in  
 similarities 0.32450600883292235 to CooperativeOrganization

Suggested relationship: CreditUnion might be CONCEPTUALLY RELATED to  
 SaleByCreditCardWithLowAPR

Suggested relationship: CreditUnion might be CONCEPTUALLY RELATED to  
 CooperationEvent

Suggested relationship: CreditUnion might be CONCEPTUALLY RELATED to  
 Institution



**Glossary Term: CERTIFICATE\_OF\_DEPOSIT** - Also called a time deposit this is a certificate issued by a bank or thrift that indicates a specified sum of money has been deposited. A CD has a maturity date and a specified interest rate, and can be issued in any denomination. The duration can be up to five years.

Similarity: 0.446691676348582 CertificateOfDeposit - The collection of receipts for bank (or other financial institution) deposits in certificate form. #CertificateOfDeposits bear interest and are payable either on a specific date or after the passage of a specific amount of time.

Similarity: 0.22881086464541786 MakingABankDeposit - A collection of events; a subset of #MoneyTransfer. In an instance of #MakingABankDeposit, an #Agent-PartiallyTangible (or representative thereof) deposits currency, checks, or other financial tender into a bank account (usually the #Agent-PartiallyTangible's personal bank account, or the bank account of the company for which they work). An agent may make a deposit in person at the bank, but it might also be made by mail, over the telephone, electronically via modem, at an ATM machine, etc.

Similarity: 0.22208788198712837 Certificate-IBT - The collection of actual hardcopy (#CertificateDocument) or electronic (#DigitalCertificate) instances of a certificate. Examples include instances of #Passport and #BirthCertificate.

-----  
Suggested relationship: CertificateOfDeposit might be EQUIVALENT to: CertificateOfDeposit

Suggested relationship: CertificateOfDeposit might be SUBCLASS with difference in similarities 0.5160952892511015 to MakingABankDeposit

Suggested relationship: CertificateOfDeposit might be SUBCLASS with difference in similarities 0.3852179136529344 to Certificate-IBT

**Glossary Term: DEALER\_LOAN** - Overnight, collateralized loan from a money market bank made to a dealer financing his position by borrowing.

Similarity: 0.3892853352816869 FinancingByBorrowing - The collection of #BorrowingSomething events in which money is borrowed. Examples include borrowing money from a bank under an authorized #LoanAgreement (a #Loan-WrittenLegalAgreement) and borrowing money from a friend under a #Loan-ByInformalOralAgreement. Specializations of #FinancingByBorrowing include #BridgeFinancing and #EnteringIntoAMortgageAgreement.

Similarity: 0.1616247547044806 LoanAgreement - A collection of #TemporaryUserRightsAgreements. In each #Agreement, one party will give something of value (usually money) to a second party, and the second party will pay the money back, often paying interest as well. Among other options, the parties can agree that there will be repeating #Paying events, with the same amount paid each time; or they may agree that all the money will be paid back all at once.

Similarity: 0.1493517014071036 BorrowingSomething - A collection of events; a subcollection of #\$TemporaryChangeOfUserRights. In an instance of #\$BorrowingSomething, an agent takes temporary control of something, usually with the permission of its owner(s). Generally, the lending agent expects the borrowing agent to use the object for one of its normal functions (see #\$intendedBehaviorCapable).

-----  
Suggested relationship: DealerLoan might be SUBCLASS with difference in similarities 0.7407791027576881 to BorrowingSomething  
Suggested relationship: DealerLoan might be SUBCLASS with difference in similarities 0.5195975775666641 to LoanAgreement  
Suggested relationship: DealerLoan might be CONCEPTUALLY RELATED to FinancingByBorrowing

## Appendix 2: OntoPlus Applications in Fisheries & Aquaculture Domain

Appendix 2 shows the application of the **OntoPlus** methodology in the fisheries & aquaculture domain. For each fisheries & aquaculture glossary term (glossary term name – glossary term description), the most similar related concepts from the Cyc ontology (similarity; the Cyc concept denotation – the Cyc concept comment) are provided. In addition, for each fisheries & aquaculture term, a list of potential equivalent, hierarchical and associative relationships with the Cyc ontology concepts is shown below.

### **Glossary Term: ALLOZYMES - Enzymes**

Similarity: 0.11431397786957938 EnzymeMolecule - This is the collection of (individual) enzyme molecules, which are globular protein molecules that can act as biological catalysts in a broad spectrum of biochemical reactions.

-----

Suggested relationship: Allozymes might be SUBCLASS with difference in similarities 0.5030667769204702 to EnzymeMolecule

### **Concept: ZINC - Heavy metals Ferromanganese nodules Metalliferous sediments Zinc compounds Zinc isotopes**

Similarity: 0.716015620694725 Zinc - A piece (i.e., specific collection of nearby molecules) of #Zinc.

Similarity: 0.23687802675660904 Metal - An instance of #TangibleStuffCompositionType. Every instance of #Metal is a tangible object having certain characteristic physical and chemical properties. Instances of #Metal are good conductors of electricity and heat, and most instances of #Metal are solids at room temperature (although all instances of #Mercury, for example, are liquids at room temperature). Solid instances of #Metal have a shiny luster, and are highly malleable. Specializations of #Metal include #Mercury, #Potassium, #Brass, #Lead, and #Iron.

-----

Suggested relationship: Zinc might be EQUIVALENT to Zinc

Suggested relationship: Zinc might be SUBCLASS with difference in similarities 0.694670148702147 to Metal

### **Concept: CHLOROPLASTS - Cells Chlorophylls Chromatophores Photosynthetic pigments**

Similarity: 0.5913888161282679 Chloroplast - A green plastid with an internal membrane system incorporating the pigment molecules that are essential to photosynthesis.

Similarity: 0.3900282501626628 Pigment - instances are pieces of a chemical compound that impart color.

-----  
Suggested relationship: Chloroplasts might be EQUIVALENT to Chloroplast  
Suggested relationship: Chloroplasts might be CONCEPTUALLY RELATED to  
Pigment.

## Appendix 3: OpenCalais to Cyc Mappings

Appendix 3 demonstrates mappings between OpenCalais entity, fact and event types and the Cyc ontology elements (concepts, predicates etc.). For instance, for each OpenCalais entity type synonymous collection from the Cyc ontology is provided. As well, for each OpenCalais relation type, a number of assertions connecting synonymous concepts from the Cyc ontology are provided.

### Entity Types Examples:

#### City

```
Constant: City.
In Mt: CalaisOntologyMappingMt.
isa: CalaisOntologyEntityType.
In Mt: CalaisOntologyMappingMt.
f: (synonymousExternalConcept City TheCalaisOntology "City").
```

#### Industry Term

```
Constant: IndustryType.
In Mt: CalaisOntologyMappingMt.
isa: CalaisOntologyEntityType.
In Mt: CalaisOntologyMappingMt.
f:      (synonymousExternalConcept      IndustryType      TheCalaisOntology
"IndustryTerm").
```

#### Political Event

```
Constant: PoliticalEvent.
In Mt: CalaisOntologyMappingMt.
isa: CalaisOntologyEntityType.
In Mt: CalaisOntologyMappingMt.
f:      (synonymousExternalConcept      PoliticalEvent      TheCalaisOntology
"PoliticalEvent").
```

#### Position of Person in the Organization

```
Constant: PersonTypeByPositionInOrg.
In Mt: CalaisOntologyMappingMt.
isa: CalaisOntologyEntityType.
In Mt: CalaisOntologyMappingMt.
f:      (synonymousExternalConcept      PersonTypeByPositionInOrg
TheCalaisOntology "Position").
```

**Geographical Region**

```

Constant: GeographicalRegion.
In Mt: CalaisOntologyMappingMt.
isa: CalaisOntologyEntityType.
In Mt: CalaisOntologyMappingMt.
f: (synonymousExternalConcept    GeographicalRegion    TheCalaisOntology
    "Region").

```

**Event/Fact Types Examples:****Bonus Shares Issuance**

```

In Mt: OpenCalaisFactExtractionMt.
Direction: :forward.
f: (implies
    (and
        (extractedOpenCalaisFactEvent    ?FACTEVENT    (OpenCalaisFactEventFn
"BonusSharesIssuance"))
        (extractedOpenCalaisAttribute    (OpenCalaisAttributeFn "company" ?C)
?FACTEVENT))
    (and
        (isa ?FACTEVENT    IssuingASecurity)
        (isa ?C Business)
        (isa ?C LegalCorporation)
        (actors ?FACTEVENT ?C))).

```

**Business Relation**

```

In Mt: OpenCalaisFactExtractionMt.
Direction: :forward.
f: (implies
    (and
        (extractedOpenCalaisFactEvent    ?FACTEVENT    (OpenCalaisFactEventFn
"BusinessRelation"))
        (extractedOpenCalaisAttribute    (OpenCalaisAttributeFn "company" ?C1)
?FACTEVENT)
        (extractedOpenCalaisAttribute    (OpenCalaisAttributeFn "company" ?C2)
?FACTEVENT)
        (not (equals ?C1 ?C2)))
    (and
        (isa ?FACTEVENT    BusinessEvent)
        (isa ?C1 Business)
        (isa ?C1 LegalCorporation)
        (isa ?C2 Business)
        (isa ?C2 LegalCorporation)
        (actors ?FACTEVENT ?C1)
        (actors ?FACTEVENT ?C2)
        (businessPartners ?C1 ?C2))).

```

**Company Affiliates**

In Mt: OpenCalaisFactExtractionMt.

Direction: :forward.

```
f: (implies
  (and
    (extractedOpenCalaisFactEvent ?FACTEVENT (OpenCalaisFactEventFn
      "CompanyAffiliates"))
    (extractedOpenCalaisAttribute (OpenCalaisAttributeFn
      "company_affiliate" ?C1) ?FACTEVENT)
    (extractedOpenCalaisAttribute (OpenCalaisAttributeFn
      "company_parent" ?C2) ?FACTEVENT))
  (and
    (isa ?FACTEVENT BusinessEvent)
    (isa ?C1 Business)
    (isa ?C1 LegalCorporation)
    (isa ?C2 Business)
    (isa ?C2 LegalCorporation)
    (actors ?FACTEVENT ?C1)
    (actors ?FACTEVENT ?C2)
    (affiliatedWith ?C1 ?C2))).
```

**Company Competitor**

In Mt: OpenCalaisFactExtractionMt.

Direction: :forward.

```
f: (implies
  (and
    (extractedOpenCalaisFactEvent ?FACTEVENT (OpenCalaisFactEventFn
      "CompanyCompetitor"))
    (extractedOpenCalaisAttribute (OpenCalaisAttributeFn "company1"
      ?C1) ?FACTEVENT)
    (extractedOpenCalaisAttribute (OpenCalaisAttributeFn "company2"
      ?C2) ?FACTEVENT))
  (and
    (isa ?FACTEVENT BusinessEvent)
    (isa ?C1 Business)
    (isa ?C1 LegalCorporation)
    (isa ?C2 Business)
    (isa ?C2 LegalCorporation)
    (actors ?FACTEVENT ?C1)
    (actors ?FACTEVENT ?C2)
    (competitors ?C1 ?C2))).
```

**Company Customer**

In Mt: OpenCalaisFactExtractionMt.

Direction: :forward.

```
f: (implies
```

```

    (and
      (extractedOpenCalaisFactEvent      ?FACTEVENT      (OpenCalaisFactEventFn
"CompanyCustomer"))
      (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn
"company_provider" ?C1) ?FACTEVENT)
      (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn
"organization_customer" ?O) ?FACTEVENT)))
    (and
      (isa ?FACTEVENT  Buying)
      (isa ?C1 Business)
      (isa ?C1 LegalCorporation)
      (isa ?O Organization)
      (actors ?FACTEVENT ?C1)
      (actors ?FACTEVENT ?O)
      (customers ?C1 ?O))).

```

### Company Founded

In Mt: OpenCalaisFactExtractionMt.

Direction: :forward.

```

f: (implies
  (and
    (extractedOpenCalaisFactEvent      ?FACTEVENT      (OpenCalaisFactEventFn
"CompanyFounded"))
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn "company" ?C)
?FACTEVENT)
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn "year" ?Y)
?FACTEVENT))
    (and
      (isa ?FACTEVENT  Event)
      (isa ?C Business)
      (isa ?C LegalCorporation)
      (foundingDate ?C (DateFromStringFn ?Y))
      (actors ?FACTEVENT ?C))).

```

### Company Product

In Mt: OpenCalaisFactExtractionMt.

Direction: :forward.

```

f: (implies
  (and
    (extractedOpenCalaisFactEvent      ?FACTEVENT      (OpenCalaisFactEventFn
"CompanyProduct"))
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn "company" ?C)
?FACTEVENT)
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn "product" ?PR)
?FACTEVENT))
    (and
      (isa ?C Business)

```



```

(isa ?C LegalCorporation)
(isa ?PR ProductTypeByBrand)
(makesProductType ?C ?PR)).

```

### Employment Relation

In Mt: OpenCalaisFactExtractionMt.

Direction: :forward.

```

f: (implies
  (and
    (extractedOpenCalaisFactEvent      ?FACTEVENT      (OpenCalaisFactEventFn
"EmploymentRelation"))
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn
"person_employer" ?PE) ?FACTEVENT)
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn
"person_employee" ?PEML) ?FACTEVENT))
  (and
    (isa ?PE Person)
    (isa ?PEML Person)
    (employer ?PEML ?PE))).

```

### Person Career

In Mt: OpenCalaisFactExtractionMt.

Direction: :forward.

```

f: (implies
  (and
    (extractedOpenCalaisFactEvent      ?FACTEVENT      (OpenCalaisFactEventFn
"PersonCareer"))
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn "person" ?P)
?FACTEVENT)
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn "position"
?POS) ?FACTEVENT)
    (extractedOpenCalaisAttribute      (OpenCalaisAttributeFn "company" ?C)
?FACTEVENT))
  (and
    (isa ?P Person)
    (isa ?C Business)
    (isa ?C LegalCorporation)
    (isa ?POS PersonTypeByPositionInOrg)
    (positionOfPersonInOrganization ?P ?C ?POS))).

```

### Product Release

In Mt: OpenCalaisFactExtractionMt.

Direction: :forward.

```

f: (implies
  (and
    (extractedOpenCalaisFactEvent      ?FACTEVENT      (OpenCalaisFactEventFn
"ProductRelease"))

```

```

    (extractedOpenCalaisAttribute (OpenCalaisAttributeFn "product" ?P)
?FACTEVENT)
    (extractedOpenCalaisAttribute (OpenCalaisAttributeFn "company" ?C)
?FACTEVENT)
    (extractedOpenCalaisAttribute (OpenCalaisAttributeFn "datestring"
?D) ?FACTEVENT))
    (and
    (isa ?FACTEVENT Event)
    (isa ?C Business)
    (isa ?C LegalCorporation)
    (isa ?P ProductTypeByBrand)
    (makesProductType ?C ?P)
    (performedBy ?FACTEVENT ?C)
    (dateOfProductRelease ?P (DateFromStringFn ?D))))).

```

## Appendix 4: Example of Financial News Analysis

Appendix 4 contains an example of particular financial news analysis. In the process of financial information analysis, the Cyc ontology is extended with relevant financial terms occurring in news (for instance, such as *Emerging Market*). Furthermore, OpenCalais is used for entity, event and fact extraction from text and Cyc is populated with new concept and relation instances obtained from financial news. Appendix 4 also provides answers to a possible query related to the information from the specified financial news.

### News Content:

Oil sets new record near \$121 a barrel: Financial News

Supply disruptions in Nigeria, where a strike and attacks by militants has hit production, has supported a market that is nervous about any threats to supply.

Tensions with Iran ratcheted higher when the world's fourth-biggest oil producer refused to accept intrusive inspections of its nuclear program that the West fears could be linked to weapons.

U.S. light crude for June delivery was up 7 cents at \$120.04 a barrel, by 7:55 a.m. EDT after earlier touching a record high of \$120.93.

London Brent crude was up 33 cents at \$118.32 a barrel, after an earlier record of \$119.07.

Gold was also strong, as oil's advance helped spur a rebound from a four-month low last week. But gold is still some way below a record of \$1,030.80 an ounce reached on March 17.

"The downward move in oil last week now seems like only a correction," said Christopher Bellew, senior vice president at Bache Commodities.

"The effect of the credit crisis in the United States is reducing people's disposable incomes and you'd expect this to have an impact on the oil price, but it's not having any impact."

Demand from emerging markets such as India and China is more than compensating for the U.S. downturn, he said.

Goldman Sachs predicted oil could soar towards \$150-\$200 a barrel because of a lack of adequate supply growth.

"The possibility of \$150-\$200 per barrel seems increasingly likely over the next 6-24 months, though predicting the ultimate peak in oil prices as well as the remaining duration of the upcycle remains a major uncertainty," the bank said.

The U.S. investment bank had predicted back in 2005 that oil was entering a "super-spike" period.

Oil prices further into the future have also risen sharply, with prices out to 2016 above \$110 a barrel.

#### VULNERABLE

Oil has nearly doubled in the past year and is up by a quarter since the start of 2008 partly due to the problems in Nigeria, plus weakness in the U.S. dollar, which has boosted the price of commodities denominated in the U.S. currency.

Last week, oil retreated almost \$10 a barrel, partly due to a reduction in speculative positions and as strikes affecting Nigeria and the North Sea came to an end.

Exxon Mobil (NYSE:XOM - News) said on Tuesday it had returned oil output in Nigeria to normal levels after an eight-day strike, but Shell (LSE:RDSA.L - News) said its production there was still down by about 164,000 barrels a day due to recent militant attacks.

"A lot of this is supply-driven, with the market very vulnerable to any disruption in supplies," said Mark Pervan, a senior commodities analyst at the Australian & New Zealand Bank.

"We're seeing large oil-producing countries coming up as a question mark," he said.

U.S. President George W. Bush is expected to talk with officials from Saudi Arabia about the effects of high fuel prices on the U.S. economy on his trip to the world's top exporter later this month.

Bush has called on the Organization of the Petroleum Exporting Countries to raise output to help bring down prices.

The U.S. dollar, whose decline in the past months has been driving speculative investments in dollar-denominated crude and other commodities, was weaker versus the euro on Tuesday on continued doubts about the health of the U.S. economy despite upside surprises from recent economic indicators.

Later in the week on Wednesday, traders will watch the weekly U.S. government report on fuel inventories, which is expected to show a 1.8 million-barrel build in crude stocks, a 1.1 million-barrel rise in distillate inventories and a 100,000-barrel fall in gasoline stocks.

(Additional reporting by Baizhen Chua in Singapore; editing by James Jukwey)

(1 row)

#### **Example of Cyc extension with financial terms from news:**

**Glossary term occurred in news:** Emerging market

**Glossary term comment:** Countries in the process of building market-based economies are broadly referred to as emerging markets.

Related to Cyc Concepts:

Similarity: 0.15292962681763106 Market

Similarity: 0.14681682020166828 ProductPlanningAndDevelopment

Similarity: 0.14625050469620562 Country

EmergingMarket is SUBCLASS of Market

#### **OpenCalais Entities, Events and Facts Extracted:**

##### **Entities:**

###### **Company:**

Australian & New Zealand Bank  
Bache Commodities Limited  
Exxon Mobil Corporation  
THE GOLDMAN SACHS GROUP, INC.  
The Shell Transport & Trading

###### **Country:**

China  
India  
Iran  
Nigeria  
Saudi Arabia  
Singapore

##### **Events & Facts:**

###### **Company Ticker:**

Exxon Mobil Corporation, XOM, NYSE  
The Shell Transport & Trading Company  
Plc, RDSA.L, LSE

###### **Person Career:**

George W. Bush, President, United States, political, current  
Christopher Bellew, senior vice president, Bache Commodities Limited, professional, current  
Mark Pervan, senior commodities analyst, Australian & New Zealand Bank,

United States

professional, current

**Currency:**

cent

USD

**Industry Term:**

bank

investment bank

large oil-producing countries

oil

oil last week

oil price

oil prices

oil producer

**Natural Feature:**

North Sea

**Generic Relations:**

it, oil, return

George W. Bush, talk

oil producer, intrusive inspections of its nuclear program, accept

We, large oil-producing countries, see George W. Bush, the Organization of the Petroleum Exporting Countries to raise

output, call on

oil, retreat

oil, soar

THE GOLDMAN SACHS GROUP, INC., oil, predict

investment bank, predict

oil, enter

large oil-producing countries, come

**Organization:**

Organization of Petroleum-Exporting Countries

U.S. government

**Quotation:**

Mark Pervan, We're seeing large oil-producing countries coming up as a question mark

George W. Bush, the Organization of the Petroleum Exporting Countries to raise output to help bring down prices

**Person:**

Christopher Bellew

George W. Bush

James Jukwey

Mark Pervan

Christopher Bellew, senior vice president, Bache Commodities, The downward move in oil last week now seems like only a correction.

Christopher Bellew, Demand from emerging markets such as India and China is more than compensating for the U.S. downturn

Mark Pervan, senior commodities analyst, Australian & New Zealand Bank, A lot of this is supply-driven, with the market very vulnerable to any disruption in supplies

**Position:**

oil producer

President

senior commodities analyst

senior vice president

**Examples of Cyc population by mapping entities, events and facts to Cyc using OntoPlus:**

"Bache Commodities Limited" is INSTANCE of Cyc concept **Business**

"Bache Commodities Limited" is INSTANCE of Cyc concept **LegalCorporation**

"United States" is EQUIVALENT to Cyc concept **UnitedStatesOfAmerica**

"cent" is EQUIVALENT to Cyc concept **Cent-UnitedStates**

"U.S. government" is EQUIVALENT to Cyc concept **#UnitedStatesFederalGovernment**

"Mark Pervan" is INSTANCE of Cyc concept **Person**

"Christopher Bellew" is INSTANCE of Cyc concept **Person**

"senior vice president" is EQUIVALENT to Cyc concept

**SeniorVicePresident-CorporateOfficer**

Christopher Bellew, senior vice president, Bache Commodities Limited, professional, current → Cyc Assertion: **(positionOfPersonInOrganization ChristopherBellew BacheCommoditiesLimited SeniorVicePresident-CorporateOfficer)**

**Possible Query:**

- Find corporate officers in Bache Commodities Limited

**Answer:**

ChristopherBellew  
SeniorVicePresident-CorporateOfficer

**Creation Time :** after 0.009 seconds (at 17:27:20 today)  
**Steps to This Answer :** 8

**Justifications :** [Full]

● (positionOfPersonInOrganization ChristopherBellew  
BacheCommoditiesLimited SeniorVicePresident-CorporateOfficer) in  
OpenCalaisFactExtractionMt

● (isa genls TransitiveBinaryPredicate) in UniversalVocabularyMt

● (genls SeniorVicePresident-CorporateOfficer VicePresident-CorporateOfficer) in UniversalVocabularyMt

● (genls VicePresident-CorporateOfficer CorporateOfficer) in UniversalVocabularyMt

● (positionOfPersonInOrganization ChristopherBellew  
BacheCommoditiesLimited SeniorVicePresident-CorporateOfficer) in  
OpenCalaisFactExtractionMt

● (isa genls TransitiveBinaryPredicate) in UniversalVocabularyMt

● (genls SeniorVicePresident-CorporateOfficer VicePresident-CorporateOfficer) in UniversalVocabularyMt

● (genls VicePresident-CorporateOfficer CorporateOfficer) in UniversalVocabularyMt

## Appendix 5: Example of News Annotation with Cyc Tagger

Appendix 5 shows how the Cyc tagger (a tool, which allows text annotation with the Cyc ontology elements) is used for financial news tagging. Cyc concepts are provided in curves and start with “#\$” symbols.

```
Oil (##$Oil) sets (##$PuttingSomethingSomewhere) new (##$NewArtifact)
record (##$Record-ExtremeValue ##$PhonographRecord ##$PhonographRecord-LP
##$Record) near $ (##$Dollar-UnitedStates)121 a barrel (##$Barrel-
UnitOfVolume ##$Barrel-Container): Financial News (##$News)
Supplydisruptions in Nigeria (##$Nigeria), where a strike and attacks
(##$Attack-MilitaryOperation) by militants (##$Militant) has (##$possesses)
hit (##$ShootingAndHittingSomething ##$HittingAnObject) production
(##$Production-Generic), has (##$possesses) supported
(##$SupportingSomething-TransferringPossession ##$CorroborationEvent) a
market ((##$MarketTypeByFocalProductTypeFn ##$FinancialAsset)
##$GroceryStore ##$ProductTypeByMarketCategory ##$Market) that is nervous
((##$MediumToVeryHighAmountFn ##$Nervousness)) about any threats (##$Threat
##$threatToTypePossiblyPresentInRegion ##$sitTypePosesThreatToType
##$SubcollectionOfWithRelationToTypeFn ##$Collection
##$sitTypePosesThreatToType ##$Collection)) to supply (##$GivingSomething
##$MakingSomethingAvailable). Tension (##$Tension ##$Stress-Feeling) with
Iran (##$Iran) ratcheted higher when the world (##$PlanetEarth)'s fourth-
biggest oil (##$Oil) producer (##$Producer-Movie ##$Producer
##$ManufacturingOrganization) refused (##$Refusing-CommunicationAct) to
accept intrusive inspections (##$Inspecting) of its nuclear program
((##$ResearchingAndDevelopingFn ##$NuclearRelatedMaterial)) that the West
fears (##$Fear) could be linked (##$ConnectingTogether) to weapons
(##$Weapon). U.S. (##$UnitedStatesOfAmerica) light crude (##$Petroleum-
LightCrudeOil) for June (##$June) delivery (##$BirthEvent) was up (##$Up-
Generally) 7 cents (##$Cent-UnitedStates) at $ (##$Dollar-
UnitedStates)120.04 a barrel (##$Barrel-UnitOfVolume ##$Barrel-Container),
by 7:55 a.m. EDT after earlier touching (##$AffectingSomething
##$TouchingSomethingBriefly ##$Touching-Handling-Management) a record
(##$Record-ExtremeValue ##$PhonographRecord ##$PhonographRecord-LP
##$Record) high (##$highAmountOf) of $ (##$Dollar-UnitedStates)120.93.
London ((##$CityNamedFn "London" ##$Ontario-CanadianProvince)) Brent crude
(##$BrentCrudeOil) was up (##$Up-Generally) 33 cents (##$Cent-UnitedStates)
at $ (##$Dollar-UnitedStates)118.32 a barrel (##$Barrel-UnitOfVolume
##$Barrel-Container), after an earlier record (##$Record-ExtremeValue
##$PhonographRecord ##$PhonographRecord-LP ##$Record) of $ (##$Dollar-
UnitedStates)119.07. Gold (##$Gold-SubIndustrySP) was also strong
(##$Strong), as oil (##$Oil)'s advance (##$Improvement-Transformation)
helped (##$HelpingAnAgent) spur (##$Spurring-PromotingSomething) a rebound
(##$Rebounding) from a four-month low last week (##$CalendarWeek
##$WeeksDuration). But gold (##$Gold ##$GoldColor) is still some way
(##$Path-Customary) below a record (##$Record-ExtremeValue
##$PhonographRecord ##$PhonographRecord-LP ##$Record) of $ (##$Dollar-
UnitedStates)1,030.80 an ounce (##$Ounce ##$Ounce-Troy ##$Ounce-
UnitOfVolume) reached (##$Reaching ##$ArrivingAtAPlace) on March (##$March)
17. "The downward move (##$MovementEvent) in oil (##$Oil) last week
(##$CalendarWeek ##$WeeksDuration) now (##$Now-Indexical) seems like only
(##$Only-NLAttr) a correction (##$Correction)," said (##$Informing
##$Speaking) Christopher Bellew, senior vice president
```

(#SeniorVicePresident-CorporateOfficer) at Bache Commodities  
 (#CommodityProduct). "The effect (#SeventOutcomes  
 (#SubcollectionOfWithRelationFromTypeFn #Situation #SeventOutcomes  
 #Event)) of the credit (#CreditAccount) crisis (#Crisis) in the  
 United States (#UnitedStatesOfAmerica) is reducing (#DecreaseAction  
 #ReducingALiquid #WeightLosingProcess #DecreaseEvent) people  
 (#Person)'s disposable income (#ApproximatePay #Income  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryRate #Income  
 #SocialBeing)) and you'd expect ((#HavingPropositionalAttitudeFn  
 #Expects)) this to have (#Possesses) an impact (#Colliding) on the  
 oil (#Oil) price (#BasicPrice #TotalCharge #Cost  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue #BasicPrice  
 #TemporalThing) (#SubcollectionOfWithRelationFromTypeFn  
 #MonetaryValue #TotalCharge #TemporalThing)  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue #Cost  
 #TemporalThing)), but it's not having (#Possesses) any impact  
 (#Colliding)." Demand (#Need-SystemCondition #Demanding-  
 CommunicationAct) from emerging (#TransferOut) markets  
 ((#MarketTypeByFocalProductTypeFn #FinancialAsset) #GroceryStore  
 #ProductTypeByMarketCategory #Market) such as India (#India) and  
 China (#China-PeoplesRepublic) is more than compensating  
 (#MakingReparationsForSomething) for the U.S. (#UnitedStatesOfAmerica)  
 downturn (#Downturn), he said (#Informing #Speaking). Goldman Sachs  
 (#GoldmanSachsGroup) predicted (#MakingAPrediction) oil (#Oil) could  
 soar (#Gliding) towards \$ (#Dollar-UnitedStates)150-\$ (#Dollar-  
 UnitedStates)200 a barrel (#Barrel-UnitOfVolume #Barrel-Container)  
 because of a lack (#Deficiency) of adequate supply  
 ((#MakingAvailableFn #Provisions) #Supplies) growth (#Tumor  
 #GrowthEvent). "The possibility of \$ (#Dollar-UnitedStates)150-\$  
 (#Dollar-UnitedStates)200 per barrel (#Barrel-UnitOfVolume #Barrel-  
 Container) seems increasingly likely over the next 6-24 months  
 (#CalendarMonth), though predicting (#MakingAPrediction) the ultimate  
 peak (#Mountain) in oil (#Oil) prices (#BasicPrice #TotalCharge  
 #Cost (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue  
 #BasicPrice #TemporalThing) (#SubcollectionOfWithRelationFromTypeFn  
 #MonetaryValue #TotalCharge #TemporalThing)  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue #Cost  
 #TemporalThing)) as well as the remaining duration (#Time-Quantity) of  
 the upcycle remains a major uncertainty," the bank (#Bank-  
 Topographical) said (#Informing #Speaking). The U.S.  
 (#UnitedStatesOfAmerica) investment bank (#InvestmentBank) had  
 (#Possesses) predicted (#MakingAPrediction) back in 2005 that oil  
 (#Oil) was entering (#TransferIn) a "super-spike" period  
 (#TimeInterval). Oil (#Oil) prices (#BasicPrice #TotalCharge #Cost  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue #BasicPrice  
 #TemporalThing) (#SubcollectionOfWithRelationFromTypeFn  
 #MonetaryValue #TotalCharge #TemporalThing)  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue #Cost  
 #TemporalThing)) further into the future (#TheFuture-Generic) have  
 (#Possesses) also risen (#AscendingEvent #IncreaseEvent  
 #GettingUpFromBedAfterSleeping) sharply (#SharpEdged), with prices  
 (#BasicPrice #TotalCharge #Cost  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue #BasicPrice  
 #TemporalThing) (#SubcollectionOfWithRelationFromTypeFn  
 #MonetaryValue #TotalCharge #TemporalThing)  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue #Cost  
 #TemporalThing)) out to 2016 above \$ (#Dollar-UnitedStates)110 a  
 barrel (#Barrel-UnitOfVolume #Barrel-Container). VULNERABLE  
 (#VulnerableThing) Oil (#LubricatingSomething) has (#Possesses)  
 nearly double (#PairFn) in the past (#ThePast-Generic) year  
 (#YearsDuration #CalendarYear) and is up (#Up-Generally) by a quarter  
 (#CalendarQuarter) since the start (#StartingPoint  
 (#SubcollectionOfWithRelationFromTypeFn #TimePoint #StartingPoint





#TemporalThing)) on the U.S. (#UnitedStatesOfAmerica) economy  
 (#EconomicSystem) on his trip (#Travel-TripEvent) to the world  
 (#PlanetEarth)'s top exporter (#exports  
 (#SubcollectionOfWithRelationToTypeFn #GeopoliticalEntity #exports  
 #FirstOrderCollection)) later this month (#CalendarMonth). Bush  
 (#Bush-MusicGroup) has (#possesses) called on the Organization  
 (#Organization) of the Petroleum (#Petroleum-CrudeOil) Exporting  
 (#ImportExportEvent) Countries (#Country) to raise (#LiftingAnObject  
 #RaisingLivingThings #Resurrection) output (#outputsRemaining  
 #outputsCreated #outputs (#SubcollectionOfWithRelationFromTypeFn  
 #SomethingExisting #outputs #CreationOrDestructionEvent)  
 (#SubcollectionOfWithRelationFromTypeFn #SomethingExisting  
 #outputsCreated #CreationEvent)  
 (#SubcollectionOfWithRelationFromTypeFn #PartiallyTangible  
 #outputsRemaining #PhysicalTransformationEvent)) to help  
 (#HelpingAnAgent) bring (#Conveying-Generic) down prices (#basicPrice  
 #totalCharge #cost (#SubcollectionOfWithRelationFromTypeFn  
 #MonetaryValue #basicPrice #TemporalThing)  
 (#SubcollectionOfWithRelationFromTypeFn #MonetaryValue #totalCharge  
 #TemporalThing) (#SubcollectionOfWithRelationFromTypeFn  
 #MonetaryValue #cost #TemporalThing)). The U.S.  
 (#UnitedStatesOfAmerica) dollar (#Dollar-UnitedStates), whose decline  
 in the past (#ThePast-Generic) months (#CalendarMonth) has  
 (#possesses) been driving (#RelentlesslyCoercingAnAgent  
 #DrivingAGolfBall #TransportInvolvingADriver) speculative investments  
 (#InvestmentVehicle) in dollar-denominated crude and other commodities  
 (#CommodityProduct), was weaker versus the euro (#Euro) on Tuesday  
 (#Tuesday) on continued (#Continuation) doubt (#Doubt) about the  
 health of the U.S. (#UnitedStatesOfAmerica) economy (#EconomicSystem)  
 despite upside surprise (#Surprise) from recent economic indicators  
 (#EconomicIndexPredicate). Later in the week (#CalendarWeek  
 #WeeksDuration) on Wednesday (#Wednesday), traders (#Trader) will  
 watch (#WatchingSomething) the weekly (#Weekly) U.S. government  
 (#UnitedStatesFederalGovernment) report (#WrittenReportOnSituation  
 #WrittenReportOnSituation-CW) on fuel (#CombustibleFuelSubstance)  
 inventories, which is expected ((#HavingPropositionalAttitudeFn  
 #expects)) to show (#DisplayingSomething) a 1.8 million-barrel build  
 (#AnimalTypeByPhysicalBuild) in crude stocks (#Stock  
 #BiologicalSubspecies #Stock-GunPart #StockTypeByBusinessAndClass  
 (#UnitOfCountFn #Stock)), a 1.1 million-barrel rise in distillate  
 inventories and a 100,000-barrel fall (#FallingEvent) in gasoline  
 (#GasolineFuel) stocks (#Stock #BiologicalSubspecies #Stock-GunPart  
 #StockTypeByBusinessAndClass (#UnitOfCountFn #Stock)). (Additional  
 reporting (#RegisteringAComplaint #Reporting) by Baizhen Chua in  
 Singapore (#Singapore #CityOfSingapore); editing (#TextEditing  
 #EditingOfCW) by James (#James-MusicGroup) Jukwey) (1 row  
 (#RowOfObjects #DisputeEvent))

## Appendix 6: Publications

Appendix 6 contains a list of research publications by Inna Novalija.

### Journal Papers:

Novalija, I.; Mladenčić, D. Ontology Extension Towards Analysis of Business News. *Informatica journal* **34**, 517–522 (2010). *Informatica is an international refereed journal*.

Bradeško, L.; Dali, L.; Fortuna, B.; Grobelnik, M.; Mladenčić, D.; Novalija, I.; Pajntar, B. Contextualized Question Answering. *Journal of Computing and Information Technology* **18/4**, 325-332 (2010). doi:10.2498/cit.1001912. *Journal of Computing and Information Technology is an international refereed journal*.

Novalija, I.; Mladenčić, D.; Bradeško, L. OntoPlus: Text-Driven Ontology Extension Using Ontology Content, Structure and Co-occurrence Information. *Knowledge-Based Systems journal* **24/8**, 1261-1276 (2011). doi:10.1016/j.knosys.2011.06.002. *Knowledge-Based Systems is an SCI and SCI expanded indexed journal*.

### Conference Papers:

Novalija, I.; Mladenčić, D. Extending Ontologies for Annotating Business News. In: *Zbornik 11. mednarodne multikonference Informacijska družba - IS 2008, Proceedings of the 11th International Multiconference Information Society - IS 2008*, volume A, 186-189 (Ljubljana: Institut Jožef Stefan, 2008).

Novalija, I.; Mladenčić, D. Semi-Automatic Ontology Extension Using Text Mining. In: *Zbornik 12. mednarodne multikonference Informacijska družba - IS 2009, Proceedings of the 12th International Multiconference Information Society - IS 2009*, volume A, 214-217 (Ljubljana: Institut Jožef Stefan, 2009).

Novalija, I.; Mladenčić, D. Building a Concept Shell: Ontology Population with Facts from WWW. In: *Zbornik 13. mednarodne multikonference Informacijska družba - IS 2010, Proceedings of the 13th International Multiconference Information Society - IS 2010*, volume A, 165-168 (Ljubljana: Institut Jožef Stefan, 2010).

Novalija, I.; Mladenčić, D. Content and Structure in the Aspect of Semi-Automatic Ontology Extension. In: *Proceedings of ITI 2010 32nd International Conference on Information Technology Interfaces*, 115-120 (Zagreb: University of Zagreb, SRCE University Computing Centre, 2010).

Bradeško, L.; Dali, L.; Fortuna, B.; Grobelnik, M.; Mladenčić, D.; Novalija, I.; Pajntar, B. Contextualized Question Answering. In: *Proceedings of ITI 2010 32nd International Conference on Information Technology Interfaces*, 73-78 (Zagreb: University of Zagreb, SRCE University Computing Centre, 2010).

Below, we present a view on the online version of the publication “OntoPlus: Text-Driven Ontology Extension Using Ontology Content, Structure and Co-occurrence Information” from Knowledge-Based Systems journal. The paper version of this publication can be found in Knowledge-Based Systems journal, 24/8 from December 2011.

AbstractArticleFigures/TablesReferences



Knowledge-Based Systems  
Volume 24, Issue 8, December 2011, Pages 1261–1276

doi:10.1016/j.knsys.2011.06.002 | How to Cite or Link Using DOI

Cited By in Scopus  
(0)


Permissions & Reprints

### OntoPlus: Text-driven ontology extension using ontology content, structure and co-occurrence information

Inna Novalija  , Dunja Mladenić, Luka Bradeško

Artificial Intelligence Laboratory, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

Received 1 July 2010; revised 20 April 2011; Accepted 1 June 2011. Available online 12 June 2011.

 Purchase

Rent the full-text article on DeepDyve

For just \$3.99  
24 hour access  
Read-only  
Non-printable

**Abstract**

This paper addresses the process of semi-automatic text-driven ontology extension using ontology content, structure and co-occurrence information. A novel **OntoPlus** methodology is proposed for semi-automatic ontology extension based on text mining methods. It allows for the effective extension of the large ontologies, providing a ranked list of potentially relevant concepts and relationships given a new concept (e.g., glossary term) to be inserted in the ontology. A number of experiments are conducted, evaluating measures for ranking correspondence between existing ontology concepts and new domain concepts suggested for the ontology extension. Measures for ranking are based on incorporating ontology content, structure and co-occurrence information. The experiments are performed using a well known Cyc ontology and textual material from two domains – finances and, fisheries & aquaculture. Our experiments show that the best results are achieved by combining content, structure and co-occurrence information. Furthermore, ontology content and structure seem to be more important than co-occurrence for our data in the financial domain. At the same time, ontology content and co-occurrence seem to have higher importance for our fisheries & aquaculture domain.

**Highlights**

- We address the process of semi-automatic text-driven ontology extension.
- Proposed **OntoPlus** methodology uses ontology content, structure, co-occurrence data.
- For experiments we exploit Cyc ontology and textual information from two domains.
- Content and structure are more important than co-occurrence in financial domain.
- Content and co-occurrence have higher value in fisheries and aquaculture domain.


**Keywords:** Knowledge engineering methodologies; Ontology extension; Large-scale ontology; Text mining; Semantic technologies


**Article Outline**

1. Introduction
2. Related work
3. Problem definition
4. Methodology
  - 4.1. Extension of Cyc Knowledge Base
5. Evaluation
  - 5.1. Data description
  - 5.2. Experimental settings
  - 5.3. Results
    - 5.3.1. Concept ranking
    - 5.3.2. Relation ranking
    - 5.3.3. Examples of Cyc KB extension
6. Discussion
7. Conclusion

Acknowledgements

References

 Corresponding author. Tel.: +386 41298345; fax: +386 14773315.

 Purchase

Copyright © 2011 Elsevier B.V. All rights reserved.

Knowledge-Based Systems  
Volume 24, Issue 8, December 2011, Pages 1261–1276

## **Appendix 7: Biography**

Inna Novalija (Koval) was born in Kiev, Ukraine, on July 10, 1983.

She received her Bachelor of Science degree in computer science at the Faculty of Informatics of National University of “Kyiv-Mohyla Academy” (Ukraine) in 2003.

She also holds Master of Science degree in Intellectual Systems of Decision Taking (Ukraine, 2005) and Master of Arts degree in Economics (Hungary, 2007).

In 2007-20011 she was working in the area of knowledge technologies and was enrolled in the PhD program at Jozef Stefan International Postgraduate School (Slovenia). Her research interests include Text Mining, Ontologies and Reasoning, Language Technologies, Economic and Business Data Analysis.

She participated in conferences on Data Mining & Data Warehouses and Information Technology Interfaces and took part in European FP7 projects, such as ACTIVE (ACTIVE - Enabling the Knowledge Powered Enterprise) and EURIDICE (EURopean Inter-Disciplinary research on Intelligent Cargo for Efficient, safe and environment-friendly logistics).

