<table>
<thead>
<tr>
<th>Title</th>
<th>Semantic Paths in Business Filings Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>O’Riain, Seán</td>
</tr>
<tr>
<td>Publication Date</td>
<td>2012-06-18</td>
</tr>
<tr>
<td>Item record</td>
<td><a href="http://hdl.handle.net/10379/3001">http://hdl.handle.net/10379/3001</a></td>
</tr>
</tbody>
</table>
Semantic Paths in

Business Filings Analysis

Seán Ó Riain B.Sc., M.Sc.,

Digital Enterprise Research Institute
College of Engineering and Informatics
National University of Ireland, Galway

A dissertation submitted for the degree of

Doctor of Philosophy

June 2012

Supervision: Dr. Paul Buitelaar
The research presented in this thesis was performed at the Digital Enterprise Research Institute, College of Engineering and Informatics at the National University of Ireland, Galway. The research was financially supported by Science Foundation Ireland under grant number SFI/08/CE/11380 (Lion-2).
Abstract

Semantic Paths in Business Filings Analysis

Seán Ó Riain

Digital Enterprise Research Institute
College of Engineering and Informatics, National University of Ireland, Galway

A dissertation submitted for the degree of

Doctor of Philosophy
June 2012

Supporting competitive business analysis of financial reports through the automated analysis and interpretation of their natural language sections, presents specific challenges including information that can be ambiguous, camouflaged, or tacitly hidden within the narrative. These sections present terminology and structural challenges for information extraction that require the application of linguistic and heuristic based domain modelling to identify the information requirement.

This thesis investigates a modelling approach that incrementally builds the business analysts information requirement as a series of Semantic Paths grounded in domain linguistic and user heuristics. A Competitive Analysis Ontology (CAO) is defined to provide semantic representation of the information requirement necessary to drive linguistic analysis and information extraction.

The evaluation of the CAO within the financial sub-domain of competitive analysis is investigated, through the development of the Analyst Work Bench (AWB), is presented. The AWB linguistically analyses a Form 10-Q’s disclosure sections, automatically populates the CAO and provides the analyst’s information requirement. The AWB leverages the CAO Semantic Paths for information search and extraction capability, to support an analyst perform a competitive analysis, with reduced manual effort.

Evaluation based on design-science principles, use methods from information retrieval and information system success to determine CAO performance and usability. A controlled experiment that compares competitive analysis performance using the AWB, against its manual performed equivalent, reported a 37% performance increase using the AWB to identify relevant information. Usability evaluation further found that CAO use contributed to task structuring, and structured information provision in a manner that directly supported task performance.
Acknowledgements

This thesis had its genesis in discussions with Hewlett-Packards Business Process Outsourcing (BPO) Unit, as to whether the information needed by business analysts could be automated, but within the context of ‘what they needed to do their job’. John Collins, the BPO manager deserves special thanks for his interest and contribution to the resulting use case requirements, the essential transfer of analyst know how, and facilitating preliminary evaluations of the work bench demonstrator. For introduction to the areas of ontologies and specifically the principles and usage considerations of the DOGMA modelling methodology, I would like to express thanks to Peter Spyns. For organising subject matter experts from the HP BPO team to participate in the evaluation, Mike Turley deserves thanks. For on-going, thought provoking and challenging discussion on scientific discourse and evaluation methodology Edward Curry, Paul Buitelaar and Ronan Fox deserve mention.

Without support from both my immediate and extended family, completing this work would have been difficult. Thanks to my parents Seamus and Frances who have always encouraged further education and always enquired as to progress. Thanks to Adrienne for her on-going support and understanding, and Caoimhe and Eóin who develop a keen sense of humour, with the recursive question of ‘is it finished yet’?

Finally thanks must once again go to Edward and Paul for guidance through the process of dissertation structuring and writing.
# Table of Contents

Table of Tables ............................................................................................................. iv

Table of Form 10-Q Report Extracts ............................................................................. v

Table of Equations ........................................................................................................ vi

Table of Figures .............................................................................................................. vi

Chapter 1 Introduction .................................................................................................. 1

1.1 Motivation ............................................................................................................... 1

1.2 Research Questions & Hypothesis ......................................................................... 4

1.3 Research Methodology ......................................................................................... 5

1.4 Research Contribution ......................................................................................... 6

1.5 Thesis Organisation ............................................................................................. 7

1.6 Summary ............................................................................................................... 8

Chapter 2 Business Filings Analysis ............................................................................. 10

2.1 Information Extraction ......................................................................................... 11

2.1.1 Approaches ..................................................................................................... 12

2.1.2 Considerations for Usage ............................................................................... 13

2.2 Business Information Extraction ......................................................................... 14

2.2.1 Extraction from EDGAR Filings ...................................................................... 16

2.2.2 Extraction of Business Information from Other Web Sources .................... 25

2.3 Business Lexical Resources ............................................................................... 26

2.4 Semantic Web and Open Data ............................................................................. 32

2.5 Summary ............................................................................................................... 35

Chapter 3 Knowledge Capture and Ontologies ............................................................ 38
Appendix IV  DOGMA Ontology Modelling Methodology ........................................... 166

Appendix V  Deloen & McLean Success Measures .................................................. 167

Appendix VI  Evaluation Questionnaire .................................................................. 168

Bibliography  □ □ □ ................................................................................................. 169

Table of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Example Financial Information Providers utilising Edgar Filings</td>
<td>15</td>
</tr>
<tr>
<td>2-2</td>
<td>Addressing Information Extraction Task Challenges from Edgar Filings</td>
<td>24</td>
</tr>
<tr>
<td>2-3</td>
<td>Business and Financial Standards Summary</td>
<td>29</td>
</tr>
<tr>
<td>2-4</td>
<td>Financial Information Ontology Summary</td>
<td>31</td>
</tr>
<tr>
<td>3-1</td>
<td>Knowledge Use and Ontologies</td>
<td>45</td>
</tr>
<tr>
<td>4-1</td>
<td>Model of Meta-discourse in Academic Texts (adapted from Hyland 2004)</td>
<td>61</td>
</tr>
<tr>
<td>4-2</td>
<td>Discourse Analysis Term List Extract</td>
<td>62</td>
</tr>
<tr>
<td>4-3</td>
<td>Concept Mapping Proposition Template Listing</td>
<td>66</td>
</tr>
<tr>
<td>4-4</td>
<td>DOM Use in CAO Development</td>
<td>74</td>
</tr>
<tr>
<td>4-5</td>
<td>Competitive Analysis Ontology Breadth and Depth Overview</td>
<td>78</td>
</tr>
<tr>
<td>4-6</td>
<td>Competitive Analysis Ontology, Sales Context, Extract</td>
<td>80</td>
</tr>
<tr>
<td>5-1</td>
<td>Sales Lexon (selected extract)</td>
<td>84</td>
</tr>
<tr>
<td>5-2</td>
<td>Grounding of Sales Lexons (selected extract)</td>
<td>86</td>
</tr>
<tr>
<td>5-3</td>
<td>Creation of Sales Meta-Lexons (selected extract)</td>
<td>87</td>
</tr>
<tr>
<td>5-4</td>
<td>Selected Rule and Associated Class Type Annotation</td>
<td>89</td>
</tr>
<tr>
<td>5-5</td>
<td>Class Instance Templates (Example 1)</td>
<td>92</td>
</tr>
<tr>
<td>5-6</td>
<td>Class Instance Templates, (Example 2)</td>
<td>92</td>
</tr>
</tbody>
</table>
TABLE 5-7 MULTIPLY OCCURRING INSTANTIATED SEMANTIC PATHS FOR RC4002-C4018, RC4002-C4019 94

TABLE 5-8 SEMANTIC PATHS CONTAINING CLASS TERMS: [COMPANY, DEFENDANT, LITIGATION] ............ 95

TABLE 6-1 SET REPRESENTATION OF PRECISION, RECALL (BAEZA-YATES 1999) ................................. 105

TABLE 6-2 CAO EVALUATION IMPACT ON CRANFIELD ASSUMPTIONS .................................................. 107

TABLE 6-3 NLP CRITERIA FOR REFERENCE STANDARD ESTABLISHMENT (FRIEDMAN 1995) ................. 111

TABLE 6-4 NLP CRITERIA FOR MINIMISING BIAS, EXTRACTED FROM (FRIEDMAN 1995) ....................... 112

TABLE 6-5 CAO AND CORPUS STATISTICS SUMMARIES ........................................................................ 113

TABLE 6-6 MANUAL ANNOTATOR PERCENTAGE OF AGREEMENT AND CONSISTENCY ESTIMATES ........... 114

TABLE 6-7 PERFORMANCE IR METRICS ACROSS ALL ANNOTATORS ..................................................... 118

TABLE 6-8 PERFORMANCE COMPARISON AGAINST SIMILAR EDGAR EXTRACTION SYSTEMS ............... 119

TABLE 6-9 SEMANTIC PATHS SUMMARY PER CONTEXT AREA ................................................................. 121

TABLE 6-10 MANUAL VS. SYSTEM USAGE SEMANTIC PATH INSTANTIATIONS SUMMARIES (PART 1) ....... 123

TABLE 6-11 MANUAL VS. SYSTEM USAGE SEMANTIC PATH INSTANTIATIONS SUMMARIES (PART 2) ..... 124

TABLE 6-12 MANUAL VS. SYSTEM USAGE SEMANTIC PATH INSTANTIATIONS SUMMARIES (PART 3) ...... 125

TABLE 7-1 AWB SUCCESS DETERMINATION QUESTIONNAIRE ............................................................... 143

TABLE 8-1 TOP 6 IDENTIFIED RELATIONSHIPS WITH ARGUMENTS .................................................... 152

TABLE 8-2: INSTRUMENT SUMMARY BASED UPON eCommerce BIASED SUCCESS (SEDDON 1996) ...... 167

Table of Form 10-Q Report Extracts

FORM 10-Q EXTRACT 4-1, CITRIX QUARTERLY PERIOD SEPTEMBER 30, 2003 (1) ................................. 53

FORM 10-Q EXTRACT 4-2 CITRIX QUARTERLY PERIOD SEPTEMBER 30, 2003 (2) ................................. 53

FORM 10-Q EXTRACT 4-3 CITRIX QUARTERLY PERIOD SEPTEMBER 30, 2003 (3) ................................. 54

FORM 10-Q EXTRACT 4-4 BEA QUARTERLY PERIOD 30 OCTOBER, 2004 ............................................. 59

FORM 10-Q EXTRACT 4-5 CITRIX QUARTERLY PERIOD SEPTEMBER 30, 2003 (4) ............................... 59
Table of Equations

EQUATION 6-1 IR PERFORMANCE METRICS (van-Rijsbergen 1979) ....................................................... 105

Table of Figures

FIGURE 4-1 EXAMPLE ANALYST ANNOTATION OF FORM 10-Q ............................................................. 52
FIGURE 4-2 EXAMPLE ANALYST ANNOTATION FORM 10-Q WITH DUAL CONTEXT INTERPRETATION ........ 54
FIGURE 4-3 FORM 10-Q CONTENT STRUCTURE ..................................................................................... 55
FIGURE 4-4 DISCLOSURE SECTION HTML SOURCE EXTRACT .............................................................. 56
FIGURE 4-5 XBRL INSTANCE DOCUMENT FINANCIAL INSTRUMENT EXTRACT ........................................ 56
FIGURE 4-6 XBRL INSTANCE DOCUMENT TEXT BLOCK EXTRACT .......................................................... 57
FIGURE 4-7 SEMANTIC PATHS SALES CONTEXT CONCEPT MAP/SEMANTIC NETWORK ...................... 65
FIGURE 4-8 SEMANTIC PATHS DEPICTION IN A MULTI-CONTEXT DIRECTED LABELLED GRAPH ............. 68
FIGURE 4-9 DOGMA MODELLING METHODOLOGY - DOMAIN CONCEPTUALISATION (SPYNs 2007) ....... 70
FIGURE 4-10 AUTOMATED CLASS TYPE INSTANCE ASSIGNMENT NUMBERS PER CONTEXT AREA ........ 81
FIGURE 5-1 AUTOMATICALLY ADDED CLASS TYPE ANNOTATIONS TO FILING HTML SOURCE .......... 90
FIGURE 5-2 CANDIDATE DETERMINATION FOR SEMANTIC PATH INSTANTIATION ................................. 96
FIGURE 6-1 ANNOTATION SETS SYMMETRIC DIFFERENCE OVERVIEW .................................................. 108
FIGURE 6-2 COMPETITIVE ANALYSIS PERFORMANCE METRICS OVERVIEW .................................. 109
FIGURE 6-3 MANUAL VS. SYSTEM USAGE SEMANTIC PATHS SUMMARY TOTALS PER CONTEXT AREA ....... 120
FIGURE 6-4 MANUAL VS. SYSTEM USAGE SEMANTIC PATHS RATER AGREEMENT .................................... 125
Semantic Paths in Business Filings Analysis

Figure 7-1 AWB, Main In-Text Visualization Area ................................................................. 131
Figure 7-2 AWB, Semantic Path Navigator ................................................................. 132
Figure 7-3 AWB, Report with Class Type Annotation Mark-up ................................................ 133
Figure 7-4 AWB, Semantic Path In-Text Class Type Instantiation ........................................ 133
Figure 7-5 AWB, Summary Extraction Report of Relevant Text Segment ................................. 134
Figure 7-6 AWB High Level Component Architecture .......................................................... 135
Figure 7-7 Automated Semantic Annotation Output ............................................................ 136
Figure 7-8 AWB XML Generated Navigator ........................................................................... 137
Figure 7-9 AWB Schema ........................................................................................................ 138
Figure 7-10 AWB Information Flow ........................................................................................ 139
Figure 7-11 D&M IS Success Model (DeLone. 2003) ............................................................. 140
Figure 7-12 System Quality - Assessment of Operational Characteristics .............................. 144
Figure 7-13 Information Quality - Content ............................................................................. 145
Figure 7-14 Information Quality - Context and Linkage ........................................................... 145
Figure 7-15 User Satisfaction – Extent of Positive or Negative Feelings Towards Artefact .... 146
Figure 7-16 Perceived Benefit – To the User .......................................................................... 146
Figure 7-17 Usage of the System – Extent of Usage ............................................................... 147
Figure 8-1 Semantic Paths Marketing Context Concept Map .................................................. 159
Figure 8-2 Semantic Paths Acquisitions Context Concept Map ............................................. 160
Figure 8-3 Semantic Paths Relationship Context Concept Map ............................................ 161
Figure 8-4 Semantic Paths Headcount Context Concept Map ............................................. 162
Figure 8-5 Semantic Paths Profit Context Concept Map ....................................................... 163
Figure 8-6 Semantic Paths R&D Context Concept Map ......................................................... 164
Chapter 1

"If we knew what it was we were doing, it would not be called research, would it?"
Albert Einstein

Introduction

Innovative use of data to derive business insight such as developing competitive advantage, risk monitoring, fraud investigation or identification of new revenue opportunity, involves information integration and some process that transforms data into usable knowledge (Economist 2010). Developing what are essentially knowledge management strategies that help business professionals define and manage their information need, to enhance relevant information discovery within a sub-domain of this business ecosystem, namely competitive analysis is the focus of this research.

1.1 Motivation

Competitive analysis is the process used by companies to understand their market, identify competitors, determine strengths and weaknesses, and predict their next strategic and tactical moves (Zahra 1993; Sheng 2005). As an activity it involves elements of competitive intelligence to collect data followed by its analysis and interpretation in support of executive decision making. Results from competitive analysis provide insight into a competitors operations and strategy (Zahra 1993; Sheng 2005), but with associated manual data collection and analysis from sources such as the U.S. SEC Form 10-Qs\(^1\), accounting for more than 75% of the overall analysis effort expended, there is minimal time dedicated to the data interpretation effort (Zahra 1993). Such corporate reports are considered of huge importance by analysts and in terms of informational analysis, the Form 10-Q comprises financial accounts and statements from the CEO. The statements themselves concern

\(^1\) Quarterly un-audited business report submitted to the U.S. Security & Exchange Commission.
corporate performance, are viewed as a promotional medium to present a corporate image and are important in building credibility and investor confidence. With regulatory requirements and the fact that many companies are value tied to intangibles (e.g. patents, brands or people (Korman 1998)), these statements attract enhanced analysis emphasis, and are actively searched for key information blocks (Van-Deemter 2000; Pfeifer 2007). While they provide the best possible estimates of, and prediction about future conditions and performance, the writing style\(^2\) used and the fact that it purposely attempts to make alternate interpretations of the information presented, that deviates from the companies reported position difficult (Hyland 1998), represents a major problem for the analyst community. So prevalent is this, that the Form 10-Q has been referred to as the “quintessential overload document” containing “semantic camouflage” (Korman 1998).

Accounting, ‘the languages of business’ provides an organisation with the ‘technology’ to present facts, events, intentions and goals to its stakeholders (Davidson 1974). Known for its verboseness and rhetoric, accounting bodies themselves provide specialised courses to ensure professionals possess the prerequisite vocabulary to read financial reports (Crowther 2006). Yet, even professional analysts with an in-depth understanding of accounting knowledge and practices, the provision of consistent appraisals will be subject to the constraints of personal subjective view and fatigue (Li-Yen 2009). An analyst performing a competitive analysis will therefore have to grapple with language, hidden meaning and information volume. Assisting an analyst in data gathering and collection will therefore require:

i) an understanding of accounting terminology studied linguistically through its use in practice

ii) capability to deal with competitive analysis terminology as encountered in unstructured filings narrative sections and

iii) some means of generating relevant information threads based on the analysis task itself.

\(^2\) Within the meaning of Section 27A of the Securities Act (1933), Section 21E of the Securities Exchange Act (1934). Forward looking statements may include terminology such as may, will, expects, plans, anticipates, goals, estimates, potential, or, continue, negative thereof or other comparable terminology regarding beliefs, plans, expectations or intentions regarding the future.
As Nonake however observes, such tacit ‘know how’ resulting from personal experience and insight is difficult to articulate and codify (Nonaka 1995). Eliciting this domain knowledge and its application as text mining in finance (here specifically business information) requires a three-dimensional framework involving data, domain models, and applications (Zhang 2004). Davenport categorized the transformation process in analytic and decision making terms by the level of structured, unstructured or semi-structured analysis questions that the data is capable of answering (Davenport 2001). Structured analysis questions are typically unambiguous, relatively uncomplicated and require little interaction between analyst and decision makers. Traditional transaction orientated data warehousing approaches (OLAP), that transform and consolidate multidimensional data, allow data processing and analysis as standardized reporting or form based output, fall into this category (Henrike 2010). Other analytic approaches draw from data mining, statistical inference, clustering and predictive analytics, based on quantitative data (e.g. stock prices and market indexes), for use in areas such as future trends prediction, behaviour patterns (Zhang 2004; 2009), or building competitive strategies. The application of text analysis using word frequency methods to online financial news articles (e.g. stock market prices (Schumaker 2009) and currency exchange rates predictions (Peramunetilleke 2002)), have also being used as additional source of financial facts albeit presented in a manner such as ‘euro expected to rally against the dollar’, that once interpreted have a direct bearing on analytic parameters and outcomes.

Competitive analysis activities reflect unstructured analytic situations where decision makers lack specific questions, or where an ability to specify an information need in advance may not be clear cut. Analyst activities that depend largely on text based interrogation with questions such as ‘what else could impact revenue negatively?’, make specification of information need, what relevant data to access and from where, time consuming and ambiguous. An ability to traverse this information space in a manner complimentary to the analysis task offers to remove the burden of recognising such camouflaged phrases (e.g. “litigation” may be represented as “legal proceedings” or “court tangles”), within text sections where the most interesting information appears. While the importance of quick and accurate identification of information within these disclosures is evident (Debreceny 2001), those within financial statements present particular difficulty as they are textual, lack
structure and have no common format (Grant 2006). Competitive analysis would benefit from natural language processing and text mining to help make sense of the text to support financial decision making (Zhang 2004), and infused with domain knowledge and insight would make the approaches more effective (Davenport 2001; Davenport 2009). Systems providing such capability to process financial statements related text will therefore remain attractive to a range of stakeholders, from securities brokers to business analysts (Grant 2006).

This thesis addresses these issues with the introduction of the idea of Semantic Paths as a means to model the competitive analysis information requirement. Domain modelling based upon discourse and task analysis identifies the terminology of interest. Information threads identify how discourse terms are associated allowing the development of Semantic Paths. Formalised as the Competitive Analysis Ontology (CAO), the ontology provides a schema to drive natural language processing based extraction from financial text against, and instantiated as a software artefact, a semantic interpretation platform for propositional content identification and appraisal. Central to the approach is the notion of Semantic Paths and their application as the CAO.

1.2 Research Questions & Hypothesis

The research questions addressed are:

[RQ1] Whether information threads (here termed Semantic Paths) are identifiable within the disclosure Sections of EDGAR Form 10-Q filings?

[RQ2] If their formal modelling as the Competitive Analysis Ontology (CAO) could be used to represent information need for competitive analysis?

[RQ3] Could the CAO help satisfy the analyst’s information need by guiding extraction and organising the provision of relevant information for competitive analysis pursuit?

The research hypothesis sought to establish that:

[RH] The Competitive Analysis Ontology as part of an information extraction system assists the identification and extraction of text segments relevant to the performance of a competitive analysis.
1.3 Research Methodology

The research methodology reflects the investigation, approaches taken and methods applied to answer the initial research questions and allow hypothesis evaluation. The methodology adheres to the design-orientated information systems research process guidelines and principles from (Osterle 2011), and design-science research guidelines from (Hevner 2004). The overall methodology adhered to follows the four basic process phases of analysis, design, evaluation and diffusion (Osterle 2011) with:

i) *analysis* catered for in Chapter 1, Chapter 2 and Chapter 3

ii) *design* presented in Chapter 4, Chapter 5 and Chapter 7

iii) *evaluation* method and results in relating to performance in Chapter 6 and utility in Section 7.4

iv) *diffusion*, discussed in Chapter 8.

The methodology was also used for thesis organisation (cf. Section 1.5). The methodology supported the incremental investigation of research questions and hypothesis evaluation through:

[RQ1] Verifying the existence of the information threads within SEC EDGAR 10-Q disclosure sections, with assistance from domain experts guided by contextual enquiry of the competitive analysis task performance (cf. Section 4.1).

[RQ2] In conjunction with domain experts develop the notion of Semantic Paths using concept mapping as the means of domain knowledge capture and initial modelling (cf. Sections 4.2, 4.3). Theoretical introduction of the DOGMA ontology modelling methodology (cf. Section 4.4), outlines its ability to bridge the knowledge divide between domain linguistic knowledge and ontological representation. This was used as background for Competitive Analysis Ontology development (cf. Section 4.6) based on the Semantic Paths, with worked examples presented Section 5.1. Availability of lexical resources to support domain terminology recognition was also investigated (cf. Section 2.3).

[RQ3] A Nearly-New Information Extraction System (ANNIE) of the General Annotation Text Engine (GATE) framework (cf. Section 5.2.1), was used to perform linguistic analysis of, and
extraction from, the textual disclosure Sections of the 10-Q, using the CAO as guiding schema. Extraction templates (cf. Section 5.2.2) were used to create knowledge repository tuples and ontological instantiation determined on the basis of Semantic Path membership and proximity within the disclosure sections. Original filings were semantically annotated using the CAO allowing visualised in-context lookup and traversal. From the research perspective, the combination of linguistic processing and the CAO as the framework for information extraction, semantic interpretation and in-context visualisation of relevant information, represented our experimental platform, the Analyst Work Bench (AWB), described in Chapter 7.

[RH] Use the experimental platform to evaluate the research hypothesis within the spirit of the Cranfield tradition. Financial text segments identified using the platform were compared to a manually created reference standard of class annotations generated against the same corpus (cf. Chapter 6). The Delone and McLean Success Model used within the Business Information Systems area was used to inform on qualitative usability. Both the performance and usability experiments were based on performance of the competitive analysis task using a domain expert focus group (Section 4.1).

### 1.4 Research Contribution

The research makes contributions to the areas of:

*Business Information Systems* with the development of the:

i) notion of Semantic Paths that linguistically models the business analyst information requirements for competitive analysis, as a series of concept-centric information threads

ii) Competitive Analysis Ontology based on Semantic Paths that formally models the business filings disclosure sections and provides the schema to support:

   a. automated structured provision of the analyst information requirement

---

3 Originator of the informational retrieval metrics of precision and recall and experimental approach based on a set of central assumptions that allowed systems evaluation (cf. Section 6.1)
b. enhanced task based search and identification of relevant information within the filing

iii) Analyst Work Bench (AWB) information system (cf. Chapter 7) that automates CAO instantiation based on the linguistic analysis of the U.S. SEC Form 10-Q, and allows an analyst search for, and identify relevant information in the filings disclosure sections as part of performing a competitive analysis

The AWB demonstrator was delivered to the Business Process Outsourcing team of Hewlett-Packard’s European Software Centre, as part of a use case (cf. Section 7.1) that provided analysts with semantic products to assist with information identification for competitive analysis. The AWB and its research resulted in a joint intellectual property (IP) licencing agreement with HP.

iv) ontology use as a framework to automate the structuring of an unstructured analysis problem. In information system terms ontological instantiation can serve as the point of articulation (Mason 1969), allowing consideration of structuring the task driven nature of decision support in terms of the information necessary to improve the effectiveness of the decision making capability.

Semantic Text Analysis: with the provision of a business argumentation taxonomy of concept associations usable for sentence targeting and relationship identification.

1.5 Thesis Organisation

Chapter 1, introduces the research context, motivating factors, challenges and opportunity, research questions and hypothesis, thesis contribution and outlines the research methodology.

Chapter 2, looks at related work of financial information extraction from EDGAR filings. The wider use of business taxonomies and ontologies for business information extraction and domain lexical resources for terminological support is discussed. Related emergent areas of Semantic Web and the growing importance of Open Data usage to general business filings analysis, are also covered.
Chapter 3, for the business audience, introduces the relationship between ontologies and knowledge capture, and **ontology contribution** to knowledge extraction and problem structuring. Practical considerations for ontologies use is also discussed.

Chapter 4, outlines the **formulation of the Semantic Paths** notion by first detailing difficulties associated with text segment extraction from business filing. Contributions from domain discourse analysis and domain experts allowing taxonomy modelling is introduced. The Semantic Paths notion is formally defined and the ontological modelling methodology employed to engineer the CAO ontology using the Semantic Paths described.

Chapter 5, discusses the practicalities of Semantic Path modelling in terms of CAO construction, through its domain coverage, and translation from linguistic model to application. Semantic Paths extraction using the CAO schema to guide linguistic analysis against and instantiate the knowledge repository is then outlined.

Chapter 6, details the **research experiment** fundamentals, evaluation methodology, including research data set establishment, experiment setup, measuring instruments and measures employed. **Experimental accuracy results** are presented and discussed.

Chapter 7, details the **Analyst Work Bench** use case demonstrator, its component architecture and information flow. A usage scenario introduces the task driven nature of competitive analysis, and a **usability evaluation** provides insight into the use of the CAO, as a framework in an application setting.

Chapter 8, provides general discussion, **conclusions and future research** topic suggestions.

### 1.6 Summary

Automated analysis and interpretation of the natural language disclosure sections of financial reports present particular difficulty in situations where the information requirement is ambiguous and the information sought semantically camouflaged within the narrative. The main extraction challenges faced are the:

- lack of an extraction schema that is both cognisant and representative of the information need in domain linguistic terms
• lack of any real semantic interpretation to assist with extracted information interpretation within the context of the original information need, and

• a means of facilitating the revisiting of identified relevant information, within its original context

Addressing the research challenges requires a multidisciplinary approach that combines both business and technical dimensions. Supporting this notion the business dimension contributes domain knowledge consisting of domain expertise, insight and heuristics, domain linguistics and the reasoning elements necessary for task understanding (information threads), through detailing what to look for and its particular relationship role in the tasks information requirement. The technical contribution addresses the schema deficiency by semantically modelling the information threads as Semantic Paths, before collectively modelling them as the competitive analysis ontology. Natural language processing guided by the ontology, provides a semantic based solution that leverages the ontology as a unifying platform, and once instantiated provides a level of semantic interpretation for propositional content assessment.

The research reported on reflects this multidisciplinary combination of both business and technical dimensions, as part of a novel approach to information extraction from corporate filings, in support of a competitive analysis.
Chapter 2

Business Filings Analysis

In broad terms Business Intelligence (BI) can be considered as a set of processes, technologies, methodologies, and architectures that transform raw data into meaningful and useful information (Evelson 2008). Traditionally used to support decision making and enable more effective tactical, strategic and operational insights, BI is fundamentally driven by the collection and merging of information from multiple sources. Originally internal structured data was focused on, but over the last decade in particular, enterprises have demonstrated increasing interest in looking to source freely available external data to enhance understanding and complement existing knowledge repositories. Web based source examples are international organisations (e.g. World bank, IMF), government (e.g. US SEC, Eurostat, irl.gov), news syndicators (e.g. Bloomberg, Reuters), company websites and specialised directories (e.g. dbpedia, TechCrunch, Yahoo!Company, Fund Index). BI analysis techniques have been largely developed to target financial numerical fact extraction based upon quantitative analysis techniques from structured data sources. Textual executive officer comment, financial figures qualification or analyst briefs, found within the confines of semi-structured and unstructured reports, present particular difficulty that the BI community continues to struggle with.

Natural language processing (NLP) and text processing are solution approaches capable of transforming unstructured content into structured representations suitable for further processing (Friedman 1995; Popov 2003; Bontcheva 2004; Maynard 2005; Maynard

---

4 http://www.imf.org/external/data.htm#data
Information Extraction (IE), is one of the key NLP technologies responsible for automated text extraction and tuple creation in some knowledge repository. Extraction accuracy and usefulness requires specification as to what constitutes meaningful semantic extraction and interpretation post extraction. An ontology representing a semantic schema, provides this level of semantic guidance and additionally acts as a lexical resource that supports both extraction and refined analysis.

Within this section we discuss related work in terms of applicability to business filings analysis. First an introduction to the broader area of information extraction along with approaches and considerations for its use are presented in Section 2.1. Section 2.2 then discusses specific information type extraction attempts from EDGAR filings, and more generally from non-EDGAR business information sources. Business lexical resources that consider business and financial related taxonomies and specifications as potential resources to guide extraction, are covered in Section 2.3. Lastly Section 2.4 Semantic Web and Open Data, discusses the emergence of freely available web based data and its exploitation with vocabulary based semantic web approaches.

2.1 Information Extraction

Deriving disambiguated data from natural language texts to cater for some pre-defined information need is a process often referred to as text or data mining\(^7\). The activities of the Message Understanding Conferences (MUC) assisted in creating the widely accepted definition of this activity as being ‘a technology based on analysing natural language in order to extract snippets of information’ (Hirschmann 1998), and adopting the now standard label of Information Extraction (IE). As a process on a primitive level, it involves taking unstructured text input and outputting formatted unambiguous data (Cunningham 2006). The data may then be used directly by applications, databases or indexing procedures for informational retrieval or by stakeholders for further analysis. The importance of such a capability to the enterprise, led Gartner to predict, that it would become part of commercial information retrieval products bundling (Gartner Research 2007).

---

\(^7\) Specifically data mining refers to the activity of applying algorithms for detection and extraction of patterns, and text mining, the process of analysing text to extract information. In the literature both are used interchangeably.
### 2.1.1 Approaches

Appelt and Israel categorized the main engineering approaches for extraction system as being either knowledge engineering (KE) or automatically trainable/machine learning (ML). KE is dependent upon a knowledge engineer, whom with domain relevant texts and assisted by domain experts, will introduce a sublanguage of terms relevant to the application domain. These are then translated into rules used to automatically identify and extract information. Examples of rule based approaches that characteristically involve an iterative process with progressive rule tuning to enhance system performance are AutoSlog (Riloff 1993), CRYSTAL (Soderland 1995), Rapier (Califf 1999), FASTUS (Appelt 1993) TextPro (Pianta 2008) and ANNIE (Cunningham 2005a; Cunningham 2005b).

ML has a lesser dependence on domain familiarity. The approach requires that a selection of corpus texts be semantically annotated in accordance with information sought for extraction. The activity can be performed by someone with *enough* domain knowledge for one system component at a time, such as named entity recognition (NE) or co-reference resolution. Higher degrees of annotation accuracy require review and increased inter-annotator agreement. Training algorithm applied to these annotated texts can generate information extraction rules. Systems using these statistical methods can seek user interaction as part of rule learning refinement. BBNs’ IdentiFinder⁸ and InfoXtract (Rohinik 2006), use the widely accepted statistically based Hidden Markov Models (HMM) for named entity recognition. HMM is based upon event sequence probability (sequential words in text), generated from annotated training data in combination with a hidden system state model. It computes the probabilities of various state transitions to determine the probable sequence of words (Miller 1998). Support Vector Machines (SVM) are another example of a general supervised learning algorithm, that has achieved state of the art performance on NE recognition (Isozaki 2002; Bovee 2005; Gao 2005; Schumaker 2009).

In general, IE system’s syntactic analysis identifies events, entities and basic (at best non-complex) relationships, between events for extraction and produce simple segment parses, which a finite state grammar can formalise robustly and quickly. Encountering

---

language constructs (such as pre-propositional phrases, modifiers or modals), cause parsing difficulty for any follow-on semantic analysis, limiting it to finding predicate structure within a reduced propositional set. Sentence constituents are identified but not their internal structure or role.

Applying domain heuristics is possible where attachment decisions are used for a smaller domain relevant vocabulary. Application to non-domain specific sentences for correct analysis and extraction are largely irrelevant, apart from the analysis and extraction of relevant clauses. This compromise known as shallow parsing, pattern matching or chunking (non-overlapping spans of text), has as it motivation, the location (or ignoring) of information and is similar in concept to computer language lexical analysis. Targeting simple domain specific information, which is correct most of the time, combined with the importance of having a fast and robust approach, a consensus has emerged that shallow finite state analysis represents an alternative to the more error prone full parsing (Appelt 1999). The most successful approaches from MUC-6 (e.g. FASTUS (Appelt 1993) and SRA (Krupka 1998) saw the uptake of pattern matching as a popular extraction technique. The trend has continued with the emergence and adaptation of light weight pattern analysis techniques based upon regular expressions (e.g. RegExTest\(^9\)) applied to areas such as the detection of email fraud (Gao 2005), business reports (Coffin 2001), event based matching (Hasan 2012), and shallow NLP analysis techniques, such as those used by the GATE and KIM system (Cunningham 2002; Popov 2003) becoming pervasive.

For a more in-depth discussion on data extraction methods in general we refer the reader to (Appelt 1999; Baeza-Yates 1999) and for data mining techniques and issues specifically involved in financial applications to (Zhang 2004).

### 2.1.2 Considerations for Usage

Considerations towards construction of new, or the customisation of existing IE systems/frameworks, influence the decision as to suitability of new and emergent scenarios, based on Appelt (Appelt 1999) are:

\(^9\)http://sourceforge.net/projects/regextest/
adequate specification of IE requirement: Is IE a suitable approach for the information task? Does the implementation cost outweigh expected results and does the enterprise have the required technical competency in the first instance? Would system predictability be better than random or might information retrieval results prove comparable or better? What is the resource overhead associated with pro-active rule maintenance, lexical resources and preparation of training data?

availability of knowledge resources: Is there availability of training data with sufficient quantity and quality? If not what is the cost of training data creation, accessibility of domain experts and availability of knowledge engineers for activities such as rule construction? Are there knowledge linguistic resources (e.g. term bases, machine readable dictionaries, taxonomies, ontologies or business specification) that can be exploited?

dealing with multiple text types: What are the relevant source formats and text types that have to be gathered, processed and indexed? Will new domain information, language identification features, text genres and multilingual capability be required?

adaptivity / reusability: Will any created dictionaries, term lists or training data be adaptable to different business tasks?

scalability: What type of processing response is looked for: Real time, off-line processing, or parallel?

2.2 Business Information Extraction

Recognising the business benefits of IE, the financial community has actively developed in-house applications to extract from financial sources such as business filings, news articles and company web sites (Leinnemann 2001; Gerdes 2003; Bovee 2005; Grant 2006). Third party information providers using the same underlying data, distinguish themselves by offering value add business information features and services for specific verticals such as risk management, industry watch events, market trends and financial analysis services (Gerdes 2003). A sample of the more widely known business information providers are presented in Table 2-1 below. Yahoo!Finance, FreeEdgar, Hoovers, EdgarScan and SECNet are examples of providers that limit (free) financial data access to income statements and the balance sheet.
Table 2-1 Example Financial Information Providers utilising EDGAR filings

<table>
<thead>
<tr>
<th>Provider Name</th>
<th>Type of Information/Service Provision (Verified 10.05.12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sec Info</td>
<td>Registrant, filings, relationships to filings, associated names, topics and sites. Categorises report filings allowing company search based on CIK, industry, sector, business</td>
</tr>
<tr>
<td>Hoovers</td>
<td>Information on U.S. companies/industries such as industry watch, business insight, executive bios, company index and business intelligence.</td>
</tr>
<tr>
<td>Morning Star</td>
<td>Internet tool suite providing insight into portfolio management, investment planning, market update and stock analysis for investment advisors, portfolios monitoring and cost alternatives.</td>
</tr>
<tr>
<td>Yahoo! Financial</td>
<td>Stock quotes, news, portfolio management resources, international market data and mortgage rates. Allows company profile browsing. Content is licensed from Morning Star.</td>
</tr>
<tr>
<td>EDGAROnLine</td>
<td>Provider of interactive business and financial data such as SEC filings, fundamental data, institutional holdings, insider trades, registrations and annual reports etc.</td>
</tr>
<tr>
<td>LexisNexis</td>
<td>Service on strategy and business development, financial analysis and risk and information analytics. Has an alliance with EdgarOnLine</td>
</tr>
<tr>
<td>Edgar-Pro</td>
<td>Provider of business, financial and competitive information about public companies, real-time updates</td>
</tr>
<tr>
<td>FreeEdgar</td>
<td>Provides SEC filing free text search, alerts and IPO searches</td>
</tr>
<tr>
<td>SECnet</td>
<td>Offering a variety of securities publications, access to SEC filings and customised on-demand research needs of legal and business professionals.</td>
</tr>
<tr>
<td>Thomson Reuters</td>
<td>Provider of company information, news items, alerts</td>
</tr>
<tr>
<td>Dow Jones Factivia</td>
<td>Factiva.com delivers the news and business information with search tools that fit information need for the intelligent enterprise</td>
</tr>
<tr>
<td>Google Finance</td>
<td>Real-time stock quotes &amp; charts, financial news, currency conversions, ability to create personal portfolio tracking</td>
</tr>
</tbody>
</table>

The others offer services that range from basic company information and EDGAR filings lookup, to more sophisticated market analysis, financial analytics, risk analytics and portfolio management.
2.2.1 Extraction from EDGAR Filings

In terms of IE task challenges, the EDGAR filings represent a mixture of financial statements presented in tabular semi-structured consolidated balance sheets, also referred to as financial items, and unstructured textual financial comment sections, known as disclosures (cf. Section 4.1 for 10-Q filings structure). To highlight extraction complexity introduced by domain specific language found in disclosure sections, and to provide a general understanding as to the important considerations and supporting tasks required of IE for business filings, we classify the challenges as:

- **extraction task supported**, the financial or business sub-domain activity that the information extracted will directly support the execution of
- **text Type**, the type of filings text extracted from to support the extraction task
- **solution approach**, the knowledge engineering or machine learning approach used for extraction
- **language Model**, where applicable the model used to define domain linguistic and support actual extraction
- **lexical Resource Engineering**, the development or use of existing domain lexical resources to support the solution approach

Related EDGAR filings extraction systems are next individually discussed, under the categories of task, approach, model and lexical resources, before being discussed collectively.

The EDGAR2XML (2001) software agent (Leinnemann 2001), was an early attempt to extract stock market investor items from 10-K and 10-Q filings using regular expressions and keyword based wrappers. Recognised keywords, surrounding text and structure were transformed into an XML Document Object Model (DOM) element and progressively added to a DOM tree. Extracted balance sheet items represented by these DOM elements generated a representation of the text structure as an XML output stream. The resulting report was then analysed manually for stock market trading decision making. Catering for variation in filing format resulted in a large number of regular expressions, making extraction difficult to manage. To improve extraction quality, the authors recommended the
use of an ontological model to cater for financial synonyms. Extension of extraction scope to include consolidated statements of income and cash flow was also recommended (Stampert 2008).

The Extraction Agent for SEC Edgar Database (EASE) (Stampert 2008) automatically identifies consolidated balance sheet sections from 10-Q/10-Ks’, using a modified version of the basic vector space model and regular expressions. Filings are split into segments, each treated as document equivalents, and similarity of search queries (e.g. balance sheet, <table>) to documents measured, providing weighted term to document mappings. Constructed domain synonym lists provided the search terms used to detect balance sheet financial items. Incorrect identification by the similarity measures, used to indicate the probability of some segment containing the balance sheet, led to the extension of the vector space model with ordered constructed pairwise search terms (e.g. assets follow assets and liabilities). EASE builds a DOM representation of the underlying HTML report. Navigation within the DOM tree using XPath and minimal regular expressions, identify the balance sheet sections, thereby avoiding the maintenance overhead encountered with (Leinnemann 2001) use of regular expressions. Evaluation using a thirty company sample, bench marked against EDGAR2XML, reported a 97% balance sheet segment identification rate for 10-Qs and 100% for Form 10-Ks.

EDGAR-Analyser (Gerdes 2003) adopted a two stage process to first identify and then extract Y2K remediation efforts in 10-K disclosure sections. At a conceptual level, the filings are treated as a composite of short discussions that address different topics, with underlying factors relevant to the topic, assumed to be in relatively close proximity to the discussion. Adhering to this hierarchical functional dependency model allows general higher level objects be recursively constructed into more specific objects as part of a tiered search strategy. To support this two stage analysis, two sets of keywords were generated, the first, a primary set of ‘Issue Defining’ keyword terms located relevant document topic section areas e.g. millennium bug or Y2K and the second, a set of ‘Critical Factor’ keywords, related to factors associated with the targeted subjects e.g. imbedded chip or cost. The terms were manually constructed from news, research and academic articles, refined and a sliding relevancy scale applied. Main topic terms were first searched for, the context (paragraphs) extracted, before a second pass re-analysed the reduced text segment list for terms relating
to contributing sub issues. The final reduced paragraph list is then left to the user to sort through manually. The first stage seeks to eliminate non relevant records (known as type II errors or false positivess), treating their removal as greater cost than the expense of missing records of interest. The second stage does the opposite, treating the cost of overlooking records (known as type I errors or false negatives) as outweighing the cost of processing more and possibly irrelevant records. For situations where the information requirement cannot be easily codified, identifying segments in a reference standard, that are not found by the system (false positives), or system found segments that are not found in the reference standard (false negatives), are equally important measures to capture. Discussions suggest the incorporation of regular expressions to cater for linguistic variations (such as term stemming and morphology), and incorporation of the WordNet lexical resource to enhance domain specific synonyms listings and system accuracy. Output was presented as a listing of ranked text segments. There was no additional attempt to reduction manual effort through efforts such as related segment linking, based on content, or the adoption of visualization capability. Using a combination of keyword search and relevancy ratings, non Y2K articles were eliminated, reducing those articles that requiring analysis by 42.6% and subsequently the volume of overall text manual processing by 96%.

The Financial Reporting and Auditing Agent with Net Knowledge, (FRAANK, 2005) (Bovee 2005) analyses the semi-structured balance sheet income statements, cash flows statements and disclosure notes of the Form 10-Q and 10-K. External web information sources are also queried to extract stock prices\(^\text{10}\) and earnings per share\(^\text{11}\), which when integrated with filing financial statements allow financial analysis, the calculation of accounting ratios, metrics and z-score (measure of bankruptcy risk). Soft information such as client, auditor and report type is also mined from the 10-K and auditors report along with SOX 404\(^\text{12}\) for indicators of financial control effectiveness. FRAANK represents a newer version of the earlier EdgarScan (Ferguson 1997), from Price Waterhouse Coopers and EDGAR Agent financial statement parsers. Accounting figures are identified and extracted by matching labels of line-items against tag synonyms from the XBRL Commercial and Industrial

\(^{10}\) Quote.com
\(^{11}\) Quickens.com
\(^{12}\) Sarbanes–Oxley Section 404: Assessment of the adequacy of the company's internal control over financial reporting
Semantic Paths in Business Filings Analysis

taxonomy (C&I) 2000\textsuperscript{13}. Training sets are used to improve parsing and tagging logic (i.e. matching labels to the synonym tags of the C&I taxonomy). Strong pattern matching capability for filings structure identification and intelligent natural language text interpretation, are provided with Perl based heuristically generated regular expressions. Learning over time is accommodated through knowledge base synonym addition based upon exceptions found. Where financial statements cannot be matched to taxonomy tags FRAANK’s sentence classification is used to identify possible deficiencies with, or extensions to the taxonomy. The use of machine translation to automate acquisition of new financial synonyms and improve accuracy was suggested by (Bovee 2005). Performance statistics for 10-K processing, reported overall reliability for capturing accounting figures and their labels in the filing body, as 94.7% for the training dataset, and 88.4% for the test dataset. The most recent version of FRAANK marks up extracted financial statements using the XBRL 2.1 taxonomy.

More recently the EDGAR Extraction System (EES, 2006) (Grant 2006) extracted pro forma\textsuperscript{14} net income and earnings per share information, along with options fair value\textsuperscript{15} from the financial statement and disclosure sections of a 10-K. A non-annotated domain specific training corpus and knowledge experts were used to develop knowledge based rules (described as analysis algorithms), used in turn to support extraction wrapper development. A training corpus was analysed using the CMU toolkit\textsuperscript{16} for Statistical Language Modelling to construct a bi- and tri-gram language model of relevant word phrase frequencies. Using term frequencies, terms were compared with terms from the Statement of Financial Accounting Standards (SFAS) No. 123 guidance document for reporting stock options based compensation and the list enhanced with similar terms. Review by domain experts refined their language model into forty relevant terms, which were incorporated into regular expressions and used to target search and extraction. EES processes 10-K’s to produce a financial tagged file (FTF) containing financial statements and a text file (TF), containing

\begin{itemize}
\item[$\textsuperscript{14}$] Pro-forma income statement provides insight into ability to generate revenues in the short and long terms
\item[$\textsuperscript{15}$] Accounting concept used as certainty of asset (or liability) market value in the absence of an established market price or an inability to determine one.
\item[$\textsuperscript{16}$] http://www.speech.cs.cmu.edu/SLM/toolkit.html
\end{itemize}
disclosure and untagged footnotes. Expression patterns found to produce the best results were those based on keyword search with term associations, e.g. extracting stock options pro forma and fair value information used “as reported | pro.forma*net” and for assumptions “risk-free|dividend yield|volatility|expected life”. The TF was searched using these patterns in combinations such as “fair.value” with “as reported” and if “as reported” appears in the same block, the text block will be written to report. Likewise “risk.free” will only be written to report if a “%” appears in the same block. Overall precision, recall and F-measures reported for stock option and fair value information extraction results, were 72.62%, 82.71% and 77.34% respectively. To improve accuracy and better cater for specific language domains, Grant recommended the use of corpus machine learning and enhancing inter-rater annotator agreement. To cater for emergent financial standards, extraction from XBRL was also recommended (Grant 2006).

The executive compensation retrieval system (ECRS) (Ding 2006), was developed to assist analyst investor investigate corporate governance. SEC DEF 14A proxy filings were mined for executive compensation plan data contained in the filings summary compensation table and free text areas using senseNet$^{17}$, a variation of the Hidden Markov Model. senseNet provided pattern matching (with regular expressions), word classification and context analysis. A personal name ontology constructed from the US Census Bureau was used for filings content analysis. Pattern matching, the personal ontology and context analysis were merged to providing prediction of word categorisation using probabilistic weights and confidence levels. A retrieval rate of 81% from the summary compensation table was reported. The use of WordNet to expand the positions of the personal ontology and XML topic maps to keep it comprehensive were mentioned as future work.

Midas extracts and aggregates facts from structured and unstructured SEC (Form DEF 14A, 10-K, 10-Q , 8-K, 13F, SC, 13D) and Federal Deposit Insurance Corporation (FDIC) periodic filings, to help build a network of connections between financial firms that: i) identify critical banking hubs as part of systemic risk analysis; and ii) provide individual entity (e.g. key executives, insider transactions, relationships to other companies) drill down capability (Hernandez 2010). Midas maps and fuses executive data, insider transactions

$^{17}$ http://www.sensenet.com/
information and company relationships, into a pre-defined entity centric schema. Entity resolution is used to identify real-world entities across all. Co-lending relationships are used to construct a network (graph) of relationships between banking institutions, and network analysis determines the institutions that pose the greater systemic risk. Loan co-investment pair-wise relations are scored and aggregating graph edges for each bank provides the total pair-wise picture across all loans. Representing the relationships in a lending adjacency matrix allows centrality measure calculation from the matrix’s principle eigenvector/eigenvalue, which provides the loading factor of each bank on the lending network. Thresholds applied to the matrix remove banks that are minimally active. The banks centrality scores can be used by regulators to rank banks in terms of system risk contribution. Scalable rule based IE on unstructured text was performed using SystemT\textsuperscript{18} and entity resolution used to integrate unstructured personnel and company data, as part of lending exposure analysis. SystemT also provided the object model repository for structured analysis. Extraction from the DEF 14A filings biographies reported 91% precision and 51% recall and their annotations 87% precision and 49% recall respectively. Entity resolution received an 82.29% recall for people matched to biographies based on CIK\textsuperscript{19}. Post tuning entity resolution recall increased to 97.38% and precision measured using data sampling, was reported as close to 100%.

(Chakraborty 2010) researched the extraction of pension plan information from the footnotes of 10-K statements. Lacking any classification standard the footnotes were restructured and a hierarchical clustering algorithm applied to semi-automatically create a taxonomy structure. The SFAS\textsuperscript{20} 87 and 158 guidelines were used to create a list of pension domain specific terms and address terminological variations. To assist hierarchical formulation historical data taxonomy structures were compared to the XBRL US-GAAP\textsuperscript{21} and found to contain additional terms in different hierarchical locations. Taxonomy differences were attributed to company reporting trends which added greater levels of aggregated information, new terms and sections addition to the footnotes, as opposed to the

\textsuperscript{18} http://www.almaden.ibm.com/cs/projects/systemt/
\textsuperscript{19} Central Index Key used by the U.S. SEC to identify the filings of a company, person, or entity.
\textsuperscript{20} Statement of Financial Accounting Standards issues by the U.S. Financial Accounting Standards Board (FASB)
\textsuperscript{21} Generally Accepted Accounting Practice
disaggregated structure followed by XBRL. The ability to map pension disclosure data to taxonomy tags was used to evaluate taxonomy comprehensiveness. Overall parsing module evaluation across pension headers, section identifiers and line items reported a success rate of 97% on training data sets and 95% on test data.

The EDGAR filings extraction systems presented (cf. Table 2-2) can be categorised in information extraction terms, as KE (Leinnemann 2001; Gerdes 2003; Bovee 2005; Grant 2006; Hernandez 2010), ML (Chakraborty 2010; Hernandez 2010), or containing elements of both, could be considered as hybrid approaches (Ding 2006; Grant 2006; Stampert 2008; Hernandez 2010). The majority achieve financial item or segment extraction solely using regular expression or in combination with statistical or natural language processing methods (Leinnemann 2001; Bovee 2005; Grant 2006; Stampert 2008; Chakraborty 2010; Hernandez 2010). Overall their solution approaches can be said to be equivalent to the use of template slots. Apart from (Chakraborty 2010), each system relies upon on the use of domain heuristics (Leinnemann 2001; Bovee 2005; Grant 2006; O’Riain 2006; Stampert 2008; Chakraborty 2010) with domain task specific lexical resources such as regulatory guidelines, standards and taxonomies, utilised to both create, and augment term and synonym lists for use in pattern detection (Gerdes 2003; Bovee 2005; Grant 2006; O’Riain 2006; Chakraborty 2010). Taxonomies and standards are useful for synonym identification where term variations are readily identifiable e.g. “current liabilities = current total liabilities”. Similarly lexical resources such as WordNet are usable where domain language models are comparable to general language, but where domain discourse introduces considerable variance e.g. for an investor company “products” indicate a revenue stream and products can be also referred to as “saleable items” and/or “goods”, domain heuristics will be necessary for synonym identification.

An interesting approach used to enhance extraction accuracy was the use of term pairing and term proximity. EASE implemented term pairing with hierarchical (pair wise) dependency for balance sheet extraction (Stampert 2008), e.g. assets follow assets and liabilities. EDGAR2XML paired term sets by first defining terms to locate relevant topic section areas in documents before applying a second set of keywords to identify targeted subjects associated with the topics (e.g. Y2K as topic and imbedded as subject) (Gerdes 2003). More tightly coupled keyword search with term associations was used by the EDGAR
Extraction System based on term proximity, e.g. the occurrence of associated terms “fair.value” and “as reported” occurring in proximity with the filing, used to identify relevant text blocks containing stock options fair information (Grant 2006).

In terms of evidence of language models used to help target extraction against, term pairing served as the de-facto model for (Gerdes 2003; Grant 2006; Stampert 2008). Others developed concept models and schemas for data mapping and fusion (Bovee 2005; Hernandez 2010) or simply relied on object representations provided by programming languages (Leinnemann 2001). ECRSs’ name ontology is the closest evidence of any formal representation model used to drive extraction (Ding 2006). While the language models are useful for syntactic rule enhancements, none focus on the provision of any additional semantic interpretative capability, or some means of revisiting extracted information within its originating context. Loss of subtlety can cause an analyst revert to the original financial statements to determine why a company, for example made particular decisions (Debreceny 2001). Direct hyperlinking of specified financial line items provided directive search (for a decision maker) from a defined set, as opposed to the next item selection for financial statement analysis in a sequential search, being found to impact the decision time and predictions made (Dull 2003). A more recent development in the area is entity-centric visualised output allowing abstracted insight such as that for systemic risk and/or detailed drill down (Hernandez 2010), from financial fact based relationships or taxonomies that provides hierarchical formalization to simplify terminology and structure challenges (Chakraborty 2010).

An ontology modelled on the application domain task would provide the conceptual schema to drive extraction, the ability to introduce and apply semantic interpretation and with an inherent hierarchy, a framework that supports directive search and access to a semantically enhanced data set.
### Table 2-2 Addressing Information Extraction Task Challenges from EDGAR Filings

<table>
<thead>
<tr>
<th>Text Type</th>
<th>System Name</th>
<th>Extraction Task Supported</th>
<th>Solution Approach</th>
<th>Supporting Extraction Model</th>
<th>Lexical Resource Engineering</th>
<th>Output Representation Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-Structured</td>
<td>EASE (Stampert 2008)</td>
<td>Balance sheet Sections</td>
<td>Vector space; Xpath; Reg. Exp.</td>
<td>Hierarchical pairwise search terms</td>
<td>Term synonym list&lt;sup&gt;22&lt;/sup&gt;</td>
<td>HTML, ASCII</td>
</tr>
<tr>
<td>Semi-Structured</td>
<td>EES (Grant 2006)</td>
<td>Stock options, pro forma and assumptions, fair value assessment</td>
<td>Statistical(n-grams); Reg. Exp.</td>
<td>Relation, keyword term association and proximity</td>
<td>Term List; SFAS 123</td>
<td>Structured Report, XML</td>
</tr>
<tr>
<td>Semi-Structured, unstructured</td>
<td>EDGAR (Gerdes 2003)</td>
<td>Y2K issues</td>
<td>Keyword search</td>
<td>Hierarchical functional dependency; Tiered search, DOM</td>
<td>EDGAR filer spec.</td>
<td>Text blocks, ordered</td>
</tr>
<tr>
<td>Semi-Structured, unstructured</td>
<td>ECRS (Ding 2006)</td>
<td>Executive compensation data for investigating corporate governance</td>
<td>HMM; Reg. Exp.</td>
<td>Not specified, but direct table mining used</td>
<td>SenseNet Personal Ontology</td>
<td>HTML</td>
</tr>
<tr>
<td>Unstructured</td>
<td>Midas (Hernandez 2010)</td>
<td>Systemic risk analysis of banking hubs</td>
<td>Reg. Exp.; Adjacency matrix; Graph</td>
<td>Relational database</td>
<td>SystemT</td>
<td>Visualised graph</td>
</tr>
<tr>
<td>Unstructured</td>
<td>(Chakraborty 2010)</td>
<td>Pension planning taxonomy</td>
<td>Clustering Algorithm</td>
<td>Not specified</td>
<td>SFAS 87, SFAS 158; XBRL Taxonomy</td>
<td>Taxonomy</td>
</tr>
<tr>
<td>Unstructured</td>
<td>AWB (O’Riain 2006)</td>
<td>Competitive analysis</td>
<td>Reg. Exp. based NLP</td>
<td>Ontology</td>
<td>Term synonym list, taxonomy</td>
<td>In-text visualised ontology hierarchy, text report</td>
</tr>
</tbody>
</table>

<sup>22</sup> Literature specifically distinguishes between terms and term synonyms
2.2.2 Extraction of Business Information from Other Web Sources

The previous section discussed the use of EDGAR filings, probably the most widely known and used web accessible business data catalogue. The filings are part of a more widely available business ecosystem of web based information sources that span news syndicators, competitor web sites, other governmental information and financial institutions. Across the literature there is ample evidence of both handcrafted (Bovee 2005; Gao 2005; Schumaker 2009) and automatically generated (Crescenzi 2001) wrappers that extract data from web sources. As noted in Section 2.1.1 the decade has seen an increasing trend towards the use of customisable IE architectures for semantic annotation, indexing, and retrieval platforms as a cost effective measure for specific information extraction requirements.

Financial news articles and feeds are a popular source of information that can be used to reinforce analysis from consolidated reports, help identify new or more holistic insights (Bovee 2005) or identify entirely new information (Saggion 2007). SCISOR (Jacobs 1990) extracts corporate merger and acquisition data from on-line financial news stories using a combination of top down conceptual interpretation, with bottom up linguistic analysis. Similarly, joint ventures and company agreements to investigate business relationships and market trends were extracted by FASTUS (Appelt 1993). The natural language processing framework LOLITA (Large-scale Object-based Linguistic Interactor Translator and Analyser) was used by (Costantino 1996) to extract company restructuring and general macroeconomic information from financial news based on financial activity templates. The Flexible Information extRaction SysTem (FIRST) (Conlon 2007) automatically extracts, structures and semantically tags corporate earning facts from Reuters and Wall Street Journal news feeds with linguistic analysis. Training data identified extraction patterns and WordNet used for lexical semantic relations. Extraction rules identified financial status ‘optimal proximity’ to a financial term (e.g. sales declined, declining sales), n-grams their instances, and a keyword in context (KWIC) index relationships between potential keywords. GATE (Cunningham 2002; Bontcheva 2004), a framework and graphical development environment has gained popularity with its extensible facility for development and deployment of language engineering components and resources. GATE has been used to support ontology based information extraction (OBIE) for business applications (Saggion 2007) and sub-domains such as market monitoring and technology watch (Maynard 2005),
along with natural language support for information integration within BI environments (Maynard 2007). The Knowledge and Information Management system (KIM) is an automatic semantic annotation framework allowing various IE modules (based on GATE), semantic repositories and information retrieval engines, and on-line front ends to be plugged in. Breaking financial news is used to source stock market prediction using a statistical approach for text analysis (Popov 2003). On-To-Knowledge is a content driven knowledge management platform based on an evolving handcrafted ontology, focused on web sources for content skills management, customer relationship management and virtual enterprise (Fensel 2000). The platform comprises an ontology editor, inference engine and user interface, that allows access to extracted content. A recent development is the large scale Information extraction and Integration Infrastructure for supporting financial decision making (also using the acronym FIRST)\(^{23}\), that is developing methods such as sentiment classification and semantic feature extraction using machine learning, qualitative modelling and visualization techniques to support financial decision making. Mining of online social sources such as blogs and bulletin boards, to support market surveillance, risk management and online retail brokerage and investment management use cases, is planned.

For further reading on semantic annotation for knowledge management we refer the reader to a survey by (Victoria 2006) or for discussion on the methodologies involved with traditional IE systems and wrapper generation systems to (Kaiser 2005).

### 2.3 Business Lexical Resources

Business and financial specifications, standards and taxonomies can be considered as potential sources of vocabularies that support content classification, that either concentrate on content structure, or focus on transactions and business processes. General industrial product and service categorization providing standardised descriptions also used within finance are ISO 4217 (currencies), ISO 3166 (country codes), and UNSPSC (classification hierarchy for products and services). More specific financial classification standards are the ISCS\(^{24}\) for economic data, ISO 10383 for exchanges and trading platforms market

\(^{23}\) [http://project-first.eu/news-events](http://project-first.eu/news-events)

\(^{24}\) International Standard Industrial Classification of All Economic Activities
identification codes and ISO 10962 for financial instruments. Over the last decade a number of business information and financial content XML standards such as FpML, IRML, MDDL, RIXML, OFX, FIX, swiftML, ISO 20022, eCl@ss, RosettaNet, ebXML, NewsML and MarketsML, have emerged that focus on business processes and their transactions (refer to Table 2-3 for summary). RIXML, eOTD and XBRL are notable for their term vocabularies, usable for describing content structure rather than process and transaction. Unlike XBRL however eOTD, NewsML and RIXML would require significant vocabulary extensions to deal with financial information (Pablo Castells 2004). XBRL provides companies with a means to prepare and distribute business reports via the web in a cost effective and universal manner. It allows consolidated balance sheet financial item content mark-up to jurisdictional specific accounting practices and to a lesser extent, the accompanying disclosure sections text blocks. There are also online general language lexical resources such as WordNet (Fellbaum 1998) which provides unique term senses accompanied by explanatory information. As previously noted however (Section 2.2.1), finer grained domain specific knowledge is required for creating usable financial sub-domain lexicons.

Ontologies are another increasingly used source of financial vocabularies. Used to represent semantic information ontologies have previously seen uptake to support the automated acquisition and extraction from business documentation (Embley 1998; Liddle 1999). Popular application areas useful for business process understanding have been:

**business modelling:** for enterprise value chains using the REA ontology to describe a generic accounting systems model (McCarthy 1982), strategic partnerships using the e3value ontology (Gordijn 2003) and the definition of e-business issues and interdependencies, such as product innovation, infrastructure management, customer relationship and corporate finance, through the business model ontology (Osterwalder 2002);

**enterprise architecture:** REA enterprise ontology (Gailly 2008), enterprise ontology (Fox 1992; Grüninger 1995; Uschold 1998);

**e-commerce:** the good relations ontology describing product and service offerings by defining the relationships between web resources, their offerings, prices, terms and conditions and legal aspects (Hepp 2008), which has received broader community interest and uptake (Ashraf 2011);
**e-Collaboration:** supporting collaboration between and within enterprises for performance management related aspects such as KPI generation. With its events interchange description capability REA has also been used as a framework for an event based service system ontology (Sicilia 2009) and has received conceptual extension suggestions to position it as specific business domain ontology (Gailly 2008).

Financial ontologies have emerged that model complex specialisation areas and focus on data content or its management as part of some established task related workflow. Financial domain ontologies for investment fund analysis (Lara 2006), forensic investigation of fraudulent emails (Gao 2005), eBanking stock market ontology (Alonso 2005) or Vanderlindens financial ontology on financial instruments (Vanderlinden 2012), involving parties, processes and procedures in securities handling ontologies, are examples that concentrate on semantic enhancement of content. Other federated ontologies were developed as frameworks that tightly couple data management and a level of semantic interpretation. The financial ontology of the Montific project accommodates a multi-lingual framework for financial control assessment (MONTIFIC 2008), and the Financial Exchange Framework ontology supports the exchange of financial entities (products, transactions, services and monitoring positions) (The IFIP Ltd. 2003). An ontology guided financial fraud management system assists analysts create and manage financial fraud policies through a financial fraud ontology (Yoon 2009), while another fraud ontology knowledge-enables existing information systems structure knowledge management among fraud examiners, intelligence analysis and decision support systems. The Musing project used ontologies to validate next generation semantic based business intelligence solutions that focused on management of credit risk within the context of Basil II\(^{25}\) (Declerk 2008).

\[^{25}\text{Banking laws and regulation recommendations issued by the Basel Committee on Banking Supervision, http://www.bis.org/bcbs/}\]
### Table 2-3: Business and Financial Standards Summary

<table>
<thead>
<tr>
<th>Name</th>
<th>Usage Category</th>
<th>Standards Description</th>
<th>Standards Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO 10383</td>
<td>Classification</td>
<td>Market identification codes</td>
<td>- X</td>
</tr>
<tr>
<td>ISCS</td>
<td>Classification</td>
<td>Economic activities</td>
<td>- X</td>
</tr>
<tr>
<td>ISO4217</td>
<td>Classification</td>
<td>Currencies</td>
<td>- X</td>
</tr>
<tr>
<td>ISO 10962</td>
<td>Classification</td>
<td>Financial instruments taxonomy hierarchy</td>
<td>- X</td>
</tr>
<tr>
<td>MarketsML</td>
<td>Info. Exchange</td>
<td>Product opportunities</td>
<td>- X</td>
</tr>
<tr>
<td>MDDL</td>
<td>Info. Exchange</td>
<td>Market data exchange for financial instrument events</td>
<td>- X</td>
</tr>
<tr>
<td>ebXML</td>
<td>Info. Exchange</td>
<td>Electronic business information between trading partners</td>
<td>- X</td>
</tr>
<tr>
<td>eCl@ss</td>
<td>Classification</td>
<td>Product description between customers and suppliers</td>
<td>- X</td>
</tr>
<tr>
<td>UNSPSC</td>
<td>Classification</td>
<td>Products and services</td>
<td>- X</td>
</tr>
<tr>
<td>RosettaNet</td>
<td>Info. Exchange</td>
<td>eBusiness processes between supply chains</td>
<td>- X</td>
</tr>
<tr>
<td>eOTD</td>
<td>Info. Exchange</td>
<td>Individuals, organizations, locations, goods, services, processes and regulations</td>
<td>X -</td>
</tr>
<tr>
<td>FpML</td>
<td>Info. Exchange</td>
<td>Derivative negotiation</td>
<td>- X</td>
</tr>
<tr>
<td>ISO 20022</td>
<td>Info. Exchange</td>
<td>Interoperable financial messages between institutions, used in securities trading</td>
<td>- X</td>
</tr>
<tr>
<td>RIXML</td>
<td>Info. Exchange</td>
<td>Schema for tagging investment research content</td>
<td>X -</td>
</tr>
<tr>
<td>swiftML</td>
<td>Info. Exchange</td>
<td>Describing financial messaging exchange, general business processes</td>
<td>- X</td>
</tr>
<tr>
<td>OFX</td>
<td>Info. Exchange</td>
<td>Business banking bill payment, investments, financial planning and insurance</td>
<td>- X</td>
</tr>
<tr>
<td>FIX</td>
<td>Info. Exchange</td>
<td>Trade related messaging between banks, brokers, vendors for financial instrument trading</td>
<td>- X</td>
</tr>
<tr>
<td>IRML</td>
<td>Info. Exchange</td>
<td>Investment risk monitoring and reporting</td>
<td>- X</td>
</tr>
<tr>
<td>XBRL / IFRS</td>
<td>Info. Exchange</td>
<td>Consolidated financial item tagging</td>
<td>X -</td>
</tr>
<tr>
<td>NewsML</td>
<td>Info. Exchange</td>
<td>News exchange</td>
<td>- X</td>
</tr>
<tr>
<td>SFAS</td>
<td>Classification</td>
<td>Statement of Financial Accounting Standards e.g. SFAS 123, 87, 158</td>
<td>- X</td>
</tr>
</tbody>
</table>
Knowledge representations that concentrate on content classification and structure whether presented as some standard, taxonomy or ontology, are readily usable as rich vocabulary sources for their particular application areas. For application purposes such as financial metric generation (Bovee 2005), that derive information from consolidated financial statements, there is a probability that vocabularies such as the general purpose financial reporting Commercial and Industrial Companies (C & I)\textsuperscript{26} taxonomy, may already exist and be sufficiently aligned to generate domain terminology from. Semantic technology and its web counterpart, the Semantic Web (Berners-Lee 2001), have also used lexical resources to assist making underlying data sources semantic representation explicit, with the use of an ontology. The Musing project used XBRL taxonomies as background semantic resource and made them semantically explicit based on information extracted from business reports for financial risk management (Declerk 2008). XBRL and Semantic Web areas were meshed by producing OWL\textsuperscript{27} based ontologies of XBRL Schemas that map XBRL to RDF instance data, as a mechanism for semantically enhancing XBRL based financial data.

Where an application area is sufficiently specialised (Leinnemann 2001; Grant 2006; Lara 2006; MUSING 2006; O’Riai 2006; Declerk 2008; Stampert 2008; Yoon 2009), such that vocabularies do not exist, or are useful only to define higher level concept definition, domain knowledge and heuristics are required to define the lower level domain lexicons and terminology. Where the area looks to the use of softer intangible heuristically based information, derived from experience and insight (e.g. competitive analysis), domain knowledge is essential. The ontology with its semantic hierarchy can also be employed as a framework to assist with information structuring, integration and semantic interpretation (FF Poirot 2002; Alonso 2005; Lara 2006; Declerk 2008; MONTIFIC 2008; Yoon 2009). A summary of the financial and business ontologies discussed is presented in Table 2-4 below. Table construction is based on dimensions from the Uschold and Jasper (Uschold 1999) ontology classification framework, with the widening of its ontology role definition to include Grubers category type of task (or domain specific nature) (Gruber 2003).

\textsuperscript{26} http://www2.xbrl.org/in/npmpxbrl.aspx?id=336
\textsuperscript{27} Ontology Web Language, http://www.w3.org/TR/owl-features/
<table>
<thead>
<tr>
<th><strong>Name / Origin Description</strong></th>
<th><strong>Application Purpose</strong></th>
<th><strong>Ontology Role</strong></th>
<th><strong>Implementation Actors</strong></th>
<th><strong>Supporting Technology</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanderlinden's Finance Ontology, (Vanderlinden 2012)</td>
<td>Financial instruments, involved parties, processes and procedures in securities handling</td>
<td>Modelling of securities instruments, actors and related tasks</td>
<td>Business analysts, modellers</td>
<td>Top Braid composer, OWL</td>
</tr>
<tr>
<td>FFO, <em>Financial Fraud Ontology</em>, (Yoon 2009)</td>
<td>Financial fraud knowledge management system to create and manage financial fraud policies</td>
<td>Modelling of fraud analysts knowledge (credit card, cheque, online and ATM)</td>
<td>Fraud analysis</td>
<td>Protégé, OWL</td>
</tr>
<tr>
<td>FFD, <em>Fortune from the Dead</em> (Gao 2005)</td>
<td>Forensic investigation of fraudulent emails</td>
<td>Facilitating of structured analysis</td>
<td>Forensic/fraud analysts/ICT</td>
<td>DOGMA, AKEM</td>
</tr>
<tr>
<td>Investment Funds Ontology (Lara 2006)</td>
<td>Investment funds analysis</td>
<td>Modelling and funds analysis framework</td>
<td>Investment analysts</td>
<td>OWL</td>
</tr>
<tr>
<td>MUSING, <em>Multi-industry, Semantic-based next generation business INtelligence</em>, (Declerk 2008)</td>
<td>Business Intelligence</td>
<td>Knowledge representation of credit risk management (Basil II)</td>
<td>Business analysts</td>
<td>OWL</td>
</tr>
</tbody>
</table>
2.4 Semantic Web and Open Data

The last few years has seen the emergence of a web of data fuelled by Open Government transparency initiatives that have resulted in significant amounts of public sector information being made freely available, for use and redistribution without restriction. Responding to requests for greater transparency and reduction of administrative burden, collected financial, economic and legal data sets have received availability mandates (European Information Society 2003) for integration and innovative reuse, to serve purposes such as new products and services development (European Information Society 2006), or for companies to become better informed on their own business circumstance, or those of stakeholders (Gross 2009).

Known as Open Data, notable economic and financial examples are EuroStat (EU), data.gov.uk (UK), sec.org, data.gov, recovery.org (US), the World Bank and IMF (International). Open Data can also include generalised business news, marketing information and competitor data available from an assortment of web sites. Examples are: DBpedia\(^{28}\), a community initiative that extracts general-purpose information from Wikipedia and makes it freely available to query; Freebase\(^{29}\), harvests information from multiple sources and allows querying of data about people, places and organisation. The SECs EDGAR repository provides public access to corporate filings in multiple formats and is arguably the more widely known source of business and financial Open Data. Presenting potential for re-use, EDGAR has been heavily exploited and re-used by third party information providers and aggregators such as Hoovers, Morning Star, Yahoo! Financial and EDGAR Online, to provide enhanced market driven value add service offerings (cf. Section 2.2).

Using the web to publish Open Data has made information more accessible, but varying information formats (e.g. XBRL, XLS, CSV, PDF, RDF, text) ensure that consumption remains difficult. Examples of standardisation attempts to enhance quality and reduce consumption costs of Open Data Sets are the use of RSS-CB\(^{30}\) by central banks to publish

\(^{28}\) [http://wiki.dbpedia.org/Datasets](http://wiki.dbpedia.org/Datasets)


exchange rate data and the Statistical Data and Metadata Exchange format\textsuperscript{31} (SDMX) used by Eurostat, is also under consideration for use by the OECD, World Bank and United Nations.

Linked Data\textsuperscript{32} based on the W3C Resource Description Framework (RDF) standard, caters for multiple formats by providing a common interoperable format and model for data linking and sharing on the web. The Linking Open Data (LOD) Cloud\textsuperscript{33} represents a large number of these interlinked RDF data sets within the wider ecosystem that are being actively used by industry, government and scientific communities. RDF is the basic machine-interpretable information representational format used by the Semantic Web that represents a general schema-less and self-describing method for encoding graph-based data, which does not follow a predictable structure. Graph labels and relations describe the meaning of the data. Data and facts are specified as statements, also called triples, represented by the atomic constructs of a subject, predicate and object. Linked Data uses these RDF triples as the fundamental building blocks, with identifiers used to uniquely identify concepts and relationships. The deployment practice adhered to, and promoted by, the Semantic Web community for good data publishing, can be classified along a rating scheme\textsuperscript{34} which advises to:

\begin{itemize}
  \item[i)] make data available on the web in any format
  \item[ii)] make the data available as structured data
  \item[iii)] use non-proprietary formats
  \item[iv)] use W3C standards such as RDF when publishing data, so that others can point to your data
  \item[v)] link your data to other people’s data (via URIs) to provide context.
\end{itemize}

Publishing data which adheres to these basic principles provide a common standards based model for data access and inter-linkage, ensuring ease of access and usage of the data for consumers. Where data is made available in non-proprietary formats, wrappers and

\textsuperscript{31} http://sdmx.org/
\textsuperscript{32} http://linkeddata.org/
\textsuperscript{33} http://lod-cloud.net/
\textsuperscript{34} http://www.w3.org/DesignIssues/LinkedData.html
converters, that map information from their source format into RDF triples, can be employed on an as-needed basis, e.g. the extractor used by the Road Runner project to automate data extraction from large web sites\textsuperscript{35}.

For businesses, integrating information is a key concern. Most domains have their own particular vocabulary which provides distinct descriptions and definitions of concepts used, their schema and recommendations on how to model. Semantic Web vocabularies support the mapping of vocabulary identifiers, supporting alignment and schema level shared understanding (O’Riain 2012a). XBRL on the other hand (cf. Section 2.3), provides direct reference capability into its instances documents, for financial instrument property and value extraction. However comparing information between alternate XBRL taxonomies requires extension or alternate representation (Wenger 2011; O’Riain 2012b). The principle of ontology use for semantic inter-operability and financial taxonomy mapping between XBRL formats has recently been proposed (Wenger 2011). But the financial statement sections (foot notes, management discussion or non-financial disclosures) were omitted and left to modeller discretion.

IR has also exploited the ontology across a diverse range of topic areas such as enhanced search (e.g. medical (Abulaish 2005), employment (Maynard 2004), national security (Sheth 2003)), and interoperability using the framework that a common ontology can provided for heterogeneous information integration (Victoria 2006). The Semantic Web model proposes the use of domain ontologies and their semantic content to assist with document annotation (Berners-Lee 2001). Semantic Web has also renewed interest in OBIE, involving the use of ontologies combined with information templates to assist information analysis (Bontcheva 2004). This more recent step views template slots as instances of the domain ontology concepts or relationships, where the template can lend itself to further addition such as reasoning, casting the template as a frame in the tradition of Shank (Schank 1977). Automating the annotation process using linguistic analysis enables new types of applications such as highlighting, indexing and retrieval, categorization, the generation of more advanced metadata, and traversal between unstructured text and relevant structured legacy knowledge (Kiryakov 2004).

\textsuperscript{35} http://www.dia.uniroma3.it/db/roadRunner/
The current lack of a global entity references for semantic financial information on the web, impacts both RDF and XBRL. Entity identification is a key requirement for data combination and even for common structured data formats (such as XBRL and RDF), different sources can vary in how the same fact is stated (Wenger 2011). Follow-on activities such as querying, analysis and reporting are also highly dependent on entity attributes identification.

Despite the fact that Semantic Web formalisms have been identified as a natural selection for heterogeneous data source integration (Bao 2010), evidences of leveraging XBRL with Open Data are largely confined to academic use cases. With the prominence and uptake of both standards set to increase, there is sufficient basis for the two standards organizations, XBRL International36 and the W3C37 to collaboratively engage business, government, technical and social communities to address the communication and interconnect issues. Both XBRL and W3C organisations have expressed interest in jointly addressing the absence of an entity naming schema (W3C and XBRL 2009; XSB 2010).

2.5 Summary

In terms of IE task challenges, the EDGAR filings represent a mixture of structured financial items and unstructured textual financial disclosure sections. Previous automated extraction using business information schemas38 faced difficulties relating to limited tagging, formatting errors, content and terminology inconsistency (Gerdes 2003), along with a lack of uptake, consistent application and effective use (Debreceny 2001). With the adoption of the XBRL standard based on jurisdiction specific accounting principles, these issues have been largely addressed for consolidated financial reporting. Despite XBRL facilitating uniform generation and automated extraction from consolidated financial filings, the textual disclosure notes, lacking structure and common format, remain problematic (Grant 2006).

Domain specific language found within the disclosure sections presents challenges for: the solution approach adopted; domain linguistic definition and the development of

36 http://www.xbrl.org/Home/
38 Here taken to mean some form of conceptual map that suggests structure and organization whether it be hierarchical or not, e.g. vocabularies, taxonomies or ontologies.
domain lexical resources. The systems investigated reflect knowledge engineering, machine learning and hybrid solution approaches. With the acceptance of shallow finite state analysis, pattern matching and shallow NLP analysis techniques, have emerged as popular and good enough light weight extraction techniques for rule based approaches.

Across the filings extraction systems there was no evidence of domain linguistic definition adhering to some specific model. Language resources use to support extraction were financial filings, sector specification and taxonomies from which term lists were developed. In some instances terms were hierarchically organised to assist search efforts (Gerdes 2003; Stampert 2008), or relational databases employed to provide the logical model (Bovee 2005; Hernandez 2010). The closest found ‘ontology usage’ was the development of a visualised graph based on the construction of an adjacency matrix for systemic risk analysis of banking hubs (Hernandez 2010). Overall the EDGAR systems extract either:

i) entire report sections

ii) financial statements as facts or

iii) financial statements used for calculation based analysis, such as ratio generation and systemic analysis

None report adherence to a repeatable methodological approach to define domain linguistic that additionally attempts to represent possible information interest. The wider financial literature does however contain examples of successful ontological modelling and application that target specific financial sub-domains, listed in Table 2-4. Characteristic of these is the central role played by the ontology in supporting information management and analysis.

Investigating whether text extraction from financial disclosure notes can be automated is foremost dependent on formalising the information requirement and modelling of the information space from a linguistic and semantic perspective. An ontology modelled on the application domain task would provide the conceptual schema to drive extraction, a level of semantic interpretation and a unifying semantic platform, from which a directive or sequential approach to financial statement analysis is possible (Dull 2003).
Development of a competitive analysis ontology, and its application to EDGAR filings disclosure sections, based on formal linguistic representation and modelling of the information space, offers the possibility of automating the information requirement for analysis.

We next introduce the relationship between knowledge, knowledge capture and ontologies, necessary as background information before describing in detail, the formulation of domain linguistics.
Chapter 3

Knowledge Capture and Ontologies

Throughout our research and literature review, information and data were found in use interchangeably and knowledge where encountered, was assumed as some form of background knowledge. As our investigations leverages the use of domain knowledge (whether as experience, insight, opinion or operational know-how), fundamental understanding of the relationship dynamic between an ontology and knowledge articulation, is warranted as preparatory background for modelling domain knowledge (Chapter 4) and Semantic Paths (Chapter 5).

Section 3.1 provides the Business Information System audience with an ontology definition that aligns with knowledge derivation and representation. Section 3.2 introduces and describes these knowledge types and view fundamentals. Section 3.3 then progresses to applicable knowledge capture paradigms and their use in enhancing knowledge understanding. Conceptual foundations and taxonomies that contribute to the knowledge definition and knowledge articulation discussion are taken from information science (Polanyi 1967; Nonaka 1994; Nonaka 1995; Alavi 2001). Within this context the ontological contribution to knowledge capture and problem understanding are then introduced. Section 3.4 introduces practical consideration for ontology use in formalising domain linguistics. Finally Section 3.5 engages in discussion and concludes with a summary.

3.1 Ontology Definition

Nenches provided one of the earlier definitions of an ontology as that which “defines the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to this vocabulary” (Neches 1991). Guarino however is accredited with the widely referenced definition of an ontology as being

“To know that we know, what we know, and to know that we do not know what we do not know, that is true knowledge.”
Nicolaus Copernicus
“an explicit specification of a conceptualization” (Gruber 1993; Guarino 1998) with conceptualization being interpreted as “an abstract model of some phenomenon in the world” and consists of entities and their relationships, that hold within a domain of interest (Genesereth 1987). The relationships themselves can be interpreted as role(s) played between objects (Halpin 2001). As an artefact an ontology has a shared vocabulary, describing entities in a domain of interest, in addition to a set of assumptions about the intended term meanings of the vocabulary (Guarino 1998). The ontologies universe of discourse is a declarative formal vocabulary about known knowledge and represents an implicit or explicit commitment to a conceptualisation (Fridman-Noy 1997). Gruber more restrictively defines the semantic agreement between entities and relationship’s necessary for agent communication, as a commitment (Gruber 1993). Further formalising the relationship between ontologies and knowledge, an ontology was noted as “providing the means for describing explicitly the conceptualization behind the knowledge represented in a knowledge base” (Schreiber 1995). Neches additionally noted that ontology construction comprises not only constituent terms, but also knowledge, that can be derived from the combination of terms (Neches 1991). Adding domain specific knowledge at the lower levels, or additional general concepts at the higher level, to existing ontologies, is another means of creating additional ontologies or knowledge bases (Swartout 1996). In general terms an ontology that adheres to greater restrictive modelling on domain semantics is referred to as a heavy weight ontology, while those without axiomatic constraints, are referred to as lightweight (Gomez-Perez 2004).

For our research we adhere to a composite ontology definition that support our need for:

i) domain knowledge representation, catered for with Schreiber’s observation that the ontology provides the conceptualisation behind the knowledge representation in a knowledge base (Schreiber 1995), and also provides the context for follow on

ii) knowledge derivation, noted by Neches that an ontology additionally comprises knowledge that can be derived from the combination of its terms (Neches 1991).
Both principles will be used to underpin our notion of a **Semantic Path** (cf. Section 4.4) and develop the ontology base and commitment layer of the Competitive Analysis Ontology (cf. Section 5.1).

### 3.2 Knowledge Fundamentals

Within the IT and knowledge management literature the prevalent view of knowledge is that of a conceptual hierarchy derived from information, and information in turn from data (Davenport 1998b). Data is commonly thought of as simple facts and information as processed data. Information when authenticated through interpreted, put into context or has meaning added becomes knowledge (Tuomi 1999)\(^{39}\). Tuomi alternatively argues that the reverse hierarchy of knowledge-information-data should be considered. Knowledge he proposes, must exist before information can be formulated and data emerges from information where meaning, structure or semantics have been added and used for information representation (Tuomi 1999). For Tuomi raw data as such does not exist but can be produced detached from any meaning, its constituent pieces already influenced by the thoughts and knowledge processes that led to its identification and collection (Tuomi 1999; Alavi 2001). External to the *knower*, knowledge therefore does not exist but is rather the result of individuals cognitive processing of information or inter-personal validation (Earl 1994). For the reverse hierarchy information derives from knowledge that has been articulated in some representational format (Tuomi 1999; Alavi 2001). Implications for this perspective of knowledge are: i) that individuals must share knowledge bases, if the same understanding of information and data is to be had (Alavi 2001), and ii) the instrument used to generate the data determines its meaning (Tuomi 1999).

For Polanyi and Nonaka organizational knowledge revolves around the dynamic relationship between tacit and explicit knowledge\(^{40}\). Polanyi considers tacit knowledge as internalised within the unconscious mind and represents a level of understanding that cannot be externalised or articulated, in effect “we know more than we can tell” (Polanyi 1967). Nonaka and Takuchi on the other hand consider tacit knowledge consists of cognitive

---

\(^{39}\) Although presumption of hierarchy when evaluation along dimensions such as structure, content, utility or accuracy, rarely survives (Alavi 2001).

\(^{40}\) For a more complete listing of knowledge types and taxonomies we refer the reader to (Alavi 2001).
(e.g. personal experiences, mental models, beliefs and perspectives) and technical elements (e.g. know how skills), that although difficult to articulate, can however be converted to explicit knowledge and codified through a social knowledge conversion process (Nonaka 1994; Nonaka 1995). De Leenheer applied this inter-dependent conversion processes of socialisation, externalisation, internalisation, and combination to the ontology evolution processes (community grounding, rendering, alignment and commitment) allowing stakeholder agreement be mapped to detailed system interoperability requirements, and vice versa (De Leenheer 2009b). The application formed the basis of the DOGMA MESS methodology (De Leenheer 2008) (cf. Section 4.5) which was deployed in business organizations and communities (De Moor 2006).

Polanyi requires tacit knowledge as a pre-condition for meaningful focal knowledge\textsuperscript{41}, whereas Tuomi notes that explicit knowledge cannot emerge without some changeable underlying tacit meaning structure\textsuperscript{42}, that allows tacit knowledge become focal relative to the remainder, through inclusion of background knowledge - in effect the socialisation portion of Nonaka’s knowledge conversion (Tuomi 1999). Knowledge emergent through inter-personal validation (Earl 1994) and overlap in knowledge spaces\textsuperscript{42}, is necessary for individuals to understand one another’s knowledge (Alavi 2001) and community knowledge is required to make the socialisation, articulation and externalisation phases of Nonaka’s knowledge conversion process possible (Tuomi 1999). Both viewpoints agree on the importance of a tacit knowledge stock, within which information while inseparable from its background, is none-the-less required for information interpretation (Nonaka 1995). For Davenport the infusion of human cognition allows knowledge refinement from information. Individuals values and beliefs are integral, helping to determine what can be observed, concluded or inferred, knowledge therefore additionally contains judgement (Davenport 1998b).

Tuomi’s reversal hierarchy begins with individual’s knowledge articulation from cognitive effort, to create the meaning structure that then allows tacit knowledge become structured and focal. Inserting information into these pre-defined data structures that

\textsuperscript{41} Knowledge that is used to assist with and improve what is in focus

\textsuperscript{42} Also referred to as knowledge stock, background or knowledge base within the information systems literature
defines meaning, generates data. An implication is that knowledge articulated with automated processing in mind, has to have information meaning represented in sufficiently atomic (data) elements, that there is no further requirement for additional meaning, as part of the processing activity (Tuomi 1999). The reverse hierarchy is representative of data model development as part of an overall knowledge management or related application. The assistance of domain experts (tacit knowledge) is required for problem understanding and model construction (meaning structures), that with technology assist, provides information for consumption, through data element population. Conversely the prevalent hierarchical view adheres to the typical information systems knowledge management application, where IT performs a data processing function that through some dedicated portal with pre-defined meaning, provides the user with information, that when used in context generates knowledge.

3.3 Knowledge Capture

Codification of, and model construction from domain knowledge requires clear problem identification and understanding. Within information systems, Gorry and Morton were among the first to provide a problem solving characteristics framework based on understanding management activities\(^{43}\) as a pre-requisite for effective system design and decision support (Gorry 1971). Their problem (or activity) type classification of structured, semi-structured or unstructured has since become widespread within decision systems theory. Problems that required judgement, evaluation and insight for problem definition are termed unstructured. Characteristic of unstructured problems is their ambiguity of problem definition, lack of formalism for dealing with problem solving phases, and the absence of a routine procedure for dealing with them. Davenport framed the problems associated with analytical capability construction, in term of iterative analytic and decision making processes, where the analytical process was based on the level of question structuring that the underlying data was designed to answer (Davenport 2001; Davenport 2009). For Davenport unstructured questions are unclear, if the question is in fact known, and requires considerable analyst involvement for information need and decision definition.

\(^{43}\) Decision categories were operational control, management control and strategic planning
Structured problems will have these phases catered for by the system, allowing the majority of decision making to be automated (Gorry 1971). Similarly structured questions are those that are organisationally defined, are relatively uncomplicated, require little decision maker interactions and where required data is readily accessible and interpretable (Davenport 2001; Davenport 2009). Where only some of the problem phases can be automated, the remainder will be left unstructured and presented to the user with varying degrees of computational and display support (Gorry 1971). Information system constructed to deal with structured decisions will therefore be different to those constructed to support unstructured decision making in terms of information supported and technology applied (Gorry 1971).

Problem solving can be decomposed into the basic phases of identification, solution and implementation. For decision-centred approaches this involves an iterative hierarchical process with stages of intelligence activity (gathering relevant information), design (analysis of possible courses of action), and choice (selection of course of action), which map to the generalised identification and solution stages (Gorry 1971). As the level of problem understanding and decision making understanding increases, unstructured problems move towards structured decisions, allowing the system takes over and stakeholders concentrate on further unstructured areas, where the more interesting problem lie. The level of structure varies with decision maker involvement, analysis requirements and degree of technology automation possible. The greater the unstructured nature of the question or decision, the greater the need for decision maker validation and establishment of information relevance (Davenport 2001; Davenport 2009).

Competitive analysis demonstrates the characteristics of both frameworks descriptions of an unstructured problem, with each of the problem solving phases also being unstructured. The objective is therefore to move the competitive analysis problem from being unstructured to structured, through the capture where possible of analyst tacit knowledge and judgement (detailed in Section 4.1). To align ontology contribution to the structuring process, the unstructured elements of both frameworks using the categories of problem structuring, knowledge capture and representational requirements, summarised below in
Table 3-1 are discussed.

Problem structuring. Unstructured problems are characterized by knowledge uncertainty, normative element disagreement (e.g. objectives, questions) and cognitive element disagreement (e.g. divergent perceptions) (Hisschemöller 1995). Leveraging uncertainty in the form of stakeholder perceptions, insight, analysis and data interpretation, and its conversion into knowledge are key elements for problem formulation. Both frameworks centre around this notion of a knowledge duality hierarchy, that depends on transformation individual knowledge space data into explicit knowledge that is understood and accepted. Specifically the amount of contextual information necessary to enable individual or group knowledge sharing is at issue (Tuomi 1999; Alavi 2001). Both require stakeholder involvement and assistance.

Knowledge Capture. Decision makers vary from those that have the required information, but need new processing methods to better understand it, to those that have difficulty in establishing need to begin with. The frameworks address such issues through decomposition, segregating those aspects of the decision process or question answering that can be automated through model development and process formalisation. Model development is dependent on interaction with decision makers and modellers. Where full structuring cannot be achieved models that assist develop insight into the relationships of decisions to goals should be provided (Gorry 1971).

Representational requirement. Model development has to reflect the leveraging of individual tacit knowledge stocks. The model has to have an accompanying meaning structure that allows tacit knowledge become structured and explicit with supporting technology. When knowledge is given structure and embedded in artefacts, it can become an information object that can be shared and revisited. Implicit knowledge (such as judgement and inference) can be further brought to bear on structured focal information to derive further knowledge through insight and interpretation.
**Table 3-1 Knowledge Use and Ontologies**

<table>
<thead>
<tr>
<th></th>
<th>Question Answering (Davenport 2001)</th>
<th>Decision Making (Gorry 1971)</th>
<th>Representational Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem Structuring</td>
<td>Analysis requirement unclear</td>
<td>Ambiguity in problem definition</td>
<td>Capability to capture and model tacit knowledge related to know how, process, and end goals.</td>
</tr>
<tr>
<td></td>
<td>Analysis questions require type identification and definition</td>
<td>Clarification required of decision type</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Requires analyst involvement</td>
<td>Requires decision maker involvement</td>
<td></td>
</tr>
<tr>
<td>Knowledge Capture</td>
<td>Question formulation and analytical process structuring</td>
<td>Intelligence, design and choice activity phase decomposition structuring</td>
<td>Capability to flexibly model for hierarchy and particular views.</td>
</tr>
<tr>
<td></td>
<td>Difficulty establishing information relevance and need</td>
<td>Information need known, difficulty with understanding and processing of what’s available, time consuming</td>
<td>Model (and data structures) that support capture of perceived information need and also allows future tacit knowledge introduction for further insight</td>
</tr>
<tr>
<td></td>
<td>Interaction required with decision maker for judgement and evaluation</td>
<td>Interaction with decision makers to define views</td>
<td>Populated model should assist stake holder interaction for information interpretation</td>
</tr>
<tr>
<td>Solution Requirements</td>
<td>Define information need, type of question to support</td>
<td>Formularise problem solving phases</td>
<td>Identify technologies that support problem structuring through information provision, based on interaction with information needs model</td>
</tr>
<tr>
<td></td>
<td>Require model development and formulisation</td>
<td>Require model development and formulisation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Introduce technology support routines</td>
<td>Introduce technology support routines</td>
<td></td>
</tr>
</tbody>
</table>

The ontology assists problem structuring by providing a representational formalism and framework. Ontology development supports the capture of tacit knowledge that helps formalise problem definition and information requirement through context and enhanced semantic meaning capture (cf. Section 5.1). The resulting semantic data structures can be automatically instantiated, and provide information in context (of the original problem) ready for assessment. In this regard the ontology adheres to Tuomi’s reverse knowledge hierarchy and representative of both problem definition and analyst information requirement.

As a *model* with linguistic analysis support, the ontology assists the migration of the unknown (tacit) problem definition and information need from uncertainty and
unstructured-ness to one of definition and structure. In doing so the level of decision making understanding regarding the information requirement increases, leaving the stakeholder dedicate more time to remaining unstructured activities such as developing further interpretation and insight. This point at which the unstructured problem transforms into a structured problem, Mason, refers to as the point of articulation (Mason 1969). The sooner this point is reached the sooner the effectiveness of the decision making capability questions can be improved. Investigation of this point and its impact is outside the scope of this research.

As a framework the ontology facilitates the co-existence of multiple knowledge perspectives (Alavi 2001). The first is knowledge as a hierarchy (discussed in Section 3.1); knowledge as an object, that allows the ontology and its atomic constructs be stored, shared and manipulated (cf. Section 5.2) and; knowledge as a condition of having access to information (cf. Section 7.2), that allows domain stakeholders further exploit the information provided through the model, with the application of their tacit knowledge for further explicit knowledge generation. Although an interesting topic the investigation of the co-existence of multiple perspectives and their impact on information systems design, is also outside the scope of this research.

3.4 Ontology Usage

Considerations towards ontology usage and development adhered to a two stage process. The first sought to establish that an ontology offered an appropriate medium to represent the information disclosures section as an information space (Sections 2.2, 2.3), while the second considered the potential methodologies\textsuperscript{44} for ontology development, from which DOGMA was selected (described in Section 4.4).

Conceptual schemas associated with traditional information systems, allow the modelling of real world application domains and supports the capture of their semantics. When translated to the physical data model (or schema), much of the inherent semantics becomes obscured, resulting in shortcomings for the domains semantic interpretation.

\textsuperscript{44} For detailed discussion on other ontology modelling methodologies we refer the reader to (Gómez-Pérez 2002; 2004).
Understanding the true nature of data relationships necessary to support any follow-on task driven activity (e.g. competitive analysis), then becomes problematic. An ontology supports the capture of domain semantics and provides a semantic interpretative mapping of the information systems semantic domain (Meersman 1999). Overcoming these limitations, the ontology integrates its conceptual and physical models to allow easier semantic exploitation and information space traversal.

Ontology usage within the last decade has seen wide spread acceptance and adoption from an initial information extraction and discovery function, primarily within the medical domain (e.g. the RiboWeb system incorporates ontology constructs for three-dimensional model knowledge representation and storing of structured molecular information (Chen 1997)), to multiple applications across multiple domains, facilitating knowledge sharing and reuse (Natalya 1997). Diverse examples within the business and finance environment are: eProcurement (De Nicola 2008), eCommerce (Fensel 2001), (semantic) extraction, integration and navigation (Sheth 2003; McGuinness 2004), and legal domain (Casanovas 2007). Domains which are increasingly incorporating ontology usage are: business process management (Abramowicz 2009; Filipowska 2009), knowledge management and knowledge process support (Staab 2001), environmental decision-support systems (Ceccaroni 2004), business intelligence (O’Riain 2006; Saggion 2007), market technology watch (Maynard 2005) and emergent semantics (Cudre-Mauroux 2006). Promoting agreement, ontologies are used prescriptively for communication and interoperation standards, promoting a common representation medium for knowledge and the basis of metadata or semantic annotation definition (Sheth 2003). Frameworks such as OntoText’s KIM take advantage of this, offering IE based metadata creation and storage capability (Popov 2003; Gao 2005).

Representing meaning is an important factor in the successful application of ontologies. With agreement on the specification of terms conceptualization, the idea of what an ontology is can become less formal and vague (Studer 1999). The simplest ontologies consist of a simple taxonomy of terms. Meaning is supplied by a single relation (usually the subclass specialisation relationship), which can also be a composite of the other relationship such as ‘part-of’. At the other end of the spectrum are formally rigorous and axiomatised ontologies such as TOVE (Fox 1992). The meaning that an ontology captures
varies in what is being represented and its degree of formality. As the level of meaning increases so too does the restriction on possible interpretation, with a corresponding reduction in the level of ambiguity. The formality of the representation language can vary from natural language to a variety of formal logics. (Uschold 1999) noted four formality continuums of:

- **highly-informal**, where the representation is expressed loosely in natural language
- **structured-informal**, expressed in a restricted and structured form of natural language, which reduced ambiguity by increasing clarity, e.g. the text version of the Enterprise Ontology (Uschold 1998)
- **semi-formal**, expressed in a formally (artificial) defined language; e.g., the Ontolingua version of the Enterprise Ontology or the Semagix project (Sheth 2003)
- **rigorously formal**: where terms are defined with formal semantics, theorems and proofs of such properties as soundness and completeness (e.g. TOVE (Fox 1992)).

Semi-formal as opposed to highly expressive formal ontologies were noted as being the more practical in application (Gruber 2003). Sheth also found this to be the case observing that semi-formal ontologies based upon limited expressive power, that did not claim formal semantics, were found to be the more practical (Sheth 2003). The Semagix project demonstrated that real world applications can be developed with semi-formal ontologies which can easily accommodate incomplete or impartial information, to meet the challenge of information integration, all achievable with limited semantics (Sheth 2003).

<table>
<thead>
<tr>
<th>Consideration: Semi-formal ontologies with limited expressive power and semantics are more practical</th>
</tr>
</thead>
</table>

Ontology construction has been previously considered more of an art form than systematic engineering processing (Jones 1998). A survey of ontology engineering practice observed that the status quo has moved little with “achieved results remaining restricted to a small community of experts affiliated to academia” (Paslaur 2006). Across the ontology engineering methodologies, no single methodology covers all required ontology engineering aspects (Gomez-Perez 2004; Paslaur 2006). The combination of different methods and alternate method engineering approaches, for different stages of the ontology development
 process, includes the provision of selection support for knowledge engineers, as a suggested viable alternative to new methodology construction (Paslaur 2006). In less specific terms no single methodology should be relied upon (Jones 1998) and practical experience to build upon is limited (Perez and Mancho 2003).

Overall findings suggest limited impact from methodologies in real-world ontology projects with practitioners not adhering to any particular methodology or systematic approach for ontology building (Gomez-Perez 2004; Paslaur 2006). Those found in use were not integrated into classical business process model (Paslaur 2006). For example, multilingual ontologies lack a scientifically based comprehensive methodological cookbook (Spyns 2007). The existing ontology management tools have however reached a feasible level of functionality to be useful (Paslaur 2006).

Consideration: No single methodology covers all required aspects and a mix-and-match approach may be necessary for different stages of ontology development

3.5 Discussion and Summary

Use case requirements (cf. Section 6.1) and related practical considerations (cf. Section 3.4) influenced domain modelling methodology and knowledge repository selection. The methodology had to cater for tacit knowledge capture of domain linguistics from business filings disclosure sections. Specifically the complex linguistic issues relating to task problem definition, language use, meaning variation and information association based on analyst operational and analytical knowledge have to be codified. Fact-oriented and layered approaches such as DOGMA have been successful in facilitating stakeholders representing and understanding semantically stable (structured) representations, while emphasising reusability and scalability. Across the methodologies investigated DOGMA stands out with its distinguishing characteristic of grounding the linguistic representation of knowledge (Jarrar 2002; Spyns 2002) and the methodological separation of domain vs. application conceptualization, termed double articulation (Spyns 2002) (cf. Section 4.4). Unlike many of the other ontology modelling methodologies DOGMA is not confined to a particular representation language (Jarrar 2002; De Leenheer 2005a; De Moor 2006). DOGMA with its
origins in database research allows the use of database modelling techniques (e.g. ORM, Entity-Relationship), for conceptual schema development, making ontology implementation in a relation model, a logical selection. The DOGMA modelling methodology and stages used, which fits with Tuomi’s reverse knowledge hierarchy (Tuomi 1999), along with Nonaka and Takuchis’ social knowledge conversion processes, is elaborated on in Section 4.4.

The previous chapter established the acceptance of shallow natural language processing analysis as a ‘good enough’ approach for information extraction. Findings from applications that base information extraction on an ontology, or Ontology Based Information Extraction (OBIE) report performance results improvements (Halpin 2001; McGuinness 2004). Providing a means of rich knowledge representation, the ontology addresses the lack of an information extraction schema, serving as a bridge between the business information requirement and extraction application. Linguistic analysis requires a three-dimensional framework that considers the data, domain model, and application artefact (Zhang 2004). The framework represents the phases of problem structuring that migrates the competitive analysis problem towards one of structure and greater certainty.

We next introduce issues associated with the free text disclosure data, the idea behind information threads and the origins of Semantic Paths. Specifically we exploit the semantically rich ontological structure to introduce the idea of providing particular information views based upon an elementary level of intuitive reasoning throughout the information space. The views collectively termed Semantic Paths (cf. Chapter 4 ) combine business domain knowledge with the operational knowledge of the analyst’s analytical process. The former represents well defined but tacit relationship between financial statements and the latter the reasoning processes behind domain knowledge, used to appraise statement quality. Both are key to problem definition and structuring. Similar combination approach’s to domain knowledge have been used by ontology based expert system for corporate financial ratings (Li-Yen 2009). Semantic Path modelling based on the DOGMA Ontology Modelling Methodology, used to construct the Competitive Analysis Ontology is then discussed.
Chapter 4

Semantic Paths

We introduce the notion of a Semantic Path through its motivational use in competitive analysis and contextual use within business filings. Section 4.1 details the difficulty of sourcing information within disclosure sections and motivates the requirement for automated information provision. Modelling the analyst information requirement introduces the notion of Semantic Paths, their origins in domain discourse and contribution from domain knowledge elicitation, in Section 4.2. Discourse analysis based upon domain linguistics, and combined with tacit knowledge capture relating to the competitive analysis task, allows fundamental knowledge representation using concept mapping to take place. The Semantic Paths are then formally defined (Section 4.3), before the modelling methodology used to develop the paths into an ontology, is introduced (Sections 4.4, 4.5), complete with examples of domain conceptualisation (Section 4.6).

4.1 Origin in Business Information Extraction

The U.S. SEC mandates that all publically quoted companies submit their financial reports and statements to the EDGAR filing repository. Filings are submitted in a proscriptive format as so-called forms\(^45\), which outline for different report types, expected information content, disclosure obligations, and in some instances the descriptive language terms allowed\(^46\). The forms are intended to address financial dissemination, analysis of time sensitive information and provide a level of financial transparency and comment that benefits shareholders and

---

\(^45\) Refer to http://www.sec.gov/info/edgar/edmanuals.htm for a full listing of EDGAR Forms and their descriptions

\(^46\) Within the meaning of Section 27A of the U.S. Securities Act (1933), Section 21E of the Securities Exchange Act (1934). Forward looking statements may include terminology such as may, will, expects, plans, anticipates, goals, estimates, potential, or, continue, negative thereof or other comparable terminology regarding beliefs, plans, expectations or intentions regarding the future
investors as a whole. The U.S. SEC annual report is of type Form 10-K and its quarterly counterpart, the Form 10-Q. Competitive analysis can be thought of in broad terms as an activity that helps determine a picture of a company’s health. Corporate filings such as the Form 10-K and 10-Q consist of consolidated financial information and unstructured disclosure statements are its primary information source. Other sources include company web sites, analyst reports, third party analysis providers, news articles, industry journals, marketing collateral, and social media such as blogs (Davidson 1997). Used for the identification of investment opportunity such as mergers, partnerships or equity stakes, competitive analysis can also be used for the identification of company candidates for outsourcing opportunities. It is the softer management statements within the filings, rather than the structured financial figures, that this research is directed towards.

Business analysis in activity terms has been categorized as comprising both mechanical and analysis tasks (Debreceny 2001). Mechanical is the critically important manual preliminary work performed to locate and correlate key text segments that actually inform the analysis taking place. An analyst performing a competitive analysis activity will systematically go through a business filing looking for such text segments of interest, annotating any found. Figure 4-1 is an example of this process in action with identified trigger terms and their proximity to other such terms.

*Figure 4-1 Example Analyst Annotation of Form 10-Q*

Davidson groups the information sources into categories of recorded data, observable data and opportunistic data.
While the importance of quick and accurate identification of information within these disclosures is evident (Debreceny 2001), those within financial statements present a particular difficulty for automated processing, as they are weakly structured, and have large areas of narrative comment from corporate officers in unstructured text. Issues encountered include the volume of narrative information, duplication of information and the fact that the most valuable is often camouflaged and difficult to find. The following example extract implicitly suggests future product revenue implications, but is it due to a problem with either: i) getting the product to market; ii) having a product that there is limited market for or; iii) issues with development and release cycles. Such ‘hints’ remain open to interpretation and pose further questions that warrant more thorough investigation.

Form 10-Q Extract 4-1, Citrix Quarterly Period September 30, 2003 (1)

“Our future success could be hindered by: delays in our introduction of new products; delays in market acceptance of new products; or new releases of our current products.....”

The next excerpt, is an example of duplicated instances found both within the same report, and across multiple consecutive reporting periods, that warns of general market conditions for possible future revenue reductions, but should it be interpreted as due to the effect of: i) companies generally reducing software purchases; or ii) maybe the companies type of product offering?

Form 10-Q Extract 4-2 Citrix Quarterly Period September 30, 2003 (2)

“The stock market in general, The Nasdaq National Market and the market for software companies and technology companies in particular, have experienced extreme price and volume fluctuations. These broad market and industry factors could materially and adversely affect the market price of our stock, regardless of our actual operating performance”

With insufficient information to provide insight on any underlying issue an analyst will intuitively, and logically, manually associate additional text segments to try and gain that insight. When taken in close proximity to previous annotations, the missing context can often be provided, as illustrated by the next excerpt, which implies that a larger than necessary workforce will likely keep future operating costs high.
For an analyst, excessive operating cost suggests a direct impact on future revenue generation. Depending on the nature of the analysis this could represent an attractive candidate to offer outsourcing services to, or an unattractive investment opportunity. This approach of manually associating and annotating related text segments, provides part of an analysis picture (corporate insight), which incrementally builds to provide a more holistic competitive analysis picture.

In this sense all text segments are of potential relevance but within different contexts. Context for competitive analysis may be thought of as the contribution that may be gained for particular areas of corporate business activities such as Sales, Acquisitions, Relationships, where each provides different understandings and insights into corporate activity (cf. Section 4.6.1). The analyst annotated text block in Figure 4-2 illustrates this dual context, which allows multiple interpretations to be made. On a domain linguistic level terms carry meaning, which are used by the analyst to first develop an awareness of report content, before inferring what the actual information intention of the report creator was (Winograd 2001).

Here from an R&D context, the interpretation can be that the product is failing to meet customer and therefore market expectation, but from a Sales context, an interpretation can point to issues with product sales and consequential revenue impact. This
notion of context introduces an additional level of complexity, ensuring that disclosure information overload and information being overlooked remains a constant companion.

A Form 10-Q’s content structure outlined in Figure 4-3, comprises two parts subdivided into a number of item areas, some of which if deemed not applicable by the reporting organisation, can be considered optional.

Figure 4-3 Form 10-Q Content Structure

PART I: FINANCIAL INFORMATION

ITEM 1. Condensed Consolidated Financial Statements (Balance Sheets, Statements of Income, Statements of Cash Flow Notes)

ITEM 2. Management discussion and analysis of Financial Condition and Results of Operations

ITEM 3. Quantitative and Qualitative Disclosures About Market Risk

ITEM 4. Controls and Procedures

PART II. OTHER INFORMATION

ITEM 1. Legal Proceedings

ITEM 2. Unregistered Sales of Equity Securities and Use of Proceeds

ITEM 3. Defaults Upon Senior Securities

ITEM 4. Submission of Matters to a Vote of Security Holders

ITEM 5. Other Information

ITEM 6. Exhibits and Reports on Form 8K

Related work on extraction from business filings (Section 2.2.1), analyse the basic structure of the filing and targeted particular areas for processing. Balance sheets for example are sectioned under PART 1, ITEM 1, Condensed Consolidated Financial Statements. Filing submission formats were initially SGML, followed later by HTML and PDF. All lacked a common set of tags and named properties, making automating processing difficult. The 2007 EDGAR filing manual did introduce SEC filing tags (similar to HTML tags), in an attempt to standardise the named properties contained in the header elements, but most companies continued to deviate. From September 2010, the U.S. SEC mandated (beginning with the 500 largest companies) that the reporting element and submission

---

48 Refer to Appendix III for an example extract from a HTML formatted filing used in this research
49 www.sec.gov/info/edgar/edgarfm-vol2-v5.pdf
format would be the eXtensible Business Reporting Language (XBRL) (Jones 2008; Sicilia 2009). XBRL comprises taxonomies and schemas that describe financial instruments, their calculations and presentation format. XBRL addresses the tags problem by allowing elements to be directly accessed and values extracted (refer to Section 2.3 for discussion on business taxonomies). Whether SEC or HTML based filings, the free text disclosure sections are found enclosed in property tags, such as:

**Figure 4-4 Disclosure section HTML source extract**

```html
<P STYLE="margin-top:0px;margin-bottom:0px; text-indent:4%">
  <FONT FACE="Times New Roman" SIZE="2">
Our business has grown rapidly. This has placed, and any future growth would continue to place, a significant strain on our limited personnel management and other resources. Our ability to manage any future growth in our business will require us to:
  </FONT>
</P>
```

XBRL provides labels with associated ids that link directly to financial facts or instance data blocks. The next sample extract provides financial figures from three different reporting periods, (in bold) for the proceeds from “investment sales in securities” in different context references.

**Figure 4-5 XBRL instance document financial instrument extract**

```xml
<hit:ProceedsFromSaleOfInvestmentsInSecuritiesAndOther
  contextRef="eol_PE6445----1020-F0005_STD_365_20100331_0"
  unitRef="iso4217_JPY"decimals="6">27410000000</hit:ProceedsFromSaleOfInvestmentsInSecuritiesAndOther>

<hit:ProceedsFromSaleOfInvestmentsInSecuritiesAndOther
  contextRef="eol_PE6445----1020-F0005_STD_365_20090331_0"
  unitRef="iso4217_JPY"decimals="6">112771000000</hit:ProceedsFromSaleOfInvestmentsInSecuritiesAndOther>

<hit:ProceedsFromSaleOfInvestmentsInSecuritiesAndOther
  contextRef="eol_PE6445----1020-F0005_STD_366_20080331_0"
  unitRef="iso4217_JPY"decimals="6">104878000000</hit:ProceedsFromSaleOfInvestmentsInSecuritiesAndOther>
```

For the reporting period "eol_PE6445----1020-F0005_STD_365_20100331_0” the LegalReserveAndRetainedEarningsAndDividendDisclosures text block are referenced, holding qualified comment on the proceeds from sale of investments in
securities. Text blocks unlike the directly references figures above require further processing.

*Figure 4-6 XBRL instance document text block extract*

```
<hit:LegalReserveAndRetainedEarningsAndDividendsDisclosureTextBlock contextRef="eol_PE6445----1020-F0005_STD_365_20100331_0">
    -----<-----
    (13)Legal Reserve and Retained Earnings, and Dividends
    -----<-----
    Dividends during the years ended March 31, 2009 and 2008 represent dividends declared during those years. For the year ended March 31, 2010, the Company did not pay any dividends. On March 18, 2010, the Board of Directors decided not to pay a dividend for the second half of the year ended March 31, 2010.
    -----<-----
</hit:LegalReserveAndRetainedEarningsAndDividendsDisclosureTextBlock>
```

With the majority of financial comment being text based, automated processing and interpretation of these text blocks, remains difficult to exploit in the absence of insightful semantics and better representation formats. XBRL does offer assistance for financial figures allowing direct element reference and extraction, but like the previous filing formats, does not offer assistance with narrative sections. With accounting standards requiring lengthy disclosures, problems associated with disclosure sections (either future or legacy) are set to remain, ensuring that even with the uptake of XBRL based systems, extraction capability from unstructured text will remain useful (Grant 2006). Freely available from the U.S. SEC’s EDGAR Database, and openly regarded as one of the richest and trusted sources of information for the financial profession (Adams 2001), the filings available in either HTML or PDF formats, offer neither semantic nor syntactic assistance, ensuring that softer financial content remains difficult to find, and even more difficult to automatically retrieve (Debreceny 2001). Addressing this lack of standard reporting elements and terminology in an attempt to eliminate many of the problems associated with the transfer and use of financial data (Coffin 2001), the SEC in 2010, introduced the Extensible Business Reporting Language. With current accounting standards requiring filings to have lengthy discussion sections, problems associated with searching these sections are set to continue, as the XBRL vocabulary tags used to label financial data items such as (e.g. nett profit, gross sales), for
content classification, do not enhance the level of semantic meaning associated with the statement contents.

Research questions [RQ2] and [RQ3] (Section 1.2), target how unstructured financial disclosure sections can be automatically transformed into structured financial information thereby reducing the required manual effort. The fundamental question next discussed is whether the analyst’s information need, can be expressed in business discourse terms?

4.2 Foundations of Semantic Path Modelling

Researching the problem motivation mentioned in the previous section requires knowledge as to the type of terms being sought, and an expectation of what is likely to be found. Domain discourse analysis enables term lists to be generated providing a basis for determining what is being sought. Additional domain insight from domain experts can then assist with concept identification, their association and hierarchy formulation, to cater for what is likely to be sought after. Concept mapping for knowledge representation then provides a basis for formal domain linguistic modelling (also referred to as lexicalisation).

Concept maps reflect a combination of business domain and operational knowledge of the analyst’s analytical process. Their proposition templates with inherent context specific semantics, we refer to as Semantic Paths, as they provide an information association and navigation map within the information space. Rich in domain knowledge the paths construction represents a reasoning task and their traversal a cognitive analysis.

4.2.1 Discourse Analysis

Hyland defines meta-discourse as a linguistic tool that creates a textual structure, that goes beyond the statement of the subject matter, and gives clues as to the purpose and attitude of the writer (Hyland 2004). In effect, meta-discourse consists of text tokens that do not contribute to the propositional development of the text, but serves to guide the reader in interpretation and response to the text (Hyland 1998). In describing the functional categories of meta-discourse, Hyland classifies meta-discourse as comprising the functional categories of interactive and interactionable (Hyland 1998; Moreale 2004). Interactive refers to the writers attempts at constraining the text, to their preferred interpretation and goals. Interactive resource usage is concerned with how discourse is organized and the extent to
which the text is structured, with knowledge of the reader and their needs in mind. Interactive resources are used by the writer to organize propositional information, in a manner that the reader is likely to find coherent and convincing, effectively managing the information flow. Interactionable refers to the level of writer intrusion into the text by way of comment, opinion and evaluation (Hyland 2004). Interactionable resources are employed to anticipate, acknowledge, challenge or negate alternate interpretations being drawn, in effect, restricting opportunities for alternate view development in the first instance. The following example illustrates this, suggesting that taking longer to receive sales revenue, is due to a strategy that targets existing customers to increase business as opposed to either: i) the sales force not really functioning correctly or; ii) customers are not informed on the product or even; iii) company not pursuing new business in addition to enhancing existing.

Form 10-Q Extract 4-4 BEA Quarterly Period 30 October, 2004

“Selling and marketing a platform product also requires significant investment in marketing to educate customers and prospects, and significant investment to educate our sales force. One goal of the platform approach is to sell more broadly inside customer organizations, and also to sell at higher levels in the customer organizations. This approach may have the effect of lengthening sales cycles and increasing deal sizes, both of which make it harder to predict the timing and size of large deals”

Another example from BAE, provides an explanation for the decline in deferred revenue as either: i) not being the companies fault; and ii) attempts to draw attention away from the fact that there may be internal issues in service provision or; iii) possibly that services are declining as the product is losing market share? None of which seem to be the fault of the company.

Form 10-Q Extract 4-5 Citrix Quarterly Period September 30, 2003 (4)

“Although deferred license revenue declined for the three months ended October 31, 2004, the primary driver in the decline in deferred revenues was deferred customer support revenues. Deferred revenues will fluctuate in the future, as a function of the timing and terms of particular transactions and will not necessarily correlate with revenue growth in any given quarter and fluctuations in deferred revenues are not an indicator of future license revenue”

50 Revenue not earned until the delivery of goods or services
Table 4-1 below details the meta-discourse categories and their functional descriptions. Hyland uses the proposition meta-discourse distinction as a starting position for academic meta-discourse exploration, but others have included propositional content as part of meta-discourse (Crismore 1990; Hyland 2004). To further blur the issue meta-discourse and propositional elements can occur within one sentence and what is considered propositional in one context, is meta-discourse in another. Results from the study of meta-discourse within CEOs letters, indicate that the functional devices of transitions and hedges, together account for 66% of all discourse items (Hyland 2004). Within the area of business postgraduate studies, this figure rises to 90% (Crismore 1990).

Disclosure notes from the financial statements present particular difficulty as they are textual, lack structure and have no common format (Grant 2006). Written in the language of business discourse ensures that an analysts faces the problem of first identifying text segments of interest, before attempting to interpret their actual business meaning. In tackling identification within this context, we adopted the idea of the meta-discourse *hedges* category, as an assist in the construction of a domain rich term set, comprising words and word phrases, usable as part of a lexicalisation process for concept formulation. As any single concept or text segment can typically provide only part of the overall proposition being made, and the level of company commitment to it, the text segments are often grouped for consideration.
Table 4-1 Model of Meta-discourse in Academic Texts (adapted from (Hyland 2004))

<table>
<thead>
<tr>
<th>Interactive Resource</th>
<th>Category</th>
<th>Function</th>
<th>Device lexicalisation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transitions</td>
<td>Express semantic relation between main clauses</td>
<td>And/or/but /in addition/thus</td>
</tr>
<tr>
<td></td>
<td>Frame Markers</td>
<td>Draw attention to discourse goals or indicate topic or argument shifts</td>
<td>Finally/to conclude/my purpose here is to/ I argue here/ Well now</td>
</tr>
<tr>
<td></td>
<td>Endophoric Markers</td>
<td>Refers to information in other parts of the text</td>
<td>See Section X/noted above/see Fig X/In Section X</td>
</tr>
<tr>
<td></td>
<td>Evidentials</td>
<td>Refer to source of information in other parts of the text</td>
<td>According to X/ 2005, X/X states</td>
</tr>
<tr>
<td></td>
<td>Code Glosses</td>
<td>Assist reader in interpretation of ideational material</td>
<td>Such as/for instance/in other words/namely/e.g.</td>
</tr>
<tr>
<td></td>
<td>Interactional Devices</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hedges</td>
<td>Withhold writers full commitment to proposition</td>
<td>Possible/might/perhaps/about</td>
</tr>
<tr>
<td></td>
<td>Boosters</td>
<td>Emphasises force or writers certainty in proposition</td>
<td>In fact/it is obvious/definitely/clearly/it is clear that</td>
</tr>
<tr>
<td></td>
<td>Emphatics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attitude markers</td>
<td>Express writers attitude to proposition</td>
<td>Unfortunately/I agree/agreement/surprise/surprisingly</td>
</tr>
<tr>
<td></td>
<td>Engagement markers</td>
<td>Explicitly refer to or build relationship with reader</td>
<td>Consider/note that/you can see</td>
</tr>
<tr>
<td></td>
<td>Self mentions</td>
<td>Explicit reference to author(s)</td>
<td>I/we/my/our</td>
</tr>
</tbody>
</table>

This type of ‘grouping’ required a level of semantic association, which we drew from the meta-discourse transition category. Both of these meta-discourse principles also assisted in later introducing cognitive aspects of the task through business logic, and domain knowledge. Table 4-2 provides a sample listing of the terms identified during discourse analysis and Appendix I may be referred to for a full listing of terms. Both of these meta-discourse principles also assisted in later introducing cognitive aspects of the task through business logic and domain knowledge.
Table 4-2 Discourse Analysis Term List Extract

<table>
<thead>
<tr>
<th>access</th>
<th>change</th>
<th>Detain</th>
<th>Headcount</th>
</tr>
</thead>
<tbody>
<tr>
<td>acquire</td>
<td>channel</td>
<td>detain.. market</td>
<td>Hold</td>
</tr>
<tr>
<td>acquisition</td>
<td>channel partner</td>
<td>detain .. market acceptance</td>
<td>hold .. market</td>
</tr>
<tr>
<td>act on</td>
<td>charge</td>
<td>detain .. release</td>
<td>hold .. market acceptance</td>
</tr>
<tr>
<td>act upon</td>
<td>code</td>
<td>detain ... Introduce</td>
<td>hold .. release</td>
</tr>
<tr>
<td>action</td>
<td>collaboration</td>
<td>difficult .. predict</td>
<td>hold ... introduce</td>
</tr>
<tr>
<td>adapt</td>
<td>competition</td>
<td>Direct</td>
<td>Holdup</td>
</tr>
<tr>
<td>administration</td>
<td>competitive</td>
<td>Director</td>
<td>holdup .. market</td>
</tr>
<tr>
<td>adopt</td>
<td>competitor</td>
<td>Discharge</td>
<td>holdup .. market acceptance</td>
</tr>
<tr>
<td>advise</td>
<td>compliancy</td>
<td>Dispose</td>
<td>holdup .. release</td>
</tr>
<tr>
<td>against</td>
<td>compliant</td>
<td>Divest</td>
<td>holdup ... Introduce</td>
</tr>
<tr>
<td>agree</td>
<td>comply</td>
<td>Divide</td>
<td>impact .. ability .. sell</td>
</tr>
<tr>
<td>agreement</td>
<td>concentrate</td>
<td>Economic</td>
<td>Increase</td>
</tr>
<tr>
<td>alter</td>
<td>conditional</td>
<td>Employee</td>
<td>Industry practice</td>
</tr>
<tr>
<td>amalgamate</td>
<td>constitute significant</td>
<td>entitlements</td>
<td>industry standards</td>
</tr>
<tr>
<td>amortized</td>
<td>contracts</td>
<td>Growing</td>
<td></td>
</tr>
<tr>
<td>amount</td>
<td>department</td>
<td>Growth</td>
<td></td>
</tr>
</tbody>
</table>

The following example outlines lack of commitment to the stated reason for reduced future revenue, namely that the company needs either: i) a new product strategy for their products or; ii) to get the product to market quicker.

Form 10-Q Extract 4-6, Citrix Quarterly Period September 30, 2003 (5)

“Our future success could be hindered by: delays in our introduction of new products; delays in market acceptance of new products or new releases of our current products; ... ”

Across the text segment we have terms such as product, delay in market acceptance, delay in introduction, introduction, and release. With analyst verification, this domain rich term set, was subsequently added to domain knowledge and used with linguistic analysis for concept identification and association.
4.2.2 Domain Knowledge Contribution

The importance of capturing and using an experts tacit knowledge for further knowledge creation within any company is long recognised (Nonaka 1995). Concept maps, used for organising and representing knowledge are a formalism that can help capture, archive and codify expert tacit knowledge and identify gaps in knowledge structure (Novak 2008). Concept maps are based on Ausubel previous work in cognitive psychology, and subscribe to the idea that a learners cognitive structure has a concept and proposition framework, that added to, assimilates the learning process (Ausubel 1963; Ausubel 1978). Comprising concepts and propositions, they are the building blocks for knowledge within any domain (Novak 2008). Concepts are defined as perceived regularity in events or objects, or records of events or objects with propositions being statements about some object or event. Propositions contain two or more concepts connected using linking words or phrases, also referred to as semantic units (Novak 2008). A characteristic is that concepts are represented in a hierarchy, in a descending order with the key (more general) concepts at the top and subordinate concepts below, depending on the context in which that knowledge is being applied or considered (Novak 2008). The context ensures that the map is best constructed with an explicit focus question, for tacit meaning to be made explicit. Our focus question used to start the process of knowledge extraction with the business analysts was:

“What core concepts (term based), do you use as triggers and how are they associated, such that possible insight for competitive analysis is achieved?”

Domain specific terms capture the knowledge of a domain and reflect it as a formal agreement on the meaning of terms commonly accepted by the community, facilitating information interaction and exchange. Where a corpus lacks standard syntactic formulation and presents a vocabulary rich in idiomatic expressions (such as business disclosure statements), the processing of term extraction is challenging. With domain expertise, the terms of the previous section were formulated into a series of concepts associated with other concepts. Concept mapping allowed concept associations to be represented as a series of term based proposition templates grounded in a number of context areas. Concept mapping further allowed the establishment of a propositions hierarchy (semantic
associations), each with their own inherent semantics, ensuring that each individually could contribute in part or full to a particular insight. Figure 4-7 below represents the concept map developed for the Sales context area. As previously noted the proposition templates with inherent context specific semantics, that provide an information association and navigating map within the information space, we refer to as Semantic Paths (O’Riain 2006).

Table 4-3 represents the insight introduced from domain experts providing the analyst interpretation on what is being sought from the proposition. The proposition template Produce [unfilled orders] backlog, as an example should be interpreted as an interest in finding information that involves the concept Products and Backlog51, as they might provide insight on the total value of orders unfulfilled for a particular product or process/manufacturing issues.

We have not found evidence of this type of approach for the domain of competitive analysis anywhere else in the literature.

---

51 Backlog is an accounting concept used by analysts to help more accurately estimate future earnings and performance
Figure 4-7 Semantic Paths Sales Context Concept Map/Semantic Network
### Table 4-3 Concept Mapping Proposition Template Listing

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Interpretation with Domain Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product [being] introduce</td>
<td>Competition, new revenue, new technology, new market, new geography?</td>
</tr>
<tr>
<td>Product [something happening with] announce</td>
<td>Movement, positive (new market, geography, market) or negative (delay, end of life, product issues, issues with sales falling sort of targets)</td>
</tr>
<tr>
<td>Product [change / comment] revenue</td>
<td>Interruption to revenue stream, revenue increase or decrease?</td>
</tr>
<tr>
<td>Produce [unfilled orders] backlog</td>
<td>Problem collecting money or with future performance or process/manufacturing problems?</td>
</tr>
<tr>
<td>Revenue [coming from] percentage</td>
<td>Source percentage change in revenue origin (where is it coming from)?</td>
</tr>
<tr>
<td>Revenue [generated from] royalties</td>
<td>Disproportionate or not, new/ declining royalties</td>
</tr>
<tr>
<td>Revenue [increasing costs] cost of goods sold</td>
<td>What percentage of? Why the increase/decrease is it labour/material/distribution related?</td>
</tr>
<tr>
<td>Revenue [talk about] profit / loss</td>
<td>Qualification of balance sheet, what are the year on year major changes?</td>
</tr>
<tr>
<td>Revenue [outgoings] contractual payment</td>
<td>Dependency on, increases to, percentage of costs, is it viable??</td>
</tr>
<tr>
<td>revenue [reduction / deferred] split substantially</td>
<td>What caused the reduction, deferral?</td>
</tr>
<tr>
<td>revenue [need to diversify]</td>
<td>Over dependency on single stream, is the business model broken, what is the strategy to diversify?</td>
</tr>
<tr>
<td>Revenue [implication for] price</td>
<td>What is the reason for the change in pricing, market or economic pressure, competition, old product requiring new version or replacement?</td>
</tr>
<tr>
<td>Announce [mentions] release</td>
<td>Release date, what markets, what does it replace, competitive announcement</td>
</tr>
<tr>
<td>Announce [indicate] delay in market</td>
<td>Market acceptance of product problems, shift in market requirements?</td>
</tr>
<tr>
<td>Announce [indicates] delay</td>
<td>Problems with development/distribution/market acceptance / pricing / competition from better products?</td>
</tr>
<tr>
<td>Announce [informs on] competitive</td>
<td>Reason for competitive issue</td>
</tr>
<tr>
<td>Announce [suggests] cancel</td>
<td>Change of offering in revenue streams / development issues/ market acceptance issues / competition there first, with better offering?</td>
</tr>
<tr>
<td>Announce [mentions] litigation</td>
<td>Managed or not, strategy issues. Positive; product is available, negative; delay, development issues</td>
</tr>
<tr>
<td>Announce [indicates] planned</td>
<td>Managed or not, strategy issues. Positive; product is available, negative; delay, development issues</td>
</tr>
<tr>
<td>Announce [mentions] schedule</td>
<td>New technology offering, revenue anticipated, reaction to market?</td>
</tr>
<tr>
<td>Introduce [reference to] 3rd party</td>
<td>Ahead of competitors or not, license from 3\textsuperscript{rd} party, develop, manufacture, distribute with 3\textsuperscript{rd} party?</td>
</tr>
<tr>
<td>Introduce [problems with] unable</td>
<td>Problems with product development or supply chain? Same as delay above, technology may not be working</td>
</tr>
<tr>
<td>Introduce [problem with] litigation</td>
<td>Hold up with/possible litigation, possible cost, what’s the issue; license, manufacturing, distribution, IP, patent?</td>
</tr>
<tr>
<td>Introduce [indicates] delay</td>
<td>Problems with product development or supply chain, same as announce-delay, introduce-delay</td>
</tr>
<tr>
<td>Introduce [implements / does not] industry standards</td>
<td>Non adherence to standards indicates short term future, low revenue, technical/design issues, R&amp;D cost issues</td>
</tr>
<tr>
<td>Introduce [implication for] price</td>
<td>Discounted, too high, changes to, or competitive pressures and price is a competitive advantage?</td>
</tr>
<tr>
<td>Introduce [with] forecast</td>
<td>Projected revenue figures, over what period, forecast greater/lesser revenue/margin vs. previous forecast</td>
</tr>
</tbody>
</table>
4.3 Defining Semantic Paths

So far our discussion has been on the definition and use of concepts and relationships. In order to help define Semantic Paths we introduce the abstraction between them and graph theory. With reference to Figure 4-8 consider a directed labelled graph $G$ defined as $G = (V, E, L)$, with $V$ the set of vertices $\{v_1, \ldots, v_n\}$, connected by a set of edges $E$, in the edge space, with labels $L$. Then $E \subseteq V \times V$ and some function $\delta$, maps the edges to a set of labels

$$\delta : E \rightarrow \mathcal{P}(L)$$

Next introduce colours/contexts to the graph such that

$$G = (C, V, E, L)$$

Where $C$ is the set of contexts $\{c_1, \ldots, c_n\}$, defined by $P(C)$. As each vertex can occur within multiple contexts, a function $\xi$ maps the set of vertices into the set of all contexts (non empty).

$$\xi : V \rightarrow \mathcal{P}(C) - \emptyset$$

A Semantic Path $\Pi = <v_1, \ldots, v_n>$ can be defined on $G$ as an ordered sequence of $V$ such that

$$(v_i, v_{i+1}) \in E \text{ for } i = 1, \ldots (n - 1)$$

As paths (like vertices), can also occur within multiple contexts for the set of all Semantic Paths $\Pi_G \in G$,

$$\theta : \Pi_G \rightarrow \mathcal{P}(C) - \emptyset \text{ as } V \rightarrow C$$

$$\theta : (v_1, \ldots, v_n) = (\xi(v_1), \ldots, \xi(v_n))$$

In general terms the context directed graph, containing $\Pi_G$ may therefore be considered as:

$$G = (V, E, C, L, \delta, \xi)$$
Semantic Paths are constructed from a union of all paths concepts based on selective ordering of those concepts, informed by domain expert’s tacit knowledge (cf. Section 4.2.2, 5.1). For each \( \mathcal{P} \) the context \( c_j \in C \) is shared along the path, and supports domain specific intentional interpretation. Each \( \mathcal{P} \), therefore represents a real information sequence and taken in context, is a means to facilitate information identification and assessment within the application domain.

### 4.4 Semantic Paths Modelling Methodology

Having established a taxonomy for business argumentation categories, and an initial knowledge representation through the use of concepts maps (cf. Section 4.2), the research question became whether the identification and extraction of information, could be automated? Bridging the knowledge gap between domain linguistic knowledge and ontology design we selected the *Developing Ontology Grounded Methods and Applications* (DOGMA) - Ontology Modelling Methodology (DOM²) (Meersman 1999; Meersman 2001; De Leenheer 2005a; Spyns 2005a; Spyns 2005b; Spyns 2007), that has its roots grounded in
database semantics and model theory. Inspired by the principles of Object Role Modelling (ORM) (Halpin 2001) and aN Information Analysis Method (NIAM) (Verheyen 1982), DOGMA looks to the application and refinement of an integrated approach, based upon database experience, linguistic insight and social science collaborative aspects.

DOGMA bases its framework on a set of seven core principles that draws from best practice from the following ontology modelling methodologies (Spyns 2007):

- Methontology (Fernandez 1997), for the inclusion of management activities
- TOVE (Fox 1992), for inclusion of competency questions to scope the domain and evaluate its conceptualization
- Enterprise Ontology (Uschold 1998), for brainstorming, middle out approach and term grouping
- Unified Process for Ontology Building Method (De Nicola 2008), using users to scope the domain of interest, along with generalised competency questions that are not heavily formalised
- OnToKnowledge (Sure 2002), for feasibility study inclusion before any ontology development begin
- CommonKADS (Schreiber 1999), for documenting development process deliverables activity.

Its methodology comprises the two main activities of a preparatory stage and an ontology engineering stage (cf. Appendix V). For a detailed description on the DOGMA formalism we refer the reader to (De Leenheer 2005a). Gruber previously identified five principles (listed in Section 4.6.1) for designing ontologies for knowledge sharing, similarly categorised, but with activities of adequate ontology coverage through knowledge expression as ontological blocks, and ontological construction (Gruber 1995). Figure 4-9 outlines the breakdown of each DOM stage into its component sub-stages, from which we are specifically interested in the lexon engineering stage associated with domain conceptualisation. Lexon engineering (cf. Section 4.6) constitutes the core DOGMA

---

52 Familiarity with ORM and its use in Relational Database Design is assumed.
component used to formally model our CAO. Lexons can be considered as lexical representation of a conceptual relationship between two concepts (De Leenheer 2007).

Figure 4-9 DOGMA Modelling Methodology - Domain conceptualisation (Spyns 2007)

An overview of DOGMA’s more important principles, equatable with Gruber’s principles of:

- clarity, the effective communication of the intended meaning of defined terms
- minimum encoding bias: “conceptualization should be defined at the knowledge level without depending on a particular symbol-level encoding”
- minimum ontological commitment, based on the consistent use of the vocabulary
- definition only of terms required for knowledge representation (Gruber 1995)

Ar (Spyns 2007):

i) the transposition of the principle of data independence into the principle of meaning independence

---

A commitment is the agreement to use a vocabulary in a coherent and consistent, manner defined by Guarino as a function that “links vocabulary terms with a conceptualisation” (Guarino 1998)
ii) in relation to its universe of discourse, the ontology should be based on plausible interpretation-less facts

iii) the ability to generate multiple views of the same stored conceptualisation allowing addition of application specific semantic restrictions, based on needs and usage

Overall DOGMAs principles support its core notion of double articulation, which decomposes an ontology into an ontology base and a separate commitment layer (Spyns 2002; Spyns 2005b). The ontology base is used to introduce Semantic Paths concept constituents, defined formally as part of the ontology, using the commitments layer. The theory behind both is introduced in Section 4.6, with instantiated examples of their use in CAO construction, presented in Section 5.1.

4.5 Related DOGMA Evolution

DOGMA-MESS (Meaning Evolution Support System) is a community based extension to the DOGMA methodology catering for distributed meaning negotiation and consensus for common ontology definition (De Moor 2006). In a preliminary stage, a defined meta-ontology captures the foundation of the ontology building process. The common meaning distillation process produces the common ontology. The ontology elicitation stage models and defines knowledge. Semantic analysis refines previous definitions based on difference detection. The various individual and common definitions are then compared in the meaning negotiation stage. The outcome of this process is a set of common definitions which provide either: the final common ontology; or the input for the next iteration (De Moor 2006).

Both DOGMA and DOGMA-MESS have been extended and enhanced to broaden its applicability to different application uses. To support the context necessary to formalise and reason about knowledge between ontology elements and relationships (De Leenheer 2007), extended the DOGMA-MESS framework to manage multiple context dependency types and operators (i.e. articulation, application, specialisation, and revision). Context dependency was defined through formal description and decomposition of context driven processes (lexical disambiguation, contextualisation, alignment, versioning), in terms of the process primitives for selecting, linking and changing knowledge elements (De Leenheer 2007).
Context definition was intended to constrain possible relations between entities and their context, making negotiation and application less vulnerable to ambiguity. Semantic Decision Tables (SDT) have also been layered into the DOGMA framework, assisting with concept to domain ontology linkage (Tang 2007a). (Van de Maele 2008) used the framework to construct an ontology based crawler, for discovery and matching between different data and document sources on the Semantic Web. A topic model ontology was developed to perform the mapping between DOGMA concepts and Semantic Web document concepts using a string matching similarity measure. In looking to address the challenges presented from isolated ontologies and unstructured data as part of a more meaningful next generation internet, a Domain Name Space Information solution framework, for disparate data sources and service discovery reconciliation, was proposed by (De Leenheer 2009a). The framework further extended DOGMA-MESS with step-wise semantic reconciliation and application procedures. Within it, the ontology base, providing language neutral patterns grounded in informal meaning description, allows providers unlock new knowledge through data silo reconciliation and requestors querying the name space, for particular service concepts, that can be linked and combined to satisfy community needs. More recently the framework has been used to reconcile contextualised vocational training competency models with the development of a Vocational Competency Ontology (De Leenheer 2010). Fact orientated business semantic representation and management were grounded using the frameworks agile semantic reconciliation and application cycle. DIY environments that target the management of intelligent components data semantics based on previously integrated SDT (Tang 2007a; Tang 2007b), for human resource management, have also being reported (Tang 2010; Tang 2011).

4.6 Domain Conceptualisation using DOGMA

Investigating the practices of ontology engineering, (Paslaur 2006) noted knowledge elicitation relating to domain analysis, as the most time consuming ontology building process task. Methodologies that handled these at a general level, were not encountered for domain analysis and a key findings was that the methodology frameworks should be able to cater for customizable process models, that allow ontology engineers combine differing
methods, for different stages of the ontology engineering process on an as-required basis (Paslaur 2006).

The FF POIROT project (FF Poirot 2002) developed the Application Knowledge Engineering Methodology (AKEM) for practical knowledge system development as part of a use case that targeted fraudulent email activity (Gao 2005). AKEM adopts the DOGMA approach to ontology engineering, but with a simplified life cycle of scoping, analysis, development and deployment (Gao 2005; Yoon 2009).

Using this approach we developed DOGMA’s knowledge elicitation and breakdown using contextual enquiry (cf. Table 4-4), with multiple domain experts. Contextual enquiry involves a one-on-one based work practices observation and series of interviews sessions with the user (business analyst), where routines involving processes, and tasks are closely followed (Beyer 1998). As an approach contextual enquiry seeks to first establish why something is done, before then understanding how/how not it is done. For our research contextual enquiry was a particularly useful tool to establish the ‘design’ of the information need (why), and information combination approach (how), behind the competitive analysis task as detailed in Section 4.2. Contextual enquiry provides the setting for discourse analysis to take place and the codifying of domain knowledge through the use of concepts maps. Its outputs were a discourse term lists and a series of concepts maps and their proposition templates, representing business argumentation categories. The concept maps and proposition templates representing Semantic Paths were treated as verbalised elementary sentences and input directly into the lexon engineering phase.
Engineering lexons involves the progressive transformation of natural language verbalised facts, from initial conceptualisation, into more formal language independent statements with informal meaning. From Figure 4-9, lexon engineering comprises lexon creation, grounding and meta-lexon generation. Lexons represent semi-formal linguistically determined propositions of the domain of discourse and apart from intuitive linguistic interpretation, are not intended to have any formal semantics or interpretation (Meersman 2001). They are written as sextuples $<(\gamma, \zeta): \text{term}_1, \text{role}, \text{co-role}, \text{term}_2>$, where informally a lexon is a fact that may hold for some domain, expressing that within the context $\gamma$, and for the natural language $\zeta$, the $\text{term}_1$ may plausibly have $\text{term}_2$ occur in role with it, and inversely $\text{term}_2$ maintains a co-role relation with $\text{term}_1$ (Spyns 2005a; Spyns 2007). Lexons are independent of specific applications and should cover relatively broad domains (linguistic level).

Key to our usage of DOGMA for domain modelling and development of the ontology base, is our interpretation of role, context and commitments. Formal ontology roles have
previously been used to generate formal semantic mappings and annotations between decision items and lexons (Tang 2007a). Role categorisations were described as belonging to either: a: is-a taxonomical role; part-of; instance-of, property-of or equivalent role. Although our role is formalised as a binary role (cf. Section 4.2.2, 5.1), with property similar to part-of or relates-to, we specifically consider it more robustly defined using is-associated-with, as it more accurately captures the intended relationship between our concept associations.

Role categorisation between concepts should be interpreted as

is-associated-with

Meta-lexons, formally noted as a triple $<\text{concept}_1 - \text{relationship} - \text{concept}_2>$ are abstracted from lexons. Meta-lexons are language-neutral and context-independent (conceptual level). Natural language terms are associated, via the language and context combination, to a unique word sense represented by a concept label (e.g. the WordNet (Fellbaum 1998) identifier person#2). With each word sense, a gloss or explanatory information is associated that describes that notion. To account for synonymy, homonymy and translation equivalents there is an m:n relationship between natural language terms and word senses (Spyns P. 2005a).

Converting lexons into meta-lexons allowed the translation from the language level to the concept level

The lexons base consisting of intuitively plausible conceptualisations of a real world domain, i.e. specific binary fact types, and constituted by lexons grouped by context and language (Spyns 2002), forms the DOGMA ontology base. Contexts in the ontology base are introduced as organising principle by grouping related lexons. A context can be considered as an abstract identifier of a possible world, leading to specific possible world semantics (e.g. Kripke models) and mapping between them (Sowa 2000). The identifier refers to implicit and tacit assumptions in a domain that maps a term to its intended concept identifier (or meaning). (De Leenheer 2007) argues that context underpinned by these
assumption, does not represent explicit formal knowledge, and context is in fact defined through reference to a source, assumed to contain the necessary assumptions (Jarrar 2003).

Domain expert tacit knowledge brought to bear (discussed in Section 4.2.2), provided the environmental context from which explicit knowledge can be generated.

Lexons and meta-lexons are meant to reach a common and agreed understanding about the domain conceptualisation (important/relevant notions and how they are expressed in one or more natural languages (Meersman 1999)), and assist human understanding.

The layer of ontological commitments, or commitment layer, mediates between the ontology base and its applications. Each commitment is a consistent set of rules (or axioms) that add specific semantics to a selection of meta-lexons of the ontology base (Jarrar 2002). The commitment layer, with its formal constraints, is meant for interoperability issues between information systems, software agents and web services, as is currently promoted in the Semantic Web area (Spyns 2005a). The constraints are mathematically founded and concern typical DB schema constraints e.g. cardinality, optionality etc. Within DOGMA, commitment describes semantic constraints in terms of paths and for which role/co-role label pairs are interpreted in which ontological relationship (De Leenheer 2007). The selection of meta-lexons to form a commitment rule represents formal constraints to instantiate the particular commitment. Selected in this manner, the ontological commitment represents an explicit instance, of an intentional logical theory interpretation, for some application (Tang 2007a) and resonates well with our notion of a Semantic Path. (Tang 2007a) used the ontology base and commitment layer to construct SDT’s, that allowed decision items in the table, be mapped and linked to concepts in a human resources domain ontology. Formalised as \( <\Gamma, C; A; R_i> \), the context identifier \( \Upsilon \in \Gamma \), references the original decision table, \( C \) and \( A \) are the sets of lexon conditions defined with the same context identifier \( \Upsilon \), \( R_i \) represents a rule set that includes commitment axioms \( r_c \in R_c \) and semantically grounded decision rules \( r_o \in R_o; \{ R_o U R_o \} \subseteq R_i \). Extending meta-lexons and commitment rules has also been applied to generate a mappings element \( M (c, w) \) from \( <m_id, \Upsilon(z, t_i), R, c, sc(c, c)> \) between DOGMA concepts \( c \in C \) and Semantic Web document...
concepts, \( c_j \in C_w \). \( m_id \) is a mapping rule id, \( \Upsilon(\xi, t_i) \) is responsible for lifting DOGMA term \( t_i \) to concept \( c_i \) from the DOGMA commitment \( \Omega \). \( R \) is the linguistic relationship between the mapping rule concept labels, and \( sc(c_u, c_j) \), a normalised similarity score. In subsequent work that looked at human resource management related decision making, the lexon base was constructed to contain decision and action lexons and commitment layer the semantically grounded decision rules (Tang 2007b; Tang 2011).

The concatenation of selected meta-lexons to form a commitment (Semantic Path, defined in Section 4.4):

i) is driven by domain expert cognitive ordering, based on well-defined interpretation of real information sequences.

ii) adheres to an ordering that imposes an explicit constraint removing the necessity for axiomised constraints.

iii) represent a sentence comprised these multiple meta-lexons

Semantic Path instantiation is automated with relevance determined by path member distance. Simply stated, the domain model as represented in the ontology base can be richer than the actual content of the Semantic Paths. The development of lexons and meta-lexons used to generate the CAO and its use as a schema to target information extraction against is discussed in Section 5.2, complete with worked examples

4.6.1 Domain Coverage

The concept maps (introduced in Section 4.2.2) were the first step in modelling the task aspect of the competitive analysis. Proposition templates codify tacit domain knowledge, reflecting how an analyst logically constructs information threads (Semantic Paths), from related information items. Analysis of filing sources identifies actual domain specific terms and relationships behind the threads. Business information topics specifically targeted (cf. Section 7.1), were cost base growth, revenue, employees, supply chain difficulties, acquisitions, R&D and internal restructuring.

To incorporate practically grounded ontology design and construction we considered guidelines that might match users expectation, to breadth (topic areas) and depth (hierarchy), in taxonomy construction (Gartner Research 2003b). Usability is directly
influenced in ease of comprehension and readability (cf. Section 7.4), by the number of taxonomy nodes (concepts) wide and deep. Research result from the Ford Motor Company best practice replication program (Dixon 2000), found that a nine-node wide and four-node deep design principle worked best for both a technical and process environment. As a function of domain complexity, breadth of less than twenty nodes was suggested if under user control and greater if dictated by external influences. Greater than fifteen will however exceed a user’s ability to retain a mental model and navigate within it (Dixon 2000). Depth experts agree, should be a hierarchy not greater than five levels deep, or a user’s willingness to drill down a standard business taxonomy will be impacted (Gartner Research 2003b). Ontocopi, an ontology analysis tool for communities of practice (Alani 2002), used a link threshold to limit the depth and length of discoverable paths while (Sheth 2003), opted not to limit semantic associations, finding that attempts to hide relationships can have the opposite effect and result in longer paths.

The type of business information were further categorised as context areas, where each context provided a different information contribution to competitive analysis. The context areas identified and used were Sales, Profits, Disposal, Headcount, Acquisition, Relationships, Research and Development and Market. Each context was organised into an information hierarchy, comprising a series of levels which established and ontological depth of four nodes and breadth of eight nodes (depicted in Table 4-5), consistent with the Ford recommendations.

Table 4-5 Competitive Analysis Ontology Breadth and Depth Overview

<table>
<thead>
<tr>
<th>Context (y)</th>
<th>Hierarchy - Term (t_n)</th>
<th>Level 0</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Profits</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Disposal</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Headcount</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Acquisition</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Relationships</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>R &amp; D</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Markets</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>
We also found that due to the interconnected nature of the information being looked at, domain experts themselves found that moving beyond four levels became counterproductive, leading to the class associations becoming circular in nature and path position within the information, becoming blurred. Table 4-6 provides the hierarchical view breakout of the Sales context area (cf. Section 4.2.2), with the more general but important subordinate concepts of [Product, Service, License], while for Product, its subordinate concepts are [Announce, Introduce, Revenue, Backlog] with each in turn having subordinate concepts [Delay, Cancel, Planned, Scheduled, Competitive, Development, Release, Litigation]. The maximum level depth for the ontology including the context area, used to reduce the circular nature of the path, was four. Some context areas as indicated by Table 4-5 were only three deep.

From an analyst’s perspective each tuple corresponds to some Semantic Path representing an elementary information insight. The path Product :: Announcement :: Delays within competitive analysis is interpreted as:

- when investigating company Sales, associations\textsuperscript{54} to Products are of interest.
- to attempt gain an insight into events/movements within the Product space, associations with Announcements are useful
- lastly for Announcement, associations with items that indicate what the event is about such as Development, Release, Delay and Plan are important.

\textsuperscript{54} Description used with domain experts to help simplify extraction of operational knowledge of the analyst’s analytical process. Associations represent tacit binary relationships between concepts.
Table 4-6 Competitive Analysis Ontology, Sales Context, Extract

<table>
<thead>
<tr>
<th>Context (γ)</th>
<th>Concept Hierarchy (t_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Level 1</td>
</tr>
<tr>
<td>Sales</td>
<td>Product</td>
</tr>
<tr>
<td>Sales</td>
<td>Product</td>
</tr>
<tr>
<td>Sales</td>
<td>Product</td>
</tr>
<tr>
<td>Sales</td>
<td>Product</td>
</tr>
</tbody>
</table>

For the domain knowledge contribution that outlines the possible business interpretation behind these associations and their use, refer to Section 4.2.2. Taking a sample of nineteen business filings across four enterprises selected for their business operations in the software products and services provision space, we looked at results of auto generating the number of class type annotations (meta-lexons\(^{55}\)) instances per ontology context area against the two Form types 10-Q and 20-F\(^{56}\). The results are presented in Figure 4-10. Sales [C1000] and Headcount [C7000], as expected are particularly well represented, due the majority of narrative discussion sections commenting on these topics.

The remainder of Profit [C2000], Acquisition [3000], Relationship [4000] and market [C5000] are grouped with equitable presence due to their occurrence less frequently within statement discourse. R&D [C8000] class type’s instances are less represented, receiving the smallest instance numbers, due to taking up less discussion. Disposal [C6000] did not feature owing to the absence of specific terms such as {Intangibles, technology, policy, fair value, intend to continue, subsidiary}, from corpus filings text.

---

\(^{55}\) Nomenclature used for meta-lexons was Cnumber, where ‘C’ indicated a concept and number a value assigned per concept in a particular context area.

\(^{56}\) Annual and transition report of foreign private issuers submitted to the U.S. SEC.
4.7 Summary

Despite the fact that some of the best pieces of business information contained within reports are often purposely “semantically camouflaged” from a linguistic and semantic perspective, familiarity with discourse analysis develops an expectation on what to look for and what can be expected of the text. Combining this expectation with domain knowledge, allows the generation of concept maps, providing the basis for domain linguistic modelling (lexicalisation) and ontology building. Ontologies used in information systems offer a useful step in formalising the semantics of the information represented and semantic domain of the information system, in a very concrete and useful manner (Meersman 1999). Although the generation of OWL Ontologies from concept maps has been tried in shallow domains (Borrajo, Castillo et al. 2007), and having considered various ontology development methodologies we selected the DOGMA modelling methodology (DOM), for its linguistic modelling support, database origins and translation capability from the language to the concept level (cf. Section 3.5).
Key to our use of DOGMA for CAO modelling purposes was the:

- overlay of knowledge elicitation techniques with those already established within the use case environment (Table 4-4, Section 4.5)
- leveraging of its lexon-engineering stage for domain conceptualisation
- ability to introduced our notion of role and context necessary for ontological commitment definition

The approach of making the underlying source semantic representation explicit, with the use of an ontology, is also used by semantic technologies and its web counterpart the Semantic Web (Berners-Lee 2001). Overall the Semantic Paths expressed through the CAO supports the capture of domain semantics and provides a semantic interpretative mapping of the information systems semantic domain (Meersman 1999). Having an IE schema that helps target extraction, allows movement towards its combination with natural language based linguistic processing to automate extraction.

We next present CAO development and linguistically supported Semantic Path extraction.
Chapter 5

Semantic Path Modelling and Extraction

Having established a taxonomy for business argumentation categories, and initial knowledge representations, through the use of concepts maps (cf. Section 4.2), the research question advanced to whether the identification and extraction of the information could be automated? Addressing this requires formal semantic representational modelling supported by information extraction techniques. An ontology provides both the schema to help target relevant information for extraction and a unifying model useful for post extraction semantic interpretation. With an ontology in place, information extraction mechanisms that directly support linguistic analysis can be considered.

The chapter structured adheres to these areas of competitive analysis ontology development by first considering practical ontology design considerations, relating to domain coverage, before discussing ontology construction in Section 5.1. The methodology follows that outlined in Table 4-4 (Section 4.5) and details the use and importance of DOGMA’s lexon-engineering stage. Section 5.2, details Semantic Path extraction through rule based linguistic processing of the filings disclosure sections and ontology instantiation.

The ontology acts a unifying platform that once instantiated with business text segments, provides a framework allowing semantic interpretation for propositional content assessment and evaluation.

5.1 Competitive Analysis Ontology

The business argumentation taxonomy represented an initial attempt to define and model an expectation on what an analyst can find in the text. From a semantic and linguistic perspective this information can be difficult to find unless guided by domain insight. In terms of the semantics of an information system, an ontology can represent a data model
for a narrow application domain and also provides a useful step in formalising the semantics of the represented information (Meersman 1999). Representing lexicalisation on rich semantics, the ontology helps to translate between the language and conceptual level (De Leenheer 2005a; Spyns 2007) and automate information identification and extraction. The domain coverage catering for the competitive analysis task is next discussed, along with the DOGMA ontology modelling methodology used for its construction.

Applying the DOGMA philosophy as outlined in Section 4.4 to the discourse analysis term results from Section 4.2.1, lexon engineering involving lexon creation, grounding and mapping to meta-lexons was undertaken. Engineered lexons composed from the sextuplet <(γ, ζ): t₁, r₁, r₂, t₂>\(^57\), (Spyns 2005b; Spyns 2007) consisting of intuitively plausible conceptualisations of the competitive analysis domain, are not intended to provide formal semantics or interpretation (Meersman 2001). Table 5-1 provides lexon fact examples, expressing that within context γ (Sales) and for language ζ (English), term₁ (Product) has term₂ (Announcement) occur in role r₁ with it and inversely term₂ (Announcement) maintains a co-role r₂ with term₁ (Product). Grouped by context and language, they form the ontology base (Spyns 2002). Related lexons were grouped as an organising principle to introduce the contexts previously mentioned. The terms and roles of Table 5-1 also contain meta-lexon labels as superscripts, included here to assist with explanation, used later in the section.

Table 5-1 Sales lexon (selected extract)

<table>
<thead>
<tr>
<th>Context (γ) = Sales, Language (ζ) = UK English</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>term (t₁)</strong></td>
</tr>
<tr>
<td>Product (^{C1002})</td>
</tr>
<tr>
<td>Product (^{C1002})</td>
</tr>
<tr>
<td>Announcement (^{C1005})</td>
</tr>
<tr>
<td>Announcement (^{C1005})</td>
</tr>
<tr>
<td>Announcement (^{C1005})</td>
</tr>
<tr>
<td>Announcement (^{C1005})</td>
</tr>
</tbody>
</table>

\(^{57}\) previously defined in Section 4.4
Term disambiguation by articulation $ct$ on the lexon base $ct(\gamma, t_1) = c_{t_1}$ allowed the introduction of abstract concept and relationship identifiers, natural language explanations, glosses and synsets. As the application domain requiring deep background knowledge of the discourse used, uniform use of synonyms provided by lexical resources (such as WordNet), was not possible due to variation in intended meaning. For example ‘delays [some-stop-word] introduction’ or ‘delays in market acceptance’ could be rephrased with a delay synonym of postpone, neither could be rephrased with other WordNet senses for delay, such as wait/stay/check, and still retain their intended business meaning. As discussed in Section 2.3, other domain related lexical resources such as FrameNet or XBRL taxonomies, were not of use to either generality or applicability to specific financial sub-domains. The difficulty in establishing relationship interdependencies between the CAO concepts, limits understanding of the subtlety of business text. Relationships would serve to better understand these further, enabling automatic business information extraction and interpretation leading to enhanced business insight. To date however, modelling such relationships for complex activities such as competitive analysis has had little success. Lexon grounding (Table 5-2) for each concept and relationship established abstract labels, natural language explanations and gloss synonym sets (where applicable). Labels used for lexons were neither existing language expressions (Meersman 1999) or have any specific meaning. Meta-lexons (listed in Table 5-3) adhering to the triple format of concept-relationship-concept $<c_1, r_1, c_2>$, were abstracted from lexons using the concept and relationship abstract labels. Conversion of lexons into meta-lexons copes with ideas expression in several ways, e.g. by morphology (whether inflection or conjugation allowing different forms of the same word), by synonymic wording (lexicology / terminology), or by syntax such as organising a sentence using a noun instead of a verb. The abstract label used to ground lexons, and introduced as superscript in the lexon creation table, provide both provenance and clarity from lexon to meta-lexon progression across Table 5-1 - Table 5-3.

As Form 10-Q’s are specific to the U.S. jurisdiction, English was the only language of interest. Lacking requirement for multi–lingual capability, the principle of having ontologies

---

58 http://framenet.icsi.berkeley.edu/
transcend in so far as possible, specific linguistic influences, was not been strictly adhered to. Due to the difficulty of such an exercise, some authors argue for language neutral representations rather than language independent ones (Guarino 1999). Resultantly schema mapping issues with different language usage, where words in one language do not have a direct equivalent lexical translation in another, necessitating paraphrasing, were not of concern and allowed collapsing the role and co-role into a single semantic relationship. For detail on dealing with multi-linguality within DOGMA we refer the reader to (De Bo 2003). Due to issues with synonyms introduction previously mentioned and the lack of requirement to cater for multi-lingual aspects, the combination of language and context normally associated with term disambiguation was used as the means of moving from the language to concept level.

Table 5-2 Grounding of Sales lexons (selected extract)

<table>
<thead>
<tr>
<th>Context</th>
<th>Explanation / Gloss</th>
<th>Gloss</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1002</td>
<td>Item manufactured/made</td>
<td>Sold to general public</td>
<td>Saleable item; item, goods</td>
</tr>
<tr>
<td>C1005</td>
<td>Public statement made for public consumption</td>
<td>Scheduled event</td>
<td>Inform, proclaim, advise</td>
</tr>
<tr>
<td>C1014</td>
<td>Delay in market indicating reduction in average order in take</td>
<td>Reduction in order intake</td>
<td>Market uptake</td>
</tr>
<tr>
<td>C1017</td>
<td>Date of new offering or product becoming available for general sale</td>
<td>Date of new product or offering availability</td>
<td>Discharge, waiver</td>
</tr>
<tr>
<td>C1016</td>
<td>Life cycle creation of a new product or offering</td>
<td>Investing in the creation of a new product</td>
<td>formulate; produce</td>
</tr>
<tr>
<td>C1012</td>
<td>Statement that the company will carry out a specific action</td>
<td>Statement of intent</td>
<td>Intention program</td>
</tr>
</tbody>
</table>
Table 5-2 Grounding of Sales lexons (selected extract), continued

<table>
<thead>
<tr>
<th>Label</th>
<th>Explanation</th>
<th>Lexical Triggers / Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1004</td>
<td>A plan to bring a new offering to the market place</td>
<td>A set of; a program to expand; delay; our intention; the postponement; enhance</td>
</tr>
<tr>
<td>R1014</td>
<td>Signifies that the market has been slow to accept new offering after initial announcement</td>
<td>Rate of customer acceptance; cancel or reduce size of orders; order; timely basis</td>
</tr>
<tr>
<td>R1015</td>
<td>Could signify slip in release date</td>
<td>Will occur in the future, statement of intent</td>
</tr>
<tr>
<td>R1019</td>
<td>Could signify that further development is required and that release date cannot be met</td>
<td>Longer product; cycle; resources; market acceptance; market</td>
</tr>
<tr>
<td>R1020</td>
<td>Could signify that the announced release date has not been adhered to</td>
<td>Cancelled orders; delay; new</td>
</tr>
</tbody>
</table>

Table 5-3 Creation of Sales meta-lexons (selected extract)

<table>
<thead>
<tr>
<th>C_1</th>
<th>DOGMA Meta-lexons</th>
<th>Corresponding internally used Semantic Path label</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1002</td>
<td>R1004 C1005</td>
<td>RC1002-C1005</td>
</tr>
<tr>
<td>C1002</td>
<td>R1002 C1005</td>
<td>RC1002-C1005</td>
</tr>
<tr>
<td>C1005</td>
<td>R1014 C1014</td>
<td>RC1005-C1014</td>
</tr>
<tr>
<td>C1005</td>
<td>R1019 C1016</td>
<td>RC1005-C1016</td>
</tr>
<tr>
<td>C1005</td>
<td>R1020 C1017</td>
<td>RC1005-C1017</td>
</tr>
<tr>
<td>C1005</td>
<td>R1015 C1012</td>
<td>RC1005-C1012</td>
</tr>
</tbody>
</table>

DOGMAs core principle of double articulation decomposes an ontology into an ontology base and a separate commitment layer (Spyns 2002; Spyns 2005b). Within DOGMA the selection of meta-lexons to form a commitment rule, represents the formal constraints used to add specific semantics to that selection of meta-lexons (Jarrar 2002) and instantiate the particular commitment. The direction of the lexon has also been previously applied to constraints such as ‘one driver has a constraint of at most one driver license’, for the lexon \(<\gamma, \text{driver}, \text{has}, \text{is issued to}, \text{driver’s license}>\)\(^{59}\), by applying a uniqueness constraint, UNIQ, to some path \(p\), for commitment construction, \(p = [\gamma, \text{driver}, \text{has, is issued with, driver’s license}]\).

---

\(^{59}\) i.e. one driver has drivers licenses and a driver’s license is issues to a driver
Meta-lexon selection in this manner corresponds to a Semantic Path where task based cognitive ordering, imposes an explicit constraint, in the same manner than an axiomatised selection of meta-lexons does.

Each meta-lexon tuple presented in Table 5-3 represents a **Semantic Path**

The commitment layer is responsible for mediation between the ontology base and its applications. We use it to map between the formal knowledge representation, semantic annotations and class type template instances, as part of our extraction approach, using linguistic processing, next discussed.

### 5.2 Semantic Path Extraction

Semantic Path extraction adhered to an Ontology Based Information Extraction (OBIE) approach that used linguistically supported extraction methods. Linguistic analysis was an enabler for:

i) semantic mark-up of corpus sources, used to assist functionality provision for our use case demonstrator and experimental platform

ii) MUC styled template population of Semantic Paths instances and their properties

#### 5.2.1 Semantic Path Linguistic Analysis

The General Annotation for Text Engineering (GATE)\(^6\) (Cunningham 2002; Bontcheva 2004), is a component-based general architecture and graphical development environment for natural language engineering. It allows selective inclusion of languages, processing and visual resource components such as tokenisers, semantic taggers, verb phrase chunkers and sentence splitters. GATE’s Java Annotation Patterns Engine (JAPE) provides finite state transducers over annotations, allowing grammar rule specification and recognition of regular expressions within these annotations. The annotations organised as a graph, are modelled as java sets of annotations that allows manipulation through the annotations.

---

\(^{6}\) For further information refer to http://gate.ac.uk
The last decade has seen the emergence and adaptation of light weight pattern analysis techniques based upon regular expressions, such as RegExTest, applied to the identification of Nigerian fraud emails (Gao 2005). Shallow NLP analysis techniques, in many cases involving pattern analysis and regular expressions such as the JAPE, have become mainstream (Cunningham 2002). GATE removes issues associated with format uniqueness, completeness and HTML quality, allowing effort to be concentrated on rule construction (Childovskii 1997) and follow on activities such as OBIE, e.g. technology watch (Maynard 2005). Automated class type identification, was implemented as a set of JAPE grammar rules, based on GATEs Porter stemmer for English, with synonym expansion to identify the inflected word form for annotation.

Grammar rules provided the implementation vehicle for the meta-lexons corresponding to the Semantic Path elemental building blocks. Table 5-4 provides examples of JAPE rules for recognising class types Partner, Product and Delay in market acceptance.

<table>
<thead>
<tr>
<th>Rule / Label (Annotation)</th>
<th>Term Annotation Description / Lexical Trigger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule C1005 (Partner): C1005-C1005</td>
<td>Joint venture, partnership, amalgamation, collaboration</td>
</tr>
<tr>
<td>Rule C1002 (Product): C1002-C1002</td>
<td>Product, item, goods, saleable item</td>
</tr>
<tr>
<td>Rule C1014 (DELAY) (MARKET) (ACCEPT): C1014-C1014</td>
<td>Delay in market [acceptance] Market uptake</td>
</tr>
</tbody>
</table>

The JAPE rules cope with morphological variations. General synonym expansion using WordNet produced “noise” in intermediary results, due to issues of general language knowledge applicability. Resultantly synonym expansion introduction was only possible with analyst selection and agreement, e.g. Product has synonym set {Item, Goods, Saleable item}. After this annotation stage more complex JAPE rules operating on the concept level, could be applied to add specific tags for meta-lexon instances, which indicate partial Semantic Paths. We opted however to not search for implicit relationship, leaving
them to the ontological association stage where they are based upon both semantic and business logic.

Figure 5-1 below illustrates the results of automated linguistic analysis and semantic annotation of class type instances, using the meta-lexon abstract labels. The top portion of the figure provides the source HTML produced and bottom figure, included to assist with clarity, the web browser viewable version. The first section presents the filings annotated HTML source adhering to the mark-up format of "<ahref="meta_lexon-report_line_number">". Filings report line 403, contains the class type instances for Company, Defendant, Litigation and report line 404, class type instances for Difficult to predict, Company, Management, Financial and Cash. Properties Exclude and Style are both used to support report viewer functionality (cf. Section 7.2).

Figure 5-1 Automatically added class type annotations to filing HTML source

Although it is difficult to predict the ultimate outcome of these cases, the management believes, based on discussions with counsel, that any ultimate liability would not materially affect the financial position, result of operations or cash flows.
5.2.2 Semantic Path Instance Extraction

Moving from class instance identification to Semantic Path instantiation involved:

i) template construction and population

ii) knowledge base population

iii) instantiation of Semantic Paths considered relevant.

The EDGAR extraction systems (discussed in Section 2.2) utilise template or template equivalent approaches for extraction. Template elements and interaction between, were investigated by looking at IE templates previously employed by the MUC-6 management task experiments (Chinchor 1998; Hirschmann 1998). The tasks describe ‘succession events’ outlining the levels of information required as:

i) the document containing references to the event

ii) relationships

iii) the entities (or objects) related to the template element subtask

iv) specific relationship and object information provided as features.

For the experiments, feature extraction is normally the result of intensive corpus based search, and relationships are driven from the ontology, rather than the text corpus. In effect MUC-6 looked at template construction from the perspective of the three fundamental elements of objects, relationships and their features. As Semantic Paths minimally considered features, element categories were mapped to document class and class instances. Template slots essentially act as database records, lending themselves readily to SQL operations within an RDB context, or RDF triples if using an RDF store, e.g. (Broekstra 2002; Caroll 2004; Harth 2007). An entity relationship diagram outlining the CAO physical model is given in Section 7.3.

Construction of closed Semantic Paths with specific meta-lexons in this manner ensures inherent semantics and a bounded view. In this regard the templates themselves and the Semantic Paths correspond closely to DOGMA commitments, bridging the gap between ontology base, and application layer and represent a particular view of the ontology. Here the ontology concepts themselves as defined by a Semantic Path perform
the constraint function, removing the necessity for formalised rules, as only a selection of relevant meta-lexons are needed. Essentially the Semantic Paths are themselves their own commitments. Table 5-5 below provides examples of extraction templates used for ontology population (described in Section 5.2.2).

Table 5-5 Class Instance templates (example 1)

<table>
<thead>
<tr>
<th>Class Type :: Legal</th>
</tr>
</thead>
<tbody>
<tr>
<td>DocId</td>
</tr>
<tr>
<td>ReportName</td>
</tr>
<tr>
<td>Term</td>
</tr>
<tr>
<td>AnnotationOffset</td>
</tr>
<tr>
<td>ReportLineNo</td>
</tr>
<tr>
<td>InstanceID</td>
</tr>
<tr>
<td>InfoItemText</td>
</tr>
<tr>
<td>LinkedConceptHead</td>
</tr>
<tr>
<td>LinkedRelationship</td>
</tr>
<tr>
<td>LinkedConceptTail</td>
</tr>
</tbody>
</table>

Table 5-6 Class Instance templates, (example 2)

<table>
<thead>
<tr>
<th>Class Type :: Delay in market acceptance</th>
</tr>
</thead>
<tbody>
<tr>
<td>DocId</td>
</tr>
<tr>
<td>ReportName</td>
</tr>
<tr>
<td>Term</td>
</tr>
<tr>
<td>AnnotationOffset</td>
</tr>
<tr>
<td>ReportLineNo</td>
</tr>
<tr>
<td>InstanceID</td>
</tr>
<tr>
<td>InfoItemText</td>
</tr>
<tr>
<td>LinkedConceptHead</td>
</tr>
<tr>
<td>LinkedRelationship</td>
</tr>
<tr>
<td>LinkedConceptTail</td>
</tr>
</tbody>
</table>

Instantiated templates are built for multiple class type instances, from which specific IE actions and expert system reasoning (here competitive business analysis) can be
performed, e.g. checking of whether or not a company is involved or may be involved in legal dispute. Moving from instantiated extraction template to Semantic Paths instantiation had to accommodate situations where the same or very similar sentences are reused in multiple areas, within the same filing, and also observed across different reporting periods for the same company.

A difficulty faced, is catering for occurrences of the same class type across multiple context areas and hierarchies within the ontology. Table 5-7 based on the first and second sentence from Figure 5-1, illustrates both scenarios that requires evaluation of multiple possible combinations of Semantic Paths. Head and tail terms are represented by their conceptual meta-lexon label, along with the text segment, that both were found to co-occur within. Considering the annotated text segment for the first tuple, split into the three sentences:

i) “Legal Proceedings”, containing the class types: Legal [C4002]

ii) “The Company is a defendant in various matters of litigation generally arising out of the normal course of business”, containing the class types: Company [C1036], Defendant [C4018] and Litigation [C4019]

iii) “Although it is difficult to predict the ultimate outcome of these cases, management believes, based on discussions with counsel, that any ultimate liability would not materially affect the Company’s financial position, result of operations or cash flows” containing the class types: Difficult to predict [C1032], Management [C3005], Believe [C5012], Company [C1036], Financial [C5015], Position [C5021] and Cash [3029].
Table 5-7 Multiply occurring instantiated Semantic Paths for RC4002-C4018, RC4002-C4019

<table>
<thead>
<tr>
<th>Head Term</th>
<th>Tail Term</th>
<th>Rep. Line</th>
<th>Seg. Id</th>
<th>Annotated Text Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4018</td>
<td>C4002</td>
<td>403</td>
<td>508070</td>
<td>Legal Proceedings The Company is a defendant in various matters of litigation generally arising out of the normal course of business. Although it is difficult to predict the ultimate outcome of these cases, management believes, based on discussions with counsel, that any ultimate liability would not materially affect the Company’s financial position, result of operations or cash flows.</td>
</tr>
<tr>
<td>C4019</td>
<td>C4002</td>
<td>403</td>
<td>508080</td>
<td>Legal Proceedings The Company is a defendant in various matters of litigation generally arising out of the normal course of business. Although it is difficult to predict the ultimate outcome of these cases, management believes, based on discussions with counsel, that any ultimate liability would not materially affect the Company’s financial position, result of operations or cash flows.</td>
</tr>
<tr>
<td>C4018</td>
<td>C4002</td>
<td>1182</td>
<td>528440</td>
<td>Legal Proceedings The Company is a defendant in various matters of litigation generally arising out of the normal course of business. Although it is difficult to predict the ultimate outcome of these cases, management believes, based on discussions with counsel, that any ultimate liability would not materially affect the Company’s financial position, result of operations or cash flows.</td>
</tr>
<tr>
<td>C4019</td>
<td>C4002</td>
<td>1182</td>
<td>528450</td>
<td>Legal Proceedings The Company is a defendant in various matters of litigation generally arising out of the normal course of business. Although it is difficult to predict the ultimate outcome of these cases, management believes, based on discussions with counsel, that any ultimate liability would not materially affect the Company’s financial position, result of operations or cash flows.</td>
</tr>
</tbody>
</table>

Semantic Path constituent possibilities within the same sentence, are class types Company [C1036], Defendant [C4018] and Litigation [C4019]. However querying the CAO reveals a possible twenty nine Semantic Paths occurrences (Table 5-8), where the class types can validly occur, as part of four different context areas. The class type Company [1036] as an example in the first Semantic Path tuple, can be found occurring in the Semantic Path [C1010, C1036] within context areas Sales, Acquisition, Relationship, Disposal, each holding out the possibility of providing different interpretations and insight of the same data. Company does however not co-occur with either of the class types Litigation or Defendant.
Table 5-8 Semantic Paths Containing Class Terms: [Company, Defendant, Litigation]

<table>
<thead>
<tr>
<th>Context [Term Label]</th>
<th>Lexon/Term Label</th>
<th>Semantic Paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1000 [Sales]</td>
<td>C1004 [Licensing]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C1002 [Product]</td>
<td>C1010 [Agreement]</td>
</tr>
<tr>
<td></td>
<td>C1003 [Service]</td>
<td>C1006 [Introduce]</td>
</tr>
<tr>
<td></td>
<td>C3000 [Acquisition]</td>
<td>C1010 [Agreement]</td>
</tr>
<tr>
<td></td>
<td>C1010 [Agreement]</td>
<td>C1019 [3rd Party]</td>
</tr>
<tr>
<td></td>
<td>C4000 [Relationship]</td>
<td>C1019 [3rd Party]</td>
</tr>
<tr>
<td></td>
<td>C4002 [Legal]</td>
<td>C4018 [Defendant]</td>
</tr>
<tr>
<td></td>
<td>C4002 [Legal]</td>
<td>C4019 [Litigation]</td>
</tr>
<tr>
<td></td>
<td>C6000 [Disposal]</td>
<td>C1010 [Agreement]</td>
</tr>
<tr>
<td></td>
<td>C1010 [Agreement]</td>
<td>C1036 [Company]</td>
</tr>
<tr>
<td></td>
<td>C3028 [Divest]</td>
<td>C1019 [3rd Party]</td>
</tr>
</tbody>
</table>

Drawing upon previous approaches used in EDGAR filings extraction of keyword term association (Grant 2006), hierarchical functional dependency (Gerdes 2003), and hierarchical pairwise search terms (Stampert 2008), we empirically evaluated the defined heuristic of class type associations on sentence distance, as a basis to drive Semantic Paths instantiations. Termed the relationship distance range (RDR), the resulting values of (-0, +/- 1) also correlated with domain expert experiences of class term associations, based on occurrence proximity clustering. Revisiting Table 5-7, with an RDR value of -1, adds the class type Legal [C4002] to the existing set of Company [C1036], Defendant [C4018] and Litigation [C4019]. Now we find that both Defendant [C4018] and Litigation [C4019], co-occur with Legal [C4002], across multiple Semantic paths within the Relationship [C4000] context.
A data cube (Figure 5-2) generated from all possible CAO class type permutations, their extraction template properties and RDR=1 as parameter, was used to identify Semantic Path instantiation candidates. The cube generates per filing processed, the Semantic Path as a relation between: i) class types; ii) the individual class types and iii) report line where they occur within the filing and associated text segment. These represent the system generated instantiated Semantic Paths that are presented to the analyst, as potentially relevant information to search, assess and extract from.

To provide some further level of context for the end user, Semantic Path rich sentences were extracted as part of a text segment, based on the RDR range, i.e. current, previous and next sentences. The approach is similar to other EDGAR extraction systems that typically extract blocks of structured text from balance sheet Sections e.g. (Stampert 2008), other specific report section (Grant 2006) or unstructured text blocks (Gerdes 2003). The means taken to resolve the multiple sentence issue was to instantiate all paths across the various contexts and let the user decide, which to consider, using the context view functionality provided by the use case demonstrator.

Mapping between formal knowledge representation and the semantic annotations remains a problematic area with suggestions that decoupling of the IE component from the
annotation tool would provide greater flexibility to deal with this gap. A suggested solution is the use of knowledge acquisition rules (KAR), that specifically map conceptual tags, with an element of the domain ontology (Amardeilh 2006). From a methodology perspective such an approach provides linkage from domain linguistic analysis to the semantic domain model, enabling follow on annotation and ontology population. We adhered to the KAR approach by consistent use of meta-lexon abstract class type labels as a federated approach for semantic mark-up of class type instances, grammar rules descriptions, ontology base and its commitments. The approach allowed the common interlinkage of linguistic analysis, semantically annotated sources and CAO enabling Semantic Paths navigation functionality in the use case demonstrator (cf. Chapter 7) and its experimental setup, next discussed.

5.3 Summary

Linguistically modelling the competitive analysis domain adhered to a two stage incremental knowledge representation approach. A corpus of U.S. SEC form 10-Q, 10-K and 20-F filings were used to develop a taxonomy that represented eight business argumentation categories and their accompanying concepts maps. Argumentation categories were used to define 149 domain specific class types. Concept maps organised the class terms into a concept hierarchy, using business context areas to define the ontological breadth of eight, and hierarchy depth of four nodes. A key part of this progressive linguistic modelling was the introduction of domain knowledge relating to the competitive analysis task itself, and its demands and constraints on the information requirement. The Semantic Paths represent this domain knowledge capture, where domain linguistics and semantics are conceptualised, using the domain conceptualisation stage of the DOGMA modelling methodology, to develop the CAO. The CAO schema is subsequently used as the basis for extraction rule specification using GATE’s Java Annotation Patterns Engine and linguistic processing drives class type instance identification, template based extraction and Semantic Path instantiation. Instantiation is dependent on a combination of ontology hierarchy and class term proximity within the originating filing.

The ontology acting as a framework allows semantic interpretation and in-context visualisation of Semantic Paths supporting the search for and identification of relevant text segments. Fundamental to this is automating the analyst information requirement through
automated ontology instantiation supported by linguistic analysis. With analysts from the outset facing the difficulty of “only having a broad understanding as to what they want but don’t really know what that is until it’s found!”\(^{61}\), automating the information provision, in a manner that allows structured traversal through the information space, represents a huge assist in finding those relevant nuggets of information (Korman 1998; Zhang 2004), for insight and decision support activity.

Using the combination of CAO with linguistic analysis to instantiate our use case artefact demonstrator, the Analyst Work Bench provided the experimental platform for research hypothesis evaluation. Experimental methodology, fundamentals, setup and results are next presented and the implementation of the approach within the Analyst Work Bench use case demonstrator later in Chapter 7.

\(^{61}\) Quote from participating business analysts during one of the contextual enquiry sessions
Chapter 6

“*If your experiment needs statistics, you ought to have done a better experiment*”

*Ernest Rutherford*

**Experiment**

Information systems evaluation can be considered as a means of asking questions about system performance in relation to objectives, divided along system or user centred evaluation lines (Saracevic 1975; Saracevic 1995). System focussed evaluation targets retrieval method effectiveness, content and coverage, along with algorithm performance. User related evaluations, access output related characteristics, such as interaction, search, and feedback, user questioning of application to task, including fitness for purpose and social aspects through environmental impact in areas such as decision making (Saracevic 1975; Saracevic 1995). Here we concentrate on system based performance using Saracevic evaluation performance criteria as the basis for our evaluation methodology (Saracevic 1995). The criteria description along with examples and mapping to our experimental context, serves to provide a fundamental basis as to what must be addressed, before evaluation takes place. The criteria require:

i) *a system and process*, e.g. a prototype, algorithm/simulation, here the Analyst Work Bench demonstrator platform

ii) *criteria representing the system objectives*, e.g. relevance, here defined by experts within the domain of interest

iii) *measuring instrument*, e.g. relevance judgements, usability, here expressed through domain expert assessment of class type annotations

iv) *measures* based upon the criteria, e.g. accuracy, heuristics, here IR metrics of precision, recall and F-measure

---

62 A usability evaluation based upon heuristic measurement of the use case demonstrator, is presented in Section 7.4.
v) methodology for obtaining measurement and performing the evaluation, e.g. TREC procedures, task based activities, here the competitive analysis task.

The criteria are next discussed, followed by experimental design, execution and steps taken where possible to reduce experimental error. Freidmans evaluation criteria for natural language processing was also drawn upon, to ensure adequate consideration and coverage of the more general aspects relating to reference standards construction and bias introduction (Friedman 1995; Friedman 1998). Lastly experimental results and discussion are presented.

We first outline experiment fundamentals in terms of the research impact from the original use case requirement, research assumptions adhered to, and evaluation objectives.

**6.1 Experiment Fundamentals**

Due to the broad nature of competitive analysis, the case study was scoped to facilitate business requirement that closely aligned to the nature of the activity itself, with technology selection influenced by possible end user operational environment. The dimensions used to categorise the use case requirements were:

- **domain of Interest**, with a requirement to cater for US SEC consolidated financial reports. The research impact was to limit the linguistic capability to cater for the English language only, with filings preference confined to the Form 10-Q

- **training data**, with the knowledge that cost and resource availability barriers existed for training corpus creation and the requirement that an un-annotated corpus would have to be catered for. The research implication was that the knowledge engineering, grammar rule approach would be the more applicable

- **domain modelling**, with a requirement that the developed model be cognitively representative of the information requirement of the competitive analysis task. The research impact was to ensure that the selected model could cater for domain linguistic representation based on a business specific term vocabulary also used to assist with grammar rule creation

- **task complexity**, with the requirement that an experienced analyst be the expected end user focus of the system. Cost benefit analysis identified a high cost to entry for
analyst engagement in evaluation, and experimental breath would consequently have to reflect availability. Research impact was to restrict evaluation breath to ensure the maximum level of domain expert participation, that was practical to achieve

- knowledge repository, where the requirement was that the repository should fit existing technology environment and demonstrate scalability, maintainability and supportability. As triple stores at the time were immature, the research implication was to instantiate the model in an RDB.

The Semantic Paths and CAO were modelled over a period of nine months with senior domain experts and iteratively refined, until agreement was reached across contributors. In this way the CAO construction represented better than per chance agreement, negating the need to establish inter-rater agreement and allowing adherence to the following assumptions that the:

i) CAO is well developed and accurately reflects the domains linguistic model

ii) Semantic Paths inherent in the CAO reflect the information requirement for an analyst performing the subjective aspect of an analysis task

iii) Semantic Paths inherent in the CAO accurately reflect the cognitive aspects of an analysts information association activity, in performing an analysis task

iv) corpus and class type manual annotation used as the reference standard is the best obtainable, with resource limitations, but accurately reflects the manual information gathering aspect of the analysis task.

The experimental objective is the assessment of the research hypothesis [RH] from Section 1.2, here restated as whether the CAO as part of an information extraction system assists the identification and extraction of text segments relevant to the performance of a competitive analysis? To support the objective, the evaluation seeks to generate IR performance metrics (cf. Section 6.2.1), based on the use of relevance judgements as the measuring instrument (cf. Section 6.2.2).
6.2 Experiment Criteria

6.2.1 Measuring Instrument

Within IR, relevance as a concept infrequently appears within its own right and is most often encountered with the definition of precision and recall measures. Typically consideration of relevance does not extend beyond recognition that a document is relevant, if the stated information needs are addressed. The predominant model of IR subscribed to is one of system/source or user/destination (Saracevic 1975), where emphasis is placed on matching queries to information objects processed by some system. Algorithmically based, the goal sought, is to maximise retrieval of relevant information. In the strictest sense the query is representative of the user and other user considerations, such as interaction are not included (Saracevic 1975).

Used originally by the Cranfield experiments from 1950 - 1966 (Cleverdon 1967), the model and metrics have been in standard use since, most notably for accessing different aspects of Information Retrieval systems as part of the TREC (1992-2003) standardisation of retrieval evaluation (Voorhees 2005). The Cranfield methodology compares systems based upon the central assumptions, that relevance judgements are complete and relevance can be approximated by topical similarity. Cranfield expects all documents within the collection, to receive unbiased relevance judgements, for all topics from expert evaluators. Cranfield is considered as the archetypal approach for the system view of relevance (Hjorland 2010), but with the emergence of user based systems, the concept of relevance has shifted away from the system view, to one of user tied to human information behaviour (Schamber 1990). Saracevic, in 1975 provided one of the earlier and more comprehensive reviews on relevance with his subject knowledge or domain view (Saracevic 1975). The view considers accumulated human knowledge from familiarity with, and knowledge of, subject terminology, along with highly knowledgeable users or domain experts, as the central variable in achieving high agreement and relevance ratings. In synthesized relevance definitions, Saracevic subscribes to utility as one definition noting that an IR systems should provide information that has utility, rather than just being relevant (Saracevic 1975; Schamber 1990) and that only users may judge the relevance of the documents and their use in a subjective manner. Schamber re-examined relevance considering its meaning and
role within information behaviour, through its historical progression from the broader notions of system topicality based relevance, to that of user relevance based on user and user information quest (Schamber 1990). Like Saracevic, Schamber also considered the traditional algorithmic model to be too static for use as a conceptual framework, for human relevance judgment exploration, and offered an alternative dynamic situational approach where the user is also central to the relevance judgement process. Schamber closely follows Saracevic observing “that relevance depends on human judgement of relationship between certain aspects of information and information need” and that a full understanding of relevance can only be arrived at when the perceptions of the end user in real information need situations are focussed on (Schamber 1990). Schamber went on to conclude that relevance was a:

- multidimensional cognitive concept, with meaning dependent upon how users perceive information, relative to their information needs situation

- dynamic concept dependent upon users judgement of the quality of relationship between information and information need (at a certain point in time) and

- complex but systematic and measurable concept, if approached conceptually and operationally from the users perspective (Schamber 1990).

Hjørland used the subject knowledge view as a foundation to revisit the concept of relevance within IR and information science (Hjorland 2010). Hjørland considers the previous distinction between system’s relevance and user’s relevance defunct, as “relevance is only meaningful in relation to goals and tasks, and machines do not have goals". Relevance is positioned as relating to task, if it supports the fulfilment of that task and ‘something’ such as knowledge is relevant to a task, if the likelihood of achieving the goal, implies by the task is increased (Hjorland 2010). In effect knowledge relevance is dependent upon its usefulness to achieve specific goals. Like Saracevic, differentiation is made between users and subject knowledge expert. Relevance determination in relation to given tasks requires subject knowledge, which in turn is also dependent on differing viewpoints, ensuring that users of information systems should not be automatically considered competent to judge relevance.
Within our experiment, the intricacies of the information requirement and the varying interpretation that different subject matter experts can draw from the same text segments makes achieving consensus, on what is relevant and exactly how, at best difficult. In effect the situation faced when performing an evaluation is that the users understanding of what is relevant can shift depending on Semantic Path constituents. Previous discussion supports the use of subject matter experts over the general user and emphasises the importance of task contribution in relevance determination. Adhering to our interpretation of relevance, within the tradition of Cranfield, we defined *relevance* for the purpose of the analysis activity evaluation as being:

> “Any text segment identified using class annotations possessing an information value likely to be of use either individually or in association with other text segments that could provide competitive analysis insight”

Taking into account the user centrality to the judgement process, we further note, the imperative of using evaluators with domain expertise for both:

i) reference standard establishment and

ii) relevance judgement assessment to determine whether the information requirement has been satisfied in terms of task activity.

Accessing relevance in this manner will ensure its role as a significant causal factor in effectiveness measurement. We next discuss the measures used for determining performance and what they mean in practice for our evaluation.

6.2.2 Measures

The IR set based metrics of precision (P) and recall (R) were used to evaluate performance. Precision is the fraction of retrieved documents that are relevant while recall is the fraction of relevant documents that are retrieved (Baeza-Yates 1999). P-R may also be considered in terms of statistical classification task outlined in Table 6-1.
Table 6-1 Set representation of Precision, Recall (Baeza-Yates 1999)

<table>
<thead>
<tr>
<th>Relevant</th>
<th>Not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>true positives ( (t_p) )</td>
</tr>
<tr>
<td>Not retrieved</td>
<td>false negatives ( (f_n) )</td>
</tr>
</tbody>
</table>

And be alternatively represented as belonging to the class of positives (Manning 2009), allowing the metrics and their harmonic mean to be re-stated as:

\[
P = \frac{tp}{tp + fp} \quad R = \frac{tp}{tp + fn} \quad F = \frac{(1 + \beta^2)PR}{\beta^2 P + R}
\]

Where in a general context:

- true positives \( t_p \), are the number of relevant documents in the reference standard
- false positives \( f_p \), the difference between documents generated by some system and the reference standard
- false negatives \( f_n \), are the difference between reference standard documents and those generated from the system
- \( F \) or F-Measure, is the harmonic mean between precision and recall. It ranges from 0, indicating no relevant document retrieval, to 1, indicating complete relevant document retrieval. \( F \) can be interpreted as looking to find the best compromise between \( P \) and \( R \) (Baeza-Yates 1999). \( \beta^2 \) is an indication of the relative importance of either precision \( (\beta^2 < 1) \), or recall \( (\beta^2 > 1) \), to the system stakeholders.

6.2.3 Competitive Analysis Task

As discussed in Section 4.1 competitive analysis has been previously categorized as comprising both mechanical and analysis tasks (Debreceny 2001). Mechanical refers to the preliminary manual work of systematically going through a business filing to locate, annotate and correlate text segments of interest, before using them as input for the analytical phase. The process in practical terms involves the use of trigger terms based upon...
proximity as defined by the Semantic Paths. The second phase of a competitive analysis represents for the analyst, a more subjective and fluid information need and search requirement. Often possessing only a vague idea of what may be found in the data, the analyst typically performs search and discovery, as an iterative process where previously found relevant text segments, guide the direction of what may next be looked for. At a basic level measurement of retrieval effectiveness (as dictated by Cranfield) requires some document collection, an information need (expressible normally as a query), and a set of judgements on relevance of documents to queries (Manning 2009). P-R can then be based on the symmetric difference between the reference standard and a given information query. Cranfield has three major simplifying assumptions summarised in Table 6-2 (Voorhees 2001). The table also details for each, its implications and how practicality aspects of the CAO evaluation impacts on them.

In evaluation terms, automated corpus processing by the Analyst Work Bench replaces the manual aspect of the competitive analysis. Automated ontological instantiation, additionally provides a system generated annotation set reference standard. Identification of text segments through CAO navigation, within the Analyst Work Bench, and within this context is representative of the information query. Reasons outlined above, point to difficulty establishing relevance based on topic similarity. For competitive analysis, analysts typically require that information segments be progressively grouped, such that the aggregate may result in drawing out propositional content. In effect the relevance of one text segment can be dependent on another. While the user information need is bounded, it cannot be said to be static, as analysts will naturally introduce bias in text segments selection, based on perceived relevance, resulting from their training and experience. Although text segments will be assigned a relevance rating, based on binary judgement, we do not consider this as a deviation from the third Cranfield assumption, as this has been the general trend in retrieval evaluations.
Table 6-2 CAO evaluation impact on Cranfield assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Implication</th>
<th>CAO Evaluation Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Relevance can be approximated by topical similarity</td>
<td>Relevant documents (all) are equally desirable</td>
<td>Documents may have to be grouped for relevance</td>
</tr>
<tr>
<td></td>
<td>Relevance of one document is independent of another</td>
<td>Relevance of one document can be dependent on others</td>
</tr>
<tr>
<td></td>
<td>User information need is static</td>
<td>Information need is relatively unspecified and subject to change</td>
</tr>
<tr>
<td>2) Single set of judgments for a topic is representative of the user population</td>
<td>Judgements are complete and unbiased and relative to the context at hand</td>
<td>Judgements vary across experts, due to perceived importance of different topics</td>
</tr>
<tr>
<td>3) The list of relevant documents for each topic is complete (all relevant documents are known)</td>
<td>Originally Cranfield experiments used sliding (five-point) relevance scale</td>
<td>Relevance is a binary choice determination</td>
</tr>
</tbody>
</table>

A more accurate and robust approach requires comparison of competitive analysis results, based on the established manual method, to that using the Analyst Work Bench. Figure 6-1 below serves to illustrate this in terms of required and generated reference standards (sets of class based annotations)⁶³. There is the set of annotations resulting from the manual reference standard creation, the reference set auto-generated by the system and finally the sub-set resulting from analyst selection from the system reference set. For clarity, these are respectively referred to as manual, system and system usage annotation sets.

The competitive analysis activity and accompanying annotated reference sets, allows a white and black box approach to evaluation. White box evaluation, a software testing approach, concerned with an application’s internal workings, whereas black box concentrates on functionality.

---

⁶³ Relationship of P-R across the sets also also included.
We combine available reference standards, white/black box approach’s, along with analysis activities, to define our approach for performance metrics generation, outlined in Figure 6-2 below. With task activity central, the manual activity is represented as a:

- preparatory phase, where the analyst gathers potentially relevant information, corresponding to the analysis gathering activity
- usage phase, allowing analysis of collected information for relevant propositional content, corresponding to the analysis activity

White box evaluation is associated with:

- manually generated annotations relative to system generated annotations (1 in Figure 6-2) and
- system generated annotation relative to tool usage generated annotations (2 in Figure 6-2)

Having a specific dependence on system generated annotations, these metrics reflect \textit{a-priori} activity, whereas black box evaluation based on user experience that leverages system output, as \textit{a-posteriori} (3 in Figure 6-2).
6.3 Experimental Setup

Despite our evaluation not being concerned with term frequency\(^6\) it was dependent on term linguistic processing for Semantic Paths instantiation, and establishing what’s happening at the discourse level. Cunningham emphasises a dimensional approach where the parameters of text type and domain, should be considered, when creating a performance profile (Cunningham 2005b). We based our evaluation profile on the dimensions of text genre, represented by the financial filings and subject areas, represented by the competitive analysis task. Both were integrated as part of experimental setup for reference standards creation, rater consistency determination and the identification of areas where errors could be introduced.

\(^6\) Discussion with analysts established that it was no indicator of significance or trend
6.3.1 Reference Standard Creation

The domain application area required that highly specialised domain experts could only participate in the evaluation (discussed in Section 6.2.2). Due to availability restrictions (outlined in Section 6.1), the evaluation was limited to a pool of five evaluators, all of whom had prior specialised competitive analysis experience within the business processing outsourcing area. The number is sufficient to avoid overlap between system training and later stage evaluation. Using competitive analysis to analyse company filings for new business opportunity, was the use case context and application area. Evaluator background profiles reflected a compliment of business, accounting, economics, and investment analysis.

Domain experts were asked to perform a manual competitive analysis on a selected Form 10-Q business filing, by annotating relevant text segments, and highlighting key terms where appropriate. Form 10-Q Extract 6-1 illustrates manual class annotations, which in CAO abstract type and label combination contains the classes:

- C1036 [Company], C1004 [License] and C1020 [price] in the first underlined sentence and
- C1038 [Existing] C5001 [Customer], C1040 [Cancel], C1004 [License] and C1002 [Product] in the second.

Form 10-Q Extract 6-1 manual class annotation extract (1)

For example, our ability to market the Citrix MetaFrame Access Suite, and its individual products including: Citrix MetaFrame Presentation Server, Citrix MetaFrame Secure Access Manager, Citrix MetaFrame Conferencing Manager, the Citrix MetaFrame Password Manager and other future product offerings could be affected by Microsoft’s licensing and pricing scheme for client devices, servers and applications. Further, the announcement of the release, and the actual release of new Windows-based server operating systems or products incorporating similar features to our products could cause our existing and potential customers to postpone or cancel plans to license certain of our existing and future product offerings.

Form 10-Q Extract 6-2 provides further example with the circled area contains multiple classes types of C1036 [Company], C4018 [Defendant], and C4019 [Litigation], with adjoining sentence (not highlighted), the class types C1032 [Difficult to predict], C3005 [Management], C5015 [Financial] and C3029 [Cash].
Friedman’s criteria for natural language processing was used to address the more general consideration involved in reference standard creation (Friedman 1995; Friedman 1998). Table 6-3 lists the criteria for reference standard establishment and how we addressed them.

**Table 6-3 NLP criteria for reference standard establishment (Friedman 1995)**

<table>
<thead>
<tr>
<th>Reference Standard Establishment</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>If domain experts are used to determine the reference standard, there should be a sufficient number to assess variability of the reference standard</td>
<td>Five domain experts were used and rater agreement determined</td>
</tr>
<tr>
<td>The test set should be large enough in that there is sufficient power to distinguish levels of performance</td>
<td>Due to task complexities and resource availability of domain experts the corpus was restricted to a single Form 10-Q</td>
</tr>
<tr>
<td>The choice of the reference standard should be based on the objectives of the study</td>
<td>Directly linked to both research questions and hypothesis.</td>
</tr>
<tr>
<td>If domain experts are used to determine the reference standard, the type of expert should be appropriate</td>
<td>The domain experts has considerable senior level experience, understanding and insight into the targeted topic areas</td>
</tr>
<tr>
<td>The method used to determine the reference standard should be clearly described, particularly if domain experts were used</td>
<td>Outlined in this section</td>
</tr>
<tr>
<td>Describe the manner in which the test documents were chosen</td>
<td>Primary filing widely used by the community and company in question, for business opportunity identification</td>
</tr>
</tbody>
</table>

Using the same domain experts for both reference standards creation and then evaluation, based on that reference standard, has previously been shown to introduce bias relating to reference standards usage (Will 1994). Such judgement inconsistencies were addressed by ensuring a sufficient number of annotators (Friedman 1995; Friedman 1998), here a focus group of five. We also minimised bias introduction by ensuring a sufficient period had elapsed between standards creation and evaluation taking place. Freidmans criteria for minimising bias with NLP use (listed in Table 6-4 below) was also used to address
the more general aspects of bias introduction and avoid the tendency of domain experts to highlight system strengths and avoid weaknesses.

Table 6-4 NLP criteria for minimising bias, extracted from (Friedman 1995)

<table>
<thead>
<tr>
<th>Minimising Bias</th>
<th>Addressed By</th>
</tr>
</thead>
<tbody>
<tr>
<td>The developer should not see the test set of documents</td>
<td>Publically available documents from <a href="http://www.sec.gov/">http://www.sec.gov/</a></td>
</tr>
<tr>
<td>If domain experts are used to determine the reference standard they should not be developers of the system or designers of the study</td>
<td>Owing to the complexity of domain knowledge required, domain experts were consulted on study construction, but only in the context of performing the actual competitive analysis task.</td>
</tr>
<tr>
<td>The developer should not perform the evaluation</td>
<td>Explicitly performed by domain experts only, of which the system developer was not</td>
</tr>
<tr>
<td>The NLP system should be frozen prior to evaluation</td>
<td>Rule set frozen twelve months prior to the evaluation</td>
</tr>
<tr>
<td>Ideally, the person designing the evaluation study should not be a developer of the system.</td>
<td>They were, but only in a procedural and instruction capacity, and not in terms of what to identify and annotate as relevant</td>
</tr>
</tbody>
</table>

Binary relevance judgements of annotations were used as an efficient means of achieving an unbiased relevance set. Manual annotations were normalised and consolidated as class types, to generate the manually annotated reference standard. Linguistic analysis of the corpus responsible for automated class annotations, and generation of Semantic Path instantiation candidates (detailed in Section 5.2.2), also served as the system generated reference standard. Traversal of the instantiated paths using the Analyst Work Bench navigation function (cf. Section 7.2), allowed identification and annotation of relevant text segments for extraction. Here as in the manual reference standard, annotations were consolidated across evaluators, generating the system usage reference standard. The statistics summary from the CAO and automated processing of the corpus is provided in Table 6-5.
Table 6-5 CAO and Corpus Statistics Summaries

<table>
<thead>
<tr>
<th>Ontology</th>
<th>#</th>
<th>Corpus</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width (nodes)</td>
<td>8</td>
<td>Filing page size (formatted/actual)</td>
<td>48/57</td>
</tr>
<tr>
<td>Depth (nodes)</td>
<td>4</td>
<td>Tokens</td>
<td>49994</td>
</tr>
<tr>
<td>Class types</td>
<td>149</td>
<td>Marked up report lines (HTML)</td>
<td>11346</td>
</tr>
<tr>
<td>Semantic Path permutations</td>
<td>727</td>
<td>Linguistically processed lines</td>
<td>1212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automatically assigned class types</td>
<td>2469</td>
</tr>
</tbody>
</table>

6.3.2 Rater Consistency Agreement

The possibility of encountering annotator inconsistencies, when manually identifying relevant items, is representative of an environment, in which annotators are not required to assign categories to all cases within the corpus. Having five annotators ruled out the use of pair based, coefficient consistency generators, such as Pearson, Spearman, or Cohen’s kappa (Baeza-Yates 1999; Sim 2005) for establishing inter-rater agreement. The Cronbach alpha consistency estimate (Cronbach 1951), does cater for multiple raters but requiring raters assign a rating to each case, and its lack of robustness to missing data also rules out its use. Randolphins free-marginal multi-rater kappa (Randolph 2005), catering for an unlimited numbers of cases, categories and raters was used to calculate inter-rater agreement. Free-marginal kappa is suggested where category assignment to all cases is not required (Brennan and Prediger 1981). Annotator agreement and kappa were assessed by two different means. The first looked at the annotation of text segment normalisation to Semantic Path constituents, and the second normalised annotated text segment to the filings reporting line. Kappa values and percentage of annotator overall agreement, from the application of Randolphins kappa calculator, to the union of individual annotations is provide in Table 6-6. The annotation of text segments with an annotator agreement level of 20% (1 in 5) for 338 annotation, reported a percentage overall agreement of 50.76% and a free marginal kappa of 0.006. For annotation of text segments where there was a 60% (3 of 5) agreement level, representing 136 annotation tuples, a 50.44% overall agreement and

---

65 Online kappa calculator available from [http://justusrandolph.net/kappa/](http://justusrandolph.net/kappa/)
kappa of -0.17 was recorded. Using the normalised to reporting line approach, there were 70 distinct tuples with an annotator agreement of 20% (1 in 5), resulting in an overall percentage agreement of 51.1, and fixed marginal kappa of -0.013. Looking at annotator agreement levels of 60% (3 of 5), representing only 22 annotation tuples, the overall percentage agreement drops marginally to 49% and fixed marginal kappa to -0.17.

Table 6-6 Manual annotator percentage of agreement and consistency estimates

<table>
<thead>
<tr>
<th>Annotations / Segment</th>
<th>Annotator Agreement level</th>
<th>No. Tuples</th>
<th>Percentage of overall agreement</th>
<th>Fixed-marginal kappa</th>
<th>Free-marginal kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1+</td>
<td></td>
<td>338</td>
<td>50.76</td>
<td>0.006922</td>
<td>0.0153</td>
</tr>
<tr>
<td>2+</td>
<td></td>
<td>226</td>
<td>46.28</td>
<td>-0.1023</td>
<td>-0.074</td>
</tr>
<tr>
<td>3+</td>
<td></td>
<td>136</td>
<td>50.44</td>
<td>-0.1734</td>
<td>0.0088</td>
</tr>
</tbody>
</table>

A rule of thumb suggests kappa of 0.70 indicates adequate rater agreement (Landis 1977), even though 0.4 to 0.75 has also been categorised as fair to good, and below 0.4 as poor (Fleiss 2004). Previous ontology learning evaluations (Kavalec 2005), which also encountering low kappa values, looked to annotator per class agreement (Poesio 1998), to establish reasonable agreement, and create benchmarks based on full rater (100%), or substantial rater agreement (60%). With kappa falling outside of the fair to good range, per-class agreement was also considered for reasonable agreement. As rater assignment were based on individual subjective case assessment, rather than consistently assigned rater ratings, per-class agreement could not be applied. Both kappa and percentage of agreement highlight the problems faced when trying to establish agreement between annotators, for non-proscriptive subjective based retrieval type evaluations - such as competitive analysis.
Task complexity and influences from rater background profiles are contributory factors, to the best overall agreement of 51.1%, for manual reference standard creation.

With the percentage of agreement based on five expert raters, there was no concern for the statistics being artificially inflated. Reasonable agreement did however require a substantially higher level of agreement (4/5 or 80%) among raters for annotations inclusion in the manual baseline. With that in mind, we selected the second approach with its percentage agreement of 51%, and greater coverage of 142 tuples (70 distinct), for use as the manual evaluation base line.

Both kappa and percentage of overall all agreement, highlight how annotators themselves, despite being domain experts, have difficulty in achieving agreement. Operating within an environment where opinions continually vary as to what is being looked for, and without explicit annotation process guidelines, items will be overlooked and missed. There clearly is a need to automate information provision to address this.

6.3.3 Additional Error Considerations

In addition to Friedman’s criteria for generalizing text analytics evaluations (Friedman 1998), previously outlined, we actively tested for and corrected the following issues that otherwise would have effected annotation results:

i) parser errors resulting in incorrect sentence splitting (e.g. across page, treating semi-colons as line breaks), could in some instances, have contributed to class type annotations being associated with an incorrect reporting line, with implications for Semantic Path instantiation

ii) spurious annotations resulting in annotation being incorrectly made with a knock-on effect for identification of class types and text segments extraction

iii) missing annotation resulting in lack of identification of class type instances and text segment extraction. The implication here is that the wrong information or information that should have been instantiated into a path, may not have been

iv) annotation errors associated with class type link creation during template insertion into the knowledge base. The implication being that the paths may not have been instantiated and consequently not made available to the evaluator.
6.4 Experimental Results and Discussion

6.4.1 Performance Metrics

Results presented below in Table 6-7 adhere to the performance metrics overview structure from Figure 6-2. The table comprises two results sets that report on evaluation of:

- a-priori preparatory (manual vs. system generated)
- a-priori usage (system generated vs. user usage of system) and
- a-posteriori (manual vs. user usage of system) activities.

The first result reflects standard evaluation and the second a re-calculation of the first, based on a re-introduction of false positive found in the first result set.

**Standard evaluation results:** The *a-priori preparatory* phase returns a relatively low precision value of 16.7%, but maximum recall value. The result, not unexpected, reflects that the tuple manual annotation set, is less than the coverage capability of the auto-generated system annotations. False negatives, the difference between annotations in the manual reference, not found in the system generated reference, of zero, reflects this. The number of false positives is high but again expected, highlighting the differences between the system generated annotations, not found in the (lesser) number of manual annotations, and is also expected. Recall at maximum value of 100% is therefore not surprising and reflects the systems linguistic capability. Precision from the *a-priori usage* phase increases from 16.7% to 86.9%, with a corresponding reduction in false positives (63 from 349) and increase in false negatives (0 to 194). False positives in this instance, represent selections made by the analyst using the system, that are missing from the system provided reference set. Investigation revealed that analysts extracted from the system provided annotations based on other sentences in text blocks as opposed to the central sentence, indicating system assistance for the analyst. The increase in false negatives is not surprising, as here again, the system generates a greater number of annotation permutations than what the analyst, using the system will deem relevant and select from. False negatives had the knock on effect of reducing recall to 68.4 %. Both a-priori preparatory and a-priori usage indicate that the system is behaving as expected, with some indication of usefulness. The more insightful measure is provided by *a-posteriori performance*, as it directly assesses the
competitive analyst information identification activity, performed manually vs. that performed with system provided support. Precision increased to 23%, reflecting an increase of 37% from the a-priori preparatory stage of 16.7%. Here false negatives, representing the difference in manual annotations not found in the system provided annotation set. Considering the results from the previous stages, the false negative value of 17, representing that 24.2% of manual annotations were not found, we expected to be less. The result does reflect that the analysts was successful in finding the majority of relevant text segments using the system, previously annotated manually, but with 24.2% not found, the system itself may be hindering a higher number being found. False positives are the difference between annotations found using the system that were not present, in the manual annotation set. They represent for the analyst, relevant but unknown, whether hidden or overlooked, information nuggets that the manual analysis did not identify, but which system usage did help identify. False positives demonstrate that the system does assist analysts in information identification.

**False positives re-introduced.** Post evaluation inspection of these false positives by domain expert evaluators based on 100% percentage agreement, highlighted that the majority of annotations (225 of 235), missing from the manual reference standard, were relevant and clearly negatively impacted precision. Measuring something that has been learned, but which was not previously known, is a difficulty also faced in ontology learning. An approach that caters for calculating the amount of learning, is the notion of posteriori precision that reassess precision, after evaluator consultation identifies false positives to re-introduce (Kavalec 2005; Schutz 2005). Re-introducing the a-posteriori 225 false positives and recompilation of metrics as outlined in Table 6-7, leads to a significant improvement in a-posteriori precision (82.3%).

For competitive analysis, recall is of greater importance for the a-priori activities and precision, for the a-posteriori. Comparison of a-posteriori evaluation is more informative of Semantic Path contribution. Precision is enhanced with system use relative to manual increasing from 16.7% to 23%. This 37% increase is also reflected in F-measure values biased towards precision ($\beta=0.5$) which increases from 23% to 27.5%, suggesting that the CAO through the AWB, did indeed assist the analyst identify and extract, a greater level of information relevant to the performance of a competitive analysis.
### Table 6-7 Performance IR Metrics across all annotators

<table>
<thead>
<tr>
<th>Activity</th>
<th># Tuples evaluated</th>
<th>#tuples = ALL [standard]</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>(\beta=0.5)</th>
<th>(\beta=1.0)</th>
<th>(\beta=1.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>t_p</td>
<td>f_p</td>
<td>f_n</td>
<td>(t_p)</td>
<td>0.167</td>
<td>1.0</td>
<td>0.207</td>
</tr>
<tr>
<td>a priori: preparatory</td>
<td>70</td>
<td>349</td>
<td>0</td>
<td></td>
<td>0.869</td>
<td>0.684</td>
<td>0.818</td>
<td>0.765</td>
</tr>
<tr>
<td>a priori: usage</td>
<td>419</td>
<td>63</td>
<td>194</td>
<td></td>
<td>0.23</td>
<td>0.805</td>
<td>0.275</td>
<td>0.358</td>
</tr>
<tr>
<td>a posteriori: performance</td>
<td>70</td>
<td>235</td>
<td>17</td>
<td></td>
<td>0.549</td>
<td>0.869</td>
<td>0.6</td>
<td>0.673</td>
</tr>
<tr>
<td></td>
<td></td>
<td>251</td>
<td>206</td>
<td>38</td>
<td>0.869</td>
<td>0.684</td>
<td>0.818</td>
<td>0.765</td>
</tr>
<tr>
<td></td>
<td></td>
<td>419</td>
<td>63</td>
<td>194</td>
<td>0.823</td>
<td>0.937</td>
<td>0.847</td>
<td>0.876</td>
</tr>
</tbody>
</table>

Comparison against similar EDGAR filings extractors (discussed in Section 2.2) cannot be easily applied. With reference to Table 6-8, structure data extractions from balance sheet, (Bovee 2005; Stampert 2008), along with compensation tables (Ding 2006), record percentage performance of 97%, 88.4% and 81% respectively, although for the latter two, the percentages include labels and named entities. Systems that target unstructured content report recall of 78.5% for extraction of pro-forma and fair value information (Grant 2006), 51% for biographies, 49% for annotations and 97.38% for entity resolution (Hernandez 2010). Precision across these systems is also higher at 64% (Grant 2006), 91% for biographies, 87% on annotations and perfect precision on entity resolution (Hernandez 2010). Pension term extraction reports in terms of an overall 95% success factor. Across these systems, Midas (Hernandez 2010), which targets elements of its extraction against unstructured biographies and EES (Grant 2006) targeting fair value fact extraction, provides a precision scale of 91% at the higher end, and 64% at the lower. EES’s precision value is however based on extracting a maximum of 40 relevant term, as opposed to our targeting of 149 class terms. If text segment identification is equatable to items identified and extracted in Table 6-8, then our system performs well for recall only, when confined to the
originally created manual reference standard, as opposed to when false positives were re-introduced.

Table 6-8 Performance Comparison against similar EDGAR extraction systems

<table>
<thead>
<tr>
<th>System</th>
<th>Approach</th>
<th>Identification &amp; Extraction</th>
<th>Performance Results Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>EASE (Stampert 2008)</td>
<td>Reg. Exp., VS, Search terms</td>
<td>Balance sheet</td>
<td>97% for 10-Q’s, 100% for 10-K’s</td>
</tr>
<tr>
<td>EDGAR-Analyser (Gerdes 2003)</td>
<td>Keyword, relevancy ratings</td>
<td>Y2K articles</td>
<td>42% less articles for analysis, 96% less processing</td>
</tr>
<tr>
<td>FRAANK (Bovee 2005)</td>
<td>ML, NLP</td>
<td>Balance sheet line items, labels</td>
<td>94.7% for training data, 88.4% for test data</td>
</tr>
<tr>
<td>EES (Grant 2006)</td>
<td>Wrapper Reg. Exp.</td>
<td>pro-forma, fair value, systemic risk</td>
<td>(P)(^66) 64.02%, (R) 78.58%, (F) 70.56%.</td>
</tr>
<tr>
<td>ECRS (Ding 2006)</td>
<td>HMM, Reg Exp</td>
<td>compensation tables, names</td>
<td>Retrieval rate 81%</td>
</tr>
<tr>
<td>Midas (Hernandez 2010)</td>
<td>Rule based, matrix, graph</td>
<td>Biographies, entities</td>
<td>Biographies (P) 91%, (R) 51%. Annotation of (P) 87%, (R) 49%; Entity resolution (R) 97.38%, (P) 100%.</td>
</tr>
<tr>
<td>(Chakraborty 2010)</td>
<td>Clustering algorithm</td>
<td>Pension terms</td>
<td>Overall success: 97% on training data, 95% on the test data.</td>
</tr>
<tr>
<td>AWB (O’Riain 2006)</td>
<td>OBIE, NLP</td>
<td>Competitive analysis relevant text blocks</td>
<td>(P) 23%, (R) 80.5%, (F) (\beta=0.5) 27.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(P) 82.3%, (R) 93.7%, (F) (\beta=0.5) 84.4%</td>
</tr>
</tbody>
</table>

66 (P) represents precision, (R) recall, and (F) F-measure, used here due to space restrictions

6.4.2 General Observations

Correlation of cross category contributions from manual annotations based results, are detailed in Table 6-10 - Table 6-12, with those obtained from system usage in terms of Semantic Paths, presented in the bar chart of Figure 6-3 below. The bottom portion of the columns represents overlap between manual and system context areas, and the top portion that from system usage only. The correlation highlights for manual analysis, that analyst concentrated on Sales, with additional reliance on Acquisition and to a lesser degree reliance on Profit, Market and Headcount. This is also reflected with system use resulting in greater instance numbers of text segments identification.
The greater level of annotations achieved with system usage, contrasted against that achieved during manual analysis, can in part be explained with analyst reluctance to annotate anything, other than what was considered minimally necessary. The increase is also attributable to the structured approach of information provision by the ontology, making the identification of in-context information easier to find and access. The use of Relationship context, previously missing from manual analysis, is further indication of ontology utility in information provision. Overall, the analyst increases annotation and relevant information selection as reflected by precision enhancement from Section 6.4.1.

Disposal Semantic Paths were modelled as extensively as other context areas (Table 6-9), but there was no evidence of their usage in either the manual or system approaches.

Investigation found that:

i) disposal specific lexical triggers such as \{Intangibles, technology, policy, fair value, intend to continue, subsidiary\}, were in fact missing from corpus filing texts

ii) some Disposal Semantic Paths also occurred within the Sales context, which the JAPE rule engine may have associated the fired rule with (cf. Section 5.2.2)

iii) analysts used the Acquisition context Semantic Paths, and as Acquisition represented the positive information search version of Disposal, annotations were associated with Acquisition, instead of Disposal.
The lack of Disposal annotations raised the obvious question as to whether unused context areas would be candidates for ontology pruning. Subsumption would require wider corpus inclusion and evaluator investigation.

**Table 6-9 Semantic Paths Summary per context area**

<table>
<thead>
<tr>
<th>Context Area</th>
<th>Level 0</th>
<th>Level 1-Level2</th>
<th>Level 2-Level3</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>12</td>
<td>84</td>
<td></td>
<td>96</td>
</tr>
<tr>
<td>Profits</td>
<td>23</td>
<td>-</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Acquisition</td>
<td>28</td>
<td>-</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Relationships</td>
<td>31</td>
<td>-</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Market</td>
<td>28</td>
<td>-</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Disposal</td>
<td>33</td>
<td>-</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Headcount</td>
<td>7</td>
<td>63</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>R &amp; D</td>
<td>8</td>
<td>29</td>
<td>37</td>
<td></td>
</tr>
</tbody>
</table>

Decomposition of the Semantic Path summaries from Figure 6-3, into their usage across manual vs. system competitive analysis approaches, are presented in Table 6-10, Section 6.4.3. The figures reflect consistent, unbiased and greater agreement for the majority of Semantic Paths used to identify relevant information, as the analyst moves between competitive analysis session types. Accumulated rater agreement increases, from 12%, with the manual approach, to 59% with system usage, outlining system utility in information foraging across analysts, despite the earlier observation that analysts will instinctively only invest the minimum effort necessary for manual filings analysis.

**6.4.3 Correlated Results Breakdown**

Table 6-10 breaks down per approach, the Semantic Path used, and summaries for the number of instances and rater agreement found. The table allows discrimination between the most frequently used paths and overlap between approaches and usage consistency. The table should be interpreted as follows:

- **path**, identifies the particular Semantic Path
- **man Inst #.**, details the number of particular Semantic Path instances identified across all raters when performing manual annotation on filings
• **usage Inst #.** details the number of particular Semantic Path instances identified across all raters through system usage

• **man Rater #:** The number of raters that identified the Semantic Path type instances when performing manual annotations on filings. The maximum number is five

• **usage Rater #:** The number of raters that identified the Semantic Path type instances through system usage. The maximum number is five.

  A dash ‘-‘ in the column indicates lack of annotator identification and consequently accompanying instance number counts. Further correlation of the Semantic Paths rater agreement from Table 6-10 is graphed in Figure 6-4. The left most columns give the Semantic Path, and the right hand column the percentage rater agreement, for manual (bottom line) and system usage (top line).
### Table 6-10 Manual vs. system usage Semantic Path instantiations summaries (part 1)

<table>
<thead>
<tr>
<th>Path</th>
<th>Man. Inst #</th>
<th>Usage Inst #</th>
<th>Rater #</th>
<th>Path</th>
<th>Man. Inst #</th>
<th>Usage Inst #</th>
<th>Rater #</th>
<th>Path</th>
<th>Man. Inst #</th>
<th>Usage Inst #</th>
<th>Rater #</th>
<th>Path</th>
<th>Man. Inst #</th>
<th>Usage Inst #</th>
<th>Rater #</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC1002-C1005</td>
<td>5</td>
<td>21</td>
<td>4</td>
<td>RC1005-C1040</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>RC1008-C1032</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>RC1019-C4009</td>
<td>-</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>RC1002-C1006</td>
<td>22</td>
<td>20</td>
<td>4</td>
<td>RC1005-C4019</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>RC1009-C1016</td>
<td>2</td>
<td>77</td>
<td>2</td>
<td>RC1019-C4024</td>
<td>-</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>RC1002-C1007</td>
<td>14</td>
<td>61</td>
<td>5</td>
<td>RC1006-C1011</td>
<td>6</td>
<td>14</td>
<td>4</td>
<td>RC1009-C1034</td>
<td>-</td>
<td>25</td>
<td>-</td>
<td>RC1020-C1024</td>
<td>1</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>RC1002-C1014</td>
<td>28</td>
<td>24</td>
<td>5</td>
<td>RC1006-C1018</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>RC1009-C1035</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>RC1020-C3006</td>
<td>5</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>RC1002-C1019</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>RC1006-C1019</td>
<td>3</td>
<td>13</td>
<td>3</td>
<td>RC1010-C1013</td>
<td>3</td>
<td>19</td>
<td>3</td>
<td>RC1020-C3008</td>
<td>-</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>RC1003-C1005</td>
<td>-</td>
<td>21</td>
<td>-</td>
<td>RC1006-C1020</td>
<td>-</td>
<td>60</td>
<td>-</td>
<td>RC1010-C1019</td>
<td>-</td>
<td>13</td>
<td>-</td>
<td>RC1022-C1007</td>
<td>-</td>
<td>61</td>
<td>-</td>
</tr>
<tr>
<td>RC1003-C1006</td>
<td>-</td>
<td>20</td>
<td>-</td>
<td>RC1006-C1021</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>RC1010-C1036</td>
<td>2</td>
<td>87</td>
<td>2</td>
<td>RC1022-C1042</td>
<td>-</td>
<td>58</td>
<td>-</td>
</tr>
<tr>
<td>RC1003-C1007</td>
<td>-</td>
<td>61</td>
<td>-</td>
<td>RC1006-C1041</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>RC1010-C1037</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>RC1022-C1045</td>
<td>-</td>
<td>30</td>
<td>-</td>
</tr>
<tr>
<td>RC1004-C1007</td>
<td>14</td>
<td>61</td>
<td>5</td>
<td>RC1006-C4019</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>RC1010-C1038</td>
<td>5</td>
<td>53</td>
<td>3</td>
<td>RC1028-C1045</td>
<td>13</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>RC1004-C1009</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>RC1007-C1020</td>
<td>10</td>
<td>60</td>
<td>4</td>
<td>RC1010-C1044</td>
<td>-</td>
<td>24</td>
<td>-</td>
<td>RC1028-C3027</td>
<td>-</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>RC1004-C1010</td>
<td>10</td>
<td>64</td>
<td>5</td>
<td>RC1007-C1023</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>RC1019-C1004</td>
<td>5</td>
<td>66</td>
<td>4</td>
<td>RC1028-C7003</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>RC1005-C1011</td>
<td>2</td>
<td>14</td>
<td>2</td>
<td>RC1007-C1024</td>
<td>8</td>
<td>17</td>
<td>5</td>
<td>RC1019-C1010</td>
<td>-</td>
<td>64</td>
<td>-</td>
<td>RC1028-C7007</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>RC1005-C1012</td>
<td>2</td>
<td>23</td>
<td>2</td>
<td>RC1007-C1025</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>RC1019-C1036</td>
<td>1</td>
<td>87</td>
<td>1</td>
<td>RC1028-C7004</td>
<td>5</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>RC1005-C1013</td>
<td>-</td>
<td>19</td>
<td>-</td>
<td>RC1007-C1042</td>
<td>2</td>
<td>58</td>
<td>2</td>
<td>RC1019-C1037</td>
<td>-</td>
<td>16</td>
<td>-</td>
<td>RC1028-C7006</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>RC1005-C1014</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>RC1008-C1028</td>
<td>-</td>
<td>29</td>
<td>-</td>
<td>RC1019-C3016</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>RC1030-C1045</td>
<td>-</td>
<td>30</td>
<td>-</td>
</tr>
<tr>
<td>RC1005-C1015</td>
<td>-</td>
<td>7</td>
<td>-</td>
<td>RC1008-C1030</td>
<td>-</td>
<td>34</td>
<td>-</td>
<td>RC1019-C4007</td>
<td>3</td>
<td>27</td>
<td>3</td>
<td>RC1030-C3027</td>
<td>-</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>RC1005-C1016</td>
<td>1</td>
<td>77</td>
<td>1</td>
<td>RC1008-C1031</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>RC1019-C4008</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>RC1030-C7003</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>-----------------</td>
<td>-------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>-----------------</td>
<td>-------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>RC2006-C5001</td>
<td>3</td>
<td>35</td>
<td>2</td>
<td>4</td>
<td>RC2021-C7004</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>4</td>
<td>RC3003-C4015</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>RC2006-C8007</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>1</td>
<td>RC2021-C7006</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>3</td>
<td>RC3003-C4016</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>RC2016-C1028</td>
<td>2</td>
<td>29</td>
<td>2</td>
<td>3</td>
<td>RC2021-C7007</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>RC3003-C8003</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>RC2016-C1030</td>
<td>5</td>
<td>34</td>
<td>5</td>
<td>5</td>
<td>RC3001-C1024</td>
<td>-</td>
<td>17</td>
<td>-</td>
<td>4</td>
<td>RC3004-c2006</td>
<td>-</td>
<td>57</td>
<td>-</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>RC2016-C7002</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>3</td>
<td>RC3001-C3006</td>
<td>-</td>
<td>41</td>
<td>-</td>
<td>5</td>
<td>RC3004-C3017</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>RC2018-C2009</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>2</td>
<td>RC3001-C3008</td>
<td>-</td>
<td>16</td>
<td>-</td>
<td>2</td>
<td>RC3004-C3021</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>RC2018-C2020</td>
<td>-</td>
<td>9</td>
<td>-</td>
<td>2</td>
<td>RC3002-C3006</td>
<td>-</td>
<td>41</td>
<td>-</td>
<td>5</td>
<td>RC3005-C2016</td>
<td>-</td>
<td>24</td>
<td>-</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>RC2018-C3006</td>
<td>-</td>
<td>41</td>
<td>-</td>
<td>5</td>
<td>RC3002-C3008</td>
<td>-</td>
<td>16</td>
<td>-</td>
<td>2</td>
<td>RC3005-C3015</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>RC2019-C1045</td>
<td>-</td>
<td>30</td>
<td>-</td>
<td>4</td>
<td>RC3002-C3010</td>
<td>-</td>
<td>16</td>
<td>-</td>
<td>3</td>
<td>RC3005-C3024</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>RC2019-C3027</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>2</td>
<td>RC3002-C3011</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>2</td>
<td>RC3028-C1004</td>
<td>-</td>
<td>66</td>
<td>-</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>RC2019-C7003</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>RC3003-C1004</td>
<td>2</td>
<td>66</td>
<td>2</td>
<td>4</td>
<td>RC3028-C1019</td>
<td>-</td>
<td>13</td>
<td>-</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>RC2019-C7004</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>4</td>
<td>RC3003-C1016</td>
<td>5</td>
<td>77</td>
<td>3</td>
<td>5</td>
<td>RC3028-C1045</td>
<td>-</td>
<td>30</td>
<td>-</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>RC2019-C7006</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>3</td>
<td>RC3003-C1018</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>RC3028-C2016</td>
<td>-</td>
<td>24</td>
<td>-</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>RC2019-C7007</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>RC3003-C1044</td>
<td>9</td>
<td>24</td>
<td>3</td>
<td>4</td>
<td>RC3028-C3006</td>
<td>-</td>
<td>41</td>
<td>-</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>RC2021-C1045</td>
<td>-</td>
<td>30</td>
<td>-</td>
<td>4</td>
<td>RC3003-C3014</td>
<td>7</td>
<td>35</td>
<td>3</td>
<td>5</td>
<td>RC3028-C3010</td>
<td>-</td>
<td>16</td>
<td>-</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>RC2021-C3027</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>2</td>
<td>RC3003-C3015</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>RC3028-C3027</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>RC2021-C7003</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>RC3003-C3016</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>3</td>
<td>RC3028-C3029</td>
<td>-</td>
<td>18</td>
<td>-</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table 6-11 Manual vs. system usage Semantic Path instantiations summaries (part 2)
Table 6-12 Manual vs. system usage Semantic Path instantiations summaries (part 3)

<table>
<thead>
<tr>
<th>Path</th>
<th>Man. Inst #</th>
<th>Usage</th>
<th>Man. Rater #</th>
<th>Usage</th>
<th>Rater #</th>
<th>Usage</th>
<th>Rater #</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC5021-C1028</td>
<td>29</td>
<td>-</td>
<td>3</td>
<td>RC7004-C2006</td>
<td>-</td>
<td>57</td>
<td>-</td>
</tr>
<tr>
<td>RC5021-C1030</td>
<td>34</td>
<td>-</td>
<td>5</td>
<td>RC7004-C3003</td>
<td>-</td>
<td>36</td>
<td>-</td>
</tr>
<tr>
<td>RC5021-C1044</td>
<td>24</td>
<td>-</td>
<td>4</td>
<td>RC7004-C8002</td>
<td>-</td>
<td>21</td>
<td>-</td>
</tr>
<tr>
<td>RC5021-C7002</td>
<td>10</td>
<td>-</td>
<td>3</td>
<td>RC7010-C1045</td>
<td>-</td>
<td>30</td>
<td>-</td>
</tr>
<tr>
<td>RC5023-C1028</td>
<td>29</td>
<td>-</td>
<td>3</td>
<td>RC7010-C3027</td>
<td>-</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>RC5023-C1044</td>
<td>24</td>
<td>-</td>
<td>4</td>
<td>RC7100-C7002</td>
<td>-</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>RC5023-C7002</td>
<td>10</td>
<td>-</td>
<td>3</td>
<td>RC7100-C7003</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>RC7002-C1045</td>
<td>2</td>
<td>30</td>
<td>1</td>
<td>4</td>
<td>RC7100-C7004</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>RC7002-C3027</td>
<td>6</td>
<td>-</td>
<td>2</td>
<td>RC7100-C7006</td>
<td>-</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>RC7002-C7003</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>RC7100-C7007</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>RC7002-C7004</td>
<td>10</td>
<td>-</td>
<td>4</td>
<td>RC8002-C2016</td>
<td>-</td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td>RC7002-C7006</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>RC8002-C3006</td>
<td>1</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>RC7002-C7007</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>RC8002-C3030</td>
<td>-</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>RC7004-C1002</td>
<td>120</td>
<td>-</td>
<td>5</td>
<td>RC8002-C7007</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>RC7004-C1025</td>
<td>8</td>
<td>-</td>
<td>3</td>
<td>RC8002-C8003</td>
<td>-</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 6-4 Manual vs. system usage Semantic Paths rater agreement
6.4.4 Comparison with Previous Analysis Findings

As an adjunct to the main evaluation we were afforded access to the detailed results from a previously conducted competitive analysis, against the same corpus used by HP’s BPO team, for business opportunity discussions with a large multinational customer. We were particularly interested in establishing how intelligence gathering and analysis outcomes were used further along the decision process. The previous competitive analysis detailed financial metrics, technology, existing relationships, and strategic challenges. The analysis relied upon four text segments extracted from filings to support discussion. Investigation determined that three were extracted from forms not included in the evaluation corpus, and the fourth, was condensed from two text segments found on adjacent lines. 40% (2/5) of the evaluators annotated these segments when performing both manual and system based competitive analysis. A closer look at the evaluators revealed that one was involved in ontology development while the other was not. The comparison although limited, does serve to illustrate the difficult and complex nature of competitive analysis, where even when the same analysts are used to perform an identical analysis, on the same company, with the same information sources, agreement can and does vary, as does the ability to overlook previously identified relevant information.

Such an operation environment would benefit from the automated provision of the information requirement in a structured manner that would provide a repeatable and structured approach to information identification. Performance results from Section 6.4.1 indicate that this is both possible and practical to achieve, but as Section 7.4.2 will outline, the user interface is a critical part of making any resultant system effective.

6.5 Summary

Saracevic’s evaluation criteria (Saracevic 1995) for system centric information retrieval, representing a more thorough methodological approach, than the traditional basic measurement of retrieval effectiveness, as dictated by Cranfield (Manning 2009), was used as the basis for experimental approach. Experiment fundamentals established case study scope and its influence on the domain of interest, training data availability, domain
modelling and the complexity of experimental setup for competitive analysis task evaluation. Experimental criteria examined the notion of relevance, its meaning within the domain of competitive analysis and the importance of using domain experts, to establish and determine, what is relevant. The methodology used to evaluate precision introduced the notion of a-priori and a-posteriori evaluation stages. The stages were used to accommodate situation, where evaluators select from the system generated items, and provide a greater level of insight on system functionality and consequently Semantic Path utility.

Experimental setup reported an annotator percentage of agreement across five domain experts of 51.1%. Experimental error and bias introduction were minimised using Friedman’s criteria for natural language processing (Friedman 1995; Friedman 1998). Actual experimental results, comparing the competitive analyst information identification activity, performed using the system to previous manual efforts, reported an a-posteriori precision increase of 37%, from 16.7% to 23%, indicating that the Semantic Paths did contribute towards greater information identification. Analysis of false positives based on 100% percentage agreement from domain experts, highlighted, that the majority of annotations missing from the manual reference standard were in fact relevant. Re-introduction of these a-posteriori false positives, with metrics recalculation, produces a significant improvement of a-posteriori precision from 54.9% to 82.3%. Other findings from analyst use of the system found that a greater number of Semantic Paths were exploited to identify relevant information, suggesting sufficient ontology coverage of the competitive analysis domain. Additionally, greater numbers of relevant information segments (expressed as Semantic Path instances members) were extracted, indicating ontology utility in supporting the analyst move from a scenario of minimalist manual analysis annotations, to structured information provision, that supported a greater level of relevant item identification and annotation. Analyst’s rater agreement in the majority of cases for identified relevant information increased from the manually achieved agreement of 12% to 59%, again indicating ontology utility. When discussing a utility view of relevance Saracevic commented that “it is fine for IR systems to provide relevant information but the true role is to provide information that has utility – information that helps to directly resolve given problems or directly bears on given actions ...” (Saracevic 1975).
The Semantic Paths through the CAO does support structured information provision and the ontology as an artefact, contributes to utility, through competitive analysis task structuring. Both together assist competitive analysis task performance, providing an automated and uniform approach to tackling the information requirement of a complex area. Although the comparison with previous analysis of the same corpus is not statistically significant, is none the less insightful, demonstrating the difficulties faced - where only 40% of the original BPO team managed to identify their own previously used relevant text.

The CAO through its instantiation in the AWB does represent an information extraction performance trade-off between complexity and specificity (Cunningham 2005b). The greater the complexity of data targeted for extraction, the more specific the domain of discourse becomes. Considering the 149 class types specified, their association and modelling as Semantic Paths, to support automated task structuring, the precision increase achieved from 16.7% to 23%, does represent for the analyst community, a substantial advancement on their manual position.
Chapter 7

Analyst Work Bench Use Case Demonstrator

This section presents the Analyst Work Bench (AWB) artefact, that utilises linguistic analysis to instantiates the CAO and allow investigation of the original research questions, as to whether:

- information threads (Semantic Paths) are identifiable within the disclosure sections of EDGAR Form 10-Q filings?
- formal modelling as the CAO could be used to represent the information requirement necessary for competitive analysis?
- CAO use can help satisfy the analyst’s information need, by guiding extraction and organising the provision of relevant information for competitive analysis.

The section is structured as follows: the business use case environment that spurred the requirements is first given (Section 7.1). The AWB artefact is then described in detail using an analyst usage scenario (Section 7.2). The walk-through highlights the information requirement and task-orientated nature of the actual analysis. The usage scenario provides the context used to perform the competitive analysis experiment evaluation (Section 6.4). Component architecture (Section 7.3) introduces artefact components functionality; design of supporting knowledge base; and information flow to support component implementation. Finally usability evaluation (Section 7.4), establishes AWB utility, based on its usage and performance in the competitive analysis experiment from Section 6.3 and 6.4.

As mentioned previously (Section 1.4) the AWB artefact was delivered and installed with the Business Process Outsourcing team of Hewlett-Packard’s European Software Centre. The AWB was also the subject of a IP licensing agreement with HP for semantic products to support competitive analysts.
7.1 Business Use Case

Hewlett-Packard’s European Software Centre is a provider of outsourced services to Independent Software Vendors (ISV’s), within the areas of software kit manufacturing, distribution and supply chain management. Due to changing business practices, the Business Process Outsourcing (BPO) business team was sought to identify which ISV’s had not to-date outsourced and which may provide new business opportunity areas. The BPO team’s goal was to build the case for enhancing current service offerings, to different areas of the ISV’s product development cycle, where HP has considerable in-house competencies and proven delivery record. The primary means of identifying ISV’s was to gain insight into company activity through the use of competitive analysis. Of interest were ISV’s:

- whose cost base was growing at a rate greater than revenue
- where revenue (total amount) per employee was falling
- that were experiencing operational difficulties in their supply chain
- had not invested in R&D at the industry (or competitor) average
- engaged in major re-structuring
- had made recent major acquisitions.

One of the main drivers of information for this analysis process were the consolidated financial information and company officer narrative sections of the US SEC Form 10-Q. Analyses of the financial figures are readily undertaken using financial macros and used to present the company’s financial picture. Qualitative analyses of CEO statements are used to determine what is really behind the company’s financial statements. Statement analysis serves to reaffirm metrics comparison and develop insight into company problematic areas. The statements and their condensed version are further relied upon when presenting to an ISV’s management team, used to establish clarity of message and make the task of selling the proposed solutions on offer easier.

The Form 10-Q typically averaging between 50-100 report pages, represents the most accurate and trust worthy source of this type of information (Section 4.1 discusses in detail the impediments and challenges involved in using the forms narrative sections to
gather this information from). The BPO team began with an ISV evaluation intake of seventy candidate companies. Analysis of multiple filings per ISV took a total of nine months to complete resulting in the identification of five companies that received further in-depth assessment for business process outsourcing potential. From that two were approached.

Any means of automating the manual intelligence gathering activity, by providing their information requirement, in a structured fashion, which was closely aligned with the competitive analysis task itself, was of interest to the BPO analysts. Automating the structured information provision offered the possibility for enhanced search and identification of those relevant information segments that otherwise might be missed or overlooked.

7.2 Analyst Usage Scenario

The main in-text visualisation area of the AWB comprising a navigator, report viewer and Semantic Path viewing areas is presented in Figure 7-1.

Figure 7-1 AWB, Main In-Text Visualization Area
A tree view hierarchical ontological **Navigator**, allows report traversal within a business context (here ‘Sales’), through selection of terms\textsuperscript{67}, and their roles to other terms. For information provision, the ontology represents the information space, and the term-role-term, the ontology constituents traversal options (or Semantic Path).

Separated out in Figure 7-2, traversal using the navigator, provides indicators for term instantiation (such as *Product*) as to:

i) which have multiple instances in the filing, given by ‘Inst: 72’, indicating seventy two term instances found.

ii) what Semantic Paths relationships are instantiated. Here *Product* has the three instantiated relationships *Announcements*, *Introduction* and *Revenue*, as indicated by ‘Links 3’. *Announcements* in turn have four instantiated relationships, *(Planned, Competitive, Delay and Release)*.

The functionality allows the analyst to quickly assess what information type has been identified, within the filing and whether pursuing that particular information thread is worthwhile. Used in this manner the analyst can selectively navigate through instantiated Semantic Paths structures and personalise analysis.

The individual term selected for viewing in the navigator (‘Planned’), has all of its term instance synonyms highlighted in the semantically enhanced version of the filing, displayed within the Report Viewer in Figure 7-3.

\textsuperscript{67} DOGMA terminology used to refer to class type and role to relationship.
The analyst can quickly iterate through term instance and their business discourse variants. Illustrated here for term *Planned*, are its domain specific synonyms of *Intention* and *Program*, not available from a lexical resource such as WordNet. When an analyst identifies and selects an instance of interest (highlighted light background), the particular instance and any additional members of the Semantic Path being traversed via the navigator, are presented in a flattened hierarchy in the **Semantic Path Viewer**, in Figure 7-4. The Semantic Path is populated based on the original context area (‘Sales’), the Semantic Path being traversed as part of the ontology structure and the specific class instance selected. This instantiated path is then presented as an aggregated view allowing analyst assessment of that particular Semantic Path for propositional context.

With each information path representing a small part of the overall information available within the filing, the analyst incrementally builds a picture of in-context business
informational segments, removed from the overhead, of having to consider original information context and association. Text segments identified as irrelevant can be excluded from further Semantic Path instantiation and instance identification by checking the exclude box. Those identified as relevant can be selectively marked for extraction using the extract box.

A summary of all the text segments previously selected for extraction can be viewed or exported using the **Summary Report Generation** functionality (Figure 7-5). Information provenance is retained through association of the Semantic Path that was being traversed, with the term instance extracted, allowing the analyst retain the original extraction context. The summary report is sharable with other stakeholders and usable for further decision support.

*Figure 7-5 AWB, Summary Extraction Report of Relevant Text Segment*

7.3 Component Architecture

The AWB was designed with the main component (Figure 7-6) catering for the main functionality areas of:

- **Analyst Work Bench**, providing the interactive interface and report generation
- **EDGAR filings processing**, responsible for semantic information processing
- **semantic information consolidation**, integrating the extracted text segments
- **database server**, supporting ontology implementation and knowledge repository.
An overview of each is provided beginning with back end processing components and working forward to interactive front end components.

**Semantic Information Processing** is fundamentally responsible for the processing of SEC Form 10-Q’s, but will also process any text based filings (e.g. Form 10-K, 20-F). The pattern mining sub-component is responsible for linguistic processing using the ANNIE component of GATE, based on handcrafted extraction rules from the CAO schema (Section 5.2.2). Pattern mining invokes the *semantic annotation* sub-component using it to enable the:

i) *EDGAR filing annotation* sub-component to semantically annotate the filing, allowing later exploitation by the UI for navigation purposes. Figure 7-7, reproduced from Section 5.2.1, provides a sample output from the semantic annotation component. The bottom portion of the figure contains the marked up filing source, viewable from a web browser.
The exclude property is used to indicate whether the term should be excluded from selection. The background property is used to turn on/off the particular term instances, corresponding with user selection of a particular term in the report viewer. The properties contribute toward managing the structured aspect of the analysis activity. Possible future work could look to the use of RDFa\textsuperscript{68} (Resource Description Framework – in – attributes), to semantically mark-up passages and words within the text.

\textit{ii) template creation} sub-component, semantically annotates extracted text segments, populate the extraction template and passes the template to the knowledge repository for insertion, (template examples previously presented in Section 5.2.2).

\textit{Semantic Information Consolidation} maps template contents into the knowledge base ontology as candidates for Semantic Path population. Once a filing has been processed, the ontology is instantiated with relevant text segments, based upon semantic proximity and Semantic Paths membership (detailed in section 5.2.2).

\textit{Interactive Interface} provides the user interface functionality of the AWB. The CAO navigation sub-component provides Semantic Path traversal capability through ontology expansion and navigation. The navigator UI, presented to the user, is dynamically generated from the knowledge repository and loaded into a DOM (cf. Figure 7-8). Walking the DOM, represents Semantic Path traversal, allowing selection of particular paths and their

\textsuperscript{68}http://www.w3.org/TR/xhtml-rdfa-primer/
members. The sub-component is also responsible for display of the annotated filing in the report viewer and generating the consolidated view of instantiated Semantic Paths.

The text segment extraction sub-component allows specification of which text segment to extract, and the report generator sub-component, the summary report of all text segments previously annotated for extraction.

Figure 7-8 AWB XML generated navigator

Knowledge Repository. The SQL server sub-component implements the ontology as a normalised data model with a core set of tables (shaded) as presented in Figure 7-9:

- **table_ont_imp** defines ontology hierarchy and membership
- **table_ont_inst** ontology instances
- **table_info_item** the text segments associated with each instance and
- **table_mapping** the valid instantiated Semantic Paths.

The remaining tables fulfil a support role, for example:

- **table_grounding**, allows lookup of abstract class types defined in meta-lexon format (examples in Section 5.1)
- **table_profile, table_schema, table_source** track analyst profile, the ontology schema used, and source filing processed.
Stored procedures were used for CRUD (create, read, update and delete) related activities of extraction, template insertion, Semantic Paths instantiation, UI generation and summary report creation.

The information flow between components in terms of interaction with the database schema is given in Figure 7-10. Once the filing has been linguistically processed, the flow accommodates two main paths responsible for:

i) the filings semantic annotation used by the AWB report Viewer to assist information space traversal.

ii) generating the text segment extraction template used to populate term instances (_ont_inst) and generate valid Semantic Path instances (_mapping).

Instantiated Semantic Paths are re-used to generate the Navigator UI, outline populated paths, and to what degree, in addition to the display of text segments associated with the populated paths, in the AWB Semantic Path viewer.
Figure 7-10 AWB Information Flow
7.4 Usability Evaluation

As part of the design-science related evaluation, we sought to determine the utility (Hevner 2004; Osterle 2011) of the AWB artefact, through assessment using heuristic instruments. Section 7.4.1 discusses the identification of relevant dimensions of the DeLone and McLean Model (D&L) Information Systems Success (DeLone 2003) and their contribution to instrument definition and questionnaire construction. Results from the questionnaire introduced to the analyst focus group, post completion of the competitive analysis experiment (defined in Section 6.2), are presented in Section 7.4.2.

7.4.1 Measuring Instrument

The D&L model illustrated in Figure 7-11, is essentially a classification scheme of IS success measures comprising six temporal and causally interdependent dimensions. The generic nature of the model makes its framework suitable as a standard evaluation platform in multiple contexts, across different domains such as knowledge management systems, accounting, eGovernment and eCommerce (Seddon 1996; Wu 2006). Our interest is in the use of its model dimension, where applicable (cf. Appendix V for a complete listing), for the usability evaluation of the Analyst Work Bench artefact. Each dimension was assessed for applicability and either omitted, augmented or replaced, with domain specific adaptations, to define specific instruments for evaluation, with an end goal of contributing to questionnaire construction for that quality area. The resultant questionnaire may be viewed in Appendix VI.

*Figure 7-11 D&M IS Success Model (DeLone. 2003)*
Information Quality targets the semantic level with instruments of completeness, ease of understanding, personalisation, relevance and security. With security and personalisation out of scope, we concentrated on understanding and relevance in terms of information provision and contribution from its presentation mechanism. The KMS distinction between information and knowledge, as dependent upon content quality and context and linkage quality (Wu 2006), was used to develop twelve questions, also termed instruments or measures (cf. Appendix VI). The Information Quality dimension was designed with the two sub-dimensions of content quality, targeting information requirement and relevance, and context and linkage, targeting the approach used for information search and presentation.

System Quality targets higher level artefact operational characteristics. It instruments are concerned with correct functionality such as ease of use, response time, flexibility and stability. Nielsen relates the ease of user interaction to performing tasks that the system was designed for, by following specific criteria (Nielsen 1989). Ease of use was therefore also included in the eight measuring instruments developed.

Service Quality instruments of assurance, empathy and responsiveness were originally intended to measure the effectiveness of areas such as end user computing and role of service providers in eCommerce enterprises. As our evaluation was unconcerned with any aspects of artefact service or maintenance, we did not include this dimension.

KMS system use, characteristically involves the more passive activity of knowledge acquisition and active activity, of knowledge sharing and dissemination. The activities support the actual intent of a KMS to assist in problem solving, decision making or its facilitation (Wu 2006). Wu found that these two dimensions could be collapsed into one, allowing them to introduce their own, to better reflect decision making, knowledge communication and relevance. Instruments concerned with personalised knowledge sharing were not considered, as the artefact only offered functionality, that shared information extracted in the form of summary reports. Instruments that looked to consider task and decision support were included, resulting in the development of three instruments targeting the overall extent of system usage.

User satisfaction as defined by D & L, considers fiscal attributes, site visits and user surveys as indicators. As our artefact was not released to a production environment, we first considered Doll and Torkzadehs user satisfaction measure, that looked at content accuracy
(covered by information quality), format, timeliness and ease of use (Dolla 1998). Satisfaction can also be attributed to system quality (or knowledge) quality (Dolla 1998; Wixom 2001), also covered under information quality. Concentrating on ease of use, we included the Wu and Wang approach of targeting general negative or positive artefact sentiment, to guide user satisfaction instruments development (Wu 2006).

Net Benefits remains an area where the MIS community has yet to achieve consensus on objective measurement (Wu 2006). Looking at individual, organisational, social and economic consequences of usage, it is typically measured by user perception and perceived system benefits. As a belief in performance or productivity it also captures effectiveness and has become an important component of IS success (Wixom 2001). Perceived benefit is a belief measure that the user or parent organisation will have for system usage. Example are resource savings and productivity increases, which for the artefact, can be somewhat measured either directly as time saving, or support actions on task activities. The greater the users perception of the importance of a task, the more useful the system will be perceived to be (Seddon 1996). Similar to Wu and Wang, we target perceived benefit, through the extent of use and introduced five instruments, which subscribe to system quality in terms of core function, namely assisting the analysis task.

Collectively taking the instruments developed for each quality area, we prepared a questionnaire which was presented to each evaluators, post completion of the competitive analysis activity using the Analyst Work Bench system. The questionnaire constructs were measured using a seven point Likert\(^{69}\) scale ranged from strongly agree to strongly disagree with an accompanying scale range of [1..7]. Items were sequenced to reduce scoring bias among the evaluators. The full questionnaire may be found in Appendix VI.

\(^{69}\) A psychometric scale commonly used in questionnaires, and is the most widely used scale in survey research. Developed by Renis Likert (1932), “A Technique for the Measurement of Attitudes”, *Archives of Psychology* 140: pp. 1–55
### Questionnaire Results and Discussion

Table 7.1 AWB success determination questionnaire

<table>
<thead>
<tr>
<th>System Quality: Assessment of operational characteristics</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Moderately agree</th>
<th>Undecided</th>
<th>Moderately disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Respondent No.</th>
<th>Respondent total</th>
<th>Rating average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is the AWB easy to use</td>
<td>0 3 2 0 0 0 0 0</td>
<td>12 5 2.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Is it user friendly</td>
<td>0 0 3 2 0 0 0 0</td>
<td>17 5 3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I find it easy to get AWB to do what I want</td>
<td>0 0 3 2 0 0 0 0</td>
<td>17 5 3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Is it easy for me to become skillful using the AWB</td>
<td>0 4 0 1 0 0 0 0</td>
<td>12 5 2.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. I believe the AWB is cumbersome to use</td>
<td>0 0 0 1 2 2 0 0</td>
<td>26 5 5.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Using AWB requires a lot of mental effort</td>
<td>0 1 0 3 1 0 0 0</td>
<td>19 5 3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Using the AWB required a learning overhead</td>
<td>0 0 3 1 0 1 0 0</td>
<td>19 5 3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Using AWB can be confusing</td>
<td>0 0 1 4 0 0 0 0</td>
<td>19 5 3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Information Quality: Quality of the AWB output</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Moderately agree</th>
<th>Undecided</th>
<th>Moderately disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Respondent No.</th>
<th>Respondent total</th>
<th>Rating average</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. The output is presented in a useful manner</td>
<td>0 0 5 0 0 0 0 0</td>
<td>15 5 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. The content representation provided by the AWB is logical</td>
<td>0 2 2 0 1 0 0 0</td>
<td>15 5 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. The information provided is relevant and helpful for the task</td>
<td>0 5 0 0 0 0 0 0</td>
<td>10 5 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. The information content meets your information requirements</td>
<td>0 0 4 1 0 0 0 0</td>
<td>16 5 3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. The information provided is meaningful</td>
<td>0 4 1 0 0 0 0 0</td>
<td>11 5 2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. AWB makes it easy for me to extract information of use</td>
<td>1 1 3 0 0 0 0 0</td>
<td>12 5 2.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context &amp; Linkage</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Moderately agree</th>
<th>Undecided</th>
<th>Moderately disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Respondent No.</th>
<th>Respondent total</th>
<th>Rating average</th>
</tr>
</thead>
<tbody>
<tr>
<td>15. The information navigation process is useful for information identification</td>
<td>0 0 4 0 0 1 0 0</td>
<td>21 6 3.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. The information navigation process is logical and fit</td>
<td>0 0 1 2 2 0 0 0</td>
<td>21 5 4.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. Information is presented in a way that provides a useful aggregated view</td>
<td>0 0 4 1 0 0 0 0</td>
<td>16 5 3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. The representation mechanism is successful in structuring the analysis task</td>
<td>0 0 4 1 0 0 0 0</td>
<td>16 5 3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. Does the information navigation process assist the cognitive workload in task performance</td>
<td>0 2 2 1 0 0 0 0</td>
<td>14 5 2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20. Provides information in context in a manner that is understandable, accessible and applicable to the task</td>
<td>0 0 4 1 0 0 0 0</td>
<td>16 5 3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User Satisfaction: Extent of positive or negative feelings towards the tool</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Moderately agree</th>
<th>Undecided</th>
<th>Moderately disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Respondent No.</th>
<th>Respondent total</th>
<th>Rating average</th>
</tr>
</thead>
<tbody>
<tr>
<td>21. My information needs are met</td>
<td>0 0 4 1 0 0 0 0</td>
<td>16 5 3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22. I am satisfied with AWB efficiency</td>
<td>0 0 3 2 0 0 0 0</td>
<td>17 5 3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23. I am satisfied with AWB effectiveness</td>
<td>0 0 3 2 0 0 0 0</td>
<td>17 5 3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24. Overall I am satisfied with the AWB</td>
<td>0 0 4 1 0 0 0 0</td>
<td>16 5 3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25. If the AWB were available I would use it</td>
<td>0 4 1 0 0 0 0 0</td>
<td>11 5 2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Perceived benefit: The valuation of benefits of the AWB to the user</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Moderately agree</th>
<th>Undecided</th>
<th>Moderately disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Respondent No.</th>
<th>Respondent total</th>
<th>Rating average</th>
</tr>
</thead>
<tbody>
<tr>
<td>26. AWB assists the cognitive workload in task performance</td>
<td>0 0 5 0 0 0 0 0</td>
<td>15 5 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27. AWB assists the ability to efficiently and effectively complete the task</td>
<td>0 2 3 0 0 0 0 0</td>
<td>13 5 2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28. AWB assists in managing the information overload</td>
<td>0 2 3 0 0 0 0 0</td>
<td>13 5 2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29. AWB provides more time for actual analysis</td>
<td>0 2 1 1 1 1 0 0</td>
<td>16 5 3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30. AWB helps the effective management of the information space</td>
<td>0 2 2 1 0 0 0 0</td>
<td>14 5 2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usage of the system: Extend of usage</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Moderately agree</th>
<th>Undecided</th>
<th>Moderately disagree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Respondent No.</th>
<th>Respondent total</th>
<th>Rating average</th>
</tr>
</thead>
<tbody>
<tr>
<td>31. Used to help make decisions</td>
<td>0 2 2 1 0 0 0 0</td>
<td>14 5 2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32. Use to communicate information with other stakeholders</td>
<td>0 0 5 0 0 0 0 0</td>
<td>15 5 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33. Used to support business position or argument</td>
<td>0 3 2 0 0 0 0 0</td>
<td>12 5 2.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Using the questionnaire, five evaluators scored the experiences of performing the competitive analysis information gathering activity, using the Analyst Work Bench artefact. Questionnaire results and ratings are presented Table 7-1.

Numbered questions are first listed in the left most columns followed by actual rater scoring results. *Respondent number* is weighted using the scale range and *respondent total*, represents the sum of the number of raters, participating in each question. The *rating average* was calculated by dividing the weighted value, by the sum of the respondents. The mean of the scale, indicating neutral (or undecided), is 4, meaning that a response rating of 3 or less, indicates moderate to good agreement and response rating of 5 or greater, moderate to strong disagreement to the questions posed.

|--------------------|---------------|-------------------------|

To assist discussion flow the questionnaire dimensions are broken out and presented in graph form with scale range listed on the y-axis and dimension instruments on the x-axis. System Quality (Figure 7-12) suggests that the artefact is generally easy to use, but can be cumbersome. This we interpret as an initial indication of user interaction / interface usage issues. Although evaluators did receiving training they moderately agree that further training and greater familiarity would be an advantage. A greater resource requirement would have a management implication for general AWB roll out and deployment.

*Figure 7-12 System quality - Assessment of operational characteristics*

<table>
<thead>
<tr>
<th>Is the AWB easy to use</th>
<th>Is it user friendly</th>
<th>I find it easy to get AWB to do what I want</th>
<th>Is it easy for me to become skillful using the AWB</th>
<th>I believe the AWB is cumbersome to use</th>
<th>Using AWB requires a lot of mental effort</th>
<th>Using AWB required a learning overhead</th>
<th>Using AWB can be confusing</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Content quality ratings (Figure 7-13) report moderate agreement on the artefact providing information in a useful manner. Importantly there is agreement on the system
providing the majority of the analyst information requirement. There is further agreement on the information provided being usable and readily consumable by the analyst. Overall rating agreements point towards an ontology that in terms of information identification and provision is proving effective.

**Figure 7-13 Information quality - Content**

![Graph showing information quality - Content](image)

Context and linkage instruments (Figure 7-14) reports further moderate agreement on the ontologies contributing to structuring of the information requirement, necessary for competitive analysis task support. Although there is agreement that relevant information is indeed provided, the search and navigation mechanism is rated as neutral, indicating that improvement is required. We interpret these ratings to indicate that the ontology has been successful in structuring and organising task information requirement, but used as an interactive/display paradigm, it is not overly successful. Further possible consideration, is that rater background profiles (accounting, business and economics), suggest a more conservative approach to new technology usage.

**Figure 7-14 Information quality - Context and linkage**

![Graph showing information quality - Context and linkage](image)
User satisfaction (Figure 7-15) received moderate agreement (3.2) echoing ratings from the context and linkage dimension. Interestingly, raters are in agreement (2.2), that they would use the artefact if available, despite some previous reservations. Further discussion with the evaluator focus group, attributed this to the artefact being: i) better than having to do the activity manually and; ii) the absence of other artefacts in the space.

Figure 7-15 User satisfaction – Extent of positive or negative feelings towards artefact

Perceived benefits from AWB usage (Figure 7-16) report moderate agreement (3) on supporting task performance. Agreement is reported in terms of personal information management (2.6) and moderate agreement on the more general information space (business filing) management (2.8). The moderate agreement as to whether there was a time transfer element from manual information gathering to actual analysis (3.2), we attribute to the previously mentioned user interface issues. These ratings do additionally point to the ontology cognitively reflecting the analyst information requirements, and information association patterns.

Figure 7-16 Perceived benefit – to the user

System usage (Figure 7-17) reports agreement to moderate agreement on artefact usefulness, to support current business analysis findings or decision making. There was similar agreement on artefact usage as a means to share information with other stakeholders. Focus group discussion indicated that the generated report, containing
extracted information nuggets, would be the primary means of achieving this, but some form of direct access to the knowledge repository would also be welcome.

Figure 7-17 Usage of the system – Extent of usage

7.5 Summary

Usability evaluation provides additional insight as to AWB information system and through it, the CAO. A questionnaire used with the analyst focus group, assessed information quality provision; perceived benefits to the competitive analysis task and inform on general user assessment. Information quality investigates the quality of artefact output in terms of:

i) content provision, where the level of agreement between analysts indicate the ontologies effectiveness, in satisfying the majority of the information requirement, and use in supporting relevant information identification

ii) information context and linkage also highlighted artefact contribution to structuring the information requirement and organising the competitive analysis task. Results also found that using the AWB’s CAO based navigator, as the means of analyst interaction, was not as successful as anticipated and drew analyst attention away from the ontologies overall structuring contribution.

Artefact perceived benefits reports better agreement on contribution to information management, at the personal level, than at the general filings level. Perceived benefit indicates the ontologies ability to capture domain semantics and knowledge with respect to competitive analysis information requirement, but records issues with user interface usage. System usage did report artefact usefulness in supporting current business analysis practices and decision making with analyst reporting a willingness to use the AWB were it available.

Overall the CAO as an artefact does assist the analyst in performing a competitive analysis, through information requirement provision, in a manner that is representative of
how an analyst, on a cognitive and personal level, associates that information. Using the ontology as the information systems interactive medium, can result in information context confusion and introduce instances where information may be overlooked. Future work here could look to the use of different user interface interaction and visualisation approaches that abstracted the ontology away from the end user. Other comments from discussion with the analyst focus group fell into the categories of display and navigation improvements, greater levels of context retention, and web based real time report access. The ontologies utility is further indicated through its direct contribution to task structuring and structured information provision, in a manner that directly supports task performance.
Chapter 8

Conclusions and Future Work

The research presented has its origins with the fundamental business question, as to whether, there was some means of assisting analysts engaged in competitive analysis, identify “semantically camouflaged” information, within the narrative section of U.S. SEC Form 10-Q’s filings. Research challenges addressed were:

- understanding the language of business discourse, its interpretation and constraints from task performance
- modelling the intricacies of the information requirement where analysts possess broad but lack specific details as to information sought
- the effect that varying information interpretation can have on analytic parameters and outcomes.

Experiments aligned to competitive analysis, compared the activity performed manually and using the AWB. The AWB represented ontology and linguistic analysis instantiation, as an application artefact, also used as the experimental platform. Conclusions relating to: performance results are summarised under Semantic Paths Modelling, Semantic Paths Use and Semantic Paths Evaluation. Future work relating to each area and further possible enhancement of the CAO, through semantic relationships, use with XBRL, and web enablement as part of the wider financial ecosystem are also discussed.
8.1 Conclusion

8.1.1 Semantic Path Modelling

• general domain related lexical resources, proved to be of limited use due to their general language coverage. Domain heuristics relating to the specific financial sub-domain of competitive analysis was therefore used as the language lexical resource

• domain heuristics with discourse analysis, allowed an expectation of what to look for, and what can be expected of the text

• formalising business argumentation categories and concept map development provides a basis for domain linguistic modelling (Semantic Paths) and ontology construction using DOGMA

• individual Semantic Paths represent a small element of the overall competitive analysis information requirement. Expressed through the CAO, the Semantic Paths represent the overall information requirement

• the CAO provides the schema to target information for extraction and a unifying model for post extraction semantic interpretation

• in structuring the information requirement based on analyst heuristics, the CAO also served, to structure the manual information gathering activity, associated with competitive analysis

• Semantic Paths that were continually found to be un-instantiated suggest either: i) pattern mining rules that require expansion or correction; or ii) ontological areas that would benefit from tuning or re-modelling, e.g. the Disposal context area

• increased and wider exploitation of Semantic Paths using the system suggesting sufficient ontology coverage of the competitive analysis domain

• overall the CAO, instantiated in the AWB, and supported by linguistic analysis provides an automated and uniform approach to tackling the information requirement of a complex area
8.1.2 Semantic Path Evaluation

8.1.2.1 Performance

- performance evaluation results report a precision increase of 37%, between manual (16.7%) versus system based (23%) competitive analysis approaches
- the presence of false positives, in the a-posteriori evaluation results, for this experiment, represents items identified using the system, that were not previously identified during manual analysis. Although false positives experimentally represent items that should be in the reference standard, here their presence indicates that the system is functioning as designed, supporting the identification of items, that otherwise would have been overlooked
- the precision value increase is directly attributable to the increased use of Semantic Paths by the analyst when using the system

8.1.2.2 Usability

- the Semantic Paths provided by the AWB increased the level of analyst information exploitation relative to manual analysis
- rater agreement on relevant information, extracted using the system, provided Semantic Paths increased in the majority of instances
- Semantic Paths in terms of information identification and provision, is effective, but the ontology as an interactive user interface has short comings
- analysts agree that the system does support task performance and information management at the personal level

8.1.2.3 General

Evaluating precision was designed with a-priori and a-posteriori evaluation stages that:
- accommodated situations where evaluators begin with a system-generated set of potentially relevant items that they select a sub-set from
- attempted to use performance metrics to provide wider insight into Semantic Path usefulness along system functionality lines
• accommodates the identification and re-introducing of a-posteriori false positives, as a means to refine a robust reference standard, in an environment where evaluator resources are minimal and training data does not exist.

8.2 Future Work

8.2.1 Semantic Relationship Extraction

Linguistic interpretation of co-occurring CAO terms, to define heuristics for potentially relevant semantic relationships, would enhance the overall ontology. Inspection of candidate sentences for linguistic regularities using a semi-automatic process, that involves dependency parsing (de Marneffe 2006; de Marneffe 2008), and manual inspection combination could then be undertaken. With relationship defined by two arguments, the arguments can be further defined by semantic types e.g. \{B: boolean, S: string, R: range, A: amount, T: temporal, D: document-id, NE: named-entity\}. The idea is to semantically ground relationships and their arguments at the defined semantic space level through these semantic types. Preliminary results in Table 8-1, lists six of the more frequently occurring relationships found, based on the analysis of 625 co-occurring CAO terms, sentences from a sample of 32 business filings, comprising 48,782 sentences with 147,885 CAO terms. Due to the underspecified nature of the implied semantics, some of the arguments were typed with two semantic types, according to possible different interpretations of the sentence context.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Argument 1</th>
<th>Type</th>
<th>Argument 2</th>
<th>Type</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>hasCost</td>
<td>Product</td>
<td>NE</td>
<td>Cost</td>
<td>A</td>
<td>58</td>
</tr>
<tr>
<td>hasRequirement</td>
<td>Customer</td>
<td>NE</td>
<td>Requirement</td>
<td>S</td>
<td>44</td>
</tr>
<tr>
<td>includedInStatementOf</td>
<td>Revenue</td>
<td>A</td>
<td>Income</td>
<td>A</td>
<td>39</td>
</tr>
<tr>
<td>foundHasIncreasedBy</td>
<td>Product</td>
<td>NE</td>
<td>Increase</td>
<td>A</td>
<td>34</td>
</tr>
<tr>
<td>hasPrice</td>
<td>License</td>
<td>D</td>
<td>Price</td>
<td>A</td>
<td>29</td>
</tr>
<tr>
<td>hasDelay</td>
<td>Product</td>
<td>NE</td>
<td>Delay</td>
<td>T</td>
<td>27</td>
</tr>
</tbody>
</table>

Evaluating whether this combination approach of relationship identification with semantic grounding of relationship arguments, would significantly enhance the CAO, through relationship addition, has yet to be investigated. Future work could also assess whether
some of the relationships which act more like object class attributes, than relations between object classes, could be further refined.

8.2.2 Wider Business Utility

The current XBRL 2.1 specification lacks provision for informative tagging of filings disclosure sections. Using the CAO to semantically annotate disclosure sections and qualification notes would allow association of the sections with:

- financial instruments, enabling text qualification statements to be considered when analysing financial figures
- wider business information types, enabling holistic search and retrieval of related, but relevant information and
- suggest XBRL taxonomy extensions to cater for textual discussion section and foot notes categorisation and mark-up.

The investigations could leverage previous research that used an OWL formalised version of the CAO to assist with semantically enhanced passage retrieval from business filings (Thai 2008). For the wider business community if financial standard interoperability is an aspiration, then an immediate challenge requiring investigation, is the ability of Semantic Web formulism’s to represent the more complex ontological constraint (e.g. accounting jurisdictional and regulatory rules), contained within the XBRL formula and calculation link bases (O’Riain 2012c).

8.2.3 Web, Eco Systems and Dataspaces

Analysts outlined interest in web based real time access to analysis output. Making the instantiated ontology available as Linked Data, would provide a common interoperable format and model, for web based data financial data linking and sharing (O’Riain 2012a). Sharing both the instantiated CAO and analysis results, in a Linked Data format, would make the output linkable and integratable with other Open Data sources, as part of a wider financial eco system (O’Riain 2012b). Previous work that made the CAO available in OWL format (Thai 2006) could be leveraged as a first step. Recently the idea of schema-less environments with massive volumes of Open Web/closed datasets and high semantic
heterogeneity or (dataspace), has emerged. Due to the volume of information, business analysts interaction with such an environment face consumption challenges such as:

i) accommodating best-effort query responses and answer through efforts such as natural language vocabulary independent querying (Freitas 2011b; Freitas 2011c; Freitas 2012a)

ii) catering for multilingual access to financial and business data (Declerk 2008)

iii) simplified multidimensional OLAP type access to financial ecosystems and dataspaces (Kampgen 2012)

iv) vocabularies to define, processes to introduce, and frameworks to implement, provenance awareness associated with financial data and objects (Freitas 2010; Freitas 2011a; Freitas 2012b)

v) approaches and methods that address low cost curation of reusable Open Data sets (Curry 2012a).

Additionally as the importance of intangibles, such as adherence to sustainable practices, becomes an increasing fundamental part of risk analysis associated with a company’s health check, the ability to source, consolidate and meaningful present sustainable data (e.g. carbon emissions KPI’s), with business data, will become a basic requirement (Curry 2012b; Curry 2012c).

8.2.4 Evaluation

Information systems evaluation asks questions about system performance, in relation to system or user centred objectives (Saracevic 1975; Saracevic 1995). Systems focussed on IR performance metrics target algorithm and retrieval method effectiveness, but do not contribute to understanding on user and utility related characteristics. Where user activity is central to system usage, utility and usability dimensions, should be additionally considered as part of an evaluation framework that moves towards an overall quality measure. The view is shared by Davenport, whom in commenting on an effective knowledge management system, notes that such a combination of technical and human elements is required in a success measure (Davenport 1998a). Evaluating both quality elements for semantic supported technology evaluations, would move towards a more thorough and inclusive
system effectiveness measure that would also serve to ground technical evaluation with end user usage. It would progress insight as to how to evaluate information, that helps resolve given problems or directly bears on given actions (Saracevic 1975).

Cognisant of the importance of user involvement, our approach considers both qualitative and qualitative aspects to semantic evaluation. Other research work that developed a dual approach, was the reference model and evaluation model, used as part of a semantic metadata evaluation (Lei 2007). The evaluation model outlined fundamental principles and quality instruments to detect issues. There is scant evidence of the development and use of dual frameworks across the literature. Formalising our qualitative and qualitative approach, with further experiment on use cases, that require a dual approach, would advance discussion as to its wider feasibility and acceptance.

8.3 Summary

The competitive analysis ontology uses domain heuristics as the lexical resource, to formally represent analyst information requirements as a series of Semantic Paths. Its construction combines, eliciting softer and intangible domain knowledge, with discourse analysis familiarity, to develop an expectation on what to look for, and what can be expected of the text. The insight drives generation of business argumentation categories and concept maps, providing the basis for domain linguistic modelling and ontology building. Overall the Semantic Paths expressed through the CAO, supports the capture of domain semantics and provide a semantic interpretative mapping of the information requirements, for the business analyst.
## Appendix I  Discourse Analysis Term Listing

<table>
<thead>
<tr>
<th>3rd party</th>
<th>certify</th>
<th>deliver new</th>
<th>gross sales</th>
<th>major percentage</th>
<th>Permit</th>
<th>replace</th>
<th>substantial majority of</th>
</tr>
</thead>
<tbody>
<tr>
<td>access</td>
<td>change</td>
<td>Department</td>
<td>Growing</td>
<td>major portion</td>
<td>personnel</td>
<td>replenish</td>
<td>substantially greater financial</td>
</tr>
<tr>
<td>acquire</td>
<td>channel</td>
<td>detain</td>
<td>Growth</td>
<td>majority of</td>
<td>Plan</td>
<td>reputation</td>
<td>takeover</td>
</tr>
<tr>
<td>acquisition</td>
<td>channel partner</td>
<td>detain .. market</td>
<td>Headcount</td>
<td>management</td>
<td>policy</td>
<td>requirements</td>
<td>technical</td>
</tr>
</tbody>
</table>
| act on    | charge  | detain .. market acceptance | Hold | management buyout | population | research &
<p>|           |         |             |             |                  |         | and development | technology |
| act upon  | Code    | detain .. release | hold .. market | manufacture | position | research cycle | third party |
| action    | collaboration | detain .. Introduce | hold .. market acceptance | market | postpone .. Introduce | research time frame | time |
| adapt     | competition | difficult .. predict | hold .. release | market channel | postpone | resources | time scale |
| administration | competitive | direct | hold .. introduction | market percentage | postpone .. market | Restructuring | time table |
| adopt     | competitor | director | Holdup | market place | postpone .. market acceptance | Revenue | timing of |
| advise    | competitor | discharge | holdup .. market | market presence | postpone .. release | right .. use | trend |
| against   | compliancy | dispose | holdup .. market acceptance | market value | predict | rights | type |
| agree | compliant | divest | holdup .. release | method | Price | royalties | unable |
| agreement | comply | divide | holdup .. Introduce | methodology | price .. goods sold | royalty payments | uncertain |
| alter | concentrate | divide | impact .. ability .. sell | modification | price .. sale | S. &amp; D. | understand |
| amalgamate | conditional | economic | Increase | modify | proceeding | S.G.&amp;A. | unfilled |
| amortized | constitute | employee | industry\industrial practice | movement | process | sale | unsure |
| amount | contracts | entitlements | industry\industrial standards | net percentage | proclaim | sales forecast | vary |
| announcement | contractual | evolving industry\industrial norm | Inform | net\nett income | procure | scale down | versus |
| applicability | control | evolving industry\industrial practice | Infrastructure | net\nett loss | produce new | Scheduled | waiver |
| applicable | copyright | evolving industry\industrial standards | Intangible | nett\net profit | Product | scheme | well being |
| as opposed to | cost | examine | | new | profit after tax | sell | worker |
| assimilate | cost .. sale | expand | Integration | new development | profit before tax | sentiment | workforce |
| attention | cost .. goods sold | expenditure | Intellectual | new formula | program | Services |
| backlog | cost(ing) item | expenses | intend .. continue | new produce | prospective | settlement |
| belief | current asset | fair value | Introduce | order book | provision | shareholder |</p>
<table>
<thead>
<tr>
<th>belief</th>
<th>customer</th>
<th>fixed asset</th>
<th>Introduction</th>
<th>organisation</th>
<th>purchase</th>
<th>significant market presence</th>
</tr>
</thead>
<tbody>
<tr>
<td>bigger percentage</td>
<td>customer requirement(s)</td>
<td>follow in</td>
<td>Invest</td>
<td>outsourcing</td>
<td>pursue</td>
<td>significantly greater financial</td>
</tr>
<tr>
<td>bigger\big portion</td>
<td>Cut</td>
<td>follow up on</td>
<td>Item</td>
<td>overall</td>
<td>R &amp;\and D</td>
<td>similar .. features</td>
</tr>
<tr>
<td>board</td>
<td>cut back</td>
<td>forecast</td>
<td>joint venture</td>
<td>overlay</td>
<td>rank</td>
<td>spend</td>
</tr>
<tr>
<td>board .. management</td>
<td>cut down</td>
<td>formula</td>
<td>large percentage</td>
<td>pace</td>
<td>ratio</td>
<td>spending levels</td>
</tr>
<tr>
<td>break</td>
<td>Deal</td>
<td>frontlog</td>
<td>larger\large portion</td>
<td>partner</td>
<td>reduce</td>
<td>split</td>
</tr>
<tr>
<td>bring down</td>
<td>decline</td>
<td>fund</td>
<td>Leadtime</td>
<td>partnership</td>
<td>Redundancy</td>
<td>statement .. intent</td>
</tr>
<tr>
<td>bulk of</td>
<td>decrease</td>
<td>Geography</td>
<td>Legal</td>
<td>patent</td>
<td>redundant</td>
<td>static</td>
</tr>
<tr>
<td>buyout</td>
<td>defend</td>
<td>goodwill</td>
<td>licence revenue</td>
<td>payoff</td>
<td>region</td>
<td>status</td>
</tr>
<tr>
<td>cancel</td>
<td>Delay</td>
<td>greater .. financial</td>
<td>License</td>
<td>people</td>
<td>regulation</td>
<td>stop</td>
</tr>
<tr>
<td>cannot/can't</td>
<td>delay .. introduce</td>
<td>greater financial</td>
<td>license payments</td>
<td>percentage margin</td>
<td>relationship</td>
<td>Strategy</td>
</tr>
<tr>
<td>capital</td>
<td>delay .. market</td>
<td>Limit</td>
<td>percentage profit</td>
<td>Release</td>
<td>subcontract (ing)</td>
<td></td>
</tr>
<tr>
<td>cash</td>
<td>delay .. market acceptance</td>
<td>gross percentage</td>
<td>Litigation</td>
<td>perceived market value</td>
<td>release date</td>
<td>subsidiary</td>
</tr>
<tr>
<td>cease</td>
<td>delay .. release</td>
<td>gross profit</td>
<td>long sales cycle</td>
<td>permission</td>
<td>renew/al</td>
<td>substantial bulk of</td>
</tr>
</tbody>
</table>
Appendix II  Business Argumentation Category Concept Maps

Figure 8-1 Semantic Paths Marketing Context Concept Map
Figure 8-2: Semantic Paths Acquisitions Context Concept Map
Figure 8-3 Semantic Paths Relationship Context Concept Map
Figure 8.4 Semantic Paths Headcount Context Concept Map

Overall, R & D, sales, manufacture, administration, acquisition, region, subcontracts, outsource
Figure 8.5 Semantic Paths Profit Context Concept Map

- Percentage
- Pricing
- Market
- Effect
- Which
- Profit (gross/net)
- Restructure
- Redundancy
- Loss
- Infrastructure
- Cost
- Amount
- Expenditure
- S&D
- Capital
- License
- Royalties
- Control
- Employee
- SG&A
- Cost of goods sold
- Charge
- Increase
- Plan
- What is
- What
- Reduces
- Plan
- Versus
- Increase
- Plan
- Amount
- Amount
- Amount
Figure 8-6 Semantic Paths R&D Context Concept Map
Appendix III  Form 10-Q HTML Extract

During the first quarter of fiscal 2002, the Company began expensing rent in the fourth quarter of fiscal 2002 of approximately $3 million per quarter, as it renewed its lease prior to February 1, 2003 for an additional term of approximately three years. During the three months ended April 30, 2003, the Company was required to restrict approximately $21.0 million of cash as collateral in connection with a new European banking arrangement, which is included in long-term restricted cash. The Company entered into a lease agreement for the lease of approximately 40 acres of land adjacent to the Company’s San Jose, California headquarters (the Land Lease). The Company intends to construct additional corporate offices and research and development facilities on the leased land. Under the land lease agreement, the lessee will finance up to $331 million of land and associated costs. The land lease has an initial term of five years with renewal options. Rent payments commence at the beginning of the third year. Although no rental payments were being made under the land lease prior to February 1, 2003, the Company began expensing rent in the fourth quarter of fiscal 2002 of approximately $3 million per quarter, as it re-evaluated the scope and timing of the construction for additional corporate offices. The agreement qualifies for operating lease accounting treatment under Statement of Financial Accounting Standards No. 13, Accounting for Leases. The Company is subject to legal proceedings and other claims that arise, such as those arising from domestic and foreign tax authorities and employee-related matters, in the ordinary course of its business. While management currently believes the amount of ultimate liability, if any, with respect to these actions will not materially affect the Company’s financial position, results of operations, or liquidity, the ultimate outcome of any litigation or claim is uncertain, and the impact of an unfavorable outcome could be material to the Company.

© Séan O’Riain 2012
Appendix IV  DOGMA Ontology Modelling Methodology

preparation and scoping
- define user requirements
- define purpose
- identify domain experts
- compile knowledge resources
- scope knowledge resources
- select relevant passages
- define scenario

domain conceptualization
- knowledge elicitation
  - brainstorming
  - abstraction
  - compile baseline taxonomy
- knowledge breakdown
  - knowledge negotiation
- knowledge discovery
  - segmentation
  - highlighting

application specification
- verbalising elementary sentences
- lexon engineering
  - create lexons
  - refine lexons
  - ground lexons
  - create meta-lexons
- structuring
  - defining competency questions
  - define semantic constraints
  - answer competency questions
# Appendix V  Deloen & McLean Success Measures

*Table 8-2: Instrument Summary based upon eCommerce biased Success (Seddon 1996)*

<table>
<thead>
<tr>
<th>Segment</th>
<th>Model Instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td>System quality</td>
<td>Adaptability</td>
</tr>
<tr>
<td></td>
<td>Availability</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
</tr>
<tr>
<td></td>
<td>Response Time</td>
</tr>
<tr>
<td></td>
<td>Usability</td>
</tr>
<tr>
<td>Information quality</td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Ease of understanding</td>
</tr>
<tr>
<td></td>
<td>Personalisation</td>
</tr>
<tr>
<td></td>
<td>Relevance</td>
</tr>
<tr>
<td></td>
<td>Security</td>
</tr>
<tr>
<td>Service Quality</td>
<td>Assurance</td>
</tr>
<tr>
<td></td>
<td>Empathy</td>
</tr>
<tr>
<td></td>
<td>Responsiveness</td>
</tr>
<tr>
<td>Use</td>
<td>Nature of use</td>
</tr>
<tr>
<td></td>
<td>Navigation patterns</td>
</tr>
<tr>
<td></td>
<td>Number of sites visits</td>
</tr>
<tr>
<td></td>
<td>Number of transactions executed</td>
</tr>
<tr>
<td>User satisfaction</td>
<td>Repeat purchases</td>
</tr>
<tr>
<td></td>
<td>Repeat visits</td>
</tr>
<tr>
<td></td>
<td>User surveys</td>
</tr>
<tr>
<td>Net benefits</td>
<td>Cost savings</td>
</tr>
<tr>
<td></td>
<td>Expanded markets</td>
</tr>
<tr>
<td></td>
<td>Incremental additional sales</td>
</tr>
<tr>
<td></td>
<td>Reduced search costs</td>
</tr>
</tbody>
</table>
## Appendix VI  Evaluation Questionnaire

<table>
<thead>
<tr>
<th>Circle the appropriate numbers</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Moderately Agree</th>
<th>Undecided</th>
<th>Moderately Disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System Quality: Assessment of operational characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Is the AWB easy to use</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Is it user friendly</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I find it easy to get AWB to do what I want</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Is it easy for me to become skilful using the AWB</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. I believe the AWB is cumbersome to use</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Using AWB requires a lot of mental effort</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Using the AWB required a learning overhead</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Using AWB can be confusing</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Information Quality: Quality of the AWB output</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. The output is presented in a useful manner</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. The content representation provided by the AWB is logical</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. The information provided is relevant and helpful for the task</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. The information content meets your information requirements</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. The information provided is meaningful</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. AWB makes it easy for me to extract information of use</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Context &amp; Linkage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. The information navigation process is useful for information identification</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. The information navigation process is logical and fit</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. Information is presented in a way that provides a useful aggregate view</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. The representation mechanism is successful in structuring the analysis task</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. Does the information navigation process assist the cognitive workload in task performance</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20. Provides information in context in a manner that is understandable, accessible and applicable to the task</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>User Satisfaction : Extent of positive or negative feelings towards the tool</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21. My information needs are met</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22. I am satisfied with AWB efficiency</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23. I am satisfied with AWB effectiveness</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24. Overall I am satisfied with the AWB</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25. If the AWB were available I would use it</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Perceived benefit: The valuation of benefits of the AWB to the user</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26. AWB assists the cognitive workload in task performance</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27. AWB assists the ability to efficiently and effectively complete the task</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28. AWB assists in managing the information overload</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29. AWB provides more time for actual analysis</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30. AWB helps the effective management of the information space</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Usage of the system: Extend of usage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31. Used to help make decisions</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32. Use to communicate information with other stakeholders</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33. Used to support business position or argument</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bibliography


Semantic Paths in Business Filings Analysis


International Conference on Intelligent Systems for Molecular Biology, pages 84-87, AAAI Press.


Gartner Research (2003b). Taxonomy Creation: Bringing Order to Complexity. Select Q&A.


dynamic, situational definition." *Journal of Information Processing & Management*

Associates, Hillsdale N.J.

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., Van de Velde, W.,
Cambridge Massachusetts, MIT press.

Proceedings of the National Dutch AI Conference (NAIC'95).

breaking financial news: The AZFin text system." *ACM Transactions on Information

Extension. *Proceedings of the 4th International Semantic Web Conference (ISWC).*
Galway, Ireland: 593--606.


Information Systems for Search, Integration and Analysis." *IEEE Data Engineering

documents containing accounting information of listed firms using Semantic Web

Smiple Size Rquirement." *Physical Therapy* 85(3).

dictionary. *Proceedings of the Fourteenth International Joint Conference on Artificial
Intelligence (IJCAI-95), San Mateo, California, Morgan Kaufmann.*

Foundations*, Addison-Wesley.

Foundations of Intelligent Systems, Proceedings of the 15th International Symposium
on Methodologies for Information Systems, Springer Verlag.


*SIGMOD Record Special Issue* 31(4): 12-17.


