# Computational Modelling for Bankruptcy Prediction: Semantic data Analysis Integrating Graph Database and Financial Ontology

1st Natalia Yerashenia
*School of Computer Science and Engineering*
*University of Westminster*
London, United Kingdom
w1578366@my.westminster.ac.uk

2nd Alexander Bolotov
*School of Computer Science and Engineering*
*University of Westminster*
London, United Kingdom
A.Bolotov@westminster.ac.uk

***Abstract*—In this paper, we propose a novel intelligent methodology to construct a Bankruptcy Prediction Computation Model, which is aimed to execute a company's financial status analysis accurately. Based on the semantic data analysis and management, our methodology considers Semantic Database System as the core of the system. It comprises three layers: an Ontology of Bankruptcy Prediction, Semantic Search Engine, and a Semantic Analysis Graph Database system.**

**The Ontological layer defines the basic concepts of the financial risk management as well as the objects that serve as sources of knowledge for predicting a company's bankruptcy. The Graph Database layer utilises a powerful semantic data technology, which serves as a semantic data repository for our model.**

**The article provides a detailed description of the construction of the Ontology and its informal conceptual representation. We also present a working prototype of the Graph Database system, constructed using the Neo4j application, and show the connection between well-known financial ratios.**

**We argue that this methodology which utilises state of the art semantic data management mechanisms enables data processing and relevant computations in a more efficient way than approaches using the traditional relational database. These give us solid grounds to build a system that is capable of tackling the data of any complexity level.**

***Index Terms*—semantic data analysis, graph database, ontology, financial analysis, financial ratios, bankruptcy prediction, computational model, FIBO, Neo4j**

## I. Introduction

We are interested in the problem of computational modelling in bankruptcy prediction. It is notable that in the contemporary digital world, companies, accumulating vast amounts of financial information, need to identify those data that are substantial for this prediction. Moreover, it is necessary to ensure that the data collected are of the required quality and are consistent. We propose a concept of an intelligent, analytical system to perform the prediction of the companies' bankruptcy. The system processes financial information of a company and undertakes a comprehensive investigation of companies' financial activities during a particular designated time period. We aim at creating a *Bankruptcy Prediction Computational Model (BPCM)* which is capable of the automated construction of an expert analytical report, where various data and information are presented reliably and objectively. This will assist significantly in advancing companies' bankruptcy analysis strategy by increasing the level of its consistency, reliability, and efficiency. We apply a holistic approach, targeting a systematic, functional, technical, methodological and informational compatibility of the components of the analysis into an integrated unity.

The main feature of the proposed system is the consolidation of the information management with the decision-making process to serve the prediction. This involves modern methods of searching, processing and storing potentially large amount of heterogeneous data together with advanced machine learning methods. One of the central notions in our study is 'Semantic data' – using this term in relation to computational systems we emphasise our interest in the *meaning* of data.

On the background of a limited number of studies in comprehensive methods of bankruptcy prediction, our methodology is based upon the utilisation and integration of the following three semantic data management mechanisms: semantic search, semantic ontology, and graph database.

To the best of our knowledge, there are no existing approaches that are integrating these techniques not only in the area of bankruptcy prediction but also in the financial data analysis in general. We note that in the literature on bankruptcy prediction, single classical instruments of multivariate statistical analysis [1] are applied independently. Moreover, the problem of assessing the financial state of a company using a mathematical model, which involves semantic data analysis and management, remains well under-explored. In this paper, we address the latter defining the process of the Semantic Database System construction.

The remaining of the paper is organised as follows. Section II provides an account of related work. In Section III we describe the problem set-up and specify the development of the Semantic Database System. Section IV includes an overview of BPCM methodology. Section V describes the architecture of the Ontology and Graph Database and illustrates their functionality. Finally, in Section VI we summarise the contributions of the research provided, discuss future work, and

2378-1971/19/$31.00 ©2019 IEEE
DOI 10.1109/CBI.2019.00017

draw conclusions.

## II. Semantic data processing and prediction techniques: related works

**Statistical Methods of Bankruptcy Prediction.** Although the combination of machine learning methods and data preprocessing for bankruptcy prediction using a semantic approach, as proposed in this paper, has not yet been observed in a single methodology, it is worth to consider some related work. The idea of creating a bankruptcy prediction model became popular as early as in the 1960s. However, the statistical methods have been mostly used as the basis for prediction. In [2] the financial ratios of bankrupt firms are compared with the performance indicators of companies that remained competitive. The paper analyses a group of companies, in a five years interval. The study considered twenty coefficients and showed that there were quite significant differences in the financial ratios of the two groups of firms. Bankrupt firms had a lower return on assets and return on sales, a higher proportion of accounts receivable, lower values of current and absolute liquidity ratios, but a higher level of debt.

According to the approach proposed in [3], which considered a specific country (USA) and a dedicated time interval (the 1960s), a set of separate five financial ratios of companies was formed. The choice of these ratios was based on some preliminary expert analysis. In the 5-dimensional space created by the selected coefficients, a hyperplane is drawn, which best separates successful companies from bankrupt companies, based on the historical financial statistics. As a result, the well-known $Z$-score model appeared. The approach of [3] known as *the method of Multiple Discriminant Analysis (MDA)* was also adopted in subsequent works, for example, in [4], where a similar model has been developed for the UK companies.

A more accurate model was proposed in [5] with the analysis of the data for a significant number of companies (2163). This was one of the first works using the regression method instead of the discriminant analysis method. The latter and similar methods are not resistant to fluctuations of the original data. The main limitation of this method is that its conclusions on one particular company are based on a set of data on a multitude of other comparable companies. Thus, the individual features of the given company are not taken into account, hence one may question the reliability of the grounds for the conclusions about the likelihood of the bankruptcy.

More recently, various methods of machine learning have supplanted traditional statistical methods.

**Machine Learning Methods of Bankruptcy Prediction.** One of the core examples of machine learning techniques applied for financial analysis is a fuzzy sets theory, presented in the fundamental work [6]. The original concept of this method is to build a functional correspondence between fuzzy linguistic descriptions (such as ”low”, ”bad”, ”average”, etc.) and special functions revealing the degree to which the values of the measured parameters belong (length, temperature, weight, etc.) to fuzzy descriptions. Fuzzy set methods have been applied in economics since the late 1970s. Among relevant work, we mention [7], [8], and [9] where new aspects of fuzzy sets theory were studied, and new mathematical models for determining financial problems were formulated.

A neural network is another example of machine learning applied in bankruptcy prediction [10]. Neural networks are trained. In the process of learning, the neural network detects dependencies between the input and output data [11]. At the learning stage, synaptic coefficients are calculated in the process of solving a neural network of problems in which the desired answer is determined not by the rules, but with the help of examples grouped into training sets.

Later, after the development of learning algorithms, the resulting models were used for a variety of practical purposes: pattern recognition, control problems, and forecasting problems. Neural networks can also be used to diagnose the companies' bankruptcy level. The evolution of the application of neural networks in business is described in [11].

A comparison between MDA and neural network technologies was made in [12]. The results confirm that neural network technologies have proven to be more efficient than the MDA-based model, later research [13] also confirmed this result.

It should be noted that earlier studies of bankruptcy prediction mechanisms did not take into account data pre-processing.

**Ontologies and Graph Databases in Finance.** In the 1980s, the 'Ontology' term migrated from philosophy to computer science field when it was used by several kinds of research on Artificial Intelligence (AI). At the end of the 90s, it actively used in such areas as information integration, information search on the Internet and knowledge management. Later ontologies began to be seen as a crucial element of Semantic Web [24]. One of the pioneering papers studying ontology in the computational framework was [14]. The process of creating ontologies relating to financial data was offered in [15], [16] and [17].

Graph Database (Graph DB), on the other hand, is a relatively new development and the analysis of this method has not been paid significant attention in academic literature. Among a few studies of this topic are [37], which provides a review of the advantages and the limitations of Graph DB, [18] and [19] which compare existing development programs, and [20], [21] and [22] which are evoted to the construction of the Graph DB themselves.

To the best of our knowledge, this technology has not been applied in the financial data of companies.

## III. Problem Set-up

**Features of the Underlying Dataset - Big Four 'V' + 'R'.** We argue that the financial dataset to be analysed for our purposes, figuratively speaking, can be characterised by four 'V' and 'R'. It shares most (four out of five big 'V') of the qualities of *Big Data* – Variety, Velocity, Veracity and Value [23] being not dependent on the Volume. However, we underline the fifth, 'R', feature of these financial data – an extremely high level of Relationships. Indeed, similar to big data, in our case, we have heterogeneous data, coming from different sources. These components of a company's financial system can be (and usually this is the most common practice) described in the form of relational tables (traditional database), e.g. it is easy to present a balance sheet or income statement in such a way. However, to show the interconnections between all elements of these tables, it is necessary to create a number of tables of a different structure containing thousands of objects. In this case, the efficiency of database management and search are substantially affected. For example, it becomes problematic to formulate a general query to several databases, because of the difference in objects and attributes of the domain or changes in objects over time. When the data are inserted, updated or deleted, the integrity constraints for the database with changing objects should be checked and assured that the data will be consistent after all modifications [24]. Also, as it was mentioned before, there is a problem of the integration of new nodes into the system. When adding a new node, it is essential to check the data and the data schema for consistency with the information already available in the system [25].

Although traditional *Relational Databases* still dominate among data storage facilities, these systems would not be suitable for the purposes of our financial analysis being unable to tackle the requirements of the *'Big Four V + R'*.

There are *NoSQL systems* that extend the capabilities of traditional databases by allowing to deal with the four 'V'. It would have been possible to utilise solutions developed for big data management, such as Scribe, HBase, Cassandra, etc. [26]. However, it would bring unnecessary complications as these solutions have been developed to tackle the Volume of the big data, the feature that our target data would not have. Moreover, these solutions were not designed to tackle the 'R' feature of our datasets.

These observations bring additional argument to use a particular class of the non-relational NoSQL repositories, *Graph DB* - their ability to tackle interconnectedness [37].

**Semantic Approach to Data.** While the Value feature with respect to Big Data, normally denotes the worthiness, usefulness of the data (a pragmatic meaning), we bring here the semantic approach as a specific framework in which the desired usefulness will be tackled.

We argue that the range of problems related to integrating financial data can be resolved within the framework of a *semantic approach* to data modelling. The semantic approach in our case is used to measure the connotative content of information, i.e. it makes information retrieval more accurate and relevant to the query [24].

We believe that the representation of information in the form of a graph is very beneficial for this approach. Indeed, when objects are correlated with the nodes of a graph and their relationships are associated with the edges, it is possible to add data from different sources into one structure effectively. The graph structure is the most convenient for representing complex engineering information when we need to constantly develop and maintain the data model throughout its life cycle [27]. Information in a semantic form is easily rebuilt and expanded when new sources become available, without the need for primary processing of the storage system, as in the case of traditional databases. The semantic approach supports flexible and extensible information models, allowing to combine financial information and avoiding the termination of the editing of the information model with each iteration of extension. The format and volume of the processed data can be specified as information requirements evolve as well as the knowledge of the needs of the participants in the life cycle develops.

Additional advantages of semantic approach are provided by the use of *ontological standards*, which allows not only to obtain information from different sources in one flexible and expandable format but also to interpret it in the same way [28].

Further, the semantics of the obtained data can be clarified from previously known sources (reference data libraries) using the same standardised technologies and tools that are used for data exchange. The combination of semantic and ontological standards helps to organise the exchange and comparison of data, the identification of conflicts and the harmonisation of contradictions. Semantic and ontological standards are also suitable for data processing tools of varying degrees of data complexity. Besides, publicly accessible digital data storage specification called *open data format* is free from licensing restrictions [29].

Hence, to solve the project tasks concerning the Semantic Database System, it is necessary to address problems connected with bridging the research gaps as shown in the 'problem set-up diagram' in Fig. 1. The gaps considered in this paper are shown as green arrows; the gaps which we will address at the next steps of the research are shown as red arrows.

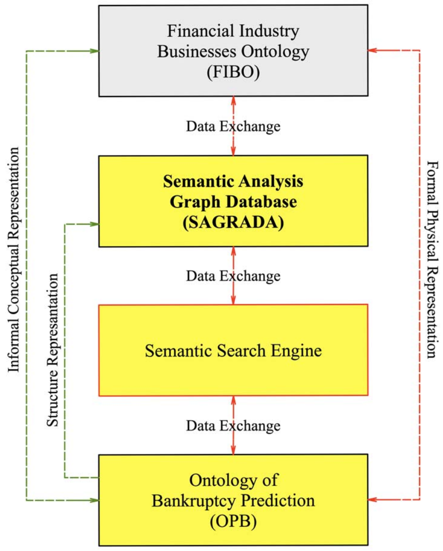
Financial Industry Businesses Ontology (FIBO)
Data Exchange
Semantic Analysis Graph Database (SAGRADA)
Data Exchange
Semantic Search Engine
Data Exchange
Ontology of Bankruptcy Prediction (OPB)
Informal Conceptual Representation
Structure Represantation
Formal Physical Representation

Fig. 1. Problem Set-up Visualisation Graph

**Ontology.** First, it is necessary to create an *Ontology* containing all the reference information used for the financial analysis of a company. This problem can be solved in two stages: the creation of the ontology structure itself, the so-called descriptive/explanatory model. Here the following should be created:

(1) an informal conceptual catalogue of all the terms (objects), their types and relationships between them represented in the form of a graph;
(2) a logical-semantic functional model using the resource metadata description language, Resource Description Framework (RDF); the primary purpose of RDF is to present statements about resources in the form that is equally well perceived by both human and machine.

This new ontology - the *Ontology of Bankruptcy Prediction (OBP)*, is based on the principles of the existing financial ontology *Financial Industry Business Ontology (FIBO)* [30]. However, in comparison with FIBO, the OBP ontology is far more compact. Besides, it purely concentrates on the information used to assess the level of financial solvency of a company, while FIBO deals with the financial sector in general.

**Graph Database.** The second problem is the creation of a *Semantic Analysis Graph Database (SAGRADA)* linked to the OBP Ontology. We utilise the Neo4j environment [31], equipped with its own declarative query language Cypher [32] – a Graph Database analogue to SQL.

Additionally, one of the significant problems in developing the Sematic Data System for BPCM model is to identify the most efficient format of data exchange between the SAGRADA and the OBP. However, we will address this issue after completing the development of both the Ontology and the Graph Database.

**Notation.** We will use the following terminology. The Bankruptcy Prediction Computation Model will be abbreviated as 'BPCM'; the Semantic Database System will be abbreviated as 'SDS', for the Ontology of Bankruptcy Prediction and the Semantic Analysis Graph Database we will use 'OBP' and 'SAGRADA', respectively. However, to simplify reading, we will often annotate these abbreviations, writing 'BP Model' for BPCM,'Semantic DB' for 'SDS', 'OBP Ontology' for 'OBP' and 'SAGRA DB' for SAGRADA.

## IV. Methodology of Bankruptcy Prediction Computational Model

In Fig. 2 we illustrate a flowchart of a BPCM Model, which is a visual representation of our research methodology.

**Semantic Search Engine.** To enable financial indicators (which can be either qualitative or quantitative) to efficiently detect company's bankruptcy level using an intelligent bankruptcy prediction module, it is necessary to carry out a highly accurate data processing. First, data must be collected from various sources, by means of *Semantic Search Engine (SDS)*. Here, the search is based upon the tehniques that tackle the contextual (semantic) value of the requested information, instead of the vocabulary definitions of individual words or expression as it is found in a traditional search query [33]. The situation is additionally complicated by a possibly rapid change of the data which requires to constantly monitor and check this information for relevance and accuracy.

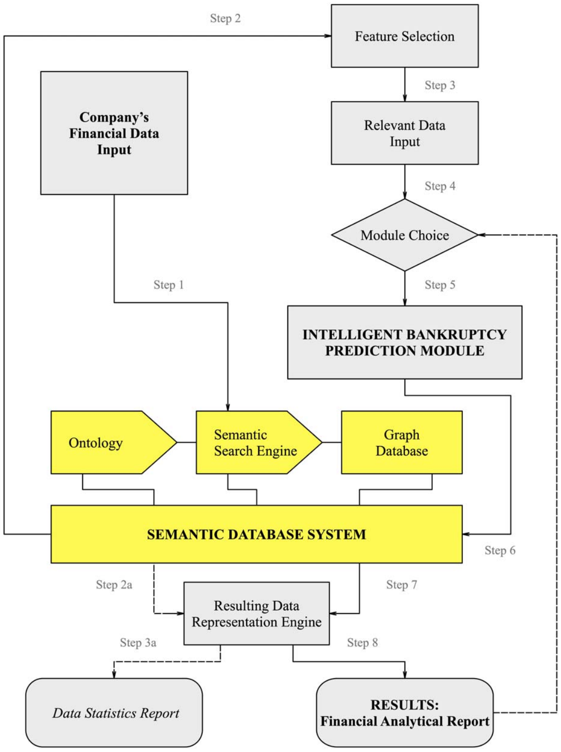

Fig. 2. Bankruptcy Prediction Computational Model flowchart

**Graph Database.** Data gained should be subsequently stored in a convenient form enabling its efficient management including search. As the storage of input and output financial data, instead of the traditional (SQL) database, we use *Graph DB* [34] as a more efficient approach than not only standard DB but also than other so called *NoSQL databases*. It is known that NoSQL databases are more efficient than the standard relational databases allowing to quickly transmit data over the network [35]. However, this is achieved by aggregating and dealing with data of a specific type, and these aggregates are disconnected. Without using the Graph DB overcoming this problem is expensive [36], while a Graph DB handles the connectedness and relationships. Both are obviously necessary features of the financial data.

Indeed, for the financial analysis, it is particularly important to identify the relationships between indicators, and we chose a Graph DB as a tool for financial data pre-processing [36]. Unlike relational databases, where the consideration of relationships intensively reduces the performance of queries on large data sets, the performance of Graph DB remains unchanged with an increase in the amount or variability of the stored data. One of the main reasons for choosing a Graph Database is its ability to significantly boost the processing of interrelated data – in a Graph DB requests are localised in a specific part of the graph. The execution time of each query depends on the size of a part of the graph that needs to be bypassed to satisfy the given request, and not on the total size of the graph [37].

Another useful feature of a Graph DB is its capability to expand. It is possible to easily add new types of interconnections, new nodes, labels and subgraphs to the existing structure, without violating the existing requests and functionality [37]. The data structure should correspond to the changing needs of the financial analysis, and not be imposed in advance and remain unchanged. Due to the flexibility of the graph model, it is not necessary to pre-simulate the task in detail and to take into account each of the potential prospects in advance; the model can be adjusted for a particular company when using BPCM model in real-time.

Finally, Graph DB are equipped with the tools required for the development and system maintenance [36]. In particular, built-in graphical data models, in combination with built-in user-friendly software interface (API) and query language (e.g. Cypher) with clear semantics for each query, allow developing applications effectively.

Later on, to asses financial distress risk, we are aiming to use not only qualitative but also quantitative input data. The tests confirm that graph databases significantly outperform relational databases in dealing with data structure of qualitative nature [38] making this technique a better option for tackling the bankruptcy prediction.

**Ontology.** To ensure a homogeneous, unified presentation of data from various sources we utilise a *Ontology of Bankruptcy Prediction* resembling Financial Industry Business Ontology (FIBO). The ontology, which is based on the semantic approach, helps to extract and integrate information from data repositories, to prepare data for further processing, and to enable communication in natural language.

**Semantic Database System.** The OBP Ontology, the Semantic Search Engine and the SAGRA DB are integrated to form a core of our computational model - the Semantic Database System. The SAGRA DB is a vital part of the model which ensures the efficiency of the communication between other system's components.

**Intelligent Bankruptcy Prediction Module.** An Intelligent Bankruptcy Prediction Module is another core component of our computational model. We envisage exploring here several modern machine learning algorithms that can be picked depending on a particular situation; among them are Fuzzy Set models, Neural Network models and Bayes classifiers. After the data are successfully entered into the system, a feature selection is carried out [39]. This procedure performs a range of genuinely significant indicators from a broad array of information, which is significantly complicated by the presence of the unnecessary (noisy) information. Among the most common modern feature selection methods are Genetic Algorithm (GA), Principal Component Analysis (PCA), T-testing, Backward selection method, Correlation analysis, etc [39]. Next, classification, ordering and standardisation of data are performed. Financial analysis and decision making can be carried out through machine learning algorithms, which significantly improve the accuracy of predictions of the state of the company (see [40], [41], [42], [43]).

## V. Semantic Database System and its Components

### A. Developing a Financial Ontology

In the world of *Semantic Web*, the data can be characterised by a particular structure or meta-information [15]. Such a presentation of data allows creating intelligent semantic information analysis systems. A model in the semantic world is comprised of an ontology and/or a set of taxonomies [44]. By an ontology, we understand a model of knowledge in a particular area (in our case, it is the financial analysis), which promotes the integration of heterogeneous resources at the conceptual level, providing a unified approach to the description of their semantics. One of the advantages of semantic technologies is the opportunity of analysing the *triple graph (traverse)* and define an inference engine. That is, based on the semantics of relations, logical conclusions can be drawn, or the other ties between concepts can be discovered.

The ontology presentation format defines the mechanisms to store concepts and their relationships in the library; it is a method of transmitting ontological descriptions to other consumers and a method of processing its concepts. Specific ontology presentation languages have been developed as ontological description formats. The most famous of these are OWL, RDF, KIF [34].

*Resource Description Framework (RDF)* provides the ability to formulate statements in a form suitable for computer processing. RDF is a metadata description model, describing resources in the form of a directed labelled graph, so each resource has properties, which in turn can also be resources or their combinations [16]. Thus, with the help of RDF, it is possible to describe both the structure of the resource and the related subject area. In this case, a model aims to standardise definitions and the use of metadata that describe Web resources. This language uses XML syntax; however, in contrast to XML, the RDF data model is a graph, which allows defining relationships between entities.

According to [16], RDF standard consists of two main parts – defining resources (†), and schema (‡). The former (†) defines a simple model for describing an object, which is considered as a resource, and links between resources in terms of named properties and values. The latter (‡) describes resources and serves for the task of structuring the subject area. It is similar to a class diagram in UML and is called an RDF Schema [45].

The basic building block in RDF is a triple ”subject (entity) - predicate (attribute) - object (value)” [46]. Such a link can be represented as an edge with the label $P$, which combines two nodes, $S$ and $O$:

$$[S] \xrightarrow{P} [O] \tag{1}$$

An interested reader can found detailed descriptions of the financial ontologies in [15], [17], [44], [47], and [49].

Ontologies are used as data sources for many software applications such as information retrieval, text analysis, knowledge extraction, and other information technologies, allowing more efficient processing of complex and diverse information. This way of representing knowledge enables applications to recognise those semantic differences that are obvious to people but not known to the computer [44].

There are initiatives in many businesses, to create industrial ontologies. Such common ontologies can serve as a basis for data exchange contracts between industry companies. In the financial world, the most well-known ontology at the moment is FIBO, which is developed by the Enterprise Data Management (EDM) Council [50]. A number of financial systems already support it, which allows businesses to exchange data corresponding to FIBO, in the semantic format [48].

The developers of FIBO call it a 'Rosetta stone' of finance [50], as it defines semantic relationships between various financial concepts, namely financial instruments and their interconnections, as well as their relationships with issuers [51]. Moreover, FIBO includes a set of basic legal, contractual and organisational concepts.

In fact, FIBO is not a single ontology, but a collection of a large number of ontologies, divided into modules and submodules [48]. The separate large modules include Financial Business and Commerce (FBC), Business Entities (BE), Securities (SEC), Indices and Indicators (IND), etc.

The basis of all FIBO ontologies is a top-level abstraction ontology called FIBO Foundations (FND) including the other sections, which in turn comprise a description of various types of basic entities. For example, one of the sections of FND is Accounting, which is a set of ontologies regarding general accounting concepts. FIBO is freely available at the EDM Council website [50].

The main and most crucial component of the financial risk management of a company is the knowledge base. Our approach to building an ontology describes the basic concepts of financial analysis, as well as the objects that serve as sources of knowledge for predicting a company's bankruptcy. It also contains the concepts and relationships required for the formation of a hierarchy of knowledge fields and the subsequent use of this hierarchy by various applications (in our case, SAGRA Database and BPCM model). In addition, expert rules and regulations can be described in terms of ontology, which significantly increases their level of succinctness and transparency for the users.

*OBP Ontology* developed for the BPCM model project is presented in Fig. 3. The structure and the content of the OBP Ontology are based on the experience of analysts specialising in the theory and practice of bankruptcy prediction [52]. This hierarchy reflects a number of the most popular indicators used to conduct a financial analysis of a company, as well as their origin (documents and concepts to which they relate) and the relationship of these indicators to each other. Financial analytic factors form the penultimate row of the hierarchy, while the principal generalising object is the concept of Companys Financial Records. The last row in the hierarchy contains linguistic variables that will be later involved in the development of machine learning computational modules.

Below we will provide an example of one of the indicators, which will give more insight into the OBP Ontology as a good illustration of a semantic search approach.

*Return on Equity (ROE)* is responsible for the company's productivity (linked to Productivity language variable) and is formed from the Income statement and Balance Sheet data. ROE is a crucial indicator for business owners. It allows determining how effectively the capital invested in the business was used [53].

The graph shows that this ratio depends on the following two indicators - 'Profit After Interest and Tax' and 'Shareholders' Funds'. In turn, Profit After Interest and Tax are two leading indicators of the Income Statement, which is formed after Interest and Tax (not shown in the graph) are deducted from Gross Profit. For the Shareholders' Funds or Equity, the liabilities indicator refers to a companys Balance Sheet.

Unlike a similar ratio Return on Actives (not presented in the scheme, as it is rarely used for bankruptcy prediction models), ROE defines the efficiency of using not all the company's capital, but only that part of it which belongs to the business owners. At the same time, ROE is indirectly related to *Gearing* (measures financial risk), since the formation of these two ratios requires Shareholders' Funds indicator. In practice, this relationship between two ratios can be traced in the well-known model – *Du Pont formula* [53] – the higher the return on equity, the better. Nevertheless, a high value of the ratio may be a consequence of the value of Gearing being too high, i.e. a large proportion of borrowed capital and a small share of its own, which negatively affects the financial stability of the business. It reflects the primary law of business – more profit, more hazard.

The working version of the OBP Ontology, given in this study, is an informal conceptual representation model, which is an initial step of the proposed approach.[1]

### B. Developing a Graph Database

*Graph DB* (for instance, Neo4J) are an example of *NoSQL databases* aimed at representing semantical data [35]. Graph databases are used for storing, processing and automated visualisation of standard structural elements. A typical Graph DB usually contains some reference information regarding objects [36]. Therefore, the user/designer does not have to spend time searching fo this information in the DB directories. It also reduces the number of possible human factor related errors. Graph DB enables to create standard elements automatically, which significantly reduces the design time [37].

As a rule, Graph DB contains two main components:

1 a set of parametric programs that create the necessary images (usually in the dialogue mode);

2 a set of batch files in which all reference and auxiliary information on the drawn elements are stored [36].

[1] At the moment, our work concerns with supplementing the structure of this ontology, as well as with the development of its formal physical representation model utilising the OWL/RDF environment.

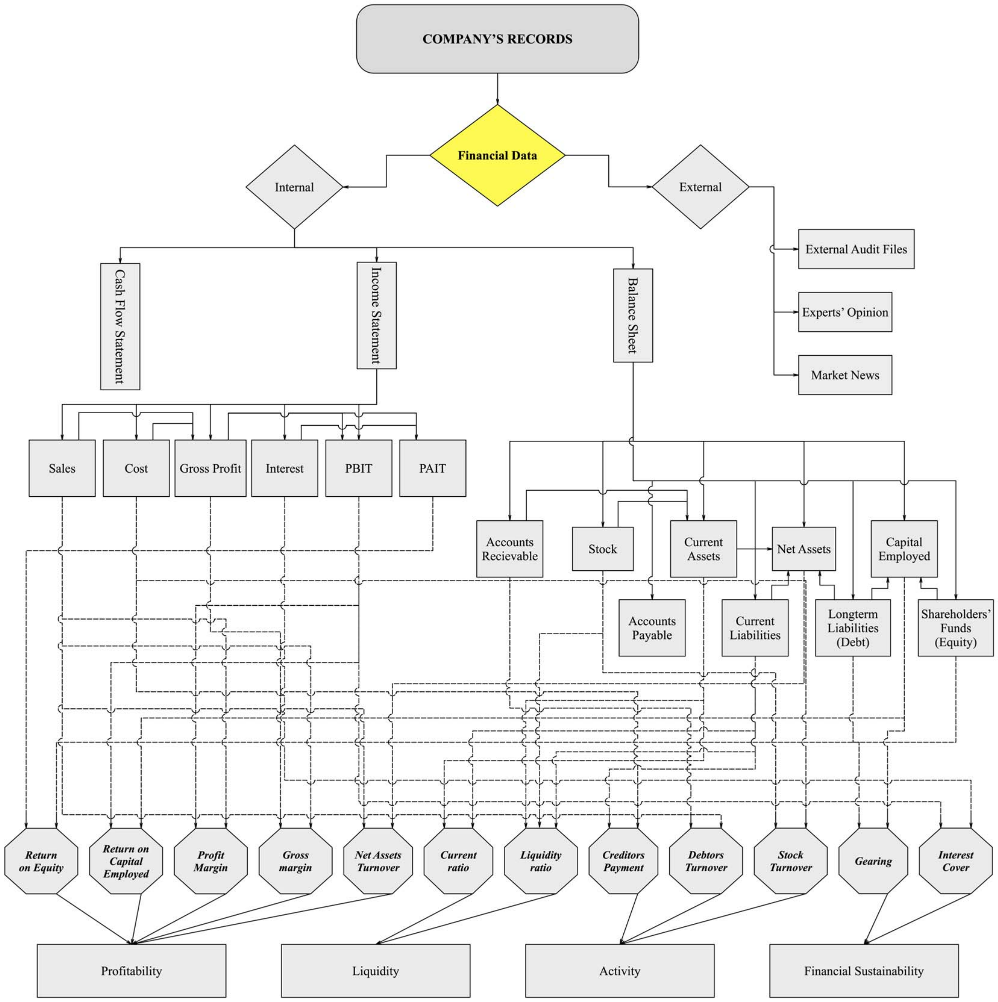

Fig. 3. Ontology of Bankruptcy Prediction

The most common applications for creating Graph DB are *Neo4j*, *DEX*, *Titan* and *OrientDB*, which are very similar in use. However, according to the empirical evaluation of these four applications using a graph database benchmark tool on different types of workloads, it was revealed that Neo4j is preferable in many respects [18]. Also, an experimental comparison of these applications was conducted using BlueBench Architecture which assesses the systems on some operations such as CreateIndexes, LoadGraphML, Traversal. ShortestPath, etc [19]. This test had revealed that Neo4j, followed by DEX and Titan, outperformed other systems because of specialisation of their backbends for exactly this type of queries.

*Neo4j*[2] is an open source Graph Database management system implemented in Java. Its developer is Neo Technology [31]. This Graph DB environment stores data in a proprietary format specifically adapted for the presentation of graph information; this approach, in comparison with the modelling of a graph database, using a relational *Databases Management Systems (DBMS)*, allows for additional optimisation in the case of data with a more complex structure. Neo4j uses its own query language, Cypher[3], though the queries can be done in other ways, for example, directly through the Java API [32]. Cypher is not only a query language but also a data manipulation language, as it provides CRUD functions for graph storage.

Although as an open source Neo4j found many applications in industrial implementation of Graph DB [54], we are not aware of any projects that directly utilise this framework in the financial analysis of a company. At the same time, the developments in *Enterprise Content Management with Neo4j* [55] and *Project Management* [56] clearly reveal how

[2] https://neo4j.com/product/

[3] https://neo4j.com/developer/cypher/

the most popular Graph DB queries are constructed.

We emphasise that the OBP Ontology structure is an excellent basis for the Semantic Analysis Graph Database which is used as a repository of the financial data for BPCM model. So, we intend to apply an existing solution of creating and managing Graph Databases and integrate it into our novel approach.

We have implemented a prototype Graph DB, a SAGRA DB, in Neo4j. The basic concepts in a Graph DB are nodes (an object of the database), relations (graph edges) and their properties. In our case, the nodes of the graph are financial ratios, financial indicators, and the documents containing them. Our graphical repository has 29 nodes divided into three categories – Ratio, Criteria (financial indicator), Statement, and 52 relationships between them (of two types – direct and inverse). Besides, 85 properties were set.

The graph representation of this setup is shown in Fig. 4. The steps of building a graph model can be considered from tracking the dependence Return on Equity (ROE) ratio, which was described above. The first step is to create the nodes of all the indicators involved. ROE is calculated by dividing ”Profit After Interest and Tax” by ”Shareholders' Funds”. Thus, we need to build three nodes:

```
N1 Create (rp1:Ratio {ratioID:  "Return on
   Equity", normative_value:"1",
   linguistic_variable: "Productivity"})
N2 Create (i6:Criteria criteriaId: "PAIT",
   normative_value:">0", year:"2018")
N3 Create (b9:Criteria criteriaId:
   "Shareholders Funds", year:"2018",
   type:"Liabilities")
```

Then, the related statements ”Balance Sheet” and ”Income Statement” should be created:

```
S1 Create (s1:Statement statementID:"Balance
   Sheet", year:"2018", tagline:'A statement
   of the assets, liabilities, and capital
   of a business or other organisation at a
   particular point in time, detailing the
   balance of income and expenditure over the
   preceding period.')
S2 Create (s2:Statement statementID:"Income
   Statement", year:"2018", tagline:'The
   statement displays the companys revenue,
   costs, gross profit, selling and
   administrative expenses, other expenses
   and income, taxes paid, and net profit, in
   a coherent and logical manner.')
```

Additionally, we define the relation between the ratios:

```
R1 Create (rp1)-[:directly_related_to] →(i6)
R2 Create (rp1)-[: inversely_related_to] →(b9)
```

Next, we define types of indicators the statements contain:

```
I1 Create (i6)-[:directly_related_to] →(s2)
I2 Create (b9)-[:directly_related_to] →(s1)
```

Finally, the result can be shown:

```
MATCH (n) RETURN n;
```

Neo4j also allows fulfilling complicated queries. For example, to choose Ratios suitable for criterion Shareholders Funds:

```
MATCH (a:Ratio)⟶(b:Criteria
criteriaId:"Shareholders Funds") RETURN a;
```

As a result, the system will show one Ratio – Return on Equity.

**Intermediate Evaluation.** The proposed Semantic Database System reflects the 'big four V + R' (see Section III). Notably, as we can see from the results, the SAGRA DB takes good care of the Variety (as well as heterogeneity), Velocity and Relationships of Financial Data.

## VI. Conclusions and Future Work

**Conclusions.** The advantage of bankruptcy prediction models is the simplicity in the interpretation of results, as well as an accurate, and less time-consuming, estimation of the probability of bankruptcy. Intelligent financial analysis systems can be used by entrepreneurs (to test the practical feasibility and cost-effectiveness of their ideas); managers of the companies (to assess the impact of strategic and operational decisions on the financial performance of the firm); third-party investors (to assess the effectiveness and risks of the project and decide on participation in it), as well as creditors (to decide on the provision of borrowed funds).

The contributions of the paper are the following.

1) Based on the analysis of various modern approaches to the processing and storage of the heterogeneous data related to the financial analysis, we proposed a novel intelligent methodology to construct a Bankruptcy Prediction Computational Model. Our methodology is based upon the utilisation and integration of the semantic data management methods.
2) Following this methodology, we have introduced a novel layered architecture for this Computational Model (see Fig. 2), which integrates the Semantic Database System and a set of modern machine learning algorithms.
3) The Semantic Database System is, in turn, a novel development, which comprises the Ontology of Bankruptcy Prediction (OBP), the Semantic Search Engine (SDS) and the Semantic Analysis Graph Database (SAGRADA).
4) We have implemented the principles of the new Ontology of Bankruptcy Prediction and the Semantic Analysis Graph Database on the example of a company financial record.
5) A roadmap for the implementation of the Semantic Database System was established: the informal conceptual representation of the OBP Ontology (shown in Fig. 1) was designed and described; the code of SAGRA DB for the BP model was built using Cypher query language and Neo4j environment (the resulting graph is presented in Fig. 4).

The developed BPCM methodology allows mot only to process a substantially larger amount of data than conventional statistical methods but also potentially a complex qualtative dataset.

In the future we envisage to utilise the machine learning approach which significantly reduces the processing time, increases the accuracy of the result, and eliminates human errors. Here all data necessary for the analysis are searched

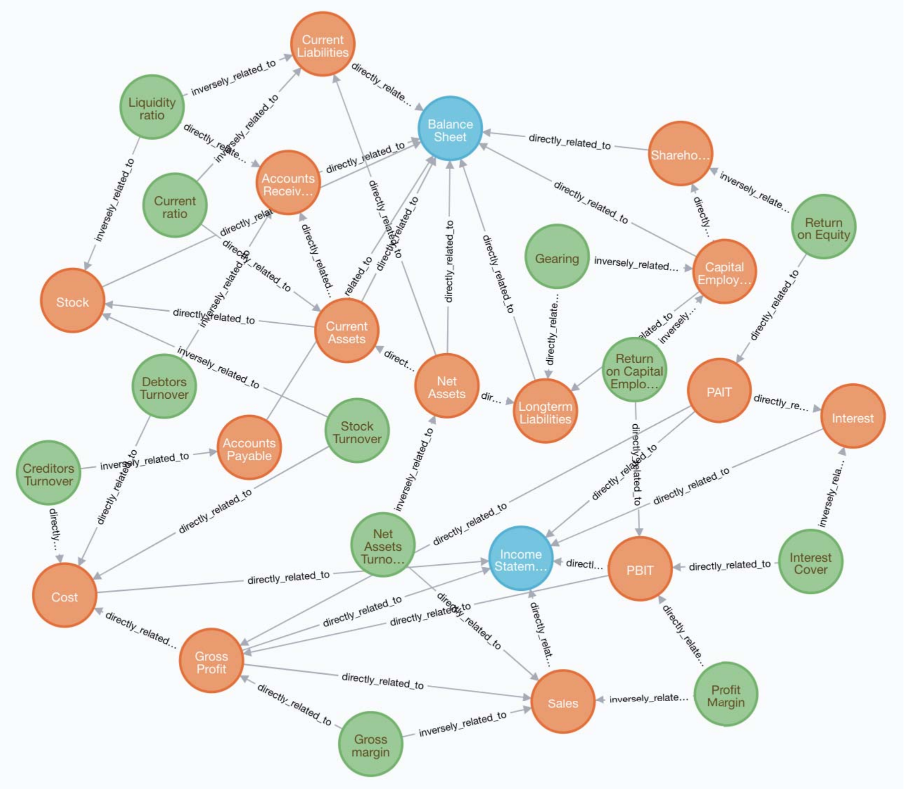

Fig. 4. OBP Ontology captured in SAGRA DB.

and selected automatically. The only way the human factor intervenes in the process is the decision making on the choice of the computation method (which leads to different kinds of results), or the type of the final report. This allows both expert and non-expert (for example, company executives) users to work with the system.

Addressing the issue of the final analytical report of the system, we propose a methodology which will embed machine learning to answer in detail why a company is experiencing a particular situation, what factors influence it, and which of them needs to be paid attention to prevent the company from bankruptcy in time. The developments described in this paper will 'pre-process' data for the input to Machine Learning layer. The Graph DB representation will enable easier and more efficient search of relevant information. Besides, the text of the analysis will be supported by tables and graphs convenient for perception. Standard statistical methods, as well as separately used methods of machine learning, do not allow a full conclusion to be drawn about a problem to a non-expert, they can only answer the bankruptcy question unequivocally.

**Future Work.** First, we will create a Semantic Search Engine, one of the unsolved tasks discussed in Section III (and identified in 'red' in Fig. 1). This engine will automatically explore data from various sources including the Internet, and classify all information relevant for the financial analysis of a company (according to OBP Ontology) and prepares it for the input into the SAGRA Database. We will also need to solve a problem in finding a way in which Semantic Database System would support all possible dynamic data formats.

Second, we will improve the structure of the OBP Ontology creating its formal conceptual representation through OWL / RDF languages (similar to FIBO). We will also work on further enhancement of the SAGRA DB itself.

Finally, we will tackle a problem of the data exchange between the structural parts of the Semantic DB finding a way to transfer data in various directions automatically.

## VII. Acknowledgment

We are grateful to *Dr Fang He*, *Dr Gabriele Pierantoni*, *Prof Andrei Kuznetsov* and to the anonymous reviewers for their useful comments and suggestions to improve the paper.

## References

[1] Y. Wu, C. Gaunt, and S. Gray, ”A comparison of alternative bankruptcy prediction models.” Journal of Contemporary Accounting & Economics 6(1), 2010, pp. 34–45.

[2] W.H. Beaver, ”Financial ratios as predictors of failure.” Journal of accounting research, 1966, pp. 71–111.

[3] E.I. Altman, ”Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.” The journal of finance 23(4), 1968, pp. 589–609.

[4] R.J. Taffler, ”The assessment of company solvency and performance using a statistical model.” Accounting and Business Research 13(52), 1983, pp. 295–308.

[5] J.A. Ohlson, ”Financial ratios and the probabilistic prediction of bankruptcy.” Journal of accounting research, 1980, pp. 109–131.

[6] L.A. Zadeh, ”Fuzzy sets as a basis for a theory of possibility.” Fuzzy sets and systems 1(1), 1978, pp. 3–28.

[7] J.J. Buckley, ”Solving fuzzy equations in economics and finance.” Fuzzy Sets and Systems 48(3), 1992, pp. 289–296.

[8] A.I. Dimitras, R. Slowinski, R. Susmaga, and C. Zopounidis, ”Business failure prediction using rough sets.” European Journal of Operational Research 114(2), 1999, pp. 263–280.

[9] C. Zopounidis, ”Multicriteria decision aid in financial management.” European Journal of Operational Research 119(2), 1999, pp. 404–415.

[10] W.S. McCulloch, and W. Pitts, ”A logical calculus of the ideas immanent in nervous activity.” The bulletin of mathematical biophysics 5(4), 1943, pp. 115–133.

[11] M. Tk, and R. Verner, ”Artificial neural networks in business: Two decades of research.” Applied Soft Computing, 38, 2016, pp. 788–804.

[12] P.K. Coats, and L.F. Fant, ”Recognizing financial distress patterns using a neural network tool.” Financial management, 1993, pp. 142–155.

[13] S. Lee, and W.S. Choi. ”A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis.” Expert Systems with Applications 40(8), 2013, pp. 2941–2946.

[14] T.R. Gruber, ”Toward principles for the design of ontologies used for knowledge sharing?.” International journal of human-computer studies 43(5-6), 1995, pp. 907–928.

[15] P. Castells, B. Foncillas, R. Lara, M. Rico, and J. L. Alonso, ”Semantic web technologies for economic and financial information management.” European Semantic Web Symposium, pp. 473–487. Springer, Berlin, Heidelberg, 2004.

[16] B. L. Grant and M. Soto, ”Topic maps, RDF Graphs, and ontologies visualization”, in: Visualizing the Semantic Web. XML-based Internet and information visualization, second edition, V. Geroimenko, C. Chen Eds., London: Springer-Verlag, 2010, pp. 59–79.

[17] H. Tang, and L. Song, ”Ontologies in financial services: Design and applications.” International Conference on Business Management and Electronic Information, 5, pp. 364–367. IEEE, 2011.

[18] S. Jouili, and V. Vansteenberghe, ”An empirical comparison of graph databases.” International Conference on Social Computing, pp. 708–715. IEEE, 2013.

[19] V. Kolomienko, M. Svoboda, and I. Holubova, ”Experimental comparison of graph databases.” Proceedings of International Conference on Information Integration and Web-based Applications & Services, p. 115. ACM, 2013.

[20] H. R. Vyawahare, , P. P. Karde, and V. M. Thakare, ”A Hybrid Database Approach Using Graph and Relational Database.” International Conference on Research in Intelligent and Computing in Engineering (RICE), pp. 1–4. IEEE, 2018.

[21] R. De Virgilio, A. Maccioni, and R. Torlone, ”Model-driven design of graph databases.” International Conference on Conceptual Modeling, pp. 172–185. Springer, Cham, 2014.

[22] S. lvarez-Garca, B. Freire, S. Ladra, and . Pedreira, ”Compact and efficient representation of general graph databases.” Knowledge and Information Systems, 2018, pp. 1–32.

[23] X. Wang, and Y. He, ”Learning from uncertainty for big data: Future analytical challenges and strategies.” IEEE Systems, Man, and Cybernetics Magazine 2(2), 2016, pp. 26–31.

[24] G.P. Mansukhlal, and C. Malathy, ”Semantic integration with ontology based approach.” International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 2257–2260. IEEE, 2016.

[25] M. Gagnon, ”Ontology-based integration of data sources.” 10th International Conference on Information Fusion, pp. 1–8. IEEE, 2007.

[26] X. Liu, N. Iftikhar, and X. Xie, ”Survey of real-time processing systems for big data.” 18th International Database Engineering & Applications Symposium, pp. 356–361. ACM, 2014.

[27] S. Shrivastava, and S.N. Pal, ”Graph mining framework for finding and visualizing substructures using graph database.” International Conference on Advances in Social Network Analysis and Mining, pp. 379–380. IEEE, 2009.

[28] A. Hogan, ”Linked Data & the Semantic Web Standards.”, 2014, pp. 3–48.

[29] R.C. McColl, D. Ediger, J. Poovey, D. Campbell, and David A. Bader, ”A performance evaluation of open source graph databases.” Proceedings of the first workshop on Parallel programming for analytics applications, pp. 11–18. ACM, 2014.

[30] EDM Council, ”About FIBO”, 2019. Availiable at: https://edmcouncil.org/page/aboutfiboreview (Accessed: 15 March, 2019).

[31] neo4j, ”The Internet-Scale Graph Platform”, 2019. Availiable at: https://neo4j.com/product/ (Accessed: 15 March, 2019).

[32] neo4j, ”Cypher Query Language”, 2019. Availiable at: https://neo4j.com/developer/cypher/ (Accessed: 15 March, 2019).

[33] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv, ”XSEarch: A semantic search engine for XML.” 29th international conference on Very large data bases, 29, pp. 45–56. VLDB Endowment, 2003.

[34] R. Angles, and C. Gutierrez, ”Survey of graph database models.” ACM Computing Surveys (CSUR) 40(1), 2008.

[35] V.N. Gudivada, D. Rao, and V.V. Raghavan. ”NoSQL systems for big data management.” IEEE World congress on services, pp. 190–197, 2014.

[36] Z.J. Zhang, ”Graph Databases for Knowledge Management.” IT Professional 19(6), pp. 26–32, 2017.

[37] J. Pokorn, ”Graph databases: their power and limitations.” IFIP International Conference on Computer Information Systems and Industrial Management, pp. 58–69. Springer, Cham, 2015.

[38] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, and D. Wilkins. ”A comparison of a graph database and a relational database: a data provenance perspective.” 48th annual Southeast regional conference, p. 42. ACM, 2010.

[39] C.F. Tsai, ”Feature selection in bankruptcy prediction.” Knowledge-Based Systems, 22(2), 2009, pp. 120–127.

[40] K. Nagaraj, and A. Sridhar, ”A predictive system for detection of bankruptcy using machine learning techniques.” arXiv preprint arXiv:1502.03601, 2015.

[41] T. Korol, ”Fuzzy logic in financial management.” Fuzzy logic-emerging technologies and applications. InTech, 2012.

[42] H. Adeli, and S.L. Hung, ”Machine learning: neural networks, genetic algorithms, and fuzzy systems.” John Wiley & Sons, Inc., 1994.

[43] F. Barboza, H. Kimura, and E.I. Altman, ”Machine learning models and bankruptcy prediction.” Expert Systems with Applications, 83, 2017, pp. 405–417.

[44] H. Dudycz, and J. Korczak, ”Conceptual design of financial ontology.” Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1505–1511. IEEE, 2015.

[45] H. Chen, Z.Wu, H.Wang, and Y. Mao, ”RDF/RDFS-based relational database integration.” 22nd International Conference on Data Engineering (ICDE’06), pp. 94–94. IEEE, 2006.

[46] F. Antoniazzi, and F. Viola, ”RDF Graph Visualization Tools: a Survey.” 23rd Conference of Open Innovations Association FRUCT, 2018, p. 4.

[47] Q. Quboa, and N. Mehandjiev, ”Creating intelligent business systems by utilising big data and semantics.” 19th Conference on Business Informatics (CBI), 2, pp. 39–46. IEEE, 2017.

[48] G.G. Petrova, A.F. Tuzovsky, and N.V. Aksenova, ”Application of the Financial Industry Business Ontology (FIBO) for development of a financial organization ontology.” Journal of Physics: Conference Series, 803(1). IOP Publishing, 2017.

[49] K. Perera, , D. D. Karunarathne, A. Siriwardena, and D. Balaretnaraja, ”Ontology based annotation mechanism for financial documents.” International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 110–117. IEEE, 2013.

[50] EDM Council, ”FIBO Ontology file directory”, 2019. Availiable at: https://spec.edmcouncil.org/fibo/ontology/master/2018Q2.5/tree.html (Accessed: 15 March, 2019).

[51] L. Makni, N. Zaaboub, and H. Ben-Abdallah, ”Reuse of Semantic Business Process Patterns.” 9th International Conference on Software Engineering and Applications (ICSOFT-EA), pp. 36–47. IEEE, 2014.

[52] J.L. Gissel, D. Giacomino, M.D. Akers, ”A Review of Bankruptcy Prediction Studies: 1930-Present.” Journal of Financial Education, 33, 2007, pp. 1–42.

[53] W.W. Wu, ”Beyond business failure prediction.” Expert systems with applications 37(3), 2010, pp. 2371–2376.

[54] neo4j, ”Neo4j GraphGists”, 2019. Availiable at: https://neo4j.com/graphgists/ (Accessed: 15 March, 2019).

[55] P.-J. Van Aeken, ”Enterprise Content Management with Neo4j”, 2019. Availiable at: https://neo4j.com/graphgist/enterprise-content-management-with-neo4j (Accessed: 15 March, 2019).

[56] N. White, ”Project Management”, 2019. Availiable at: https://neo4j.com/graphgist/project-management (Accessed: 15 March, 2019).