Competitive Analysis of Business Filings Using Ontologies and Linguistic Analytics

Completed Research-Paper

Seán O'Riain
Digital Enterprise Research Institute,
NUI Galway,
IDA Business Park, Lower Dangan,
Galway, Ireland
sean.oriain@deri.org

Edward Curry
Digital Enterprise Research Institute,
NUI Galway,
IDA Business Park, Lower Dangan,
Galway, Ireland
ed.curry@deri.org

Robert Pinsker
College of Business,
Florida Atlantic University
777 Glades Road
Boca Raton, FL, 33431, US
rpinsker@fau.edu

Abstract

Competitive business analysis is a manually intensive process that depends on analyst heuristics to gather and interpret relevant information from sources such as the SEC filings. The continuous growth in the volume of reports, and difficulty experienced with formats such as the eXtensible Business Reporting Language (XBRL) when sourcing information, challenge the viability of the analyst to conduct manual analysis. In particular analysis of the free text discussion sections of financial reports for qualitative data requires considerable effort, making comparison of qualitative data alternatives difficult. There is a clear need for an automatic linguistic analysis capability that can add meaning, structure and provide relevant contextual information to reduce the effort needed to analyse free text discussion sections.

This paper introduces a qualitative interactive Decision Support System (QDSS), driven by an Ontology Based Linguistic Analysis Pipeline that leverages domain knowledge to drive linguistic analysis and generate structured, semantically marked-up data. The approach is evaluated within a competitive intelligence scenario where an analysis using a QDSS based on the pipeline output is compared against a manually conducted analysis. Experiment results report a 37% performance improvement in finding relevant information and usability results, on the pipelines contribution to competitive analysis task structuring and information provision.

Keywords: Competitive analysis, linguistic analysis, ontology development, ontology based information extraction, information extraction, business intelligence, semantic technologies

Introduction

Competitive analysis is used as an investigative tool by business analysts to deliver insight into several critical business processes. Those processes include: a firms or competitor's operations and strategy; understand market movement; identify competitors; determine strengths and weaknesses; and predict the next strategic and tactical moves (Shaker and Chaples., 1993; Sheng et al., 2005). Competitive analysis monitors competition or environmental factors, captures essential measures of activity, and organizes the measures to help decision makers detect and respond to changes (Sauter et al., 2005). It involves an analyst performing the following processes (Debreceny and Gray 2001):

- Sourcing, identifying and collecting relevant information, e.g. financial instruments, ratios and discussion eluding to company performance
- Interpreting the information to develop an understanding as to what may be occurring
- Generating insight to support (executive) decision making

In activity terms, competitive analysis comprises the main tasks of:

- Manually locating and correlating key information
- Analysis of the correlated data.

Evaluating quantitative financial data involves utilising widely accepted financial metrics and ratios. Once this information is gathered comparisons are straightforward. In instances where some information is messing, substitution with alternatives is possible (Debreceny et al., 2011). Evaluating the qualitative data (i.e. free text discussions) is more problematic, lacking any such common measures making normative comparisons of alternatives difficult (Sauter et al., 2005). Qualitative Decision Support Systems (QDSS) track and organize qualitative information providing systemic support for its use within decision making.

This paper addresses the lack of a common measure for relevant criteria to support qualitative aspects of competitive analysis as part of a wider QDSS. Using design-orientated information systems research process guidelines and principles from (Osterle 2011), the research methodology adhered to the process phases:

- Analysis. Problem analysis is presented as challenges to the business to perform a competitive analysis in term of information requirement and to the analyst in terms of information specification. Analyst requirements outline both research objectives and functional requirements of the developed solution.
- Design. The solution approach introduces the Ontology Based Linguistic Analysis Pipeline and its major components. The pipeline uses domain linguistic modelling and linguistic analysis to extract common measure features from the text. A qualitative interactive DSS (instantiated with pipeline output), allows the qualitative analysis of text documents (i.e. Form 10-Q filings) to support competitive analysis.
- Evaluation. Pipeline evaluation reports on the contribution to competitive analysis using domain specific performance measures and usability orientated methods. Finally research conclusions and future work are discussed.

Central to our approach for developing a common measure for qualitative analysis is the use of domain knowledge to develop a task specific ontology, modelled as a series of information trails or semantic paths. The paths in turn reflect how an analyst contextually associates information when going through the reports discussion sections. To our knowledge this approach for the financial sub-domain of competitive analysis has not been previously attempted.

Competitive Analysis Challenges

For an analyst, the competitive analysis process presents challenges relating to information interaction. The first challenge is the identification of what information to include, and how it should be prepared for ease of exploitation. SEC reports such as the quarterly filing (Form 10-Q), comprising financial accounts and statements from the Chief Executive Officer (CEO), are a major source of competitive analysis information used by the analyst. In particular, the free text management statements that comment on corporate performance and intangibles such as people, brands and patents are actively searched for key information and interpretation (Pfeifer, 2007). Manually locating and correlating key information from within financial statements is recognised as presenting particular difficulty due to their textual nature, lack of structure and lack of common format (Grant 2006). The filings' size and sheer volume, ensures that up to 75% of analyst resource availability is expended in information gathering to support analysis (Zahra 1993). Even for professional analysts continuously conducting appraisals will be influenced by the constraints of personal subjective views and fatigue (Li-Yen 2009).

Once information has been collected, correlated, and structured, the second challenge becomes the application of appropriate resources for interpretation. However, analysts are highly skilled, are always in demand and are treated as premium resources. Assisting the analyst to identify the 'information nuggets' could be supported with the automatic detection of relevant information, together with an ability to traverse the 'information space' in a structured fashion that helps with its identification (Zhang 2004). The resource intensive nature of the analysis process brings the third challenge into focus. Namely, how can an analyst generate an analysis in a time frame suitable for decision making? Competitive analysis would benefit from natural language processing and text mining to help make sense of the text to support financial decision making (Zhang 2004). Currently there is a lack of support tools that assist an analyst use and leverage of information in this regard.

Analyst Requirements

Sauter et al. (2005) note the principles of a qualitative competitive intelligence DSS include:

- The system should provide data that reflects the perceptions of a broad range of individuals and
- The system should provide a mechanism for prioritizing, reporting, and analyzing information in a manner that facilitates evaluation and judgment application (Massetti, 1996).

The principles reinforce the centrality of information and its primary concern for the decision maker. These systems provide a logically-organized vehicle through which the analyst can move from information to knowledge. Competitive analysis requirements include information acquisition, interpretation and analysis generation.

Information acquisition from sources requires that the more widely encountered data formats of the eXtensible Business Reporting Language (XBRL), Hypertext Mark-up Language (HTML), Common Separated Values (CSV) and plain text be catered for. For a more thorough discussion relating to issues with financial data integration, such as data source inter-dependencies, data mismatch and object versus schema level fact expression, we refer the reader to Curry et al. (2009) and O'Riain et al. (2011). Similar to those studies, we incorporate Form 10-Q data available in XBRL / XML serialized format as our main data source.

Information interpretation introduces the interlinked requirements of domain knowledge usage and linguistic analysis. The complexity of competitive analysis ensures any automated processing requires hand crafted approaches highly reliant on domain expertise. The capturing of domain knowledge and know-how from a range of stakeholders is required to instruct linguistic analysis on how best to deal with financial term ambiguity, language variation, alternate representations, and results organization. Linguistic analysis can then be applied to textual, structure-less, and format-less financial statements with resulting output made directly available to the analyst (Grant et al., 2006).

Analysis generation requires that an analyst be supported with interactive tools at multiple levels. The first is the document or filings level to facilitate targeted information discovery and search through the document. The second is the interactive query level, allowing the analyst conduct complex companies/sector type querying across integrated sources in an analytics knowledge-base (KB). If semantic knowledge bases are used, reasoning-based 'what if' scenarios can also be introduced. Near real-time events and alerts level is the final level that maximizes the actionable window for analysis. Alerts require that the interactive system maintain a continuous watch for available filings via dynamic monitoring by automated linguistic analysis and event-based architectures (Curry, 2004).

Solution Approach

In our solution, requirements are aligned to the stages responsible for preparatory processing of financial information and its use to support analyst investigation. Accordingly, our solution architecture, illustrated in upper area of Figure 1, reflects an analytics pipeline. The pipeline begins with inputs of a business filing and domain knowledge, and then performs a linguistics analysis that creates a semantic analysis of the filing based on the domain knowledge.

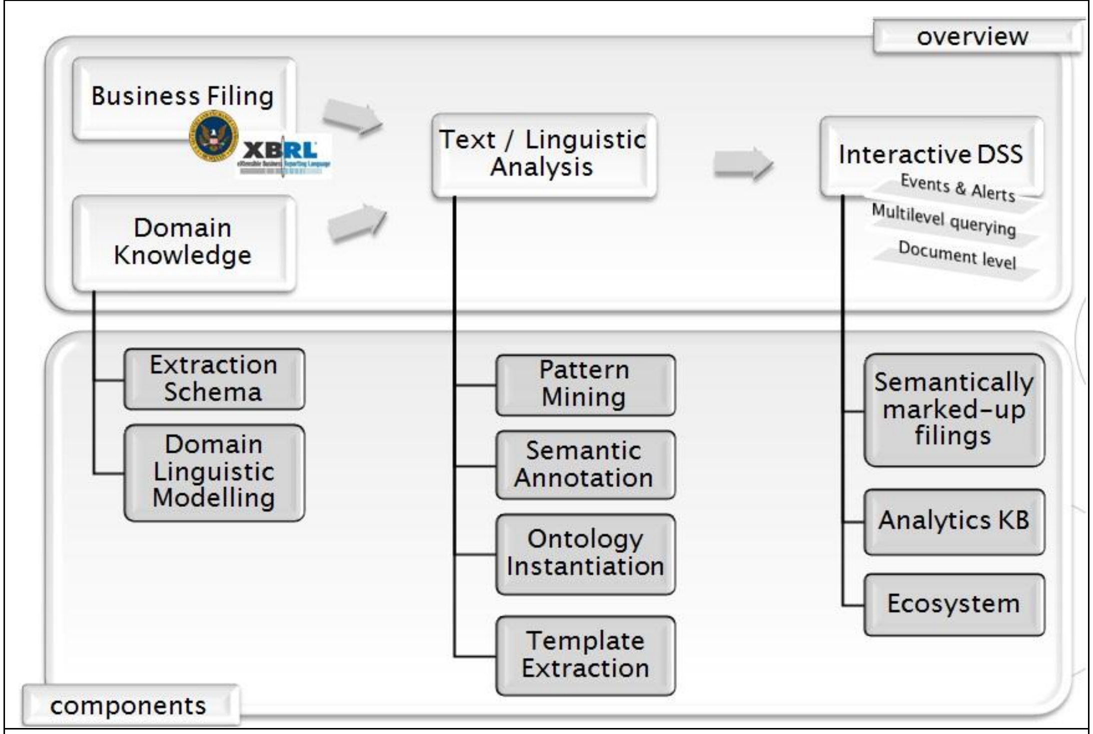


Figure 1. Solution Architecture Overview

The normal basis for a QDSS is information about a set of transactions relating to an organization. We deviate from this conventional notion of transactions, replacing them instead with the novel idea of information threads, called Semantic Paths. The paths, provided automatically by the analytics pipeline are based on the domain knowledge of the competitive analysis ontology. The construction of semantic paths contributes to the analyst's task performance by supporting structured information provision and consumption in a meaningful and succinct form. Underpinning our solution implementation is a domain specific linguistic analysis pipeline. The pipeline is further decomposed in the lower section of Figure 1. Its main components of domain linguistic modelling, linguistic analysis are next introduced, with greater detail and examples to follow in later sections.

Domain Linguistic Modelling. An analyst looking to acquire and interpret information has to negotiate domain specific (accounting and business) language, terminology variation, and hidden meanings. To navigate this linguistic landscape, the analyst draws upon experience and practice-based heuristics. Codification of this domain knowledge needs to capture the information of possible interest, how it should be filtered, condensed, and associated - essentially the knowledge required to perform the analysis task. We used the Developing Ontology Grounded Methods and Applications (DOGMA) Ontology Modelling Methodology (Spyns et al., 2008) to model this linguistic knowledge at a conceptual level. The resulting ontology for competitive analysis is used as an extraction schema to drive automated linguistic analysis of the financial filings. Termed ontology based information extraction (OBIE), it is supported by rule-based, natural language processing.

Linguistic Analysis provides the backbone of the overall analytics pipeline that acts as an enabler for process automation. The General Annotation for Text Engineering (GATE), a component-based architecture and development environment for natural language engineering, was the linguistics tool used (Bontcheva 2004). GATE uses language processing components such as tokenisers, semantic taggers, verb phrase chunkers and sentence splitters. Extraction schema rules within GATE permit automated recognition of accounting and business concepts within the filings text using regular expressions or pattern mining. GATE facilitates semantic annotation to tag individual concept instances found within the filings, making information search easier and more accurate. Semantic annotation is also used to extract text segments containing concept instances that are inserted into the KB.

Interactive DSS leverage the data output from linguistic analysis pipeline. The output is made available to the analyst using three interactive mediums, namely:

- Event driven notification service that performs a continuous data patterns and discovery analysis to alert the analyst. An example is the use of key performance indicator generation to target fraud detection (Nguyen et al., 2005).
- Multilevel query capability that allows information search and scrutiny across multiple filings content, financial instruments (i.e. metrics and ratios), and free-text discussion. This capability supports queries such as "*Who are the internet companies in Texas with revenue > $10M?*".
- Document level interrogation that allows an analyst to drill down through filings discussion sections in an interactive manner that supports the domain knowledge ontology.

The ontology used to semantically annotate the filing is also used as an information map to assist navigation within the filing. Since the ontology captures the qualitative aspect of competitive analysis information requirement, its instantiation will only capture information deemed relevant for analytics purposes. Additionally the KB can be used to query across multiple filings. Complex queries can also be issued across consolidated sources and schema (where additional data sources are included). Multilevel querying across these semantically linked and consolidated sources supports complex keyword, object orientated, path traversal and faceted queries (O'Riain, 2010).

Ontology Based Linguistic Analysis Pipeline

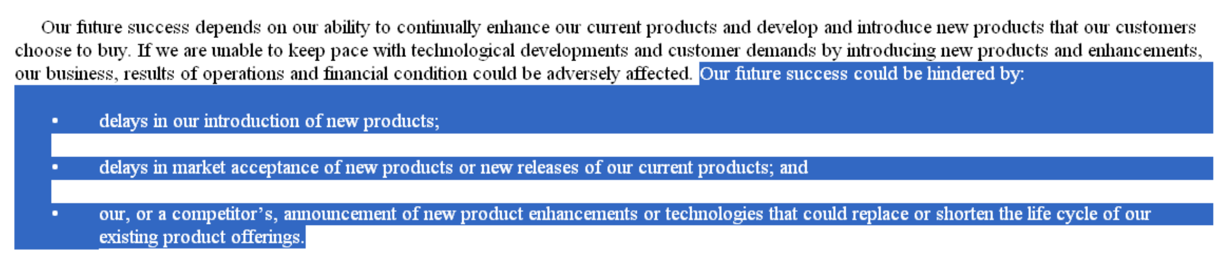
As previously mentioned, analysts tasked with developing insights face significant text and language problems when attempting to negotiate their way through filings' textual disclosure sections. The text can be misleading, non-committal, present ambiguous language and purposely employ varied terminology. Capturing and codification of domain expert tacit knowledge (as defined by Nonaka's (1994) knowledge conversion process) to help add meaning to these narrative sections in any automated way first requires:

- Domain linguistic analysis to establish an understanding of the terminology used and specification of primitives for tacit knowledge codification.
- Formal ontological modelling that reflects the information combination and use within the context of competitive analysis task performance. Ontology modelling details how the output from domain linguistic analysis is formalized as the Competitive Analysis Ontology.

The resulting ontology schema can then be used to drive automated text analysis and act as a semantic framework for information provision and navigation.

Domain Linguistic Analysis

We incorporate examples taken from Form 10-Q filings' discussion sections to illustrate the challenges faced by the analyst to motivate the solution approach. Figure 2 (see below) is an example of information being non-committal as to the actual underlying issue. The disclosure implicitly suggests possible future product revenue implications. For an analyst, this suggests potential issues with: getting the product to market; having a product that there is limited market/demand for; or, more fundamentally, development and release cycle issues.

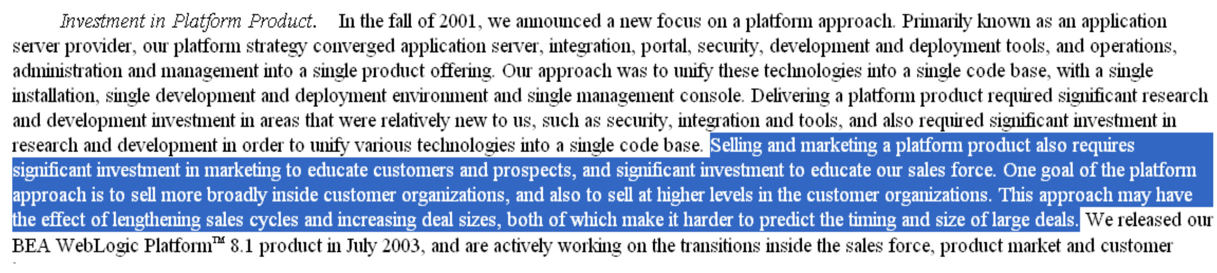


Our future success depends on our ability to continually enhance our current products and develop and introduce new products that our customers choose to buy. If we are unable to keep pace with technological developments and customer demands by introducing new products and enhancements, our business, results of operations and financial condition could be adversely affected. Our future success could be hindered by:

- delays in our introduction of new products;
- delays in market acceptance of new products or new releases of our current products; and
- our, or a competitor's, announcement of new product enhancements or technologies that could replace or shorten the life cycle of our existing product offerings.

Figure 2. Form 10-Q sample extract (a)

Conversely, Figure 3 (see below) is written to appear coherent and convincing. It leaves the reader with an interpretation that does not include the potential underlying causes. Management's disclosure suggests that the strategy of increasing business through targeting of existing customers is responsible for the receipt of sales revenue. Alternative issues with sales force performance, such as lack of customer product awareness or difficulties with a policy that favours existing business over developing new, is not disclosed.



Investment in Platform Product. In the fall of 2001, we announced a new focus on a platform approach. Primarily known as an application server provider, our platform strategy converged application server, integration, portal, security, development and deployment tools, and operations, administration and management into a single product offering. Our approach was to unify these technologies into a single code base, with a single installation, single development and deployment environment and single management console. Delivering a platform product required significant research and development investment in areas that were relatively new to us, such as security, integration and tools, and also required significant investment in research and development in order to unify various technologies into a single code base. Selling and marketing a platform product also requires significant investment in marketing to educate customers and prospects, and significant investment to educate our sales force. One goal of the platform approach is to sell more broadly inside customer organizations, and also to sell at higher levels in the customer organizations. This approach may have the effect of lengthening sales cycles and increasing deal sizes, both of which make it harder to predict the timing and size of large deals. We released our BEA WebLogic Platform™ 8.1 product in July 2003, and are actively working on the transitions inside the sales force, product market and customer

Figure 3. Form 10-Q sample extract (b)

Both examples 'hint' at further questions that although subject to interpretation, do warrant additional investigation as they have direct revenue implication if proved correct. The key challenge for an analyst is to ensure that such sections are identified and highlighted for analysis within some context setting. Previous investigations found that analysts conducting such manual information acquisition dedicated 12.5% of their available time to searching the filings introduction sections and establishing where in the filing to look for relevant information. They then spent the remaining 87.5% of their time analysing the identified sections (O'Riain and Spyns, 2006).

To capture this rich linguistic knowledge and operational know-how, concept mapping was used. Concept map construction is a method used to organizing, capture and codify tacit knowledge, in addition to identifying gaps in knowledge structure (Novak et al., 2008). Using contextual analysis, we iteratively developed concepts maps to represent domain expert's knowledge by:

1. Identifying key domain phrases and variations in the text; establishing the type of information sought.
2. Using term lists to establish concepts (objects), their association, and hierarchy formulation.
3. Organizing concepts into statements or propositions in concept mapping terms. Propositions contain two or more concepts connected using linking words or phrases; form a semantic unit (Novak et al., 2008).
4. Using proposition templates to represents domain expert insight and how different information types are associated.

Figure 4 (see below) outlines a basic concept map excerpt that deals with sales related information for competitive analysis.

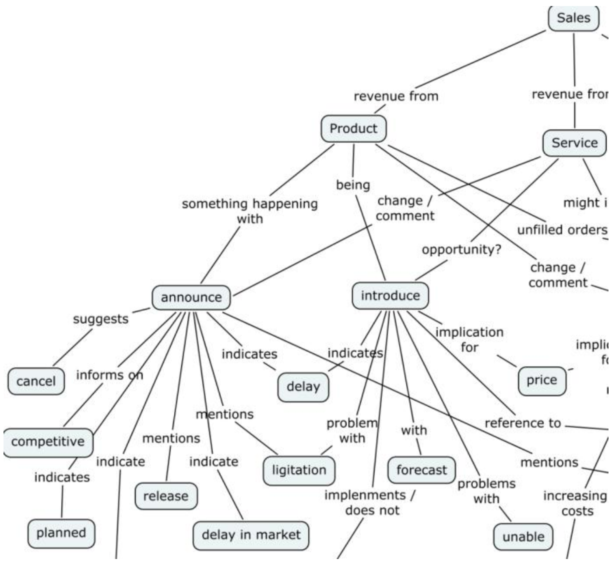


Figure 4. Sales concept map extract

Table 1 outlines a selected proposition template extract from the sales concept map in Figure 4. The propositions outline information concepts together with a domain interpretation of their importance. For example, the third proposition should be interpreted as an interest in finding information that involves the concept "product" and "announce" occurring in near proximity. For the analyst, this may be an indication of product-related announcements that are either:

- Positive, relating to product expansion or entry into new geographies or markets or
- Negative, with hints of falling product sales and product-related issues such as shipping delays or end of the product's life.

Multiple proposition templates can be combined together to build up a series of semantic paths that represent particular elements of the overall information requirement. The semantic paths *construction* represents an analyst reasoning exercise and its *traversal*, a cognitive analysis.

Overall, a total of nine concept maps covering the business context areas of sales, profits, disposal, headcount, acquisition, relationships, R&D and markets were developed. Their propositions represent the fundamental domain linguistic building blocks used to construct the competitive analysis ontology.

Table 1. Sales propositions template extract

	Proposition	**Interpretation with Domain Knowledge**
1.	`Product [being] introduce`	Competition, new revenue, new technology, new market, new geography?
2.	`Introduce [indicates] delay`	Problems with product development or supply chain, same as announce-delay, introduce-delay
3.	`Product [something happening with] announce`	Movement, positive (new market, geography) or negative (delay, end of life, product issues, issues with sales falling short of targets)
4.	`Announce [indicate] delay in market`	Market acceptance of product problems, shift in market requirements?

Competitive Analysis Ontology Development

To support automated linguistic analysis, the domain linguistic knowledge captured during concept mapping had to be formally modelled as an ontology. For this the Developing Ontology Grounded Methods and Applications (DOGMA) - Ontology Modelling Methodology (DOM; Spyns et al., 2008) was used. DOGMA facilitates:

- The transformation of natural language facts (i.e. the concept map propositions) from an initial conceptualization into more
- Formal language-independent statements with informal meaning. The statements are used to develop the ontology and map it to an abstracted level used to support automated information extraction. The statements are referred to as lexons and their abstraction as meta-lexons, respectively.

Lexons represent semi-formal linguistically determined propositions of the domain of discourse. They are written as sextuples:<(γ,ζ): term$_1$, role, co-role, term$_2$> where a lexon is some fact that may hold for some domain within *context* γ , and *natural language* ζ, the *term$_1$*, may have *term$_2$*, occur in *role* with it (and inversely *term$_2$* maintains a *co-role* relation with *term*; Spyns et al., 2007). The lexon engineering process involves initial lexon creation, grounding and generation of meta-lexon.

Lexon Creation. Each proposition (or semantic path) defined in the previous section is treated as a domain statement for the purpose of lexon creation. Accordingly, role labels received enhanced description and context that adhered to the previously introduced business topic areas. Table 2 outlines the development of lexon verbalized facts based on the Sales propositions (cf. previous section). These should be read as the "Product" (concept), "follows" (has a relationship with) "announcement" (concept). Inversely, "Announcement″ (has a relation) "proceeds" with "Product." For the analyst, this represents an interest in events related to products associated with information about product announcements. "Announcement," in turn, is "interested" in information relating to "Delays or Release or Development" (of the product).

Table 2.Sales Lexon Extract

Context (γ) = Sales, Language (λ) = English

Head term (t_1)	***Role (r_1)***	***Co-role (r_2)***	***Tail term (t_2)***
Product	Follows	Precedes	Announcement
Product	Is_described_by	Describes	Announcement
Announcement	Publicizes	Is_announced_in	Delay
Announcement	Publicizes	Is_announced_in	Release
Announcement	Publicizes	Is_announced_in	Development
Announcement	Publicizes	Is_announced_in	Plan

Lexon grounding and meta-lexon creation. Lexon grounding is used to introduce abstract concept identifiers, synonym sets, and natural language descriptions. Table 3 below illustrates the concept "Product" as being assigned a language independent identifier or meta-lexon, here C1002, which accompanies the accounting explanation, definition and alternate term representation (i.e. product can also be referred to as saleable item or goods). Within DOGMA, these lexons are grouped by context and language and represent conceptualizations of real world domains. Altogether, they constitute the DOGMA o*ntology base*.

Table 3. Concept grounding

Sales, English			
Label	***Explanation***	***Glossary***	***Synonyms***
`C1002`	Item manufactured/made	Sold to general public	Saleable item; item, goods
`C1005`	Public statement made for public consumption	Scheduled event	Inform, proclaim, advise
`R1004`	Set of Program to expand	Intention	Postponement

Selection of meta-lexons from the ontology base (for example, as guided by the information paths from Table 2 and adhering to the format *<concept$_1$ – relationship – concept$_2$>)* represents the formal constraint `<C1002, R1004, C1005>`, also known as a "commitment." Contexts (business category areas) in the ontology base are introduced as organizing principles by grouping commitments. The contexts as constructed provide an analyst with an information map for competitive analysis. The accompanying ontology schema represents the means to automate information identification.

Linguistic Analysis

GATE grammar rules are constructed from lexons in the ontology base. GATE performs pattern matching and term instance annotation within the filings. A parallel process extracts surrounding text segments and populated templates for insertion into the knowledge base. Table 4 presents the template for an instance of the commitment rule `<C1005, R1014, C1014>`. The rule represents the semantic path, that class "Announcement" occurs in proximity with the class "Delay in market."

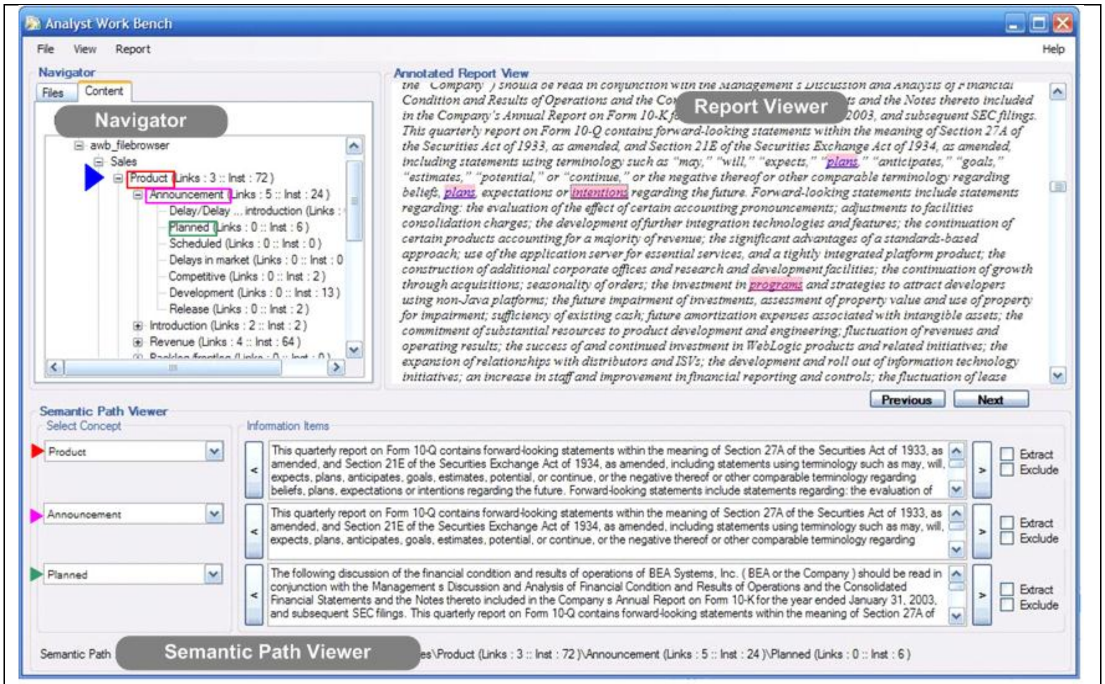
Table 4. Class instance template

Class Type :: Delay in market acceptance	
DocId	`400120`
ReportName	`Company_10-Q_2008-11-10`
Term	`C1014`
AnnotationOffset	`200140`
ReportLineNo	`915`
InstanceID	Auto generated
InfoItemText	If we are unable to keep pace with technological developments ... hindered by: delays in our introduction of new products... delays in market acceptance of new products and services or new releases... ✂...
LinkedConceptHead	`C1014`
LinkedRelationship	`RC1005-C1014`
LinkedConceptTail	`C1005`

Across the disclosure sections and filing, these paths are build up and a proximity algorithm is applied to determine whether they should be allowed instantiate the ontology, thereby making that text segment available to the analyst.

Document Level Interactive DSS

This section presents an overview of the Analyst Work Bench artefact (AWB) that leverages the analytic pipeline output to support competitive analysis performance. The AWB's in-text visualization (presented in Figure 5 below) comprises of three areas: Navigator, Report Viewer, and Semantic Path Viewer. The Navigator displays the competitive analysis ontology extracted dynamically from the knowledge base. The Navigator can be traversed to allow selection of a particular context (”sales”) and it supports drill down through the semantic paths for the ontology concept selection (“planned”). Selecting a context or concept triggers instance highlighting within the semantically annotated filing displayed in the Report viewer panel. The analyst traverses through the instances, selecting ones of interest. Based on the original business context, the ontology path traversed in the Navigator, along with the instance selection, triggers the semantic path population in the Semantic Path Viewer. The contextual informational picture together with the in-text viewing capability allows the analyst search and view relevant information paths for assessment and interpretation. Used in this manner, the semantic paths provide a guided information traversal mechanism through filings text sections to support the competitive analysis task.


Analyst Work Bench
File
View
Report
Help
Navigator
Files
Content
Navigator
awb_filebrowser
Sales
Product (Links : 3 :: Inst : 72)
Announcement (Links : 5 :: Inst : 24)
Delay/Delay ... introduction (Links :
Planned (Links : 0 :: Inst : 6)
Scheduled (Links : 0 :: Inst : 0)
Delays in market (Links : 0 :: Inst : 0
Competitive (Links : 0 :: Inst : 2)
Development (Links : 0 :: Inst : 13)
Release (Links : 0 :: Inst : 2)
Introduction (Links : 2 :: Inst : 2)
Revenue (Links : 4 :: Inst : 64)
Annotated Report View
Report Viewer
Previous
Next
Semantic Path Viewer
Select Concept
Information Items
Semantic Path Viewer

the "Company") should be read in conjunction with the Management's Discussion and Analysis of Financial Condition and Results of Operations and the Co ts and the Notes thereto included in the Company's Annual Report on Form 10-K f 2003, and subsequent SEC filings. This quarterly report on Form 10-Q contains forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, including statements using terminology such as "may," "will," "expects," "plans," "anticipates," "goals," "estimates," "potential," or "continue," or the negative thereof or other comparable terminology regarding beliefs, plans, expectations or intentions regarding the future. Forward-looking statements include statements regarding: the evaluation of the effect of certain accounting pronouncements; adjustments to facilities consolidation charges; the development of further integration technologies and features; the continuation of certain products accounting for a majority of revenue; the significant advantages of a standards-based approach; use of the application server for essential services, and a tightly integrated platform product; the construction of additional corporate offices and research and development facilities; the continuation of growth through acquisitions; seasonality of orders; the investment in programs and strategies to attract developers using non-Java platforms; the future impairment of investments, assessment of property value and use of property for impairment; sufficiency of existing cash; future amortization expenses associated with intangible assets; the commitment of substantial resources to product development and engineering; fluctuation of revenues and operating results; the success of and continued investment in WebLogic products and related initiatives; the expansion of relationships with distributors and ISVs; the development and roll out of information technology initiatives; an increase in staff and improvement in financial reporting and controls; the fluctuation of lease

Product

This quarterly report on Form 10-Q contains forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, including statements using terminology such as may, will, expects, plans, anticipates, goals, estimates, potential, or continue, or the negative thereof or other comparable terminology regarding beliefs, plans, expectations or intentions regarding the future. Forward-looking statements include statements regarding: the evaluation of

□ Extract
□ Exclude

Announcement

This quarterly report on Form 10-Q contains forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, including statements using terminology such as may, will, expects, plans, anticipates, goals, estimates, potential, or continue, or the negative thereof or other comparable terminology regarding

□ Extract
□ Exclude

Planned

The following discussion of the financial condition and results of operations of BEA Systems, Inc. (BEA or the Company) should be read in conjunction with the Management s Discussion and Analysis of Financial Condition and Results of Operations and the Consolidated Financial Statements and the Notes thereto included in the Company s Annual Report on Form 10-K for the year ended January 31, 2003, and subsequent SEC filings. This quarterly report on Form 10-Q contains forward-looking statements within the meaning of Section 27A of

□ Extract
□ Exclude

Semantic Path es\Product (Links : 3 :: Inst : 72)\Announcement (Links : 5 :: Inst : 24)\Planned (Links : 0 :: Inst : 6)

Figure 5. AWB In-Text Visualization

Related Work

Recognising the benefits of IE to retrieve and supply information, the financial community has been actively developing applications to extract from wider business ecosystem sources such as business filings, news articles and company web sites (Gerdes 2003; Schumaker and Hsinchun 2009, Bovee et al., 2005; Grant et al., 2006). Financial news articles in particular have a long history of being mined for assorted information such as company restructuring and general macroeconomic information (Costantino et al., 1996), earning facts (Conlon et al., 2007) or stock market prices (Schumaker and Hsinchun, 2009).

US SEC filings extraction systems organize their discussions around the type of information targeted. EDGAR2XML (Leinnemann and Schlottmann., 2001) was an early attempt that targeted Y2K remediation efforts in 10-K corporate disclosures using keyword searches. Identified test segments were extracted based on proximity within the text and output presented as a ranked text segments list. The Extraction Agent for SEC Edgar Database (EASE) automatically identifies and extracts consolidated balance sheet sections of 10-Q/10-Ks', using the vector space model (algorithm) and regular expressions, a coding format for specifying string patterns to recognise in text (Stampert et al., 2008). The EDGAR-Analyzer software agent also extracts from 10-Q/10-Ks, using regular expressions but targets stock market investor items (Gerdes, 2006). Analyses of the semi-structured balance sheet income statements and cash flows statements, including the narrative sections of the Form 10-Q/10-K, is performed by the Financial Reporting and Auditing Agent with Net Knowledge (Bovee et al., 2005). Regular expressions were used to extract stock prices and earnings per share information for use in financial metrics generation. The EDGAR Extraction System (Grant et al., 2006), also targeting the 10-K statement and disclosure sections, extracted pro forma net income, earnings per share information, and fair value options. Statistical methods and regular expressions based on a language model of forty terms defined with domain experts were used to target search and extraction. Midas (Hernandez et al., 2010) extracts and aggregates facts from both structured and unstructured SEC and US Federal Deposit Insurance Corporation filings using a combination of statistical methods and regular expressions. A network (graph) of interconnected relationships between banking institutions is constructed for use in systemic risk analysis. Upon network analysis, the network identifies critical banking hubs as institutions that pose the greater systemic risk.

Common characteristics of the related filing extraction systems are: 1) the lack of a sharable, formal model to specify domain specific information requirements and extraction schema; 2) the use of a dedicated analytics pipeline that generates directly consumable structured output; and 3) a lack of assistance for qualitative evaluation. Ontologies modelled on domain tasks offer a shared semantic understanding and opportunity to apply semantic interpretation. With an inherent hierarchy, the ontology also offers a framework to supports directed search and access to a semantically enhanced data set.

Pipeline Evaluation

Our approach to pipeline evaluation was to observe its contribution and use as part of a qualitative document level interactive DSS. The evaluation had two distinct parts. The first investigated the pipelines contribution to competitive analysis information provision through its ontologically driven linguistic analysis capability. The second investigated the effect and impact of pipeline output consumption on the qualitative aspects of competitive analysis performance. This first was evaluated using *performance* methods from information retrieval and the second with *usability* measures from information science (described below). The criteria adopted for pipeline evaluation from both performance and usability experiments are outlined in Table 5.

Table 5 Pipeline Evaluation Criteria

Criteria	***Performance***	***Usability***
Artifact usable as basis for experiment	AWB QDSS, as a pipeline consumer, provider of relevant information	AWB, as an interactive QDSS
Criteria representing system objectives	Relevance of information provided	Usefulness and usability of information provided / DSS environment
Measuring instrument	Relevance judgment expresses as a binary weighting	Success determination using Likert scale
Measures	Precision, recall	Weighted average
Methodology for measurement, evaluation performance	Based on the competitive analysis task	Questionnaire survey of participants

The overall goal of the qualitative QDSS system developed is to assist an analyst perform the qualitative aspect of a competitive analysis, by providing relevant information for interpretation. Due to the complexity involved and resourcing constraints, five senior domain experts were used as part of an analyst focus group to perform both controlled experiments. The QDSS demonstrator and experiments were part of an industrial use case.

For the performance experiment, analysts were instructed to perform a competitive analysis on a selected corpus by identifying and annotating relevant text segments, first manually and then using the AWB DSS. Manually annotated text segments were compared with those found using the AWB QDSS and the information retrieval metrics of precision and recall used to report on performance. "Precision" determines the percentage of retrieved text segments (or classically documents) that are relevant, while "recall" determines the percentage of relevant text segments that are retrieved. "Relevance" was defined by the focus group as information of use to competitive analysis and assigned by majority weighting. With a goal of finding relevant text segments, previously missed by the manual evaluation, precision rather than recall becomes the performance metric of interest. Relevant text segments identified by the analyst using the AWB QDSS, which was not previously annotated during manual analysis, increased precision from 16.7% to 23%, representing a significant 37% improvement on manual results.

Although performance results indicate the usefulness of an ontology and linguistic analysis to enhance information provision, it does not inform on any potential contribution to the AWB DSS in terms of information displayed and consumed. The DeLone and McLean Model of Information Systems Success (2003) was modified based on Wu et al. (2006); and Nielsen (1993), to develop and introduce information-centric heuristics measures. A total of five dimensions and 33 instruments covering: system operational characteristics; information quality addressing content, and context and linkage quality; user satisfaction, dealing with DSS usage experience; perceived benefit of the DSS to the user and; overall system usage, were defined. Further insight into the use of semantic paths as an organising principle in ontology construction, and their contribution to pipeline output and DSS activity, are obtained with a report on usability results relating to information quality and perceived system benefits in Table 6.

Having completed a competitive analysis using the AWB DSS, the analyst focus group using the usability determination questionnaire, rated their experiences using a seven point scale with range from *strongly agree (1)* to *strongly disagree (7)*. Respondent number is weighted using the scale range, respondent total represents the sum of the number of evaluators (or raters) participating in the question and the weighted value divided by respondent number provides the rating average. The rating average is compared to the scale range to determine the level of agreement for the particular instrument dimension (question). A scale mean indicating neutral (or undecided) is 4; mean ratings of 3 or less indicate moderate to good agreement and; mean ratings of 5 or greater indicate moderate to strong disagreement to the questions posed.

Content quality ratings indicate moderate agreement with the system providing information in a useful manner. More importantly, there is agreement on the majority of the analyst information requirements being addressed, but with some reservation. There is further agreement that information is provided in a useful manner, is directly usable and easily consumable by the analyst. We find analyst agreement with ease of text segment extraction, which is typically used for further analysis. This result indicates that the approach of combining pipeline output (ontology and knowledge base) to support analysis works. Overall, the ratings suggest that the pipelines domain ontology in terms of information identification and provision was effective.

Further results indicate that context and linkage instruments identify further moderate agreement on the artifact contributing to support the information requirements of the competitive analysis task. While relevant information is provided, using the ontology as the main means of interactive search and navigation received a neutral rating, indicating that improvement is required. The ratings point to ontology utility in terms of task and information requirement structuring.

Perceived benefits from system usage report moderate agreement relating to supporting task performance. Agreement is achieved in terms of artifact contribution to information management at the personal level and moderate agreement on information management at the more general information space (business filing) level. The moderate agreement rating on time transfer from manual information gathering to actual analysis is attributed to previously mentioned user interface shortcomings. These ratings also suggest that the ontology does reflect analyst information requirements and information association patterns.

Overall, the performance results indicate the usefulness of an ontology based linguistic analysis pipeline to support qualitative, interactive document level competitive analysis task performance. Usability results reinforce these findings, with further indication of the additional contribution that the ontology makes to task structuring and structured information provision.

Table 6. Usability determination questionnaire and results (applicable dimensions)

	Res.No.	Res.#	Rating Avg.
Information Quality - Content Quality			
The output is presented in a useful manner	15	5	3
The content representation provided by the system is logical	15	5	3
The information provided is relevant and helpful for the task	10	5	2
The information content meets your information requirements	16	5	3.2
The information provided is meaningful	11	5	2.2
AWB makes it easy for me to extract information of use	12	5	2.4
Information Quality - Context & Linkage Quality			
The information navigation process is useful for information identification	21	6	3.5
The information navigation process is logical and fit	21	5	4.2
Information is presented in a way that provides a useful aggregate view	16	5	3.2
The representation mechanism is successful in structuring the analysis task	16	5	3.2
Does the information navigation process assist the cognitive workload in task performance	14	5	2.8
Provides information in context in a manner that is understandable, accessible and applicable to the task	16	5	3.2
Perceived benefit - *The valuation of system benefits to the user*			
The system assists the cognitive workload in task performance	15	5	3
The system assists the ability to efficiently and effectively complete the task	13	5	2.6
The system assists in managing the information overload	13	5	2.6
The system provides more time for actual analysis	16	5	3.2
The system helps the effective management of the information space	14	5	2.8

Conclusion

Overall, the AWB as a qualitative interactive DSS was found to assist an analyst perform a competitive analysis. The ontology based on analyst heuristics was found to be an effective framework to represent domain knowledge and structure the information requirement. Using the structured output from the linguistic analysis pipeline as a context basis in a QDSS application was found to be effective in terms of information provision and supporting information extraction. Additionally, the ontology-based semantically enhanced filings assisted analyst navigation within the information space. The design approach illustrates the ability to codify otherwise tacit domain knowledge to drive automated qualitative analysis and enable common measures for qualitative evaluation.

To progress efforts towards interactive querying and event provision, future work will look to pipeline enhancements in the areas of:

Domain Knowledge Inclusion. To deal with the demands of different text analytics, domain and topic specific ontologies with have to be introduced into the pipeline seamlessly. One avenue of investigation is the use of XBRL, IFSB, IASB financial and regulatory taxonomies standards as lexical resources to automatically extract ontology primitives (O'Riain et al., 2012).

Open Data Consumption. Extend the definition and range of interactive data to include the wider web of Open Data (Curry et al., 2012). Making output available in RDF as Linked Data allows linkage and interoperability with a wider financial data space. A promising area for application to qualitative financial information is the multi-level, schema-less querying of heterogeneous data sets to support best effort response across RDF data (Freitas et al., 2012).

Multilingual Support. Multilingual access and querying of financial and business reports, across differing jurisdictions (Declerck et al. 2008, 2010), with the use of a multi-lingual analytics pipeline.

Acknowledgements

The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2), and the EU FP7 MONNET project under Grant Agreement No. 248458.

References

Bovee, M. Kogan, A., Nelson,K. and Srivastave, R. 2005. "Financial Reporting and auditing agent with net knowledge (FRAANK) and eXtensible business reporting language (XBRL)," *Journal of Information Systems*, 19(1): p. 19-41.

Chakraborty V. and Vasarhelyi M.A. 2010. "Automating the process of taxonomy creation and comparison of taxonomy structures," in *19th Annual Research Workshop on Strategic and Emerging Technologies, American Accounting Association*. San Francisco, CA, USA.

Conlon, S. J., Lukose, S., Hale, J. G. and Vinjamur, A. 2007. "Automatically Extracting and Tagging Business Information for E-Business Systems Using Linguistic Analysis," *Semantic Web Technologies and E-Business: Toward the Integrated Virtual Organization and Business Process Automation*. A. F. Salam, Stevens, J. (eds.), IGI Global: 101-26.

Costantino, M., Morgan, R.G. and Collingham, R.J. 1996. "Financial information extraction using pre-defined and user-definable templates in the LOLITA system," *Proceedings of the Fifteenth International Conference on Computational Linguistics*.

Curry, E. 2004. "Message-Oriented Middleware," *In Middleware for Communications*, Q. H. Mahmoud (ed.), Chichester, England: John Wiley and Sons, pp. 1–28

Curry, E., Freitas, A., O'Riain, S. 2012. "The Role of Community-Driven Data Curation for Enterprises,*" in Linking Enterprise Data*, " (ed.), D. Wood, Springer US: 25-47.

Curry, E, Harth, A. and O'Riain, S. 2009. "Challenges Ahead for Converging Financial Data," *In W3C Workshop on Improving Access to Financial Data on the Web*, Arlington, Virginia, USA.

Bontcheva, K., Tablan, V., Maynard, D. and Cunningham, H. 2004. "Evolving GATE to meet new challenges in language engineering," *Journal of Natural Language Engineering* 10(3/4): 349-373.

Debreceny, R. and Gray, G.L. 2001."The production and use of semantically rich accounting reports on the Internet: XML and XBRL,"*International Journal of Accounting Information Systems*. 2(1), 47-74.

Debreceny R., Farewell S., Felden C., d'Eri A. and Piechocki M. 2011. "Feeding the Information Value Chain: Deriving Analytical Ratios from XBRL filings to the SEC," *22nd XBRL Intl. Conf.,* Brussels.

DeLone W. H. and McLean E.R. 2003. "The DeLone and McLean Model of Information Systems Success: A Ten Year Update," *Journal of Management Information Systems*. 19(4): p. 9-30.

Declerck, T., Krieger, Hans-Ulrich., Horacio, S., and Spies, M. 2008. "Ontology-Driven Human Language Technology for Semantic-Based Business Intelligence," *Proceedings of the 18th European Conference on Artificial Intelligence*, IOS Press.

Declerck, T., Krieger, H.U., Thomas, S.M., Buitelaar, P., O'Riain, S., Wunner, T., Maguet, G., McCrae, J., Spohr, D. and Montiel-Ponsoda, E. 2010. "Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe," *Internal Financial Control Assessment Applying Multilingual Ontology Framework*, Eds. József Roóz and János Ivanyos, Készült HVG Press Kft.

Freitas, A., Curry, E., Oliveira, J.G., O'Riain, S. 2012a. "Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends," *Internet Computing, IEEE* 16(1): 24-33.

Gerdes, J. (2003). "EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC's EDGAR database," *Decision Support Systems*, Vol. 35, No. 1, p. 7-29.

Grant, G.H. and Conlon, S.J. 2006. "EDGAR Extraction system: An automated Approach to Analyze Employee Stock Option Disclosures," *Journal of Information Systems* Vol. 20(2): p. 119-142.

Hernandez, M., Ho, H., Koutrika, G., Krishnamurty, R., Popa, L., Stanoi, I., Vaidyanathan, S. and Das, S. 2010. *Unleashing the Power of Public Data for Financial Risk Measurement, Regulation, and Governance*. WWW. Raleigh, North Carolina.

Korman, R. (1998). "Investing It", *Mining for Nuggets of Financial Data*. The New York Times.

Leinnemann C., and Schlottmann, F. 2001. *Automatic extraction and analysis of financial data from the EDGAR database*. South African Journal of Information Management. 3(2).

Li-Yen, S., Ching-Wen, C.and Shiue, W. (2009). "The development of an ontology-based expert system for corporate financial rating." *Journal of Expert Systems with Applications* 36(2): 2130-2142.

Massetti, B. 1996 "An Empirical Examination of the Value of Creativity Support Systems on Idea Generation," *MIS Quarterly*, Vol. 20, No. 1, p. 83-97.

Nonaka, I. 1994. "A Dynamic Theory of Organizational Knowledge Creation." *Organization Science* 5: pp 14 -37.

Novak, J.D. and Cañas, A.J. 2008. "The Theory Underlying Concept Maps and How to Construct and Use Them, Technical Report IHMC CmapTools," Institute for Human and Machine Cognition: Florida.

O'Riain, S. Harth, A and Curry, E. 2011. "Linked Data Driven Information Systems as an enabler for Integrating Financial Data". In Alexander Yap (ed.), *Information Systems for Global Financial Markets: Emerging Developments*, IGI Global.

O'Riain, S., Curry, E., and Harth, A. 2012. "XBRL and open data for global financial ecosystems: A linked data approach." *International Journal of Accounting Information Systems* 13(2): 141-162.

Osterle, H., Becker, J., Frank, U.,Hess, T., Karagiannis, D.,Krcmar, H., Loos, P., Mertens, P., Obrweis, A.,Sinz, and Elmar J. 2011. "Memorandum on design-oriented information systems research," *Eur J Inf Syst* 20(1): 7-10.

Pfeifer, S. 2007. "How to deconstruct the annual report". *The Sunday Telegraph*, April.

Sauter, V.L. and Free, D. 2005, "Competitive intelligence systems: qualitative DSS for strategic decision making," *SIGMIS Database*, 36 (2), p. 43-57, ACM, NY, USA.

Schumaker, R. and Hsinchun, C. 2009. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Transactions on Information Systems* 27(2): 12:1-12:19.

Shaker, A. Z. and Chaples, S.S. 1993. "Blind Spots in Competitive Analysis," *in The Academy of Management Executive* (1993-2005) Academy of Management.

Sheng, Y. P. Mykytyn, P.P. and Litecky, C.R. (2005) Competitor Analysis and Its Defences in the EMarketplace. Communications of the ACM. 48(8): p. 107-112.

Spyns, P., Tang, Y. and Meersman,R. 2008. "A model theory inspired collaborative ontology engineering methodology," *Journal of Applied Ontology*, 2007. 4(1-2), P13-39, IOS Press.

Nguyen, T.M., Schiefer, J., Tjoa, A.M. 2005. "Sense & response service architecture (SARESA): an approach towards a real-time business intelligence solution and its use for a fraud detection application," *Proceedings of the 8th ACM international workshop on Data warehousing and OLAP*, DOLAP, p. 77—86, ACM, Bremen, Germany.

Nielsen J. 1993. *Usability Engineering*. Academic Press ed.

Stampert, T., Seese, D., Weinhardt, C. and Schlottmann, F. 2008. *Extracting Financial Data from SEC Filings for US GAAP Accountants. In Handbook on Information Technology in Finance* (eds.) Bernus. P., Baewicz, J., Schmidt, G. J., Shaw, Michael J.,. Heidelberg, Springer 357-375.

Wu Jen-Her and Wang Y.-M. 2006. "Measuring KMS success: A re-specification of the DeLone and McLean's model," *Journal of Information Management*. 43: p. pp. 728-739.

Zahra, S. A., Chaples, S.S. 1993. "Blind Spots in Competitive Analysis." *The Academy of Management Executive* (1993-2005) 7(2): 7-28.