Analysis of Uses and Limitations of Mechanism of Extraction of XBRL Data in Financial Research

IGOR PUSTYLNICK, OKSANA TEMCHENKO, SERGEY GUBARKOV
School of Business
Far Eastern Federal University
Vladivostok
RUSSIA
e-mail: ipustylnick@conestogac.on.ca

Abstract—Since Extensible Business Reporting Language (XBRL) has become a language of mandatory statements submission in the U.S., it opened new opportunities for electronic parsing and auditing of these statements. We show the example of automatic extraction of the statement data and propose the improvements to the taxonomy based on Generally Acceptable Accounting Principles (GAAP) of the USA, which may be able to improve the extraction process.

Key-words: XBRL Parsing; financial statements; data extraction; data verification; XBRL data .

1. Introduction

Ever since XBRL filings have become mandatory in the United States, there were significant efforts made to produce the evidence of the viability of XBRL based financial data in auditing and financial research. The research and practical efforts reported in this paper were greatly influenced by very encouraging results, summarized in [1]. The authors showed that it was possible not only to read the report the data, but also use it in the financial analysis and auditing of the reported data, the capabilities announced earlier in [2].

Despite the potential arguments, the assumption can be made that the data presented in the digital way in the XBRL format is the real data submitted by the companies, reporting their financial results to the U.S. SEC and by proxy to the public at large. Being extracted, this data may be superior to the data, produced and exported by the financial data aggregators [3]-[5], such as Yahoo, COMPUSTAT and others.

Despite the initial expectations of the ease of extraction process from XBRL based financial documents, the further review of the results of such extraction followed. The researchers noted a relatively low quality of the data presented in the financial reports [6], as well as the excessive use of the elements (tags) of XBRL documents, defined by the companies for their own use (called ‘extensions’ in the U.S. XBRL standards), which significantly reduced the ability of the extracting algorithms to determine their accounting meaning [7].

The research efforts described in this paper were undertaken in order to obtain a significantly large set of financial data, which could be used in the accounting/financial research, representing a large random sample of the unfiltered financial data. It was expected that by using this data in the financial research in the specific areas of accounting and audit, it would be possible to obtain the results, similar to the ones, that would have been obtained with the use of the COMPUSTAT data feed.

The goal of this paper was not to produce a novelty algorithm that does not exist or not yet in use in the computing community. Here we tried to establish the limits of the use of conventional tools, such as Object Oriented Programming and XML document parsing, in the scope of use with XBRL.

At the moment the wealth of XBRL data existing in the U.S. SEC repository allows making conclusions on the quality of XBRL documents. The goal of this research was to show how much of such data can be converted into the data stream, similar to the ones provided by aggregators and how much of this data can be used in the particular research effort.

For rest of this paper: Section 2 will present the algorithm of extracting data from the XBRL based financial statements obtained from EDGAR database of U.S. Security and Exchange Commission (SEC). In Section 3 we will discuss the research method and the problems of such extraction process and we will especially point to the places in the U.S. XBRL standard, which deter the successful extraction of data. In Section 4 we will also show our method of data verification. Section 5 will elaborate on the conclusions we were able to make based on such evaluation, the potential limitations of this research and the directions, in which it can be extended.

2. Past Research

The research on the use of XBRL can be divided into two large parts: the initial efforts to describe the language and to chart the way of its use [2][8] and the evaluation of the statements submitted in XBRL format past 2011 when such submissions became mandatory in the USA [9]. At that time, the authors used the standard of XBRL as guidance and described the potential of XBRL in the area of financial reporting [10] and auditing of financial statements [11]. Later, the authors started to talk about the internal assurance of the XBRL based financial statements, mainly from the perspective of internal XBRL assurance [12].

The research, based on the existing XBRL data, submitted to the U.S. SEC concentrated on the specific directions of use of XBRL based documents, which became possible when the EDGAR database amassed a large number of statements. The researches in [13] discuss the issue of the information asymmetry in the large early adopters of XBRL and find that such asymmetry has decreased, which resulted in the larger trading volumes.

The researchers examining a large number of the U.S. filings [7] find that many companies use extensions to the XBRL standard schema (taxonomy). Such extensions may significantly impact the extraction of financial data from XBRL based financial documents [14]. In contrast with the U.S. filing standard, the European standard does not allow extensions to the country-based XBRL taxonomy, which allowed researchers and financial analysts extracting better quality data from the XBRL based statements [15].

There are various algorithms of extraction from XBRL based documents, which are presented in the recent years. The researchers in [16] discuss the possibilities of using data mining in order to detect fraud in the financial statements. The ontology approach to the data extraction is shown in [17] and it allowed the researchers to make conclusions over the financial aspect of the presented data. Overall, there has been a number of valuable efforts in extending the mechanisms of extraction of the financial data. However, the majority of these efforts are dealing with the algorithmic aspect of extraction avoiding dealing with the financial or auditing direction of data.

In the recent years, there is a number of papers demonstrating the novel extraction algorithms [18] or novel approach to fuzzy logic around building queries, which are designed to work against XBRL documents [19]. Such algorithms take into consideration the announced taxonomy of the XBRL and do not work with the data at hand.

Many researchers on the subject take an approach similar to [20] attempting to use semantic web approach to XBRL queries. While this approach is quite valid, it is also based on the assumption that XBRL documents are valid and tend to adhere well to the underlying taxonomy. It will be shown further in this research that this approach cannot be used to the full extent with the real XBRL files on hand.

The research described in [21] represents an approach, which is very similar to the one used in [6] and the one, we were using in our research. It was, however, performed on the data for the closed Italian XBRL taxonomy, where all elements are fixed and cannot be extended by the companies creating and submitting their XBRL based financial statements.

Further in this paper, we will present and analyze the algorithm of extracting XBRL based financial data from the U.S. XBRL taxonomy, which allows every company defining an infinite number of extensions to the standard elements defined in the taxonomy. At the beginning of the research we formulated two questions we wanted to get answers for: (1) is it possible to extract data from the XBRL based financial reports by using fully automated algorithm; (2) how much of this data can be used and what is the quality of the extracted data.

3. Research Method and Algorithm

Description of any mechanism of extracting the data from a particular data source must include two required elements: the data, which needs to be extracted and the algorithm, used in the extraction process. Prior to engaging in the extraction effort, we also chose the mechanism of verification of the data, which was extracted. Since the original intent was to use this data in the research on earnings management, the data and the mechanism of verification came from this area.

3.1. Data Extraction

Based on research, presented in [22], the following financial variables to be extracted are compiled in Table 1.

TABLE 1. VARIABLES TO BE EXTRACTED FROM XBRL

Balance Sheet	**Statement of Income**
Assets	Revenue
Current Assets	Cost of Goods Sold
Accounts Receivable	Operating Expenses
Property Plant & Equipment	Net Income
Depreciation	EBIT*
Current Liabilities	
Current Portion of Debt	
Long Term Debt	
Total Value of Shares	
Retained Earnings	

*EBIT is not a part of the official statement but XBRL taxonomy places it in Income Statement tree

All variables are standard accounting variables, which can be found in the financial statements, presented in any accounting textbook. However, the companies, presenting their financial statements in XBRL format to the U.S. SEC belong to various sectors of economy where the same financial variables can be presented in the variety of ways. U.S. XBRL Taxonomy presents another large challenge. The composition of the taxonomy presents all financial statements as data trees. Even if the data is not presented for the variable from Table I, it can be calculated from all values of child variables, presented in the statement. Therefore, the algorithm of extraction becomes an algorithm of traversing the data tree to find all children of the variable under review.

At the early stages of the research it was discovered that the extension elements defined in the calculation linkbase of the particular company have no connection with the elements of US GAAP taxonomy. The calculation linkbase of the company may or may not define the calculation rules, which tie the extension elements with the elements of the US GAAP taxonomy. Therefore, the authors made a decision to use only US GAAP taxonomy elements in the calculations of the values of the fields used in the research.

For each of the fields defined in Table I the tree traversing is presented in Figure 1. The tree was traversed depth-first and for each tree node on each level the tag was compared with the set of tags retrieved on the presentation level. If tag was found, the branch of the tree was sealed and the value was added to the total value corresponding to the field under review. The algorithm continued traversing the next sibling on the same tree level.

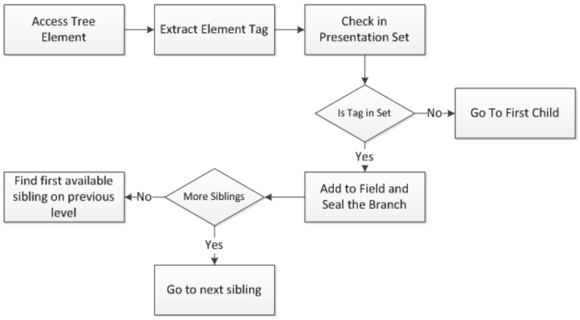

Figure 1. Algorithm of Data Extraction from XBRL

3.2. Data Verification

The initial verification of data and the algorithm was performed by comparison of the data, appearing on U.S. SEC website and the data extracted by using the described algorithm. Any failure (and subsequent adjustment) of the algorithm was performed by using the data from the same statement, which appeared to break the process of calculations. The spot checks were also performed when the data was successfully extracted using the automated process.

Since the data was retrieved from the financial statements in the 'blind' manner, the verification of such data becomes very important. For the purpose of this research the verification consisted of two stages: (1) the data was checked for the existence of information, i.e., the data in the fields contained values other than zero; (2) the data is suitable for the calculations and further verification.

For the verification of data, we used two formulas, which are often used in fraud detection efforts, namely the formula for Z-Score by Altman [23] and the formula for the calculation of M-Score by Beneish [24].

The accounting variables used in the calculation of these indicators were collected into a single data set. Z-Score indicator was calculated using original formula.

$$Z = 1.2X1 + 1.4X2 + 3.3X3 + 0.6X4 + 1.0X5 \tag{1}$$

$$X1 = \frac{WorkingCapital}{TotalAssets} \tag{2}$$

$$X2 = \frac{\mathrm{Re}\,tainedEarnings}{TotalAssets} \qquad (3)$$

$$X3 = \frac{EBIT}{TotalAssets} \qquad (4)$$

$$X4 = \frac{MarketValueOfEquity}{BookValueOfDebt} \qquad (5)$$

$$X5 = \frac{\mathrm{Re}\,venue}{TotalAssets} \qquad (6)$$

M-Score was also calculated using original formula, provided by M. Beneish.

$$\begin{aligned} M &= -4.84 + 0.920 * DSRI + 0.528 * GMI + \\ &0.404 * AQI + 0.892 * SGI + 0.115 * DEPI - \\ &0.172 * SGAI + 4.679 * TATAI - 0.327 * LEVI \end{aligned} \qquad (7)$$

The indicator variables, included in (7) are the financial ratios extensively used in accounting practice and, as such, are well known to the accountants and financial analysts.

$$DSRI = \frac{Accounts\,\mathrm{Re}\,ceivable}{\mathrm{Re}\,venue} * 365 \qquad (8)$$

$$GMI = \frac{\mathrm{Re}\,venue - COGS}{\mathrm{Re}\,venue} \qquad (9)$$

$$AQI = \frac{Assets - PP\,\&\,E}{Assets} \qquad (10)$$

$$SGI = \mathrm{Re}\,venue \qquad (11)$$

$$DEPI = \frac{Depreciation}{Depreciation + PP\,\&\,E} \qquad (12)$$

$$SGAI = \frac{OperatingExpenses}{\mathrm{Re}\,venue} \qquad (13)$$

$$TATAI = \frac{WorkingCapital - Depreciation}{Assets} \qquad (14)$$

$$LEVI = \frac{LongTermDebt + CurrentLiabilities}{Assets} \qquad (15)$$

Beneish M-Score formula uses the ratios of the values of the indicators, defined by (8) - (15), obtained in the adjacent years.

Both research papers by Altman and Beneish, mentioned earlier, give certain threshold of the values of Z-Score and M-Score respectively, which were recorded for the companies with various degrees of financial health. In order to assess the behavior of the mentioned indicators, we introduced two other samples: one assembled from the companies, which were engaged in fraudulent revenue manipulations, also used in [22] and the sample of companies with exceptional liquidity, presented in [25]. The statistical comparison of the extracted data with the data from other two samples must indicate how the sample of financial data extracted from XBRL must be treated.

4. Results and Discussion

The results of this research are tailored to the needs, which were specified earlier. The authors of this paper fully understand that they are fully dependent on the choices made. However, it is possible to say that the set of variables is significantly diverse and can be used in the variety of research efforts. Table II presents the percentages the usable data for each variable. At this point, only the data that was not present, was deemed unusable. The sample of data collected for all years of mandatory submission contained over 20,000 entries. It appeared to be feasible to break the data into sub-samples for the years 2011-2015.

TABLE 2. PERCENTAGES OF USABLE DATA PER VARIABLE AND YEAR

Variable	Total	2011	2012	2013	2014	2015
Assets	92.74%	92.37%	92.34%	92.56%	93.63%	92.68%
Curr. Ass.	93.91%	93.75%	93.58%	94.02%	94.30%	93.74%
Cash	93.54%	93.19%	93.32%	93.55%	94.10%	93.51%
Acc. Rec.	60.96%	59.35%	61.28%	59.95%	63.28%	60.68%

Variable	Total	2011	2012	2013	2014	2015
Cap. Ass.	81.08%	81.26%	80.37%	79.82%	82.92%	80.96%
Deprec.	39.92%	43.23%	38.95%	39.08%	40.06%	38.20%
Liabilities	96.85%	96.65%	96.69%	96.78%	97.09%	96.95%
Curr. liab.	78.49%	76.01%	77.77%	78.60%	80.07%	79.40%
Curr. ltd	49.58%	47.33%	49.60%	48.88%	51.79%	50.22%
Debt	45.62%	41.13%	45.37%	44.46%	50.58%	47.05%

By looking at the variables, the following observations can be made. The variables, representing the largest concern, are related to long term debt, depreciation and accounts receivable. The rest of the variables have at least 70% of values available for research computations. From the accounting perspective, the values of current and long-term debt are absolutely optional. The company may not have any debts present at the particular balance sheet. The absence of the values, related to accounts receivable and depreciation is more troubling.

Upon further examination, of XBRL based statements it was revealed that in many cases companies tend to use custom tags for these two accounting variables. While this fact does not bear any significance in displaying the data, it is obviously detrimental for data extraction because the tags used are not connected to the standard tags in any way. If the regulator wants to perform the initial automatic auditing of the submitted statements, they must demand that standard tags are used for the standard statement variables whenever possible.

From further examination of the data, presented in Table 2, one can see that the majority of the variables have a very consistent percentage of usable values across the years. It means that the companies generally stay with the same submission pattern, without significant improvement and deterioration of data. In the view of this, the overall sample of data appears to be fully representative of the acquired data and can be safely used in research without the concerns of changing year-to-year data yield.

M-Score by Beneish requires the data from two adjacent years, the calculation of Z-Score usually requires data recorded for the current year. In order to be consistent with the numbers of entries for both indicators, we omitted Z-Score calculations for 2011. The calculations for Z-Score are presented in Table III.

TABLE 3. DESCRIPTIVE STATISTICS FOR Z-SCORE FOR ALL SAMPLES

Var	N	Mean	St. Dev	Q1	Median	Q3
Fraud	124	0.323	4.745	0.329	0.97	1.896
Clean	36	3.452	1.166	2.79	3.346	3.882
All	5083	1.1941	2.4899	0.4451	1.4717	2.4882
2012	1201	1.307	2.5103	0.5563	1.5591	2.5623
2013	1382	1.1891	2.5704	0.4813	1.4953	2.4781
2014	1196	1.3221	2.2215	0.5483	1.5049	2.4686
2015	1226	0.8817	2.595	0.1234	1.2364	2.3186

All samples, presented in Table 3, were subject to Kolmogorov-Smirnov normality test. All samples appear normal with at least 99% significance. T-Value comparison of the means showed that means of all XBRL samples are different from the means of fraud and clean sample. The values of medians of all XBRL based samples appear to be in 1 ≤ Z-Score ≤ 3 range, which, according to the theory, represents so-called "grey" zone for the companies with average or slightly below average liquidity. These values of Z-Score appear in the research papers by Altman, mentioned earlier. The values for 2015 may represent a slight concern from the liquidity perspective, which skew the value of the larger sample. Such anomaly may be a sign of economic troubles related to the number of other events.

The calculation for M-Score are presented in Table IV.

TABLE 4. DESCRIPTIVE STATISTICS FOR M-SCORE FOR ALL SAMPLES

Var	N	Mean	St. Dev	Q1	Median	Q3
Fraud	123	-1.682	1.358	-2.452	-1.781	-0.929
Clean	36	-1.239	1.038	-1.87	-1.38	-0.579
All	1480	-1.486	1.332	-2.385	-1.625	-0.726
2012	322	-1.520	1.333	-2.433	-1.672	-0.798
2013	422	-1.442	1.479	-2.322	-1.501	-0.707
2014	377	-1.438	1.205	-2.353	-1.616	-0.673
2015	331	-1.584	1.293	-2.492	-1.702	-0.764

All samples, presented in Table IV, were subject to Kolmogorov-Smirnov normality test. All samples appear normal with at least 99% significance. T-Value comparison of the means showed that means of all XBRL samples are different from the mean of the clean sample. The values of XBRL data do not exhibit any out-of-usual behavior, showing that there is a fair number of companies with strong performance, which have relatively high M-Score.

During the extraction, the results revealed another problem, which was not anticipated at the beginning. There was a mismatch in the presentation of precision for certain variables. Although this fact does not affect the display of the financial statements, it is clearly affecting the calculations as they appear to be shifted based on increased, decreased and/or omitted precision attribute.

Table 5 represents the percentages of usable entries for each indicator, considering that all variables included must have non-null values. The “No Outlays” columns represents the values, which do not have any out-of-norm values, such as |Z| > 10, | ΔP-ΔR| > 5 or zero values for any indicator in M-Score formula. The percentages of “No-Outlays” are based on the total number of usable values of 5441.

TABLE 5. PERCENTAGES OF USABLE VALUES OF INDICATORS

Var	Total	2012	2013	2014	2015	No Outlays
M	12.39%	12.1%	12.8%	15.6%	10.4%	27.20%
Z	49.61%	49.4%	48.5%	50.6%	49.8%	93.42%

The yield of usable statements appears to be rather low at 5441 out of 16439 in the years of 2012-2015 or 1360 per year on average. However, the similar research on earnings management performed by [26] used close to 15000 entries in 10 years or 1500 per year on average. Hence, XBRL feed yield fits the needs of a research of the mentioned magnitude.

The data from Table 2 shows that there is no unequivocal answer to the first research question. There is enough usable data in the XBRL based financial statements (see previous paragraph), but the percentage of the usable statements can only be determined when the variables for the particular research are defined.

The response to the second question appears to be more definite. The data obtained from XBRL represents the random sample of data. Calculated values for Z-Score, presented in Table 3 are clearly in so-called ‘grey zone’ (1 ≤ Z ≤ 3), where the majority of the U.S. Companies operate. The values of M-Score appear to be lower than the values for the companies with exceptional performance and higher than the ones for the revenue manipulators. This is also consistent with the expectations for a moderately well-performing random company. Therefore, it is possible to say, that if there is a sufficient volume of data, extracted from XBRL via the previously described algorithm, it can be used in financial research as a random data (similar to the one obtained from Yahoo or COMPUSTAT).

4.1. Consequences for Software Development Using XBRL

The use of XBRL for the filing is rapidly becoming mandatory in the free market countries. Being a structured language facilitating the electronic transmission of the financial data, XBRL is tested in this research for the use in the financial analysis of the data it carries. The general possibility of such extraction exists. However, over the course of performing this research the authors came across a number of significant limitations standing on the way of extracting efforts.

For many years the data in the financial statements was as significant as the labels, which describe the meaning of such data, creating an unbreakable pairing of label and data [27]. XBRL introduces the third component of the filing – the tag, which encapsulates the data and has a relationship with the label, representing the meaning of the tag to the public (hence the meaning of the data the tag contains).

For the observer, reading the statement from the regulator web site the tag is concealed. They still observe the same relationship between the data and the label. The companies, which submit the statement place more emphasis on the label than on the tag it is attached to. Presently they select the tag based on its suitability for data description. The intended use of the tags based on the company profile (Industrial, Investment, Real Estate, etc.) remains on the advisory level. The described situation makes it very difficult to create and follow the tree calculation patterns described in [1], as well as in this research.

There are various means of rectifying this XBRL filing paradox. The research by [28] proposes the use of the paradigms such as Ontology Web Language (OWL), which are suitable for the knowledge management systems. Considering that XBRL filings are containers of the financial facts, this approach is very suitable for both assembling and parsing XBRL statements.

We propose using another approach to the systematization of the XBRL data based on the containment pattern used in the software development [29]. This pattern denotes one object as a container, which owns the references to the other objects. This pattern is already used in the XBRL filings in the form of linkbases, containing references to other elements, such as facts, arcs, etc. The accounting practices define very specifically, which elements comprise certain statement or any of its parts. The containment relationship between the “top of the tree” tag for Accounts Receivable and the custom tags, depicting receivables, mentioned earlier could allow inclusion of the custom tags into the calculation of the basic variables such as Accounts Receivable. Figure 2 offers a common representation of the Containment pattern, tied to the tree pattern used in this research and in [1].

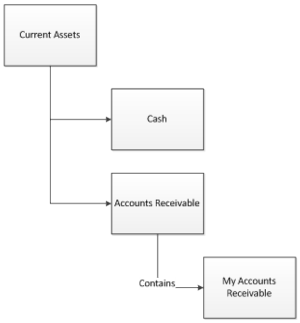

Figure 2. Containment Pattern

Figure 3 shows a software development representation of the same pattern. It is important to note that software relationship between standard element and the custom element must be of aggregation type. The composition association between the object on Figure 3 depicts the fact that fact that these two elements share a connection on the GAAP level. The share (aggregate) relationship signifies the fact that elements have relationship on the filing level only. The relationship cardinality of 0..1 shows that root elements may or may not exist in the filing. The relationship of the type 0..* shows that if the elements exist within the parent arc there may be more than one of them. This pattern can be extended towards the mentioned practice of using tags, which are not a part of the standard calculation tree.

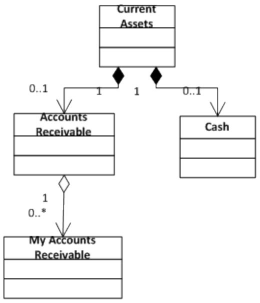

Figure 3. Object Relationship between Elements of Current Assets

The containment relationship between Current Assets and its parts already exists. It is recorded in the taxonomy calculation tree. The company should only add the element (i.e., Accounts Receivable from the taxonomy) and create a calculation relationship between its own AR tag and the tag in the taxonomy structure. Any custom tag, which does not have a taxonomy based container should be considered a 'dangling' tag and forced out by the filing verification software.

This containment relationship reflects similar relationships existing between the elements of the financial statements. Any of the three major statements contains a finite number of elements. Any financial information fed into the statement (ex. Allowable Expenses) can be always matched with an element of the statement. If the company requires showing non-standard elements on their Statement of Income they would always summarize these elements into a line, which is standard for the Income Statements across the nationally accepted GAAP.

It can be foreseen, that the use of XBRL will be extended in the coming years. However, using the largest repository of XBRL based documents (U.S. SEC XBRL storage) as an example, it can be easily concluded that blind extraction of the financial information based on data description, contained in the supporting taxonomy, is not quite possible. Any algorithms mentioned in chapters 2 and 4 will require a perfect match between the taxonomy and the files, claiming to use it.

The companies and the hired agents are using XBRL to submit the documents, which can be used to display the financial information. In the process of converting the accounting data into the financial statements two things are important: the presented data and its description. There is absolutely no requirements that the XML component of the parsing process (the tag) has one-to-one correspondence with the presented data. Such discrepancy reduces the ability to parse the data to a chance. It is very tempting to work only with the tags that have over 90% ability to be parsed (according to Table 2). Such choice cannot be made because any research or analysis effort must work with the data required and not data at hand.

5. Conclusion

The goal of this research was to show to the community interested in the advancement of XBRL based filings certain possibilities of XBRL use and the limitations of the current state of XBRL based filing process. The research shows, that even with the limited number of filings available from SEC, it is possible to obtain enough data for performing the financial research similar to the ones, which were performed by using data from concentrators, such as Yahoo or COMPUSTAT.

The research clearly shows that the data from the financial repository of U.S. SEC is marginally suitable for the financial research. Although the data is checked against the major accounting equations, the presence of the required tags is not guaranteed by the rules of XBRL Reporting

The extension tags used in the reporting may be valid for the company that uses them. The purpose of these tags from the perspective of the filing company is to display the data as it was previously pointed out. The lack of connection of this data with the main taxonomy presents an unsolvable problem to any parsing mechanism.

The research shows that in its present state XBRL filing process does not clearly support retrieving of data, which was submitted in the filing. The retrieval efforts performed by this research, which did not include text parsing of the XBRL tags, yielded around 25% of the data suitable for further financial research. The containment approach to placement of the custom tags described in this paper would increase the number of suitable filings significantly.

The accounting science does not allow for any fuzzy conclusions. Any text parsing of XML tags using fuzzy logic is prone to type I and type II errors. Using accounting data, origin of which (tag from which it was taken) cannot be 100% guaranteed is a potential source of accounting errors and wrong conclusions based on the use of such data. The lack of the rules around tag naming adds to the concerns about using OWL and fuzzy logic in queries of XBRL files.

It is possible to reduce the discarding of the filings on account of them using tags not belonging to the designated tree pattern. However, the number of tags, which can possibly substitute the required tag from the tree pattern is rather high and it would slow down the processing of the statements significantly. The enforcement of the containment pattern by the developers of the national taxonomies, allowing extensions, can greatly reduce such discarding as more filings could be read by using the tree pattern mechanism.

The U.S. XBRL taxonomy was created to appease all companies with all lines of business. This was done in order to ease the process of creating XBRL documents containing the required financial statements. The assumption was made that the accountants of the companies have similar training and certification background hence will use the same tags for the same purposes.

In this research we took a very specific area of financial research, which requires application of certain rules. The further efforts in this area can be directed towards applying this algorithm to the other areas of the financial research in order to verify how much data can be used. Out of the questions posted by the younger researchers, there is a prevalent one: how much can we trust the data submitted to U.S. SEC.

Despite the fear that the data may come from the companies perpetrating fraud, it is necessary to assert that the data coming from the XBRL based financial statements is the data that every financial analyst sees and uses in financial decisions. The data, obtained from the XBRL based statements is a close to the source as possible. The results obtained by using this data represent the true, unaltered picture of the financial situation.

One of the largest limitations of this research is that apart from [1] nobody produced a comprehensive research of blind querying the data. The data, obtained in [4], [5] is a result of manual extractions and any results obtained in [19]-[21] do not guarantee that the files were selected randomly. The results and the errors, similar to the ones published in [1] and in this research have not yet been published. The authors hope that the definitions of the limitations can be used in the similar efforts applied to the other taxonomies.

Financial reporting has rules it adheres to but even these rules do not guarantee that for every filing all elements required in the research will be present. However, event with the natural differences in the filings it must be possible to reduce the rate of discarding of the data, which cannot be used, in the particular research if the filing rules are strengthened.

REFERENCES

[1] R. S. Debreceny, A. D'Eri, C. Felden, S. M. Farewell, and M. Piechocki, "Feeding the Information Value Chain: Deriving Analytical Ratios from XBRL filings to the SEC," presented at the 22nd XBRL International Conference, Kansas University, KS, 2010.

[2] R. Pinsker, "XBRL awareness in auditing: a sleeping giant?," *Managerial Audit Journal,* vol. 18, pp. 732-736, 2003.

[3] J. G. San Miguel, "The Reliability of R&D Data in COMPUSTAT and 10-K Reports," *The Accounting Review*, vol. 52, pp. 638-641, 1977.

[4] E. Boritz and W. G. No. (2013, The Quality of Interactive Data: XBRL Versus Compustat, Yahoo Finance, and Google Finance. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2253638 [Retrieved: July, 2013]

[5] R. Chychyla and A. Kogan. (2013, Using XBRL to Conduct a Large-Scale Study of Discrepancies between the Accounting Numbers in Compustat and SEC 10-K Filings. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2304473 [Retrieved: July, 2013]

[6] R. S. Debreceny, S. M. Farewell, M. Piechocki, C. Felden, and A. Graening, "Does it add up? Early evidence on the data quality of XBRL filings to the SEC," *Journal of Accounting and Public Policy*, vol. 29, pp. 296–306, 2010.

[7] R. S. Debreceny, S. M. Farewell, M. Piechocki, C. Felden, A. Graening, and A. D'Eri, "Flex or Break? Extensions in XBRL Disclosures to the SEC," *Accounting Horizons*, vol. 25, pp. 631-657, 2011.

[8] R. S. Debreceny and G. Gray, "The production and use of semantically rich accounting reports on the Internet: XML and XBRL " *International Journal of Accounting Information Systems*, vol. 2, pp. 47-74, 2001.

[9] E. Boritz and W. G. No, "Computer-Assisted Functions for Auditing XBRL-Related Documents," *Journal of Emerging Technologies in Accounting*, vol. 13, pp. 53-83, 2016.

[10] R. S. Debreceny, A. Chandra, J. J. Cheh, D. Guithues-Amrhein, N. J. Hannon, P. D. Hutchison, *et al.*, "Financial Reporting in XBRL on the SEC's EDGAR System: A Critique and Evaluation," *Journal of Information Systems*, vol. 19, pp. 191-210, 2005

[11] E. Boritz and W. G. No, "Business Reporting with XML: XBRL (Extensible Business Reporting Language)," in *Encyclopedia of the Internet* vol. 3, H. Bidgoli, Ed., ed: John Wiley and Sons, 2004, pp. 863-885.

[12] M. Bovee, A. Kogan, R. P. Srivastava, M. Vasarhelyi, and K. Nelson, "Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL)," *Jounal of Information Systems*, vol. 19, pp. 19-41, 2005.

[13] M. A. Geiger, D. S. North, and D. D. Selby, "Releasing Information in XBRL: Does It Improve Information Asymmetry for Early U. S. Adopters?," *Academy of Accounting and Financial Studies Journal*, vol. Arden 18.4, pp. 66-83, 2014.

[14] J.-H. Lim and T. Wang, "The Impact of Service Provider Switches on XBRL Qua," in *Pacific Asia Conference on Information Systems*, 2016.

[15] M. Enachi and I. I. Andone, "The Progress of XBRL in Europe – Projects, Users and Prospects," *Procedia Economics and Finance*, vol. 20, pp. 185-192, 2015.

[16] R. S. Debreceny and G. Gray, "A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits " *International Journal of Accounting Information Systems*, vol. 15, pp. 357-380, 2014.

[17] R. Chowdhuri, V. Y. Yoon, R. T. Redmond, and U. O. Etudo, "Ontology based integration of XBRL filings for financial decision making," *Decision Support Systems*, vol. 68, pp. 64-76, 2014.

[18] E. Tsui, W. M. Wang, L. Cai, C. F. Cheung and W. B. Lee, "Knowledge-based extraction of intellectual capital-related information," *Expert Systems with Applications*, vol. 41, no. 2014, pp. 1315-1325, 2014.

[19] M. Radzimski, J. L. Sanchez-Cervantes, A. Garcia-Crespo and I. Temiño-Aguirre, "Intelligent Architecture for Comparative Analysis of Public Companies Using Semantics and XBRL Data," *International Journal of Software Engineering and Knowledge Engineering*, vol. 24, no. 05, p. 801, 2014.

[20] B. Kämpgen, T. Weller, S. O'Riain, G. Weber and A. Harth, "Accepting the XBRL Challenge with Linked Data for Financial Data Integration," in *The Semantic Web: Trends and Challenges*, 2014.

[21] B. M. Franceschetti and C. Koschtial, "Do bankrupt companies manipulate earnings more than the non-bankrupt ones? " *Journal of Finance and Accountancy* vol. 12, pp. 1-22, 2013.

[22] I. Pustylnick, "Using Z-Score in detection of revenue manipulations," presented at the 21-st International Scientific Conference Economics and Management, Brno, Czech Republic, 2016.

[23] E. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of* Finance, pp.589-609, 1968.

[24] M. Beneish, "Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance,"

Journal of Accounting and Public Policy, vol. 16, pp. 271-309, 1997.

[25] J. Escalada. (2011, August 29,2012). *45 Dividend Stocks With Good Credit Scores*. Available: http://seekingalpha.com/article/307563-45-dividend-stocks-with-good-credit-scores [Retrieved: March, 2012]

[26] I. P. Jansen, S. Ramnath, and T. L. Yohn, "A Diagnostic for Earnings Management Using Changes in Asset Turnover and Profit Margin," *Contemporary Accounting Research*, vol. 29, pp. 221-251, 2012.

[27] A. van der Hoek, "Design-time product line architectures for any-time variability," *Science of Computer Programming*, vol. 53, pp. 285-304, 2004.

[28] C. Hoffman and M. M. Rodrigues, "Digitizing Financial Reports – Issues and Insights: A Viewpoint," *International Journal of Digital Accounting Research*, vol. 13, pp. 73-98, 2013.

[29] C. Blilie, "Patterns in scientific software: an introduction," *Computing in Science & Engine*, vol. 4, pp. 48-53, 2002.