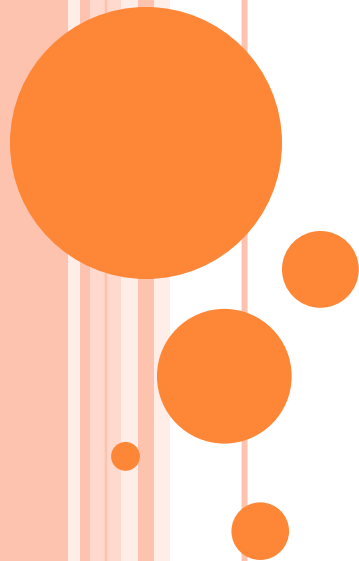


Extending Ontologies for Annotating Business News

**Inna Novalija, Dunja Mladenić
J. Stefan Institute, Slovenia**

17 October, 2008



OUTLINE

- Introduction
 - Semantic technologies in the news domain
 - Development of Financial Ontology
- Methodology
 - Design criteria for ontology development
 - Overview of the methodologies, Cyc method
- Preliminary experiments
 - Financial news domain
- Discussion & Conclusion



INTRODUCTION

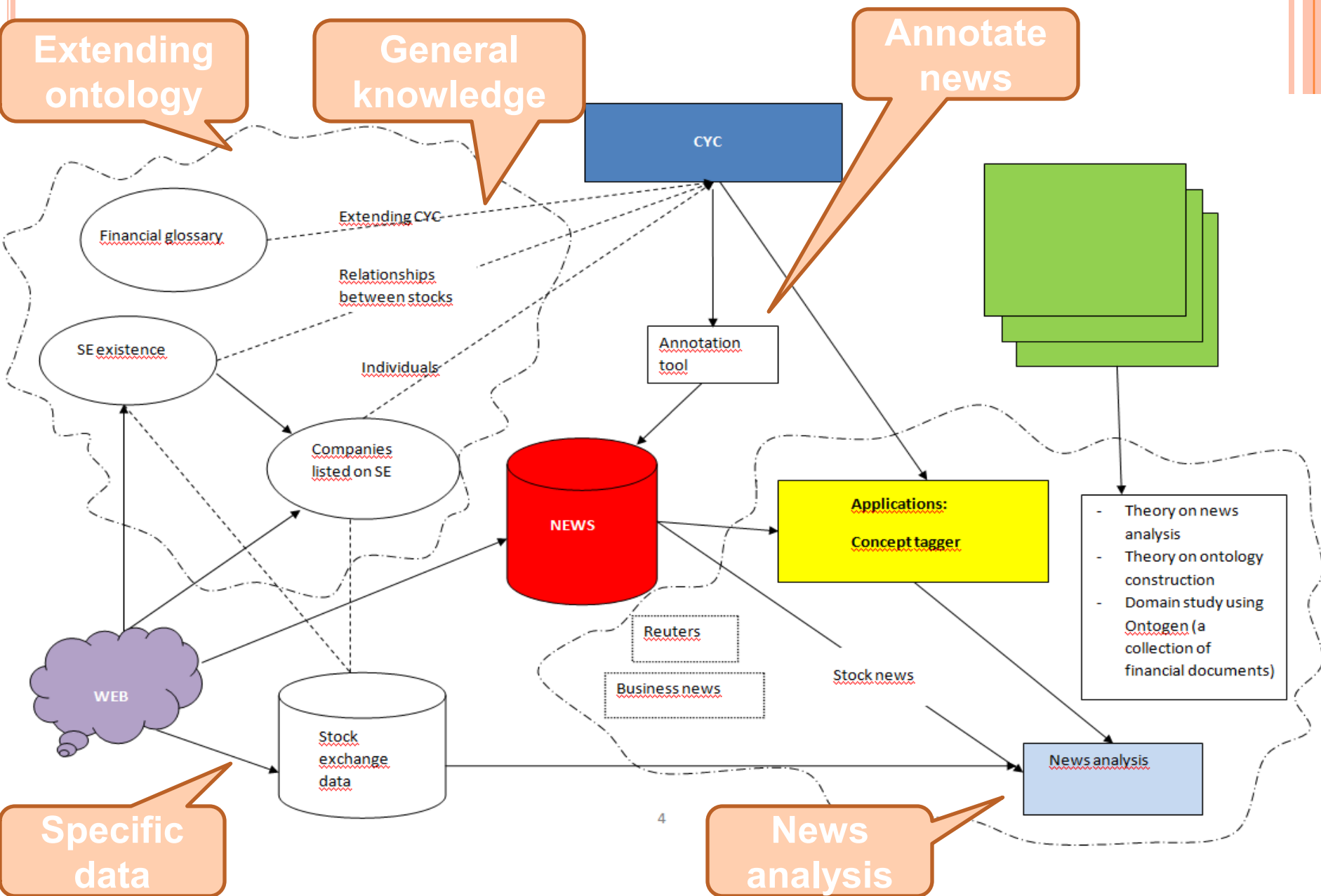
SEMANTIC TECHNOLOGIES IN THE NEWS DOMAIN

- **Focus:** manual extensions of ontologies to support the annotation of business news.
- **Findings:** proposed extensions of ontology results in annotation with better coverage of terms that are relevant for business domains.
- **Why News?**
 - News reports are one of the largest sources of information about society.
 - The analysis of news allows to make the important conclusions about trends in the society life.
 - Semantic technologies already successfully applied for news analysis.



INTRODUCTION

RESEARCH SCHEME



INTRODUCTION

SEMANTIC TECHNOLOGIES IN THE NEWS DOMAIN

Challenges while using semantic technologies in news analysis:

- News are dynamic (constantly changing).
- News are interactive.
- News are socially biased.
- News agencies produce huge amounts of content.




INTRODUCTION

DEVELOPMENT OF FINANCIAL ONTOLOGY

Challenges while building Financial Ontology:

- Nature of the financial tasks
 - dynamic, distributed, global, heterogeneous in nature
 - large amount of continually changing, and generally unorganized, information available
 - variety of all kinds of information (like market data, financial report data, breaking news, etc.)
- Slow standardization efforts
 - high complexity of the financial standards
 - high competition and dynamics of the financial sector influence the implementation of the new technologies

Challenges for our work:

- Dynamics of the financial sector → importance of the temporal aspects while building a Financial ontology
 - Heterogeneity of financial information and tasks.
- 

METHODOLOGY

DESIGN CRITERIA FOR ONTOLOGY DEVELOPING

- Design criteria for ontology development :
 - Clarity, Coherence, Extendibility, Minimal encoding bias, Minimal ontological commitment.
- Additional methodological principles:
 - Ontology double articulation, Ontology modularization principle.

Our focus: clarity, extendibility and ontology modularization



METHODOLOGY

OVERVIEW OF CANDIDATE METHODOLOGIES

Ontology development methods/methodologies:

- *Uschold and King's method* – methodology for ontology building, proposed in 1995.
- *Grüninger and Fox's methodology* – method based on competency questions, proposed in 1995.
- *METHONTOLOGY* - complete ontology development process; one of the most famous methodologies, proposed in 1997.
- *On-To-Knowledge* - developed in 2004 for introducing and maintaining ontology based knowledge management applications into enterprises.
- *Cyc method* - arises from the development of Cyc Knowledge Base, introduced in 1984.

Our research: using Cyc method



METHODOLOGY

CYC METHOD

- Phases of build the Cyc ontology:
 - **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.
- Building top-down (first top level ontology containing the most abstracts concepts)



METHODOLOGY

CYC KNOWLEDGE BASE

- One of the largest knowledge bases
- “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”
- Divided into the large number of “microtheories”, each of which represents the set of assumption for a particular knowledge domain
- Contains nearly 300 000 concepts and several dozen hand-entered assertions about/involving each of them and 15 000 different predicates.

Assertions are continually added manually as well as automatically as a product of the inference process.



METHODOLOGY

CYC

- Cyc gives an extremely powerful mechanism of creating and using different ontologies.
- **Reasons for using Cyc in our research:**
 - Extensive amount of versatile integrated information
 - Large number of assertions
 - OpenCyc & ResearchCyc
 - Flexible and convenient language (CycL)
 - Suitable interface



PRELIMINARY EXPERIMENTS

NEWS ANNOTATION - FINANCIAL NEWS DOMAIN

○ **Experiment:**

- Random samples of ten news articles from two news datasets: Reuters and Yahoo Finance
- Manually annotated for financial terms (for evaluation)
- Cyc annotator applied on the news :
 - using the original Cyc ontology
 - simulating extension of the original Cyc ontology

○ **Reuters news:**

- Set of 1450 news from 1996 that were labeled as financial and business services
 - *18 FINANCIAL AND BUSINESS SERVICES*
 - *181 BANKING AND FINANCIAL SERVICES*
 - *182 INSURANCE*
 - *1831 FINANCIAL SERVICES*
 - *184 RENTING AND LEASING EQUIPMENT*
 - *185 REAL ESTATE DEALING*

○ **Yahoo Finance news:**

- News for the last three months (mid. May – mid. August 2008)
- 26000 news articles
- Uncategorized news materials with financial connotation



PRELIMINARY EXPERIMENTAL RESULTS ON ONTOLOGY EXTENSION BASED ON YAHOO GLOSSARY OF FINANCIAL TERMS

Precision: 84% compared to 56%

Recall: 63% compared to 41%

Table 1. Financial news tagged by Cyc (Reuters)

Article name	Total words	Fin. Terms	Fin. Terms tagged	Fin. Terms tagged correctly	Precision, %	Recall, %	Precision after adding terms from Yahoo glossary, %	Recall after adding terms from Yahoo glossary, %
Doc. 1 Lloyd's of London serves notice of emergency stay.	648	14	13	8	62%	57%	93%	93%
Doc. 2 UK Lloyd's moves to ward off doubts on recovery.	489	13	9	6	67%	46%	75%	69%
Doc. 3 CANADA: Canadian banks poised for higher third-quarter profits.	612	45	30	17	57%	38%	69%	40%
Doc. 4 Malaysia's Intria buys into two construction firms.	432	24	15	9	60%	38%	80%	50%
Doc. 5 Slovak PM sees banks releasing funds for bad debts.	156	10	10	2	20%	20%	90%	90%
Doc. 6 <u>ArgentBank</u> to buy Assumption Bank & Trust.	64	9	7	6	86%	67%	100%	78%
Doc. 7 Nationwide, Halifax bid for UK defense sale - paper.	123	9	5	4	80%	44%	80%	44%
Doc. 8 Australia: Current Australian Takeovers (A to E) - Aug 26.	481	19	17	8	47%	42%	82%	74%
Doc. 9 AUSTRALIA: RTRS-Australia's COAL in A\$300 mln float - paper.	365	18	13	4	31%	22%	100%	83%
Doc. 10 China state firms form new insurance company.	67	11	7	6	86%	55%	100%	64%
Avrg.	344	17	13	7	56%	41%	84%	63%

Manually identified

Tagged by Cyc

PRELIMINARY EXPERIMENTS

RESULTS

Results of the experiment:

- Annotating using the original Cyc
 - Reuters news: precision 56% and recall 41%
 - Yahoo financial news : precision 69% and recall 57%
- Publicly available Yahoo Financial Glossary
 - Contains about 50% of the financial terms untagged or tagged incorrectly by Cyc (54% Reuters, 52% Yahoo financial news)
- Annotating simulating extension of Cyc by Yahoo Glossary increases the average precision and recall
 - Reuters news: precision 84% and recall 63%
 - Yahoo financial news : precision 82% and recall 73%

Hypothesis to be tested in future work:

- Extension of Cyc ontology by the terms from Yahoo Financial Glossary
→ improved annotation and analysis of the financial news

DISCUSSION & CONCLUSION

- Cyc gives powerful mechanisms for creation and extension of ontologies:
 - Cyc knowledge base contains a large number of assertions
 - Available for public/research: OpenCyc and ResearchCyc
 - Flexible and convenient language (CycL)
 - Suitable interface
- Financial Ontology developed within Cyc as basis for financial news analysis



Thank You for the Attention!

