AN ONTOLOGICAL APPROACH FOR MONITORING AND SURVEILLANCE SYSTEMS IN UNREGULATED MARKETS

A Thesis submitted to the University of Manchester for the degree of PhD in Business Administration

In the Faculty of Humanities

2013

Mohamed Zaki

Faculty of Humanities
Table of Contents

Acknowledgements ...................................................................................................................................... xi
Preface .................................................................................................................................................. xii

1 Introduction ........................................................................................................................................ 1
  1.1 Background and motivation ........................................................................................................ 1
  1.2 What are the drivers in issuing securities in an unregulated market? ................................... 2
  1.3 The vulnerability of the unregulated market ........................................................................... 3
    1.3.1 Manipulative practices in unregulated markets .............................................................. 4
    1.3.2 Manipulation Participants ............................................................................................... 5
  1.4 Market surveillance in unregulated markets ........................................................................... 7
  1.5 Aims and Objectives .................................................................................................................. 9
  1.6 Research Questions ..................................................................................................................10
  1.7 Structure of the thesis .............................................................................................................. 11

2 Literature Review ............................................................................................................................. 13
  2.1 OTC and Pink Sheets Market Structure and Trading Process Background .......................... 13
    2.1.1 Market Tiers .................................................................................................................... 14
    2.1.2 The Trading Process ....................................................................................................... 18
  2.2 Review of financial economics literature .............................................................................. 20
  2.3 Market Manipulation: Importance, Definitions, and Types .................................................. 23
    2.3.1 Market Manipulation Types ............................................................................................ 26
    2.3.2 Information-Based Manipulation ....................................................................................31
  2.4 Market monitoring and surveillance systems: Definition and technologies .......................... 39
    2.4.1 Technologies which support market monitoring and surveillance systems ..................39
    2.4.2 Previous research on BI for fraud detection ................................................................. 46
    2.4.3 Existing systems and frameworks ................................................................................. 49
  2.5 Financial Market Ontology ...................................................................................................... 54
    2.5.1 Existing Financial Ontologies ......................................................................................... 57
  2.6 Gaps in the literature .................................................................................................................. 61

3 Research Methodology ...................................................................................................................... 64
  3.1 Description of the methodology ............................................................................................... 64
    3.1.1 The design science research cycle ................................................................................... 65
    3.1.2 Steps and outputs of design research methodology ....................................................... 66
    3.1.3 The philosophical grounding of design science research ............................................... 69
  3.2 Justification of the chosen methodology ................................................................................... 70
3.3 Methodology application .................................................................................................. 75

4 Financial ontology for fraud purposes .................................................................................. 78

4.1 The process of constructing the financial domain ontology ............................................... 80

4.2 Financial Ontology Methodology analysis ......................................................................... 84

4.2.1 Stage 1: Unitizing ......................................................................................................... 85

4.2.2 Stage 2: Categorizing ................................................................................................. 91

4.2.3 Stage 3: Filing in patterns ......................................................................................... 108

4.2.4 Stage 4: Member checking ....................................................................................... 110

4.3 Ontology Layering Architecture ..................................................................................... 114

4.3.1 Lexical Ontology Layer ............................................................................................ 115

4.3.2 Data Source Layer ................................................................................................... 118

4.3.3 Domain Layer .......................................................................................................... 123

4.3.4 Application Layer .................................................................................................... 126

4.4 Ontology Mapping .......................................................................................................... 146

4.5 Summary ......................................................................................................................... 151

5 “Stock-touting” Spam E-mails Text-Mining Application ....................................................... 154

5.1 Background ..................................................................................................................... 154

5.2 Instantiation of the Stock Spam E-mail Text-Mining Application ..................................... 155

5.2.1 Application Ontology .............................................................................................. 158

5.2.2 Ontology Mapping .................................................................................................. 176

5.3 Summary ......................................................................................................................... 178

6 BI Financial Market Service system for a financial market monitoring and surveillance system .............................................................................. 180

6.1 Background ..................................................................................................................... 180

6.2 Financial Ontology Instantiation as BI Financial Market Service system ....................... 182

6.2.1 Data Source Layer ................................................................................................... 184

6.2.2 Application Layer (BI Financial Market Service System) ........................................ 187

6.3 Case Study ....................................................................................................................... 189

6.3.1 Data Sources ........................................................................................................... 191

6.3.2 Information Management ...................................................................................... 192

6.3.3 Economic Analysis ............................................................................................... 201

6.3.4 Output Management .............................................................................................. 202

6.4 Summary ......................................................................................................................... 202

7 Discussion ........................................................................................................................... 204

7.1 Financial ontology for fraud purposes ............................................................................ 206
List of Tables

Table 2-1 Market Manipulations, Categories, Definitions and Types (adapted from Cumming et al. 2008; Diaz 2011; FINRA 2010; Investopedia 2011a; U.S. Securities and Exchange Commission 2011c) ...................................................... 29
Table 2-2 Review of Spam e-mails approaches .................................................... 34
Table 3-1 Philosophical Viewpoints of the three research paradigm perspectives (Vaishnavi and Kuechler 2004/5) .......................................................................... 70
Table 3-2 Characteristics of the research adapted from IS Design Research Framework (Hevner, March et al. 2004) ............................................................... 74
Table 4-1 Focus Group Rounds Process Overview .............................................. 83
Table 4-2 Total number of annotated concepts per case .................................... 88
Table 4-3 Pump and Dump Categorization Analysis .......................................... 95
Table 4-4 Insider Trading Case Categorization Analysis .................................... 99
Table 4-5 Marking the Close Categorization Analysis ....................................... 103
Table 4-6 Front Running Categorization Analysis .......................................... 108
Table 4-7 Analysis of the main Classes of the Financial Ontology ................. 110
Table 4-8 'Civil Case No' Analyzer ............................................................... 135
Table 4-9 Timeline Event Analyzer Output Sequences .................................... 145
Table 5-1 Macro Values .................................................................................. 163
Table 5-2 Regular Expression Symbol Ticker patterns .................................... 173
Table 5-3 Stock Quote Patterns Development ............................................... 176
Table 7-1 Summary of case studies ............................................................... 213
List of Figures

Figure 1-1 Manipulation Participants................................................................. 6
Figure 2-1 OTC Market Structure........................................................................ 15
Figure 2-2 Adapted Taxonomy of Market Manipulations and Abuse (Cumming and Johan 2008; Putnins 2009; Diaz 2011)................................................................. 27
Figure 2-3 Conceptual Architecture of the SONAR System (Kirkland and Senator 1999)..................................................................................................................... 50
Figure 2-4 Surveillance Workflow of the ThaiBMA (Mongkolnavin and Tirapat 2009)..................................................................................................................... 51
Figure 2-5 The Combination of Economic and Behavioural Analysis Approaches to Trading Analysis (Mongkolnavin and Tirapat 2009)................................................................. 52
Figure 2-6 Market Monitoring Framework (Díaz, Zaki et al. 2011)................ 53
Figure 2-7 FF POIROT model layers using Wigmore charts (Kingston et al. 2004) ................................................................. 59
Figure 3-1 Research Mapping adapted Design Research. Cited from (Vaishnavi and Kuechler 2004/5)............................................................................................ 66
Figure 3-2 Information Systems Design Research Framework (Hevner, March et al. 2004)............................................................................................................... 68
Figure 3-3 Research Mapping adapted Design Research. Cited from (Vaishnavi and Kuechler 2004/5)............................................................................................ 77
Figure 4-1 A taxonomy of elements in the manipulation scenario (Diaz 2011) .... 79
Figure 4-2 Constant comparative method (Grove 1988)........................................ 85
Figure 4-3 Examples of inconsistent annotations.............................................. 87
Figure 4-4 The Proposed Financial Ontology ..................................................... 112
Figure 4-5 Financial Ontology produced in PoolParty environment ............... 114
Figure 4-6 Financial Ontology Layering Architecture adapted from (Mikroyannidis and Theodoulidis 2012) ................................................................. 115
Figure 4-7 Investopedia Dictionary ................................................................. 117
Figure 4-8 Distribution of Investopedia Classes................................................. 117
Figure 4-9 The SEC data source ontology.......................................................... 119
Figure 4-10 Enforcement Class ........................................................................ 122
Figure 4-11 Litigation Releases Concept .............................................................. 123
Figure 4-12 Domain Ontology Implementation.................................................... 125
Figure 4-13 Semantic Search and Tag Events in PoolParty................................. 126
Abstract

Ontologies are a key factor of Information management as they provide a common representation to any domain. Historically, finance domain has suffered from a lack of efficiency in managing vast amounts of financial data, a lack of communication and knowledge sharing between analysts. Particularly, with the growth of fraud in financial markets, cases are challenging, complex, and involve a huge volume of information. Gathering facts and evidence is often complex. Thus, the impetus for building a financial fraud ontology arises from the continuous improvement and development of financial market surveillance systems with high analytical capabilities to capture frauds which is essential to guarantee and preserve an efficient market.

This thesis proposes an ontology-based approach for financial market surveillance systems. The proposed ontology acts as a semantic representation of mining concepts from unstructured resources and other internet sources (corpus). The ontology contains a comprehensive concept system that can act as a semantically rich knowledge base for a market monitoring system. This could help fraud analysts to understand financial fraud practices, assist open investigation by managing relevant facts gathered for case investigations, providing early detection techniques of fraudulent activities, developing prevention practices, and sharing manipulation patterns from prosecuted cases with investigators and relevant users.

The usefulness of the ontology will be evaluated through three case studies, which not only help to explain how manipulation in markets works, but will also demonstrate how the ontology can be used as a framework for the extraction process and capturing information related to financial fraud, to improve the performance of surveillance systems in fraud monitoring. Given that most manipulation cases occur in the unregulated markets, this thesis uses a sample of fraud cases from the unregulated markets.

On the empirical side, the thesis presents examples of novel applications of text-mining tools and data-processing components, developing off-line surveillance systems that are fully working prototypes which could train the ontology in the most recent manipulation techniques.
Declaration and Copyright Statement

Declaration

I hereby swear that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

The following three notes on copyright and the ownership of intellectual property rights:

i. The Author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the "Copyright") and he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and exploitation of this thesis, the Copyright and any Intellectual Property Rights and/or Reproductions described in it may take place is available from the Head of School Manchester Business School (or the Vice-President) and the Dean of the faculty of Humanities, for faculty of Humanities’ candidates.
To my family, my parents Elsayed Helmy, and Wegdan Elatar. They taught me everything that is essential in life. They gave me values, inspiration and support and raised me to be the person that I am. For all their love and sacrifice I thank God almighty. Special thanks to my lovely sisters Yara and Sara for their love and supporting me through difficult time. Also, I cannot forget my lovely nephews Mohamed and Ziad and my nieces Lama and Lojayna. Despite their age, they gave me the love and kindness.

This thesis is dedicated to all people who died in the Arab spring revolutions. In particular, Egyptians who are sacrificed their lives for the sake of liberty, justice and dignity.
Acknowledgements

First and foremost, I thank my academic supervisor Dr Babis Theodoulidis for his immense wisdom, constant support and guidance since the day I started my MSc. dissertation in Manchester Business School. I also thank my academic co-supervisor Dr Pedro Sampaio, who was always accessible and offered me insightful advice every time I needed it. Babis and Pedro are not only great supervisors, but also amazing people. It has been a pleasure and an incredible experience to work with and learn from them, and without doubt I will remember the time under their supervision as the most challenging, but also the most enjoyable and experienced years of my life.

I thank my colleague Dr David Diaz for his advice and friendship during the long hours we dedicated to our research. Also to my colleagues, Dr Stefanos Strickland, Azar Shahgholian, Abioye Olukorede, Hussain Al-Arrayed, Liwei Liu, Sahar Elkaremi, Ons Shamaileh, and Guilherme Biesek and other fellow research students, and friends at the Manchester Business School.

I gratefully acknowledge all the support from the people and colleagues in the University of Zagazig, Business administration department. My interest in research started while I was still an undergraduate student and I was motivated and encouraged at a very early stage by Professor Tarek Eldomiatty and Dr Al Kammar. I am grateful to Professor Rashed who gave me the chance to travel to the UK and for all his support in the first 3 years of my study. I owe thanks to the Head of the Department of Business Administration in Zagazig University Dr Ayman El Taher for his support.

Big thanks to all my friends in Manchester who supported me in the four years. Especially, Nazar Hamood, his wife Yara Mohamed, their kids Youssuf and Yazan. Also I would not forget Yara’s father Mr. Mohamed, Yara’s mother Shafika Abdulghafoor who were always like my parents here in the UK.

Finally, I thank my friends who were always there to share joys and sorrows and were a source of encouragement throughout my work on this thesis, many and special thanks to Mahmoud Marzouk, Ahmed Derballa, Mohamed Ghaly, Mohamed Abdulelah, Wael Moustafa, Bassem Elhalawany, Safwan and his wife Saima, Moutaz Elmasry, Yousry Ahmed and Hisham Omara.
Preface

Some of the work in this thesis has been presented as joint work with my supervisors and colleagues at the Manchester Business School. Case study 2, together with parts of Chapter 2, was presented as a conference paper under the title *Using Text Mining to Analyze Quality Aspects of Unstructured Data: A Case Study for ‘stock- touting ’ Spam Emails*, at the Americas Conference on Information Systems (AMCIS) 2010, (*Proceedings*, p. 364).

Case study 3, together with parts of Chapter 2, was presented as a conference paper under the title of *Financial Market Service Architectures: A “Pump and Dump” Case Study* at the SRII 2012 Global Conference, 23th-27th of July, San Jose, California, USA; the proceedings will be published by IEEE in a forthcoming special issue of the journal *SRII/IEEE*.

The researcher published papers on the application of data mining for the detection of manipulations in the stock markets that were not included here, as these were related, but not essential part of the main topic of this thesis. A paper was published in the *Journal of Manufacturing Technology Management* Special Issue on Intelligent Management Systems and Operations. Previous version of this paper were also presented at the 16th International Conference on Information and Software Technologies, IT 2010, Kaunas, Lithuania, 2010. Also, the researcher also co-authored a working paper on the application of data mining under the title of 'A Systematic Framework for the Analysis and Development of Financial Market Monitoring Systems' at the SRII 2011 Global Conference, March 30-April 2, San Jose, California, USA, and it is expected that proceedings papers will be published by IEEE in a forthcoming special SRII/IEEE journal.

The researcher also co-authored several papers on the application of text mining for analyzing customer feedbacks under the title “Analysing Customer Feedback Using Text Mining: A Service-Dominant Logic Perspective” paper is submitted to the *Journal of Service Research* on March 2012. The paper is in the second round review. Previous version of this paper was also presented at Applied Human Factors and Ergonomics, San Jose, California, U.S. under title “Service Modeling of Compliments and Complaints and Its Implications for Value Co- Creation”. Another version at Frontiers in Service Conference 2012 under title “Analysing Customer Feedback on Resource Utilisation: A Case Study Review of Compliments and Complaints through Text Mining.”
Publications related to this work

Journal papers:


Conference Papers:


1 Introduction

1.1 Background and motivation

Since the beginning of trading in Amsterdam in the 1600s, financial markets have been considered one of the main sources to raise capital. In fact, financial markets have many features that attract investors such as liquidity, transparency, and trading mechanisms. Liquidity is considered to be an important motivation for investment in the financial market, compared with less liquid investments such as real estate. Stock markets provide liquidity because investors can sell securities quickly and easily. Transparency refers to how much information the current and potential market participants hold about the trading process. Thus, exchange markets try to operate in transparent and efficient ways to guarantee fair execution of the trading and to protect investors from non-publicly visible bids or offers, and inaccurate financial reports (Hasbrouck 2007).

There are two forms of financial market: regulated and unregulated. Financial institutions registered with regulated markets such as the National Association of Securities Dealers Automated Quotations (NASDAQ) are subject to certain regulations and supervision handled either by self-regulatory organizations (SROs) or non-SROs. An SRO is an organization that enforces certain industry standards and requirements related to securities trading over financial organizations on behalf of the government, to protect investors; for example, the Securities and Exchange Commission (SEC).

Unregulated or unstructured financial markets include the over-the-counter market (OTC), which provides an alternative to stock exchange listings for the securities of issuers who either choose not to be listed on the regulated market or do not meet the relevant listing and financial reporting requirements. These differences in information environment, market structure, and type of investors present challenges to regulators and market participants in collecting and auditing information, and in analyzing the trading (Ang, Shtauber et al. 2011). Many investors and regulators consider the OTC market is the “Wild West” of the securities market because of its less regulated nature and the poor disclosure environment (Bollen and Christie 2009). Serious efforts need to be made to improve the stability, transparency and oversight of the OTC market (European Commission 2010), and regulators face challenges in providing a fair, transparent,
and robust regulatory and monitoring framework to protect unregulated markets from fraudulent activities. Thus, the continuous improvement and development of financial market surveillance is essential to guarantee and preserve an efficient market, detecting fraudulent patterns and also assuring a transparent, accurate and high-quality level of information.

This research focus on the OTC market, examining its structure and considering the different market manipulation types associated with it.

1.2 What are the drivers in issuing securities in an unregulated market?

One of the key drivers supporting the issue of securities in the unregulated market is that issuers do not need to register their securities with the SEC, nor are they obliged to submit current information in their reporting requirements. Securities quoted and traded in the OTC market, especially in the Pink Sheet tier, are classified into four categories.

First are the securities which list companies that are economically distressed or at an early stage of development; they may have been delisted from regulated markets, such as Enron and Adelphia communications (Bollen and Christie 2009). Other companies with distrustful management decide to “go dark” and cease reporting to the SEC while continuing trading in the OTC market. Around 480 companies decided to go dark between 1998 and 2004, issuing press releases stating that their motivation was to reduce compliance costs from disclosure requirements (Ang, Shtauber et al. 2011). Sometimes these companies “reinvented” themselves, by changing either the company name and ticker symbol or their business model (to a gold mining or petroleum company, for example) to take advantage of a specific industry’s high prices. Therefore, they can attract potential investors interested in a specific sector, or clean their previous record if they have legal issues or poorly performing investments. Furthermore, they are able to set new prices for the new securities (Lease 2010).

The second category comprises microcap companies who are not qualified for listing in the regulated market; they raise capital and trade under the penny stock scheme. The OTC market is the target of many microcap companies with low capital, normally less than $10 million in assets. Many of these companies eventually go bankrupt and go out of business.
The third category refers to foreign issuers who have stocks listed in their home country and want to trade via American Depositary Receipts (ADRs) in the OTC market, such as Nestle, Volkswagen, Heineken and Nintendo. Generally, these stocks avoid trading in regulated exchanges because of the expense of filing documents with the SEC, and also because this market has fewer regulations.

The last category refers to companies that trade occasionally and do not fully rely on the stock market to raise capital (Bollen and Christie 2009). Although some companies may show some level of disclosure they are still not covered against quality or investment risks. In particular, there are companies that list their securities in the OTCQB tier which should file reports with the SEC, yet these companies demonstrate a lack of transparency when transmitting public information as there are no financial or qualitative standards for this tier.

Due to the lack of reliable information and reports, it is difficult for investors to make decisions about the company’s management, products, services and finances. Furthermore, investors face an extremely illiquid market where any changes in demand, either real or artificial, could have a significant impact on stock prices and volume and, therefore, on investors’ positions (Lease 2010). Thus, in such a market, investors usually rely heavily on unofficial sources of information such as press releases, forums, spam e-mails and others to build their investment decision (U.S. Securities and Exchange Commission 2011).

However, some securities are quoted and trade in both the regulated and unregulated markets, and these must abide by the SEC rules, including the registration requirement, currency in financial reporting and monitoring of securities trading in both markets (Bollen and Christie 2009).

1.3 The vulnerability of the unregulated market

Historically, stock market manipulation has been an important issue for both the regulation of trading and market efficiency. Given that most exchanges have moved to electronic trading platforms to enable investors to trade across markets and jurisdictions, manipulative practices are found not only in single markets but extend to cross exchanges and markets (Cumming and Johan 2008). One of the reasons for establishing the SEC was to protect and reduce stock market manipulation (Aggarwal and Wu 2006).
Although manipulative practices in the regulated market seem to have declined, there is still scope for market manipulation in the developed and emerging unregulated stock markets, and this is a serious issue in the OTC market (Aggarwal and Guojun 2003). Regulated market established many restrictive rules and regulations to protect the market, improve transparency and limit fraudulent attempts. Furthermore, they invested heavily to adopt market surveillance systems. For example, NASDAQ bought SMARTS surveillance platform from Smarts Group International Ltd to detect fraudulent activities (Aitken, Harris et al. 2010).

Generally, the SEC issues warnings to investors through its website, advising on the vulnerability of the market and securities prices may be subject to fraud. In particular, microcap stocks have become the target of stock promoters and manipulators. Existing studies show that most manipulation cases occur in unregulated markets, which are relatively small, illiquid and require fewer disclosures of listed firms; in other words, they are subject to a less regulatory framework (Aggarwal and Wu 2006).

Therefore, investors face an extremely low liquid market; any change in demand, either real or artificial, may result in investors encountering difficulties in selling, especially, after the manipulators have dumped the prices and left the market. Indeed, the unregulated market still suffers from frequent attempts at fraud (Lease 2010). This research therefore aims to improve the current financial market surveillance systems in order guarantee and preserve an efficient market by detecting manipulation activities taking place in the unregulated market.

1.3.1 Manipulative practices in unregulated markets

Fraudsters can make profits through using various manipulative practices that distort prices at the expense of other investors and create information asymmetries (Cumming and Johan 2008). These manipulative practices have been classified as trade-based manipulation and information-based manipulation (Aggarwal and Guojun 2003).

Trade-based manipulation refers to a trader who engages in transactions to give the impression of stock price movements within a single market or across markets, by marking the close, layering bid/asks, washing trades, matching orders, spoofing etc. (Cumming and Johan 2008). For example, from 2004 to 2008 George
Georgiou was involved in a series of fraudulent schemes to manipulate the market for the common stock of four microcap companies in the OTC market, namely, the Avicena group Inc., Neutron Enterprises Inc., Hydrogen Hybrid Technologies Inc. and Northern Ethanol Inc. Georgiou orchestrated and directed manipulative trading in these stocks by using numerous nominee accounts held at offshore broker-dealers, and using a variety of traditional manipulative techniques, such as matched trades, wash sales, and marking-the-close (U.S. Securities and Exchange Commission 2009).

Information-based manipulation refers to fraudsters using Internet bulletin boards, newsletters, spam e-mails and chat rooms to post messages advising investors to react to specific stocks based on “inside” information. This scheme uses many fraudulent strategies to urge investors to buy stocks and make huge profits after dumping and selling their shares into the market; as the price of the stock falls, investors lose their investments. Indeed, the Internet has been used as the main manipulation channel of information dissemination because it is fast, high volume, reaches a mass audience and is cheap (U.S. Securities and Exchange Commission 2011).

These messages tend to have a professional structure using an "infallible" combination of "inside" information, fabricated or exaggerated information (hard to identify as fiction) about the company’s sales, revenue projections, acquisitions or new products and services stock market data (Lease 2010). Several studies have reported that most of the touted stocks are nearly always found listed on unregulated markets such as Pink Sheets and sometimes the OTC market (Frieder and Zittrain 2007). For example, Jonathan Lebed, a teenager, manipulated stocks by posting messages on Yahoo Finance forums and made profits of $800,000 (Aggarwal and Wu 2006).

1.3.2 Manipulation Participants

The manipulation scheme could be organized collaboratively with many participants, as shown in figure 1-1; they cooperate with each other in order to violate the securities. Many are publicly traded companies, represented by top shareholders or executives who directly or indirectly pay stock promoter companies to tout specific microcap securities using spam e-mails, misleading press releases and newsletters. The releases contain fictitious information about
microcap stocks, disseminated through legitimate financial news portals on the Internet (U.S. Securities and Exchange Commission 2011).

![Figure 1-1 Manipulation Participants](image)

Due to the lack of information about promoted companies, it is difficult to determine the extent of their involvement in fraud cases. However, their main manipulation strategy is creating “hype” and attracting investors to buy promoted stocks. Regularly, promoters and targeted companies collaborate with each other to mislead either naïve investors or investors who want to take advantage and the share profits of the manipulators. Manipulators could reinvest the money into their own business or deposit the profits into their personal accounts. Sometimes the money is used to liquidate the position of top shareholders or executives. In most cases, promotions facilitate and support ‘pump and dump’ manipulation schemes (Lease 2010).

Ideally, the business model of promoted companies increases the securities’ volume of stocks to satisfy their customers. Thus, they engage in an overabundance of manipulation strategies to gain profits. For example, promotion companies prefer to disseminate their fraudulent information after the close of markets or near their opening, to excite investors and make them react to the news. Another common strategy is the use of spam e-mails to get investor’s attention, and with information such as “be ready for tomorrow’s hot stock pick“ (Lease 2010). Earlier work has found that over 80% of all e-mail traffic is classified as spam, with 15% of these messages related to stock touts (Frieder and Zittrain 2007).

Although, brokers are the main market makers in unregulated markets, they have been known to abuse their power and manipulate the market through a variety of classic techniques, such as placing matched orders, wash sales, boiler room and
marking the close (U.S. Securities and Exchange Commission 2011). Rather than allow the markets to set the price of the issuers’ securities through the usual interchange of supply and demand, brokers deliberately or carelessly violate the market by participating in and furthering a market manipulation scheme to drive up the prices of penny-traded microcap stocks. Generally, brokers use communication services such as e-mails and instant messaging (IM) services to communicate with purchasers and sellers (U.S. Securities and Exchange Commission 2011).

Investors are another key agent in the trading process. Given the number of Internet fraud events that have occurred recently, investors should already know better how to identify advice from spammers. Investors may be inexperienced (naïve) or experienced. The former pursue fast wealth and, although they may be unsure about the reliability of the information in Internet messages, they choose to buy because they underestimate the liquidity risk associated with target stocks. Manipulators only need to find a few such naïve investors in order to make the campaign successful. However, experienced investors participate in the manipulation process to encourage the campaign and share the profits with the manipulators (Hanke and Hauser 2008).

1.4 Market surveillance in unregulated markets

Ideally, authorities and regulators, either exchanges or securities commissions employ sophisticated market surveillance mechanisms to detect such manipulative activities within their own market, or across markets. Normally, surveillance is conducted on equity securities markets by market regulators with in-house surveillance departments or by SROs. When exchanges become commercial entities and are deemed to be profit centres, they maintain their role to develop appropriate regulations and ensure the integrity of their markets to attract new companies, increase trading activity and market capital. Furthermore, in terms of competition, investors and issuers gain confidence in a marketplace that shows strong surveillance and monitoring capabilities (Cumming and Johan 2008). Generally, these SROs will either perform their own investigations and enforcement actions or refer matters to security commissions when unusual trading patterns are recognized (Canadian Securities Administrators 2011).
The opaque environment of unregulated markets and the absence of legal trade data disclosure requirements mean that OTC markets are less directly monitored than the regulated market. Furthermore, regulators have limited ability to detect inappropriate market behaviour and conduct in the market. Therefore, this could become a threat to the market’s integrity and efficiency (Francesca Taylor 2010).

Currently, regulators do not have full access to OTC trading data and market participants for analysis. Moreover, there is no collaboration agreement between domestic and international regulators to share the information essential for monitoring and surveillance purposes (Canadian Securities Administrators 2011). OTC markets have some concerns regarding the lack of granularity and accuracy of the trading information. Moreover, other issues such as conflicts of interest, business requirements, duplication in reporting, post-trade reporting and transaction reporting requirements depend on the reliability of the market (European Commission 2010).

Thus, in order to guarantee transparency, enhance investor confidence, gain higher market capitalization and mitigate abuses in these markets, regulators have to improve current market surveillance in order to minimize systemic risks, monitor, detect, and prevent potential market violations. Generally, fraud cases are complex and growing rapidly, making it difficult to keep knowledge of the fraudulent patterns up to date. Furthermore, the sheer volume of data, especially the unstructured sources, requires proficient surveillance and monitoring systems with high analytical capabilities to capture the fraud (Canadian Securities Administrators 2011).

The current financial market monitoring systems suffer from the absence of a financial ontology (Hilary, Yi-Chuan et al. 2009). There is thus an urgent need to build a financial ontology to combat fraud by efficiently managing vast quantities of financial data, providing better communication and knowledge sharing among analysts, providing a mechanism to demonstrate knowledge of the processes of financial fraud, understanding and sharing financial fraud logic operations, managing relevant facts gathered for case investigations, providing early detection techniques of fraudulent activities, developing prevention practices, and allowing reuse of these knowledge resources in different financial contexts (Kingston, Schafer et al. 2004).
1.5 Aims and Objectives

Previous work and the discussion above have exposed the weaknesses in the structure of the unregulated market. In particular, many studies have demonstrated the potential for contagion manipulation activities arising from the interconnectedness of market participants and the limited transparency between different counterparty relationships. Both Canadian and European securities commissions (European Commission 2010; Canadian Securities Administrators 2011) have recommended that regulators must conduct further research into ways to improve the surveillance systems. However, very few researchers have looked at OTC trading.

The aim of this research is the construction of a financial ontology for fraud purposes that contains a large set of financial concepts and definitions originating from multidisciplinary communities. The ontology serves as a reference map for researchers and practitioners to position their work in the context of market manipulations and market misbehaviours in general. In particular, the ontology acts as a semantically rich knowledge base in market monitoring systems that specialize in information management. Thus, a financial ontology for fraud purposes could play a central role between information management and business intelligence analysis components.

The ontology simulates the unregulated market in a machine-processable form, providing an information management layer in the framework of a semantically rich knowledge base that can be used for the semi-automatic interpretation of relevant unstructured resources. Based on the semantics of ontologies, information and patterns can be extracted from natural language texts and, after processing, knowledge can be extracted that will help the business intelligence (BI) analysis components to trigger the alarms.

In order to achieve this aim, the thesis examines the structure and trading operation of the unregulated market, the recent consultations on its structure and governance, as well as future usage scenarios and emerging technologies. The usefulness of the ontology has been evaluated through three case studies, which not only help to explain how manipulation in unregulated markets works, but also demonstrates how the ontology can be used as a framework for the extraction process and capturing information related to financial fraud, to improve the
performance of surveillance systems in fraud monitoring. More specifically, the objectives of the research are:

i. To examine the unregulated market structure, market tiers, trading operations and different market manipulation types associated with such markets.

ii. To build a financial ontology containing a large set of financial concepts and definitions to demonstrate the processes of financial fraud. The proposed financial ontology will act as a framework to guide the extraction process and capture financial fraud concepts from unstructured sources.

iii. To evaluate the financial ontology through specific instantiations which demonstrate through the three case studies how the ontology can help fraud detection. The ontology acts as a knowledge management repository system to tackle the information management problem, exploring the existing manipulation patterns of prosecuted cases in a coherent way.

iv. On the empirical side, to present examples of novel applications of text-mining tools and data-processing components, and to develop offline surveillance systems that are fully working intelligent systems. These systems will be built and tested using a unique sample of prosecution cases, reporting on their individual and combined characteristics and performance. In order to demonstrate its functionalities and capabilities, each prototype will be presented within an individual case study and contextualized using one or more scenarios using real data (unstructured and structured) from the OTC market and other relevant sources.

v. To identify gaps in financial research studies, as none of the previous works have considered the use of text-mining techniques for the analysis of "stock touting" in unregulated markets through adopting a “design approach”. In this context, text mining could help in raising alarms when traditional data-mining analyzers do not predict correctly or generate weak signals of suspicious trading behaviour. In particular, this could help to minimize ongoing information-based manipulation.

1.6 Research Questions

In order to achieve these objectives, the study answers the following research questions:
1. What are the gaps and limitations in existing market monitoring and surveillance systems in relation to unregulated markets?

2. What are the underlying ontological foundations for the body of knowledge in market monitoring and surveillance systems?

3. How to build a financial ontology for fraud purposes from a corpus? Which architecture will be used to address the integration challenges between ontology layers? What methodology will best analyze the domain ontology? Which ontology environment the ontology will be deployed?

4. Is it possible to build financial fraud ontology as a framework for the extraction process and to capture information related to financial fraud? How will the extracted information be presented in an organized and coherent way, to be considered as a knowledge base for different users?

5. How will the proposed ontology be instantiated and evaluated under different case study scenarios? How to design intelligent systems to extract information from unstructured sources based on the financial ontology?

6. How the ontology and its instantiations will help in raising alarms or open investigation?

1.7 Structure of the thesis

The next chapter presents the literature review and discusses existing market monitoring and surveillance systems. It reviews relevant financial topics, financial ontologies, and provides background information on text-mining and data mining tools used in the detection of financial fraud.

Chapter 3 presents the research design, discussing the chosen 'Design Science and Information Systems' methodology and the 'Information Systems Design Research Framework' that support this work.

Chapter 4 presents the financial ontology for fraud purposes, describe its ontology layering architecture, the process of constructing the domain ontology from a corpus, and including one case study to demonstrate empirically how text mining can be integrated with the financial fraud ontology to extract financial concepts from the unstructured sources and demonstrate the published prosecuted cases in an appropriate knowledge base.
Chapter 5 describes the case study of a stock spam e-mail as an unstructured data source, demonstrating another instantiation of the proposed financial fraud ontology as a text-mining application.

Chapter 6 demonstrates the business intelligence (BI) service system for a financial market monitoring and surveillance system in which different components interact in a coordinated way to produce proactive alarms for possible securities fraud cases.

Chapter 7 discusses the role of the financial fraud ontology in the fraud detection context and presents a summary of the three instantiations.

Chapter 8 summarizes this thesis and its contributions, its limitations and future work, and discusses how the thesis has answered the research questions.
2 Literature Review

By examining relevant literature from the financial domain, this chapter reviews the background and existing knowledge of financial market monitoring and surveillance systems. It begins with an overview of the current structure and trading processes of an unregulated market (OTC and Pink sheets) in order to highlight the key differences between the unregulated market and the regulated markets. Next, a review of market manipulations in the financial context is presented, along with description of the different types of manipulation. More specifically, the researcher focuses on information based manipulation and reviews the effect of different textual sources used by manipulators to hype investors and pump the prices. It then gives a brief description of current market monitoring systems and the technologies involved in such systems. Finally, the chapter reviews existing attempts at financial ontology and concludes with a summary of limitations and research gaps in existing market monitoring and surveillance systems.

2.1 OTC and Pink Sheets Market Structure and Trading Process Background

This section gives an overview of the current structure and trading processes of an unregulated market (OTC and Pink sheets) in order to highlight the key differences between the unregulated market and the regulated markets. In 1904, the Pink sheets market was established as a quotation service for market makers in the OTC market. The OTC market offers issuers, market makers and investors an exclusive set of rules and regulations, unlike the regulated markets.

In particular, OTC market provides an alternative means of listing securities on a stock exchange for issuers that either choose not to list their company in regulated stock exchanges or whose company failed to meet the relevant listing requirements. Thus, market participants do not need to register with regulators such as SEC, nor it is necessary for them to be “current” in their reporting requirements (Bollen and Christie 2009).

In 1999, the market provided participants with an electronic quotation service to cope with demand for real time quotations over the internet. At that point, market makers could still negotiate and execute orders via telephone conversations. Then in 2003, the OTC Markets Group introduced an electronic platform (OTC Link)
through which market makers could negotiate and execute their orders. Currently, investors can trade almost 10,000 equity and debt securities through the broker of their choice. (OTCmarkets 2011).

The structure of the OTC Market is very different from that of regulated markets, particularly in its trading process, as shown in figure 2-1. OTC Market activities are operated by OTC Markets Group Inc. which organises the OTC marketplace into three tiers based on the level of disclosure, quality, risk, and information that companies choose to provide to investors: OTCQX (premier tier), OTCQB and OTC Pink. The OTC market is regulated by the Industry Regulatory Authority (FINRA), SEC and various state securities regulators (OTCmarkets 2011). However, OTC Markets Group Inc. is neither a stock exchange nor a self-regulatory organisation (SRO) (U.S. Securities and Exchange Commission 2011).

Given that, there is a group of OTC securities which are registered in regulated markets such as the National Association of Securities Dealers Automated Quotations (NASDAQ), and these stocks are obliged to abide by the rules, regulation, and registration and financial reporting requirements of the relevant regulators (Bollen and Christie 2009).

The market has main five participants, namely; companies, investors, broker-dealers, regulators, inter-dealer quotation/trading systems represented by the OTC Link platform and FINRA’s OTC Bulletin Board. Each participant performs a different role; thus, the researcher shows how these participants work together to gain an understanding of how the market operates and exemplify its complexity (OTCmarkets 2011).

2.1.1 Market Tiers

As shown in figure 2-1, companies can issue and sell securities in the market to raise capital, complete an acquisition and allow selling shareholders to withdraw from investments if necessary. Companies choose to join different tiers in the market based on the level of disclosure they are willing to provide. OTC securities are allocated to a market tier based on the reporting scheme and level of disclosure, which is divided into three categories: namely Current, Limited or No Information. Regarding the current information category, companies have to submit filings publically, either to regulators or OTC Disclosure and News Services.
For limited information disclosure, companies are allowed to either submit financial reports from the last six months through the OTC Disclosure and News Services, or use the SEC’s EDGAR system to provide information such as registration statements, periodic financial reports, and other forms. However, the no information category is for companies that will not commit to providing any type of disclosure to any relevant counterparties (regulators, OTC Market, the public) (OTCmarkets 2011).
Within the OTC market group Inc. there are three main tiers, namely, OTCQX, OTCQB, and OTC Pink. OTCQX Market Place is the premium tier of the OTC market, representing a trillion dollars of market capitalisation. Currently, 4,268 securities listed in this market use a quality controlled platform, submit and meet the highest financial standards, and go through a qualitative review to offer transparent information through the market makers. Furthermore, these securities are required to have Current disclosure in addition to the QX listing requirements. Theoretically, companies can achieve a greater level of liquidity, rise in capital, price development and efficient trading mechanism by providing a high level of transparency (OTCmarkets 2011).

This market tier provides investors with two kinds of service: global, through a platform called OTCQX international, and local, through OTCQX US. OTCQX international offers multinational companies a visible cross listing in the US without the duplicative regulatory encumbrances of a normal US exchange listing. The local service offers domestic companies a quality controlled listing and shareholder services on the OTC market at significant cost savings compared to traditional markets. These services increase the flow of information, improve price discovery, raise the profile of, and increase trading and liquidity in, OTC market companies such as Roche (RHHBY), Adidas (ADDYY), Air France-KLM (AFLYY), and Deutsche Telekom (DTEGY) which all have securities on the OTCQX premium market tier (OTCmarkets 2011).

OTCQB market place is the middle tier of the OTC market which helps investors to identify companies that are up-to-date in their reporting obligations with the SEC or report to a U.S. banking or insurance regulator. The OTCQB market was established in 2010 due to the increase in U.S. reporting companies. The market was designed to help investors identify OTC traded companies that report to the SEC, Federal Deposit Insurance Corporation (FDIC) or the U.S banking regulator, and must be current in their disclosure. Currently, 3,430 securities listed in this market. These securities do not qualify for OTCQX as there are no financial or qualitative standards for this tier. OTCQB securities are quoted on the OTC link quotation and trading system, and a very small percentage are also quoted on OTCBB (OTCmarkets 2011).

The bottom tier, which attracts large numbers of late or non-SEC reporting companies, is called the OTC Pink market place. Currently, this market lists 6,080
companies that are in financial distress, in early the stages of business development or sometimes decisively dark to US investors. Thus, the market is risky and speculative as it has no minimum financial standards or reporting requirements, and features companies without audited financials (Bollen and Christie 2009).

Moreover, companies registered in this market choose the level of information disclosure (current, limited, no information) they wish to offer their investors. Those companies which choose to register with *OTC Pink Current information* follow international reporting standards or submit filings through OTC Disclosure. However, this level of disclosure does not include quality or investment risk.

*OTC Pink limited information* is aimed at companies which have financial reporting problems, are in economic distress, could announce bankruptcy, or are reluctant to meet OTC markets’ reporting standards. Generally, this category is for companies which offer inadequate financial information from the last six months on the OTC disclosure and news services or did not submit the SEC’s EDGAR form in the previous six months.

The last category of this tier is *OTC Pink no information*, which includes companies which do not intend to provide disclosure to the public markets, regulator or exchange market. Generally, this category does not publish any information or news via OTC Market news service and in the rare instances that they do provide any information; it is generally over six months old. Thus, this kind of securities is highly risky for investors because companies do not provide sufficient or accurate information. Furthermore, investors face extremely illiquid markets and any change in demand, either real or artificial, could have a significant impact on stock prices and volume, and therefore, on investor positions.

OTC Securities that are not listed or traded or quoted on any U.S stock exchange or on the OTC Market may be traded in the Grey market, where no bid and ask information is available. Grey market securities are reported to SRO by brokers and the SRO distributes the trade data to market data and financial websites to enable investors to follow security prices and volumes. Due to the lack of pre-trade data where bids and offers are neither available nor collected in the central market, the market is completely non-transparent, relatively illiquid market and any change on demand, either real or artificial, investors may encounter difficulty in their selling position (OTCmarkets 2011). The SEC reported
23,419 manipulation cases have been prosecuted from the OTC market, 2,793 of these cases occurred in the OTC pink market place (Pink sheets) (U.S. Securities and Exchange Commission 2011).

2.1.2 The Trading Process

Investors in the OTC market vary according to their knowledge, behaviour, and experience; however, they share a common objective of generating profits from their investments. They have to decide which of the two types of order they want their brokers to execute: limited orders, which allow investors to specify the exact price they are willing to buy or sell at, and market orders, which allow brokers to execute orders immediately to either buy or sell at the current market price (best bid or offer). All OTC securities transactions, whether buy or sell orders, are executed through FINRA-registered broker-dealers (OTCmarkets 2011).

The role of the broker is to find ‘match’ orders, which means they execute orders internally when an investor is willing to sell a particular security for the same price another investor is willing to buy the same security. In some cases brokers cannot, or choose not to, trade internally and instead trade marketable orders externally. An order is described as ‘marketable’ when the price specified can immediately be executed in the market. Furthermore, limited orders are marketable if the limit price is better than or equal to the bid price (for sell orders) or ask price (for buy orders). Thus, brokers may buy or sell their own (principle) account at their own risk, and have to ensure that orders are matched, even if this means finding another broker willing to trade that security. In cases where the order is not marketable, the broker can create or edit its existing quote on the OTC Link quotation system to reproduce a new price or size.

Other brokers in the market are acquainted with the price they are willing to buy or sell at. Brokers do not normally show the entire order to other brokers, because other brokers could lower their prices or cancel the order execution as a result of this information. Once brokers have updated their quote, they may monitor the market until the limit price is satisfied, or they may send an order to another broker as an external trading process for non-marketable orders. In addition, they may receive an order against their standing quote for a different price and size. For example, another broker may have a marketable order. Therefore, the broker may
accept, reject or counter (send a different price or size) the order (OTCmarkets 2011).

One of the main differences between the OTC market and other equity markets such as the New York stock exchange (NYSE) is that OTC has no central exchange to match/execute orders. All trades are by negotiation and agreed directly between the brokers. Thus, brokers are liable for their quote price and volume; otherwise they will be penalized by FINRA. Brokers have to communicate and trade directly with each other by posting their quotes either on the OTC Link or FINRA’s OTC Bulletin Board (OTCBB) quotation services. Traditionally, investors relied on FINRA’s OTC Bulletin Board (OTCBB) quotation service that shows real time quotes for equities, last sales prices, and volumes.

Market makers had to use the telephone to communicate and make any trades in OTCBB. Furthermore, only SEC-reporting or bank/insurance-reporting companies are eligible for quotation on this platform. However, brokers can now communicate with each other using the OTC Link electronic messaging and trade negotiation system. Applications such as OTC Markets Group, OTC Dealer or OTC Fix allow brokers to access the OTC Link and enable them to view and trade the targeted OTC securities quotes. Recently, 95% of OTC securities quotes have become available in OTC link. Once brokers agree an order, they are required to report their trade to FINRA within 90 seconds of the order’s execution. FINRA then disseminate this information to the market (OTCmarkets 2011).

In summary, this section examined the current structure and trading processes of an unregulated market (OTC and Pink sheets). The discussion shows that the OTC market is operated differently from that of the regulated markets. The market provides an alternative means of listing securities on a stock exchange for issuers. One of the key differences is the market tiers which classified into three tiers based on the level of disclosure, quality, risk and information that offered to investors: OTCQX (premier tier), OTCQB and OTC Pink.

Market participants is another difference, the market has main five participants, namely; companies, investors, broker-dealers, regulators, inter-dealer quotation/trading systems represented by the OTC Link platform and FINRA’s OTC Bulletin Board. The OTC market has no central exchange to match/execute orders. All trades are by negotiation and agreed directly between the brokers. Despite the fact that market offers different level of disclosure in each tier, still the
market is risky and speculative as it has no minimum financial standards or reporting requirements, and features companies without audited financials. Thus, the market suffers from many contagion manipulation activities arising from the interconnectedness of market participants and the limited transparency between different counterparty relationships.

In particular, 23,419 manipulation cases have been prosecuted by the SEC. The absence of data disclosure requirements mean that that OTC markets are less directly monitored than is the regulated market. Furthermore, regulators have limited ability to detect inappropriate market behaviour and conduct in the market. Therefore, this could become a threat to the market’s integrity and efficiency. Thus, the continuous improvement and development of financial market surveillance systems with high analytical capabilities to capture the fraud is essential to guarantee and preserve an efficient market.

2.2 Review of financial economics literature

The impetus to effectively and systematically address stock market efficiency, including factors such as stock price manipulation, has long presented a very dynamic challenge to academia, the industry and relevant authorities. In this context, (Fama 1970) introduced the efficient market hypothesis (EMH), which states that stock prices correlate with known market information, and the market is information oriented.

Consequently, normal return on investment could be obtained from the stock market and the market should react instantly to any information release. This demonstrates that new information plays an important role in changing stock prices and volumes. However, this theory has been criticised by financial rationalists. For example, if the EMH is correct, we can assume that closing prices should reflect all information available during trading. However, this does not appear to hold be reflected in the empirical evidence. According to (Harris 1989) work, transaction prices rise at closing hours even if there was no new information available at that moment.

Many authors have tested the impact of non-competitive behaviour in the stock market, and verified the possibilities of market manipulation. In particular, Vila (1989) modelled market manipulation using the game theory concept and presented two cases with asymmetric information forms. The technique helped to
simplify the manipulation scenarios and demonstrated the role of private information in the manipulation (Vila 1989). Similarly, (Benabou and Laroque 1992) developed a specialised model that proved the notion that numerous types of agents who possess private information can manipulate prices through misleading advertisements. Kumar et al. (1992) introduced an asymmetric information model to investigate the susceptibility of future markets to price manipulation, and verified that manipulators can profit from their strategy.

In one of the most influential works, Allen and Gale (1992) introduced three categories of market abuse which were described as follows: Action-based manipulation, i.e. actions that change the actual or perceived value of the assets; information-based manipulation, involving the release of false information or rumours to tempt investors and mislead their trading decisions; and finally, trade-based manipulation, which occurs when traders buy and sell stocks to mislead the markets by artificially moving the price to a level that suits their purposes.

A subsequent study by Van Bommel (2003) argued that traders with access to valuable information are the source of rumours, and that they try to manipulate prices to increase their profits by providing investors with vague or inaccurate trading advice. The paper demonstrated how followers and rumourmongers use different information-based ‘pump and dump’ manipulation strategies to maximise their profits. These strategies vary from sending simple buy/sell advice to investors, or ‘bluffing’ by spreading rumours despite not having any information, or cheating by spreading false information. The paper showed that rumourmongers can benefit from a manipulation twice; firstly, when they send a rumour and investors act upon it, and secondly, when rumours have been overshot and manipulators can trade in the opposite direction.

Eren and Ozsoylev (2006) introduced a theoretical indicator into a standard market microstructure model, considering the discreet time and multi-period model to explore the different economic conditions under which hype and dump manipulation could result in an equilibrium outcome. The paper argued that an informed trader could be dishonest and decide to disseminate rumours about specific securities to influence the investment decisions of uninformed traders. The ‘uninformed’ traders could include sophisticated traders with possible relationship between the rumour and the touted securities. The presence of sophisticated traders discourages rumour dissemination campaigns by informed traders. In
contrast, naïve traders help informed traders to execute their strategy because they believe the rumours are true and act upon them. The main findings of the paper are that hype and dump manipulation (informed traders buy when prices are low and then sell at an inflated price) is attained in equilibrium when naïve traders in the market react and the cost of the rumour campaign is relatively high.

Later, Aggarwal and Wu (2006) extended Allen and Gale (1992) findings by presenting a theoretical framework for profitable market manipulations, and provided empirical evidence using a comprehensive dataset of manipulation cases which occurred in the US stock markets and were published in SEC litigation releases from 1990 to 2001. Their research found that informed parties such as corporate insiders, brokers, underwriters, large shareholders and market makers were likely to be manipulators. Furthermore, they showed that stock prices go up throughout the period of manipulation and then fall once the period has passed. Prices and liquidity are higher when the manipulator sells, then go down when the manipulator buys. In addition, throughout the duration of sells, prices are higher when liquidity and volatility are greater. As a result, stock market manipulation may have a significant impact on market efficiency. The analysis showed that most manipulation cases occur in unregulated and emerging financial markets, such as the OTC and Pink sheets, which are relatively small and illiquid, have lower disclosure requirements and are less bound by regulators’ rules and enforcements.

Interestingly, since 1995 more than 7,300 firms have delisted from the regulated market (Macey, O'Hara et al. 2008). These delisted firms continue to trade, sometimes actively, in the unregulated market. For example, companies could be delisting because they violate one or several of regulated market’s listing requirements, but continue trading in the OTC or pink sheets market.

As very few researchers have looked at OTC trading and Pink Sheets, the literature here is relatively limited. Studies such as Angel et al. (2004) and Macey et al. (2008) evaluated trading costs of firms delisted to the OTC or pink sheets markets. Both papers suggested that stock delisting processes bring with a decrease in liquidity, rise in trading costs, share prices falling to approximately half their previous value and increased volatility. These high volumes suggest that unregulated markets create valuable trading opportunities by delisting securities from regulated markets. Furthermore, the studies argued that a large proportion of
the delisting decisions taken by NYSE or NASDAQ were applied inconsistently, with arbitrary implementation and sometimes appeared aggressive. Some of these companies left willingly for reasons such as a merger. However, most of them were forced to leave by regulated exchanges on which they listed. As a result, many delisted companies choose to move to unregulated markets which are illiquid and impose less listing requirements (Macey, O'Hara et al. 2008).

Other works, such as Bollen and Christie (2009), have studied the microstructure of unregulated markets and assessed the ability of existing theories to capture their outstanding features. The authors used the microstructure theory to understand and demonstrate quoting and trading behaviour on the pink sheets market. The tremendous increase in stock spam and the involvement of most stocks in ‘pump and dump’ manipulation schemes motivated the authors to test how such unregulated and unmonitored markets can function effectively despite these manipulation attempts, using microstructure theory.

In summary, this economic literature review presents an overview of how firms have delisted from the regulated market and continue to trade in the unregulated market. Previous studies show that many delisted companies choose to move to unregulated markets because they either forced to leave the regulated market or left the market willingly. Furthermore, the aforementioned literatures show how the problem of market manipulation has been historically defined and investigated by the financial and economic community, and has highlighted that most manipulation cases occurred in unregulated and emerging financial markets. In particular, how followers and rumourmongers use different information-based ‘pump and dump’ manipulation strategies such as stock spam e-mails to generate profits.

### 2.3 Market Manipulation: Importance, Definitions, and Types

The aforementioned literature demonstrated that stock price manipulation is one of the most widely discussed topics in the stock market. Furthermore, manipulators can artificially increase securities prices and make profits using various strategies, from classic manipulative trading practices that influence prices, to the sophistication of spam and scam manipulation using various internet channels. Unfortunately, such schemes prevail, and this affects the market’s integrity and efficiency (Allen and Gale 1992). In October 1998, the SEC announced that 44
companies and individuals had manipulated securities using internet channels and therefore violated the market (Baker 1999). Despite the long history of stock market manipulation, the definition of such a concept still needs further clarification. The definition and scope of “manipulation” or “manipulative” becomes a controversial subject across different disciplines. In particular, finance and economics research still uses the term ‘manipulation’ imprecisely (Kyle and Viswanathan 2008). Though some legislative institutions have sought to create various manipulation provisions, the definition of manipulative practices has generally been left to the courts on a case-by-case basis (Putnins 2009). Since the passing of the Securities Act in 1930, there is evidence that market manipulation has a significant impact on the efficiency of the securities market (Rogoff 1976).

Thus, in order to suppress manipulation on the stock markets, US congress drafted the Securities Exchange Act of 1934 (“Exchange Act”). The Act was passed by US Congress as a crucial corrective measure through which to regain public confidence in the securities market in light of the abuses which contributed to the Great Crash in 1929 (Rogoff 1976). The Exchange Act established broad provisions, such as: it is unlawful “to use or employ, in connection with the purchase or sale of any security … any manipulative or deceptive device or contrivance” (Putnins 2009). Another provision was directed at brokers and dealers operating in the market: it is unlawful “to effect any transaction in, or to induce the purchase or sale of, any security…. otherwise than on a national securities exchange, by means of any manipulative, deceptive, or other fraudulent device or contrivance”. Other provisions in the Exchange Act prohibit the use of fraudulent activities related to the sale of securities (Rogoff 1976; U.S. Securities and Exchange Commission 2010).

Furthermore, the European Union commission drafted a statutory market abuse law called “Market Abuse Directive 2003”, which reads as follows: “market manipulation shall mean transactions or orders to trade which give, or are likely to give, false or misleading signals as to the supply of, demand for, or price of financial instruments, or which secure … the price of one or several financial instruments at an abnormal or artificial level” (Commission 2005; Putnins 2009). Manipulative practices in the regulated and unregulated markets are under the purview of federal securities laws and regulations. The SEC imposes these rules,
and regulations, and brings sanctions upon those who try to induce artificial price increases or any other manipulative activities. The rules were designed to protect investors from market violations and to maintain competitiveness in the financial markets (Rogoff 1976).

Despite the existence of these authoritarian regulations, William S. Laufer\(^1\) noted that fraud cases are seldom prosecuted or tried by the courts as criminal cases. Therefore, prosecutors find it challenging to prepare a case with appropriate evidence, and the gain for an exchange of a successful prosecution would be small in comparison to the efforts and resources necessary to bring a criminal case. Moreover, the SEC has a very limited sphere of authority and investigative operations; thus, it can only bring a civil case and then escalate it to the FBI to appoint attorneys and deal with it as a criminal case. From a different perspective, many fraudsters can afford to hire experienced lawyers who are skilled in finding holes in the cases. At worst, a manipulator may be fined or gets a black mark on his trading record and temporary suspension from trading in the exchange. Securities cases can be extremely complex and difficult to demonstrate to juries. That is why regulators only select certain cases for prosecution and prioritise instances of organised manipulation. Furthermore, few courts have experience in trying securities fraud cases. The US Attorney in New York City is the only one that has a specialised securities fraud team to deal with such cases (knowledge@wharton 2000).

Therefore, there is an urgent need to build a financial ontology to combat fraud by efficiently managing vast quantities of financial data, providing better communication and knowledge sharing among analysts, providing a mechanism to demonstrate knowledge of the processes of financial fraud, understanding and sharing financial fraud logic operations, managing relevant facts gathered for case investigations, providing early detection techniques of fraudulent activities, developing prevention practices, and allowing reuse of these knowledge resources in different financial contexts.

Some concerted efforts to define the term “manipulation” precisely (Putnins 2009) can be found in finance and economics literature. For example, (Fischel and Ross 1991) argued that there was no objective definition of manipulation, and that

\(^1\) William S. Laufer is Professor of Legal Studies and Business Ethics, Sociology, and Criminology at Wharton School, University of Pennsylvania.
existing definitions were meaningless and inaccurate. From their point of view, manipulation could be defined as a trader’s dishonest intention to move stock prices in a specific direction. They also analysed the relationship between manipulation and fraud, for example, asserting that spreading false information should be classified as fraud (Putnins 2009). Some years earlier, Jarrow (1992) investigated the power of large traders to manipulate prices to generate profits at zero risk. However, Kyle and Viswanathan (2008) work proposed that trading strategy should not be considered “illegal price manipulation” unless manipulators intend to use manipulation strategy to weaken the economic efficiency by way of price accuracy (prices are less accurate) or market liquidity (less liquid market).

Despite the fact that slight efforts have been devoted to build financial ontologies for fraud purposes as a way of representing and capturing financial knowledge and demonstrate the relationships between these concepts. The ontology could benefit from the great deal of information provided by the financial community. Ontology could be constructed to acquire, share, organize, reuse, and maintain this knowledge. In particular, the market manipulation classification presented by (Cumming and Johan 2008; Putnins 2009; Diaz 2011) will be used as part of the domain ontology layer in the financial ontology proposed in this thesis.

2.3.1 Market Manipulation Types

In order to provide an overview of the different types of manipulation strategies, the thesis adapts the simple taxonomy (Figure 2-2) introduced by Putnins (2009) and Diaz (2011) to be part of the concepts will be demonstrated in the proposed ontology. According to Putnins (2009), manipulations could be broadly classified into ‘runs and raids’, ‘contact-based manipulations’ and ‘abuse of market power techniques’. Within these main areas, further sub-categorisations could follow those introduced by Allen et al. (1992), namely, ‘information-based’, ‘trade-based’, and ‘breach of fiduciary duty’, which replaces ‘action-based’ (Diaz 2011).

In the runs and raids category, manipulators can inflate and deflate securities prices by either buying or short selling stocks while attracting liquidity to the stock. In such schemes, manipulators try to hold their position at the inflated or deflated price to make a profit from naive investors. The most common types of ‘run’ manipulation are ‘pump and dump’ schemes which include various strategies such as wash sales, pooling, and rumour dissemination.
In this type of scheme, manipulators try to promote the stock by disseminating an infallible combination of stock market data and misleading statements about the company, in order to attract investors. Once the price has been pumped up, manipulators seek profit by selling stock and dumping the prices into the market (U.S. Securities and Exchange Commission 2011). ‘Bear raids’ are considered to be one form of ‘run’ manipulation in which manipulators short sell stocks and influence the price by motivating others to sell, while protecting their own positions at a lower price. Profits are gained from the manipulated market when investors buy at inflated prices or sell at deflated prices (Putnins 2009). ‘Runs and raids’ manipulation is extended to information-based manipulations, for example, in a ‘boiler room’ or ‘cold calling’ situation, a group of dishonest brokers hire
salespeople to make telephone calls to sell securities to potential investors. This could generate the same effect and violate stock prices in the market (U.S. Securities and Exchange Commission 2011).

In contrast, manipulators use contract-based manipulations strategy to make profits externally. For instance, ‘marking–the-close’ or ‘portfolio pumping’ manipulation involves fund managers who try to place large numbers of orders to influence the value of the fund and drive securities prices up, in order to charge their customers extra commission. Such schemes are highly destructive for investors and the market because the gain is temporary and the stocks will fall in value once the scheme has ended (Investopedia 2011).

‘Abuse of market power’ manipulations are similar to ‘run’ manipulations, though the former take place in commodity and goods markets. A classic example of ‘squeeze’ would be manipulators controlling or changing the supply and demand for a specific security or commodity to generate profit. Despite the importance of trading cross-market or cross-produce, manipulators can abuse the system and make a profit by manipulating the price of underlying assets and holding the whole inventory of the commodity. For example, in ‘corners’ manipulations, fraudsters could control derivative and commodity markets by placing orders which drive the commodity price up, and once it has taken effect and influenced investors to buy, fraudsters cancel their orders (Diaz 2011).

The fourth category, ‘breach of fiduciary duty’, refers to abuses in which market participants do not comply with the contractual terms and regulations of an exchange. For example, using the ‘front running orders’ manipulation strategy, brokers could use information about orders placed by investors to their own benefit. Thus, brokers could abuse the market by placing buying orders just before investors’ orders, influencing the securities prices (Diaz 2011).

The discussion set out above provides an overview of the wide variety of practices and strategies which can be used to manipulate the stock market. It would be challenging to draw up an exhaustive list of manipulative behaviours. For example, some are considered to be legal practices in some formal sources rather than manipulative practices. For example, the concept of *scalping* is a trading strategy that aims to make a profit from small changes in prices. Manipulators try to act like specialists - usually bloggers, TV presenters or newsletter reporters - who recommend investors to buy at a specific bid price and sell at a specific ask price.
to gain the bid/ask difference. Scalpers normally gain credibility and the trust of investors, thus investors act upon their recommendations and violate prices (Diaz 2011).

Given that our research interest lies solely in demonstrating the different scenarios of market manipulation in unregulated market, Table 2-1 summarises the most common market manipulations strategies in unregulated market discussed in the existing literature and the SEC.

Table 2-1 Market Manipulations, Categories, Definitions and Types (adapted from Cumming et al. 2008; Diaz 2011; FINRA 2010; Investopedia 2011a; U.S. Securities and Exchange Commission 2011c)

<table>
<thead>
<tr>
<th>Manipulation Known as</th>
<th>Category</th>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Abnormal prices of related products at delivery locations</td>
<td>Abuse of market power</td>
<td>Trade-based</td>
<td>The expiring futures price and the spot price at the delivery market are abnormally high relative to prices at other, non-deliverable locations; the prices of related products; and prices of non-deliverable grades of the same commodity.</td>
</tr>
<tr>
<td>2. Advancing [decreasing] the bid [ask]</td>
<td>Runs and raids</td>
<td>Trade-based</td>
<td>Influence the security or derivatives either by increasing or decreasing the bid or ask.</td>
</tr>
<tr>
<td>3. Boiler room pumping</td>
<td>Runs and raids</td>
<td>Information-based</td>
<td>Brokers use boiler room tactics to motivate customers about specific stocks by hiring high-pressure salespeople to give positive information to investors and using words like “it’s a sure thing” or “opportunities like this happen once in a lifetime” to pump the prices and violate the market.</td>
</tr>
<tr>
<td>4. Churning Churn and burn, twisting, overtrading</td>
<td>Breach of fiduciary duty</td>
<td>Trade-based</td>
<td>Brokers violate market by excessively trading in customer’s account in order increase their commissions.</td>
</tr>
<tr>
<td>5. Dissemination Touting, spamming, e-mail blasting, Internet fraud, hype and dump manipulation, pump and dump schemes</td>
<td>Runs and raids</td>
<td>Information-based</td>
<td>Dissemination of false or misleading market information such as spam e-mails, Internet bulletin boards (forums or blogs), newsletters, questionable press releases, and chat rooms</td>
</tr>
<tr>
<td>6. Front running client orders Intervention</td>
<td>Breach of fiduciary duty</td>
<td>Trade-based</td>
<td>A transaction to the detriment of the order giver on the basis of, and ahead of, an order which he is to carry out for another.</td>
</tr>
<tr>
<td>7. Front running research</td>
<td>Breach of fiduciary duty</td>
<td>Information-based</td>
<td>Misuse of price or volume confidential and sensitive information as a result of research activities.</td>
</tr>
<tr>
<td>8. Hype and dump Pump and dump; pumping, touting</td>
<td>Runs and raids</td>
<td>Information-based</td>
<td>A scheme that attempts to boost the price of a stock through recommendations based on false, misleading or greatly exaggerated statements. The perpetrators of this scheme, who already have an established position in the company's stock, sell their positions after the hype has led to a higher share price and make profits. Pump and dump schemes often occur on the Internet where it is common to see messages posted that urge readers to buy a stock quickly or to sell before the price goes down.</td>
</tr>
<tr>
<td>No.</td>
<td>Description</td>
<td>Type</td>
<td>Basis</td>
</tr>
<tr>
<td>-----</td>
<td>----------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>9.</td>
<td>Insider trading</td>
<td>Pump and dump; pumping, touting</td>
<td>Breach of fiduciary duty Information-based</td>
</tr>
<tr>
<td>10.</td>
<td>Marking the close</td>
<td>&quot;Painting the tape&quot;; &quot;stabilisation&quot;, &quot;Portfolio Pumping&quot;</td>
<td>Contract-based Trade-based</td>
</tr>
<tr>
<td>11.</td>
<td>Marking the open</td>
<td>&quot;Painting the tape&quot;; &quot;stabilisation&quot;, &quot;Portfolio Pumping&quot;</td>
<td>Contract-based Trade-based</td>
</tr>
<tr>
<td>12.</td>
<td>Matched orders</td>
<td>&quot;Pools&quot;, &quot;collusion&quot;</td>
<td>Runs and raids Trade-based</td>
</tr>
<tr>
<td>13.</td>
<td>&quot;Off-shore&quot;</td>
<td></td>
<td>Runs and raids Information-based</td>
</tr>
<tr>
<td>14.</td>
<td>Pre-arranged trade</td>
<td>&quot;Pools&quot;; &quot;collusion&quot;</td>
<td>Runs and raids Trade-based</td>
</tr>
<tr>
<td>15.</td>
<td>Pyramid schemes</td>
<td></td>
<td>Runs and raids Information-based</td>
</tr>
<tr>
<td>16.</td>
<td>&quot;Risk-free&quot; or &quot;guaranteed&quot;</td>
<td></td>
<td>Runs and raids Information-based</td>
</tr>
<tr>
<td>17.</td>
<td>Scalping</td>
<td></td>
<td>Runs and raids Information-based Trade-Based Manipulation</td>
</tr>
<tr>
<td>18.</td>
<td>Short sales</td>
<td>Bear raid; free riding</td>
<td>Runs and raids Trade-based</td>
</tr>
<tr>
<td>19.</td>
<td>Slur and dump</td>
<td>Pump and dump; pumping, touting</td>
<td>Runs and raids Information-based</td>
</tr>
<tr>
<td>20.</td>
<td>Spoofing/painting the tape</td>
<td></td>
<td>Runs and raids Trade-based</td>
</tr>
<tr>
<td>21. Squeeze</td>
<td>Short Squeeze</td>
<td>Abuse of market power</td>
<td>Trade-based</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------</td>
<td>-----------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>22. Stub quoting</td>
<td>Breach of fiduciary duty</td>
<td>Trade-based</td>
<td>According to SEC, a 'stub quote' is an offer to buy or sell a stock at a price so far away from the prevailing market that it is not intended to be executed, such as an order to buy at a penny or an offer to sell at $100,000. A market maker may enter stub quotes to nominally comply with its obligation to maintain a two-sided quotation at those times when it does not wish to actively provide liquidity. A 'stub quote' differs from 'stuff quoting' because the former are offers made by an official market maker, while the latter are made by HFTs that are not obliged to do so.</td>
</tr>
<tr>
<td>23. Trade through</td>
<td>Breach of fiduciary duty</td>
<td>Trade-based</td>
<td>The completion of a client’s order at a price inferior to the best posted bid or ask. This is not per se considered manipulative, but many commentators (and the surveillance authorities themselves) consider it manipulative because the market maker who receives the order is unable or unwilling to fill it at the best posted bid or ask price, and hence the trade is instead executed at the market maker's price.</td>
</tr>
<tr>
<td>24. Trading away</td>
<td>Breach of fiduciary duty</td>
<td>Trade-based</td>
<td>Refers to brokers trading for their personal account through a brokerage other than their employer. It can also involve private placement transactions structured between a broker and a client without the knowledge of the employer.</td>
</tr>
<tr>
<td>25. Wash sale</td>
<td>'Pools'</td>
<td>Runs and raids</td>
<td>Trade-based</td>
</tr>
</tbody>
</table>

2.3.2 Information-Based Manipulation

This section describes how information-based manipulation has been used extensively and creatively to target specific stocks as part of the classic "pump and dump" schemes which enable fraudsters to generate profits in the unregulated market. Most commonly, manipulators state that they have ascertained private information about stocks, such as investment advice, and encourage a specific investment decision which they disclose along with financial terms and recent price quotes. Thus, stock promoters speculate on positive price models of the traded stocks, disseminating messages to potential investors to drive the price of the touted stock upwards or downwards (Zaki, Theodoulidis et al. 2011). Furthermore, the internet helps fraudsters to communicate easily, cheaply and reach investors en masse. It is difficult for investors to differentiate between legitimate claims and false ones because the information enclosed in messages is fine printed and claims to present valuable information (Baker 1999).
The most common uses of the internet in securities fraud include: spam e-mails, internet fraud (including internet bulletin board messages, online investment newsletters and chat rooms), paid promoters, boiler rooms, and questionable press releases (U.S. Securities and Exchange Commission 2011).

2.3.2.1 Spam E-mail Messages

The existing literature (Table 2-2) shows that touting campaigns are an effective way of manipulating market prices, and that there are agents who react to such information. Despite the availability of numerous anti-spam software programmes, the amount of spam e-mails is growing exponentially (Hu, McInish et al. 2010). Earlier studies have found that over 80% of all e-mail traffic is classified as spam, with 15% of these messages related to stock touts (Frieder and Zittrain 2007). In 2008, McAfee found that unprotected users received an average of 70 spam messages per day (BBC 2008). There are numerous examples to show the impact these messages have on price and trading volume and how destructive they have been for the financial markets. For example, a major spam eruption hit the internet in June 2007 which led to a 20% increase in securities trading volume. This is because the messages were designed professionally and 5 billion of them were disseminated that day. Many cases have been prosecuted by US authorities, and charges have been brought against promoters for “pump and dump” scheme manipulations (Bullguard 2007). Therefore, the SEC suspends trading of any company that has been heavily touted by spammers as a result of a pump and dump spam campaign.

Numerous studies suggest that “pump and dump” campaigns can yield positive abnormal returns and an increase in stock volatility. For example, Mei, Wu et al. (2004) present some empirical evidence of an equilibrium model to demonstrate how “smart money” can profit from other investors’ irrational behaviours. The model is based on a sample collected from cases which have been prosecuted by the SEC. The suggested model classified investors into behaviour-driven investors, arbitrageurs, and a manipulator. The manipulator’s strategic action, together with other investors’ behavioural biases can lead to higher returns, an increase in volatility and trading volumes, short-term price contribution and long term price reversal. Thus, manipulators have the power to manipulate the market through biased messages. Furthermore, investors could follow these messages hoping that the senders are insiders who have superior information they want to
share publically. In most cases, opportunistic individuals (corporate officers, financial journalists, or “gurus”) are less than honest and want to reveal information to mislead the market and reap large profits from those investors who acted upon their announcements (Roland and Laroque 1992). These rumours and misleading announcements facilitate market manipulation (Bommel 2003).

In particular, the effect of these rumours and touting messages is more evident in the unregulated market as they endogenously drive pricing of several securities up and affect the market’s integrity (Mei, Wu et al. 2004). Various studies provide evidence that touted spamming works and that there are recipients who read it and act upon it. For example, Böhme and Holz (2006) found evidence of the harmful effect of spam messages on the financial markets. The study employed a multiplicative multivariate regression model and a classical event study methodology, focusing on the effect of spam e-mails on the return and volume of the target stocks.

Frieder et al. (2007) evaluated and analysed the impact of touted spam on the trading activity of specific stocks using a broad sample from the Pink Sheets market. The paper compared the touted stock with another control sample (not touted) during the same period. Evidence was found of a significant positive return on days where heavy spam touting took place. Furthermore, the volume of trading was shown to correspond positively to heavy spam touting.

Hanke and Hauser (2008) began their analysis by describing the common characteristics of the advertised stock, such as price level and average turnover. Besides investigating the effect of touting spam on returns and volume, they also measured the effect of spam email on other variables such as excess returns, turnover and intra-day volatility. The research showed that stock touting had a significant positive impact on the securities prices. Furthermore, liquidity is one of the foremost factors in determining the success of spamming campaigns. Lastly, repetitive spam sent on consecutive days continued to increase demand for the targeted stock, which strengthens the spammer’s position and affords them more time for liquidation.

Nelson, Price et al. (2009) studied the 'attention effect' in the market response to spam messages. The research showed that returns and abnormal volume were highest when spam e-mails cited stocks with extreme target prices and copied press release information. Furthermore, the 'attention role' helped spammers to
select the target stocks. The study emphasised that a successful touted campaign has an influential impact on investor attention and affects the asset prices. Regression analyses findings showed messages with a target price trigger a significant market response when accompanied by information from an issued press release.

Bouraoui (2009) used the event study methodology to evaluate market efficiency by analysing the reactions of the market toward a stock spam event. In particular, the research assessed the impact of the stock spam on returns, considering the evolution of volatility during the period. As a result, stock spam e-mails affect the behaviour of investors who act upon the information cited in the messages. Moreover, there is evidence of positive abnormal returns during the first three days following the spam email’s dissemination. Given that spam emails work perfectly in the unregulated market, the results of the study affirm that information-based manipulation schemes are thriving and spammers generate profits from investors who still believe such information.

Hu, McInish et al. (2010) analysed the contents of promoted stock spam e-mails to identify the main attributes (target price, message length, e-mail source, incentives, and touting international business) that affect investor’s reactions and trading behaviour. The analysis shows that a short price target attribute makes a significant difference to market reactions in terms of abnormal returns and trading volume, compared to other factors. Interestingly, the study shows that spam e-mails touting US stocks had abnormal returns; however, touting other non-US stocks did not have this effect.

**Table 2-2 Review of Spam e-mails approaches**

<table>
<thead>
<tr>
<th>Paper</th>
<th>Objectives</th>
<th>Data source</th>
<th>Method and Models</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mei et al. (2004)</td>
<td>Shows how smart money can strategically take advantage of investors' behavioral biases and manipulate the price process to make profit</td>
<td>“pump-and-dump” cases prosecuted by SEC from January 1980 to December 2002.</td>
<td>Discrete-time economy Model</td>
<td>“pump-and-dump” Manipulation affect the market with higher return, increased volatility, larger trading volume, short-term price continuation and also long-term price reversal during the manipulation period</td>
</tr>
<tr>
<td>Böhme et al. (2006)</td>
<td>Examines the effects of spam e-mails on the return and volume of the target stocks</td>
<td>Collected from Richardson’s Stock Spam Effectiveness Monitor (SSEM) archive between November 2004 and February 2006</td>
<td>Multiplicative multivariate regression model and classical event study methodology</td>
<td>Stock prices react significantly and positively to spam e-mails</td>
</tr>
<tr>
<td>Frieder et al. (2007)</td>
<td>Assess the impact of spam that touts stocks upon the trading</td>
<td>Database of 1,802,016 spam messages from newsgroup</td>
<td>Pooled OLS Regressions with clustered standard</td>
<td>1. Significant positive return on days prior to heavy</td>
</tr>
</tbody>
</table>
activity
news.admin.net-abuse.sightings (NANAS)
errors
touting via spam
2. Trading Volume responds positively and significantly to heavy touting
3. Significant negative return in the days following touting

Hanke et al. (2008)
Investigate the effects of stock spam e-mails on excess returns, turnover, and intra-day price range
Collected spam e-mails from Crummy database website (year 2005)
Panel regression
Spam e-mails have a significant impact on all of these variables

Bouraoui (2009)
Examines the impact of stock spam messages on returns of firms od study.
A sample of 110 firms of penny stocks from February 2006 to June 2008
Methodology of event to compute abnormal returns, introduced a modelling GARCH
Stock spams affect the behaviour of financial participants who react favourably to the messages

Hu et al. (2010)
Investigates how the content of spam e-mails (price, message length, incentives, and source) affected the price and trading volume of the touted stocks.
40,000 spam messages collected from Crummy database from November 2004 to August 2007
Event study methodology to determine market reactions associated with each spam attribute
Abnormal return, turnover, and risk of the market were significantly higher for spam e-mails containing a target price

2.3.2.1 Internet message boards

The internet provides investors with information that they use to make investment decisions. Nowadays, investors can access information on different internet boards from paid sources, such as ValueLine and Zacks, as well as free sources, such as stock message boards and chat rooms. Message boards such as Yahoo Finance, Raging Bull and TheLion.com are popular free sources which help investors to learn and enable them to search for information, recommendations, and profitable investment ideas about specific stocks. Furthermore, investors achieve greater self-reliance to make investment decisions based on their research and trust in these online resources (Vilardo 2004).

Financial message boards on the internet can be divided into two main categories: namely, chat rooms and bulletin boards. Chat rooms are live forums where members can talk about “hot” stocks; however, they lack historical archives which would allow offline members to participate in the discussion. Bulletin boards provide a more organised mechanism to allow members to post or reply to messages at any time (Tumarkin and Whitelaw 2001). However, many of these internet message boards or chat rooms could contain fallible information and fraudsters normally hide their identities and post messages urging investors to buy stock to improve their positions and trading profit at the expense of other investors (U.S. Securities and Exchange Commission 2011). Thus, the SEC and Federal Trade Commission (FTC) are eager to follow activity on online message boards to
protect investors from these manipulation attempts (Sabherwal, Sarkar et al. 2011).

Previous studies report that online talk is not just noise and that there are significant predictive effects of the number of messages on trading activities. One such work is by (Wysocki 1999), who undertook two research studies to evaluate the impact that stock message boards hosted by Yahoo! have on the stock market. The initial paper Wysocki (1998) examined the main determinants for message-posting volume (low/high profile) on stock message boards. The study provided an explanation of why investors like to post information on these boards. The explanation can be broadly divided into four categories as follows: 1) meaningless random noise, 2) “herding” behaviour by unsophisticated investors, 3) investor disagreement due to inadequate information about a company from normal sources, and 4) investor disagreement due to uncertainty about company projections. The study found that unsophisticated investors follow these message boards to trace “hot stocks” that have potential high prices or abnormal financial performance. In particular, investors are attracted to message postings that discuss information about start-up companies with uncertain prospects and risky characteristics. Finally, the analysis showed that there is a strong relationship between daily trading action and message board activity. For instance, the price and volume of touted stock is influenced by a large increase in the volume of overnight message activity.

Wysocki’s (Wysocki 2000) subsequent paper studied in detail the attributes of individuals (active, casual) who posted messages on the board (“short sellers and message-postings activity on the web”). The analysis showed company’s message board jumps during high short-sale months. Interestingly, casual posters appear more during high short-sale months. This suggests that there are possible manipulators who appear as short sellers and spread false information on stock message boards to hype prices and gain profits.

Tumarkin and Whitelaw (2001) examined the relationship between internet message board activity and abnormal stock returns and trading volume, using event study and the vector auto-regression (VAR) method. In particular, the study followed Wysocki (1999) in terms of its analysis perspective but placed greater emphasis on examining opinions expressed by participants as a determinant of stock prices. The study found that there was significant correlation between days
which featured high posting activity and changes in participants’ opinions, and abnormal returns and trading volume. However, the analysis showed that posting activity cannot predict securities future returns and volume.

Antweiler and Frank (2004) used data from Yahoo! and Raging Bull to study the impact of messages in the market. They considered the bullishness of the messages to predict returns, whether disagreement between participants was associated with a higher volume of trades, and the impact of message levels or bullishness on market volatility. The study concluded that internet message boards present relevant financial information and the discussions which take place on them are not noise. The analysis showed that message board posting predicts a negative return on the following day. Despite the fact that evidence of bullishness or disagreement on volatility is weak, the study suggested that message boards help to predict market volatility. Furthermore, they found that differences of opinion are associated with more trades.

The aforementioned research primarily concerns itself with examining the relationship between message boards and stock price volatility and returns. In contrast, a study by Pleis (2009) involved an experiment to test the impact of message postings on investment decisions made by nonprofessional investors. The study compared investment decisions before and after participants was exposed to different types of message postings (no message postings, positive, negative, and mixed). Surprisingly, the study found that only negative postings influenced investment decisions, which is contrary to the SEC’s beliefs about the “Pump and Dump” model. However, the study had many limitations, for instance they only tested one stock as an initial investigation and participants were provided with limited information on the basis of which to make the investment decision. Interestingly, this study is consistent with the assumption that there are many investors who cannot understand or interpret financial information provided by companies; thus, they are more likely to seek help in interpreting this information from other sources such as internet message boards (Elliott, Hodge et al. 2006).

Sabherwal, Sarkar et al. (2011) extend the literature on the information contained on stock message boards to provide better understanding of the effect these messages have on trading behaviour. The paper argued that an online stock message board could be used by fraudsters as a direct method by which to violate stock prices with small market capitalisation and weak fundamentals. The analysis
shows a significant positive abnormal return (two day pump) on the heavily touted
day and the preceding day, followed by a two day dump stock manipulation
pattern among online traders. Though the aforementioned studies argued that
online messages are not just noise, it is still uncertain whether online messages
influence trading activities. Furthermore, none of the earlier studies had evaluated
the impact of sentiment on returns. Therefore, Sabherwal, Sarkar et al. (2011)
contributed by examining three posting activity factors (sentiment index,
disagreement index and the number of messages) to test their impact on returns.
The research found that sentiment index is the dominant factor and an important
predictor that helps explain returns, volatility, and the proportion of volume in
small-sized trades.
In summary, the dramatic growth of the internet has transformed the way investors
receive information about publicly traded companies. As is evident from the earlier
literature, several studies provide anecdotal evidence that touting campaigns using
spam e-mails and postings on internet message boards work, cause substantial
fluctuations in price, volatility, returns and trading volume during stock campaigns,
and that there are recipients who read the messages and act upon them.
However, the analysis and evaluation performed by previous researchers followed
a traditional approach and relied only on manual classifications.
Given that the amount of spam e-mails and internet message boards is increasing
exponentially, there is a clear need for a more efficient, dynamic, and incremental
way to extract key and sentiment information from these textual sources, and link it
with structured data to provide better analysis and an accurate securities fraud
detection model. In particular, SEC is concerned about the growth of these
sources of internet fraud because they could represent successful attempts to
manipulate stock prices as part of ‘touting campaigns’.
Therefore, the financial market needs an efficient and automatic approach that
considers these information-based manipulation sources and demonstrates a
mechanism to detect fraudulent patterns raised from these sources to alert
investors and regulators about potential manipulative practices. Therefore, this
thesis aims to provide significant cross-fertilization between financial research
studies and information technology as it attempts to incorporate text mining
techniques for the analysis of “stock touting” in unregulated markets through
adopting a “design approach”. In this context, text mining could help in raising
proactive alarms which could help to minimize on-going information-based manipulation. Furthermore, the proposed financial ontology will act as a framework to guide the extraction process and capture financial fraud concepts from these information-based manipulation sources.

2.4 Market monitoring and surveillance systems: Definition and technologies

For the purpose of this thesis, market surveillance follows the definition stated in the work of Polansky et al. (2004). Market surveillance is defined as “systems to encompass the processes and technologies that support the detection and investigation of potential trading rule violations, whether defined in statute or marketplace rules”. Thus, the definition explains that market monitoring and surveillance systems should have processes and technologies that support the investigation and detection of any potential trading violations with special consideration given to the rules of the marketplace and codes of conduct of exchanges and trading venues. The system should include components such as detection, investigation, and enforcement which are supported by different processes and technologies in order to strengthen the enforcement actions. The term “enforcement” means ‘initiating a formal due process proceeding to charge a violation against a legal or natural person and, if found guilty, to impose some form of sanction’ (Polansky, Kulczak et al. 2004).

2.4.1 Technologies which support market monitoring and surveillance systems

Polansky et al (2004) recommended the development of a management information system (MIS) platform to raise the alarm about potential market abuses. These systems are currently used in limited fashion and act as a reporting archive for multiple functions within the exchange. Market surveillance could use these current archive systems (market and non-market data) to retrieve specific data as needed, whether structured (transactional data, quotes, trades, etc.) or unstructured data (news materials, reports, emails, Internet bulletin-messages, fillings, disclosures, etc.) in a timely manner. This would provide inputs for the event detection and analysis components (using BI techniques such as data mining and text mining as described in the following subsections). Finally, the
system would quickly generate alerts and report to analysts about abusive behaviours.

The objective of the systems is to reduce the amount of material that analysts have to scan manually, and allow them to rapidly and easily identify manipulative scenarios which require investigation. These systems could be separated into two categories: offline batch system and real-time systems. The difference between the two batches is how quickly regulatory action could be taken. Despite the fact that the offline batch will infer some delay in the detection process, the real time approach does not produce significant regulatory benefits relative to the costs involved. However, the real-time batch is valuable in the detection of schemes related to time urgency such as marking-the-close.

The current financial market monitoring systems suffer from the absence of a financial ontology (Hilary, Yi-Chuan et al. 2009) that could help in managing vast quantities of financial data, providing better communication and knowledge sharing among analysts, providing a mechanism to demonstrate knowledge of the processes of financial fraud, understanding and sharing financial fraud logic operations, managing relevant facts gathered for case investigations, providing early detection techniques of fraudulent activities, developing prevention practices, and allowing reuse of these knowledge resources in different financial contexts (Kingston, Schafer et al. 2004).

Furthermore, the ontology could act as a semantically rich knowledge base in market monitoring systems that specialize in information management. In particular, the financial ontology could play a central role between information management and business intelligence analysis components. Based on the semantics of ontologies, information and patterns can be extracted from natural language texts and, after processing, knowledge can be extracted that will help the BI analysis components to trigger proactive alarms.

2.4.1.1 Background to data mining

Over the last decade, data mining has come to be considered as one of the most important emerging fields in information technology. The term “data mining” refers to the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules (Michael J. Berry and Linoff 2004). Data mining emerges from several disjointed fields such as applied statistics, information systems, machine learning, artificial intelligence, knowledge discovery, and data
engineering (Cao, Yu et al. 2009). Despite the fact that thousands of algorithms and methods have been developed, domain driven data mining is not actively applied in the business environment (Cao, Yu et al. 2009). One of the most enticing areas for the application of these emerging technologies is finance, which is becoming more amenable to data-driven modelling as large sets of financial data become available. In particular, the sheer volume of data on financial markets has been increasing exponentially and has reached extraordinary levels of data generation compared with other areas. Thus, finding a suitable way to analyse and investigate these data is a vital part of monitoring and surveillance systems. The application of analytical models generated from the data mining could help provide surveillance systems with alternative means of identifying market manipulations instead of using the traditional statistical and economic models (Diaz, Theodoulidis et al. 2011).

Data mining projects can follow a three-stage process: data preparation, which includes data cleaning and data pre-processing operations such as removing noise if appropriate, and dealing with outliers and missing data fields; the model derivation stage, which takes account of choosing suitable algorithms to identify data patterns and considering the sampling process (partitioning the data in training and testing) to identify the algorithms; the usage and maintenance stage involves monitoring database updates through significant iterations and continuing to validate the patterns (Kopanakis and Theodoulidis 2003).

There are two techniques of data mining - supervised and unsupervised. Generally, supervised data mining models attempt to explain particular field in the model, whereas unsupervised data mining identifies patterns or similarities between groups of records without using particular target fields. Classification, estimation and predictions are examples of supervised data mining techniques where the model attempts to explain and find the value of a target field. Clustering is a good example of unsupervised data mining where the model uncovers new patterns from the data without considering any target fields (Michael J. Berry and Linoff 2004).

Classification is one of the most common data mining tasks and is characterised by well-defined classes with a training sample from a dataset which has pre-classified patterns for learning. This should encourage application of the model to unclassified data in order to provide better classification. Typical examples of a
classification task include classifying credit applicants, identifying fraud patterns, direct marketing and technical analysis. Association rules are another common analysis that builds a model based on groups of elements which appear together. This task generates rules from the data based on two elements which frequently occur together. Market basket analysis, cross selling opportunities, disease detection, gene analysis, and web mining are typical applications of association rules. Clustering analysis involves segmenting a varied population into a number of more uniform subgroups or clusters. Clustering is distinct from classification as it does not require any predefined classes. The model is built based on elements grouped together on the basis of self-similarity (Michael J. Berry and Linoff 2004).

2.4.1.2 Background to text mining

In a manner analogous to data mining, text mining aims to extract particular information through the identification and exploration of interesting patterns. The proliferation of documents and digitised media available nowadays has resulted in the need to develop new mechanisms to enable humans to search, process, and analyse increasing amounts of data from multiple information sources. This problem is exacerbated by the unstructured format of sources such as web pages, emails, reports, articles, etc. that make up the majority of the data that humans deal with on a daily basis (Feldman and Sanger 2007).

Studies indicate that 80% of company information is contained in text documents. These documents exist in the form of descriptive formats and could contain industry specific terms and abbreviations. Therefore, both technical and manual efforts are required to deal with these unstructured sources, extract insightful information and discover useful patterns within them (Ur-Rahman and Harding 2012).

However, manual attempts to correlate data across documents, map complex relationships between concepts, or identify new patterns in the sheer volume of document collection are, at best, extremely labour-intensive and time consuming and associated with a lower degree of accuracy. Thus, there is a need for automatic methods of identifying patterns, discovering insightful information, and exploring data relationships which could dramatically enhance the speed and efficiency with which these textual sources can be analysed (Feldman and Sanger 2007).
Text mining is a relatively recent technological development that aims to address the information management problem of unstructured data, called documents, through the use of techniques from areas such as data mining, machine learning, natural language processing, information retrieval, and knowledge management. The term text mining is a little different to the general form of data mining and can be defined as the process of analysing unstructured textual information in an attempt to discover structure and implicit meanings “hidden” within the text (Karanikas, Tjortjis et al. 2000).

Both data and text mining might share common methods of collecting information as raw data, and processing it through the application of data mining. Thus, text mining involves the pre-processing of document collections, called corpus, through techniques such as text categorisation, information extraction, term extraction, and storage of the intermediate representations in a structured format, as well as the techniques to analyse these intermediate representations, which include distribution analysis, clustering, trend analysis, association rules, and visualisation of the results.

The pre-processing of the document collections provides the necessary structure for analysis in a similar way to the structured data used for data mining. In text mining, analysis is more challenging as document collection can either be static or dynamic (i.e. characterised by the addition of new or updated documents over time). In fact, one can describe pre-processing as a procedure by which documents are converted into structured data, where each document is described through a set of features or attributes, called concepts, and could be represented as columns in a table or Excel file where each row corresponds to a document (Feldman and Sanger 2007).

The techniques used during the pre-processing of document collections could be classified into linguistic and non-linguistic categories, or even a combination of the two. Linguistic techniques take into consideration the natural language characteristics of the text in a document e.g., syntax, grammar, dictionaries, etc. In contrast, non-linguistic techniques view documents as a series of characters, words, sentences, paragraphs, etc. Non-linguistic techniques treat each document as a list of words, count the number of times that specific terms (single words or sets of words) appear within a document or corpus, and calculate their proximity to other related terms by taking into consideration their physical proximity within the
document or their presence in related documents. Since it is based on these two types of technique, text mining is sometimes described as linguistics-based text mining or non-linguistics-based text mining (alternatively rule-based text mining or statistical text mining) (IBM 2009).

Non-linguistic techniques that could be used during text mining include *distribution analysis*, which examines how specific terms are distributed across a corpus or in relation to an individual document; *frequency analysis*, which examines the extent to which a set of terms is represented in the document collection with co-occurrences at or above a minimal support level; or statistics-based unsupervised techniques such as Bayesian networks, neural networks, support vector machines (SVM) and latent semantic analysis (LSA). These techniques are also based on the presence of certain terms in a document and they use built-in mechanisms to improve the accuracy of their analysis based on previous results, and feature ways to remove irrelevant results (IBM 2009).

Linguistic techniques involve many linguistic processes such as tokenisation, part-of-speech, segmentation, spelling and grammar, translation, summarisation, stop words, term extraction, and pattern-based information extraction. Tokenisation is the process of separating the set of characters in a document into words or other meaningful elements (text spans) called tokens. It also involves assigning a position for each token relative to the start of the document, for indexing purposes. Stemming/lemmatisation is the process of reducing words to their base word or “stem” by removing suffixes like ‘-s’, ‘-ed’ and ‘-ing’, prefixes like ‘pre-’ and ‘pro-’, or grammatical nuances. Part-of-speech tagging is the process of assigning a valid linguistic category for each token in a document. The valid categories depend on the natural language, and typically include the following: noun, verb, adjective, adverb, participle, coordinator, determiner and preposition. In addition, part-of-speech (POS) tagging could associate tokens with their context, i.e. sentences, paragraphs and documents. Segmentation is the process of separating the text of a document into linguistic segments such as sentences and paragraphs, and furthermore, it can remove certain characters or sequences of characters, or replace them with spaces. Spelling and grammar checking is the process of automatically correcting spelling and/or grammar errors in order to standardise documents for further analysis. Sometimes, however, it might be desirable to analyse documents in their uncorrected format, so spelling and grammar checking
may not be appropriate. Translation is the process of converting a document from one natural language to another. This is desirable when documents from different languages are analysed together. The translation engines available nowadays work surprisingly well both for formal and informal language forms. Summarisation is the process of reducing the size of a document to a short description that outlines its contents in a similar way to the abstract of a journal article. The summarised document does not need to be well-formed in terms of grammar and can sometimes consist simply of a set of keywords (or tags). The term stop words refers to the process of identifying words with little or no meaning for the purposes of differentiating between two documents, the so-called stop words. In English, stop words include ‘the’, ‘and’, ‘of’ or less common ones such as ‘nevertheless’ and ‘differently’. Term extraction is the process of identifying the unique terms in a given document or corpus. Terms can be single words (uni-terms) or sets of words (multi-terms) and they are meaningful concepts within the context of a specific natural language i.e., are terms found in a dictionary of this language (Froelich and Ananyan 1999; Feldman and Sanger 2007).

Pattern-based information extraction refers to the process of extracting information from documents using human-authored linguistic patterns (or rules) to recognise basic linguistic elements of interest, and complex patterns and relationships between various linguistic and non-linguistic elements. This process utilises a pattern (or rule) language to allow users to specify the appropriate patterns as a set of rules (also called regular expressions). Patterns are classified into entity extraction and event extraction groups. Entities correspond to basic linguistic elements such as person name, company name, product name, location, etc; they are sometimes called named entities. Entities could also correspond to non-linguistic elements such as a web link, email address, social security number, phone number, etc. Events are complex patterns that define extraction patterns consisting of named entities, non-linguistic entities or other concepts found within a document (IBM 2009).

Linguistic techniques very often make use of external resources to help with the analysis. These resources may relate to the specific natural language being used in the documents, such as a dictionary or thesaurus. However, it is also possible that concepts in a particular document might also refer to a specific subject area (called domain) such as financial services, biology etc., in which case it may also
be appropriate to use domain-specific resources such as lexicons, taxonomies and ontologies. Some domains also have significant subdomains which may themselves be appropriate for text mining applications. For example, in the case of financial services, there are a number of subdomains, such as corporate finance, securities trading, and commodities that represent a significant body of knowledge with their own linguistic resources.

On that basis, text mining can be defined as domain-independent or domain-dependent. Domain-independent text mining can still involve the use of natural language resources, but their use is independent of any specific body of knowledge or domain and thus, it could be argued that this applies to all documents in any language. In general, text mining applications are most effective if some level of domain-specificity is incorporated into the analysis. It can be argued that leveraging information from resources belonging to a specific domain can improve the identification of the most appropriate concepts for that domain, thus making the analysis more meaningful and effective.

Therefore, the research makes a further contribution through the design of an artefact demonstrating an automatic and efficient method of using text mining technique to extract key attributes and characteristics of different unstructured sources generated as part of "touting campaigns". The financial ontology will provide the underlying framework for the extraction process and the capture of information related to financial fraud. Another very important contribution of the ontology is that it will address the need to reuse the text-mining process in other parts of the domain, and to integrate the extracted information with other systems within the organization.

2.4.2 Previous research on BI for fraud detection

This section provides a review about studies used data mining and text mining to investigate and detect fraudulent behaviours and ascertaining evidence of potential cases of fraud within different financial markets (Phua, Lee et al. 2005). In fact, these systems could support financial organisations to proactively detect transactions where market abuse is suspected, thus minimising the circulation of rumours or illegal information (Donoho 2004). Therefore, this research aims to contribute to a better understanding of the market manipulation problem and
provide part of a unified framework for the design and analysis of market manipulation systems.

The fraud detection literature outlines a number of different data mining techniques that have been utilised to detect fraudulent activities in different domains such as telecommunications, stock exchanges, credit-cards, insurance companies, and banks and securities firms. Focusing specifically on the market manipulations domain, Kirkland and Senator (1999) described how the National Association of Securities Dealers (NASD) Inc. used a fraud detection system [called Advanced Detection System (ADS)] to monitor trades and quotations in the NASDAQ Stock Market. The main objective of the system was to detect and identify any suspicious trading behaviour for further investigation and raise the level of surveillance from issue-based to firm-based patterns and practices. The system uses many artificial intelligence techniques such as visualisation, data mining techniques (association rules, decision trees), time sequence pattern, and pattern recognition to cover market surveillance, trading violations and to help the regulatory parties protect the market from any breaches. Goldberg, Kirkland et al. (2003) describe the Securities Observation, News, Analysis and Regulation system (SONAR), also developed by NASD. The system’s main purpose is monitoring NASDAQ transactions in the stock market to detect and identify any potential insider trading and any falsification of news stories for the purposes of fraud. SONAR uses a range of techniques including data mining, text mining, statistical regression, fuzzy matching, and rule based inference to mine news wire stories and US Securities and Exchange Commission (SEC) filings. Furthermore, the system evaluates price and volume models for various securities within the market, and each day generates flag alerts for further investigation that could support the prosecution of those involved in manipulation. Donoho (2004) utilised data mining techniques (C4.5, decision tree, neural network, K-mean clustering, and logistic regressions) for the early detection of insider trading manipulation schemes before the news broke within the option market. The researcher carried out case studies for two companies which took into account the news released by the companies themselves. The news indicated negative trading behaviour. In addition, the study compared the results of the different algorithms used in the experiment. Furthermore, the research concluded
that the early detection of insider trading using these data mining techniques could protect the market and its investors from potential losses.

Xia (2007) generated a conceptual framework to identify the individuals (and their communities) involved in trade-based manipulation using data mining. Specific data mining techniques such as Euclidian Distance (ED), Shared Nearest Neighbour (SNN), density-based algorithm (DBSCAN) and graph-partitioning algorithm (METIS) were used, based on the scenarios of trade-based manipulation. The researcher developed a software tool which could implement these proposed techniques to detect the manipulation and identify the suspects and their communities. The experiment is based on real transaction data obtained from the stock exchange market. The conceptual framework had some success in detecting trade-based manipulations and the suspects within their communities.

Diaz, Theodoulidis et al. (2011) employ a data mining approach to analyse two cases of manipulation in the New York Stock Exchange. The researcher used the decision trees technique to distinguish between manipulations and normal trading and to improve organisational fraud detection systems. The paper analysed intraday data to study abnormal returns, volume patterns and their relationship with the manipulations and reactions to the news releases. The research reports that a higher proportion of abnormal returns and volume outliers are directly related to the manipulation sample. Moreover, changes in volumes, either daily or compared to previous days, are key elements when investigating manipulations. Furthermore, news and events are important moderators of these indicators, and they are preferred over alternative control variables such as different preconditions. This allowed the decision rules to be more general, and potentially valid for other types of manipulation schemes.

Zaki, Theodoulidis et al. (2011) described a case study on fraud detection using data mining techniques that help analysts to identify possible instances of touting based on spam emails. Various data mining techniques such as decision trees, neural networks and linear regression are utilised in this emerging domain. The application of these techniques is demonstrated using data from the Pink Sheets market. Results strongly suggest the cumulative effect of "stock touting" spam emails is key to understanding the patterns of manipulations associated with touting email campaigns, and that data mining techniques can be used to facilitate fraud investigations of spam emails.
In terms of gaps in the research, several market violations are dependent upon the existence of news or other textual resources such as emails and internet bulletins that are disseminated to the public in order to influence market prices. Designing analysis to identify these types of resources and capture the information which is relevant to surveillance will contribute to the detection systems, especially in manipulation schemes such as “pump and dump” and “insider trading”. Many markets generate substantial quantities of textual resources, that require labour intensive manual processing, which is expensive and associated with a lack of accuracy (Polansky, Kulczak et al. 2004).

On the empirical side, the thesis will present examples of novel applications of text-mining, data mining tools and data-processing components, developing several off-line surveillance systems that are fully working prototypes which could train the ontology in the most recent manipulation techniques such as ‘pump and dump’. In particular, the thesis demonstrates the impact of using automated, linguistics-based text mining techniques to automate analysis and classification of large amounts of financial textual sources, identify interesting patterns and relationships between them and help fraud analysts to proactively investigate possible manipulation cases. This could help regulators and market participants to minimise the risks from manipulation schemes such as ‘pump and dump’ schemes.

2.4.3 Existing systems and frameworks

This section presents a review of the existing market monitoring surveillance systems currently in place in stock markets. The business department responsible for the surveillance of stock markets regulated by NASD is the market regulation department (MRD). Historically, MRD established an automated surveillance system in 1988; MRD developed a system called stock watch automated tracking (SWAT) that was responsible for detection of insider trading fraud (Goldberg, Kirkland et al. 2003). In 1996, MRD developed the advanced detection system (ADS) that was designed to identify possible patterns of violation by brokers (Kirkland and Senator 1999). SWAT was replaced by the securities observation news analysis & regulation (SONAR) framework, which was developed within the system to act as a knowledge-based automated surveillance system (Senator 2000; Goldberg, Kirkland et al. 2003).
Goldberg et al.'s (2003) work presents one of the most complete conceptual architectures for the SONAR system (figure 2-3). Detection of market manipulation depends on collecting and analysing relevant data to help analysts investigate fraud cases. One of the major objectives of the system is to automate evidence gathering from different sources, including structured data (market data represented in quotes, trades, and orders), unstructured sources (corporate filings, news, etc.), and numerous internal financial documents such as SEC referrals, complaint data and disciplinary history data. This data will be fed into analysis components and linking processes to identify potential violation patterns which will later be reviewed by analysts to determine whether there is an explanation for the apparent violation. SONAR provides comprehensive surveillance activities with respect to insider trading and fraud. In particular, SONAR combines text mining components to detect patterns from news wires and SEC filings. Furthermore, a data mining component is designed to detect characteristic “events” in the context of prices and volumes activities in the market. The system uses post-extraction and consolidation analysis to combine and link pieces of evidence with each other in a meaningful way in order to demonstrate the possible violation episodes and link them to the correct market activities.

Figure 2-3 Conceptual Architecture of the SONAR System (Kirkland and Senator 1999)
The work of Mongkolnavin and Tirapat (2009) introduced a monitoring system used on the Thai Bond Market, which was commissioned by the Thai Bond Market Association (ThaiBMA). As shown in figures 2-4 and 2-5, the market uses a real-time approach to monitor transactions, investigate any unusual ones, and notify regulators where enforcement action is required. The system includes a trading analysis component which uses both economic analysis and behavioural analysis to identify fraudulent patterns. Economic analysis relies on traditional financial econometrics and compares current trading prices with the historical time series properties of the bond prices using GARCH models (Bollerslev 1986). This approach is not the best option because the bond market in Thailand is relatively low compared to other developed markets. Thus, the system includes a behavioural analysis approach which employs an association rules data mining application to analyse the behaviour of traders. The system integrates the two approaches (figure 2-4) to raise alarms and issue warning signals to the analysts for further investigation.

The inputs into this system consist of data in the form of prices, volumes, trader information and collection of market integrity rules. The output of the system is passed to an information unit to continue the monitoring process and decide whether the flagged trade requires further investigation, which may be followed by suitable enforcement action according to the law and market rules.

Figure 2-4 Surveillance Workflow of the ThaiBMA (Mongkolnavin and Tirapat 2009)
Díaz, Zaki et al. (2011) proposed a market monitoring framework, comprising of the analysis components, tasks and flows of information of a complete financial market monitoring system. The framework is designed to have a past time or reactive monitoring engine, which is fed with either structured or unstructured data sources. The framework considers the possibility of potential manipulation across two markets, represented in figure 2-6 as markets A and B. The information management component is dedicated to keeping and integrating the various types of data required for the analysis component. Furthermore, the information management component includes all data preparation and pre-processing necessary to render the data suitable for analysis. Generally, the analysis component is divided into two categories, namely behavioural analysis and economic analysis. In this context, behavioural analysis employs different BI techniques such as data mining, text mining and social network analysis to analyse the actions and characteristics of the agents involved in the manipulation scheme. However, economic analysis employs econometric and financial models, besides data mining and text mining technique, to analyse the impact of the manipulation on the market, represented by price and volume violation. The output of the engine and the information management layer in any form will be recursively stored in a Knowledge Base and made available to the rest of the system. The
results from the different analyses are used to generate warning signals and alarms for the markets affected by the manipulation scheme. The output management layer will be used differently, according to the type of user. For example, regulators could flag trades and initiate a further investigation process, which may be followed by enforcement actions based on laws and market rules. The framework considers the real time approach using a Demand-driven Active Mining of Data Streams, embedded in the real time monitoring engine as in (Fan, Huang et al. 2004).

![Figure 2-6 Market Monitoring Framework (Díaz, Zaki et al. 2011)](image)

According to Aitken, Harris et al. (2010), some exchanges such as the London Stock Exchange and DirectEdge, have launched Enhanced Liquidity Provider Programmes (ELPs) which offer an integrated view of both displayed and dark pool order books. There are plans for these exchanges to gradually adopt real-time market surveillance using the SMARTS surveillance platform from Smarts Group International Ltd. a NASDAQ OMX company², which is regarded as the most complete and advanced technology available. To date, SMART systems have been used by more than 50 national securities, exchanges and regulators around the world.

²http://www.smartsgroup.com
Given that the Market Monitoring Framework introduced by Díaz, Zaki et al. (2011) suffers from the absence of a financial ontology. This research extends the Market Monitoring Framework, to build the ontology that will simulate the unregulated market in a machine-processable form, providing an information management layer in the framework of a semantically rich knowledge base that can be used for the semi-automatic interpretation of relevant unstructured resources. Based on the semantics of ontologies, information and patterns can be extracted from natural language texts and, after processing, knowledge can be extracted that will help the BI analysis components to trigger the alarms. Furthermore, previous studies focus only in regulated market.

This thesis contributes by identifying a number of additional components that address requirements for the proactive identification of possible fraudulent behaviours in the unregulated market, such as ‘pump and dump’ scenarios, through the analysis of textual information resources. This could provide the regulators and market participants with efficient fraud detection services that could play a vital role in monitoring, detecting, and deterring against abuses in the unregulated markets.

2.5 Financial Market Ontology

This section presents a review of the existing financial ontologies. It describes the meaning of ontologies, different types of ontologies and evaluates previous work that has been carried out in financial ontology-related research areas. Ontologies are widely perceived to play a vital role in fields such as Artificial Intelligence (AI), Business Intelligence and Knowledge Management. The philosophical meaning of the ontology is “Study of existence or being or reality” (Mizoguchi and Ikeda 1998). Generally, the objective of the ontology is to describe the classes (terms and vocabularies), and define their relationships according to specific domain rules, in order to describe the knowledge in a generic way.

The most widely quoted definition of ontology is that provided by Gruber (1993) “ontology is a formal, explicit specification of a shared conceptualisation”. Conceptualization referred to the abstract model of objects in the world and shared is assumed there is a community who can accept the definitions and relationships described in the ontology. The IEEE Standard Upper Ontology working group defines an ontology as “similar to a dictionary or glossary, but with greater detail
and structure that enables computers to process its content. Ontology consists of a set of concepts, axioms, and relationships that describe a domain of interest” (IEEE 2002).

Since the early 1990s, ontologies have become a popular research topic that has been studied by several artificial intelligence research communities, including knowledge engineering, natural language processing, and knowledge representation (Mikroyannidis 2007). Ontologies are also a key enabling technology for the Semantic Web, as they offer a way to give information a common representation and semantics. They mainly constitute “a shared and common understanding of a domain that can be communicated between people and application systems” (Davies, Fensel et al. 2003). Consequently, they can be deployed to bridge communication gaps between humans, software systems, or between humans and software systems (Mikroyannidis 2007).

There are three different levels of interpretation for any ontology, namely: conceptualisation, specification, and representation. Generally in the conceptualisation level, humans capture and construct the concepts and relationships. At the specification level, the ontology objects of the conceptualisation level are defined in a precise way through adding and constructing the defined concepts and relationship hierarchies accurately. The representation level contains the actual formalisation, i.e. ontology language that represents the ontology specification (Klein 2004).

There are various types of ontologies providing different perspectives of building ontologies, namely, top-level, domain, and specific domain ontologies. Top-Level ontologies aim to provide general categories at the highest level of ontology such as Penman Upper Model, Cyc/OpenCyc, EuroWordNet, IEE Standard Upper Ontology, and others. Core domain ontologies aim to cover a specific topic or subject area such as finance, commerce, and others (Brewster 2008). Specific domain ontologies are more specialized ontologies such as financial fraud ontology, financial news ontology, and others. Navigli et al. (2003) among others argue that designing specific domain ontology is challenging because many domains evolve quickly and cannot rely on general core domain ontologies which have general resources such as WordNet, Investopedia, etc.

Information management stances various challenges for building ontologies. Ontologies are a key factor of Information management as they provide a common
representation to the domain. However, the approach adopted in most information management systems such as SHOE, OntoView, Sesame & OMM, PROMPT, and KAON is simplistic because these systems follow the flat architecture Ontology layering. The disadvantage of the flat architecture ontologies are managed independently which makes it challenging to integrate them, especially when multiple ontologies are introduced. Mikroyannidis and Theodoulidis (2012) proposed the multi-layer architecture for building ontologies. The advantage of using the multi-layer architecture is building number of layers containing ontologies for different purposes developed by different author groups. The architecture could help to improve the manageability of the technologies and demonstrate the integration between different ontologies presented through intra-layer (same layer) or inter-layer (different layer) ontology mapping.

The purpose of the ontology engineering is “to provide a basis of building models of all things”. The essential components of ontology engineering are basic issues in philosophy and knowledge representation, ontology design describes whether the ontology is general or domain or specific, standardization such as EDI, reuse and sharing of knowledge, media integration such as documents, design methodology describes the methodology and support environment, ontology evaluation (Mizoguchi and Ikeda 1998).

This thesis focuses only on seven components of the ontology engineering. Firstly, basic issues in philosophy and knowledge representation as this research adapted Mikroyannidis and Theodoulidis (2012) multi-layer architecture for building ontologies. Secondly, the ontology is a domain specific ontology which aims to build financial fraud ontology from corpus. Thirdly, the ontology relies on different data sources to build the knowledge of the proposed ontology such as SEC litigation releases. Fourthly, based on the architecture adapted from Mikroyannidis and Theodoulidis (2012), the ontology follows KAON standardisation (RDF/XML). Fifthly, the proposed ontology integrated the SEC litigation releases as relevant documents where fraud cases are described in details. Sixthly, the research used the focus group method as top-bottom approach to build the specific domain ontology. Furthermore, constant comparative method introduced by (Grove 1988) is used as a methodology to analyse this focus group data. The ontology used PoolParty Thesaurus management as a support environment to present the proposed ontology (http://poolparty.biz/). At last, the ontology is evaluated through
specific case studies to show the applicability of the ontology to help users in the market monitoring surveillance systems.

2.5.1 Existing Financial Ontologies

The finance domain has historically suffered from a lack of efficiency in managing vast amounts of financial data, a lack of communication and knowledge sharing between analysts, the lack of a mechanism to resolve synchronisation problems when multiple users access data, and the lack of an automated system to publish the reports (Hilary, Yi-Chuan et al. 2009). Hilary et al. (2009) demonstrated an ontology-based approach for BI applications of financial knowledge management systems (FKMS), specifically employing statistical and data mining analysis to manage daily operations, make better trading decisions and support investment decisions. FKMS consists of five layers: the resource layer, data storage and management layer, knowledge/trend/pattern layer, and finally the user process layer. Various data sources are converted into a time series database based on a predefined schema, then loaded into its data warehouse, using the OLAP tool to allow users to retrieve periodical reports and generate sets of cubes for analysis. The result of the OLAP process is fed into the statistical and data mining techniques to run valuation models of the bond rating classification. Finally, FKMS provides users with financial and economic data, analysed data with different decision support modules, and experiments with various sets of parameters. Furthermore, the ontology of knowledge management adds value to the FKMS by collecting, classifying and sharing this knowledge with users.

The European Commission funds research projects which design and develop aspects of financial ontology, such as the Data, Information and Process Integration with Semantic Web Services (DIP) project that developed stock market ontology. The ontology was evaluated by analysing an eBanking case study (Bankinter). The project used On-to-knowledge methodology to construct the main concepts and relations, and classify them using natural language processing. The ontology resources were based on reusing existing public financial domain sources such as extensible business reporting language (XPRL is used to report the periodic information of companies to the stock exchange market authorities), existing stock market websites, stock brokers, and other relevant sources such as Spanish bank websites. This ontology aimed to help bank customers understand
complex operations in the stock market by covering cases from most of the European stock markets and the national market, considering main concepts, relationship between concepts and operations in the stock market (Alonso, Bas et al. 2005).

The impetus for building financial fraud ontology arises from the growth of financial fraud in a number of sectors, and the massive losses that those sectors are suffering as a result of such schemes. Generally, fraud cases are challenging, complex, and involve a huge volume of information. Gathering facts and evidence is often particularly complex. Therefore, there is an urgent need to build a systematic ontology which aims to promote prevention practices and knowledge of financial fraud processes, managing massive amounts of relevant facts gathered for case investigations, early detection techniques for fraudulent activities, policies, laws and regulations (Kingston, Schafer et al. 2004).

Thus, the ‘financial fraud prevention oriented information resources using ontology technology project’ (FF POIROT) was funded to supplement the efforts made by EU member states to combat financial fraud. The project aimed to design a set of user requirement specifications that define the functionality of the financial fraud ontology. In particular, the project demonstrated a legal modelling approach of detailed criminal fraud ontology, which contains a combination of concepts from various laws, prevention practices, and knowledge of processes of financial fraud within the EU states. Furthermore, the research focus is to examine the legal approach of cross-border value added tax fraud within the EU, and investment fraud on the Internet. The ontology is a semantic representation of mining concepts from unstructured or semi structured resources and internet sources. Furthermore, the ontology was compiled using several European languages in order to leverage a comprehensive knowledge repository. The project proposed knowledge modelling based on using inference networks of law to show the direction of reasoning or the direction of probabilistic influence among nodes on the network. The model can help investigators to break fraud cases into main three models, using Wigmore’s argument chart, as shown in figure 2-7. Firstly, a proposition layer (major hypothesis layer) includes any suspected fraud cases. Secondly, the law layer includes appropriate laws and regulations that match the fraud case. Thirdly, the evidence layer embraces all facts and evidence that supports and proves the fraud case (Kingston, Schafer et al. 2004). Existing legal
ontologies are not detailed enough to model the factual materials of fraud law. Thus, the model has been complemented with SUMO’s financial ontology and ontology of services and McCarthy’s REA ontology.

The ‘FF POIROT’ ontology application is used for filtration and evidence recognition to help investigators identify fraud patterns on manipulated securities and gather all justifiable evidence for prosecution. The ontology was developed based on the Developing Ontology-Guided Mediation for Agents (DOGMA) approach which is concerned with traceability of decision-making in the development stage. Natural language text was used as a medium of fraud forensics to extract relevant terms from lexons and extract relationships between the terms (Zhao, Kingston et al. 2004). The case study selected was an email fraud application which uses the financial ontology repository and its multilingual lexical resources to recognise different types of e-mail fraud such as phishing, Nigerian advance fee fraud and lottery scams. The research demonstrates the development cycle of such an application, which contains four tracks, namely: language engineering, terminology engineering, knowledge engineering and system engineering. The application is considered to be an intelligent system that extracts information from emails or web pages based on the financial ontology models provided by financial fraud detection domain experts and analysts. The application relied heavily on text clustering technique and regular expression rather than developing sophisticated natural language processing analysers. However, this could limit the research because of the dynamic and deceptive nature of such fraud emails, and sentiment models would be needed for more detailed and deeper analysis. Overall, the email fraud application added value to
the ontology application by training the model with the most recent manipulation techniques (Kerremans, Tang et al. 2005).

In summary, fraud cases are notorious for their complexity because the facts and evidence surrounding a fraud can be highly complex. ‘FF POIROT’ followed a bottom-up approach and mostly used legal-driven ontology modelling applications. The elements of the ontology are justified based on analysis conducted from real life cases and knowledge engineering activities. The ontology could act as a specialised financial legal model that could help in solving a specific problem in a particular context. It lacks generality and does not demonstrate the conceptual system of the knowledge structure of domain experts; in particular, the common financial vocabulary used by the community and the essential financial business logic is not included in the application ontology. Therefore, this thesis contributes by resolving these limitations of ‘FFPOIROT’ through building an ontology using a bottom-up and top-bottom approach and is mostly domain expert driven in order to cover the common vocabulary and business logic processes of the financial market (Zhao, Leary et al. 2005).

Through examining the existing approaches in these areas and identifying their limitations, the research objectives were refined and a number of requirements for the proposed ontology architecture have been identified. Therefore, this research uses the multi-layer architecture to build the proposed financial ontology adapted from (Mikroyannidis and Theodoulidis 2012) because of the aforementioned reasons. The proposed ontology is integrated and aligned with other ontologies from different sources to design a comprehensive conceptual system that could enable its wide resources coverage to help users understand the financial fraud logic operations and allow reuse of these knowledge resources in a different financial context. Specific instantiations will be demonstrated through case studies to show the applicability of the ontology to help users in the market monitoring surveillance framework detect fraud cases and identify fraudulent patterns in the stock market. Furthermore, the ontology could add value in proactive fraud monitoring or help in open investigation by acting as a knowledge management repository system that manages and controls the masses of data that can be gathered during financial fraud and share existing manipulation patterns from prosecuted cases among investigators and relevant users.
2.6 Gaps in the literature

This section discusses and comments on the limitations and gaps of the literature in order to come up and identify a set of requirements needed to build financial fraud ontology for fraud purposes. The criterion for selecting certain research areas for inclusion in this chapter was their relevance to the objectives of this work, as these have been stated in the Introduction of this thesis.

The discussion shows that previous studies focus only in analysing the fraudulent activities occur in the regulated market. However, the unregulated market is operated differently from that of the regulated markets. Furthermore, the opaque environment of unregulated markets and the absence of legal trade data disclosure requirements mean that OTC markets are less directly monitored than is the regulated market. Thus, many firms choose to move to unregulated markets because they either forced to leave the regulated market or left the market willingly. The consequence of these limitations is that most manipulation cases occurred in the unregulated markets.

Thus, the continuous improvement and development of financial market monitoring and surveillance systems with high analytical capabilities to capture the fraud is essential to guarantee and preserve an efficient market. This thesis argues that market monitoring systems require financial ontology in order to efficiently manage the vast quantities of financial data, provides better communication and knowledge sharing among analysts, provides a mechanism to demonstrate knowledge of the processes of financial fraud, understands and shares financial fraud logic operations (practices of different manipulation types), manages relevant facts gathered for case investigations, provides early detection techniques of fraudulent activities, develops prevention practices, and allows reuse of these knowledge resources in different financial contexts.

Despite the fact that there are few attempts to build financial ontologies for fraud purposes, the existing ontologies serve the legal purposes. In particular, these ontologies contain detailed criminal fraud concepts from various laws, prevention practices, and knowledge of processes of financial fraud within the EU states. Furthermore, the existing ontology lacks generality and does not demonstrate the conceptual system of the knowledge structure of domain experts; in particular, the common financial vocabulary used by the community and the essential financial business logic is not included in the application ontology.
Therefore, this thesis contributes by resolving these limitations of ‘FFPOIROT’ to an extent through building an ontology using a bottom-up and top-bottom approach and is mostly domain expert driven in order to cover the common financial vocabulary and understand different manipulative activities within the unregulated financial market.

The discussion set out above provides an overview of the wide variety of practices and strategies which are used to manipulate the stock market such as ‘pump and dump’ manipulation strategies. In particular, several studies provide anecdotal evidence that touting campaigns using spam e-mails and postings on internet message boards work in the unregulated market, cause substantial fluctuations in price, volatility, returns and trading volume during stock campaigns, and that there are recipients who read the messages and act upon them. This could become a threat to the unregulated market’s integrity and efficiency.

Therefore, this thesis aims to provide significant cross-fertilization between financial research studies and information technology as it attempts to incorporate text mining techniques for the analysis of "stock touting" in unregulated markets through adopting a “design approach”. In this context, text mining could help in raising alarms which could help to minimize on-going information-based manipulation. Furthermore, the proposed financial ontology will act as a framework to guide the extraction process and capture financial fraud concepts from these information-based manipulation sources.

The analysis presented in the information-based manipulation studies followed traditional financial approach and relied on manual classifications. Given that the amount of spam e-mails and internet message boards is increasing exponentially, through the instantiations of the ontology this research makes a further contribution through the design of an artefact demonstrating an automatic and efficient method of using text mining technique to extract key attributes and characteristics of different unstructured sources generated as part of "outing campaigns". These developed artefacts act as off-line surveillance systems that are fully working prototypes which could train the ontology in the most recent manipulation techniques such as ‘pump and dump’. Furthermore, it will show the applicability of the ontology to help users in the market monitoring surveillance framework in order to detect fraud cases and identify fraudulent patterns in the stock market.
Chapter 3 begins to address these issues by describing the methodology used to design an information system artefact capable of addressing the gaps identified above, which is presented in extension in Chapter 4, 5 and 6. Chapter 4 presents the financial ontology for fraud purposes, describe its ontology layering architecture, the process of constructing the domain ontology from a corpus, and including one case study to demonstrate empirically how text mining can be integrated with the financial fraud ontology to extract financial concepts from the unstructured sources and demonstrate the published prosecuted cases in an appropriate knowledge base. Chapter 5 describes the case study of a stock spam e-mail demonstrating another instantiation of the proposed financial fraud ontology as a text-mining application. Chapter 6 demonstrates the case study of BI system for a financial market monitoring and surveillance system.
3 Research Methodology

This chapter discusses the methodological approach used to design an innovative solution with which to address the gaps identified in the literature review and achieve the research objectives of this thesis. Design science will be used to achieve those objectives concerned with building ontology for fraud purposes. This chapter begins by providing a general definition of design research, outlines its philosophical and epistemological underpinnings, followed by a justification of why this methodology is appropriate for this research. It continues with a detailed description of how the methodology was applied.

3.1 Description of the methodology

This study uses the ‘design science research’ (DSR) approach which is part of the analytical techniques and orientation (complementing the positivist and interpretive perspectives) used to conduct research into information systems (IS). The DSR research method is employed to offer learning through constructing artefacts. It could be defined as “an approach that involves analysis of the use and performance of designed artefacts to understand, explain and, very frequently, to improve on the behaviour of aspects of IS” (Orlikowski and Barley 2001). Generally, these artefacts are diverse and include solutions such as human/computer interfaces, information retrieval system design methodologies (Vaishnavi and Kuechler 2004/5). Artefact design is an activity which distinguishes the professions from the sciences (Simon 1996; first published in 1969; IS Association 2009) and focuses on design activity from an intellectual level perspective (Simon 1996). (Vaishnavi and Kuechler 2004/5) differentiate between "natural science" and the "science of the artificial" (design science). A natural science is “a body of knowledge about some class of things or objects or phenomenon in the world (nature or society) that describes and explains how they behave and interact with each other”. However, a science of the artificial is “a body of knowledge about artificial (man-made) objects and phenomena designed to meet certain desired goals” while considering the internal and external environment forces that may affect the artefact. The power of IS research when considering the convergence of people, organisations and technology as it is examines the technological system, the social system, the two systems side by side, and also investigates the
phenomena which emerge when the two systems interact (Davis and Olson 1985; Lee 1999; Hevner, March et al. 2004). Thus, the process should occur in a cycle in which knowledge is used to create works and works are evaluated to build knowledge. This thesis assumes that there is an interdependent relationship between design science and behavioural research (which aims to develop and verify theories that explain or predict human or organisational behaviour). In this relationship, relevance is achieved by extending the boundaries of human and organisational capabilities by developing artefacts and theories that address important and relevant problems (Hevner, March et al. 2004).

3.1.1 The design science research cycle

Takeda et al. (1990) demonstrated the general design cycle, describing the reasoning activities undertaken in design research. The design cycle shows how knowledge is generated using a sequence of reasoning activities. This view demonstrates how DSR supports problem solving that continually shifts perspectives between design processes (a sequence of expert activities which produce an intellectual product) and designed artefacts, to tackle the same complex problem (Hevner, March et al. 2004).

Generally, the design starts with awareness of problem which emphasises the need to improve research, address a deficiency, solve a problem, and improve the performance of research activities. In the next process, a suggestion for a solution could be captured from the existing knowledge, research, and theories that could accommodate the problem area. Next, an initial trial is conducted to implement the artefact through a development process based on the suggested solution. Partial or full trial artefacts are then evaluated based on the functional specification recommended in the suggestion stage. Normally, the suggestion, development, and evaluation stages are iterative until they offer a reasonable version of the artefact. The conclusion stage indicates the end of the design of a specific project, accompanied by the termination of the iterative process (Vaishnavi and Kuechler 2004/5).

Knowledge is constantly produced through the process of artefact design, which is represented by two feedback loops called the circumscription process and operation and goal knowledge process. The circumscription process generates understanding that could only be gained from the specific act of construction. The
DSR *learns* and *discovers* when activities *don’t work*, with the possibility of either complementing or contradicting existing knowledge or theories. On the other hand, *the operation and goal knowledge* produces understandings through artefact operations. The original attempts at an artefact may be incomplete and do not give a clear understanding of the artefact objectives. However, these attempts could drive the need for further investigation and exploration in this area. Thus, the design, development, and evaluation is iterated a number of times before the final design artefact is produced (Markus, Majchrzak A. et al. 2002). Finally, when the artefact becomes mature enough for use, all ambiguous events will provide explanations as to the reasons why the artefact worked.

3.1.2 Steps and outputs of design research methodology

(Vaishnavi and Kuechler 2004/5) summarise the general research methodology that is basically related to the process of reasoning in the design cycle, as illustrated in figure 3-2. These two processes are *building* and *evaluating*. Each process is associated with a research output.

![Figure 3-1 Research Mapping adapted Design Research. Cited from (Vaishnavi and Kuechler 2004/5)](image-url)
As demonstrated in figure 3-2, an awareness of problem may be identified by different multidisciplinary communities, with the possibility of contributing new findings to the researcher’s field. The output of this phase is either a formal or informal proposal for a new research project. This is followed by the suggestion phase which produces a tentative design. Proposal and tentative design are linked with each other and presented in a single box to highlight the importance of a good proposal and good design; if they are not of a sufficient standard, both should be rejected in the early stages of the research.

At the development stage, implementation of the tentative design begins, and various techniques with which to construct the artefact are considered. Normally, novelty is more significant in the design of artefact than its state-of-practice. (March and Smith 1995) proposed four general outputs for design science research: constructs models, methods, and instantiations. (Vaishnavi and Kuechler 2004/5) proposed a fifth output of design research: better theories.

- **Constructs** are the conceptual vocabulary of a problem/solution of a research area. Constructs come into effect during the conceptualisation of the problem and develop throughout the design cycle.

- **Model** is a group of defined propositions or hypotheses describing relationships between constructs. (March and Smith 1995) identify models with the design problem and its potential solutions.

- **Methods** are a group of processes that are used as a guideline to perform a task. Methods are goal oriented and have various exemplifications such as mathematical formulae, computational algorithms, software solutions, informal descriptions, or 'best practice' approaches, including frameworks and industry standards.

- **Instantiations** demonstrate the possibility of implementing constructs, models and methods in the operational environment, representing the practicability of applying the artefact, and evaluate its performance (Hevner, March et al. 2004).

- **Better theories** could be generated by the methodological construction of artefact phase that may act as experimental proof of method, an experimental exploration, or both. On the other hand, relationships may develop between the components of the artefact which make certain behaviours more visible than they were during the construction or
evaluation stages. This could therefore increase understanding of these components and potentially complement or contradict previous conceptions or theories. Therefore, design research can make significant contributions to the development of new theories or disproving existing ones.

Once the artefact has been designed, an evaluation process should take place based on the criteria proposed at the *awareness and problem* stage. This provides feedback and better understanding of the problem, in order to enhance the quality of the artefact and the design process (Hevner, March et al. 2004). Some deviations from the original hypothesis are to be expected. However, these should be considered and tentatively explained either to confirm or contradict the original hypotheses. Finally, the *conclusion* stage is the end of the research life cycle and is represented by the results the artefact produces.

Hevner et al. (2004) demonstrate a conceptual framework, as shown in figure 3-3, for understanding, executing, and evaluating IS research combining behavioural science and design science paradigms.

![Figure 3-2 Information Systems Design Research Framework (Hevner, March et al. 2004).](image)

The main objective of design research is to solve the domain problem through *building* and *evaluation* of artefacts. The artefacts should consider the business needs of the organisation, which include its goals, tasks, problems and opportunities as they are perceived by people within the organisation. These
business needs should be reviewed to meet the organisation's objectives, considering strategies, structure, culture and processes. Furthermore, existing technologies within the organisation, such as infrastructure, applications, communications and architecture, should be reviewed. As a result, IS research utilises the knowledge base which contains existing foundations (e.g. theories, frameworks, etc.) and methodologies (e.g. analysis and measures for evaluation) to achieve the research objectives (Hevner, March et al. 2004).

Thus, design research directs its efforts to addressing “wicked problems” and designing intellectual artefacts to solve these problems in innovative ways. However, routine design relies on existing knowledge to solve organisational problems such as constructing a financial or marketing information system using best practice artefacts (constructs, models, methods, and instantiations) which already exist in the knowledge base.

3.1.3 The philosophical grounding of design science research

Generally, the four basic beliefs in research studies are ontology (which describes the nature of reality), epistemology (explores the nature of knowledge), methodology and axiology (study of values). From an ontological point of view, design science research changes the state of the world through the novelty of designing artefacts. (Bunge 1985) asserted the significant impact made by DSR when researchers shift between pragmatic and critical realistic research perspectives, guided by a pragmatic evaluation in the design cycle. Thus, DSR researchers are flexible, and have alternative contextually situated and multiple world-states at their disposal, compared to researchers in other paradigms, such as positivism and interpretive. Furthermore, a composite socio-technical system is enabled as a unit of analysis and the hypotheses are subject to modification in the design cycle.

From an epistemological perspective, the artefact is constructed based on factual information through the process of circumscription. Knowledge is generated through the creation and construction of the design to deal with a specific research area. The reasoning and meaning of this knowledge is explained through the design cycle and iterations. Moreover, DSR researchers may change their assumptions during the iterative process and act as observers to record the system’s behaviour and compare it to the theory used in the adductive phase. The
degree to which the artefact functions predictably (what the artefact means is what it does) gives the DSR a similar epistemology to natural-science research. DSR researchers pursue the truth and an understanding of the problem and have a greater tolerance of vagueness and ambiguity than positivist researchers. Many researchers have stated that the findings of their research are not widely understood but still consider them to make a contribution to the community (Hevner, March et al. 2004). Much DSR research work is considered in terms of “wicked problems” which contradict existing theories and require further exploration. Generally, “wicked problem” research highlights and codifies the problem and disseminates it to the community as a contribution that will lead to further investigation (March and Smith 1995).

Table 3-1 Philosophical Viewpoints of the three research paradigm perspectives (Vaishnavi and Kuechler 2004/5)

<table>
<thead>
<tr>
<th>Basic Belief</th>
<th>Positivist</th>
<th>Interpretive</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology</td>
<td>A single reality. Knowledge, probabilistic</td>
<td>Multiple realities, socially constructed</td>
<td>Multiple, contextually situated alternative world-states. Socio-technologically enabled</td>
</tr>
<tr>
<td>Epistemology</td>
<td>Objective; dispassionate. Detached observer of truth</td>
<td>Subjective, i.e., values and knowledge emerge from the researcher-participant interaction</td>
<td>Knowing through making. Objectively constrained construction within a context. Iterative circumscription reveals meaning</td>
</tr>
<tr>
<td>Methodology</td>
<td>Observation, quantitative, statistical</td>
<td>Participation; qualitative, hermeneutical, dialectical</td>
<td>Developmental. Measure artefactual impacts on the composite system</td>
</tr>
<tr>
<td>Axiology: what is of value</td>
<td>Truth: universal and beautiful, prediction</td>
<td>Understanding: situated and description</td>
<td>Control, creation; progress (i.e. improvement); understanding</td>
</tr>
</tbody>
</table>

3.2 Justification of the chosen methodology

The validity of design science in IS was established through the work of (March and Smith 1995), who developed a framework with which to present the relationships, activities and outputs of design and natural science research. Its research paradigm makes a significant contribution by understanding the “wicked problem” domain, as well as its solutions, with engagement from complementary research cycles such as behavioural science, which are required in the building and application of useful intellectual artefacts. In particular, Hevner et al. (2004) argue that design science is especially concerned with problems “...characterised by unstable requirements and constraints based upon ill-defined environmental
contexts; complex interactions among subcomponents of the problem and its solution; inherent flexibility to change design processes as well as design artefacts (i.e., malleable processes and artefacts); a critical dependence upon human cognitive abilities (e.g., creativity) to produce effective solutions; and a critical dependence upon human social abilities (e.g., teamwork) to produce effective solutions.'

The research problem of this thesis is considered as a “wicked problem” because it is ill-defined, since financial markets have evolved significantly over several hundred years and their structure has changed. These changes have affected the number and type of markets and financial products, as well as their operations and rules. In particular, unregulated markets still suffer from frequent fraud attempts because they are relatively small and illiquid, have lower disclosure requirements for listed firms, and are subject to a less stringent regulatory framework. Thus, design research is the appropriate methodology with which to deal with this problem because the nature of the problem is constantly evolving due to the rapid changes in financial business models, manipulation schemes, technology and regulations.

From another perspective, design research is expressly suited to problems that involve analysis of the use and performance of designed artefacts to comprehend, describe and improve information management (Vaishnavi and Kuechler 2004/5). The unregulated market structure experiences information management problems represented by a lack of data granularity, inaccurate trading information, and other issues such as conflicts of interest, business requirements, duplication in reporting, and transaction reporting requirements which are subject to market inadequacy.

Therefore, it is essential to propose an appropriate solution with which to collect and audit information, and analyse its effect on the efficiency and integrity of trading within the market to improve the stability, transparency and oversight of the market. This thesis proposes financial ontology for fraud purposes to manage vast quantities of financial data efficiently, provide better communication and knowledge sharing among analysts, provide a mechanism to demonstrate knowledge of the processes of financial fraud, understand and share various manipulative practices, manage relevant facts gathered for case investigations, provide early detection techniques of fraudulent activities, develop prevention
techniques using BI technology, and allow reuse of these knowledge resources in different financial contexts.

The ontology will act as a semantically rich knowledge base in market monitoring systems that specialize in information management. Thus, the ontology could play a central role between information management and business intelligence analysis components. The ontology will simulate the unregulated market in a machine-processable form, providing an information management layer in the framework of a semantically rich knowledge base that can be used for the semi-automatic interpretation of relevant unstructured resources.

Generally, most of the existing information management systems that use ontologies utilize a flat architecture for ontology management. However, flat architecture ontologies are managed independently, making integration difficult, especially when multiple ontologies are introduced. In order to overcome this problem, this research introduces the multi-layer architecture of financial ontology adapted from the work of (Mikroyannidis and Theodoulidis 2012).

Due to the scarcity of accurate information in the unregulated market, investors rely heavily on unofficial sources such as press releases, forums, spam e-mails. Thus, manipulators can abuse the market and generate profits through various manipulative practices that influence stock prices, either real or artificially, at the expense of other investors, giving rise to information asymmetries.

The solution to the problem still depends on people's cognitive and social abilities to make trading decisions, and regulatory changes, to perform surveillance and detect market manipulations. There are no 'right' or 'wrong' solutions; only 'good enough' solutions could be proposed. The unregulated market requires an ontological framework which supplements current market surveillance and can harmonise market operations, to ensure integrity and transparency within the single market or cross-market.

In this thesis, the proposed ontological approach is not following the standard ontology engineering approach because of its domain specificity. The ontology used the focus group approach as a top-bottom method to interview financial domain experts to analyze the SEC litigation releases. The focus group was seen as an appropriate and advantageous means of building the domain ontology because it allows participants’ perspectives to be revealed through discussion and build financial dictionary for fraud purposes. This ontology should consider
challenges such as the sheer number of data sources (structured and unstructured), resources, expertise, practices, emerging trends and the analytical capabilities (human and technological) required to develop proficient surveillance and enforcement systems.

Therefore, the financial ontology contains a comprehensive concept system that can act as a semantically rich knowledge base for a market monitoring system that could help users understand financial fraud practices and allow the reuse of these knowledge resources in a different financial context. The ontology could help in fraud monitoring or assist open investigation by acting as a knowledge management repository system for the masses of data gathered during financial fraud, as well as sharing manipulation patterns from prosecuted cases with investigators and relevant users.

The proposed ontology is evaluated through specific instantiations and demonstrated through three case studies. The applicability of these case studies will demonstrate the functionality of market monitoring surveillance systems in fraud detection. The three applications are considered as intelligent systems that extract information from unstructured sources based on the financial ontology.

On the empirical side, the thesis presents examples of novel applications of text-mining tools and data-processing components, developing off-line surveillance system that are fully working prototypes which could train the ontology in the most recent manipulation techniques. Each case study will follow the design research methodology to create independent artefacts that solve particular problems.

In Case Study 1, the domain ontology relies on a suitable and coherent corpus (SEC Litigation releases) that represents financial fraud. The emphasis of the text-mining task was to demonstrate the published prosecuted cases in an appropriate knowledge base.

In case study 2, the ontology will be evaluated through the automatic categorization and classifications of stock spam e-mails. In this case, stock spam e-mails have been chosen as a data source because manipulators use it extensively to ‘hype’ and attracts investors to buy the promoted stocks. Furthermore, ‘pump and dump’ scheme are heavily used in the unregulated market to violate securities’ prices. In this context, the proposed financial fraud ontology plays a vital role in providing the underlying framework for the extraction process and capturing information on touted stocks.
In Case study 3, another instantiation of the proposed financial fraud ontology, addresses requirements for proactively identifying fraudulent “pump and dump” scenarios through the analysis of textual information resources. The instantiation analyzes the impact of ‘stock-touting’ spam e-mails and misleading press releases on trading data in the unregulated market.

The evaluation of the proposed ontology will be carried out with reference to a real case from the OTC market which was prosecuted by the SEC. It explains how the financial fraud ontology might be incorporated into existing fraud analysis processes, the extent to which the processes can be automated, the relationships with other types of analysis and the role that the fraud analysts play. Table 3.2 demonstrates the different characteristics of the research problem from design science perspective, using the IS design research framework introduced by (Hevner, March et al. 2004).

Table 3-2 Characteristics of the research adapted from IS Design Research Framework (Hevner, March et al. 2004)

<table>
<thead>
<tr>
<th>Design Science and IS research components</th>
<th>Addressed research</th>
</tr>
</thead>
</table>
| **Environment Context**                 | People: This research considers the different stakeholders affected by the manipulation, such as investors, traders, general public. Furthermore, the research considers fraud analysts to monitor the market behaviours and detect fraudulent patterns for further investigation (as shown in section 1.3.2).  
Organisations: The research studied the structure of unregulated markets such as OTC and pink sheets (as shown in section 2.1).  
Technology: This research evaluates the existing infrastructure, applications, technologies and capabilities of different monitoring and surveillance systems currently operating in some exchanges and regulators (as shown in section 2.4.4, 2.4.5, 2.4.6). |
| **Knowledge Base**                      | Foundations: The research reviewed and evaluated previous relevant research from various disciplines (Finance and economics, IS and BI, market manipulation and fraud detection) that corresponds to the research problems. Existing theories, market monitoring frameworks, and financial ontologies are used as a good foundation upon which to build and develop this research (as shown in section 2.2, 2.3, 2.4.6, 2.5.1).  
Methodologies: The research used different BI techniques such as text mining and data mining to evaluate the quality and effectiveness of the artefact. The validation and measurement criteria of these techniques are reviewed in section 2.4.4.1 and 2.4.4.2. Moreover, the methodology considers the complex nature of the environment in which the solution is developed and the previous knowledge that is available when performing the task. Complexities in the environment include, for example, the rapid developments in manipulative activities and the relevant technologies used to facilitate their dissemination. |
| **Business need**                       | • The opaque environment of unregulated markets and the absence of legal trade data disclosure requirements mean OTC markets are less directly monitored than regulated markets. Furthermore, regulators have a limited ability to detect inappropriate conduct in the market. Therefore, this could develop into a threat and build systematic risk in the OTC market and the same effect could reach other linked regulated markets because of the absence of a harmonised, robust market.  
• These differences in information environment, market structure and investors could create differences and challenges for regulators and market participants in the collection and audit of information, and in the analysis of the effect on the efficiency and integrity of trading. Yet, there are still significant efforts to be made to improve the stability, transparency and oversight of the OTC market (European Commission 2010).  
• In order to guarantee transparency, enhance investor confidence, gain higher market capitalisation and mitigate abuses in these markets, regulators have to develop an ontological framework which supplements current market surveillance in order to harmonise market operations, ensure integrity with other markets, minimise systemic risks, and monitor, detect, and prevent potential market violations within single markets or cross-market. |
Build:
- Build financial fraud ontology for the financial market to gain better understanding of financial fraud practices and provide the information management layer in the MMF with a semantically rich knowledge base that could help the BI analysis components to trigger the alarms.
- Create specific artefacts will be demonstrated through case studies to show the applicability of the ontology to help users in the market monitoring surveillance framework detect fraud cases and identify fraudulent patterns in the stock market.
- In particular, the artefacts demonstrate an automatic and efficient method of using text mining technique to extract key attributes and characteristics of different unstructured sources generated as part of “touting campaigns” within the unregulated market.
- The ontology could add value in proactive fraud monitoring or help in open investigation by acting as a knowledge management repository system that manages and controls the masses of data that can be gathered during financial fraud and share existing manipulation patterns from prosecuted cases among investigators and relevant users.

Evaluate
- Analytical procedures (precision and recall method is used to evaluate the text mining extraction, ROC lift to evaluate the data mining techniques)
- Case studies (SEC manipulation cases)
- Prototypes to demonstrate the functionalities and capabilities of the text mining artefact, framework and the ontology. Instantiations demonstrate how the prototypes could be used by users or independent fraud detection service providers to raise proactive alarms as to a manipulation scheme in the unregulated market.

3.3 Methodology application

This thesis adapts the ‘IS Systems Research framework’ introduced by (Hevner, March et al. 2004) and ‘Design Science Methodology’ as shown in figure 3-4 (Vaishnavi and Kuechler 2004/5). As such, chapters and sections can be mapped directly to that framework and that methodology. In particular, Chapter 1 extensively describes the environment, business need and different characteristics of the research problem ‘wicked problem’. Parts of Chapter 2 continue discussion of environment; specifically the organisation, represented by OTC market, and the technology used in existing market monitoring services. Chapter 2 also covers the knowledge base by reviewing and analysing previous work from other relevant disciplines including empirical findings, and definitions and types of market manipulation. Furthermore, the chapter reviews the existing theories, market monitoring frameworks, fraud detection technologies, and financial ontologies which constitute a solid foundation on which to build and develop this research. Chapter 3 describes and justifies the methodology adapted in this research and reviews another important part of the knowledge base. Chapter 4 demonstrates the design of artefact, namely financial ontology. In particular, it exemplifies the design of the financial ontology; Constructs are also discussed to show the architecture of the ontology including the main classes and relationships between concepts.

Part of Chapter 4, Chapter 5, and Chapter 6 present individual instantiations of the artefact by means of case studies. Each case study demonstrates the functionalities and capabilities of the artefact using real data (unstructured and
structured) from the OTC Market and other relevant sources. In addition, the ontology instantiations show how the prototypes could be used by regulators or independent fraud detection service providers to raise proactive alarms about information-based manipulation schemes. Evaluation runs case by case, judging the utility of the proposed ontology for increasing understanding of a part or the whole of the problem and the use of the ontology model demonstrated in Chapter 4.

Chapter 7 discusses the role of the financial fraud ontology in the fraud detection context and presents a summary of the three instantiations. Furthermore, Chapter 7 summarises how each case study develops from independent artefacts that solve the problems that are introduced in that specific case. Conclusions and results are exposed on two levels. At the lower level, each case study is presented with its own results and conclusions, and knowledge gained is abstracted to be integrated later into a final Chapter 8 'conclusions and lessons learnt' section following the circumscription and operation and goal knowledge loops.

Overall, all eight chapters consider the business need of this research with a rigorous review of the applicable knowledge as represented by respective arrows in figure 3-3. According to (Gregor 2006) this thesis should be classified under Type V: Theory for Design and Action which is “how to do something. It is about the principles of form and function, methods, and justificatory theoretical knowledge that are used in the development of IS”. The reason for this is that the thesis provides a reasonable description of the phenomena of market manipulation in unregulated markets and the role of market surveillance in protecting the market from different types of fraudulent activity, particularly information based schemes. One of the features of ‘Type V’ theory is going beyond basic description to analyse outstanding attributes of phenomena and the relationships among the phenomena. The proposed financial ontology could be considered to fulfil that role as it may help provide better understanding of financial fraud practices in the financial market with an attempt to provide a clear definition of the main classes of this ontology and describe the relationships between the concepts and classes identified. To assess the knowledge contribution, the artefact is assessed through case study instantiations which act as an evaluation scheme of the ontology and framework to the degree required to satisfy criteria such as completeness of
models and methods, their simplicity, consistency, ease of use, interestingness as a valid claim, and the quality of results obtained through use of the method.

Figure 3-3 Research Mapping adapted Design Research. Cited from (Vaishnavi and Kuechler 2004/5)
4 Financial ontology for fraud purposes

Much work has been done in building ontologies as a way of representing and capturing knowledge. Simply, ontology can be defined as a set of concepts concerning some domain of human knowledge and a specification of the relationships between these concepts that enables computers to process the content. With the existence of so much information that could be turned into knowledge, ontologies are constructed to acquire, share, reuse, and maintain knowledge. Various studies have demonstrated “knowledge elicitation” techniques which are used especially for the creation of expert systems. These techniques include interviews, focus groups, process tracing and conceptual methods that are used for ontology learning (Brewster 2008).

Communities such as finance have access to vast repositories and enormous volumes of unstructured and semi-structured texts from various sources. In practice, manual techniques are challenging and increasingly unviable in satisfying the required analysis tasks. Moreover, financial knowledge is rapidly evolving, making it difficult to keep any knowledge base up to date. Since ontology represents the knowledge of the domain, and humans can communicate their knowledge successfully through textual sources, many studies argue that it is possible to construct an ontology for a specific domain based on processing a large collection of text (corpus) (Brewster 2008).

Given that there is a strong argument that it is beneficial to re-use an existing ontology or taxonomies for a domain if one is available (Brewster 2008), construction of the domain ontology used the main classes of the taxonomies proposed by (Diaz 2011).

Diaz (2011) constructed a collection of taxonomies in an attempt to build a large set of concepts and definitions commonly used in financial markets. Furthermore, these taxonomies followed a ‘top-down’ approach based on complete literature reviews undertaken to construct concepts that could explain a common understanding of market monitoring and surveillance systems. These concepts have been organized into three major categories: taxonomy of basic elements and entities present in different manipulation scenarios; taxonomy of the interacting agents in financial market systems; and taxonomy of the sub-systems, components and information management concepts of a model detection engine. The first two categories’ concepts are only used as a first version to build the
domain ontology. These categories demonstrate the market manipulation scenario in eight main concepts as shown in figure 4-1. The taxonomies were separated and they have in total 274 concepts.

![Diagram of domain ontology]

**Figure 4-1 A taxonomy of elements in the manipulation scenario (Diaz 2011)**

The ‘Agent’ concept represents the fraudster that performs a manipulation by taking an action. The agent’s ‘Action’ could be classified as information-based (related to the release of false information or rumours to tempt investors and mislead their trading decisions); trade-based (occurs when traders attempt to manipulate a stock by buying and selling stocks, misleading the markets by moving the price artificially to levels that are convenient to their purpose); or action-based (changing the actual or perceived value of the assets).

Furthermore, the ‘Venues’ concept describes where investors and manipulators organize their trade manipulations. Agents could target any type of asset, such as stocks, penny stocks, money markets, bonds, derivatives, or structured notes. The taxonomy included a temporal dimension to describe the ‘time’ of manipulation.

The consequences of the manipulation are described in the ‘Effects’ concept to show the patterns that could be identified in the behaviour of the agents, or in trading orders and executions. The ‘Market Manipulation Types’ concept describes the different manipulation types existing in the financial market.

At the second level of the ontology, manipulation types are classified into four main classes: ‘Abuse of market power’, ‘Contract-based’, ‘Breach of fiduciary duty’ and ‘Runs and raids’. However, the work of Diaz (2011) helped to build an organized body of knowledge from the literature; the study has not been extended to construct a full ontology for the domain.
Thus, this research contributes to building a financial ontology for fraud purposes with general classes that contain a large set of financial concepts and definitions. Further, the domain ontology relies on a suitable and coherent corpus (SEC Litigation releases) relevant to financial fraud to propose additions and alterations to the existing taxonomies. Also, the class hierarchy has been developed using a combination of ‘top-down’ and ‘bottom-up’ approaches.

In practice, the role of domain experts are to identify the knowledge that lies in the texts to construct an ontology with relevant financial fraud concepts; the ontology will be able to answer questions similar to those asked by users and analysts reading the texts themselves. Furthermore, they help to provide a certain degree of depth to the organization of the financial concepts within the ontology.

Some key questions that the ontology should answer include: Who is (are) the agent(s) involved in the manipulation? Which asset is being targeted? In which venue is the manipulation taking place? Which action has been performed or is planned? Is it a trade-based or an information-based action? Which patterns are associated with this manipulation? When was this manipulative action performed? Where is the manipulator getting his profit from?

The ontology class hierarchy has been developed using a combination of top-down and bottom-up approaches. Each class includes general descriptions, links to different sources including linked data and instances as particular exemplars to fill the slot values of the class. Moreover, all sub-classes inherit the slots/properties of their super-class. To a certain extent, the challenge lies in finding ways to identify the types of relationship between concepts and to label these relations. The whole process of building ontology was iterative in order to produce a good enough version that could help building a common understanding of financial market fraud.

4.1 The process of constructing the financial domain ontology

This section introduces a proposed method to construct a financial ontology for fraud purposes directly from a corpus of finance-related text or documents as shown in table 4-1. Ideally, with sufficient exemplars of financial concepts identified by financial experts participating in this research, the ontology can be used to construct an accurate representation of its usage and meaning for users or analysts. The financial domain experts acted as coders and annotated interesting
financial concepts from the litigation releases which are a good corpus that could help to build the domain ontology.

The research used the focus group as a method of interviewing financial domain experts to analyze SEC litigation releases. The focus group was seen as an appropriate and advantageous means of building the domain ontology because it allows participants’ perspectives to be revealed through discussion, with questions and arguments that are different from individual interviews, for example. The focus group aimed to identify interesting financial concepts lying in the SEC litigation releases to learn the domain ontology with relevant financial fraud concepts which will enable the domain ontology to answer similar questions that could be asked by ontology users who read those releases.

Furthermore, the participants helped to provide a certain degree of depth to the organization of the financial concepts within the ontology. In particular, they helped to build a dictionary of financial concepts for fraud purposes and map these concepts to the relevant ontology classes. Moreover, they were allowed to add new classes to the ontology and suggest better naming for the existing classes. The output of the focus group was to produce a satisfactory version of the domain ontology based on the analyzed SEC litigation releases.

This research uses the notion introduced by Livingstone and Lunt (1994:181) called the theoretical saturation criterion which means: once your major analytical categories have been saturated, it would be appropriate to bring data collection to halt. The research selected number of litigation releases that should be analyzed by the participants. The four litigation releases have been selected based on the four main manipulation types; these are ‘insider trading’, ‘pump and dump’, ‘front running’, and ‘marking the close’.

The total size of the focus group was ten participants, as suggested by Morgan (1998a). However, Morgan recommended using smaller groups when participants are likely have a lot to contribute and when the topics are complex and controversial. Thus, this focus group was held in two rounds. The first round had two participants to analyze one of the releases (a pump and dump case) as a pilot and evaluation of the process. The second round had eight participants, divided into four groups of two; each pair shared and analyzed one litigation release.

All the participants have a finance background, studying for either a Master’s or a PhD in finance. Two facilitators were involved in the focus group to guide the
session by presenting a general overview of the process and the topic. The facilitators also played an important role in identifying differences of opinion and exploring with participants any ambiguous concepts they did not understand. However, the facilitators were careful about how far their involvement should go, in order not to influence the conduct of the focus group. Table 4-1 shows the focus group agenda for the two rounds. The objectives of the focus group were to:

1. Analyze four SEC litigation releases.
2. Identify interesting financial concepts from this corpus.
3. Create a financial dictionary for fraud purposes based on the SEC litigation releases.
4. Discover the ways in which the participants discuss the identified financial concepts from the corpus, and how they map these concepts to the proposed ontology classes with the possibility of adding new classes if necessary and suggesting better naming for the existing classes.
5. Produce a satisfactory version of the ontology based on the analyzed cases.
### Table 4-1 Focus Group Rounds Process Overview

<table>
<thead>
<tr>
<th>Items</th>
<th>1st round</th>
<th>2nd Round</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Number of facilitators</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Number of Analyzed Cases</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Manipulation type</td>
<td>Pump and dump</td>
<td>Pump and dump, insider trading, marking the close, front running</td>
</tr>
<tr>
<td>Length of the case</td>
<td>34 pages</td>
<td>By order: 35, 18, 37, 14 pages</td>
</tr>
<tr>
<td>Litigation Release Number</td>
<td>21024</td>
<td>By order: 21024, 19929, 20899, 21446</td>
</tr>
<tr>
<td>Number of sessions</td>
<td>1 session</td>
<td>Break the focus group into 2 sessions</td>
</tr>
</tbody>
</table>

#### Session Agenda

- A presentation demonstrates the research and focus group objectives
- Focus Group Tasks
- Some guidelines about the coding and mapping to the ontology classes

#### Session Tasks

- Read the given document (SEC Litigation release) to have an overview about the document structure
- Annotate interesting concepts
- Discuss with your colleague the annotated concept
- Map the annotated concepts to ontology classes
- Adding new ontology classes if necessary or providing a better naming for the existing ontology concepts
- Load the annotated concepts and ontology mapping to the NVIVO 9:0

#### Session Duration

- 5 hours
- Session 1: 20 minutes
- Session 2: average 4.30 hours

#### Session objectives

- Trial and evaluating the process of the focus group

#### Tape recorded

- Yes

#### Case Analysis completion

- No
4.2 Financial Ontology Methodology analysis

This research used the constant comparative method as a methodology to analyse the focus group data. The method was introduced by (Glaser and Strauss 1967) and expanded by (Lincoln and Cuba 1985). Originally, the method was developed for use in grounded theory methodology, but now is widely used to analyse qualitative data gathered from different media such as interviews, focus groups, examination of documents, and experiments. In particular, the methodology explains the process of organizing data such as words, sentences, and paragraphs into either existing or unidentified categories (Grove 1988). This thesis adapted the methodology process guidelines suggested by (Lincoln and Cuba 1985) to analyze the focus group data and build the domain ontology. As shown in figure 4-2, the methodology suggests four distinct stages: Unitizing, in which researcher identifies meaningful pieces of information such as words, sentences or paragraphs in the data and codes these sets of information with a description about their source, the type of participants, etc. Categorizing groups similar pieces of information together; new categories may emerge if the existing categories are not suitable for particular pieces of information. Categories are labelled with a suitable name to express their meaning, and provisional rules for inclusion are established. This process may require some revision of relevant information units, provisional rules and category names in order to avoid overlapping, incomplete or missing categories. At this stage the researcher should determine any possible relationships between or among categories. The Filing in patterns stage aims to flesh out the categories after extensive data collection and unitizing; the researcher checks that all incidents are classified into relevant categories and that these categories have become saturated. The Member checking stage allows the researcher to reconstruct the data to check whether it is a reasonable representation of reality. As a further check, the whole process can be evaluated by an external auditor for validation (Grove 1988). This research uses the four stages to analyze the focus group data in order to build the domain ontology with a good representation of financial reality.
4.2.1 Stage 1: Unitizing

4.2.1.1 1st round analysis

This stage analyzes the process of annotation by the participants. As discussed above, the annotation process and the ontology mapping process were carried out in the first round in a single session. In this round, the participants analysed the ‘pump and dump’ case. The annotation process was not completed in the first round because of the time restrictions and the fact it was a trial session intended to check the process; thus, it was not necessary to ask participants to annotate the whole document. The facilitator gave some instructions regarding the coding and the annotation process with examples from the given litigation release. To some extent, the participants followed the instructions concerning annotating concepts as either uni-terms or multi-terms. Both participants in the first round made their annotations directly in the NVIVO, which helped the researchers to analyse the process. In total, participant (A) annotated 99 concepts and participant (B) annotated 78 concepts.

Generally, both participants tried to identify concepts to describe the litigation release. For example, both were concerned about annotating the names of agents who were the defendants in the litigation release. Furthermore, they both considered ‘friends, family, and persons’ as part of an agent’s group as a third party needing to be annotated, because they helped the agents to manipulate the market by showing the existence of investors who would buy the shares. However, they are generally not prosecuted because it is hard to prove their intentions. Also,
the participants were concerned about whether concepts implied any suspicious descriptions of the role of the agent in his company, such as ‘a sole partner and owner’, ‘sole associates’, or ‘a sole owner and employee’.

Participants identified concepts describing the manipulation actions carried out by the manipulators to violate the stock and pump the prices to attract investors, such as ‘material false, misleading statements, issuing a series of press releases that contains baseless projections of future revenue, misleading overstatements of business relationships and product development, and identified contracts for future business that did not exist, opinion mill, dump more than two million restricted shares of Mobile Ready for approximately $70,000 in profits’, issuing shares of Complete Identity to themselves and to friends, family, and persons’, ‘sold more than 2 million restricted shares of Mobile ready’. Despite the fact that these concepts vary in terms of manipulation action classifications, the participants considered them as concepts that should be included in the ontology to train this layer with financial domain patterns and actions, the better to describe the different methods applied by manipulators.

Interestingly, one of the participants considered the effect of the legal registration process as a pattern for manipulation. For example, companies can issue shares without complete registration; however, those shares are restricted, not freely tradable and cannot be sold on any stock market. They are, therefore, of limited value to those who hold them until the restrictions are removed. After a certain amount of time, the restrictions can be removed if an attorney issues an opinion letter stating that all the legal requirements for removing the restrictions have been met. Also, other terms related to the asset, the venue of the manipulation represented in the trading or exchange market, and the timeline of the manipulation were considered by the participants for annotation.

4.2.1.2 Lessons learnt from the 1st round

The facilitator observed that the speed of the annotation process varied from one participant to the other. Furthermore, the facilitator noticed that the participants were not consistent in some parts of their annotation. For example, as shown in figure 4-3, in annotating agents’ names the sometimes considered aliases, as “Sandra B. Masino ("Masino")” and sometimes not, as “Albert J.Rash ("Rash") or
Kathleen R. Novinger ("Novinger"). The facilitator considered this inconsistency issue and in the second-round participants were made aware of the problem. Another observation was that neither participant paid attention to the specific terms that needed to be annotated as they were annotating a whole sentence. As a general observation, neither participant was interested in the legal side of the litigation release, perhaps because of their backgrounds. For example, one participant was interested in highlighting the different laws and acts stated in the litigation release to accuse the defendants of the different violations. Another observation: there were many terms of the participants had no knowledge, so the facilitator explained the terms and allowed them to use the Internet to discover the meaning from sources such as www.investopedia.com. The facilitator notified the second-round participants to consider this in their annotation process.

4.2.1.3 2nd round Analysis

In the second round, the focus group of eight participants analysed four cases; each litigation release was shared between two participants. Firstly, the participants read the cases and annotated the interesting concepts in their own time. At the second meeting, each pair of participants discussed their case and the annotated concepts to agree which concepts should be recorded in NVIVO and later mapped to the ontology classes. From the different cases, these annotated concepts will be used to build a financial dictionary for fraud purposes. Table 4-2 shows the number of concepts annotated and agreed per case. In this round, the participants considered the ontology classes and annotated the concepts that could describe the case and the patterns that could help the ontology to explain similar manipulation types in the future. The number of concepts varies for many reasons, such as the size of the case, different views between participants in analyzing the case, the nature of the manipulation scheme, and the information provided for each case. An analysis of
each case will be discussed to highlight the differences in annotation among the participants.

<table>
<thead>
<tr>
<th>Case</th>
<th>Participant</th>
<th>Total number of annotated concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pump and Dump</td>
<td>C,D</td>
<td>144</td>
</tr>
<tr>
<td>Insider Trading</td>
<td>E,F</td>
<td>111</td>
</tr>
<tr>
<td>Marking the Close</td>
<td>G,H</td>
<td>129</td>
</tr>
<tr>
<td>Front Running</td>
<td>I,J</td>
<td>87</td>
</tr>
</tbody>
</table>

**1. Pump and Dump Case**

In the Pump and Dump case, participants C and D largely agreed with the annotations from the first-round participants. However, they considered, for example, the financial laws and regulations referred to by courts in the manipulators’ cases. Most of the laws, e.g. Securities Act Rules 144 and 504 with specific sections such as section 5(a), 5(c) and 17(a) were annotated in the belief that this would help the ontology users to search for the kinds of laws and regulations breached by particular types of manipulation. The mapping between ontology layers could help users to understand the meaning and description of these rules and regulations. In particular, the lexical ontology layer has a full description of these laws and regulations. Thus, the relationship between concepts could illustrate in detail the financial concepts included in the different layers. Furthermore, the participants were eager to annotate the description of agents, either through individual profiles such as the agent’s name, e.g. ‘Sandra B. Masino (“Masino”)’, taking into consideration the advised right annotation, the agent’s address, e.g. ‘resident of Costa Mesa, California’, and the agent’s title, e.g. ‘owner of the law firm’. Furthermore, both participants considered the description of the corporation involved in the case such as ‘Mobile Ready is a non-reporting, publicly traded company’, ‘Market software applications for mobile devices’. Overall, participants C and D annotated 144 concepts.

**2. Insider Trading case**

In the Insider Trading case, participants E and F were more concerned to identify not only the manipulators but also other participants involved in the manipulation, such as intermediaries who facilitated the process or victims who suffered from the scheme. For example, three banks were identified as victims: the ‘Sun Country Bank’, ‘Valencia Bank’ and ‘Monetary Bay Bank’ with ‘compensation and benefits
consulting firm "C&B consulting firm" as an intermediary who helped the agent, 'Bank owned life insurance BOLI sales representative Robert J. Gallivan ("Gallivan").

Furthermore, participants noted that the agent had been involved in different manipulation cases and had been suspended from the Commission for 12 months, prohibiting work with any brokers, dealers, or investment. Thus, they annotated all the related background concepts about the agent’s history to help the ontology users avoid dealing with this agent in the future.

In this case, the manipulator used non-public material to tip friends and relatives to trade on the basis of his position in BOLI, which is illegal conduct. Thus, all concepts related to the evidence and patterns, including the legal one, that show how the agent acquired the information and did the manipulation were considered by the participants. Interestingly, the participants decided to highlight the timeline of the manipulation before, during and after the manipulation. Overall, participants E and F annotated 111 concepts.

3. Marking the Close

Participants G and H annotated 129 concepts from the Marking the Close case. They were concerned to identify the evidence and facts about patterns that could be used in the ontology to identify potential wash sale, matched order and marking the close manipulation cases. In particular, structured evidence was more dominant in this case as participants annotated how manipulators artificially inflate the stock prices to gain profit, which then decline again. Annotations included ‘artificially inflating Avicena’s stock price or to create false appearance of an active and liquid market’, ‘this closing price represented an almost 100 percent from previous day’s close’, ‘Georgiou’s manipulation scheme started losing momentum’, and ‘declined from $5.40 per share on June 5th to $0.88 on November 6, 2007’.

In this case, the agent used the help of other nominees and compensated them for allowing him to dictate the trading. Thus, the participants tried to annotate concepts to show the means of manipulation and the different beneficiaries from the scheme, e.g. ‘using numerous nominee accounts held at offshore broker dealers’, ‘Georgiou directly paid nominees’, and ‘he allowed them to sell shares at specific time in order to profit in exchange for entering trades at Georgiou’s direction’.
Interestingly, the participants differentiated between the methods used by the agent, either to provide clients with false or misleading information or conceal information about his association with nominee accounts, e.g. ‘Georgiou attempted to conceal his association with the nominee account’, ‘he avoided US financial institutions and moved cash in and out of his foreign account via wire transfers to minimize any paper trail’, and ‘Georgiou never disclosed that the value of these securities was artificially inflated as a result of his manipulative activities’. Also, the participants considered that the agent’s different ways of communication in directing his manipulative trades were worthy of annotation, such as ‘Georgiou often directed the manipulative trades only by cell phone or sent and received messages by way of BlackBerry PIN’, and ‘he often used email accounts in other’s people names to disguise his identity’.

4. Front Running Case

Participants I and J analyzed the Front Running case and annotated 87 concepts. They were concerned with describing the evidence and patterns used by the agent to manipulate the trading system. For example, they annotated ‘Vianna improperly misused his access to Maxim’s order management system to divert customer A’s profitable trade to Creswell and Creswell’s unprofitable trade to Customer A’. Furthermore, they annotated the number of times the agent performed the manipulation action, e.g. ‘At least 57 times between July 2007 and March 2008, Vianna simultaneously entered orders in the accounts of customer A and Creswell to trade the same amounts of the same stock’. Also, the case was not classified by the court as front-running research, so the participants annotated equivalent concepts to describe the manipulation scheme, such as ‘This case arises from a fraudulent scheme orchestrated by Vianna a former broker in a New York brokerage firm, to divert dozens of profitable trade from one of his customers, “Customer A”, to another of his customers “Creswell”.

The participants were concerned to describe fully the orders handled by the agent, e.g. ‘Vianna entered a new order to sell 60,000 of NVLS for Creswell’s account’. In terms of benefits from the scheme, the participants annotated related concepts such as ‘Vianna received ill-gotten gains through the fraudulent scheme, including at least $125,000 in commission paid to him with respect to the corresponding trades in customer A’s and Creswell’s accounts’.
4.2.2 Stage 2: Categorizing

The participants were asked to map the annotated concepts to the ontology classes. The facilitator presented the different classes and their definitions and description, as well as hard copy of the ontology with a descriptive note for each class. In this stage, the participants mapped the annotated concepts to suitable classes in the ontology, considering adding new classes if necessary or suggesting better names for existing classes.

Thus, to an extent the ontology uses the annotated concepts to build a financial dictionary for fraud purposes and answer some key questions about the litigation releases such as: Who is(are) the agent(s) involved in the manipulation? Which asset is being targeted? In which venue is the manipulation taking place? Which action has been performed or planned? Is it a trade-based or an information-based action? Which patterns are associated with this manipulation? When was this manipulative action performed? Where is the manipulator getting his profit from?

1. Pump and Dump Case

Table 4-3 shows the categorization analysis of the Pump and Dump case. The participants classified the ‘pump and dump’ concept under ‘Market manipulation\Information-based \Runs and Raids\hype and dump’. They categorized ‘Mobile ready Entertainment Corp.’ and ‘Complete Identity, Inc’ under ‘Asset/Financial /Equity/Pink Sheets’. They were not sure whether ‘Complete Identity’ was a public company until they found evidence in the case, e.g. ‘Mora had been working at Complete Identity, INC. (“Complete Identity”), a corporate branding services provider that they took public through a reverse merger with an existing Pink Sheets shell company in August 2005’.

The analysis shows that the ‘Actions (What)’ class had the highest percentage, 6.36%. The participants agreed that the case had evidence that could be classified as ‘information-based’, and ‘trade-based’. However, the first-round participants identified more concepts that could be categorized under ‘action-based’, e.g. “The fraud includes all of them [:00:32:0]…It seems they did everything” [2:53]. The ‘Information-based\Misleading information’ class had 1.73% of concepts to describe how the agent disseminated misleading press releases and statements to violate the stock prices, such as ‘made materially false and misleading statements’, ‘misrepresented critical facts with no reasonable basis’, ‘issued materially false and misleading press releases’, ‘issuing a series of press releases’,
that contains baseless projections of future revenue, misleading overstatements of business relationships and product development, and identified contracts for future business that did not exist' and 'never accurately disclosed the nature of its business and the products and services it offered'.

The manipulator also disseminated false information to the investors, e.g. ‘identified contracts for future business that did not exist’, and ‘issued materially false and misleading press releases’, and the participants classified 0.99% of these concepts as ‘Information-Based\False Information’. The case showed that manipulators used a company called 144opinions to manage their legal letters. However, this company helped the manipulator to fraudulently facilitate the sale of securities in violation of the registration provisions of the federal securities laws. The participants therefore suggested adding classes called ‘Opinion Mill’, and ‘Legal Opinion letters’ to describe any legal violation evidence [03:17 – 04:39].

In this case, the manipulator hyped the stock then sold the shares to gain profits. Thus, in the ‘Trade-based’ class, the participants mapped concepts such as ‘dump more than two million restricted shares of Mobile Ready for approximately $70,000 in profits’, ‘issuing shares of Complete Identity to themselves and to friends, family, and persons’, and ‘sold more than 2 million restricted shares of Mobile ready’ to the ‘Trade-based\Trade\Sell’ class, 0.88%.

They also identified ‘never registered any of the offering’ as a trade-based concept, although they claimed that it better matched the legal and administrative part of share issuance. Thus, they suggested adding a class called ‘Other related trade Activities’ with sub-classes ‘non-registered offering’ and ‘Registered offering’. For example, they classified ‘never registered any of the offering’, and ‘Restricted Stock Service’ to the ‘non-registered offering’ [25:42 - 31:11].

As with insider trading and front running, the participants suggested adding a class to describe the means of manipulation through the ‘means’ class, but in this case they suggested ‘regulation specific actions’ with 1.57% of the annotated concepts. They also added a ‘Financial regulation’ class with 1.88% concepts to include all the laws and regulations used by courts against manipulators [8:41- 9:00].

The Agent (Who) class had 1.28% of the concepts that described the manipulators. However, the participants were concerned to add classes describing their profiles and characteristics. The participants of the first round classified ‘Michael H.Magolnick’ and ‘Craig A.Mora’ who were the CEOs of Mobile Ready
under the ‘membership\Insider’ class. However, ‘Sandra B.Masino’, ‘Albert J.Rash’, ‘Kathleen R. Novinger’, and ‘friends, family, and persons who provided start-up capital’ were classified under ‘Membership/outsider’. Also, ‘friends, family, and persons who provided start-up capital’ were classified as ‘Benefit\Third Party’. The second-round participants considered classifying all manipulators’ names (0.23%) under ‘membership\Insider’. Based on the case, the first-round classification was correct, because the outsiders were helping the insider manipulators to violate the stock prices. The participants agreed to add a class called ‘characteristics’ with two sub-classes: ‘individual’ which had 0.69% of concepts describing the agent’s profile such as ‘resident of Costa Mesa, California’, and ‘sole partner and owner of the Law Firm of Albert 1 Rasch and Associates ("Rasch and Associates"); and ‘corporation’, with 0.29% of concepts, describing company profiles [1:33:50 -1:34:39] such as ‘market software applications for mobile devices’, and ‘non-reporting, publicly-traded company’. The Effects (consequences) class had 1.00% of the annotated concepts describing the different evidence and patterns identified by the participants. In the first round, concepts such as ‘dump more than two million restricted shares of Mobile Ready for approximately $70,000 in profits’, ‘daily trading volume increased by more than 500% and Mobile Ready’s per share price increased by more than 50%’, ‘Magolnick and Mora each individually sold more than 2 million restricted shares of Mobile Ready’, and ‘while their friends, family and early investors were responsible for the illegal sale of more than 18 million restricted shares into public market’ were categorized under ‘Structured (Effects (Consequences)/Evidence and Patterns/Financial Economic)’. For ‘Unstructured’ (Effects/(Consequences)/Evidence and Patterns/Financial Economic), they identified ‘Magolnic and Mora began issuing a series of press releases that contained baseless projections of future revenue, misleading overstatements of business relationships and product development, and identified contracts for future business that did not exist’. The participants noted the effect of the legal registration process, which was a pattern for manipulation. They discussed how companies could issue shares without complete registration; however, those shares were restricted and not freely tradable, and they could not be sold on any stock market. They are, therefore, of limited value to those who hold them until the restrictions are removed. After a certain period of time, the
restrictions can be removed if an attorney issues an opinion letter stating that all the legal requirements for removing the restrictions have been met. According to the SEC, the defendants in this case issued such letters even though the prerequisites for removal of the restrictions had not been satisfied. Thus, the participants suggested adding a class called ‘Evidence and patterns\Legal’ with 0.69% patterns [01:09:17- 01:09:41]. For example, under ‘Legal’ patterns were ‘this opinion is based on, but not limited to Information obtained from the Securities and Exchanges [SIC] Electronic Data Gathering and Retrieval System’, ‘Legal Opinion Mill’, and ‘Disgorgement of illegally obtained funds together with prejudgment interest’.

The ‘Financial Market stakeholders’ class had 0.54% of concepts describing different stakeholders: ‘investors’ ‘shareholders seeking to obtain exemptions under Rule 144’, and ‘market’ such as ‘over-the-counter securities market’, ‘Pink OTC Markets, Inc. (the "Pink Sheets")’, ‘Regulator’ and ‘Commission’.

Regarding the Time (When), the participants commented that they found specific periods mentioned in the manipulation. They suggested adding a class called ‘Period’. Also, they suggested adding a class related to ‘Market Cycle’ with subclasses ‘Bull’ and ‘Bear’ to describe upward and downward financial market trends [1:16:34 - 1:16:51]. The participants categorized 0.06% of the concepts to ‘Time of year’, e.g. during 2007, Masino, who is not an attorney, and her company 144 Opinions, prepared, and Rasch and Novinger, both attorneys’.
Table 4-3 Pump and Dump Categorization Analysis

<table>
<thead>
<tr>
<th>Main Classes</th>
<th>Node</th>
<th>Percentage coverage</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions (What)</td>
<td>Nodes[Financial Ontology]/FinancialFraud[Actions (What)]/Action-Based[Regulation specific actions]</td>
<td>1.57%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Actions (What)]/Information-Based[False Information]</td>
<td>0.99%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Actions (What)]/Information-Based[Legal Opinion Letters]</td>
<td>1.05%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Actions (What)]/Information-Based[Misleading Information]</td>
<td>1.73%</td>
<td>6.36%</td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Actions (What)]/Information-Based[Opinion Mill]</td>
<td>0.04%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Actions (What)]/Trade-Based[Other Related Trade Activity/Non-registered Offerings]</td>
<td>0.08%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Actions (What)]/Trade-Based[Trade]/sell]</td>
<td>0.88%</td>
<td></td>
</tr>
<tr>
<td>Financial</td>
<td>Nodes[Financial Ontology]/FinancialFraud[Financial Regulation]</td>
<td>1.88%</td>
<td></td>
</tr>
<tr>
<td>Regulation</td>
<td>Nodes[Financial Ontology]/FinancialFraud[Agent (Who)]/Benefit[Third party]</td>
<td>0.07%</td>
<td></td>
</tr>
<tr>
<td>Agent (Who)</td>
<td>Nodes[Financial Ontology]/FinancialFraud[Agent (Who)]/Characteristics[Corporation]</td>
<td>0.29%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Agent (Who)]/Characteristics[Individual]</td>
<td>0.69%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Agent (Who)]/Membership[Insider]</td>
<td>0.23%</td>
<td></td>
</tr>
<tr>
<td>Effects (Consequences)</td>
<td>Nodes[Financial Ontology]/FinancialFraud[Effects (Consequences)]/Benefit[Direct]</td>
<td>0.24%</td>
<td>1.00%</td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Effects (Consequences)]/Benefit[Indirect]</td>
<td>0.07%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Effects (Consequences)]/Evidence and Patterns[Legal]</td>
<td>0.69%</td>
<td></td>
</tr>
<tr>
<td>Financial Market Stakeholders</td>
<td>Nodes[Financial Ontology]/FinancialFraud[Financial Market Stakeholders]/Investors]</td>
<td>0.07%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Financial Market Stakeholders]/Market]</td>
<td>0.22%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Financial Market Stakeholders]/Regulators]</td>
<td>0.04%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes[Financial Ontology]/FinancialFraud[Financial Market Stakeholders]/Support Organisations]/Services[Financial]Data Provision[Structured]</td>
<td>0.21%</td>
<td></td>
</tr>
<tr>
<td>Time (When)</td>
<td>Nodes[Financial Ontology]/FinancialFraud[Time (When)]/Time of Year]</td>
<td>0.06%</td>
<td>0.06%</td>
</tr>
<tr>
<td>Asset (Target)</td>
<td>Nodes[Financial Ontology]/FinancialFraud[Asset (Target)]/FinancialEquity|Pink sheet]</td>
<td>0.07%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Market manipulation (Types)</td>
<td>Nodes[Financial Ontology]/Marketmanipulation[types]/</td>
<td>Runs and Raids|Hype and Dump]</td>
<td>0.06%</td>
</tr>
</tbody>
</table>

2. Insider Trading case

For the insider trading case, table 4-4 shows that class ‘Time (When)’ got the highest percentage of mapping concepts, 12.01% in total. The participants suggested adding a sub-class called ‘Events Timeline’ in order to describe the manipulation events of the case especially before, during and after the manipulation [VN5200493:5:51-6:45].

In particular, ‘During Fraud\Period’ which describes a specific date during the fraud received 6.15% of the annotated concepts; and ‘During Fraud\Time of Day\Morning’ 0.26% of concepts referring to morning events such as ‘Gallivan’s office manager set up the call with Gallivan for the morning of July 10th, and then

3 VN520049: Name of the recording file for the focus group
contacted the C&B Consulting Firm to request certain benefit information concerning Valencia's SERPs'. ‘During Fraud\Trading Hours\Inside\Close’ had 0.23% of concepts, highlighting the trading hours such as ‘Based on Mid Valley stock's closing price of $19.60 on September 16, 2003’; and ‘Before Fraud\Period' had 5.37% of concepts describing events that happened before the manipulation, such as ‘On July 9, 2002, a Union employee e-mailed Gallivan's Minnesota office manager asking her to set up a due-diligence conference call with Gallivan for July 10, 2002’.

In total, the ‘Effects (Consequences) class was allocated 7.63% of the annotated concepts, which included the different evidence and patterns used by the manipulators to violate the market. In this case, 4.64% of these patterns were mapped to ‘Legal’, such as ‘the CEO of the California Holding Company admonished Gallivan and the other attendees that the meeting was confidential and that a confidentiality agreement was already in place’, and ‘Gallivan agreed to keep the matter confidential. The Minnesota Investor further expected Gallivan to maintain confidentiality based on his own history of confiding in Gallivan’; and 0.95% were structured patterns which describe any manipulation trading or price violation such as ‘Gallivan purchased 3,070 shares of Sun Country at a price of $11.10 per share for a total cost of $34,082’ based on the non-public information he obtained, and ‘Gallivan's purchase represented over 38% of the 8,000 share total volume in Sun Country common stock on February 19’. Furthermore, 2.04% of annotated concepts were classified under ‘Benefit/Direct’, which describes the direct benefit of such a scheme to manipulators, e.g. ‘On August 6th, Gallivan reaped an unlawful profit of approximately $9,455 on the shares that he purchased during this time period’.

For the Actions (What) class, the participants classified concepts describing the type of manipulation. In total, this class had 6.73% of the annotated concepts, with 4.07% for the ‘Information-Based\Material’ which is a new class added by the participants to differentiate between public materials and non-public materials to describe insider trading materials used in the manipulation [VN520048: 13:06-18:36]. Examples of the concepts classified in this class are ‘Gallivan obtained material, nonpublic information by virtue of his position as a BOLI sales representative’, and ‘tipped friends and relatives to trade, on the basis of material, nonpublic information’. Also, manipulators used some trading actions of buying
and selling stocks to violate the market and take advantage based on the non-
public information recommended by the main manipulator, e.g. 'The Ohio Friend,
the Ohio Friend's wife, and the C&B Consulting Firm Friend collectively purchased
3,000 shares of Monterey stock on the basis of Gallivan's recommendation' and
'Gallivan sold these shares on April 30, 2003, the day of the merger
announcement, at the closing price of $12.00 per share, reaping illegal profits of
approximately $2,758'. Therefore, the ‘Trade-Based\Trade\buy’ class got 1.89% of
the annotated concepts and ‘Trade-Based\Trade\sell’ got 0.46% of concepts.
The participants suggested adding a new sub-class called ‘means’ with the small
percentage of 0.31% to describe how the manipulator had the opportunity to get
involved in such manipulation at the first place, e.g. ‘Gallivan consulted with banks
and their executives regarding benefit plans, such as supplemental executive
retirement plans ("SERPs"), and BOLI, which is often used by banks as an
accounting mechanism for funding SEWS’ [VN520048: 1:03:20 – 1:08:36].
To describe the agents involved in the manipulation, participants allocated 4.40%
of the annotated concepts to the ‘Agent (Who)’ class. In detail, 1.62% of the 4.40%
were used to describe the network of relatives and friends that assisted the
manipulator to execute the scheme as ‘Assistance\In Collusion’, e.g. ‘Gallivan
recommended the purchase of Valencia stock to a friend of his who resides in
Ohio ("Ohio Friend"), a cousin who resides in Minnesota ("Minnesota Cousin"), a
son who resides in Minnesota ("Minnesota Son7"), and a friend who works for the
C&B Consulting Firm ("C&B Consulting Firm Friend")’; and their share of the
benefits from this assistance, e.g. ‘The Ohio Friend's wife, the Minnesota Cousin,
the Minnesota Son, and the C&B Consulting Firm Friend collectively purchased
3,850 shares of Valencia stock on the basis of Gallivan’s recommendation,
between July 11 and August 5, 2002’; thus 1.05% of the 4.40% was classified as
‘Benefit\Third party’.
The participants suggested a new class called ‘Agent's Background’ to define the
historical profile of the agent and check whether he was involved in any prior
manipulation scheme (e.g. Gallivan consented to the entry of a Commission order
against him in the proceeding In The Matter Of Denis McCauley and Co., Inc., et
al., Administrative Proceeding File No. 3-4425, Release No. 11365 (April 22, 1975)
[VN520048:1:28:42 – 1:29:05].

97
The litigation case classified the manipulation mainly as an insider trading scheme. Therefore, the participants associated 2.94% of the annotated concepts to describe the scheme as ‘Information-Based\Breach of fiduciary duty\Insider Trading’, e.g. ‘Gallivan breached a duty of trust and confidence to his employer the C&B Consulting Firm by trading on the basis of confidential information that he had obtained in the course of performing his duties as the C&B Consulting Firm representative on Valencia's executive compensation plans’. However, the participants also classified the case as a ‘hype and dump’ manipulation because the agent tried to disseminate the information to manipulate the prices e.g. ‘Gallivan called Harbor's CEO on the telephone several times in the weeks following the October 6 meeting to encourage Harbor to pursue a transaction with California Holding’.

The participants suggested a new class called ‘Manipulation Participants’ to describe not only the agent but also victims and intermediaries who might be involved indirectly in the manipulation [VN520048: 34:00-39:00]. For example, ‘Valencia Bank 8 Trust (“Valencia”)’, and ‘Monterey Bay Bank (“Monterey”) are considered as victims and they are classified under ‘Victims\Organisations’. In this case, participants classified the target manipulated stock (Sun Country, Mid Valley, Harbor stock) as ‘Asset (Target)\Financial\Equity\Common stock’, e.g. ‘Gallivan purchased 3,070 shares of Sun Country at a price of $11.10 per share, for a total cost of $34,082’. The investigator body here was represented by the SEC who investigated the case.

The concepts related to the investigation process were classified under ‘Financial Market Stakeholders\Regulators\Tasks\Investigation’, e.g. ‘SEC's investigation into Gallivan's possible insider trading activity’. The participants found that it would be beneficial if they added a class containing all the laws and regulations referred to by the court against the manipulators. Thus, they suggested a main class called ‘Laws and regulations’ [VN520048:51:54-56].
Table 4-4 Insider Trading Case Categorization Analysis

<table>
<thead>
<tr>
<th>Main Classes</th>
<th>Node</th>
<th>Percentage coverage</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (When)</td>
<td>Nodes\Financial Ontology\Financial Fraud\Time (When)\Events Timeline\During Fraud\Period</td>
<td>6.15%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Time (When)\Events Timeline\Before Fraud\Period</td>
<td>5.37%</td>
<td>12.01%</td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Time (When)\Events Timeline\During Fraud\Time of Day\Morning</td>
<td>0.26%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Time (When)\Events Timeline\During Fraud\Trading Hours\Inside\Close</td>
<td>0.23%</td>
<td></td>
</tr>
<tr>
<td>Effects (Consequences)</td>
<td>Nodes\Financial Ontology\Financial Fraud\Effects (Consequences)\Evidence and Patterns\Legal Benefit\Direct</td>
<td>4.64%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Effects (Consequences)\Evidence and Patterns\Financial Economic\Structured</td>
<td>2.04%</td>
<td>7.63%</td>
</tr>
<tr>
<td>Actions (What)</td>
<td>Nodes\Financial Ontology\Financial Fraud\Actions (What)\Information-Based\Material\Non Public</td>
<td>4.07%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Actions (What)\Trade-Based\Trade\buy</td>
<td>1.89%</td>
<td>6.73%</td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Actions (What)\Trade-Based\Trade\sell</td>
<td>0.46%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Actions (What)\Agent</td>
<td>0.31%</td>
<td></td>
</tr>
<tr>
<td>Agent (Who)</td>
<td>Nodes\Financial Ontology\Financial Fraud\Manipulation Participants\Agent (Who)\Agent Background</td>
<td>1.73%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Manipulation Participants\Agent (Who)\Assistance\In Collusion</td>
<td>1.62%</td>
<td>4.40%</td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Manipulation Participants\Agent (Who)\Benefit\Third party</td>
<td>1.05%</td>
<td></td>
</tr>
<tr>
<td>Manipulation Type</td>
<td>Nodes\Financial Ontology\Financial Fraud\Market Manipulation (Types)\Information-Based\Breach of fiduciary duty\Insider Trading</td>
<td>2.25%</td>
<td>2.94%</td>
</tr>
<tr>
<td></td>
<td>Nodes\Financial Ontology\Financial Fraud\Market Manipulation (Types)\Information-Based\Runs and raids\Hype and Dump</td>
<td>0.69%</td>
<td></td>
</tr>
<tr>
<td>Asset (Target)</td>
<td>Nodes\Financial Ontology\Financial Fraud\Asset (Target)\Financial\Equity\Common stock</td>
<td>1.58%</td>
<td>1.58%</td>
</tr>
<tr>
<td>Manipulation Participants</td>
<td>Nodes\Financial Ontology\Financial Fraud\Manipulation Participants\Victims\Organisations</td>
<td>0.53%</td>
<td>0.53%</td>
</tr>
<tr>
<td>Financial Market Stakeholders</td>
<td>Nodes\Financial Ontology\Financial Fraud\Financial Market Stakeholders\Regulators\Tasks\Investigation</td>
<td>0.26%</td>
<td>0.26%</td>
</tr>
<tr>
<td>Laws and Regulations</td>
<td>Nodes\Financial Ontology\Financial Fraud\Laws and Regulation</td>
<td>0.21%</td>
<td>0.21%</td>
</tr>
</tbody>
</table>

3. Marking the Close

Table 4-5 shows the categorization analysis of a case involving marking the close, wash sales, and matched order manipulation activities. The participants associated 4.86% of the annotated concepts to Actions (What) to describe the form of manipulation. In this case, the manipulator used mixed actions including information-based, action-based and trade-based activities to violate the market and stock prices.

For the action-based methods, the participants associated 0.46% of the concepts with this class. For example, ‘controlled a large percentage of the unrestricted stock’, ‘Although Georgiou opened these accounts, he never used his own name on any account records’, and ‘Georgiou did actively manipulate Northern Ethanol by coordinating an illicit kickback scheme with an individual who, unbeknownst to
Georgiou, was the UA, and conducting matched orders’ are concepts classified as ‘action-based’. As in the insider trading case, the participants suggested adding a class called ‘Means’ to describe the manipulation actions [41:00-43:00]. However, the only difference is that these participants added this class to ‘Action-based’. In total, 2.21% of the annotated stocks were associated with the ‘Means’ class, e.g. ‘using numerous nominee accounts held at offshore broker-dealers’, ‘Georgiou asserted direct control over several nominee accounts for which he was able to issue trading and wiring instructions directly to broker-dealers’, and ‘He appointed three consecutive board members, who routinely leaked confidential information to him’. In the same class, they suggested adding another sub-class called ‘action technique’ to describe the terminology and technique used among fraudsters to execute the manipulations [2:51:04 - 2:52:24]. For example, the case described some manipulation tactics and techniques used to manipulate the market, e.g. ‘Georgiou later acknowledged that this “A to B to C” tactic was effective early on in the Avicena manipulation’. Interestingly, for the information-based actions, the participants added one class called ‘concealed Information’ to describe the hidden information used by the manipulators to violate the market [1:50:00 -1:50:39]. In this case the manipulator did not disclose information as part of his manipulative activities. 0.45% of the annotated concepts are classified under ‘concealed information’, e.g. ‘Georgiou attempted to conceal his association with the nominee accounts’, ‘they also discussed that the purchases would be made in such a way as to avoid regulatory scrutiny’, and ‘Georgiou never disclosed that the value of these securities was artificially inflated as a result of his manipulative activities’. Also, the manipulator engaged in unauthorized trading in client accounts and provided a client with false and misleading information. Thus, 0.40% of the annotated stocks were classified under ‘misleading information’, such as ‘creating a false or misleading appearance of active trading’, and ‘creating a false or misleading appearance with respect to the market for any such security’. Given that these two examples mentioned that the manipulator used false and misleading information, the participants also classified these two specific concepts under ‘False Information’. The ‘false Information’ class has 0.28% of annotated concepts, e.g. ‘Georgiou falsely represented to representatives of Accuvest that
his investment group would be transferring in significant assets to cover the margin deficit'.

Another sub-class called ‘communication’ was added to describe communication between manipulators [1:50:38-1:54:21]. This class has 0.34% of the annotated concepts, such as ‘Georgiou often directed the manipulative trades only by cell phone or sent and received messages by way of BlackBerry PIN’, and ‘When he did use e-mail, he often used email accounts in other people’s names to disguise his identity’.

For the trade-based actions, the manipulators engaged in buying and selling orders and trades to artificially inflate the stock prices to gain profit before they declined again. Concepts such as ‘influence the closing price of a security by executing purchase or sale orders at or near the close of normal trading hours’, and ‘order to buy or sell securities that is entered with knowledge that a matching order on the opposite side of the transaction has been or will be entered’ are classified under ‘Trade-based\order\buy’ and ‘Trade-based\order\sell’ with percentages 22% and 33% respectively.

The participants classified the evidence, patterns, beneficiaries and facts under ‘Effects (Consequences)’, with 1.15% of the annotated concepts. The ‘Financial Economic\Structured’ includes 0.56% of the patterns used by the manipulators to violate the market, such as ‘artificially inflate or depress the closing price for the security and can affect the price of “market-on-close” orders’, ‘artificially inflating Avicena’s stock price or to create the false appearance of an active and liquid market’, ‘decline from $5.40 per share on June 5th to $0.88 on November 6, 2007’, and ‘By June 2007, Georgiou’s manipulation scheme started losing momentum’.

The ‘Asset’ class includes 0.19% of the stocks manipulated in this case, e.g. ‘manipulate the market for the common stock of four microcap companies – Avicena Group, Inc. (“Avicena”), Neutron Enterprises, Inc. (“Neutron”), Hydrogen Hybrid Technologies, Inc. (“Hydrogen Hybrid”), and Northern Ethanol, Inc. (“Northern Ethanol”)’. The manipulation resulted in direct benefit and illgotten gains to the manipulators. 21% of relevant concepts were classified under ‘Benefit/Direct’, such as ‘realized at least $20.9 million in illgotten gains’, and ‘withdraw cash that he wired to offshore bank accounts’.

The ‘Agent (who)’ class had 48% of concepts to describe the manipulators and how they cooperated with each other to violate the stock prices. In particular,
'Assistance\In Collusion' had 0.28% of the annotated concepts to describe these manipulators in the network, e.g. 'an individual who is now a Cooperating Witness (the “CW”), was at all times communicating with Georgiou and placing trades from within this district', 'the individual was, in fact, an undercover FBI agent (the “UA”), 'using another individual, who purported to have access to a network of corrupt brokers, to bribe brokers to buy Avicena stock and “park” or hold the stock in their unknowing clients’ accounts’, and ‘Georgiou told the CW that he had hired a “group” to participate in the manipulation’. The participants suggested adding a sub-class called ‘other associates’ to categorize 0.19% of concepts describing the nominees who helped the manipulator to dictate their trading [1:45:00- 1:49:15], e.g. ‘Georgiou compensated the nominees for allowing him or his associates to dictate their trading’, ‘Georgiou directly paid nominees’, and ‘he allowed them to sell shares at specified times in order to profit in exchange for entering trades at Georgiou’s direction’. The ‘Time/Period’ class represents the timeline event and shows when the manipulation took place, e.g. ‘On May 9, 2006, at Georgiou’s direction, the CW executed a wash sale for 100,000 shares between two of his accounts’. The manipulated stock is a common stock and the manipulation took place on the Pink Sheets market, e.g. ‘Avicena’s common stock is quoted on the Pink Sheets, an inter-dealer electronic quotation and trading system in the over-the-counter securities market which is operated by Pink OTC Markets, Inc., (“Pink Sheets”) under the ticker “AVCE” (formerly “AVGO”). Thus the participants allocated this information to ‘Asset (Target)\Financial\Equity\Pink Sheets with 0.18% concepts and ‘Venue (Where)\Trading Mode\Electronic\Traditional’ with 0.22%. The case was investigated by the SEC: ‘The Commission seeks a final judgment ordering Georgiou to pay civil money’; therefore, the information was classified under ‘Financial Market Stakeholders\Regulators\Tasks\Enforcement’ with 0.44 %. As in the insider trading case, the participants suggested adding a ‘Laws’ class to include all laws and regulations referred to by the court in judging such manipulation, e.g. ‘Section 10(b) of the Securities Exchange Act of 1934 (“Exchange Act”) [15 U.S.C. § 78j(b)]’. The class had 0.58% of the annotated concepts.
4. Front Running Case

Regarding the Front Running case, the ‘Time (When)’ class had the highest percentage of annotated concepts, 10.70%. As in the insider trading case, the participants suggested adding sub-classes ‘before’, ‘during’ and ‘after’ in order to describe the timeline of manipulation events. Table 4-6 shows that ‘During\Period’ had 9.66% of the annotated concepts compared to ‘Before\Period’ with 1.05%. For example, the participants classified ‘Between July 2007 and March 2008, Vianna simultaneously entered orders in the accounts of Customer A and Creswell to trade the same amounts of the same stock’, and ‘Starting in July 2007 and continuing until March 2008, acting in concert with Employee A, Vianna carried out a fraudulent scheme to divert trading profits from Customer A to Creswell’ under ‘During\Period’, and ‘In October 2003, Customer A opened an account with Vianna...
at Maxim to engage in proprietary trading. Vianna was the registered representative for the Customer A account; Employee A had authority to trade the account, and ‘At the end of April 2007, Employee A referred Creswell to Vianna at Maxim’, ‘Creswell opened a Maxim account on June 20, 2007, and ‘Vianna was the registered representative for the Creswell account’ are classified under ‘Before\Period’.

The participants were concerned about describing the evidence and patterns used by the agent to manipulate the stock. Thus, they classified 10.63% of the concepts under Effects (consequences) to list all the evidence and patterns associated with this scheme. In particular, the ‘Evidence and Patterns’ class has 7.95% of the concepts that describe the evidence and patterns used to prosecute the agent, such as ‘these orders were executed by Maxim at $32.36 per share’, ‘Vianna accessed Maxim's order management system and switched the customer accounts on the original orders that had been entered at 10:19 a.m., so that Creswell became the buyer of 60,000 shares of NVLS at $32.36 per share and Customer A the seller’, and ‘Creswell was a seller of those shares at $32.94 per share, realizing trading profits of almost $32,000’.

The participants classified 1.23% of the concepts under ‘Benefit\Direct’, such as ‘The effect of this scheme was to transfer all trading risk from Creswell to Customer A, since Creswell profited whether the stock price went up or down’, Creswell realised over $3.3 million in profits from trades in its Maxim account that were executed against a corresponding trade in Customer A’s account’ and ‘Vianna received ill-gotten gains through the fraudulent scheme, including at least $125,000 in commissions paid to him with respect to, the corresponding trades in the Customer A’. The participants suggested adding a new sub-class called ‘Evidence and Patterns\trading system’ to describe any manipulation that could occur in the trading system [58:00-60:00]. The class has two sub-classes: structured (any manipulation associated with numerical data) and unstructured (any manipulation associated with textual data).

In this case, the agent misused the trading system, changing the record of two customers and diverting one customer’s trade to another, e.g. ‘Every time the market moved to make Customer A's trade profitable and Creswell's trade unprofitable, Vianna improperly misused his access to Maxim's order management system to divert Customer A's profitable trade to Creswell and Creswell's
unprofitable trade to Customer A by changing Maxim's records to inaccurately reflect the account for which the orders were entered
.

In order to describe the manipulation activities, the participants classified 9.83% of the annotated concepts in the 'Actions (What)' class. Given that the agent was manipulating the trading systems, the participants suggested adding a class called ‘Other Related Trade Activity\Manipulating trading system’ to describe the scheme and its actions [31:47-32:02]. This class has 5.12% of the concepts, e.g. ‘Vianna carried out this scheme in violation of his fiduciary duties to Customer A by manipulating Maxim's order entry system and falsifying the records of orders to purchase and sell securities that the executed on behalf of Customer A and Creswell’, and ‘He placed a buy order in one customer's account and a sell order in the other customer's account’.

Also, the participants annotated all the order transactions executed by the agent and classified 0.37% of the concepts under ‘Trade-Based\Buy’, e.g. ‘Creswell was a buyer of 60,000 NVLS at $32.36 per share and a seller of those shares at $32.94 per share, realizing trading profits of almost $32,000’. Also, the participants suggested a ‘Trade-Based\Order\order description’ class to describe the details of the orders, such as the volume of shares and price [1:53:04-1:54:09]. This class had 1.20% of the concepts, e.g. ‘the corresponding orders were for blocks of stock, ranging from 20,000 to 310,000 shares’, ‘Vianna entered a new order to sell 60,000 shares of NVLS for Creswell's account’, and ‘Creswell was a buyer of 60,000 NVLS at $32.36 per share and a seller of those shares at $32.94 per share, realizing trading profits of almost $32,000’.

Also, the case showed that manipulators entered simultaneous orders in the two customers’ accounts to trade the same number of shares of the same stock. Thus, the participants decided to add a new class called ‘Trade-Based\Order\simultaneous buy and sell’ to describe this activity [1:50:34-1:51:41], with 1.98% of the concepts, e.g. ‘Vianna entered simultaneous orders in the Customer A and Creswell accounts to trade the same number of shares of the same stock. In each instance, Vianna entered an order to buy the stock for the account of one customer and entered an order to sell the stock for the account of the other customer’. The manipulator misled his customers, stating that ‘Vianna entered corresponding orders for the Customer A and Creswell accounts, the market moved so that Creswell's original trade was profitable and Customer A's
original trade was unprofitable: In each such instance, Vianna allowed the orders to remain as they originally had been entered’. Thus, the participants classified this annotated concept under ‘Misleading Information’.

For the ‘Financial Market stakeholders’ class, the participants classified 5.37% of the concepts and categorized, for example, the broker dealers involved in this case as ‘Intermediators\Yes-Intermediaries\Broker-Dealers’ with 1.41% concepts, such as ‘Vianna has been associated with another registered broker-dealer. Prior to joining Maxim, Vianna was employed as a registered representative of a number of broker-dealers in New York’, and ‘During the period relevant to this complaint, Creswell maintained and continues to maintain accounts with broker-dealers in the Southern District of New York’. Regarding the ‘Regulators\Tasks\Enforcement’ class, the participants classified 3.96% of the concepts related to penalties and civil actions imposed by the court against the agent. For example, ‘the Commission brings this action pursuit to the authority conferred upon, seeking permanently to engage in the acts, practices and courses of business alleged herein’, ‘the Commission also seeks a final judgment requiring Vianna to disgorge ill-gotten gains, if any, with prejudgment interest thereon, and to pay civil money penalties’, and ‘The Commission seeks a final judgment requiring Creswell, as relief defendant’.

The case was not clearly classified as a front running scheme. Therefore, the participants annotated and classified 2.87% of the concepts under ‘\Market Manipulation (Types)\Trade-Based\Breach of fiduciary duty\Front-Running’, which could help to define the scheme based on the information provided within the case. 1.24% of the concepts helped to define the scheme, such as ‘to divert dozens of profitable trades from one of his customers to another of his customers’, ‘Vianna carried out this scheme in violation of his fiduciary duties to Customer A by manipulating Maxim’s order entry system and falsifying the records of orders to purchase and sell securities that he executed on behalf of Customer A and Creswell’.

However, the participants added two classes, one called ‘Frequency’ to describe the frequency of manipulations, e.g. ‘At least 57 times between July 2007 and March 2008, Vianna simultaneously entered orders in the accounts of customer A and Creswell to trade the same amounts of the same stock’ [46:33-48:17]. Another class called ‘action description’ classified 1.19% of the concepts, such as ‘he
placed a buy order in one customer's account and a sell order in the other customer's account. Every time the market moved to make Customer A's trade profitable and Creswell's trade unprofitable, Vianna improperly misused his access to Maxim's order management system to divert Customer A's profitable trade to Creswell and Creswell's unprofitable trade to Customer A by changing Maxim's records to inaccurately reflect the account for which the orders were entered' [22:10-23:29].

For the Agent (Who), the participants classified 2.44% of the concepts to list the agents involved in the manipulation under ‘Assistance\In Collusion’ class, such as ‘a fraudulent scheme orchestrated by Vianna, a former broker at a New York brokerage firm named Maxim Group LLC (“Maxim”),’ and ‘acting in concert with Employee A, Vianna carried out a fraudulent scheme to divert trading profits from Customer A to Creswell’. The participants suggested adding and ‘agent details’ class, as in previous cases, to give a more detailed description of the agents, such as ‘Vianna, age 38, is a resident of New York’, ‘Creswell is a British Virgin Islands corporation with an office in Geneva, Switzerland’ [22:20 – 25:58], and ‘The Agent ran this manipulation in the Southern District of New York’. Thus, participants classified 1.60% of the concepts under ‘Venue (Where)\Other crossing system\region’. Furthermore, the manipulated asset was classified under ‘Financial\Equity\Common stock’.
Generally, these actions are associated with timeline events to describe the manipulator actions related to information-based, trade-based or action-based. Furthermore, the litigation releases themselves have different sections describing the different actions taken by manipulators, especially in the summary and the facts sections. Thus, participants saturated this class with 27.78% concepts from manipulator actions related to information-based, trade-based or action-based.

4.2.3 Stage 3: Filing in patterns

This stage aims to flesh out the categories after extensive data collection and unitizing. In this stage, all annotated concepts are checked to see whether they have been classified in relevant categories and whether these categories have become saturated. Table 4-7 shows how many concepts have been categorized under the main classes of the ontology. It is clear that participants were concerned about describing the agent’s ‘action’ which represents the highest percentage of categorization in the ontology, with 27.78%.

Furthermore, the litigation releases themselves have different sections describing the different actions taken by manipulators, especially in the summary and the facts sections. Thus, participants saturated this class with 27.78% concepts from manipulator actions related to information-based, trade-based or action-based.

Generally, these actions are associated with timeline events to describe the...
temporal dimension of the manipulation activity. Therefore, the participants decided to categorize concepts describing when the manipulators took the action to execute the manipulation. Thus, the ‘Time (When)’ class has the second highest percentage, by 23.02%, of concepts have annotated by the participants. In order to describe the consequences and effects of these actions on the efficiency of the market, the participants classified 20.41% of patterns and evidence that describe the effect of agents' behaviour, or trading orders and action’s executions on the market integrity.

The ‘Agent (Who)’ class represents the fraudster who performs a manipulation. The participants classified 8.60% of the agents involved in the manipulation with a brief profile description of their status, address, age, and job description. Sometimes, there are companies involved in the manipulation to help the agent to execute his manipulation. This class with its nodes helped the participants to classify most of the concepts that list and describe the agents (insider or outsider), their associates (insider or outsider) and the benefits resulting from the manipulative activities. Generally, the different agents involved in the case were mentioned in different sections in the litigation releases such as ‘Summary’, ‘Defendants and relevant persons and entities’, and ‘Facts’. In Insider Trade case participants suggested an alternative general class called ‘Manipulation participants’ which should include not only the agent (Who) but also ‘victims’ either ‘\Victims\individual’ or ‘\Victims\Organization’ which describe the victims used by the agents to facilitate their manipulative actions. They also introduced an ‘Intermediary’ subclass which describes the network of people or organizations that act as facilitators to enable him to execute the manipulative activity or help him to acquire the appropriate information that could give him an advantage over other investors. Overall, this class has 0.53% of relevant concepts from only insider trading.

Furthermore, agents normally target any type of assets such as stocks, penny stocks, money markets, bonds, derivatives, and structured notes while they perform their manipulation. The ‘Asset (Target)’ class has 2.41% of concepts describing whether the asset is ‘real’, i.e. commodities or financial which includes equity, derivatives and debt. Based on the type of manipulation, agents could target one asset or a group of assets to violate the market and gain profits. Generally, targeted assets are mentioned in the ‘Summary’ and ‘Facts’ sections.
The ‘Market Manipulation Types’ class describes the different manipulation types existing in the financial market. Given that the four analyzed cases were selected based on the four sub-classes of the ‘market manipulation (types)’ ‘abuse of market power’, ‘contract-based’, ‘breach of fiduciary duty’ and ‘runs and raids’, the participants classified 6.03% of the concepts that describe these types. In some litigation releases, the plaintiff describes the manipulative activities without defining the type of manipulation. The participants covered this by identifying suitable concepts to define the manipulation type.

Overall, the ‘Financial Market Stakeholders’ class had 6.61% of concepts describing the different stakeholders involved in the cases, such as investors, intermediaries (e.g. brokers), regulators (e.g. fraud detectors, enforcement agencies, and investigators). Furthermore, the ‘Venues (Where)’ concept has 1.82% of concepts describing where investors and manipulators organize their trade manipulations. The litigation release considered a legal document, one of whose sections states the laws and regulations imposed by the court against the manipulators. Thus, most participants agreed on adding a new class called ‘Laws and regulations’ to include 2.67% of acts and laws mentioned in the case. However, some participants were not interested in such laws because of their financial background.

Table 4-7 Analysis of the main Classes of the Financial Ontology

<table>
<thead>
<tr>
<th>Main Classes</th>
<th>Insider Trading Case</th>
<th>Marking The Close Case</th>
<th>Front Running Case</th>
<th>Pump and Dump Case</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions (What)</td>
<td>6.73%</td>
<td>4.86%</td>
<td>9.83%</td>
<td>6.36%</td>
<td>27.78%</td>
</tr>
<tr>
<td>Agent (Who)</td>
<td>4.40%</td>
<td>0.48%</td>
<td>2.44%</td>
<td>1.28%</td>
<td>8.60%</td>
</tr>
<tr>
<td>Asset (Target)</td>
<td>1.58%</td>
<td>0.18%</td>
<td>0.58%</td>
<td>0.07%</td>
<td>2.41%</td>
</tr>
<tr>
<td>Effects (Consequences)</td>
<td>7.63%</td>
<td>1.15%</td>
<td>10.63%</td>
<td>1.00%</td>
<td>20.41%</td>
</tr>
<tr>
<td>Financial Market Stakeholders</td>
<td>0.26%</td>
<td>0.44%</td>
<td>5.37%</td>
<td>0.54%</td>
<td>6.61%</td>
</tr>
<tr>
<td>Market Manipulation (Types)</td>
<td>2.94%</td>
<td>0.16%</td>
<td>2.87%</td>
<td>0.06%</td>
<td>6.03%</td>
</tr>
<tr>
<td>Time (When)</td>
<td>12.01%</td>
<td>0.25%</td>
<td>10.70%</td>
<td>0.06%</td>
<td>23.02%</td>
</tr>
<tr>
<td>Venue (Where)</td>
<td>0.00%</td>
<td>0.22%</td>
<td>1.60%</td>
<td>0.00%</td>
<td>1.82%</td>
</tr>
<tr>
<td>Laws and Regulations</td>
<td>0.21%</td>
<td>0.58%</td>
<td>0.00%</td>
<td>1.88%</td>
<td>2.67%</td>
</tr>
<tr>
<td>Manipulation Participants</td>
<td>0.53%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.53%</td>
</tr>
</tbody>
</table>

4.2.4 Stage 4: Member checking

This stage enables the researcher to reconstruct the data to check whether it is a reasonable representation of reality. As a further check, the whole process could be evaluated by an external auditor for validation (Grove 1988). The current
version of the domain ontology has nine concepts: action, asset, effects, financial market stakeholders, laws and regulations, manipulation participants, market manipulation types, time, and venue.

4.2.4.1 Ontology Producing

Based on the analysis from the focus group, the domain ontology has been built and developed using the PoolParty Thesaurus management system, as shown in figure 4-4. The web link for the developed financial ontology at the following: 

There are more than 50 ontological editors such as protégé, GALEN Case Environment (GCE), ICOM Integrated Ontology Development Environment, IsaViz, JOE, KAON (including OIModeller) and others. This research used PoolParty because it has many features such as it makes use of the RDF/SKOS specification (Simple Knowledge Organisation System) developed by the World Wide Web Consortium (W3C). The system helps to build and maintain ontologies, vocabulary and taxonomy using a web-based environment integrated with text mining and linked-data technologies. It provides powerful functionalities such as data integration from different sources (structured and unstructured) based on
Figure 4-4 The Proposed Financial Ontology
flexible metadata models, semantic search engines, recommender systems (similarity search), annotation and tag recommender systems, auto-complete (the system suggests words or phrases the user might want, without the user actually typing them in completely), queries (based on the SPARQL query interface) and a semantic index builder. It is especially designed to be used by domain experts. PoolParty uses a thesaurus to control vocabulary as it consists of a collection of concepts. Several labels such as synonyms, abbreviations and spelling variants can be associated with the concept. A thesaurus is highly structured, as it describes the relationships between concepts which are organized in hierarchical or associative relationships. Hierarchical relationships are used to indicate concepts which are narrower or broader in scope.

PoolParty employs SKOS (Simple Knowledge Organization System) to define ontology classes and properties based on the Resource Description Framework (RDF). RDF is one of the fundamental semantic web specifications and was developed within the W3C framework. RDF uses graphs as its data structure. This means every concept is a node; edges (properties) connect these nodes to create a graph. Every statement in RDF encodes information in triples. A triple has the form “subject > predicate > object”, e.g ‘Information-based Manipulation’ > SKOS:narrower > ‘Market Manipulation Types’.

Each concept, property and class in SKOS has a unique identifier called a URI or http URIs. For example, the URI/URN of the ‘Insider trading’ concept is displayed as a link ‘http://open.poolparty.punkt.at/FinancialOntology/67’ to the respective concept in the PoolParty Frontend (HTML format) as shown in figure 4-5. Each top concept has metadata that contains general information for the concept scheme, such as author, title, description, creation date and contributors. The PoolParty interface has a search feature to enable ontology users to find concepts quickly; beginning with entering a search string in the search bar, the system will look for concepts that contain that string in any label and will suggest matching one using the autocomplete feature.

As shown in figure 4-5, financial ontology has four main super-class concepts: application ontology, data source ontology, domain Ontology and lexical ontology layers. Each super-class concept has its own hierarchical concepts. An associative relationship is of a more general nature; it can be used to indicate any
kind of relation between concepts. Concepts and relationships are explained in greater detail by means of examples and visualizations in the following sections.

![Financial Ontology produced in PoolParty environment](image)

**Figure 4-5 Financial Ontology produced in PoolParty environment**

### 4.3 Ontology Layering Architecture

Most of the existing information management systems that use ontologies utilize a flat architecture for ontology management. Generally, flat architecture ontologies are managed independently which makes it challenging to integrate them, especially when multiple ontologies are introduced. Thus, the research introduces the multi-layer architecture of the financial ontology adapted from (Mikroyannidis and Theodoulidis 2012). The advantage of using the multi-layer architecture is establishing number of layers containing ontologies for different purposes developed by different author groups. Furthermore, the architecture helps to improve the manageability of the technologies and demonstrate the integration between different ontologies presented through intra-layer (same layer) or inter-layer (different layer) ontology mapping.

The layers are presented in a pyramid diagram as shown in figure 4-6 to demonstrate a range of basic and generic financial concepts in the lower layer, to
the more specific domain concepts of the upper layers. Each layer is developed and maintained by a different group of authors, according to the expertise required.

From bottom to top, the lexical layer uses the comprehensive online financial dictionary provided by (www.investopedia.com). The domain layer uses and integrates the different taxonomies developed by (Diaz 2011) which is considered as a fair attempt to build an organized body of knowledge for financial market manipulation. The data source ontology layer is based on the website of the US Securities and Exchange Commission (SEC) (www.sec.gov) which has published litigation releases and prosecuted cases. The application layer has the text mining application that automates the process of extracting financial concepts from texts and provides an appropriate knowledge base about financial market manipulation. The integration of the ontology pyramid layers is achieved with the use of ontology mapping between ontologies belonging to the same layer (intra-layer), or different ones (inter-layer).

![Figure 4-6 Financial Ontology Layering Architecture adapted from (Mikroyannidis and Theodoulidis 2012)](image)

4.3.1 Lexical Ontology Layer

This layer contains a domain-independent ontology of a purely lexicographical nature. The lexical ontology (www.investopedia.com) is used because of its comprehensive online financial dictionary and other financial contents such as articles, investment tutorial and videos offered to visitors and the financial community. Indeed, it is one of the biggest websites that is considered as a well-respected financial source, and attracts on average millions of visitors wanting to
improve their financial understanding. Forbes said “Investopedia focuses on markets, investment tools, commentary and online communities. They have over two million unique page views per month” (Edmonton 2007). The website contains over than 6,300 articles, around 13,255 dictionary entries, and 750 pages of tutorials covering aspects of finance and investing.

The Investopedia dictionary has 28 main concepts (such as Accounting, Stock, Taxes, etc). Each concept has relevant financial terms including its definition. Figure 4-8 shows a distribution analysis demonstrating the total number of terms per concept. Investopedia differentiates itself from other online dictionary sources by offering “Investopedia Definitions” which are plain English interpretations of difficult financial concepts, as shown in figure 4-7. Each financial concept is associated with related concepts and their definitions. As shown in figure 4-8, the two concepts with the highest percentage of terms are ‘General Finance’ with 14.31% and ‘Accounting’ with 12.07%. However, concepts such as ‘Currencies’, ‘Forex trading’, ‘Formulas’, ‘Hedge Funds’, ‘Investor Relations’ have the lowest percentage of terms, less than 1%. On average, other classes have terms around 3.5%.

To some extent, Investopedia has managed to build a financial taxonomy with many relevant financial terms and definitions, and associated with related concepts. However, Investopedia has some limitations, such as not considering financial market manipulation as a main concept, which should be uniquely identified. For example, it does not have a classifier for market manipulation; terms related to the types of manipulation are distributed over different concepts, e.g. ‘Pump and Dump’ has been classified under ‘financial BuzzWords’.

Furthermore, many terms are duplicated and classified under more than one concept; e.g. ‘Securities and Exchange Commission- SEC’ is classified into three concepts: ‘Acronyms’, ‘Investor Relations’ and ‘Stocks’. Interestingly, the acronym is attached to its term, although there is a concept called ‘Acronym’ that should be included among all financial acronyms and linked to its own main terms. For example, an acronym such as SRO is currently attached to term Self-Regulatory Organizations as (Self-regulatory Organizations-SRO). Finally, Investopedia is not up to date, with many manipulation terms such as ‘Front Running Research’, ‘Scalping’, ‘Commodity Flow at Delivery Points’, ‘Corner’, ‘Squeeze’, ‘Stuff Quoting’, ‘Flash Quoting’, ‘Trade Away’, ‘Contract Prices at Different Expirations’,
‘Pre-arranged Trade’, ‘Advancing/Decreasing Bid (Ask)’, ‘Option Expiration Date Stock Price or Volume Changes’ and ‘Mini Manipulation’.

Figure 4-7 Investopedia Dictionary

Figure 4-8 Distribution of Investopedia Classes
4.3.2 Data Source Layer

Data sources, either structured or unstructured, are vital in any BI systems. Depending on the domain, structured sources are normally in numerical format, such as the quotes or trades of specific stock. However, unstructured sources are normally in a textual format such as most website content, reports, e-mails and notes. Whilst the Domain Ontology layer describes the general principles of a domain, the Data Source Ontology layer specializes in the organization of information from these sources. In the proposed financial ontology, the relations between the concepts in the domain and the data source ontology would be represented as the mappings between the two layers (Mikroyannidis and Theodoulidis 2012).

In order to improve the reorganization of the information provided by the SEC website, the data source ontology layer exploits the semantic aspects of the web by addressing the structure of the website ontology. The SEC website ontology is strongly related to the topology of the SEC site, since its concepts represent the thematic categories covered by the pages of the site. All concepts are instantiated in the website pages. Each page, depending on its content, is an instance of one or more concepts. In the data source ontology layer, it is assumed that the website ontology is initially constructed by the webmaster, according to his perception of the thematic organization of his website. The website ontology then evolves over time, based on extracted navigation paths, in order to follow trends in the usage of the website by its visitors. The data source ontology layer is based on the website of the US Securities and Exchange Commission (SEC) (http://www.sec.gov).

This section demonstrates the data source ontology layer and how the classes are instantiated from the SEC website pages, as shown in figure 4-9. More detailed information about the enforcement class is presented, the topic of this research.

**About:** This class contains web pages describing the profile of the SEC including their mission and objectives. Also, it includes articles about the SEC’s history, responsibilities, reports, activities, events, securities laws, organization and operation. Furthermore, it contains information about the commissioners who are appointed to interpret federal securities laws, issue new rules, and oversee the inspection of securities and coordinate securities regulation with federal, state, and foreign authorities.
Divisions: This class describes the five main SEC divisions: the Corporate Finance division assists the Commission to supervise corporate disclosure of important information to the investing public; the Enforcement division investigates possible violations of securities laws, recommends action by the Commission when appropriate, either in a federal court or before an administrative law judge, and negotiates settlements; the Risk, strategy and financial innovation division provides the Commission with sophisticated analysis that integrates economic, financial, and legal disciplines; the Investment Management division regulates investment companies such as mutual funds, closed-end funds, etc.; and the Trading and Markets division establishes and maintains standards for fair, orderly, and efficient markets, regulating the major securities market participants, including broker-dealers, self-regulatory organizations (such as stock exchanges, FINRA, and clearing agencies), and transfer agents.

Figure 4-9 The SEC data source ontology

Regulation: This class provides information regarding SEC rulemaking activity, including concept releases, proposed rules, final rules, interpretive releases, and policy statements. Moreover, it has announcements concerning SRO rulemaking, instructions for Exchange Act Exceptional Applications, other Commission notices, and public petitions for rulemaking submitted to the SEC.
**Education:** This class provides a huge amount of financial educational material to help investors’ to invest wisely and avoid fraud.

**Filings:** This class describes information about the complete list filings (companies are required to file registration statements) available through EDGAR and instructions for searching the EDGAR database.

**News:** This class contains the latest in press releases, news, speeches, public statements, special studies, events, developments, and updates to the SEC website.

**Enforcement:** This research focuses on this class because it contains information about SEC enforcement actions, Commission opinions, administrative proceedings, trading suspensions, litigation releases, appellate court briefs, commission amicus and notices concerning the creation of investors’ claims funds in specific cases. Furthermore, the class has additional information related to corporation agreement and information about how the SEC investigations are conducted.

The enforcement division assists the Commission by suggesting the commencement of investigations of securities law violations. The Commission can bring civil actions in federal court or as administrative proceedings before an administrative law judge, and by prosecuting these cases on behalf of the Commission. With a formal order of investigation, the division collects evidence of possible violations of securities laws from many sources such as market surveillance activities, investor tips and complaints, SROs, other divisions of the SEC, other securities industry sources, and media reports. Following an investigation, the enforcement division presents their findings to the Commission for its review.

The Commission can authorize the staff to file a case in federal court, ask the court for a civil action against manipulators or bring an administrative law judge (ALJ), who is independent of the Commission to check the evidence provided by the enforcement division and make a decision that includes findings of fact, legal conclusions and recommended sanctions. Administrative sanctions include cease and desist orders, suspension or revocation of broker-dealer and investment advisor registrations, censures, bars from association with the securities industry, civil monetary penalties, and disgorgement. However, in many cases, the Commission agrees with the charged party to settle the case without going to trial.
Figure 4-10 shows the list of different enforcement action concepts that has been added to the enforcement class:

- **Administrative Proceedings**: This concept contains a list of notices and orders concerning institution or settlement of administrative proceedings.

- **Commission Opinions and Adjudicatory Orders**: This concept contains a list of opinions and orders issued by the Commission adjudicating either appeals from initial ALJ decisions or disciplinary or adverse action taken by SROs such as FINRA or motions in connection with these proceedings.

- **Accounting and Auditing Enforcement Releases**: This concept provides financial reporting-related enforcement actions concerning civil lawsuits brought by the Commission in federal court and notices and orders concerning the institution and/or settlement of administrative proceedings.

- **Trading Suspensions**: This concept provides a list of recent SEC trading suspensions. According to the federal securities laws, the SEC has authority to suspend any stock for up to 10 trading days when it determines the necessity of protecting investors and is required for public interest.

- **Office of Administrative Law Judges (ALJ)**: This concept contains the ALJ decisions and ALJ orders for the cases that have been sent by the Commission to determine whether the allegations in the order are true to issue an initial Decision in a specified period of time and may order sanctions that include suspending or revoking the registrations of securities.

- **Commission Amicus and Appellate Court Briefs**: These two concepts contain lists of some of the legal briefs which the Commission’s staff submitted in various court actions.

- **Litigation Releases**: This concept provides the list of litigation releases concerning civil lawsuits brought by the Commission in federal court.
4.3.2.1 Litigation releases

Each litigation release has a release number, release publication date, and action that include the defendants' names; and most of the releases have an external link to SEC complaint. The SEC complaint documents are the reports produced by the US district court that describe violations cases of securities laws. In these documents the court describes in detail all the evidence and facts that make its decisions to prohibit the acts or practices that were the results of violation of the law or commission rules. Based on the case, the court might order some sanctions against defendants such as suspension of the defendants from their positions or/and returning any illegal profits (disgorgement) or/and issuing any civil penalties or/and defendants being subject to additional fines and imprisonment.

Each document has a format and structure as shown in figure 4-11. The document format includes the ‘ComplainID’ which is equivalent to the litigation release number, document type ‘PDF’ or ‘HTML files, the file size, and the link to the document. The document structure consists of five main sections: ‘Civil case action no’ which is the file number the court allocates to the document; ‘district court name’, ‘court clerk’s office stamp’ which includes his name, signature, title and filing date, ‘title’ including the names of the defendant and the plaintiff, and the main ‘document sections’ which generally include a summary of the complaint, list
of defendants and relevant person entities involved in the violation, jurisdiction and 
the venue of the court, facts that describe the manipulation scheme and all the 
evidence that has been collected by the Commission to prosecute the defendants, 
‘fraud for reliefs and violation’ includes all the acts and laws that defendants 
violated in such case.

Figure 4-11 Litigation Releases Concept

4.3.3 Domain Layer

Based on the analysis from the focus group, the domain ontology has been built 
and developed using the PoolParty Thesaurus management system, as shown in 
figure 4-12. Each concept holds sub-concepts to build the hierarchy of concepts. 
The hierarchy is formed by stating that concepts are in ‘narrower’ or ‘broader’ 
relationships. For example, the ‘Insider trading’ concept has a broader concept 
‘Breach of Fiduciary Duty’ and narrower concepts such as ‘Gallivan breached a 
duty of trust and confidence to the Minnesota Investor, who had a history, pattern, 
and practice of sharing confidential information with him’, and ‘Gallivan breached a 
duty of trust and confidence to his employer the C&B Consulting Firm by trading 
on the basis of confidential information obtained in the course of performing his
duties as the C&B Consulting Firm representative on Mid Valley’s executive and director compensation plans’.

Each concept has a preferred label which describes the main word or phrase that is used to identify the concept, while multiple alternative and hidden labels are allowed. For example, the concept ‘insider trading’ is the preferred label to describe the name of this manipulation. Furthermore, definitions are included to describe the meaning of the concepts. For example, the ‘insider trading’ concept is defined as *The buying or selling of a security by someone who has access to material, nonpublic information about the security. In other words, the trading becomes illegal when the material information is still nonpublic because trading while having special knowledge is unfair to other investors who don't have access to such knowledge.* This definition is quoted from the Investopedia website. Thus, ‘financial sources\Investopedia source’ is built into a metadata section to include a link to the Investopedia hyperlink. For example, ‘insider trading’ has a metadata web link to Investopedia ‘http://www.investopedia.com/terms/i/insidertrading.asp’.

There are three ways to build semantic relations in PoolParty, namely, related, exact-matching concepts, and close-matching concepts. Related concepts have been used to link concepts to each other. This does not put concepts in a hierarchical relationship, but in an associative one. It is rather a semantic association which is used to express any kind of relationship between concepts. For example, the ‘Insider Trading’ concept has an associative relationship with the ‘Benefit’ and ‘Agent\Agent characteristics\individual’ concepts because of the annotated concept ‘This case concerns unlawful insider trading and tipping by securities law recidivist and bank-owned life insurance (BOLI) sales representative, Robert J. Gallivan (Gallivan)’.

Exact matching concepts have been used to link the concepts with linked open data features to enrich the concept description. DBpedia resources (the semantic version of Wikipedia) have been used to describe the concept and are displayed as exact matches in the ontology, as shown in figure 4-12. For example, the ‘insider trading’ concept has been linked to DBpedia resources and its subjects, definitions, types, pictures etc. The DBpedia definitions are used as definitions for the insider trading concept. Close-matching concepts have concepts from different concept schemes or linked data which are nearly identical. This relation type has
not been used in the ontology because it was important to have a precise definition for each concept.

Each concept scheme uses RDF triples specifications based on W3C, grouped into two sets. The first set displays all triples where the concept scheme is used as subject. However, the second set displays all triples where the concept is used as an object. For all objects and subjects representing concepts a clickable URN or URI is displayed. Following those links the triples of these concepts are displayed as shown in figure 4-12.

Figure 4-12 Domain Ontology Implementation
The domain ontology expanded its thesauri with the annotated concepts analyzed by the four litigation releases. The litigation releases are uploaded and processed in the ontology project to extract and map the domain-specific terms to the relevant domain ontology classes. These extracted concept events are stored with the document and have been classified to relevant classes in the domain ontology. A full-text search feature is enabled with a build-document index in order to update the index of documents and have the four litigation releases included in the domain ontology. For example, concepts such as ‘Valencia Bank 8 trust (Valencia)’ has been annotated, stored, tagged, and classified under ‘Market Participants\Victim\Organization’, as shown in figure 4-13.

![Figure 4-13 Semantic Search and Tag Events in PoolParty](image)

In order to validate the ontology from any violations of SKOS principles, some quality queries have been executed to check the ontology’s integrity. These queries validated the SKOS thesaurus from the completeness of its graph structure, to be non-cyclic (e.g. no circularity in the broader-narrower hierarchy, it has no disjoints between related and hierarchal paths, check whether multiple concepts have identical labels, find labels with fewer than two characters, or concepts without a definition).

4.3.4 Application Layer

The fact that the SEC website provides access to vast repositories and volumes of unstructured and semi-structured texts from various sources makes the analysis of these sources interesting, because there are clear opportunities for rich financial fraud contexts to emerge, impacting how fraudsters practise the fraud and
manipulate the market. Indeed, manual techniques are challenging and increasingly impracticable to satisfy the required analysis tasks. Therefore, text-mining technology is used to automate the process of extracting financial concepts from texts and provide an appropriate knowledge base about financial market manipulation. Text mining has been identified as the most appropriate technological area, allowing automatic analysis of large quantities of documents through the development of linguistic and non-linguistic patterns.

In particular, the application layer used the SEC case study to provide empirical evidence of how text mining could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of extracting financial concepts and demonstrating the published prosecuted cases in an appropriate knowledge base. Moreover, this layer demonstrates significant cross-fertilization across Information technology and finance as it attempts to incorporate the text-mining process within the proposed financial fraud ontology and explore the potential efficiency and effectiveness benefits for analyzing the SEC litigation releases. Thus, domain-specifcity is incorporated through the development of linguistic resources of text mining for the domain of financial fraud in order to make the analysis more meaningful and effective.

The proposed financial fraud ontology played a vital role in providing the underlying framework for the extraction process and capturing information related to financial fraud. Moreover, the ontology helped to demonstrate the extracted information in an organized and coherent way that is as a knowledge base to facilitate users’ acquisition, maintenance and access to knowledge, improving search results in the SEC enforcement portal. Another very important contribution of the ontology is that it addresses the need to reuse the text-mining process in other parts of the domain, and to integrate the extracted information with other systems within the organisation.

Therefore, the application ontology layer (figure 4-14) constructed for the SEC case study represents the organization of the developed text-mining components. This layer adapts the application ontology for information extraction (IE) tasks introduced by Mikroyannidis et al. (2012). The application layer can process all document types such as text, audio and video. Thus, a DocumentFormat class is added to the application layer with the following sub-classes: AudioFormat, MultimediaFormat, and TextualFormat. This research concerns textual resources;
thus the TextualFormat class will have concepts related to documents used in the application, such as the SEC case study, the SEC litigation release and RSS feed. In addition, each document instance could be related to one or more annotations used to develop the linguistic patterns.

Two main types of resource are used in the text-mining components, namely language resources and processing resources. Language resources contain resources such as a thesaurus, list of terms, concepts, synonyms, and types (semantic groupings of concepts). Processing resources incorporate analyzers, generators, recognizers (e.g. speech transcribers, handwriting recognizers), and retrievers (e.g. search engines).

Generally, processing resources utilize language resources such as tokenization and morphological analysis to extract terms from the text and match them with the appropriate named entities. Tokenization analysis parses documents into characters and words that are called tokens. Advanced algorithms used in the text-mining application are classes added to the application layer to deal with language challenges such as inconsistent punctuation, and special characters such as dash symbols and apostrophes. For the morphological analysis, stemming algorithms are additional classes to identify the roots of terms listed in the text-mining application. Text-mining applications integrate stemming algorithms on the tokenization output to conflate the tokens into an orthogonal set of distinct morphological groups that are used to train the extraction engine to group similar forms (singular or plural) of terms and add them to the dictionary.
4.3.4.1 Document Format

The text-mining application used different textual sources collected from the SEC website, such as RSS format, HTML and PDF files. The RSS format is a simple XML-based standardized format used to download the recent litigation releases published in the SEC website. This is used to prepare the collected textual data for the text-mining process. Most of the litigation releases have an HTML link that contains a short description as a summary of the case followed by a detailed description of the cases. Some of these litigation releases have an external link to an SEC complaint. The SEC complaints are PDF documents produced by the district courts to provide a full description of the prosecuted cases. In particular, the text-mining application analyzed the third quarter of 2012 which contains 62 litigation releases, as shown in figure 4-15.
4.3.4.2 Language Resource

The ‘corpus’ concept has the actual litigation releases of the third quarter of 2012 that have been analyzed in the application, such as ‘LR-22420’ and ‘LR-22421’. The ‘document’ class contains the ‘insider trading’ case study which was originally a PDF file but has been converted to the ‘docx’ extension. This case will be used to demonstrate how text mining could automate the process of classifying financial concepts in the proposed classes in the financial fraud ontology. As mentioned above, the financial fraud ontology acted as the underlying framework to identify the concepts that need to be captured and extracted from the different sources. The text-mining application was developed on the basis of the results of the manual annotation process provided by the focus group participants.

With respect to the lexical ontology layer, the lexicalResource concept in this application layer is used to incorporate the Investopedia financial dictionary as a financial lexical resource for the text-mining application. Mozendra WebCrawler software was used to download the 28 main concepts with a total of 13,255 financial concepts. Furthermore, the application incorporated each concept definition (general definition and Investopedia definition) and related terms to the concept, as shown in figure 4-16. In practice, this part of the application simulates
the Investopedia website’s functionality. In particular, the application allows users to query specific financial terms, understand their meaning and explore other related terms. In practice, the text-mining application should capture financial concepts from the documents and provide a full explanation of the term. For example, a user might want to do a simple query to explore the meaning of the terms 'Shares', 'Insider Trading', 'Securities and Exchange Commission' and 'Securities Exchange Act of 1934' that have been extracted from the document. Figure 4-16 shows the output, the definitions of the terms. Also, the user may want to know what other concepts are related to the term ‘insider trading’ for example. Figure 4-17 shows that ‘Expert Network’, ‘FLSA’, ‘Insider Director’, ‘Insider Information’, ‘Insider Trading Act of 1988’, and ‘Market Surveillance’ are related terms to the ‘Insider Trading’ concept.

It is important to emphasize the role of ontology in adopting an evolutionary dynamic open-learning model that can update the new financial terms that emerge from the domain. The text-mining application can be updated by these lexical resources to increase the level of automation expected for accurate extraction.

The financial domain, for instance, evolves new terms as products, services, manipulation types and techniques are introduced; this emphasizes the need for a dynamic open-learning model to ensure the continuous improvement and appropriate accuracy levels that are expected from the text-mining components.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Definition</th>
<th>Insider/Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insider Trading</td>
<td>Definition of Insider Trading: The buying or selling of a security by someone who has access to material, non-public information.</td>
<td></td>
</tr>
<tr>
<td>Securities and Exchange Commission</td>
<td>Definition of Securities and Exchange Commission: An independent commission created by Congress to regulate the sec.</td>
<td></td>
</tr>
<tr>
<td>Shares</td>
<td>Definition of Shares: A stock or ownership interest in a corporation or financial asset.</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4-16 Terms definitions output**

<table>
<thead>
<tr>
<th>Terms</th>
<th>Related Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insider Trading</td>
<td>Audibility</td>
</tr>
<tr>
<td>Insider Trading</td>
<td>Insider</td>
</tr>
<tr>
<td>Insider Trading</td>
<td>Insider Director</td>
</tr>
<tr>
<td>Insider Trading</td>
<td>Insider Information</td>
</tr>
<tr>
<td>Insider Trading</td>
<td>Fair Labor Standards Act - FLSA</td>
</tr>
<tr>
<td>Insider Trading</td>
<td>Market Surveillance</td>
</tr>
<tr>
<td>Insider Trading</td>
<td>Insider Trading Act of 1988</td>
</tr>
<tr>
<td>Insider Trading</td>
<td>Expert Network</td>
</tr>
</tbody>
</table>

**Figure 4-17 Insider Trading Related Terms**

The developed text-mining application uses different components to develop the advanced linguistic patterns, which could contain the following components: sub-classes from the developed library, synonyms, macros, and word gaps. In particular, ‘FinancialFraud Library’ is another sub-class added to the
‘lexicalResource’ concept which has 20 classes, as shown in figure 4-18. This library was developed to help the text-mining application to extract concepts from the litigation releases, especially the fraud-related concepts. Based on the annotations provided by the focus group participants in the insider trading case, the library has over 223 concepts that are classified and mapped to classes. This library is used to develop and construct the advanced linguistic patterns annotated by the focus group participant. For example, “Confidentiality” includes a list of terms related to confidential information such as “confidential information, confidentiality, confidential advice, confidentiality agreement, confidentiality policy, code of ethical conduct, confidential, etc”. “OperationType” contains terms related to trading operation such as “acquired, sold, obtained recommended, sell, buy, purchase, etc”. “AgentNetwork” includes all terms representing the type of relatives who help manipulators to execute the fraud, such as “son, cousin, friend's wife, friend, relative, etc”. “Employees” has all terms related to employees such as “employee, manager, chairman, board, office manager”. “District Court Name” holds a list of different state courts such as “federal district court, eastern district of New York, middle district of Florida, district of New Jersey”. Furthermore, the library includes synonyms which associate two or more concepts that have the same meaning. In particular, synonyms have been used to resolve the issue of misspelled concepts, and concepts having the same meaning, e.g. “Securities and Exchange Commission” and “Commission”.

The macros is another class added to the lexicalResource concept, which represents reusable patterns; it is used to simplify the appearance of literals and word strings needing to be extracted, e.g. prepositions, articles, and verbs. Overall, the application includes six macros that support the extraction process and pattern development. For example, in figure 4-18, “mPreposition” includes a list of tokens related to prepositions such as “to, from, for, of, on, at, with, about, into, etc ...”. “mArticles” includes a list of tokens related to English language articles such as “a, the, an, etc”. “mVtobe” is another macro that holds tokens related to concepts concerning the verb ‘to be, e.g. “is, are, was, were”, etc.
4.3.4.3 Processing Resource

This concept demonstrates the text-mining process and the advanced linguistic pattern approach employed on the basis of natural language processing (NLP) in order to linguistically analyze the litigation releases. 60 advanced linguistic patterns were developed to analyze the litigation releases sentence-by-sentence and to apply focus group participants’ recommendations. This section demonstrates the ‘Insider Trade’ analyzers developed to automate the process of extracting information in the context recommended by the ontology to explain the fraud cases.

A. Data Source Analysis

Following the data source ontology, the text-mining application aims to extract key information from the data source, in this case the litigation release page in the
SEC website. As shown in figure 4-19, the target concepts are ‘litigation release number’, ‘release publication dates’, ‘agents’ which are the defendants’ (individual or organization) names, ‘document format type’, the ‘document link’, ‘civil case no.’ which is the allocation number issued by the court, ‘district court no.’ including state courts or federal courts, and the ‘plaintiff’. The text-mining application used some of the predefined libraries incorporated in the IBM PASW 14 software, such as date, time, person, organization. However, these patterns did not capture all related concepts within the document. Thus, extra patterns have been developed to cover the gap and to increase the accuracy of extraction. Furthermore, other analyzers were developed from scratch to match the specific structure of linguistic patterns.

Table 4-8 explains the patterns developed by the ‘Civil Case No’ analyzer. Each court has a different pattern and in order to capture most of them, 18 linguistic patterns using regular expressions were developed. These patterns have been numbered sequentially from 1–18 to avoid any break in numbering which might cause suspension or conflict when processing the document. For example, in the first pattern in Table 4-9, “6-12-CV-00932-JA-GJK”, the regular expression was developed as [0-9]{1, 2} to match a digit repeated exactly one or two times [e.g. 6], followed by specific character “–”, similarly, s[0-9]{1, 2} to match the two digits, followed by “-”, followed by two character [case sensitivity is considered] [a-zA-Z]{1,2}, followed by five numbers [0-9]{3, 5}, followed by 2 characters [a-zA-Z]{1,2}, followed by three character [a-zA-Z]{1,3}. The “Civil Case No” analyzer successfully captured 100% from the 62 litigation releases used in the application.
From the 62 litigation releases, the analyzers’ ‘litigation release number’, ‘release publication dates’, ‘actions’, ‘document format type’, ‘document link’, ‘short description’, ‘detailed description’, and ‘plaintiff’ have a high accuracy level of almost 100% precision. Regarding ‘Civil case no.’, the analyzer achieved 98.93% precision, and for ‘District court Name’ 93.55% precision. In five releases the court names were not included due to the pending status of the allegation or administrative proceeding status, or were missed.

Despite the size of the documents, the analyzers are good enough to demonstrate how text mining could be used for automating the analysis of SEC litigation releases. Figures 4-20 and 4-21 show the final output of the target concepts captured by the data source analyzers. For example, in litigation release number ‘LR-22396’ published on 20th June 2012, the defendants are ‘Gary J.Mortal’, ‘Martel Financial Group’, and ‘MFG Funding’.

The user can find short or detailed descriptions of the release via the link http://www.sec.gov/litigation/litreleases/2012/lr22396.htm. Therefore, the only information provided by the SEC is through the link as the release does not yet have a complaint file issued by the respective court. The Security and Exchange Commission, the plaintiff in this release, sent the case to the federal district court. The civil case number of the release is ‘12-cv-11095’.

<table>
<thead>
<tr>
<th>Item</th>
<th>Regular Expression developed Patterns</th>
<th>Examples of ‘Civil case no.’ Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>regex1=[0-9][1,2][0-9][1,2][a-zA-Z][1,2][0-9][3,5][a-zA-Z][1,2][a-zA-Z][1,2][0-9][3,5]</td>
<td>Civil Action No. 6-12-CV-00932-JA-GJK</td>
</tr>
<tr>
<td>2</td>
<td>regex2=[0-9][1,2][0-9][1,2][a-zA-Z][1,2][0-9][3,5][a-zA-Z][1,2][a-zA-Z][1,2][0-9][3,5]</td>
<td>Case No. 2:12-cv-03794-JLL-MAH</td>
</tr>
<tr>
<td>3</td>
<td>regex3=[0-9][1,2][a-zA-Z][1,2][0-9][3,5][a-zA-Z][1,2][a-zA-Z][1,2][0-9][3,5]</td>
<td>Case No. 09-cv-7594-KBF-THK</td>
</tr>
<tr>
<td>4</td>
<td>regex4=[0-9][1,2][a-zA-Z][1,2][0-9][1,4][a-zA-Z][1,2][0-9][1,3]</td>
<td>Civil Action No.97-CV-1643-D</td>
</tr>
<tr>
<td>5</td>
<td>regex5=[0-9][1,2][a-zA-Z][1,2][0-9][1,4][a-zA-Z][1,2][0-9][1,3]</td>
<td>Case No.12-CV-6421</td>
</tr>
<tr>
<td>6</td>
<td>regex6=[0-9][1,2][a-zA-Z][1,2][0-9][1,3][a-zA-Z][1,2][0-9][3,5][a-zA-Z][1,2][0-9][3,5]</td>
<td>Civil Action No. 1:09-cv-7594-KBF-THK</td>
</tr>
</tbody>
</table>

Table 4-8 ‘Civil Case No’ Analyzer
B. SEC Complaint Document Analysis

The text-mining application analyzed the SEC complaint document produced by the US district courts that describes cases of violations of securities laws. In these documents the court describes in detail all the evidence and facts that make the decisions to prohibit the acts or practices that were the results of violation the law or commission rules. The application contains 11 analyzers to extract the annotated financial concepts recommended by the domain experts (focus group participants). Following the domain ontology layer, the analyzers analyzed the document sentence-by-sentence and applied the experts' recommendations. In total, 60 advanced linguistic patterns were developed to extract information related
to financial fraud and classify this information in the appropriate ontology classes, as guided by the domain ontology layer. These analyzers with ontology integration improve the efficiency and effectiveness of extracting information and demonstrating the SEC case in the appropriate knowledge base. Furthermore, users are assisted in their acquisition, maintenance, and access to knowledge, with improved search results from the SEC website. This section demonstrates the analyzers that were developed, and shows how they provide answers to the key ontology questions.

1. Manipulation participants analyzer

This analyzer represents the manipulator who performs the manipulation, and whether the manipulator acts by him or has networks of other agents who helped him to execute the manipulation. Furthermore, the analyzer extracts the information describing benefits behind such manipulation, whether they accrue to the manipulator or to others. Finally, it checks whether the manipulator has any previous records or history of manipulations or violations. In total, the analyzer has nine linguistic patterns to answer these questions and describe the manipulator and his social network profile; see figure 4-22.

![Figure 4-22 Manipulation Participant Analyzer](image)

The first three patterns show the agent who performed the violation and the manipulation activity type, as shown in figure 4-22. The patterns automatically analyze the sentence, extract the concept ‘Robert J. Gallivan’ and classify it under the <Person> sub-category. The concept ‘defendant’ is classified under the
<LegalTitle> sub-category, the concepts ‘breached a duty of trust and confidence’ and ‘insider trading activity’ under the <Insider Trading> sub-category, and the concept ‘the C&B consulting Firm’ under ‘organization’. Using the regular expression the analyzer retrieves the dates corresponding to the manipulation activity. The patterns automatically classify these sentences as the manipulator who performed the manipulation and map it to the ‘Manipulation Participants\Agent\Agent Characteristics\Individual’, as shown in figure 4-20. The line width and node sizes in a concept graph represent the global frequency counts of the extracted concepts from the document. For example, apparently the concepts ‘Robert J. Gallivan’ and ‘breached a duty of trust and confidence’ were mentioned in the document several times, represented by the thickness (Global count 5) of the line as shown in figure 4-23.

Generally, from a linguistic perspective, there are some literals or word strings that are important to the analysis, while others can be excluded. For example, the prepositions “to”, “from”, and others are not essential. Furthermore, some words or tokens can be excluded from the extraction process and are not important to highlight in the pattern output. Thus, the analyzers use word gap “e.g. @ {0, 1}” to guide the pattern dealing with the existence of the token, without the necessity of showing it in the pattern output.

Figure 4-23 Who is (are) the agent(s) involved in the manipulation?

In order to check whether the manipulator has a previous violation record, three patterns are developed to automatically analyzes the complaint document and extracts the concepts that describe the manipulation history of the manipulator, as
shown in figure 4-24. The Commission found that Gallivan, who was affiliated with a broker-dealer at the time of the scheme, wilfully violated Section 17(a) of the Securities Act of 1933. In 1975, without admitting or denying the Commission's findings, Gallivan consented to the entry of a Commission order against him in the Proceeding File No. 3-4425.

![Diagram showing Gallivan and affiliated broker-dealer]

**Figure 4-24 Manipulator Violation History**

As shown in figure 4-25, the last three patterns in the manipulation participants’ analyzer are developed to extract the information that addresses the manipulator’s social network, which helped him to violate the securities Harbour, Mid valley and Valencia securities (Target Assets). In this case Gallivan recommended the purchase of different stocks to his relatives, friend, and cousin to gain unlawful and combined profits reaching $58,453. The patterns automatically classified these sentences to the ontology class as the agent was in collusion with other agent networks (Assistance\ln collusion). Furthermore, both the manipulator and his networks received benefits from the manipulation (Benefit\Own Benefit and Benefit\Third Party), as shown in figure 4-21.
2. Investigation, Enforcement and Regulatory Analyzer

This analyzer has been imposed by the court. The analyzer aims to extract information related to the manipulation type in the case, and also the different market-surveillance and market-monitoring activities of regulators, such as detection, investigation and enforcement tasks. To complete the regulatory framework, the regulator sends the case to the court pursuing civil penalties, and enforcement action is imposed by the courts. In total, the analyzer has nine linguistic patterns to describe the manipulation scheme and its legal aspects, as shown in figure 4-25.

The different patterns automatically analyze the sentences, extract the concepts and classify them under a different part of the domain ontology, as shown in figure 4-26. In particular, three patterns extract the concepts relating to the plaintiff, to investigate and enforce the case to the ‘Financial Market stakeholders’ class. Another three patterns extract all concepts related to the laws and regulations that the manipulator violated and map them to the ‘Laws and Regulations’ class. The last three patterns are concerned with capturing the type of manipulation the agent executed to violate the market; they classify it under the ‘Market manipulation Types’ class.
For example, the \(<\text{Plaintiff}> + \text{Regulator Tasks}+ \text{Person}+ \text{Insider Trading}>\) pattern extracts the key concepts from the sentence “SEC's investigation into Gallivan's possible insider trading activity”. The analyzer shows how the developed patterns can extract all the relevant information expressing the role of regulator to investigate any manipulation activity. As shown in figure 4-27, the plaintiff in this case is the Security Exchange Commission, which is investigating an insider trading activity executed by Gallivan. The SEC seeks civil penalties against the manipulator, and thus the case has been sent to the court with all evidence to judge and impose penalties based on the laws and regulations framework. Under sections 21(a) 21(d), 21(e) and 27 of the Exchange Act, the court found that Gallivan violated the securities laws and regulations based on insider trading and tipping activity.
Figure 4-27 Type of Manipulation and laws and regulations imposed by court

3. Timeline Manipulation Events and Actions Analyzer

This analyzer combines three analyzers ‘Actions’, ‘Effects’, and ‘Time’. This section demonstrates the patterns developed for the three analyzers. The analysis indicates a strong relationship between the three analyzers because they describe the facts and nature of the manipulation activity executed by fraudsters. These actions could be related information-based activities such as obtaining non-public information and breach of confidence or trust, or could be trade-based activities such as buying and selling stocks to stimulate the market and violate prices. In this case, each action is associated with the temporal dimension and explains the period in which manipulator performed these manipulative activities.

These actions have an effect on the manipulated assets represented in direct or indirect benefits and unlawful profits. In this case, patterns and evidence such as legal, financial and economic unstructured information are used to trace the behaviour of manipulators and show the consequences of their behaviour on the market.

This analyzer contains 50 advanced linguistic patterns to extract information related to facts and the actions executed by the manipulator associated with the timeline. As shown in figure 4-28, nine patterns are classified under ‘Actions’ classes, 20 patterns extract concepts related to the ‘Effects’ of manipulation, and 21 patterns are classified under the ‘Time’ class which describe the events before, during and after the fraud.
Figure 4-28 Timeline Manipulation Events and Actions Analyzer

Figure 4-29 demonstrates the output of the 50 patterns developed for this analyzer applied to the ‘insider trading’ case study. The output shows the complexity of the manipulated activities executed by the agent. As shown in the figure, most of the events are interconnected and interrelated, such as date, agreements and
confidentiality, stocks prices, amount of purchases, manipulated assets, agent and his networks, amount of combined profits collected by the agent, the way of communication and meeting to obtain the non-public material to take advantage, other evidence and patterns either legal or related to economic structure or trading used by the agent to violate the market.

The analyzer can model the sequence of the manipulative events by involving all the patterns related to the time. Table 4-10 shows the sequence of events based on the ‘Date’ and ‘Target Assets’ concepts. In this case, the manipulator violated the prices of four stocks: Valencia Stock, Sun Country Stock, Mid valley Stock, and Harbor Stock. The manipulation actions between the four stocks are similar, as the manipulator ‘Robert J. Gallivan’ has breached a duty of trust and confidence of his company ‘C&B Consulting Firm’. Based on his role in the company, the manipulator attended meetings and set up calls with companies and investors which gave him an opportunity to obtain non-public material. ‘Robert J. Gallivan’ used this information to buy these stocks and recommended them to his social network to combine profits from these trading transactions. Indeed, the agent agreed and signed that he would keep the matter confidential, but this was not the case and he breached this trust and violated the securities based on the insider information he acquired.

Table 4-9 shows how the linguistic patterns from different analyzers work together to model the timeline of the agent’s actions.
For example, the first sentence could be extracted through the following pattern: 
[<Date>+<Contextual>+<agreements>+<Institute>+<Institute>]. The patterns are run in sequence to achieve the required extraction and analysis.

Table 4-9 Timeline Event Analyzer Output Sequences

<table>
<thead>
<tr>
<th>Target Assets</th>
<th>Timeline Events</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Valencia Stock</strong></td>
<td></td>
</tr>
<tr>
<td>14/06/2002</td>
<td>after receiving written indications of interest union bank</td>
</tr>
<tr>
<td>union</td>
<td>submitted a second written indication of interest 24/06/2002</td>
</tr>
<tr>
<td>09/07/2002</td>
<td>a union employee e-mailed robert j. gallivan minnesota office manager</td>
</tr>
<tr>
<td>robert j. gallivan office manager</td>
<td>set up the call for the morning Jul-10</td>
</tr>
<tr>
<td>on the morning</td>
<td>10/07/2002 robert j. gallivan placed a call union's offices</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>acquired material, nonpublic information union 10/07/2002</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>bought 3,228 shares valencia</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>stock purchase 0.25% valencia</td>
</tr>
<tr>
<td>closing price</td>
<td>USD33.1 per share</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>unlawful profit USD9,455</td>
</tr>
<tr>
<td>Jul-10</td>
<td>25/07/2002 robert j. gallivan breached a duty of trust and confidence the c&amp;B Consulting Firm</td>
</tr>
<tr>
<td>Jul-16</td>
<td>25/07/2002 robert j. gallivan breached a duty of trust and confidence union</td>
</tr>
<tr>
<td>material, nonpublic information</td>
<td>received union's employees</td>
</tr>
<tr>
<td>Jul-10</td>
<td>04/08/2002 robert j. gallivan recommended the purchase valencia</td>
</tr>
<tr>
<td>friend</td>
<td>purchased 3,850 shares</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>combined profits USD8,636</td>
</tr>
<tr>
<td><strong>Sun Country Stock</strong></td>
<td></td>
</tr>
<tr>
<td>18/02/2003</td>
<td>the minnesota investor dinner meeting scheduled sun country chairman</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>acquired material, nonpublic information 18/02/2003</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>purchased 3,070 shares sun country</td>
</tr>
<tr>
<td>total cost</td>
<td>USD34,082</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>purchase 0.38% sun country</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>sold shares 30/04/2003</td>
</tr>
<tr>
<td>reaping illegal profits</td>
<td>USD2,758</td>
</tr>
<tr>
<td><strong>Mid Valley Stock</strong></td>
<td></td>
</tr>
<tr>
<td>Aug-03</td>
<td>premiwest offered purchase mid valley</td>
</tr>
<tr>
<td>mid valley</td>
<td>employee subsequently called the c&amp;B consulting firm 08/09/2003</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>purchased 5000 shares mid valley</td>
</tr>
<tr>
<td>purchase of</td>
<td>mid valley USD156,760</td>
</tr>
<tr>
<td>mid valley</td>
<td>closing price USD19.6 Sep-16</td>
</tr>
<tr>
<td>robert j. gallivan</td>
<td>illegal profit USD39,750</td>
</tr>
<tr>
<td>mid valley</td>
<td>an expectation confidentiality</td>
</tr>
</tbody>
</table>
The different layers of the financial fraud ontology were mapped to each other in order to integrate them into the ontology architecture shown in figure 4-6. The main type of mapping used was concept mapping. In each ontology, a number of classes were mapped to classes of other layers via semantic relationships which resulted in inter-layer mapping. Also, each layer has a relationship between concepts in the same layer called ‘intra-layer’ mapping. Inter-layer and intra-layer mapping were demonstrated using the PoolParty semantic relationship called ‘related (R)’. This section shows a number of ontology intra-layer and Inter-layer mappings. Figure 4-28 shows the corpus used in this ontology and how it is linked to different concepts within the layer or between layers.

The intra-layer mapping of the ‘litigationrelease’ concept is the ‘IE_Application\DocumentFormat\TextualFormat\SEC Litigation Releases RSS Feed’ and ‘IE_Application\DocumentFormat\TextualFormat\html’ which represents the document format used to evaluate the ontology. Given that some litigation releases have SEC complaints document, another semantic relationship is added as shown in figure 4-30. Furthermore, the litigation releases were analyzed through a text-mining application using different analyzers, such as ‘Action (defendant)’, ‘Date’, ‘Document Type’, ‘Plaintiff’, ‘Release No.’, ‘Short Description’, ‘Detailed Description’, and ‘Civil Case No.’.

In terms of Inter-layer mappings, the concept ‘Litigation Releases’ has relationships with the concept ‘SEC Website\Enforcement\litigation releases’ and ‘DocumentFormat\DocumentType\html’ in the data source ontology layer.
The navigation in PoolParty is intuitive, pressing the ‘SEC Complaint Document' concept in order to know what relationships are associated with it. Two concepts are considered as intra-layer mappings: ‘IE_Application\LanguageResource\Corpus\LitigationRelease’ and ‘Processing\Analyzer\See also\SEC Complain Document Analyzer’. Inter-layer mapping is expressed by one concept from the data source ontology layer ‘SEC Website\Litigation Release\See also\SEC Complain Document’.
Generally, each complaint document has five different sections describing the fraud case: complaint summary, defendants and relevant person or entities; facts; fraud for reliefs and violation; jurisdiction and venue; and signatures. The first section starts with ‘Complaint Summary’ which has given a brief description of the case, including information about the fraudster, the type of manipulation, manipulated securities and the market. Figure 4-32 describes the relationships of the ‘Complaint Summary’ concept with concepts from the domain and application layers. Inter-layer mapping is applied by linking the ‘Complaint Summary’ concept with concepts from the domain layer such as ‘Manipulation participant\Agent Characteristics\Individual’, ‘Manipulation participant\Agent Characteristics\Organization’, ‘Market manipulation types’, ‘Venue’ and ‘Asset’. Regarding the application layer, the concept ‘Complaint Summary’ was analyzed using analyzers such as ‘Individual’, ‘Organization’, ‘Market manipulation Type’, ‘Venue’, and ‘Asset’ analyzers.

Figure 4-32 Complaint Summary Section Relationships

In order to get details about the manipulation activities and actions used by the defendants to execute the manipulation, as shown in figure 4-33, the ‘Facts’ concept which contains the facts section of the SEC complaint document is used. This section includes how the defendants manipulated the securities, the benefits and effects of these actions, targeted and manipulated assets, the market and the
temporal dimension of these actions. This concept also has inter-layer mapping relationships with the domain and application layers. According to the domain layer, the concept has a semantic relationship with ‘Action’, ‘Venue’, ‘Time’, ‘Period’, ‘Organization’, ‘Individual’, ‘Effects’ and ‘Asset’. Two analyzers are used to analyze this section; thus, the concept has links to ‘Manipulation, Actions and timeline Events Analyzers’ and ‘Agent Characteristics Analyzer’.

The ‘Defendants and relevant person or entities’ concept has a relationship with the ‘Manipulation participant’ concept in the domain ontology and ‘Manipulation participant Analyzer’ in the application layer that extracts the key manipulators in the case with their social networks.

![Figure 4-33 Facts Section Relationships](image)

The timeline events in any SEC cases are important because they describe the period of each activity executed by manipulators. Generally, the manipulative activity is associated with the time dimension, the manipulator who performed the action, materials used by the manipulators, the trading of shares, and the benefits resulting from these actions. Thus, the ‘period’ concept illustrated in figure 4-34 has intra-layer mappings with concepts in the domain layer such as ‘Agent Characteristics’, ‘Asset’, ‘Assistance’, ‘Benefit’, ‘Material’, ‘Third party’, ‘Structured’
and ‘Morning’. The analyzer in the application layer that is used to analyze these concepts is the ‘Manipulation Actions and timeline events analyzer’.

![Figure 4-34 Period Concept Relationships](image)

The section of fraud for reliefs and violations has information about the different laws and regulations that have been referred to by the court against the defendants. Thus, as shown in figure 4-35, the concept ‘fraud for reliefs and violation’ has an inter-layer relationship across the architecture. The concept is mapped to ‘Laws and regulations’ concepts in the domain ontology, the lexical layer and using the Investigation, Enforcement and Regulator analyzer in the application layer to extract the concepts related to securities violations. ‘Jurisdiction’, ‘venue’ and ‘signatures’ concepts describe the court that investigated the case and charged the manipulators. This concept is mapped to the ‘District Court Name’ in the ontology and data source ontology layers and the ‘district court name analyzer’ in the application layer.
4.5 Summary

In summary, this chapter has presented a proposed method to construct a financial ontology for fraud purposes directly from a corpus. The researcher invited financial experts to participate in constructing the required financial concepts to build an accurate representation of the ontology. In practice, those experts identified the knowledge within the texts to construct an ontology with relevant financial fraud concepts, in order to enable the ontology to answer similar questions that could be asked by users or analysts reading those texts. Furthermore, they helped to provide a certain degree of depth to the organization of the financial concepts within the ontology.

Despite the fact that most ontologies use a flat architecture, this research proposed multi-layer architecture ontology adapted from Mikroyannidis et al. (2012). The architecture improves the manageability of the technologies and demonstrates the integration between different ontologies presented through intra-layer and inter-layer ontology mapping. Four layers are introduced: the lexical layer, domain layer, data source layer, and application layer. The lexical layer contains a domain-independent ontology of a purely lexicographical nature; it used the ‘Investopedia’ website as a lexical financial dictionary. Investopedia has
13,255 sub-concepts classified under 28 main concepts. The data source ontology layer exploits the semantic aspects of the SEC web and addresses its topology. This research focuses on the enforcement class because it contains information about the litigation releases, the topic of this thesis. The litigation releases provide lists of prosecuted cases concerning civil lawsuits brought by the Commission in the federal court. Each litigation release contains information, such as release number, release publication date and action, and most of the releases have external links to SEC complaints. The SEC complaint documents are the reports produced by the US district court that describes violation of securities laws cases. In these documents the court describes in detail all the evidence and facts that lead to its decisions against manipulators.

The domain ontology layer is constructed to model the financial domain for fraud purposes, using the main classes of the taxonomies proposed by Diaz (2011). However, Diaz' work has not been extended to construct a full ontology for the domain. Thus, this research contributes to building a financial ontology for fraud purposes based on a coherent corpus (SEC Litigation releases) which is relevant to financial fraud. The class hierarchy used a combination of ‘top-down’ and ‘bottom-up’ approaches. The financial domain experts acted as coders and annotated interesting financial concepts from the litigation releases. The research used a focus group methodology to allow participants’ perspectives to be revealed through discussion, questions and arguments.

Based on the analysis of the focus group, the domain ontology was built and developed using the PoolParty thesaurus management system. Each class includes general descriptions; links to different sources include linked data and instances as particular exemplars to fill the slot values of the class. Moreover, all sub-classes inherit the slots/properties of their super-class. The whole process of building the ontology was iterative in order to produce a version that could help building a common understanding of financial market fraud.

The application layer has the text-mining application that automates the process of extracting financial concepts from texts; it provides empirical evidence of how text mining could be integrated with the financial fraud ontology to improve the efficiency and the effectiveness of extracting financial concepts and demonstrating the published prosecuted cases in an appropriate knowledge base. This layer adapts the application ontology for information extraction tasks introduced by
Mikroyannidis et al. 2012). The lexical part of the application layer simulates the Investopedia website functionality. In particular, the application allows users to query specific financial terms, learn their meaning and explore other terms related to the specific term. In practice, the text-mining application should capture financial concepts from the documents and provide a full explanation of terms.

An advanced linguistic pattern approach is employed on the basis of natural language processing in order to linguistically analyze the litigation releases. 60 advanced linguistic patterns were developed in order to analyze the litigation releases sentence-by-sentence and to apply the focus group’s recommendations. Several analyzers were developed to provide answers to the key ontology questions: Who is (are) the agent(s) involved in the manipulation? Which asset is being targeted? In which venue is the manipulation taking place? What action has been performed or is planned? Is it a trade-based or an information-based action? Which pattern is associated with this manipulation? When was this manipulative action performed? Where is the manipulator getting his profit from?

Finally, the different layers of the financial fraud ontology were mapped to each other, intra- or inter-layer, for integration into the ontology architecture.
5 “Stock-touting” Spam E-mails Text-Mining Application

This chapter presents the instantiation of the proposed financial fraud ontology though demonstrating another possible text-mining application of the ontology. The ontology played a vital role in providing the framework for the extraction process and capturing information related to touted stocks. Moreover, the ontology demonstrates the extracted information in an organized and coherent way, a knowledge base for users to facilitate fraud detection and proactively monitor the behaviour of these stocks. This case study shows how to use the ontology architecture to map its resources into the different ontology layers. Furthermore, it describes the case study of a stock spam e-mail as an unstructured data source that demonstrates the process of refining linguistic resources to extract relevant, high quality information including stock profiles, financial keywords, stock and company news (positive/negative), and compound phrases from stock spam e-mails. The context of this study is to identify high-quality information patterns that can be used to support relevant authorities in detecting and analyzing fraudulent activities.

A version of this case study, together with parts of Chapter 2, was presented as a conference paper under the title Using Text Mining to Analyze Quality Aspects of Unstructured Data: A Case Study for ‘stock-touting’ Spam Emails, at the Americas Conference on Information Systems (AMCIS) 2010, (Proceedings, p. 364).

5.1 Background

The volume of available data on the stock market has increased exponentially over the last ten years. The data is both structured, represented by quote and trade data; and unstructured, such as spam e-mails, misleading press releases, forums, chat-rooms and blogs. In particular, manipulators and some promotion companies play a critical role in misleading investors, using different manipulation strategies based on textual resources to ‘hype’ and attracts investors to buy the promoted stocks. Stock spam e-mails have been used extensively and creatively to target specific stocks in order for fraudsters to gain unlawful profits. Most commonly, spammers claim that they have ascertained private information about stocks. The e-mails contain fine-print messages claiming valuable information such as investment advice and specific investment decisions disclosed with
financial terms and recent price quotes. Thus, stock spammers speculate on positive price models of the traded stocks and send thousands of e-mails to possible investors to drive the price of the touted stock upwards or downwards.

In this context, this research contributes to identifying gaps in the financial research as no previous work has considered the use of text-mining techniques for the analysis of "stock touting" in unregulated markets. Furthermore, the traditional techniques are time- and human resource-intensive, as well as being error-prone and largely ineffective in dealing with the massive volume of spam e-mails. Thus, this research demonstrates a novel application of text-mining analyzers that could automatically and efficiently extract key attributes and characteristics of different unstructured sources generated as part of "outing campaigns".

The proposed financial fraud ontology plays a vital role in providing the underlying framework for the extraction process and capturing information on touted stocks. Moreover, the ontology helps to present the extracted information in an organized and coherent way, effectively a knowledge base for users to facilitate fraud detection and monitor the behaviour of these stocks. In particular, the developed text-mining analyzers may help in raising proactive alarms in situations where traditional data-mining analyzers fail to predict correctly, or generate only weak signals of suspicious trading behaviour. Furthermore, this could help regulators and market participants to minimize the on-going “pump and dump” schemes that rely on circulation of rumours or illegal information.

5.2 Instantiation of the Stock Spam E-mail Text-Mining Application

This section describes the instantiation of the financial fraud ontology to automate the categorization and classification of stock spam e-mails. In particular, stock spam e-mails are used as another data source to evaluate the proposed financial ontology. Furthermore, the case study demonstrates empirical evidence of how text mining could be integrated and used with the financial fraud ontology to improve the efficiency and effectiveness of extracting concepts from these messages. The text-mining application employed the dictionary and advanced linguistic patterns (NLP) to extract information of financial fraud interest. This stock spam e-mail application produces lists of touted stock cited in these messages, associated with other information to support regulators in watching the behaviour of these stocks.
The case study evaluates the proposed financial fraud ontology through introducing another data source called ‘stock spam e-mails’. The spam e-mails dataset was collected from Richardson’s Stock Spam Effectiveness Monitor (SSEM) archive (Crummy 2006). This site monitors and filters spam e-mails through spam-trap addresses that receive an enormous amount of spam. About 15% of the spam is promoting some company or other. The site was set up in 2004 and closed down in early 2008 due to the difficulty of maintenance. A script was run each evening to find new stock spam in various inboxes and grab the latest price charts for those stocks and annotates them. It then used the accumulated data to write out reports. The scientifically based website was used to investigate ways to filter spam e-mails and evaluate their impact on the stock market. The archive contains spam messages from January 2006 to February 2008.

Figure 5-2 demonstrates the spam e-mail data source ontology layer for the spam e-mail sources. The spam e-mail messages provided by Richardson’s SSEM archive incorporates two concepts: metadata and body. The data source layer provides information about the key terms required to be extracted from these messages for better understanding of the spammers and spam e-mails. The research commenced by using 100 stock-touting spam messages for manual annotation from the large quantity of collected unsorted spam e-mail. Within the scope of this research, manual analysis was conducted: spam tout volume
analysis considering date and time, and content analysis. The manual analysis was utilized to produce the taxonomy of these messages. This taxonomy is based on key financial concepts, phrases and content analysis to help in developing the pattern and classifications that are used in the automatic process.

As shown in figure 5-2, the metadata concept deals with key header information such as date, time, sender, receiver, and subject. The body concept deals with named entities such as stock profiles, which includes information about the touted stock such as symbol ticker, the company holding the stock, and the sector that the company is related to. Furthermore, the spammers use some techniques to encourage investors to trade, such as price or volume projections, trading date, advice, recommendations, and some financial investment indicators, either long or short term; also, any financial signals such as buying signals. The ontology considers phrases that indicate whether the cited news is related to the stock (e.g. trading description) or the company itself (e.g. agreements). This feature may give positive or negative feelings to investors and encourage them to respond and trade following the recommendations given by the spammers.
Figure 5-2 Spam E-Mail Data Sources Ontology Layer

5.2.1 Application Ontology

The application ontology layer constructed for the spam e-mail case study represents the organization of the developed text-mining components. This layer adapts the application ontology for the information extraction (IE) tasks introduced by Mikroyannidis et al. (2012). However, the text-mining process is generally defined as a process workflow which consists of user-defined pipelines of analysis tasks that can be either pre-defined or customized by the user. For the purposes of this research, the proposed IE tasks were modified in order to define the specific text-mining process as shown in figure 5.3. In this case study, IBM SPSS Modeller version 12.0 (IBM 2012) was used as a text-mining tool to implement the proposed text-mining application.

Case study analysis involves understanding the business domain and the available datasets. In the spam e-mails dataset collected from the Richardson’s SSEM archive, the messages were originally archived in the Linux e-mail format.
‘Mbox’. In this first step, the messages are converted to a uniform format ‘TXT’ that can be used for further analysis. This conversion is performed internally and does not change the original data. The messages have been added to the ‘corpus’ class to be ready for the pre-processing tasks. In this case study, manual annotation was carried out as part of the case study analysis. The purpose of the manual annotation is to exemplify how text mining would work if the process were done manually, similar to the method applied in the previous spam e-mail literature explained in section 2.3.2.1. The process of manual annotation is ideally carried out by the researcher. It is applied to 100 spam e-mail messages. Figure 5-3 shows the manual annotation of the e-mail header which identifies metadata such as ‘Date’, ‘From’, ‘To’, and ‘Subject’.

![E-mail Header](image)

**Figure 5-3 Spam E-mail Header Annotation**

Regarding the body of the e-mail messages, figure 5-4 shows the concepts highlighted for extraction purposes. In the body, spammers claim that they have ascertained private information about Remington Ventures Inc. stock (RMVN). The e-mails contain fine-print messages claiming valuable information such as investment advice, stimulation of a specific investment decision disclosed with financial terms and recent price quotes. Thus, stock spammers speculate on positive price models of the traded stocks and urge investors to buy the stock in order to violate the stock prices. The output is a set of documents together with the concepts that correspond to each one. The output of the manual annotation process is sometimes called the golden set of documents and is assumed to be 100% correctly annotated (the training dataset); it can be used to evaluate the accuracy (precision and recall) of the extraction process in the training stage before deployment.
In general, the resource involves the pre-processing of document collections, the corpus, which contains the spam e-mail messages. ‘Processing resources’ tasks are aimed at analyzing unstructured or semi-structured documents and converting them into structured texts through the extraction of appropriate concepts. Only after this is completed can the documents be used in the “deployment” stage. A key component of the processing tasks in text mining is the evaluation process (testing the dataset) to ensure that the developed text-mining models are sufficiently accurate for the purposes of the analysis.

The two metrics normally used for this are precision and recall. More specifically, precision is the ratio of the extracted concepts that are correct (called True Positives or TP) over the total number of extracted concepts. The incorrectly extracted concepts are called False Positives or FP. So, precision is calculated as TP / (TP + FP). Recall is the ratio of correct concept instances that are extracted (TP) over the total number of correct concepts. To calculate recall, we need to calculate the number of False Negatives (FN) which represents the correct concepts that have been missed. So, recall is calculated as TP / (TP + FN).
5.2.1.1 Language Resource

Linguistic techniques very often make use of external resources to help with the analysis. These resources could relate to the specific natural language being used in the documents, such as dictionaries and thesauri. However, it is also possible that concepts in a particular document might refer to a specific domain, which is in our case is financial fraud. Thus, the language resources in the stock spam e-mail application include resources such as a thesaurus, list of terms, libraries of concepts, synonyms, and types (semantic groupings of concepts). These lexical libraries are used to develop and construct the advanced linguistic patterns shown in figure 5-6. For example, the finance library includes a list of common financial concepts and the opinion library holds concepts relating to sentiments (positive/negative/neutral).

In order to cover most of the touted stocks in the e-mail messages, the text-mining application has a library called ‘Stock profile’ containing an updated list of OTC traded stocks from 2006 to 2009, with information about 23,029 traded companies collected from the OTC market. In this list, information on the company name, symbol ticker, market tier, venue, region, city, and locale were provided. For
example, ACMAT Corp. is a traded company in Pink Sheets market with a symbol ticker ACMT. The company is a domestic company located in New Britain, CT. Generally, the symbol ticker contains 4 to 5 characters. During the extraction process, the e-mails were scanned and analyzed in order to identify symbol ticker words within the dictionary and map the symbols to an event called <Normal Cases>. However, these symbol tickers are issued without consideration of the lexical meaning, and thus may create conflicts with English words. For instance, words like “every”, “this”, “auto”, are listed as stock symbols. This could lead to conflict in the extraction process as many of these words will be mapped to the <NormalCases> event. In order to overcome this conflict, an alternative English dictionary set which contains a list of 509 English terms was added to map these cases to another event called <SpecialCases>. Generally, most of the e-mails contain the venues of the touted stocks and the market type, such as primary market, derivatives market, etc. Thus, a list of 216 stock market venues were added to the <Exchange> category and 16 market types are part of the text-mining application library.

The macros, which represent reusable patterns, are used to simplify the appearance of literals and word strings needing to be extracted, e.g. calendar words, positive, trading session status, finance, financial indicator. The macros include different forms of the tokens including misspelling, punctuation errors, etc. for accurate extraction. Overall, the model includes 21 macros that support the extraction process and pattern development.
In table 5-1, “Word_cal” includes a list of tokens related to time, such as month, week and day. “Word_symb” includes a list of tokens related to the symbol ticker word such as symbol, stock symbol, symb, sym, etsymbol, o.t.c sym bol, etc. “Verbs_Word” is another macro that holds tokens related to concepts concerning the verb ‘to be’, e.g. “is, are, was, were”, etc. “FINA_INDI” includes a list of tokens related to market indicators such as market cap, short term, short, shrt trm|sh0rt, long term, grade, status, etc. “mProp” contains a list of tokens related to common prepositions. Table 5-1 shows the values of the different macros developed for the stock spam e-mail text-mining application.

**Table 5-1 Macro Values**

<table>
<thead>
<tr>
<th>Macro Name</th>
<th>Values (tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word_cal</td>
<td>month</td>
</tr>
<tr>
<td>Pos_Word</td>
<td>high</td>
</tr>
<tr>
<td>SES_STA</td>
<td>open</td>
</tr>
<tr>
<td>FINA_INDI</td>
<td>market cap</td>
</tr>
<tr>
<td>SING_INDIC</td>
<td>buy</td>
</tr>
<tr>
<td>DurationWords</td>
<td>last</td>
</tr>
<tr>
<td>Weekdays</td>
<td>saturday</td>
</tr>
<tr>
<td>word_symb</td>
<td>symbol</td>
</tr>
<tr>
<td>OTC</td>
<td>otc</td>
</tr>
<tr>
<td>PKSym</td>
<td>.</td>
</tr>
<tr>
<td>News_Word</td>
<td>news</td>
</tr>
<tr>
<td>Verbs_Word</td>
<td>is</td>
</tr>
<tr>
<td>Pronouns_Word</td>
<td>he</td>
</tr>
<tr>
<td>NotW_Word</td>
<td>notification</td>
</tr>
<tr>
<td>mDet</td>
<td>any</td>
</tr>
<tr>
<td>mProp</td>
<td>in/on</td>
</tr>
<tr>
<td>Quest_Word</td>
<td>what</td>
</tr>
<tr>
<td>vQuant</td>
<td>little</td>
</tr>
<tr>
<td>Email_Reply</td>
<td>re</td>
</tr>
<tr>
<td>Comp_word</td>
<td>inc</td>
</tr>
</tbody>
</table>

In addition, word gaps were used to define a numeric range of tokens that may be present between two concepts. These word gaps are very useful when matching similar phrases that may differ only slightly due to the presence of additional determiners, prepositional phrases, adjectives, or other similar words. For example, @{0,3} means that a match can be made between the two defined elements if there are 0, 1, 2 or 3 concepts present, but no more than three words.

5.2.1.2 Processing Resource

The techniques used during the processing of document collections are classified into linguistic and/or non-linguistic techniques. Linguistic techniques take into consideration the natural language characteristics of the text in documents, e.g. syntax, grammar, dictionaries, etc., whereas non-linguistic techniques view
documents as a series of characters, words, sentences, paragraphs, etc. Non-linguistic techniques treat each document as a list of words, count the number of times that specific terms (single words or sets of words) appear within a document or a corpus, and calculate their proximity to other related terms by taking into consideration their physical proximity within the document or their presence in related documents.

The stock spam e-mail application used both techniques, although more emphasis was given to the linguistic-based patterns in order to process the e-mail messages and improve the efficiency and the effectiveness of extracting concepts from these messages. This advanced linguistic technique performs a deeper analysis of the words, phrases and syntax inside the e-mails and, thus, it helps to uncover knowledge of the underlying language and these e-mail messages. In particular, the output of the text-mining application is a list of touted stock that has been cited in these messages associated with other information that could support regulators in watching the behaviour of these stocks. Several techniques have been incorporated such as tokenization, part-of-speech, spelling and grammar, stop words, term extraction, and pattern-based information extraction.

In particular, NLP pattern-based information extraction has been employed in this application to extract information from spam e-mail messages and recognize interesting linguistic elements of touted stock. Analyzers have been developed utilising a pattern (or rule) language to classify these patterns into entity extraction and event extraction patterns. Entities correspond to basic linguistic elements such as person name, company name, product name, location, etc; they are also called named entities. Entities could also correspond to non-linguistic elements such as currency, digit, e-mail, etc. Events in spam e-mail messages are complex patterns that define extraction patterns consisting of named entities, non-linguistic entities and other concepts.

A. Spam E-mail Metadata Analysis

Following the data source ontology, the text-mining application aims to extract key information from the spam e-mail messages. The target concepts are metadata information such as ‘Date’, ‘From’, ‘To’, and ‘Subject’. The text-mining application used some of the predefined libraries incorporated in the IBM PASW 12 software, such as person, organization, date, e-mail, and time. However, these patterns did not capture all related concepts within the document. Thus, extra patterns have
been developed to fill the gap and to increase the accuracy of extraction. Furthermore, other analyzers have been developed from scratch to match the specific structure of linguistic patterns. This analysis gives information about who sends the e-mail, the recipients and when it was sent. The analyzers produced in this analysis could help regulators to know when the touting campaign started, how long spammers continued their campaigns for specific stocks and who sent it and to whom. This alert which targeted investors to the possibility of manipulation activity could be generated by these e-mails.

1. Header Analyzer

In this analyzer three patterns were developed to capture the main metadata. The first pattern `<$vSender> <$Email> | <$Person>` aims to extract the sender of the e-mails, represented by the term “From:” and followed by the person name such as “natalia britton” or the e-mail address such as Natalia.britton@bms.com. The pattern automatically analyses the e-mails and extracts the concept “from:” and classifies it under the `<Sender>` type; the concept “Natalia Britton” is classified under `<Person>` as shown in figure 5-7. Also, there is a possibility that the e-mail of the sender appear in the header Natalia.britton@bms.com and in this case it will be classified under `<email>` type. The precision accuracy of this pattern to capture the sender from the training sample was 88%.

![Figure 5-7 Sender Pattern](image)

The second pattern `<$vRecipient> <$MEmail>` was developed to extract the recipient of the e-mail, which in this case is considered as a target investor who...
could read, act on, and help spammers to violate stock prices. The pattern automatically extracts the concept “To:” and classifies it under “Recipient”, followed by extracting the e-mail address of the recipient and classifying it under <Email> as shown in figure 5-8. The precision accuracy of this pattern to capture the recipients from the training sample was 100%.

Figure 5-8 Recipient Pattern

The last pattern in this analyzer, ‘date and time stamp’, aims to extract the date and the time of receiving these messages. This pattern could help regulators to know when spammers start their campaign for a specific stock, and how long they continue their campaigns for specific stocks. As shown in figure 5-9, the pattern <$vDateDay> @{0,7} <$Date> <$vTime> automatically extracts the concept “Date:” and classifies it under <DateDay> type, followed by date and time. Word gap is used in this pattern to exclude any word tokens from the extraction process that are not important in the pattern output. Thus, the word gap “@{0,7}” is used to guide the pattern dealing with the existence of any irrelevant tokens without the necessity of showing it in the pattern output. Also, the spam e-mails are unstructured and could be associated with a wide range of tactics to deceive this type of automatic extraction system. The precision accuracy of this pattern to capture the timestamp of the e-mails from the training sample was 85%.
2. Subject Analyzer

In the manual annotation, the subject was used as a way to get attention and attract investors to open the message and see the contents. The subject has been classified into three main categories: news related, stock related, and market related. Generally, news-related subjects are the most common category used to attract investors, as it includes words such as notification news update, good news, attention updated news, etc. Stock-related subjects contain positive announcements about the stock, such as high stock performance, this stock is amazing, what are rolling stock, stock moves up, smart money stock, etc. Market-related stock holds market words such as OTC update, market calls, portfolio, top market timing, etc.

As shown in figure 5-10, this analyzer has a total of 48 patterns to extract automatically the “Subject” associated with the spam e-mails and assign it to a category. Overall, the precision accuracy of this analyzer to capture the “Subject” from the training sample was 93%. The highest percentage of e-mail messages had “News related subjects” (46%), followed by market-related subjects (27%) and stock-related subjects (20%).

2. Subject Analyzer

In the manual annotation, the subject was used as a way to get attention and attract investors to open the message and see the contents. The subject has been classified into three main categories: news related, stock related, and market related. Generally, news-related subjects are the most common category used to attract investors, as it includes words such as notification news update, good news, attention updated news, etc. Stock-related subjects contain positive announcements about the stock, such as high stock performance, this stock is amazing, what are rolling stock, stock moves up, smart money stock, etc. Market-related stock holds market words such as OTC update, market calls, portfolio, top market timing, etc.

As shown in figure 5-10, this analyzer has a total of 48 patterns to extract automatically the “Subject” associated with the spam e-mails and assign it to a category. Overall, the precision accuracy of this analyzer to capture the “Subject” from the training sample was 93%. The highest percentage of e-mail messages had “News related subjects” (46%), followed by market-related subjects (27%) and stock-related subjects (20%).
Figure 5-10 Subject Analyzer
As shown in figure 5-11, 25 patterns were developed to extract news-related subjects. In particular, the pattern “<Subject>+<Notification>+<News>+<News>” extracts linguistic patterns related to news or press notifications from 11 e-mail messages, such as “subject: notification – information release”, “subject: notification - press release”, “subject: notification - headline news”, “subject: notification-news report”, and “subject: notification news release”. Similarly, the pattern <Subject>+<Date><News>+<News> extracts patterns related to news notifications but with a date dimension such as “subject: 01/03/06 information release”, and “subject: 2006/01/17 news release”. Patterns such as <Subject>++<Positive++<News> extract all positive news such as “subject: good news”. Figure 5-9 shows examples of different news-related linguistic patterns extracted from the subject of the e-mail messages.

16 patterns were developed to extract stock-related subjects. Generally, this subject covers a wide range of positive messages about the touted stocks to trick investors and encourage them to act upon these campaigns. Figure 5-12 shows various examples of stock-related subjects extracted from the training sample. For example, spammers used words such as smart money stocks, produce maximized return, stock investing, high stock rollers, stock screaming options, stock moves up, etc. All these positive messages may affect investors, especially the naïve ones how want to earn money quickly. In particular, small investors in markets such as OTC are easy to mislead because the size of trading is small compared to
the regulated market. In these patterns, the opinion library is incorporated to include various positive words in order to extract information relating to sentiments (positive/negative/neutral). However, it is apparent that spammers used more positive words in the subject context even though negative words, such as screaming, and red hot have been used in positive context.

Furthermore, this analysis has observed reply e-mails, indicating some communication between spammers and target investors. This confirms the assumption that there are people who read these messages, correspond about the attached information, and act upon it. Previous work has shown that the motivations vary, because there are people are naïve, while others help the spammers in his manipulation activity for profit sharing. The stock spam touting text-mining application employs analyses that could help regulators and relevant users, such as exchanges to alert their investors about these sorts of manipulative activity.

Finally, the market-related subject category has seven patterns to extract any linguistic patterns relevant to the market, including OTC update, market calls, market update, and top market timing. Figure 5-13 shows various examples from the market-related subject category. In particular, “OTC update” is frequently used by spammers as the subject of their campaigns. In the training sample, 18 messages were captured using this pattern.

Thus, <Subject>+<Market>+<TimeWords> was developed to capture this linguistic pattern from e-mail messages.
B. Spam E-mail Body Analysis

Following the spam e-mail data source ontology, the text-mining application aims to extract key information from the body of the spam e-mail messages. The target concepts are finance-related information such as ‘Stock symbol’, ‘Company name’, ‘Market’, and ‘Stock news’ with different financial indicators. As mentioned above, the e-mails contain fine-print messages claiming valuable information such as investment advice, stimulation about a specific investment decision disclosed with financial terms, and recent price quotes. Thus, stock spammers speculate on positive price models of the traded stocks and urge investors to buy the stock in order to violate the stock prices.

As shown in figure 5-14, four analyzers were developed to extract such information from the body of the messages. The analysis could help the regulator to investigate the spammer’s techniques regarding the touted stocks. Sometimes the information enclosed in the e-mails is correct, based on non-public information collected by insider trading manipulators who use ‘pump and dump’ schemes in order to increase the prices. Generally, they use a third party to send such e-mails, to hide their identity and cover their manipulation activity. The spam e-mail body analyzers could help regulators to compare the attached information with the actual trading behaviour of the touted stock and could also be an indicator of an insider trading scheme if the information is correct. Furthermore, a proactive alarm could be produced based on this analysis before the actual execution of fraud.
Figure 5-14 Spam E-mail Body Analyzers

1. Symbol Ticker Analyzer

The symbol ticker contains 4 to 5 characters. During the extraction process, the e-mails were scanned and analyzed in order to identify symbol ticker words within the dictionary and map the symbols to an event called <NormalCases>. However, as explained above, an alternative English dictionary set of 509 English terms was added to map cases of confusion with normal English words into another event called <SpecialCases>. Spammers also use tactics to deceive the automatic filtration systems by changing the way they cite the symbol, e.g. (RMVN.PK, R-M-V-N, R M V N, etc.) which obviously will not be identified by using the dictionary method. The pattern-based information extraction and regular expression could help to address these challenges by employing natural language processing (NLP), with patterns designed as a set of rules to describe structures in words, combinations of them or combinations of rules. For example, if the task is to find or recognize stock symbol tickers that have been mentioned in the e-mail body or text, it is possible to define a rule in which all collections of 4-5 capital letters are mapped to an entity called <SpecialPattern>. An advantage of this approach is that it is possible to extract not only stocks tickers that were originally listed in the dictionary, but also to extract newly issued symbol tickers that match the predefined rules. Table 5-2 shows how these
patterns are developed to extract the different formats of symbol tickers cited in the training dataset.

Table 5-2 Regular Expression Symbol Ticker patterns

<table>
<thead>
<tr>
<th>Symbol Ticker Special Patterns</th>
<th>Extraction Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>regexp1=[a-zA-Z]{1} [a-zA-Z]{1}</td>
<td>R.M.V.N</td>
</tr>
<tr>
<td>regexp2=[a-zA-Z]{1},[a-zA-Z]{1}</td>
<td>R.M.V.N.PK</td>
</tr>
<tr>
<td>regexp3=[a-zA-Z]{1}[a-zA-Z]{1}</td>
<td>R.M.V.N</td>
</tr>
<tr>
<td>regexp4=[a-zA-Z]{1},[a-zA-Z]{1}</td>
<td>R_M_V_N.PK</td>
</tr>
<tr>
<td>regexp5=[a-zA-Z]{1}[a-zA-Z]{1}</td>
<td>R_M_V_N</td>
</tr>
<tr>
<td>regexp6=[a-zA-Z]{1}[a-zA-Z]{1}</td>
<td>R-M-V-N</td>
</tr>
<tr>
<td>regexp7=[a-zA-Z]{1},[a-zA-Z]{1}</td>
<td>R-M-V-N.PK</td>
</tr>
<tr>
<td>regexp8=[a-zA-Z]{1}[,;]?[a-zA-Z]{2}</td>
<td>RMVN,PK</td>
</tr>
<tr>
<td>regexp9=[a-zA-Z]{4}</td>
<td>RMVN</td>
</tr>
</tbody>
</table>

Moreover, the pattern-based method utilizes word association analysis to identify the relationship between two or more entities. For example, during the manual annotation it was noticed that the word ‘symbol’ is generally followed by a symbol ticker, such as Symbol: RMVN, as shown in figure 5-15. Additionally, different forms of the word symbol have been considered, such as ‘stock symbol’, ‘target symbol’, “symb’, ‘sym’, ‘etssymbol’, ‘o.t.c. sym bol’, ‘o.t.c symb0l’, and others. There is the further possibility that the stock exchange name or market venue may precede the symbol ticker, such as “OTCPK: RMVN”. Again, another pattern called MKTSymbol was developed to resolve this issue, also considering different combinations of the market names like “otcP”, "otc","otc.bb", “otcbb”, “other otc”, and others. Overall, this analyzer holds six patterns aiming to extract the different cases of the symbol ticker from the e-mail messages. The precision accuracy of this analyzer to capture the “symbol ticker” from the training sample was 100%.
2. Company Name Analyzer

Company name is another concept of the stock profile required to be extracted from the spam e-mail data source ontology, and this analyzer aims to extract the company name that has been touted in the e-mail messages. The text-mining application used some of the “Organization” library incorporated in the IBM PASW 12 software to extract the company name. To increase the accuracy of capturing the company name, the text-mining application used the companies list “OTC Company” (a list of 23,029 companies collected from the OTC market) provided in the library called “Stock profile” from the language resources.

Similarly to the symbol ticker analyzer, the pattern-based one utilizes word association analysis to identify the relationship between the word ‘company:’ or ‘company name:’ it is generally followed by a company name, as shown in figure 5-16 “company: Remington Ventures Inc.”. During the manual annotation it was noticed that another pattern of the company name may come after the ‘symbol ticker’, such as “RMVN: Remington Ventures Inc”. Overall 4 patterns have been developed to extract the different forms of company name and to be classified to the stock profile class in the ontology. Overall, the precision accuracy of this analyzer to capture the “Company” from the training sample is 89%.

3. Market Analyzer

Following the spam e-mail data source ontology, “market” is another concept of the stock profile required to be extracted from the spam e-mails. This analyzer has only one pattern to extract the market mentioned in the e-mail messages, usually the ‘OTC market’. Overall, the precision accuracy of this analyzer to capture the “market” from the training sample is 100%.

<table>
<thead>
<tr>
<th>Slot1 Concept</th>
<th>Slot2 Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>symbol</td>
<td>RMVN</td>
</tr>
<tr>
<td>symbol</td>
<td>RMVN</td>
</tr>
<tr>
<td>sym:</td>
<td>RMVN</td>
</tr>
<tr>
<td>otk:</td>
<td>RMVN</td>
</tr>
<tr>
<td>symbol:</td>
<td>RMVN</td>
</tr>
<tr>
<td>symbol:</td>
<td>RMVN</td>
</tr>
<tr>
<td>symbol:</td>
<td>RMVN</td>
</tr>
<tr>
<td>symbol:</td>
<td>RMVN</td>
</tr>
</tbody>
</table>

Figure 5-15 Symbol Ticker Analyzer
4. Stock Price Quotation Analyzer

The work of (Nelson, Price et al. 2009) argued that the target price quote and copy press release information advertized in spam e-mails could trigger a significant market response. Information such as target price or volume projections, trading date expectations, recommendations, and financial signals like buying or selling signals could have a significant impact on investors’ reaction toward the touted stock.

Therefore, analyzers such as target stock price and volume, financial long- or short-term investment indicators, speculative trade date, and buy or sell signals have been developed. Overall, 21 advanced linguistic patterns were developed to extract the different patterns of stock price quotes and volume projections. This analyzer used “Digit” and “currency” entities, pre-defined non-linguistic elements incorporated with the IBM PASW 12 software. However, these non-linguistic patterns did not capture all related concepts within the document. Thus, extra patterns have been developed to close the gap and to increase the accuracy of extraction.

Furthermore, other analyzers were developed from scratch to match the specific structure of linguistic patterns. Overall, the precision accuracy of this analyzer to capture the “Stock quotes” from the training sample was 100%. The examples in Table 5-3 explain how the patterns are automatically developed and classified into a Stock price quotation analyzer. Spammers use different linguistic patterns to enclose financial information, which could be of interest to investors and motivate them to trade. These patterns include the current price of the stock, such as “last/current/target price: .17”. Also, it could include some historical prices such as “Last trade jan 4th: $1.80, Days open/close: $1.80” to increase the credibility of the information in case investors check the historical behaviour of the touted stock.
Some e-mails have some price and volume projections for specific dates, such as “Jan 25th Volume: 650 share, December 19th: $0.65”. Spammers provide some positive financial indicator strategy speculation such as “Long term: $45-$55, Short term: (1) week: $2.50 - $3.50, Price increase: 42%, lifted now to $2.75, up $0.77 (150%)”, followed by trading recommendations such as “Indicator: strong Buy” in order give the impression that they are financial professionals and have private information about the touted stocks.

Table 5-3 Stock Quote Patterns Development

<table>
<thead>
<tr>
<th>Pattern Syntax</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>$vDate $vFinance $vDigit $vFinance</td>
<td>Jan 25th Volume: 650 share</td>
</tr>
<tr>
<td>$vFinancialIndicator $vCurrRange</td>
<td>Long term: $45-$55</td>
</tr>
<tr>
<td>($vDate$Wekdays) $vFinance $vDigit</td>
<td>Friday Volume: 120000</td>
</tr>
<tr>
<td>$vCalendar $vPositive $vPercent</td>
<td>1 month increase: 265%</td>
</tr>
<tr>
<td>$vFinance $vPositive $vPercent</td>
<td>Price increase: 42%</td>
</tr>
<tr>
<td>$vPositive $vCurrency $vPercent</td>
<td>up $0.77 (150%)</td>
</tr>
<tr>
<td>$vDate $vPositive $vCurrency</td>
<td>Jan 25th High: $0.65</td>
</tr>
<tr>
<td>$vCalendar $vSessionStatus @(0,3) $vCurrency</td>
<td>Days open/close: $1.80</td>
</tr>
<tr>
<td>($vDate$Wekdays) $vSessionStatus $vCurrency</td>
<td>jan 4th/Friday open/close: $1.80</td>
</tr>
<tr>
<td>$vDurationWords $vFinance $vDate $vCurrency</td>
<td>Last trade jan 4th: $1.80</td>
</tr>
<tr>
<td>$vDurationWords $vFinance $vCurrency</td>
<td>last/current/targetprice: .17</td>
</tr>
<tr>
<td>$vDate $vCurrency $vMoneyUnit</td>
<td>December 19th: $4.62 million</td>
</tr>
<tr>
<td>$vDate $vCurrency</td>
<td>December 19th: $0.65</td>
</tr>
<tr>
<td>$vFinance $vCurrency</td>
<td>Price: .17</td>
</tr>
<tr>
<td>$vFinancialIndicator $vCurrency</td>
<td>Long term: $45</td>
</tr>
<tr>
<td>$vFinancialIndicator $vDigit $vCalendar $vCurrRange</td>
<td>Short term: (1) week: $2.50 - $3.50</td>
</tr>
<tr>
<td>$vFinancialIndicator $vPositive $vSignalIndicator</td>
<td>Indicator: strong Buy</td>
</tr>
<tr>
<td>$vFinancialIndicator $vSignalIndicator</td>
<td>Ind icator: Buy/strong buy</td>
</tr>
<tr>
<td>$vPositive $vDurationWords to? $vCurrency</td>
<td>lifted now to $2.75</td>
</tr>
<tr>
<td>$vFinancialIndicator $vWeights-Measures</td>
<td>Market cap: 1.6M</td>
</tr>
</tbody>
</table>

5.2.1.3 Deployment

A key component of the processing tasks in text mining is the evaluation process that ensures that the developed text-mining analyzers are accurate enough for the purposes of the analysis, and again, the precision metric was used for the spam e-mail text-mining application. The analyzers were evaluated through the testing dataset of 26,947 stock spam e-mail messages from Richardson’s SSEM archive. Overall, the performance of most of the analyzers was good; the precision of the symbol ticker analyzers was 99%, the company name analyzer 99.92 %, the recipient and sender analyzer 99% precision, and the date and time stamp 87% precision. However, in full deployment, the subject and stock quote analyzers require further improvement and elaboration of the design of patterns, in order to better address analysis requirements and increase the model’s accuracy.

5.2.2 Ontology Mapping

The different layers of the financial fraud ontology were mapped to each other for integration into the ontology architecture, as shown in figure 5-1. The main type of
mapping used was concept mapping. In each ontology layer, a number of classes were mapped to classes of other layers via semantic relationships, resulting in inter-layer mapping. Also, each layer has relationships between concepts in the same layer, the ‘intra-layer’ mapping. Inter-layer and intra-layer mapping were carried out in the PoolParty semantic relationship.

This section shows examples of ontology intra-layer and inter-layer mappings. Figure 5-17 shows the different relationships of the symbol ticker analyzer in the application ontology layer and how it is linked to different concepts within the layer or different layers. The analyzer has intra-layer mapping in the application ontology layer represented by a semantic relationship with ‘LanguageResource\LexicalResource\StockSpamLibrary\NormalCases’, ‘LanguageResource\LexicalResource\StockSpamLibrary\SpecialCases’, and ‘LanguageResource\ LexicalResource\ StockSpamLibrary\market’. Furthermore, the analyzer has inter-layer relationships with the data source layer and domain layer. The analyzer linked these to the “Spam e-mail archive\Stock profile\ stock symbol” concept in the data source ontology layer, this is linked to the “Asset” concept in the domain ontology.

![Figure 5-17 Symbol Ticker Analyzer Relationships](image)

The stock price quotation analyzer is another example of different relations between the application layer and the domain, lexical ontology and data source
layers. In particular, the analyzer has intra-layer mapping in the application ontology layer, represented by a semantic relationship with “LanguageResource\LexicalResource\finance”, to the “Spam e-mail archive\Body\stock quote” concept in the data source ontology layer and the “Effects\evidences and patterns\Financial economics\structured” concept in the domain ontology. Finally, the analyzer is linked to different Investopedia concepts such as finance, financial acronyms, financial buzz words and active trading.

5.3 Summary

This chapter describes a case study with stock spam e-mails as an unstructured data source, demonstrating the process of refining linguistic resources to extract relevant, high-quality information including stock profiles, financial key words, stock and company news (positive/negative), and compound phrases from stock spam e-mails. The context of this study was to identify high-quality information patterns that can be used to support relevant authorities in detecting and analyzing fraudulent activities.

In particular, manipulators and some promotion companies could play a critical role in misleading investors, using different manipulation strategies based on textual resources to ‘hype’ and attracts investors to buy the promoted stocks. The e-mails contain fine-print messages claiming stimulating specific investment decisions disclosed with financial terms and recent price quotes. Thus, stock spammers speculate on positive price models of the traded stocks and send thousands of e-mails to potential investors to drive the price of the touted stock upwards or downwards.

In this context, this research contributes to identifying the gaps in the financial research studies, as none of the previous works have considered the use of text-mining techniques for the analysis of "stock touting" in unregulated markets. Furthermore, these traditional techniques are both time- and human resource-intensive and error-prone, and they are ineffective in dealing with the massive volume of spam e-mails.

Thus, the research demonstrates a novel application of text-mining analyzers that could automatically and efficiently extract key attributes and characteristics of different unstructured sources generated as part of "outing campaigns". The proposed financial fraud ontology played a vital role in providing the framework for
the extraction process and capturing information related to touted stocks. Moreover, the ontology helped to present the extracted information in an organized and coherent way, a knowledge base for users to facilitate fraud detection and proactively monitor the behaviour of these stocks. In particular, the developed text-mining analyzers could help in raising proactive alarms when traditional data-mining analysers do not predict correctly or generate weak signals of suspicious trading behaviour. This could help regulators and market participants to minimize the on-going “pump and dump” schemes that rely on the circulation of rumours or illegal information.

The text-mining application employed the dictionary and advanced linguistic patterns (NLP) to extract information of financial fraud interest. More emphasis was given to the linguistic-based patterns in order to process the e-mail messages and improve the efficiency and effectiveness of extracting concepts from these messages. The text-mining application extracts concepts from e-mail headers: ‘Date’, ‘From’, ‘To’, and ‘Subject’. This analysis gives information about who sent the e-mail, who are the recipients and when it was sent, enabling regulators to ascertain when the touting campaign started, how long spammers continue their campaigns for specific stocks, and who sent the information and to whom it was targeted. Furthermore, regulators could warn the targeted investors and alert them to the possibility of manipulation activity generated from these e-mails.

Body analysis of the messages aims to extract key information, such as ‘Stock symbol’, ‘Company name’, ‘Market’, and ‘Stock price quotes’ with the different financial indicators, and price speculation. Thus, the spam e-mail body analyzers could help regulators to compare the information with the actual trading behaviour of the touted stock, which might be an indicator of insider trading schemes. Furthermore, a proactive alarm could be produced based on this analysis before the actual execution of the fraud. Overall, the performance of the most developed analyzers was good, in particular the symbol ticker, company name, recipient, sender, date and time stamp analyzers. However, the other analyzers need further improvement, and iterations of the design of patterns in order increase the accuracy.
6 BI Financial Market Service system for a financial market monitoring and surveillance system

This chapter presents the instantiation of the proposed financial fraud ontology though demonstrating another possible application of the ontology: the business intelligence (BI) service architecture for a financial market monitoring and surveillance system in which different components interact in a coordinated way with internal and external service providers to produce proactive alarms for possible securities fraud cases. The proposed BI service system is demonstrated through an exemplar case study of text mining and data mining to analyze the impact of ‘stock-touting’ spam e-mails and misleading press releases on trading data.

The BI service system is another application of the financial fraud ontology to show how an independent service provider could have helped in raising alarms about a possible ongoing ‘pump and dump’ scheme. The proposed service architecture extends the Market Monitoring Framework (MMF) proposed by (Díaz, Zaki et al. 2011), through incorporating the automated linguistics-based text mining required to extract the key concepts of spam e-mails and press releases, relate them to other available information and highlight the likelihood of their being part of ‘stock touting campaigns’.

The evaluation of the proposed service system is carried out through an exemplar case that relates to a real case from the over-the-counter (OTC) market and which was prosecuted by SEC. Through this, it is explained how the financial fraud ontology could be incorporated within the existing fraud analysis processes, the extent to which the processes could be automated, the relationships with the other types of analysis and the role that the fraud analysts play.

A version of this case study, together with parts of Chapter 2, was presented as a conference paper under the title of Financial Market Service Architectures: A “Pump and Dump” Case Study at the SRII 2012 Global Conference, 23th-27th of July, San Jose, California, USA; the proceedings will be published by IEEE in a forthcoming special issue of the journal SRIII/IEEE.

6.1 Background

The ‘Pump and Dump’ manipulation scheme is an ongoing market manipulation scheme that presents several challenges for investors and financial authorities.
Although rules and regulations have been toughened and authorities have taken drastic actions against fraudsters, existing work shows that most manipulation cases happen in unregulated markets such as OTC and Pink Sheets, because they are relatively small, illiquid, necessitate lower disclosure requirements for listed firms, are prone to conflicts of interest, lack accurate public information (European Commission 2010), and are subject to less regulatory oversight (Aggarwal and Wu 2006).

Indeed, accurate information is the best tool for investing wisely in the financial market, but its scarcity means that investors, especially microcap investors, have to rely heavily on unofficial sources of information such as unconfirmed press releases, forums and spam e-mails. In particular, fraudsters pump microcap stocks by releasing false messages that urge investors to buy or sell stocks; they then make significant profits by dumping the stock when the prices artificially rise. Substantial fluctuations in price, volatility, returns and trading volume are associated with stock touting campaigns, and the fact that there are recipients of the messages who act upon the information. This is especially true for unregulated markets such as OTC, where investors face extremely illiquid markets, accompanied by the additional risk of investing in companies which are mostly small in size, or start-ups.

Thus, the continuous improvement and development of financial market surveillance is essential to guarantee an efficient market. Recently, the European Commission and the US securities administrators recommended that regulators pursue further studies and research into the development of a comprehensive surveillance system to supplement current market surveillance. This research discusses a BI service system that considers the perspective of a regulating authority in the form of a market analyst.

Normally, the analyst has wide access to different sources such as market data, including unstructured sources such as spam e-mails, press releases, forums, news or financial forms and filings. The system extends the Market Monitoring Framework (MMF) (Díaz, Zaki et al. 2011) through incorporating automated, linguistics-based text-mining analysis to help raise proactive alarms and highlight potential fraudulent patterns generated as part of a ‘touting campaign’. The system would be of interest to customers who want analysis of stocks that have been touted in different unstructured sources such as spam e-mail, press releases,
forums and social media channels, in order to raise the alarm about these stocks and avoid any manipulation activity.

The proposed BI service system contributes to identifying the gaps in financial research, as none of the previous studies have considered the use of automated, linguistic-based text-mining techniques for the analysis of ‘stock touting’ in unregulated markets. In particular, this research demonstrates the impact of using text-mining techniques to automate the analysis and classification of large amounts of financial textual sources, identifying interesting patterns and relationships between them and helping fraud analysts to proactively investigate possible manipulation cases.

The system considers the current and potential characteristics of market-monitoring systems and demonstrates how the proposed financial fraud ontology could be used within the existing fraud analysis processes, the extent to which the processes could be automated, the relationships with other types of analysis and the role of the fraud analysts. This could help regulators and market participants to minimize the risks from ‘pump and dump’ schemes. The BI service system case study investigated “pump and dump” from the OTC and Pink Sheets market.

6.2 Financial Ontology Instantiation as BI Financial Market Service system

This section demonstrates the instantiation of the BI financial market monitoring system as a fraud detection service. Figure 6-1 shows that the system considers the use of different sources of financial and textual data in the data source layer, such as spam e-mails, press releases and IFDF forms. From the application perspective, the system could help fraud analysts to discover patterns in trading indicators that are related to market manipulation activities. The BI market service system combines various techniques of text mining and data mining to analyze structured and unstructured data that can be performed either inside or outside the main system boundaries.

The proposed BI system extends the Market Monitoring Framework (MMF) proposed by (Díaz, Zaki et al. 2011) to investigate potential fraud cases, analyzing “pump and dump” activity from the OTC and Pink Sheets market to provide empirical evidence of how the system could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of detecting fraud cases. In this particular case, fraudsters created an artificial demand on the Mobile Ready
Entertainment Corp. stock (symbol: MRDY) a publicly-traded company quoted on the Pink OTC Markets, Inc. (‘Pink Sheets’) through the distribution of false and misleading press releases and possibly spam e-mails. The investigation found that several press releases contained disingenuous revenue projections, and mentioned indistinct business relationships.

The information can be found at the following website: http://www.sec.gov/litigation/complaints/2008/comp20644.pdf. The MRDY case has been selected because it presents an instance of a ‘pump and dump’ microcap manipulation which matches the study interest as it highlights the need for automated, linguistics-based text mining. In addition, the case demonstrates how the BI financial market service system could play a vital role in proactively monitoring, detecting, and deterring abuses in the OTC markets.

The financial fraud ontology could play a vital role between information management and business intelligence analysis components. In this instantiation, structured data pre-processing and outlier detection tasks are performed inside the system. In particular, trading data is used to detect unusual trading volumes, jumps in prices and imbalances between bids and asks trading orders. Unstructured data is also processed internally, with the option of requesting additional external services analysis.

Thus, the analyst has the choice of integrating information provided by external services agents or relying solely on the internal capabilities of the system. Unstructured sources filing could be used to create proactive indicators, alerts and attention-grabbing devices of potential ‘pump and dump’ manipulation cases, especially if these events are accompanied with significant changes in trading behaviour.
6.2.1 Data Source Layer

The case study uses various sources such as spam e-mails, press releases, and IFDF (Form 144). The spam e-mails dataset was collected from Richardson’s SSEM archive (Crummy 2006). This site monitors and filters spam e-mails through using spam-trap addresses that receive an enormous amount of spam. The archive contains spam messages from January 2006 to February 2008. Spammation website is another source of spam e-mails are used in this case. It has web feeds of information about stocks advertised by unsolicited commercial e-mails. Figure 6-2 demonstrates the spam e-mail data source ontology layer for the spam e-mail sources adopted for the previous case study, again using spam messages from the SSEM archive.
The data source ontology layer also includes information relating to business news and financial events collected from the Factiva database. Factiva includes comprehensive news resources providing international/regional news, trade and industry publications, investment analysis, stock exchange feeds, press releases and international news feeds dating back 10 years. A wide range of detailed information about companies, industry/market trends, new product development, ad-spend, mergers and acquisitions, and economic information is supplied. Figure 6-3 shows how the news is classified into concepts such as the company cited in the news, its related sector, the subject of the news, source of the news, the region, and the language used to write the news reports. Some extra options such as look-up are incorporated to facilitate users’ searching (Factiva 2007).
A third source is from Wharton Research Data Services (WRDS), a comprehensive, Internet-based data research service which provides the user with over 200 terabytes of data across multiple disciplines including finance, marketing and economics. Figure 6-4 shows that WRDS has 23 concepts across multiple datasets compiled from independent sources that specialize in specific historical data, such as Capital IQ, NYSE Euronext, CRSP, and Thomson Reuters. In particular, Thomson Reuters, a global provider of data, analysis and information tools, offers the most comprehensive range of indispensable, market-customizable solutions to help users make better decisions, be more productive and achieve financial results (Wharton Research Data Services 2009). In particular, it provides data about 13F institutions, insider filings, mutual funds, and WRDS-Reuters DealScan.

The case analyzed in this chapter illustrates the retrieval of data from the Insider Filings Data Feed (IFDF) database (Wharton Research Data Services 2009), a source for all US insider activity, as reported on forms 3, 4 and 5, tables 1 and 2, average returns, amendments concordance and form 144. This last form must be filed with the SEC by an affiliate of the issuer as a notice of the proposed sale of securities under Rule 144, when the amount to be sold by the affiliate during any three-month period exceeds 5,000 shares or units or has an aggregate sales price in excess of $50,000.

A person filing Form 144 must have a serious intention of selling within a reasonable time the securities referred to on the form. The research collected a total of 37 items, all of them corresponding to Form-144 filings. The IFDF does not contain any other insider filing forms of Mobile Ready Co. for 2007. Finally, the
EDGAR database was used from the SEC website as a source of information regarding the proceedings actions related to the MRDY case.

6.2.2 Application Layer (BI Financial Market Service System)

This section demonstrates the BI financial market service system as an instantiation of the market monitoring framework (MMF) introduced by (Díaz, Zaki et al. 2011). Figure 6-5 presents five concepts: Data Sources, Information Management, Knowledge Base, Fraud Detection Engines, and Output Management. Data sources are those data sets that could be used in the analysis, including structured datasets such as stock quotes and stock trades, and unstructured, textual sources.

These data sources can initially be explored by an information management class where different processes for data preparation and pre-processing take place. In this layer, many BI techniques are incorporated, such as data mining, text mining, online analytical processing (OLAP), and social network analysis. The output from
the information management class is saved in the knowledge base, which may contain knowledge related to previously known cases, watch lists, or previously known patterns of manipulation.

The fraud detection engine class comprises two concepts, offline and real-time engines. In the offline engine, structured data along with updated patterns and models are used to perform a wide range of analysis, which can run automatically or on demand. Two analyses are incorporated, behavioural and economic analysis. The key difference between these two concepts is that behavioural analysis focuses on the actions and characteristics of the manipulators, and the economic analysis considers only the effects and consequences of these actions. The real-time engine includes data polling, circuit breakers and exchange rules monitoring, using for example ‘complex events processing’ components. The output management concept performs the task of updating the knowledge base, and generates investigation reports, alerts, and case management tasks. It includes two tasks: investigation and enforcement; and rapid dissemination. In investigation and enforcement, the authority decides to start investigate cases and collect evidence to follow the correct procedures of prosecution and apply market rules. Rapid dissemination involves all the required actions being reported to other key participants, including the markets involved in the manipulation.
6.3 Case Study

This section analyzes the case study to demonstrate the deployment and evaluation of the proposed BI financial market service system. The domain ontology helped to produce the timeline event of the different manipulation activities performed by the manipulator. As represented in figure 6-6, fraudsters issued two misleading press releases in January and February 2007 respectively. On March 8th 2007, and consecutive days, an unidentified agent sent several spam e-mail messages. On the 10th of March 2007, an independent stock spam tracker, called spamnation.info, reported on the presence of this touting campaign. On an undetermined date after the 10th of March the president of the company (Mora, 2008) sent a disclaimer e-mail to SPAMNATION.info stating that ... *spam promoting MRDY.PK: was sent out by an investor trying to create a market to sell to. The company was in no way responsible for nor connected to the spam mail.* After the spam e-mails campaign, the record shows that a number of Forms-144 were completed and submitted to the SEC during April and June 2007, indicating
the intention of the directors of the company to sell their shares. This registration occurred after the actual selling of the shares. Following the submission of Forms-144, a number of misleading press releases were issued on consecutive days towards the middle of July 2007. These press releases contained false statements of material fact and mentioned a contract with a company called ‘Simple Fit Inc.’

The BI financial market service system findings correspond with what was reported by the SEC investigation. In July 2008, the SEC published a case against the fraudsters stating that between January and July 2007, defendants directly and indirectly made use of sixteen separate press releases, many of which contained false statements of material facts. For example, on 19th of July 2007, manipulators issued a press release falsely claiming that Mobile Ready had been awarded an exclusive contract with Simply Fit Holdings, Inc. Furthermore, they claimed that the contract with Simply Fit Holdings could yield an excess of one million dollars of revenue for [Mobile Ready] over the next year...’ The two managers of the company had personally received 76.9 million of Mobile Ready shares. However, these shares had never been registered with the Commission or subjected to an applicable exemption from registration. While the market for Mobile Ready shares was artificially inflated, the fraudsters began to sell shares into the public market. However, the SEC case does not mention the role, if any that spam e-mails played in the touting campaign.

**Figure 6-6 MRDY Case Study Timeline**

![Figure 6-6 MRDY Case Study Timeline](image-url)
6.3.1 Data Sources

The case study uses various sources such as real trading data retrieved from the Bloomberg data service provider, as well as a collection of unstructured data sources related to the MRDY Company such as spam e-mails, press releases and IFDF forms. For ‘structured DataSource’, the research had access to MRDY closing prices, including bids and asks information, for the whole of 2007. Figure 6-7 shows an example of an MRDY spam e-mail found in the database. This case also used spam watcher lists available at spammation.info, containing a list of companies whose stocks where touted during 2007. The e-mail campaign mentioned some information regarding the current stock prices and some projections of target prices in order to motivate investors to buy the touted stock.

**Figure 6-7 An example of MRDY Spam E-mail**

Figure 6-8 presents the example of an annotated press release issued on the 19th of July 2007 targeting the particular stock ‘Mobile Ready Entertainment Corp’ in the prosecuted case analyzed in this chapter.

**Figure 6-8 Example of MRDY Press Release**
Figure 6-9 shows some attributes of Form-144 which includes ticker, insider name (owner), the date of filing (fdate), the date SEC received the form (secdate), the expected date of sale (psaledate), insider signature date of the form (sigdate), the number of securities to be sold (pshares), the estimated market value of the proposed sales (value), and the name of the executing broker (broker).

<table>
<thead>
<tr>
<th>ticker</th>
<th>owner</th>
<th>fdate</th>
<th>secdate</th>
<th>psaledate</th>
<th>sigdate</th>
<th>pshares</th>
<th>value</th>
<th>broker</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRDY</td>
<td>MAGOLNICK</td>
<td>2007-04-11</td>
<td>2007-04-02</td>
<td>2007-04-04</td>
<td></td>
<td></td>
<td>2219830 44000.000</td>
<td>NEWBRIDGE SECURITIES CORP</td>
</tr>
<tr>
<td>MRDY</td>
<td>MORA CRAIG A</td>
<td>2007-04-24</td>
<td>2007-04-02</td>
<td>2007-04-03</td>
<td></td>
<td></td>
<td>2219830 44000.000</td>
<td>NEWBRIDGE SECURITIES CORP</td>
</tr>
<tr>
<td>MRDY</td>
<td>MORA CRAIG A</td>
<td>2007-07-02</td>
<td>2007-07-03</td>
<td>2007-07-25</td>
<td></td>
<td></td>
<td>968833 25000.000</td>
<td>NEWBRIDGE SECURITIES CORP</td>
</tr>
<tr>
<td>MRDY</td>
<td>MAGOLNICK</td>
<td>2007-09-17</td>
<td>2007-08-08</td>
<td>2007-08-15</td>
<td></td>
<td></td>
<td>2205830 50658.020</td>
<td>FIRST CLEARING LLC</td>
</tr>
<tr>
<td>MRDY</td>
<td>MORA CRAIG A</td>
<td>2007-09-10</td>
<td>2007-09-05</td>
<td>2007-09-15</td>
<td></td>
<td></td>
<td>2201830 12000.000</td>
<td>FIRST CLEARING LLC</td>
</tr>
</tbody>
</table>

Figure 6-9 IFDF (Form144) of the MRDY

There are intra-mapping and inter-mapping semantic relationships between the different concepts in the ontology. For example, as shown in figure 6-10, the Query and OLAP Reporting concept has relationships in the same layer (application layer), such as data sources, press releases and IFDF. Also, there are inter-mapping relationships with the Factiva database and insider filings.

Figure 6-10 Query and OLAP Relationship

6.3.2 Information Management

In the information management, structured data pre-processing and outlier detection tasks are performed inside the system. In particular, trading data is used to detect unusual trading volumes, jumps in prices and imbalances between bids and asks trading orders. Unstructured data is also processed internally, with the
option of requesting additional external services analysis. In this sense, the analyst has the choice of integrating information provided by external services agents or relying solely on the internal capabilities of the system. Unstructured sources, like spam e-mails and filing forms, could be used to create proactive indicators, alerts and attention-grabbing devices of potential microcap manipulation cases, especially if these events are accompanied with significant changes in trading behaviour.

Before any analysis take place, the data need to be pre-processed in order to combine and match different sources, either structured such as trading data, or unstructured such as spam e-mails, spam tracker feeds, press releases and financial filings. However, as the pre-processing and extraction of key terms is a difficult task, the analyst could use externally provided analysis to cross-check the internal results. For instance, spam e-mails are considered as a totally unstructured data source and the BI system has dedicated a group of tasks to improve their quality and address any inconsistencies, inaccuracies and omissions in the original data. Furthermore, the BI system considers the option of including input data from external unstructured data analysis providers: one dedicated exclusively to the analysis of spam e-mails and e-mail traffic, and another to the analysis of news and press releases coming from formal sources.

The BI financial market service system contains both retroactive and proactive monitoring engines that utilize different types of data source. Data could initially be explored by an information management component where different processes for data preparation take place. In particular, restructured data, along with updated patterns and models, could be used in this concept to perform a wide range of analysis, running automatically or on demand. In this instantiation the different tasks and analysis components were developed following the guidelines of the CRISP Data Mining reference model (Chapman, Clinton et al. 2000) using the IBM-SPSS PASW13 data- and text-mining workbench.

As shown in figure 6-11, the information management concept includes a range of data-preparation components including normal database manipulation operation components, such as query and OLAP reporting, outlier data-mining detection, association analysis and visualization components. Furthermore, some pre-processing tasks including developed text-mining analyzers extract information from spam e-mails such as ‘Date’, ‘Current Price’, ‘News’, ‘Price Increase’.
‘Sender’, ‘Recommendation’, ‘Target Price’ and ‘Buy Signal’. In particular, the objective of this instantiation is to demonstrate how existing detection systems using data mining can be enhanced. The analysis shows how an automated, linguistics-based text-mining approach could help to proactively raise alarms for possible ‘pump and dump’ manipulation with the use of unstructured data sources.

Considering the timeline of the case, it is possible to appreciate from figure 6-12 the outlier detection component based on the proposed financial indicators: jumps in prices, jumps in volume and bid and ask prices imbalances introduced by (Zaki, Theodoulidis et al. 2011) alarms raised in close relation with jumps in volume while the trading by the insiders was executed. For example, it is possible to distinguish alarms, represented as red squares in figure 6-12, on and after the 24th of July 2007. However, it is necessary to remark that the touting campaign started early in January 2007 with the publication of misleading press releases and arguably spam e-mails.
Although fraudsters reported their intention to sell, as stated in the SEC proceedings and corroborated by the Form-144 information, the form submission occurred after the actual selling of the shares, as shown in the figure 6-13. Moreover, it is possible to imply from an inspection of figure 6-13 that an association analysis performed using the Form-144 information, shows no strong relationships between dates, insiders and brokers. Thus, the filing form analysis on its own would not have been sufficient to detect the on-going fraud scheme.

There is a clear need for automated linguistic-based text mining to extract key concepts from these textual sources, and link them to structured data to provide
better analysis and an accurate securities fraud detection model. In particular, the SEC is concerned about the growth of these Internet fraud sources because they could represent successful attempts to manipulate stock prices, therefore, this BI financial market service process analyzed these information-based manipulation sources and demonstrated how automated linguistic-based text mining helps to raise alarms about touting stocks to alert regulators about potential manipulative practices.

Stock spam e-mails are a good example of unclean unstructured data. Thus, the proposed text-mining methodology employs two forms of the information extraction process: dictionary-based and pattern-based. The dictionary-based information extraction method requires configuring and training systems with linguistic resources like thesauri, lists of terms, concepts, synonyms, and types (semantic groupings of concepts). This method could use tokenization and morphological analysis (explained in section 5.2.1.2) to extract terms from spam e-mails and match them with the appropriate named entities. Tokenization analysis parses documents into characters and words that are called tokens, as.

The lexical ontology contains a library comprising an updated list (October 2009) of 23,028 OTC traded stock (symbol ticker, company name). Moreover, the Pink Sheets market supplied a list of symbol tickers and company names issued from 2006 to October 2009. Regarding the market data, the data preparation phase covered appending and sorting activities to construct the final dataset, i.e. data that was fed into the analysis from the initial raw data. The OTC and Pink Sheets data sources were appended to create a single issued stock, symbol ticker and company list to be used in the analysis phase. Restructured market data, along with converted e-mails documents, are stored in a knowledge base to feed the economic analysis with a clean dataset. For example, the symbol ticker contains from 4 to 5 characters. During the extraction process, the e-mails are scanned and analyzed in order to identify symbol ticker words within the dictionary and map the symbols to an event called <Stock Symbol>. However, spammers could use tactics to deceive the automatic filtration systems by changing the way of citing the symbol which obviously will not be identified by using the dictionary method.

Therefore, the pattern-based information extraction could help to address these challenges by employing natural language processing (NLP). This method extracts
information from e-mails using linguistic patterns or rules to recognize features of interest, especially unrecognized concepts from the dictionary-based model. This method performs a deeper analysis of the words, phrases and syntax inside the e-mails and thus helps to uncover knowledge of the underlying language and these e-mail messages.

These patterns are designed as a set of rules that describe structures in words, combinations of words, or combinations of rules. For example, if the task is to find or recognize stock symbol tickers that have been mentioned in the e-mail body or text, it is possible to define a rule in which all collections of 4-5 capital characters are mapped to an entity called <Stock Symbol>. An advantage of this approach is that it is possible to extract not only stocks tickers that were originally listed in the dictionary, but also to extract newly issued symbol tickers that match the pre-defined rules. Figure 6-14 shows how effective were the pattern-based method in extracting stock symbols from spam e-mails in 2007.

![Figure 6-14 Stock Symbol Information Extraction Comparison](image)

Furthermore, the pattern-based method incorporates a named entity recognition analysis that identifies and extracts certain non-linguistic terms and maps them to entities such as date, person, e-mail address (sender and receiver), and IP address, which will be used to extract the metadata associated with spam e-mails. The subject of the e-mail is another interesting element to analyze, as it could be one of the attraction features to tempt investors to open and read the e-mail. Overall, analysis of the metadata of spam e-mails is useful, although the messages themselves could contain misleading and incorrect information.

The pattern-based method also uses word association analysis to identify the relationship between two or more entities. For example, during the manual annotation it was noticed that the word ‘symbol’ is generally followed by a symbol ticker, such as Symbol: MRDY, as shown in figure 6-15. Additionally, different
forms of the word symbol were considered, such as ‘stock symbol’, ‘target symbol,’ “symb’, ‘sym’, ‘etssymbol’, ‘o.t.c. sym bol’, and ‘o.t.c symb0l’.

Figure 6-15 Stock Symbol Extraction

The analysis generates the required stock spam e-mail ontology to extract the key concepts and attributes required for fraud detection. The metadata analysis deals with key header information like date, time, sender, receiver, and subject. The body analysis deals with extracting named entities of stock spam e-mails for further analysis. For example, it extracts information on stock profiles such as the symbol ticker, the company holding the stock, and the sector to which the company belongs, as well as stock quote information that could tempt investors, such as target price or volume projections, trading date expectations, recommendations, and financial signals like buying or selling signals. The application also aims to extract the concepts and phrases that indicate whether cited news relates to the stock.

Previous work argued that target price quote and copy press release information advertised in spam e-mails could trigger a significant market response. Therefore, analyzers such as target stock price and volume, financial long-term or short-term investment indicators, speculative trade date, and buy or sell signals are developed.

The output of this process is a table containing symbol and metadata information (date, sender), and body content information such as current price, target prices and speculative potential profit, as shown in figure 6-16. The MRDY spam e-mails campaign was enacted on the 8th March, 9th Match, 10th March and 7th of April. The analysis shows that MRDY was heavily touted on the 9th (24% of the total spam messages) and 10th of March (72%). Furthermore, spammers
recommended and encouraged investors to buy and hold the stock and referred investors to a misleading press release that had been issued previously. For example, they used push phrases like ‘check or see the news’, ‘the hottest news released’, ‘see bullish news online’, and ‘fresh news’.

<table>
<thead>
<tr>
<th>Date</th>
<th>Sender</th>
<th>PatternCatalog</th>
<th>Current Price</th>
<th>Target Price</th>
<th>Percentage</th>
<th>Positive</th>
<th>BuySignal</th>
<th>Check News concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:burnias@optonline.org">burnias@optonline.org</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td>巡视<a href="mailto:hux0892@ymail.com">hux0892@ymail.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, released, very hot news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:tppnnvob@msn.com">tppnnvob@msn.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:jyewet@nau.com">jyewet@nau.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, check the news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:aphindhe@xhr.de">aphindhe@xhr.de</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:zemczy@compuzone.com">zemczy@compuzone.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, check the news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:david@mail.com">david@mail.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, fresh news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:glacerz@msn.com">glacerz@msn.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, the hottest news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td>wannabuyxasw.com</td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, check the news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:wenlili@msn.com">wenlili@msn.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:aphindhe@xhr.com">aphindhe@xhr.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, the hottest news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td>hoomphil@<a href="mailto:e@llton.com">e@llton.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, released, very hot news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:wenlili123@msn.com">wenlili123@msn.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:janet@msn.com">janet@msn.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:epohio@msn.com">epohio@msn.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:david@mail.com">david@mail.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, fresh news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:glacerz@msn.com">glacerz@msn.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, the hottest news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td>wannabuyxasw.com</td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, check the news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:wenlili@msn.com">wenlili@msn.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, news</td>
</tr>
<tr>
<td>2007-02-05</td>
<td><a href="mailto:aphindhe@xhr.com">aphindhe@xhr.com</a></td>
<td>mdry</td>
<td>USD0.04</td>
<td>USD62.2</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>strong buy/hold, the hottest news</td>
</tr>
</tbody>
</table>

Figure 6-16 MRDY Text Mining Analyzer Output

Regarding the option of cross-validation, the analysis included information provided by the spamation website, including web feeds of information about stocks advertised by unsolicited commercial e-mails. In 2007, MRDY was listed as a touted stock on their database, as shown in figure 6-17. However, this website does not provide detailed analysis on spam e-mail content or metadata. Figure 6-18 shows the semantic relationship between the analyzers and data source concept in the data source ontology and application layer.

![Spamation Track List](image)

**Figure 6-17 Spamation Track List**
The press releases and news issued in 2007 relating to the Mobile Ready company were retrieved. News items retained the original XML tags, and XML analyzers were used to query key concepts and metadata from the tags such as symbols, company name, date, subject, and other cited information in the body. Based on the news analysis, figure 6-19 shows press releases were strongly represented in July 2007, especially on 16th, 17th, 18th, 19th, and 25th. As in the SEC case, defendants tried to sell their shares, and therefore issued and posted press releases to mislead investors and pump the MRDY price up then dump it again after they had sold their shares. For example, as shown in figure 6-8, they released misleading information regarding the impact of the contract. Sentences like: ‘Simply Fit gears up to capture a share of $100 billion in annual beverage industry sales’ or ‘The project could yield in excess of one million ($1,000,000) dollars of revenue for MRDY over the next year’ could be found in the news text.
6.3.3 Economic Analysis

The economic analysis uses the output from the information management concept to create a complete representation of the fraud scheme. Figure 6-20 is a graphical representation of the events using the tools available in the pattern visualization analysis component. The visualization includes the price and volume time series, plotted together with the outliers’ alarms, cumulative news, and cumulative volume sales as declared on Forms-144. From left to right, it can be seen that alarms were always reactive to the changes in volume, and occurred several days after the misleading press releases were issued. The only proactive signals available throughout the duration of the manipulative scheme were the body content analysis of spam e-mails which was sent at the beginning of March, at least two months after the first misleading press releases. In fact, there is a slight increase in the stock price associated with the spam e-mail campaign which reached a high of $0.06 on the 12th and 13th of March 2007. However, this increase in price was not accompanied by a significant increase in the volume. According to financial theory, markets respond to new information and, as expected, there was some reaction or effect associated with the press releases. Therefore, the existence of press releases on the 16th, 17th, 18th, 19th, and 25th of July led to an increase in the volume of trading that was also captured by the outlier detection engine, raising corresponding alarms.
6.3.4 Output Management

Considering all the information available, it would have been challenging to detect this manipulative scheme without the spam analyzers and, potentially, the press releases analyzer components. The output management could have raised early alarms of suspicious behaviour based on the proposed BI analyses. This could help fraud analysts to open case management immediately rather than waiting more than a year before the actual SEC prosecution was completed.

6.4 Summary

This chapter proposes the BI financial market service system as another application of the financial ontology; it addresses requirements for proactively identifying possible fraudulent “pump and dump” scenarios through the analysis of textual information resources. The BI service system was evaluated through the investigation of a real case in the Pink Sheets market during the first half of 2007. This research demonstrates that the use of additional sources of information such as press releases provided evidence for disingenuous revenue projections, and indistinct business relationships. The case shows how both spam e-mails and false press releases are used as a fast, cheap and high impact means of dissemination of false information.
The proposed BI service system uses data mining analyzers to detect outliers and patterns of manipulation by utilizing a number of financial indicators as inputs, similar to (Zaki, Theodoulidis et al. 2011). More importantly, the proposed BI service system introduces the use of automated, linguistics-based text mining to extract key concepts of spam e-mails and press releases generated as part of 'touting campaigns'. In this context, it is possible to highlight that it would have been difficult to detect this manipulative scheme without these analyzers.

In fact, given the strategy followed by the insider traders in this case, it was not possible to raise early alarms of manipulation, and the fraud-related actions were detected only when the actual manipulative trading occurred. Unstructured sources, like spam-emails and forms, could be used to create proactive indicators and alerts of potential microcap manipulation cases, especially if these events are accompanied with changes in trading behaviour. Using on-demand analysis services from external providers could help investors and analysts to raise the alarm earlier.
7 Discussion

As stated in the literature review, stock market manipulation is an important issue for both the regulation of trading and for market efficiency. In particular, regulators face challenges in providing a fair, transparent, robust regulatory and monitoring framework to protect the market from fraudulent activities. Given that most exchanges have moved to electronic trading platforms to give investors the opportunity to trade across markets and jurisdictions, manipulative practices are found not only in single markets but also extend to cross exchanges and markets (Cumming and Johan 2008).

Manipulators can make profits through the use of various manipulative practices that deform prices at the expense of other investors and create information asymmetries (Cumming and Johan 2008). These manipulative practices take place in two main ways: trade-based manipulation and information-based manipulation (Aggarwal and Guojun 2003). Information-based manipulation refers to fraudsters using Internet bulletin boards, newsletters, spam e-mails and chat rooms to post messages advising investors to react to specific stocks based on “inside” information. In this way, manipulators use many fraudulent strategies to urge investors to buy stocks and make huge profits, but after “dumping” or selling their shares in the market, the price of the stock falls and investors lose their investment.

The Internet gives manipulators an easy channel through which to disseminate information cheaply, fast, and with high volume to reach a mass audience (U.S. Securities and Exchange Commission 2011). These messages might have a professional structure using an "infallible" combination of "inside" information and fabricated or exaggerated information (hard to identify as fiction) about the company’s sales, revenue projections, acquisitions, new products or services stock market data (Lease 2010).

In some cases, promoters and targeted companies collaborate to mislead either naïve investors or investors who want to take advantage and share profits with the manipulators. Manipulators reinvest the money in their own businesses or deposit their gains in personal accounts. Sometimes the money is used to liquidate the position of top shareholders or executives (Lease 2010).

The literature concludes that there are weaknesses in the structure of the unregulated market, such as OTC and Pink Sheets, because they are relatively
small, illiquid markets, necessitating lower disclosure requirements for listed firms, and subject to a less regulatory framework and fewer rules (Aggarwal and Wu 2006). Furthermore, the market suffers from information management problems which could result in conflicts of interest, lack of granularity, and lack of accuracy of the levels of information (European Commission 2010). The consequences of such limitations expose the market to many manipulation activities such as ‘pump and dump’ schemes arising from the interconnectedness of market participants and the limited transparency of relationships between different counterparties.

One of the reasons for establishing the Securities and Exchange Commission (SEC) was to protect and reduce stock market manipulation (Aggarwal and Wu 2006). The continuous improvement and development of financial market surveillance is essential to guarantee a stable and efficient market, by detecting fraudulent patterns and assuring transparent, accurate and high-quality information. The current financial market monitoring systems lack the financial ontology that could enable the efficient management of vast quantities of financial data, provide better communication and knowledge-sharing among analysts, provide a mechanism to resolve synchronization problems when multiple users access the data, and automate the publication of reports (Hilary, Yi-Chuan et al. 2009).

Generally, fraud cases are challenging and complex, and the amount of information involved is huge. Gathering facts and evidence is an especially complex process. Furthermore, financial fraudsters’ knowledge is rapidly evolving, making it difficult to keep any knowledge base of their activities up to date. There is therefore an urgent need to build a systematic financial ontology to organize knowledge about financial fraud, explain and share financial fraud logic operations, manage the enormous number of relevant facts gathered for case investigations, provide early detection techniques of fraudulent activities, suggest prevention practices, and allow reuse of this knowledge in different financial contexts (Kingston, Schafer et al. 2004).

For these reasons, one of the most important contributions of this research is the construction of ontology for financial fraud purposes, containing a complete set of financial concepts and definitions originating from multidisciplinary communities. As discussed in the literature, although there have been attempts to build financial fraud ontologies, such as the FF POIROT project (Kingston, Schafer et al. 2004),
these ontologies lack generality and do not demonstrate the conceptual system of the knowledge structure of domain experts; in particular, the common financial vocabulary used by the community and the essential financial business logic were not included in these applications. Furthermore, these ontologies took the approach of legal models, detailing the terminology of criminal fraud rather than demonstrating the logic behind financial fraud.

7.1 Financial ontology for fraud purposes

This research developed an ontology with a class hierarchy feature developed using a combination of top-down and bottom-up approaches. Each class includes general descriptions, with links to different sources such as linked data and instances as particular exemplars to fill the slot value of each class. Moreover, all sub-classes inherit the slots/properties of their super-class.

The challenge lies in finding ways to identify the type of relationship between concepts and to label those relations. Generally, most of the existing information management systems that use ontologies utilize a flat architecture for ontology management. However, flat architecture ontologies are managed independently, making integration difficult, especially when multiple ontologies are introduced. In order to overcome this problem, this research introduces the multi-layer architecture of financial ontology adapted from the work of (Mikroyannidis and Theodoulidis 2012).

The multi-layer architecture has many features, such as establishing the number of layers containing ontologies for different purposes, developed by different author groups; improving the manageability of technologies; and demonstrating the integration between different ontologies through both intra-layer and inter-layer mapping relationships.

In this thesis, the layers are presented as a pyramid (see figure 4.1), with a range of basic and generic financial concepts at the bottom and the more specific domain concepts at the top. Each layer is developed and maintained by a different group of authors, according to the expertise required. The lexical layer contains a domain-independent ontology of a purely lexicographical nature. The data source layer specializes in the organization of information from various data sources, either structured or unstructured. The domain ontology layer describes the general principles of the domain, integrating the different taxonomies developed by (Diaz
Finally, the application layer is used to automate the process of extracting financial concepts from unstructured sources and provides an appropriate knowledge base about financial market manipulation.

In the lexical layer, the ontology used the Investopedia website which contains a comprehensive online financial dictionary of 13,255 concepts, classified into 28 main groupings. The categories used in Investopedia have some limitations, as they do not consider financial market manipulation as a main concept that should be uniquely identified. For example, ‘Pump and Dump’ has been classified under ‘financial BuzzWords’. Furthermore, many terms are duplicated and classified under more than one concept; for example, ‘Securities and Exchange Commission- SEC’ is classified under three concepts: ‘Acronyms’, ‘Investor Relations’ and ‘Stocks’. Interestingly, the acronym is attached to its term, even though there is a concept ‘Acronym’ that should include all financial acronyms linked to their own main terms. For example, the acronym SRO is currently attached to the term Self-regulatory Organizations as (Self-regulatory Organizations-SRO). Finally, Investopedia is not up to date, omitting many manipulation terms such as ‘Front Running Research’, ‘Scalping’, ‘Corner’, ‘Squeeze’, ‘Stuff Quoting’ and ‘Flash Quoting’.

The data source ontology layer is based on the website of the US Securities and Exchange Commission (SEC) (www.sec.gov) which publishes litigation releases and prosecuted fraud cases. These cases were considered as a good corpus to extract financial fraud concepts and build a financial fraud dictionary. In the data source ontology layer, it is assumed that the website ontology was initially constructed by the webmaster, according to his perception of the thematic organization of his website. The website ontology then evolves over time, based on extracted navigation paths, in order to follow trends in its usage.

The SEC website provides access to vast repositories and large volumes of unstructured and semi-structured texts from various sources. This makes the analysis of these sources interesting because there are clear opportunities for rich financial fraud contexts to emerge, impacting on how fraudsters practise the fraud and manipulate the market. Indeed, manual techniques are challenging and increasingly impracticable in satisfying the required analysis tasks.
7.1.1 Domain Ontology Methodology analysis

In the domain ontology layer, the researcher used a focus group of financial domain experts to analyze SEC litigation releases. SEC litigation releases are considered a suitable and coherent corpus relevant to financial fraud, from which additions and alterations to the domain ontology can be proposed. The financial domain experts analyzed four SEC litigation releases representing four different manipulation types: ‘insider trading’, ‘pump and dump’, ‘front running’, and ‘marking the close’. Three of the cases are from the OTC market and only the front running case is from a regulated market because we could not find a matching case in the OTC market.

The focus group was seen as an appropriate and valuable means of building the domain ontology because, unlike individual interviews, it allows participants’ perspectives and their questions and arguments to be revealed through discussion. This focus group were held in two rounds; in the first round two participants analyze the pump and dump case as a pilot and to evaluate the process; the second round involved eight participants, working in pairs. The focus group aimed to identify interesting financial concepts embedded in the SEC litigation releases, to train the domain ontology with relevant financial fraud concepts that will enable it to answer questions similar to those that might be asked by ontology users reading those releases.

The focus group allowed participants’ perspectives to be revealed through discussion and build financial dictionary for fraud purposes. In particular, the financial experts provided depth to the organization of the financial concepts within the ontology, helping to build a dictionary of financial concepts for fraud purposes and mapping these concepts to the relevant ontology classes. They were also allowed to add new classes to the ontology and suggest better names for existing classes. The output of the focus group was to produce a good version of the domain ontology based on the analyzed SEC litigation releases. Because it is domain specific, the research did not follow the standard ontology engineering approach.

In terms of analysing these focus group and building the ontology classes, the research used the constant comparative method (Glaser and Strauss 1967) as the methodology. This method was originally developed for use in grounded theory, but is now widely used to analyze qualitative data gathered from different media.
such as interviews, focus groups, examination of the documents, and experiments. The methodology suggests four distinct stages: ‘Unitizing’, ‘Categorizing’, ‘Filling in patterns’, and ‘Member checking’.

At the unitizing stage, participants annotated and highlighted interesting financial documents in the litigation release. In the first round they did this in a single session, but the annotation process was time consuming and varied from one participant to the other, leading to incomplete annotation.

In the second round, therefore, the participants annotated the litigation releases in their own time, and a second meeting was held in which each pair of participants their case and the annotated concepts, to come to an agreement about which concepts should be recorded in NVIVO and later mapped to the ontology classes.

From the four cases, 441 concepts were annotated. The participants had a different annotation focus in each case. For example, in the pump and dump case, the participants considered the financial laws and regulations referred to by the courts; this should help users searching for legal infringements. These participants also highlighted the description of the manipulators either as individuals or by organization. However, in the insider trading case, participants were more concerned to identify manipulators’ background and prosecution history, as well as other entities involved, such as intermediaries who facilitated the process or victims who suffered from the scheme. Timeline concepts proved to be important in describing the manipulation activities before, during and after the fraud. In addition to the legal evidence and patterns captured by participants, structured evidence was dominant in the marking the close case, with participants annotating how manipulators artificially inflated stock prices to gain profit, before they declined again. Finally, evidence related to manipulation of the trading system was considered by participants annotating the front running scheme.

At the categorizing stage, the participants mapped the annotated concepts to suitable classes of the ontology, adding new classes if necessary or suggesting better naming of existing classes. In each case, the participants had a different focus of interest. For example, ‘Time (When)’, ‘Effects (Consequences)’, and ‘Actions (What)’ were the classes with the highest percentage of annotated concepts in both the insider trading case and the front running case. Sub-classes ‘before’, ‘during’ and ‘after’ were added under ‘Events time line events’ because the participants wanted to describe the timeline events of the different
manipulation activities, including the different evidence and patterns used by the manipulators to violate the market.

Furthermore, ‘Manipulation Participants’ was suggested to describe not only the agent profile but also victims and intermediaries who might be involved indirectly in the manipulation. A new sub-class called ‘Material’ was added to differentiate between public materials and non-public materials used in insider trading. In the marking the close case, the participants associated the highest percentage of annotated concepts with Actions (What) and Effects (Consequences) to describe the different manipulation activities, especially the structured patterns. In the pump and dump case, Actions (What)’ class had the highest percentage; the participants also agreed that in this case evidence could be classified as ‘information-based’, ‘trade-based’, or ‘action-based’. New concepts were added to explain the different information-based manipulative events, such as ‘Opinion Mill’, and ‘Legal Opinion Letters’. All participants agreed that the ontology needed a ‘Laws and Regulations’ concept to describe the different laws and acts violated by manipulators.

At the ‘Filling in patterns’ stage, categories were fleshed out after the extensive data collection and unitizing. All annotated concepts were classified in relevant categories and these categories became saturated. The participants were clearly concerned to describe the agent’s ‘action’, and this represents the highest percentage of categorization in the ontology, at 27.78%. Generally, these actions are associated with timeline events to describe the temporal dimension of the manipulation activity, so the ‘Time (When)’ class has the second highest percentage of concepts (23.02%) annotated by the participants. In order to describe the consequences and effects of these actions on the efficiency of the market, the participants classified 20.41% of patterns and evidence that describe the effect of agents ‘behaviour, trading orders and execution of actions on the market integrity. The ‘Agent (Who)’ class represents the fraudster who performs a manipulation. The participants classified 8.60% of the manipulators with a brief profile of their status, address, age and job description. Sometimes, companies help the agent to execute his manipulation.

Furthermore, agents normally target any type of assets such as stocks, penny stocks, money markets, bonds, derivatives, and structured notes; the ‘Asset (Target)’ class has 2.41% of concepts, describing whether the asset is ‘real’ (commodities) or financial (equities, derivatives and debts). The four analyzed
cases were selected to reflect the four sub-classes of the ‘Market manipulation (types)’: ‘abuse of market power’, ‘contract-based’, ‘breach of fiduciary duty’ and ‘runs and raids’; the participants classified 6.03% of the concepts that describe these types. Overall, the ‘Financial Market Stakeholders’ class were allocated 6.61% of concepts to describe the different stakeholders involved in the cases, such as investors, intermediaries (e.g. brokers), regulators (e.g. fraud detectors, enforcement agencies and investigators). The ‘Venues (Where)’ concept had 1.82% of concepts, describing where investors and manipulators organize their trade manipulations. Finally, the ‘Laws and regulations’ concept holds 2.67% of acts and laws involved in the cases.

At the ‘Member checking’ stage, the ontology was checked to confirm that it is a reasonable representation of reality. Based on the analysis of the focus group, the domain ontology was built and developed using the PoolParty Thesaurus Management System. The current version of the domain ontology has nine concepts: actions, assets, effects, financial market stakeholders, laws and regulations, manipulation participants, market manipulation types, time, and venue. Each concept holds sub-concepts in a hierarchy. A semantic association is used to express any kind of relationship between concepts.

7.1.2 Ontology evaluation

The proposed ontology was evaluated through specific instantiations which demonstrated through three applications its applicability in improving the functionality of market monitoring surveillance systems in fraud detection. The three applications are considered as intelligent systems that extract information from unstructured sources based on the financial ontology. The financial ontology provides the underlying framework for the extraction process and the capture of information related to financial fraud. Moreover, the ontology displays the extracted information in an organized and coherent way, which can be used as a knowledge base.

The instantiations demonstrated how the ontology might help in proactive fraud monitoring or in open investigation by acting as a knowledge management repository system to control the wealth of data gathered during financial fraud cases, and to share existing manipulation patterns of prosecuted cases among investigators and interested users. On the empirical side, the thesis presents
examples of text-mining applications with fully working prototypes which could train the ontology in the latest manipulation techniques. Another very important contribution of the ontology is that it addresses the need to reuse the text-mining process in other parts of the domain, and to integrate the extracted information with other systems within the organization.

Table 7-1 presents a summary of the applications and their main characteristics:
## Table 7-1 Summary of case studies

<table>
<thead>
<tr>
<th>Name</th>
<th>Brief Description</th>
<th>Ontology Layers</th>
<th>Techniques and Analyzers</th>
</tr>
</thead>
</table>
| **Case Study One:** SEC litigation releases Text Mining Application | The case study used the SEC litigation release to provide an empirical evidence of how text mining could be integrated with the financial fraud ontology to improve the efficiency of extracting financial concepts and demonstrating the published prosecuted cases in appropriate knowledge base. | **Lexical:** Investopedia | **Text mining Analyzers:**  
- Manipulation participants Analyzer  
- Investigation, Enforcement and Regulatory Analyzers  
- Timeline Manipulation Events and Actions Analyzer |
| | | **Data Source:** SEC litigation releases |  |
| | | **Domain:** Financial Fraud |  |
| | | **Application:** SEC litigation releases Text Mining Application |  |
| **Case Study Two:** “Stock-touting” Spam E-mails Text Mining Application | The case study demonstrates text mining application that could automatically and efficiently extracting key attributes and characteristics of different unstructured sources generated as part of "touting campaigns". | **Lexical:** Investopedia Libraries such as OTC securities list, Finance, Opinion | **Text mining Analyzers:**  
- Metadata Analysis:  
  - Header Analyzer  
  - Subject Analyzer  
- Body Analysis:  
  - Symbol Ticker Analyzer  
  - Company Name Analyzer  
  - Market Analyzer  
  - Stock Price Quotation Analyzer |
| | | **Data Source:** Spam E-mails |  |
| | | **Domain:** Financial Fraud |  |
| | | **Application:** Spam E-mails Text Mining Application |  |
| **Case Study Three:** BI Financial Market Service system for a financial market monitoring and Surveillance system | The case study describes the business intelligence (BI) service system for a financial market monitoring and surveillance system in which different components interact in a coordinated way with internal and external service providers to produce proactive alarms for possible securities fraud cases. The proposed BI service system is demonstrated through an exemplar case study of text mining and data mining to analyze the impact of 'stock-touting' spam e-mails and misleading press releases on trading data. | **Lexical:** Investopedia and Libraries | **Text mining Analyzers:**  
- BuySignal analyzer  
- Current price Analyzer  
- Date Analyzer  
- News analyzer  
- Price Increase Percentage  
- Sender Analyzer  
- Stock Symbol analyzer  
- Target price analyzer |
| | | **Data Source:** Unstructured: Spam E-mails Press releases, IFDC. Structured: MRDY quotes closing prices |  |
| | | **Domain:** Financial Fraud |  |
| | | **Application:** BI Service system |  |
| | | **Data Mining Techniques:**  
- Association Analysis  
- OLAP  
- C5.0 Decision Trees Analyzer |  |
1. Case Study 1

In Case Study 1, the challenge of building a text-mining application to automate the process of extracting financial concepts from SEC litigation releases was addressed. In particular, as described throughout, the emphasis of the text-mining task was on developing several analyzers using advanced linguistic patterns (NLP) in order to provide empirical evidence of how text mining could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of extracting financial concepts and demonstrating the published prosecuted cases in an appropriate knowledge base.

The application demonstrates the applicability of extracting information from the litigation releases provided in the SEC website. Key concepts such as 'litigation release number', 'release publication dates', 'actions', 'agents' (the defendants' individual or organization names, 'document format type', 'document link', 'civil case no.' (the allocation number issued by the court), 'district court no.' (including state courts or federal courts), and 'plaintiff' were extracted from the enforcement section of the SEC website.

Furthermore, the application extracted concepts from the SEC complaint document produced by the US district court that describe cases as violations of securities laws. This application contains analyzers for market participants, investigation, enforcement and regulatory bodies, timeline manipulation events and actions. These analyzers aim to extract the annotated financial concepts recommended by the domain experts and provide answers from the ontology to the key questions enumerated in section 4.3.4.

The applicability of the system is to allow users to query specific financial terms, learn their meaning and explore other terms related to the specific term. Several analyzers were developed to provide answers to the key ontology questions from litigation releases document: Who is (are) the agent(s) involved in the manipulation? Which asset is being targeted? In which venue is the manipulation taking place? What action has been performed or is planned? Is it a trade-based or an information-based action? Which pattern is associated with this manipulation? When was this manipulative action performed? Where is the manipulator getting his profit from?
2. Case Study 2

This second instantiation of the financial fraud ontology automates the categorization and classification of stock spam e-mails, a data source chosen for further evaluation of the proposed financial ontology. The case study demonstrates empirical evidence of how text mining could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of extracting concepts from e-mail messages.

It contributes to identifying gaps in financial research studies, as none of the previous studies has considered the use of text-mining techniques for the analysis of "stock touting spam e-mails" in an unregulated market. Furthermore, the traditional techniques are especially time and human resource intensive and error-prone, and ineffective in dealing with the massive volume of spam e-mails. Thus, the instantiation demonstrates a novel application of text-mining analyzers to automatically and efficiently extract relevant, high-quality information including stock profiles, financial key words, stock and company news (positive/negative), and compound phrases from stock spam e-mails generated as part of "touting campaigns".

The proposed financial fraud ontology played a vital role in providing the framework for the extraction process. It displayed the extracted information in an organized and coherent way, as a knowledge base for fraud detection and monitoring the behaviour of stocks. In particular, the text-mining analyzers could raise proactive alarms; traditional data-mining analyzers do not predict correctly, or generate only weak signals of suspicious trading behaviour. Furthermore, the ontology could help regulators and market participants to minimize the on-going "pump and dump" schemes that rely on the circulation of rumours or illegal information.

The text-mining application extracts concepts from the e-mail headers, providing information about the sender and recipients of the e-mail and when it was sent. This can help regulators to know who started the touting campaign and when, how long the spammers continued their campaigns for specific stocks and who was targeted. Regulators could also warn targeted investors. Body analysis of the e-mails extracts key information such as ‘Stock symbol’, ‘Company name’, ‘Market’, and ‘Stock price quotes’ with its different financial indicators, and price speculation. Thus, the spam e-mail body analyzers will help regulators to compare
this information with the actual trading behaviour of the touted stock and identify insider trading schemes. Proactive alarms can be raised before the actual execution of the fraud.

3. Case Study 3

Case study 3, another instantiation of the proposed financial fraud ontology, addresses requirements for proactively identifying fraudulent “pump and dump” scenarios through the analysis of textual information resources. In particular, it describes the business intelligence (BI) service system for a financial market monitoring and surveillance system in which different components interact in a coordinated way with internal and external service providers to produce proactive alarms for possible securities fraud cases.

The proposed BI service system identifies the key concepts of spam e-mails and press releases to be extracted, relates them to other available information and highlights the likelihood of their being part of ‘stock touting campaigns’.

Furthermore, it incorporates data-mining analyzers to detect outliers and patterns of manipulations utilizing a number of financial indicators as inputs and carried out to a real case from the OTC market.

The applicability of this system is to raise proactive alarms when manipulators start their touting campaign dissemination. Analysts could use such application to add these touting stocks in watch list and monitor them in case any violation could happen. Also, market could send some notifications to their investors alerting them from acting upon such fallible information appeared in these unstructured sources.
8 Conclusions

This chapter summarizes this thesis. It begins by restating the motivation, aims and objectives, the structure and contributions of the work, and discusses its limitations. Policy implications are also discussed, before the chapter ends with suggestions for a number of future research areas based on the work presented in the thesis.

8.1 Overview of the work

The aim of the study was the development of a financial ontology for fraud purposes, to improve the functionality of marketing monitoring surveillance systems. The proposed ontology demonstrates knowledge of the processes of financial fraud, understanding and sharing financial fraud logic operations, managing the massive amount of relevant facts gathered for case investigations, providing early detection techniques of fraudulent activities, offering prevention practices, and allowing reuse of these knowledge resources in different financial contexts.

In order to achieve this aim, the researcher examined recent work on ontologies, as well as the new requirements for the operation of financial markets, especially the recent consultations on the structure and governance of unregulated markets such as OTC and Pink Sheets. The usefulness of the proposed financial ontology was evaluated through three critical case studies, which not only demonstrated through practical examples the weakness of the market and different manipulation scenarios, but also explained how the proposed solutions could improve financial market surveillance systems.

The thesis organized as follows: chapter 1 discussed the important role financial markets play in modern society and considered the different structures of financial markets. It addressed the different challenges to providing a fair, transparent, robust regulatory and monitoring framework to protect the market from fraudulent activities. In particular, the vulnerability of unregulated markets was discussed, highlighting the fact that unregulated markets suffer from information management problems which lead to various manipulative practices such as 'pump and dump' schemes arising from the interconnectedness of market participants and the limited transparency between different counterparty relationships. The chapter also discussed why unregulated markets need to be monitored, not only to meet
the objective of reducing market manipulations and fraud but also to level the playing field for all participants. Finally, it described the research proposal to address these problems.

Helping to create strong foundations for the financial fraud ontology, chapter 2 presented a review of financial fraud, ontologies, and fraud detection tools, which related to the knowledge-based aspects of the methodology. It began by presenting a brief overview of the unregulated market structure and its market tiers. This was followed by an explanation of the different market manipulation types and how they have been defined historically and investigated by the financial and economic community. It highlighted that manipulations do indeed exist, and that they can be very profitable for the manipulators but detrimental to other market participants. It also provided a review of the existing market monitoring and surveillance systems, components and processes, and the allocation of responsibility for monitoring and surveillance. The chapter discussed the different technologies that have been used to perform the tasks of fraud detection. Finally, it reviewed the existing financial fraud ontologies with an overview of their construction, limitations and the gaps which motivated this research.

Chapter 3 provided a complete review of another important part of the knowledge base, describing and justifying in detail the research methodology and its philosophical perspective. More specifically, chapter 3 discussed the method used to research and design an innovative solution that addresses the gaps identified in the understanding and functionality of market monitoring and surveillance systems. The chapter started by providing a general definition of the research approach followed by a justification of why this methodology was appropriate for the task at hand. It continued with an in-depth description of how the methodology was implemented.

In chapter 4, financial fraud ontology was introduced, specifically, the ontology architecture with different layers. The chapter discussed the methodology for constructing the proposed financial ontology, with a detailed analysis of the practical considerations. The chapter demonstrates how the ontology was built using the PoolParty environment, with detailed discussion of the different ontology mappings.

In terms of evaluating the utility of the ontology, section 5.4.1 and chapters 6 and 7 presented individual instantiations of the artefact by means of case studies.
Evaluation was carried out on a case-by-case basis, judging the utility of the proposed ontology in solving and increasing understanding of a part or the whole of the problem and the use of the models introduced previously. Each case study was presented with its own results and conclusions, re-introduced in chapter 8 on overall conclusions and in the sub-section on lessons learnt, following the circumscription and operation and goal-knowledge loops of the research methodology.

8.2 Contributions

This thesis delivers several contributions to knowledge. In relation to research question 1, it examines the context in which market monitoring and market surveillance systems are currently being used. The unregulated market structure and its different tiers were described, and the gaps in expected functionalities of such systems were addressed. The different manipulation types associated with unregulated markets were examined. The study identified that market monitoring surveillance systems lack a financial ontology that could effectively solve the information management problem in unregulated markets. Such ontology could manage vast amounts of financial data, provide better communication and knowledge sharing among analysts, provide a mechanism to resolve synchronization problems when multiple users access the data, and automate the publication of reports.

In relation to research question 2, the thesis proposed a financial ontology for fraud purposes that contains a large set of financial concepts and definitions originating from multidisciplinary communities. This ontology aims to help market monitoring surveillance systems to update the knowledge of the processes of financial fraud, understand and share financial fraud logic operations, manage massive amounts of relevant facts gathered for case investigations, incorporate early detection techniques of fraudulent activities and prevention practices, and allow reuse of these knowledge resources in different financial contexts.

In relation to research question 3, the study introduced multi-layer architecture of the financial ontology adapted from Mikroyannidis et al. (2012). The multi-layer architecture (lexical, domain, data source and application layers) has many features, such as establishing the number of layers containing ontologies for different purposes developed by different author groups, improving the
manageability of technologies, and demonstrating the integration of different ontologies presented through intra-layer and inter-layer ontology mapping relationships. In this thesis, the layers are presented in a pyramid diagram (figure 4.1) with a range of basic and generic financial concepts in the lowest layer, and more specific domain concepts at the top. Each layer is developed and maintained by a different group of authors, according to the expertise required.

In order to build the ontology, this work relied on a suitable and coherent corpus relevant to financial fraud, of litigation releases provided on the SEC website. The ontology has a class hierarchy feature developed using a combination of top-down and bottom-up approaches. The top-down approach used a focus group of financial domain experts to analyze four SEC litigation release cases. In the bottom-up approach, the research used the constant comparative method (Glaser and Strauss 1967) to analyze the data collected from the focus group and build the domain ontology.

In relation to research question 4, this thesis has demonstrated how the financial ontology provides the framework for the extraction process and the capture of information related to financial fraud. Manual techniques being increasingly impracticable for the task, text-mining technology is used to automate the process of extracting financial concepts from texts, providing a knowledge base about financial market manipulation.

In particular, the application layer used the SEC litigation releases to provide empirical evidence of how text mining could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of extracting financial concepts and demonstrating the published prosecuted cases in the knowledge base. The text mining analyzers were developed to extract the financial concepts recommended by the domain experts and provide appropriate knowledge about the players in the fraud.

In relation to research question 5, the thesis presents examples of novel applications of text-mining analyzers and data-processing components, developing several offline surveillance systems that are fully working prototypes which could train the ontology with the most recent manipulation techniques. Another significant contribution of the ontology addresses the need to reuse the text-mining process in other parts of the domain, and to integrate the extracted information with other systems within the organization. Each case study presented its own
text-mining analyzers, analyses, results and outputs, and followed the design research methodology to create independent artefacts that solve the particular problems introduced in each of them.

In relation to research question 6, this thesis presents another instantiation of the proposed financial fraud ontology that addresses requirements for proactively identifying possible fraudulent “pump and dump” scenarios in unregulated markets through the analysis of textual information resources. The case study describes the business intelligence (BI) service system for a financial market monitoring and surveillance system in which different components interact in a coordinated way with internal and external service providers to produce proactive alarms for possible securities fraud cases.

8.3 Limitations of the work

Despite the extensive work reported in this thesis, there are still challenges that need to be addressed through further investigation. With respect to the construction of the ontology, it is still possible to enhance and expand it. There is a limitation being the formal description of its structures and properties. Future work will have to analyze more cases to cover a complete range of market manipulation types. New manipulation schemes such as high frequency trading and stuff quoting need to be analyzed and considered in the ontology. Furthermore, the ontology should be evaluated by domain experts with different backgrounds, such as lawyers, financiers, investigators and analysts, in order to cover the different perspectives and views of the ontology.

Financial knowledge and concepts are rapidly evolving, which makes it difficult to keep any knowledge base up to date. This ontology lacks a means of modelling the ontology evolution process in a valid transaction time that recognizes all data referring to the temporal aspect. This could help in capturing changes in the layers and presents the different usage of the ontology and captures the changes in data. Regarding the coverage of the cases and datasets, focusing on only a few case studies can potentially skew the outcome. However, if the text-mining process is developed together with the ontology, additional training might be required, in terms of additional annotation inputs from the domain experts. The text-mining analyzers have achieved very good levels of accuracy given the level of training effort involved. The intention of the research was to prove the feasibility of
developing a text-mining model based on ontology, rather than focusing on improving the accuracy of the model. Having said that, it will be possible to optimize the accuracy over time, through an iterative process; thus trains the model to cover the gaps.

In terms of sources, the case studies covered limited number of unstructured sources. Future work will consider other types such as social media and other related financial documents could enrich the ontology knowledge. There may have been changes in the way the financial markets and fraudsters work that have not been taken into consideration in the selected cases. However, although the knowledge gained from the analysis of the cases is valuable, this thesis pursues as a higher goal the creation of a common body of knowledge, and this can still be achieved even from particular nuggets of knowledge.

8.4 Future Work

Further research should address further evaluation of the proposed financial ontology for fraud purposes, especially with respect to the different users; it should consider aspects such as completeness, usability, and usefulness. As mentioned previously, future work should analyze and include in the ontology a more comprehensive range of market manipulation cases, including new schemes such as high-frequency trading and stuff quoting. Furthermore, the ontology should be evaluated by different domain experts with different backgrounds, as outlined above.

A section of the ontology could be dedicated to deal exclusively with cases from unregulated markets, making way for further research addressing manipulation cases from the regulated markets. A benchmark analysis could be provided to compare the patterns between different markets. Real-time fraud analysis is further instantiation that would enhance the detection capabilities and automation of the ontology.

From the technological side, text-mining techniques and data mining are used to demonstrate how BI could be integrated with the ontology in order to investigate cases. Thus, in considering new technologies such as cloud-computing and big data software deployment, it would be valuable to complement the ontology with full deployment for use as crowd monitoring or fraud knowledge management portal. The portal would act as a repository for the different stakeholders using the
financial market. Using cloud capabilities, crowd monitoring could incorporate different fraud analysis to facilitate the process of detecting manipulative activities in the market. Knowledge management system would include all the patterns of manipulation considered in a temporal dimension, manipulators and their networks, manipulated assets, venues that suffer most from such manipulations, different manipulation patterns associated with the cases, the benefits, and the laws and regulations referred to by the courts.

Overall, this research has delivered a novel artefact with working tools for the analysis and development of a financial fraud ontology that could improve market monitoring and surveillance systems. The ontology presents an organized body of knowledge for proactive fraud monitoring and open investigation by acting as a knowledge management repository system to manage and control the huge quantities of data that can be gathered during financial fraud, and to share existing manipulation patterns of prosecuted cases among investigators and other users.

The study has demonstrated an intelligent system using data and text-mining tools to tackle information and market-based manipulations. Such intelligent systems could be used by regulating authorities and in self or non-self-regulated venues, as well as by other types of customer, such as individual investors, institutional investors and broker-dealer firms. They could also enable new configurations and business models for regulators. This ontology needs to be able to reuse the text-mining process and integrate the extracted information with other systems.
9 References


The Institutions, Economics, and Econometrics of Securities Trading, Oxford University Press.


Wharton Research Data Services (2009).


