An SEC 10-K XML Schema Extension

Full Paper

William Augustine
University at Albany, SUNY
waugustine@albany.edu

Sanjay Goel
University at Albany, SUNY
goel@albany.edu

Abstract

The paper addresses the limitation of the current XBRL schema in being able to extract information related to information security disclosures from the annual disclosure statements of firms. It is possible to automatically extract the security relevant information from these statements with parsing XML tags through the use of scripts that extract text based on specific delimiters occurring in text; however, this solution is difficult, inefficient, and error-prone. XBRL provides a structure based on XML for extraction of this data, however, there are gaps in the XBRL nomenclature that necessitates the use of scripting for pulling the non-XBRL components. The extensive written portions of the annual disclosure statements where the security related information is present (based on SEC guidance) tends to be poorly described in terms of tag set deployment. The extension to XBRL proposed in this paper will help gather security information from SEC reports more robustly and accurately. The paper presents the schema as well as an example of its application to a sample SEC report.

Introduction

The U.S. Securities and Exchange Commission (SEC) has mandated publicly traded firms to file an annual disclosure statement ("Form 10-K") that contains information about the performance of the firm; including financial information. These reports also contain extensive narratives from the company's leaders about future plans and market expectations. The information is meant to inform current and potential shareholders about company performance so investors can assess the risks of the investment and determine the appropriate valuation of the stock. There is a lot of valuable information in these reports both in text as well as in the financial data provided in the firms. This information has been used extensively for both academic research (detailed below) and market analysis. Section 7, which contains management discussions and analysis (MD&A), is particularly important since it provides information about qualitative and quantitative risks that the organization faces. We are specifically concerned about information security risks whose disclosure is now strongly suggested (though not mandated) by SEC through a guidance issued in 2011 (U.S. Securities and Exchange Commission 2011).

Even though it is possible to manually analyze the information, it is a lot more effective when the process of data extraction and analysis can be automated. Data analytic techniques, such as natural language processing (NLP), are making the text sections of these reports increasingly valuable because when they are accumulated and aggregated, a more complete picture of a company or market can be created; particularly when examining these documents longitudinally. While the components and order of the 10-K reports are mandated by law there is considerable variability across reports and consequently, parsing unstructured information becomes difficult. Scripting languages can be used for extraction of the information based on use of specific delimiters however this is difficult to do and makes the process error-prone. XBRL provides a structure based on XML for extraction of this data, however, there are gaps in the XBRL nomenclature that necessitates the use of scripting for pulling the non-XBRL components. We propose an extension beyond the existing XBRL schema to fill the gaps in the XBRL schema and allow for error free extraction of the data for research.

To extract just those elements of value for a particular purpose, tagging rules can be utilized. This form of employment currently exists and is well defined for the accounting (ledger) data elements but is lacking throughout the main body of the documents. Currently, the extensive written portions of these documents tend to be poorly described in terms of tag set deployment. Only the interactive (XBRL formatted) documents undergo comprehensive format and validation checks.

This paper proposes an XML schema extension to the SEC report taxonomy specification. This would be used in addition to the XBRL accounting taxonomy already in use. This tag set would label the individual subsections of the SEC report in order to allow better automated access to the data included in prose format. The rest of the paper is organized as follows: Section 2.0 provides the Literature Review related to efforts on data extraction from 10-k files, Section 3.0 presents the background of 10-K reports and SEC mandates; Section 4.0 shows our proposed schema; and section 5.0 presents a brief summary.

Literature Review

The eXtensible Markup Language (XML) is a data markup language that is descended from Standard Generalized Markup Language (SGML) and serves to semantically mark elements of data within documents. What allows for XML's extensibility is provision for customizing tag sets. A key extension is the eXtensible Business Reporting Language (XBRL). The primary schema elements used in processing SEC reports are defined within the US GAAP taxonomy. This specification defines how to tag specific accounting data and was developed by the Financial Accounting Standards Board (FASB). While, the tags have been defined comprehensively for most accounting information, research has recognized that other parts of the reports have not been tagged thoroughly (Gerdes 2003). SEC reports represent a treasure trove of free (open and unencumbered) data usable to explain a multitude of aspects about the form and operations of commercial organizations. Specifically, there is hidden information in unstructured data that can be leveraged for research. For instance, researchers from the University of Notre Dame examined 100,404 10-K reports from between 1994 and 2006 for “ethics-related terms” and whether there were any correlations with subsequent malfeasance (Loughran et al. 2009). The requisite ethics statement disclosures on forms 8-K and 10-K (in addition to other sources) were used by Rodrigues and Stegemoller (2010) to look for adherence to Sarbanes-Oxley regulations.

Consequently, there is a large need for systematically accessing unstructured textual information from 10-K files for research but the problem is non-trivial. The problem is difficult enough that some researchers have gone so far as to use genetic algorithms to parse 10-K reports (Carroll et al. 2008). Carroll and Lee described the difficulties inherent in identifying labels in the documents for use in extracting specific pieces of text; such as Items 7 or 8. “While a human reader can easily resolve subtle naming inconsistencies and distinguish cross-references from misordered segments, automated strategies are typically limited by the available training examples” (ibid, p.2). Using this highly sophisticated approach on 112 randomly selected filings from 2005 resulted in varied levels of success. These researchers obtained an f-measure of 0.834 in identifying Item 7, however identification of Part IV only obtained an f-measure of 0.207.

Bao and Datta used textual analysis on SEC annual reports to categorize risk types for which the authors developed a sentence based variation of Latent Dirichlet Allocation, they called Sent-LDA. This method was then tested in the Part 1 sections of 10-K reports. This authors recognized that "[i]t is quite challenging to extract textual risk factors in section 1A from 10K forms because they are highly unstructured" (Bao et al. 2012, p.6]. To support their research, 10-K reports from EDGAR were parsed and scraped using custom heuristic rules. The authors acknowledge that there still may be mis-extracted contents.

Kogan, et al. (2009) looked at file sizes, word counts and term frequencies as predictors of the volatility in stock performance. That work used both full files and extracted Item 7 (MD&A) sections because the authors considered the latter to be “where the most important forward-looking content is most likely to be found” (Kogan et al. 2009, p.3). Custom programming was used to identify and cull just the MD&A from the reports for analysis but because of inconsistencies in the preparation of the reports “[n]ot all of the documents downloaded pass the filter at all” (ibid). Other researchers used an analysis of 10-K MD&A narratives for an explanation of changes in inventory increases as reported in the financial results. This was done by manually reading and hand coding of the MD&A narratives of 568 sample company reports (Sun 2010).

Several products have been developed to provide access to this freely available data. Some of these are fee based software as a service (SaaS) offerings while others are provided free of charge (but often require some form of registration to use without restriction). Many allow for the use of a web based element specific RESTful API to get at the desired 10-K data. Table 1 lists a small sample of the products and

projects available beyond the SEC EDGAR portal primarily because of the difficulty in extracting meaningful information from these submission documents. The creator of the Rank and Filed tool, Maris Jensen, laments on the project's home page (http://rankandfiled.com/) that "[d]ata tagging is the red-headed stepchild of the Commission -- out of hundreds of forms, only about a dozen are filed as structured data -- and the first program to automate the selection of SEC filings for review, the Division of Economic and Risk Analysis (DERA)'s 'Robocop', has been 'aspirational' for years." Perhaps, instead of trying to solve everything at once a stepwise solution can be employed wherein small changes can be implemented quickly and in succession to create a larger solution.

Company	URL
CorpWatch	http://api.corpwatch.org/
EDGAR Online	http://www.edgar-online.com/
Kimonolabs, Inc.	http://kimonolabs.com/sec/explorer/
Last10k.com	https://dev.last10k.com/
Rank and Filed	http://rankandfiled.com/

Table 1

Due to the vast amount of data available through SEC filings researchers have employed data mining techniques to process them. Annual report data obtained from EDGAR has been used to create support vector machine classifiers for determining credit risk (Danenas et al. 2015).

SEC Mandates & Edgar

The US Congress passed the 1934 Securities and Exchange Act; a key component of which was the creation of the SEC following the Black Tuesday (October 29, 1929) great stock market crash that exacerbated the country's general economic malaise and brought forth the Great Depression. A key purpose of the SEC has been their charge of overseeing and regulating the capitalization of publicly traded companies. One mechanism used to perform this function is through the mandate for public companies to provide standardized reports of their finances and those important conditions that might influence one to invest (or not to invest) money in those companies' securities (primarily, stocks, bonds and related funds). Companies are legally required to file many types of reports and documents with the SEC including quarterly statements (form 10-Q) as well as an annual statement (form 10-K). "Section 13(b) of the '34 Act (and section 19(a) of the '33 Act) gives the SEC the power to prescribe the form and content of the financial statements filed under the Act." (Benston 1973, p.133).

Although referred to as an annual report, the 10-K is often materially different from the annual reports that companies distribute to shareholders; the latter tend to be primarily promotional material. This work will refer to the SEC 10-K reports with the understanding that the 10-Q is considered equivalent such that the evaluation and recommendations apply to both. The 10-K consists of four parts. Part I, contains Items 1 through 4 which describe the company, its business and risk factors. Part II, Items 5 through 9 are for financial data. Part III covers Items 10 through 14 and details personnel and governance. Lastly, Part IV consists of Item 15 and is for exhibits, financial statements and schedules related to those values expressed in Part II. These reports are collected by the SEC and made available to the public free of charge.

Since 1996, the SEC Office of Interactive Disclosure has provided online access to the required company submitted documentation[1] through the Electronic Data Gathering, Analysis and Retrieval (EDGAR) system. EDGAR provides search options for locating submitted filings based on entity name, stock ticker symbol or central index key (CIK). The CIK is an SEC assigned value for uniquely identifying a submitter.

1 Some exceptions exist, such as Forms 3, 4, 5, 144 and others; plus filers that apply for a hardship exemption.

Different disclosures for a submitter are kept and made available historically (back to 1996, if available). Groups of companies can be identified by searching based on standard industrial code (SIC), country or state of incorporation.

User accessibility to EDGAR files is allowed using the file transport protocol (FTP) (in anonymous, passive mode) or through a browser based interface. FTP users can download index files identifying directory locations of documents for further retrieval. A feed of real time submissions is also available using really simple syndication (RSS). Documents are stored in character format (binary data, such as images or spreadsheets, are base64 encoded and embedded as ASCII characters) and often multiple markup formats are used. For example, Table 2 shows some of the markup headers from the 2015 10-K report file for the TJX Companies, downloaded from EDGAR[2]. It is not unusual for a single 10-K file to include SGML, HTML and XML formats. The many markup formats were created for different purposes and some markup formats are more descriptive or expressive than others. In this case there are 119 documents embedded in this file; 107 HTML, 7 XBRL, 2 MS Excel spreadsheets plus some included Javascript and Stylesheets.

Line No.	Tags and Elements
1	<SEC-DOCUMENT>0001193125-15-114276.txt : 20150331
2	<SEC-HEADER>0001193125-15-114276.hdr.sgml : 20150331
3	<ACCEPTANCE-DATETIME>20150331164230
4	ACCESSION NUMBER: 0001193125-15-114276
5	CONFORMED SUBMISSION TYPE: 10-K
6	PUBLIC DOCUMENT COUNT: 21
7	CONFORMED PERIOD OF REPORT: 20150131
8	FILED AS OF DATE: 20150331
9	DATE AS OF CHANGE: 20150331
10	
24680	<FILENAME>tjx-20150131.xml
24681	<DESCRIPTION>XBRL INSTANCE DOCUMENT
24682	<TEXT>
24683	<XBRL>
24684	<?xml version="1.0" encoding="us-ascii" standalone="yes"?>
50683	<FILENAME>tjx-20150131.xsd
50684	<DESCRIPTION>XBRL TAXONOMY EXTENSION SCHEMA
50685	<TEXT>
50686	<XBRL>
50687	<?xml version="1.0" encoding="US-ASCII"?>
51491	<FILENAME>tjx-20150131_cal.xml
51492	<DESCRIPTION>XBRL TAXONOMY EXTENSION CALCULATION LINKBASE
51493	<TEXT>
51494	<XBRL>
51495	<?xml version="1.0" encoding="US-ASCII"?>
52115	<FILENAME>tjx-20150131_def.xml
52116	<DESCRIPTION>XBRL TAXONOMY EXTENSION DEFINITION LINKBASE
52117	<TEXT>
52118	<XBRL>
52119	<?xml version="1.0" encoding="US-ASCII"?>

2 http://www.sec.gov/Archives/edgar/data/109198/000119312515114276/0001193125-15-114276.txt

```
54770 <FILENAME>tjx-20150131_lab.xml
54771 <DESCRIPTION>XBRL TAXONOMY EXTENSION LABEL LINKBASE
54772 <TEXT>
54773 <XBRL>
54774 <?xml version="1.0" encoding="US-ASCII"?>

58100 <FILENAME>tjx-20150131_pre.xml
58101 <DESCRIPTION>XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE
58102 <TEXT>
58103 <XBRL>
58104 <?xml version="1.0" encoding="US-ASCII"?>
```

Table 2

Proposed XML Schema

One such change would be to use XML tagging to identify the component sections within the annual reports. This should not be an onerous burden on the filers. Even adhering to the XBRL mandate cost almost 70% of companies less than $10,000 per year (AICPA 2015). A simple document type definition (DTD) against which to validate an enhanced tagged 10-K is given in Table 3. This could also be implemented using eXtensible Schema Definition (XSD).

DTD

```
<!DOCTYPE SEC10K
[
<!ELEMENT SEC10K (Part1, Part2, Part3, Part4)>
<!ELEMENT Part1 (Item1, Item1A, Item1B, Item2, Item3, Item4)>
<!ELEMENT Item1 (#PCDATA)>
<!ELEMENT Item1A (#PCDATA)>
<!ELEMENT Item1B (#PCDATA)>
<!ELEMENT Item2 (#PCDATA)>
<!ELEMENT Item3 (#PCDATA)>
<!ELEMENT Item4 (#PCDATA)>
<!ELEMENT Part2 (Item5, Item6, Item7, Item7A, Item8, Item9, Item9A, Item9B)>
<!ELEMENT Item6 (#PCDATA)>
<!ELEMENT Item7 (#PCDATA)>
<!ELEMENT Item7A (#PCDATA)>
<!ELEMENT Item8 (#PCDATA)>
<!ELEMENT Item9 (#PCDATA)>
<!ELEMENT Item9A (#PCDATA)>
<!ELEMENT Item9B (#PCDATA)>
<!ELEMENT Part3 (Item10, Item11, Item12, Item13, Item14)>
<!ELEMENT Item10 (#PCDATA)>
<!ELEMENT Item11 (#PCDATA)>
<!ELEMENT Item12 (#PCDATA)>
<!ELEMENT Item13 (#PCDATA)>
<!ELEMENT Item14 (#PCDATA)>
<!ELEMENT Part4 (Item15)>
<!ELEMENT Item15 (#PCDATA)>
]>
```

Table 3

Once this is in place, simple XML utilities supporting XPATH queries could be used to extract the full text for each of the individual item sections for analysis and further processing. A small submission file was identified and the HTML tags were changed to make the document well-formed (this involved providing

closing markers for non-block tags, such as <hr/>, and setting attributes to be of the form attribute=”value”). Requiring this of 10-K submissions should not cause undue hardship. The <body> tag was followed by <xml> and </body> was preceded with </xml>. The blocks which encapsulate Item1 and Item 7A were tagged with <Item1>,</Item1> and <Item7A>,</Item7A> respectively. Using the xmlstarlet toolkit (http://xmlstar.sourceforge.net/), the resulting file was validated using the command

```
$ xml val test10k.html
```

and then Items 1 and 7A were effortlessly extracted using the XPATH query commands

```
$ xml sel -t -v "html/body/xml/div/div/Item1" test10k.html
```

and

```
$ xml sel -t -v "html/body/xml/div/Item7A" test10k.html
```

respectively.

Conclusion and Future Work

Creating an easy means for individuals to benefit from the value of the reports mandated by the SEC would improve the confidence of investors and the safety of their investments. In turn, that would benefit business by improving their ability to raise capital since more money will be brought into the financial markets. Working around the difficulties of getting the desired information from 10-K reports has involved academic contortions and the generation and deployment of third party tools which does not appear to mesh with the original intention of the EDGAR designers and the purpose of requiring XBRL tagging of financial data. A simple DTD to validate an XML tag set for the report item text elements would be easy to achieve and provide easy access for future NLP use.

REFERENCES

American Institute of Certified Public Accountants (AICPA). 2015. “XBRL Costs for Small Companies.pdf.,” (available at http://www.aicpa.org/InterestAreas/FRC/AccountingFinancialReporting/XBRL/DownloadableDocuments/XBRL%20Costs%20for%20Small%20Companies.pdf).

Bao, Y., and Datta, A., "Summarization of corporate risk factor disclosure through topic modeling.", *Proceedings of the 33rd International Conference on Information Systems*, Orlando, FL, 2012

Benston, G. J. 1973. “Required disclosure and the stock market: An evaluation of the Securities Exchange Act of 1934,” *The American Economic Review* (63:1), pp. 132–155.

Bryant, R., Katz, R. H., and Lazowska, E. D. 2008. *Big-data computing: creating revolutionary breakthroughs in commerce, science and society*, December (available at http://www.datascienceassn.org/sites/default/files/Big%20Data%20Computing%202008%20Paper.pdf).

Carroll, J., and Lee, T. Y. 2008. “A genetic algorithm for segmentation and information retrieval of SEC regulatory filings,” in *Proceedings of the 2008 international conference on Digital government research*, Digital Government Society of North America, pp. 44–52 (available at http://dl.acm.org/citation.cfm?id=1367843).

Danenas, P., and Garsva, G. 2015. “Selection of Support Vector Machines based classifiers for credit risk domain,” *Expert Systems with Applications* (42:6), pp. 3194–3204 (doi: 10.1016/j.eswa.2014.12.001).

Gerdes, J. 2003. “EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC’s EDGAR database,” *Decision Support Systems* (35:1), pp. 7–29.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. 2009. “Predicting risk from financial reports with regression,” in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 272–280 (available at http://dl.acm.org/citation.cfm?id=1620794).

Loughran, T., McDonald, B., and Yun, H. 2009. “A Wolf in Sheep’s Clothing: The Use of Ethics-Related Terms in 10-K Reports,” *Journal of Business Ethics* (89:S1), pp. 39–49 (doi: 10.1007/s10551-008-9910-1).

Rodrigues, U., and Stegemoller, M. 2010. "Placebo ethics: a study in securities disclosure arbitrage," *Virginia Law Review*, pp. 1–68.

Sun, Y. 2010. "Do M&A Disclosures Help Users Interpret Disproportionate Inventory Increases?," *The Accounting Review* (85:4), pp. 1411–1440 (doi: 10.2308/accr.2010.85.4.1411).

U.S. Securities and Exchange Commission. 2011. *CF Disclosure Guidance: Topic No. 2, Cybersecurity*. Retrieved from http://www.sec.gov/divisions/corpfin/guidance/cfguidance-topic2.htm