# THE ANNUAL REPORT ALGORITHM: RETRIEVAL OF FINANCIAL STATEMENTS AND EXTRACTION OF TEXTUAL INFORMATION

Jörg Hering

Department of Accounting and Auditing, University of Erlangen-Nürnberg,
Lange Gasse 20, Nürnberg D-90403, Germany

## ABSTRACT

*U.S. corporations are obligated to file financial statements with the U.S. Securities and Exchange Commission (SEC). The SEC´s Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system containing millions of financial statements is one of the most important sources of corporate information available. The paper illustrates which financial statements are publicly available by analyzing the entire SEC EDGAR database since its implementation in 1993. It shows how to retrieve financial statements in a fast and efficient way from EDGAR. The key contribution however is a platform-independent algorithm for business and research purposes designed to extract textual information embedded in financial statements. The dynamic extraction algorithm capable of identifying structural changes within financial statements is applied to more than 180,000 annual reports on Form 10-K filed with the SEC for descriptive statistics and validation purposes.*

## KEYWORDS

*Textual analysis, Textual sentiment, 10-K parsing rules, Information extraction, EDGAR search engine*

## 1. INTRODUCTION

Information Extraction (IE) can be defined as the process of “finding and extracting useful information in unstructured text” [1]. In contrast to Information Retrieval (IR), a technology that selects a relevant subset of documents from a larger set, IE extracts information from the actual text of documents [2]. Important sources for IE are unstructured natural language documents or structured databases [3] [4]. Since U.S. corporations are obligated by law to file financial statements on a regular basis with the U.S. Securities and Exchange Commission (SEC), the SEC´s Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system containing millions of financial statements is one of the most important sources of corporate information available [5] [1].Unfortunately, most of the available textual data in the SEC EDGAR database is weakly structured in technical terms [6] [7] [8] especially prior to 2002 when the use of markup

Dhinaharan Nagamalai et al. (Eds) : CoSIT, SIGL, AIAPP, CYBI, CRIS, SEC, DMA - 2017
pp. 147– 166, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70415

languages was less common [9]. A limited number of tagged items, formatting errors and other inconsistencies lead to difficulties in accurately identifying and parsing common textual subjects across multiple filings [10] [11] [7]. These issues directly affect the ability to automate the extraction of textual information from SEC submissions [10] [12] [13]. Business data providers are offering expensive commercial products (e.g. AcademicEDGAR+, Edgar Pro, Intelligize). As research in the context of textual analysis is growing (e.g. Tetlock 2007 [14]; Loughran and McDonald 2011a [15]; Jegadeesh and Wu 2013 [16]) the question occurs which particular financial statements and disclosures are publicly available for free, how to retrieve these corporate documents and how to decode the embedded textual information in order to be incorporated into investment decisions, trading strategies and research studies in financial economics [5].Today only a very limited amount of specific literature for extracting textual information from financial statements filed with the SEC and its EDGAR system is available (except Gerdes 2003 [10]; Stümpert et al. 2004 [17]; Grant and Conlon 2006 [1]; Engelberg and Sankaraguruswamy 2007 [18]; Cong, Kogan and Vasarhelyi 2007 [19]; Thai et al. 2008 [20]; Chakraborty and Vasarhelyi 2010 [21]; Hernandez et al. 2010 [22]; Garcia and Norli 2012 [5]; Srivastava 2016 [23]). This paper is based on neither of these because first, non-specialist technology is used to retrieve financial statements in an efficient way and secondly, the algorithm designed to extract textual information is platform-independent. The suggested method can compensate for expensive commercial products and help to replicate empirical research results. The paper shall serve as a technical guide on how to retrieve financial statements filed with the SEC and how to decode the embedded textual information provided by the EDGAR system for business and research purposes.

The remainder of the paper proceeds as follows. Section 2 presents the amount and variety of corporate documents distributed by the SEC´s electronic disclosure system. Section 3 demonstrates how to retrieve these documents from the EDGAR database. Section 4 describes the fundamentals of HyperText Markup Language and examines the electronic data provided by the SEC. Section 5 describes the fundamentals of regular expressions and specifies an algorithm to extract textual information embedded in financial statements. Section 6 validates the capabilities of the extraction algorithm. Section 7 presents descriptive statistics of annual reports filed with the EDGAR database. The last section concludes.

## 2. SEC´S EDGAR DATABASE

Publicly owned companies, their officers and directors as well as major investors are obligated by law (Securities Exchange Act 1934, Section 2) to file various disclosures (forms) with the SEC [10]. The main purpose of making certain types of corporate information publicly available is to improve the efficiency of security markets and to protect capital market participants [5]. “The laws and rules that govern the securities industry in the United States derive from a simple and straightforward concept: all investors, whether large institutions or private individuals, should have access to certain basic facts about an investment prior to buying it, and so long as they hold it. To achieve this, the SEC requires public companies to disclose meaningful financial and other information to the public. This provides a common pool of knowledge for all investors to use to judge for themselves whether to buy, sell, or hold a particular security” [24]. In order to protect investors, to maintain efficient capital markets and to improve access to publicly available corporate disclosures, the SEC developed the EDGAR database [10] and describes it as a system which “performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the U.S. Securities and Exchange Commission" [25].

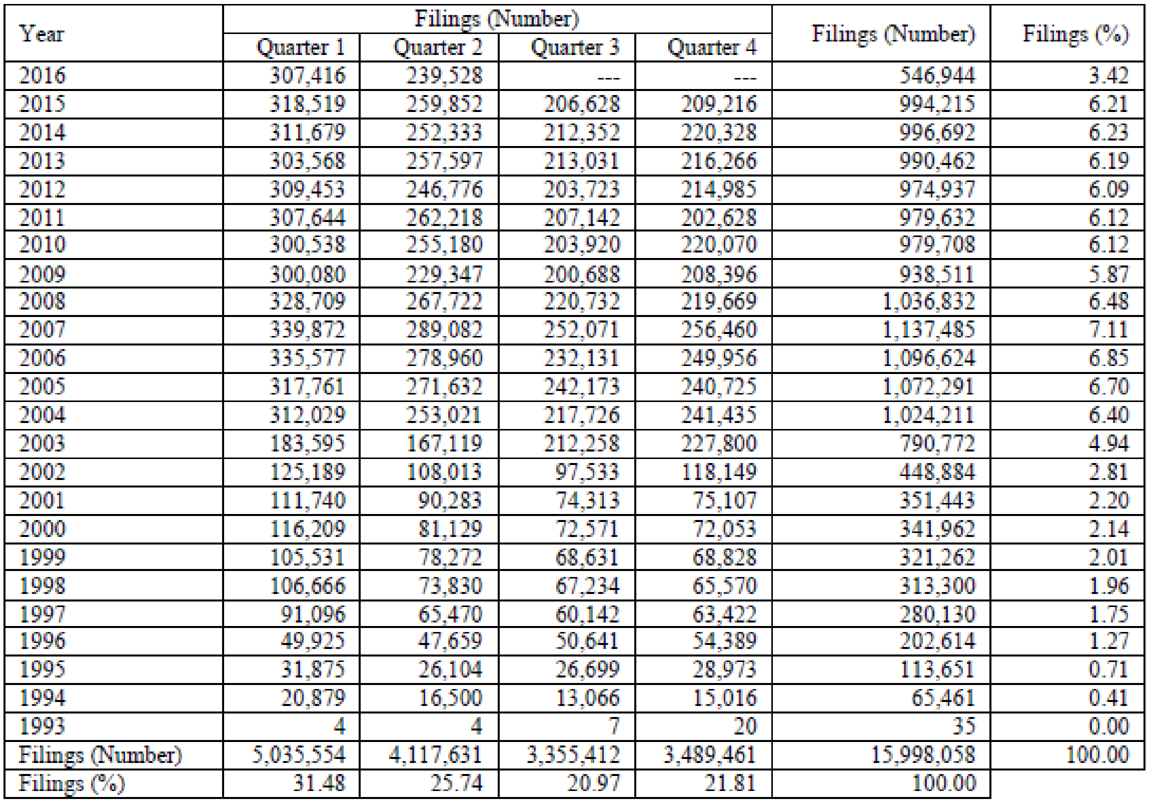
Originally the EDGAR system was developed by the SEC as a pilot system for electronic disclosure in 1983. In order to test and evaluate EDGAR´s performance the SEC requested electronic filings in 1994 after completing the phase-in of a mandated test group in December 1993 (the phase-in began on April 26, 1993) [26] [11] [27]. As of May 6, 1996 the SEC obligated all public domestic U.S. companies (issuers) to file submissions electronically through the EDGAR system [28] [11] [27] [1] except for certain filings made in paper because of a hardship exemption under Regulation S-T [29] [25]. Filing for foreign private issuers (companies organized outside of the U.S.) and foreign governments via EDGAR [26] became mandatory on May 14, 2002 [30]. The Securities Exchange Act of 1934 (Securities Exchange Act 1934, Section 13(a), (b), Section 15(d)) empowers the SEC to require (periodic) reporting of information from publicly held companies [24]. In general, all public domestic companies with assets exceeding $10 million and at least 500 shareholders become subject to Exchange Act reporting requirements (Securities Exchange Act 1934, Section 12(g)) alongside certain individuals [10]. Among other disclosures, corporations with publicly traded securities are required (Securities Exchange Act 1934, Section 13(a), (b), Section 15(d)) to file annual and quarterly reports (Form 10-K, Form 10-Q) as well as current reports (Form 8-K) on an ongoing basis with the SEC and its EDGAR system [24]. Since by law these public corporate disclosures have to be accurate (Securities Exchange Act 1934, Section 13(i)) and represent a company´s operations, they themselves represent a treasure trove of valuable information for investors and researchers [10] [18].

## 2.1. Underlying data in SEC´s EDGAR database

In order to understand the amount and variety of corporate information (e.g. financial statements) distributed by the SEC, I retrieve and analyze all form index files since the implementation of the EDGAR system in 1993. The SEC EDGAR form index files list all publicly available disclosures made through the system in a certain quarter and sort the submissions by their particular filing form type. Table 1 reports the total number of submissions that have been made with the EDGAR system for each quarter and year since the introduction of the EDGAR database.

A tremendous amount of publicly available disclosures was filed with the SEC between 1993 and 2016. In total 15,998,058 filings were submitted to the EDGAR system in order to be publicly distributed. On average 31.48 percent (5,035,554) of these filings became available in the first, 25.74 percent (4,117,631) in the second, 20.97 percent (3,355,412) in the third and 21.81 percent (3,489,461) in the last quarter of each year since 1993. Most noticeable is the overall increase in total submissions through the EDGAR system reaching its peak in 2007 with more than 1.1 million disclosures for that particular year. By analyzing the index files more precisely, investors and researchers can gain an insight into the specific type of information the SEC is making publicly available through its EDGAR system [5]. Table 2 describes the most common filing (form) types filed with the EDGAR system.

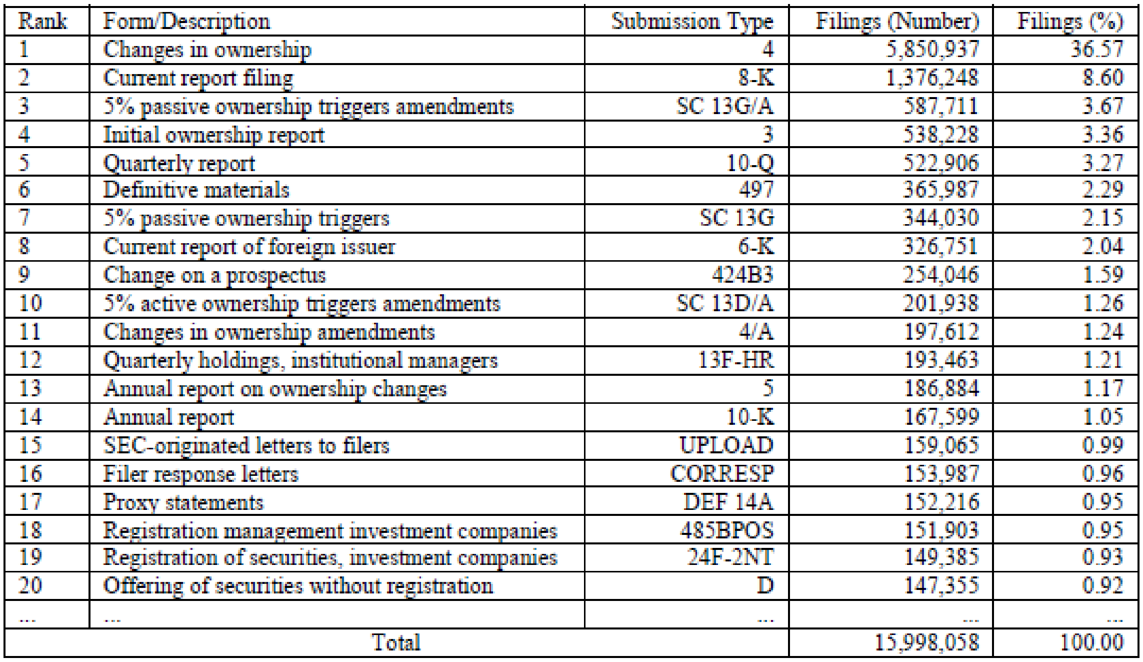
Table 1. Statistics on EDGAR submissions



| Year | Filings (Number) | | | | Filings (Number) | Filings (%) |
|---|---|---|---|---|---|---|
| | Quarter 1 | Quarter 2 | Quarter 3 | Quarter 4 | | |
| 2016 | 307,416 | 239,528 | --- | --- | 546,944 | 3.42 |
| 2015 | 318,519 | 259,852 | 206,628 | 209,216 | 994,215 | 6.21 |
| 2014 | 311,679 | 252,333 | 212,352 | 220,328 | 996,692 | 6.23 |
| 2013 | 303,568 | 257,597 | 213,031 | 216,266 | 990,462 | 6.19 |
| 2012 | 309,453 | 246,776 | 203,723 | 214,985 | 974,937 | 6.09 |
| 2011 | 307,644 | 262,218 | 207,142 | 202,628 | 979,632 | 6.12 |
| 2010 | 300,538 | 255,180 | 203,920 | 220,070 | 979,708 | 6.12 |
| 2009 | 300,080 | 229,347 | 200,688 | 208,396 | 938,511 | 5.87 |
| 2008 | 328,709 | 267,722 | 220,732 | 219,669 | 1,036,832 | 6.48 |
| 2007 | 339,872 | 289,082 | 252,071 | 256,460 | 1,137,485 | 7.11 |
| 2006 | 335,577 | 278,960 | 232,131 | 249,956 | 1,096,624 | 6.85 |
| 2005 | 317,761 | 271,632 | 242,173 | 240,725 | 1,072,291 | 6.70 |
| 2004 | 312,029 | 253,021 | 217,726 | 241,435 | 1,024,211 | 6.40 |
| 2003 | 183,595 | 167,119 | 212,258 | 227,800 | 790,772 | 4.94 |
| 2002 | 125,189 | 108,013 | 97,533 | 118,149 | 448,884 | 2.81 |
| 2001 | 111,740 | 90,283 | 74,313 | 75,107 | 351,443 | 2.20 |
| 2000 | 116,209 | 81,129 | 72,571 | 72,053 | 341,962 | 2.14 |
| 1999 | 105,531 | 78,272 | 68,631 | 68,828 | 321,262 | 2.01 |
| 1998 | 106,666 | 73,830 | 67,234 | 65,570 | 313,300 | 1.96 |
| 1997 | 91,096 | 65,470 | 60,142 | 63,422 | 280,130 | 1.75 |
| 1996 | 49,925 | 47,659 | 50,641 | 54,389 | 202,614 | 1.27 |
| 1995 | 31,875 | 26,104 | 26,699 | 28,973 | 113,651 | 0.71 |
| 1994 | 20,879 | 16,500 | 13,066 | 15,016 | 65,461 | 0.41 |
| 1993 | 4 | 4 | 7 | 20 | 35 | 0.00 |
| Filings (Number) | 5,035,554 | 4,117,631 | 3,355,412 | 3,489,461 | 15,998,058 | 100.00 |
| Filings (%) | 31.48 | 25.74 | 20.97 | 21.81 | 100.00 | |

Notes: The table presents the total number of filings made on EDGAR for each year between 1993 and 2016. Each individual filing in a particular quarter is listed in an associated EDGAR form index file on the SEC server.

Table 2. Statistics on EDGAR form types



| Rank | Form/Description | Submission Type | Filings (Number) | Filings (%) |
|---|---|---|---|---|
| 1 | Changes in ownership | 4 | 5,850,937 | 36.57 |
| 2 | Current report filing | 8-K | 1,376,248 | 8.60 |
| 3 | 5% passive ownership triggers amendments | SC 13G/A | 587,711 | 3.67 |
| 4 | Initial ownership report | 3 | 538,228 | 3.36 |
| 5 | Quarterly report | 10-Q | 522,906 | 3.27 |
| 6 | Definitive materials | 497 | 365,987 | 2.29 |
| 7 | 5% passive ownership triggers | SC 13G | 344,030 | 2.15 |
| 8 | Current report of foreign issuer | 6-K | 326,751 | 2.04 |
| 9 | Change on a prospectus | 424B3 | 254,046 | 1.59 |
| 10 | 5% active ownership triggers amendments | SC 13D/A | 201,938 | 1.26 |
| 11 | Changes in ownership amendments | 4/A | 197,612 | 1.24 |
| 12 | Quarterly holdings, institutional managers | 13F-HR | 193,463 | 1.21 |
| 13 | Annual report on ownership changes | 5 | 186,884 | 1.17 |
| 14 | Annual report | 10-K | 167,599 | 1.05 |
| 15 | SEC-originated letters to filers | UPLOAD | 159,065 | 0.99 |
| 16 | Filer response letters | CORRESP | 153,987 | 0.96 |
| 17 | Proxy statements | DEF 14A | 152,216 | 0.95 |
| 18 | Registration management investment companies | 485BPOS | 151,903 | 0.95 |
| 19 | Registration of securities, investment companies | 24F-2NT | 149,385 | 0.93 |
| 20 | Offering of securities without registration | D | 147,355 | 0.92 |
| ... | ... | ... | ... | ... |
| Total | | | 15,998,058 | 100.00 |

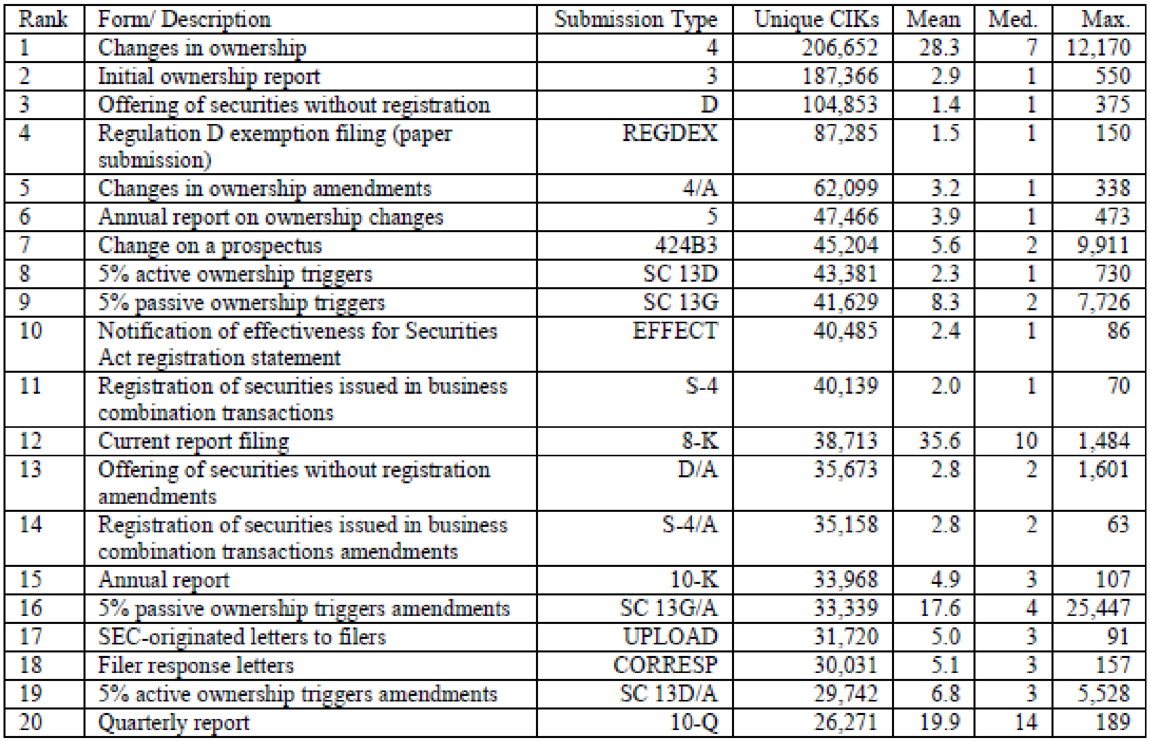
Notes: The table presents the most frequent form types filed with the EDGAR system between 1993 and 2016. The first column ranks each filing type in descending order of total submissions. The second column gives a short description of each filing form type [5]. The third column lists the form codes used on EDGAR to identify a particular filing type made with the database. The next column contains the number of total submissions of a particular filing form type. The last column shows the amount of total submissions for each filing type in relation to all submissions made with the SEC EDGAR database.

The submission type most often filed with the EDGAR system since its implementation is Form 4. Between 1993 and 2016 5,850,937 filings report purchases or sales of securities by persons who are the beneficial owner of more than ten percent of any class of any equity security, or who are directors or officers of the issuer of the security [5]. The second most frequent submission type filed with the SEC is Form 8-K. 1,376,248 filings of this submission type are listed in the EDGAR index files. The current report filing is required by companies in order to inform shareholders about certain corporate events. These events of material importance for a company include information on significant agreements, impairments, changes in management etc. [5]. Important submission types for investors and researchers such as the annual report on Form 10-K have been submitted 167,599 times. Quarterly reports on Form 10-Q have been filed 522,906 times in total between 1993 and 2016. Another important submission type is Schedule 13G (SC 13G). Investors who are not seeking control over a firm (passive investors) must file this submission type as required by the SEC when crossing the five percent ownership threshold of a company [5]. In total 344,030 filings of this particular submission type alone are reported on EDGAR.

The SEC assigns to each filer a Central Index Key (CIK) which is a unique identifier used on the EDGAR database in order to label and identify each individual filer in the system [10]. Since 1993 in total 580,225 unique CIK numbers were assigned and stored in the SEC´s electronic disclosure system. The majority of these CIKs were not assigned to publicly traded companies but to private firms, hedge funds and mutual funds as well as to private individuals who receive a CIK when filing with the SEC [5]. Table 3 reports the number of unique CIKs (unique filers) filing a certain submission type with the SEC and its EDGAR system.

Submission type Form 4 (Form 3) was submitted by 206,652 (187,366) different filers between 1993 and 2016. Annual reports on Form 10-K were submitted to the SEC by 33,968 filers. Quarterly reports on Form 10-Q can be associated with 26,271 unique filers whereas the number of CIKs assigned to current reports on Form 8-K is 38,713. On average each registrant filed 4.9 annual reports on Form 10-K and 19.9 quarterly reports on Form 10 Q with the EDGAR system in addition to 35.6 current reports on Form 8-K since 1993. AFS SenSub Corp. (CIK 1347185), an issuer of asset-backed securities, filed 107 annual reports on Form 10-K (56 on 10-K/A). PowerShares DB Multi-Sector Commodity Trust (CIK 1367306), an investment company offering several investment funds, filed 189 quarterly reports on Form 10-Q (7 on 10-Q/A). Chase Bank USA, National Association (CIK 869090) filed 1,484 Form 8-K statements (12 on 8-K/A). 730 Schedule 13D Forms were filed by Gamco Investors, INC. (CIK 807249), an investment advisory and brokerage service firm, (5,528 on SC 13D/A) whereas FMR LLC (CIK 315066), the financial services conglomerate known as Fidelity Investments, filed 7,726 Schedule 13G Forms (25,447 on SC 13G/A).

Table 3. Statistics on EDGAR filers



| Rank | Form/ Description | Submission Type | Unique CIKs | Mean | Med. | Max. |
|---|---|---|---|---|---|---|
| 1 | Changes in ownership | 4 | 206,652 | 28.3 | 7 | 12,170 |
| 2 | Initial ownership report | 3 | 187,366 | 2.9 | 1 | 550 |
| 3 | Offering of securities without registration | D | 104,853 | 1.4 | 1 | 375 |
| 4 | Regulation D exemption filing (paper submission) | REGDEX | 87,285 | 1.5 | 1 | 150 |
| 5 | Changes in ownership amendments | 4/A | 62,099 | 3.2 | 1 | 338 |
| 6 | Annual report on ownership changes | 5 | 47,466 | 3.9 | 1 | 473 |
| 7 | Change on a prospectus | 424B3 | 45,204 | 5.6 | 2 | 9,911 |
| 8 | 5% active ownership triggers | SC 13D | 43,381 | 2.3 | 1 | 730 |
| 9 | 5% passive ownership triggers | SC 13G | 41,629 | 8.3 | 2 | 7,726 |
| 10 | Notification of effectiveness for Securities Act registration statement | EFFECT | 40,485 | 2.4 | 1 | 86 |
| 11 | Registration of securities issued in business combination transactions | S-4 | 40,139 | 2.0 | 1 | 70 |
| 12 | Current report filing | 8-K | 38,713 | 35.6 | 10 | 1,484 |
| 13 | Offering of securities without registration amendments | D/A | 35,673 | 2.8 | 2 | 1,601 |
| 14 | Registration of securities issued in business combination transactions amendments | S-4/A | 35,158 | 2.8 | 2 | 63 |
| 15 | Annual report | 10-K | 33,968 | 4.9 | 3 | 107 |
| 16 | 5% passive ownership triggers amendments | SC 13G/A | 33,339 | 17.6 | 4 | 25,447 |
| 17 | SEC-originated letters to filers | UPLOAD | 31,720 | 5.0 | 3 | 91 |
| 18 | Filer response letters | CORRESP | 30,031 | 5.1 | 3 | 157 |
| 19 | 5% active ownership triggers amendments | SC 13D/A | 29,742 | 6.8 | 3 | 5,528 |
| 20 | Quarterly report | 10-Q | 26,271 | 19.9 | 14 | 189 |

Notes: The table presents the most frequent submission types made on EDGAR in descending order of unique SEC registrants filing a particular submission type. The time period is 1993-2016. The fourth column contains the total number of unique filers submitting a particular form type. Columns 5-7 present the means, medians and maxima of particular filing form types submitted by unique SEC filers

## 3. SEC EDGAR DATA GATHERING

Researchers in the field of finance and accounting often rely on programming languages (Perl, Python, R, SAS, and SPSS) to retrieve financial statements filed with the SEC. The use of a programming language as a tool is problematic for several reasons. First, many people analyzing financial reports are not familiar with these programming languages. For them it is time-consuming to apply a specific and complex coding language to obtain the corporate filings from EDGAR. Secondly, due to downloading only one filing at a time the procedure is very slow especially when obtaining massive data from the database. Thirdly, since incremental changes have to be made to the algorithm to retrieve another filing form type or filings from another company this particular method is very error-prone.

In contrast, widely used internet browsers (e.g. Mozilla-Firefox, Google-Chrome) can be easily equipped with powerful applications (e.g. DownThemAll, GetThemAll) which offer advanced download capabilities. These fully integrated browser extensions are able to identify links contained in a webpage or file and download the desired document parts simultaneously. To feed these applications only a standard MS Excel spreadsheet is necessary.Every filing made through the EDGAR system in a particular quarter between 1993 and 2016 is stored in an associated index file (file extension *.idx) [5]. The EDGAR index files therefore represent a helpful resource in retrieving massive data from the database. They list important information for each filing such as the name of the filer, the particular central index key, the date and the type of the submission as well as the particular name of the document on the SEC server. In general, four different types of index files are available sorting the filings made on EDGAR by company name, form type, central index key or by submissions containing financial statements formatted in eXtensible Business Reporting Language (XBRL) [31] [32]. When examining the form index files more precisely one can see that the index files do not only contain the name of any filing made on EDGAR but rather the (entire) server path. Table 4 illustrates an excerpt of information stated in the SEC EDGAR form index file from the first quarter of 2016. By opening the index files for example with a simple MS Excel spreadsheet (file extension *.xlsx) a Uniform Resource Locator (URL) can be created for each financial statement which is listed in a particular index file since the name of the filing and its (partial) server path (directory) is stated. To do so the protocol (https://), the hostname (www.sec.gov/) and a link to the archives directory (Archives/) have to be added to the file name from the index file. Table 5 illustrates the URL components of Coca Cola´s 2015 annual report on Form 10-K filed with the SEC on February 25, 2016. These URLs which have been composed based on the EDGAR index files can be copied into a plain text file (file extension *.txt). By opening it with the browser extensive data (financial statements) can be retrieved from the SEC and its EDGAR system in a fast and efficient way using a browser extension (however, the composed URLs can also be implemented in any other data gathering method).

This method offers various significant advantages. First, for many people composing URLs with commonly used and easy accessible computer software like MS Excel is simpler and faster than relying on complex coding languages to identify and retrieve the documents in question. Secondly, since multiple documents can be retrieved at the same time using browser extensions, the described method is again a lot faster especially when obtaining massive data from EDGAR. Thirdly, by sorting or filtering the different index files in MS Excel the proposed method can easily be adjusted to retrieve another filing form type or data from another company. The result of this procedure is validated through obtaining exactly the same financial statements investors and researchers would retrieve using a complex, slow and error-prone alternative.

## 4. HYPERTEXT MARKUP LANGUAGE IN SEC FILINGS

Because financial statements filed with the SEC are formatted in HyperText Markup Language (HTML) the fundamentals of HTML are illustrated first, followed by an examination of the data formatted in HTML provided by the SEC and its EDGAR system.

### 4.1. Fundamentals of HyperText Markup Language

HyperText Markup Language (HTML) is a universally understood digital language which is used to publish and distribute information globally. HTML is the publishing language of the World Wide Web [33]. HTML is used to create HyperText documents that are portable from one platform to another [34] due to their generic semantics as a Standard Generalized Markup Language (SGML) application [33]. HTML enables authors to publish documents online, assign a specific look or layout to document content (tagging) [35] [21] or to retrieve information online via HyperText links [33]. The World Wide Web Consortium (W3C) is maintaining and specifying the vocabulary (applicable markups) and grammar (logical structure) of HTML documents [35].

A valid HTML document is composed of three different parts [33]. First, it declares which version of HTML is used in the document through the document type declaration (<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">). The document type declaration names the document type definition (DTD) specifying which elements and attributes can be implemented into a document formatted in HTML [33]. HTML 4.01 specifies three different DTDs: HTML 4.01 Strict DTD; HTML 4.01 Transitional DTD and HTML 4.01 Frameset DTD [33]. The W3C recommends to use HTML 4.01 Strict DTD which excludes presentation attributes since these elements are supposed to be replaced by style sheets [36]. The second part of a HTML document is the document head (<HEAD>). This section contains information about the current document such as the title and relevant keywords for search engines. In general, the elements appearing in the head section are not presented by a document formatted in HTML [33]. The third and most important part of a HTML document is the body (<BODY>). This section contains the actual content of the document such as text paragraphs, images, graphics, tables, links, etc. [33]. The content in the document body can be structured in many different ways using various HTML elements (tags) to accomplish a certain look or layout to present the embedded information.

### 4.2. SEC EDGAR HTML Data

“Official” financial statements filed with the SEC have to be formatted either in American Standard Code for Information Interchange (ASCII) or in HyperText Markup Language (HTML 3.2/4.0). Financial statements formatted in Portable Document Format (PDF) or XBRL are considered “unofficial” documents (submissions formatted in PDF and XBRL may qualify as official documents as well when specific criteria are met) [34]. Due to a limited support of HTML in order to reduce the number of inconsistencies caused by HTML 4.0 implementation variances [37], the EDGAR system only accepts a subset of HTML 3.2 semantics (tags) and several HTML 4.0 attributes [34] therefore enforcing several restrictions (no active content, no external references etc.) of HTML formatting in financial statement submissions [34].The “Complete Submission Text File” (file extension *.txt) provided by the EDGAR system represents an aggregation of all information in a particular financial statement filed with the SEC. The text version of the filings on the SEC server contains the 10-K document formatted in HTML, XBRL, exhibits and ASCII-encoded graphics (“binary-to-text” encoding or “uuencoding” converts binary data files to plain ASCII-printable characters to facilitate transfer across various hardware platforms) [38] [39]. Besides the “Complete Submission Text File” several submission parts (documents) are also provided in HTML (file extension *.htm) such as the core 10-K document and the exhibits which have been submitted [38]. For example, Coca Cola´s 10-K filing on February 25, 2016 lists the core 10-K filing in HTML format, ten exhibits, eight graphic files (file extension *.jpg), six XBRL files and a single “Complete Submission Text File” containing all of these documents [40].

## 5. TEXTUAL INFORMATION IN FINANCIAL STATEMENTS

This section describes how regular expressions are used to extract textual information from financial statements filed with the SEC. First, I illustrate the fundamentals of regular expressions. Then I discuss the algorithm to extract textual information from financial statements using only regular expressions before presenting the actual text embedded in financial statements as a result of the designed algorithm. Due to their high relevance for investors and researchers an actual annual report on Form 10-K from the Coca Cola Company serves as basis for the illustration

## 5.1. Fundamentals of Regular Expressions

Regular expressions or regular sets were first used as an algebra by mathematicians to describe models developed by physiologists of how the nervous system would work at the neuron level. The first published computational use of regular expressions was in 1968 by Ken Thompson [41] who describes regular expressions as "a method for locating specific character strings embedded in character text" [42]. They are implemented not only in modern programming languages, but also in application programs that can be used for text analysis without special programming skills (e.g. RapidMiner).

Regular expressions ("RegEx"; "RegExp"; "RegExes") with a general pattern notation (pattern language) allow to process all kinds of text and data in a flexible and efficient way [41] [13]. In particular RegExes can be used to modify textual elements or to identify and extract certain information from different documents [43]. The two types (full) regular expressions are composed of are special characters (metacharacters) and normal (literal) text characters acting as the grammar and the words of the regular expression language [41] [43]. For example, RegEx: "[0-9]" identifies all digits, RegEx: "[a-zA-Z]" isolates all upper and lower-case letters (character classes) and RegEx: "." matches all of these elements (metacharacter) embedded in an underlying text document [41] [43]. Another metacharacter and counting element (quantifier) within the regular expression language is a star or an asterisk (*) which quantifies the immediately preceding item within the defined expression (match any number of the preceding element including none) [41] [43]. Counting elements or quantifiers are used to specify the search pattern of regular expressions in more detail. "Greedy" quantifiers like "*" match as much as possible whereas "lazy" quantifiers such as "*?" match as little as possible to satisfy the search pattern of a composed regular expression [41] [43].

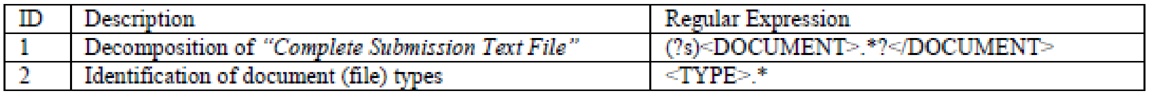
In addition, regular expressions can be modified in the way they are interpreted and applied using different regular expression modes (modifiers). These modifiers allow to change the search pattern of a particular regular expression (matching mode) in modern programming languages or in application programs. Regular expressions equipped with "case-insensitive match mode" ((?i)) ignore the letter case of the input (textual elements) during the matching process allowing the search pattern to match both upper and lower case letters [41] [43]. Since modern applications work with multiple (coding) lines regular expressions need to be modified in order to match a string across different lines. "Dot-matches-all match mode" also known as "single-line mode" ((?s)) modifies the search pattern of a regular expression in a way that it matches a character string across multiple lines [41] [43]. By designing regular expressions and implementing them into modern computer software the results of various search patterns (textual information) can be highlighted and changed or even removed from the underlying text at all [41] [43].

## 5.2. Extraction of Textual Information

Researchers in the field of finance and accounting (as well as business data providers) use the "Complete Submission Text Files" (file extension *.txt) provided by the SEC and its EDGAR system to extract textual information from financial statements. In order to delete all non-textual elements (HTML tags and their corresponding attributes) most often special text-processing programs and their predefined applications (HTML Parser) are used. This again is problematic for several reasons. First, using predefined text-processing operators to delete non-textual elements makes one platform-dependent since a specific HTML Parser can not be (easily) implemented into any other text-processing program in use. Secondly, since the extraction algorithm of the HTML-Parser is complex or not presented at all its extraction results can hardly be validated. Thirdly, because of these drawbacks empirical research results are challenging to replicate for a particular or any other data sample. Regular expressions can in fact overcome these problems in extracting textual information embedded in financial statements filed with the SEC. They offer platform-independent (research) results which can be validated and replicated for any data sample at any given time.

The proposed extraction algorithm ("Annual Report Algorithm") first decomposes the "Complete Submission Text File" (file extension *.txt) into its components (RegEx 1). In the end, the entire algorithm is validated through obtaining exactly one core (Form 10-K) document and the number of exhibits which have been embedded in the "Complete Submission Text File" for every financial statement in the data sample. Next, the "Annual Report Algorithm" identifies all other file types contained in the submission since these additional documents are not either a core document or an exhibit within the text version of the filing (RegEx 2). Table 4 illustrates the regular expressions needed to decompose the "Complete Submission Text File" of a financial statement filed with the SEC and to identify the embedded document (file) types.

Table 4. Regular expressions contained in the "Annual Report Algorithm"

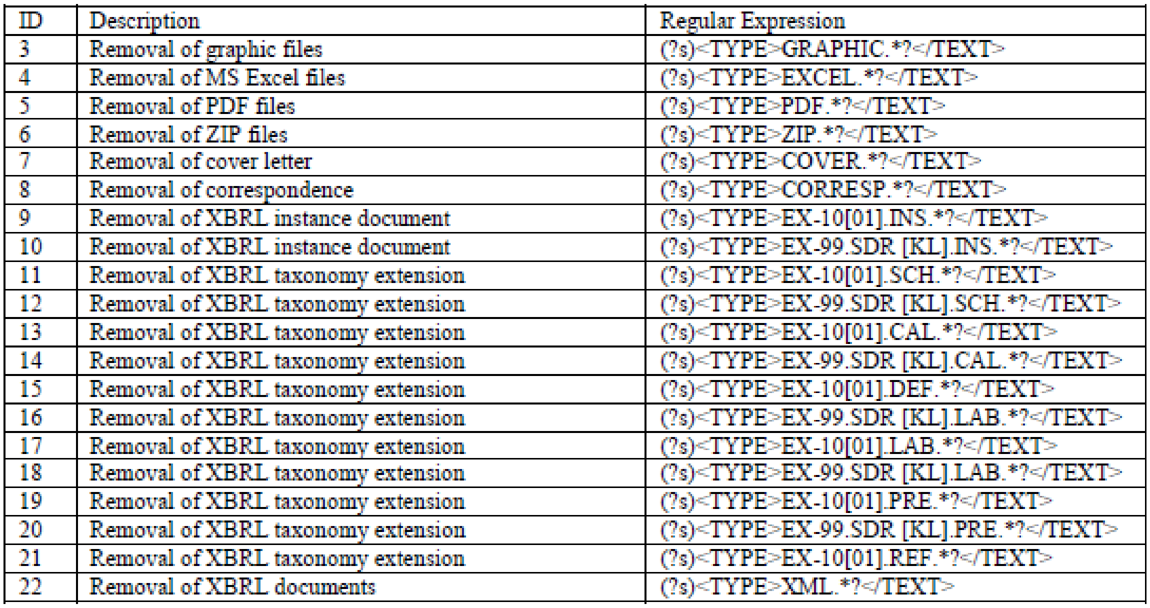


| ID | Description | Regular Expression |
|---|---|---|
| 1 | Decomposition of *"Complete Submission Text File"* | (?s)<DOCUMENT>.*?</DOCUMENT> |
| 2 | Identification of document (file) types | <TYPE>.* |

Notes: The table presents the regular expressions contained in the "Annual Report Algorithm" for extracting documents and identifying document (file) types.

In addition to the filing components described earlier (10-K section, exhibits, XBRL, graphics), several other document (file) types might be embedded in financial statements such as MS Excel files (file extension *.xlsx), ZIP files (file extension *.zip) and encoded PDF files (file extension *.pdf). By applying additional rules in the "Annual Report Algorithm" (RegExes 3-22) these documents are deleted to be able to extract textual information only from the core document and the various exhibits contained in the "Complete Submission Text File". The additional SEC-header is not supposed to be removed separately since it has already been deleted by the algorithm. Table 5 illustrates the regular expressions applied to delete document (file) types other than the core document and the corresponding exhibits.

Table 5. Regular expressions contained in the "Annual Report Algorithm"

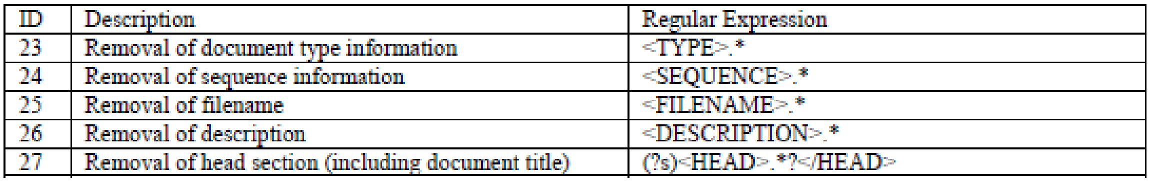


| ID | Description | Regular Expression |
|---|---|---|
| 3 | Removal of graphic files | (?s)<TYPE>GRAPHIC.*?</TEXT> |
| 4 | Removal of MS Excel files | (?s)<TYPE>EXCEL.*?</TEXT> |
| 5 | Removal of PDF files | (?s)<TYPE>PDF.*?</TEXT> |
| 6 | Removal of ZIP files | (?s)<TYPE>ZIP.*?</TEXT> |
| 7 | Removal of cover letter | (?s)<TYPE>COVER.*?</TEXT> |
| 8 | Removal of correspondence | (?s)<TYPE>CORRESP.*?</TEXT> |
| 9 | Removal of XBRL instance document | (?s)<TYPE>EX-10[01].INS.*?</TEXT> |
| 10 | Removal of XBRL instance document | (?s)<TYPE>EX-99.SDR [KL].INS.*?</TEXT> |
| 11 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-10[01].SCH.*?</TEXT> |
| 12 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-99.SDR [KL].SCH.*?</TEXT> |
| 13 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-10[01].CAL.*?</TEXT> |
| 14 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-99.SDR [KL].CAL.*?</TEXT> |
| 15 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-10[01].DEF.*?</TEXT> |
| 16 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-99.SDR [KL].LAB.*?</TEXT> |
| 17 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-10[01].LAB.*?</TEXT> |
| 18 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-99.SDR [KL].LAB.*?</TEXT> |
| 19 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-10[01].PRE.*?</TEXT> |
| 20 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-99.SDR [KL].PRE.*?</TEXT> |
| 21 | Removal of XBRL taxonomy extension | (?s)<TYPE>EX-10[01].REF.*?</TEXT> |
| 22 | Removal of XBRL documents | (?s)<TYPE>XML.*?</TEXT> |

Notes: The table presents the regular expressions contained in the "Annual Report Algorithm" for deleting nonrelevant document (file) types.

Next, the "Annual Report Algorithm" deletes all metadata included in the core document and the exhibits (RegExes 23-27). Table 6 illustrates the regular expressions for deleting metadata in SEC EDGAR documents.

Table 6. Regular expressions contained in the "Annual Report Algorithm"

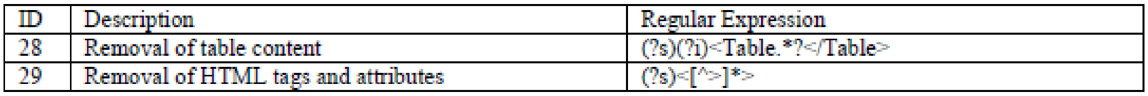


| ID | Description | Regular Expression |
|---|---|---|
| 23 | Removal of document type information | <TYPE>.* |
| 24 | Removal of sequence information | <SEQUENCE>.* |
| 25 | Removal of filename | <FILENAME>.* |
| 26 | Removal of description | <DESCRIPTION>.* |
| 27 | Removal of head section (including document title) | (?s)<HEAD>.*?</HEAD> |

Notes: The table presents the regular expressions contained in the "Annual Report Algorithm" for deleting nonrelevant document metadata.

Before deleting all HTML elements and their corresponding attributes (RegEx 29) the algorithm deletes tables since they contain non-textual (quantitative) information (RegEx 28). Table 7 illustrates the set of regular expressions applied to delete tables and HTML elements embedded in financial statements filed with the SEC.

Table 7. Regular expressions contained in the "Annual Report Algorithm"



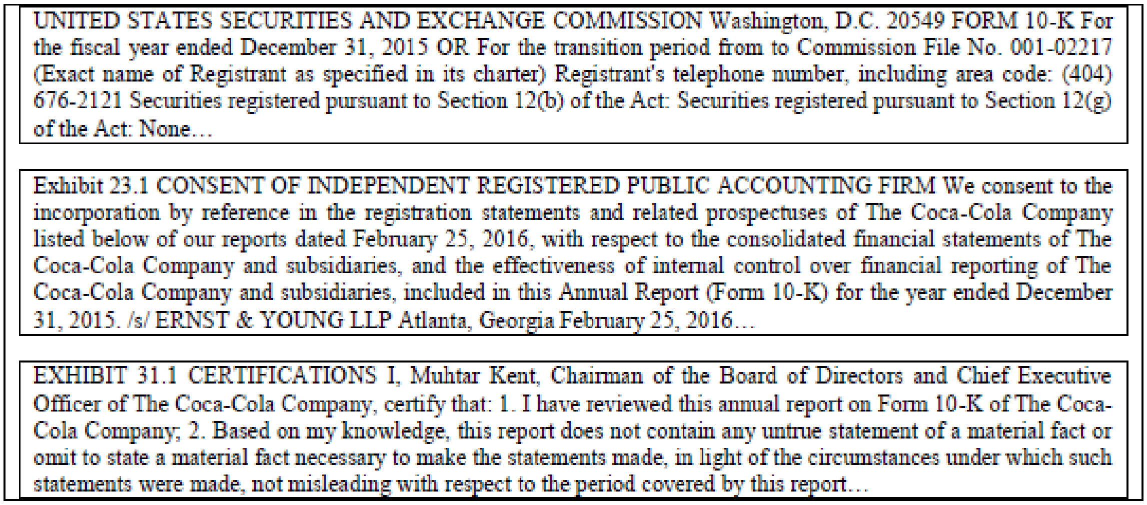
| ID | Description | Regular Expression |
|---|---|---|
| 28 | Removal of table content | (?s)(?i)<Table.*?</Table> |
| 29 | Removal of HTML tags and attributes | (?s)<[^>]*> |

Notes: The table presents the regular expressions contained in the "Annual Report Algorithm" for deleting tables and HTML elements

After extracting the core document and the exhibits as well as deleting all HTML elements, the "Annual Report Algorithm" adjusts the content embedded in the body section of each HTMLformatted document in order to extract textual elements from financial statements on the EDGAR database. According to the SEC filer manual the EDGAR system suspends financial statements which contain extended ASCII characters. However, it supports submissions with extended character references. By using ISO-8859-1/Latin-1 decimal character references or entity-names (either technique is allowed within SEC submissions) extended ASCII characters can be embedded in financial statement submissions. These extended character sets within HTML documents included in the "Complete Submission Text File" need to be decoded to be able to extract human-readable textual information from financial statements [34]. The "Annual Report Algorithm" finally decodes all extended character sets (RegExes 30-680) most likely embedded in financial statements filed with the SEC and its EDGAR system formatted in HTML 4.01 (ASCII, ANSI/Windows-1252, ISO-8859-1/Latin-1, mathematical, Greek, symbolic and special characters).

## 5.3. Extraction Results

By applying the "Annual Report Algorithm" investors and researchers are able to extract textual information from financial statements filed with the SEC for thousands of companies in a fully automated process. Based on the "Complete Submission Text File" provided by the EDGAR system the algorithm extracts the core (Form 10-K) document and the exhibits which have been embedded in the text version of a company´s financial statement. For example for Coca Cola´s 2015 annual report on Form 10-K filed on February 25, 2016 via EDGAR the algorithm extracts one core document in addition to ten different exhibits. Figure 1 illustrates partial extraction results for the 10-K section of the annual report as well as for two exhibits.



UNITED STATES SECURITIES AND EXCHANGE COMMISSION Washington, D.C. 20549 FORM 10-K For the fiscal year ended December 31, 2015 OR For the transition period from to Commission File No. 001-02217 (Exact name of Registrant as specified in its charter) Registrant's telephone number, including area code: (404) 676-2121 Securities registered pursuant to Section 12(b) of the Act: Securities registered pursuant to Section 12(g) of the Act: None…

Exhibit 23.1 CONSENT OF INDEPENDENT REGISTERED PUBLIC ACCOUNTING FIRM We consent to the incorporation by reference in the registration statements and related prospectuses of The Coca-Cola Company listed below of our reports dated February 25, 2016, with respect to the consolidated financial statements of The Coca-Cola Company and subsidiaries, and the effectiveness of internal control over financial reporting of The Coca-Cola Company and subsidiaries, included in this Annual Report (Form 10-K) for the year ended December 31, 2015. /s/ ERNST & YOUNG LLP Atlanta, Georgia February 25, 2016…

EXHIBIT 31.1 CERTIFICATIONS I, Muhtar Kent, Chairman of the Board of Directors and Chief Executive Officer of The Coca-Cola Company, certify that: 1. I have reviewed this annual report on Form 10-K of The Coca-Cola Company; 2. Based on my knowledge, this report does not contain any untrue statement of a material fact or omit to state a material fact necessary to make the statements made, in light of the circumstances under which such statements were made, not misleading with respect to the period covered by this report…

Figure 1. Examples of the extraction result of the “Annual Report Algorithm“

Notes: The figure presents extraction results from Coca Cola´s 2015 annual report on Form 10-K filed with the SEC. The first part of the figure displays the actual 10-K section embedded in text version of the submission. The second part shows the statement of the auditing firm. The certification of the annual report by the CEO is presented in the last part of the figure.

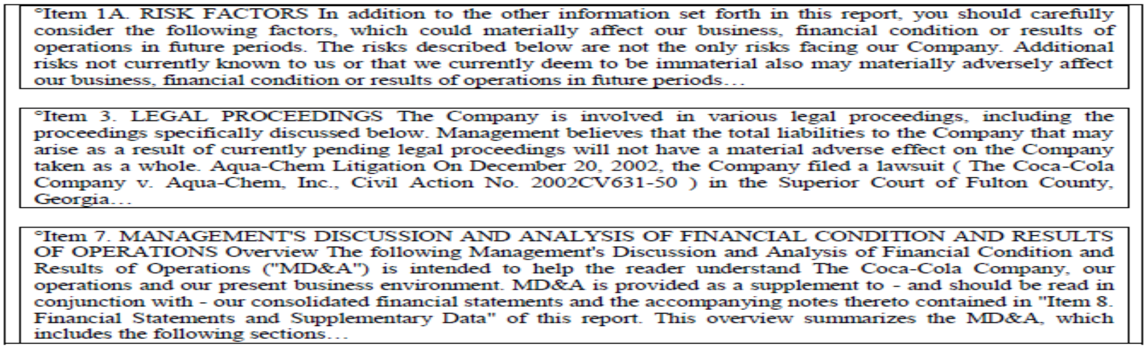
.Besides from textual content of entire documents (10-K section and exhibits) contained in the “Complete Submission Text File” investors and researchers might be interested in extracting textual information from particular sections (Items) within the core 10-K section of an annual report (like Item 1A - Risk Factors; Item 3 - Legal Proceedings; Item 7 - Management’s Discussion and Analysis of Financial Condition and Results of Operations etc.). In order to extract textual information from particular 10-K items the “Annual Report Algorithm” is modified to the “Items Algorithm”. Excluding all exhibits, the modified “Items Algorithm” isolates only the 10 K section within the SEC submission. After deleting nonrelevant information and decoding reserved characters within the document investors and researchers can extract textual information from specific 10-K items. Table 8 specifies the modified “Items Algorithm” applied to extract textual information from particular items of the annual report on Form 10-K filed with the SEC.

Using only regular expressions to extract textual information from financial statements investors and researchers can implement the designed extraction algorithms in any modern application and computer program available today. By applying either the “Annual Report Algorithm” or the “Items Algorithm” entire documents (10-K section and exhibits) or particular items from the core 10-K section can be extracted from the annual SEC submissions in order to be analyzed. More importantly, while compensating for expensive commercial products the algorithms and their extraction results can be validated and replicated for any data sample at any given time. Figure 2 finally illustrates several extraction results of the “Items Algorithm” from the annual report on Form 10 K highly relevant to investors and researchers alike.

Table 8. Regular expressions contained in the “Items Algorithm”

| ID | Description | | Regular Expression |
|---|---|---|---|
| 1.1 | Extraction of 10-K section | | (?s)<TYPE>10-K.*?</TEXT> |
| 2.1 | Removal of document metadata | | RegExes 23-28 |
| 3.1 | Removal of table content | | (?s)(?i)<Table.*?</Table> |
| 4.1 | Decoding of reserved characters | | See RegExes 30-680 |
| 5.1 | Identification and renaming of item headings (“>°Item”) | | (?s)(?i)(?m)> +Item\|>Item\|^Item |
| 6.1 | Removal of multiple empty spaces | | (?s) + |
| 7.1 | Extraction of Item 1. | - Business | (?s)(?i)°Item 1[^AB012345].*?°Item |
| 7.2 | Extraction of Item 1A. | - Risk Factors | (?s)(?i)°Item 1A.*?°Item |
| 7.3 | Extraction of Item 1B. | - Unresolved Staff Comments | (?s)(?i)°Item 1B.*?°Item |
| 7.4 | Extraction of Item 2. | - Properties | (?s)(?i)°Item 2.*?°Item |
| 7.5 | Extraction of Item 3. | - Legal Proceedings | (?s)(?i)°Item 3.*?°Item |
| 7.6 | Extraction of Item 4. | - Mine Safety Disclosures | (?s)(?i)°Item 4.*?°Item |
| 7.7 | Extraction of Item X. | - Executive Officers of the Company | (?s)(?i)°Item X.*?°Item |
| 7.8 | Extraction of Item 5. | - Market for Registrant’s Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities | (?s)(?i)°Item 5.*?°Item |
| 7.9 | Extraction of Item 6. | - Selected Financial Data | (?s)(?i)°Item 6.*?°Item |
| 7.10 | Extraction of Item 7. | - Management’s Discussion and Analysis of Financial Condition and Results of Operations | (?s)(?i)°Item 7[^A].*?°Item |
| 7.11 | Extraction of Item 7A. | - Quantitative and Qualitative Disclosures About Market Risk | (?s)(?i)°Item 7A.*?°Item |
| 7.12 | Extraction of Item 8. | - Financial Statements and Supplementary Data | (?s)(?i)°Item 8.*?°Item |
| 7.13 | Extraction of Item 9. | - Changes in and Disagreements with Accountants on Accounting and Financial Disclosure | (?s)(?i)°Item 9[^AB].*?°Item |
| 7.14 | Extraction of Item 9A. | - Controls and Procedures | (?s)(?i)°Item 9A.*?°Item |
| 7.15 | Extraction of Item 9B. | - Other Information | (?s)(?i)°Item 9B.*?°Item |
| 7.16 | Extraction of Item 10. | - Directors, Executive Officers and Corporate Governance | (?s)(?i)°Item 10.*?°Item |
| 7.17 | Extraction of Item 11. | - Executive Compensation | (?s)(?i)°Item 11.*?°Item |
| 7.18 | Extraction of Item 12. | - Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters | (?s)(?i)°Item 12.*?°Item |
| 7.19 | Extraction of Item 13. | - Certain Relationships and Related Transactions, and Director Independence | (?s)(?i)°Item 13.*?°Item |
| 7.20 | Extraction of Item 14. | - Principal Accounting Fees and Services | (?s)(?i)°Item 14.*?(°Item\|</TEXT>) |
| 7.21 | Extraction of Item 15. | - Exhibits, Financial Statement Schedules | (?s)(?i)°Item 15[^°]*?</TEXT> |
| 8.1 | Removal of HTML tags and attributes | | (?s)<[^>]*> |

Notes: The table presents the regular expressions contained in the modified “Items Algorithm” for extracting particular items from the annual report on Form 10-K. RegExes 1.1-6.1 modify the text version of a financial statement to be able to extract (clear) textual information from particular items. RegExes 7.1-7.21 represent the actual regular expressions designed to extract particular sections from the text version of the annual report.



°Item 1A. RISK FACTORS In addition to the other information set forth in this report, you should carefully consider the following factors, which could materially affect our business, financial condition or results of operations in future periods. The risks described below are not the only risks facing our Company. Additional risks not currently known to us or that we currently deem to be immaterial also may materially adversely affect our business, financial condition or results of operations in future periods…

°Item 3. LEGAL PROCEEDINGS The Company is involved in various legal proceedings, including the proceedings specifically discussed below. Management believes that the total liabilities to the Company that may arise as a result of currently pending legal proceedings will not have a material adverse effect on the Company taken as a whole. Aqua-Chem Litigation On December 20, 2002, the Company filed a lawsuit ( The Coca-Cola Company v. Aqua-Chem, Inc., Civil Action No. 2002CV631-50 ) in the Superior Court of Fulton County, Georgia…

°Item 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS Overview The following Management's Discussion and Analysis of Financial Condition and Results of Operations ("MD&A") is intended to help the reader understand The Coca-Cola Company, our operations and our present business environment. MD&A is provided as a supplement to - and should be read in conjunction with - our consolidated financial statements and the accompanying notes thereto contained in "Item 8. Financial Statements and Supplementary Data" of this report. This overview summarizes the MD&A, which includes the following sections…

Figure 2. Examples of the extraction result of the “Items Algorithm”

Notes: The figure presents extraction results from Coca Cola´s 2015 annual report on Form 10-K filed with the SEC. The first part of the figure displays Item 1A (Risk Factors) embedded in the overall 10-K section.

The last two parts of the figure show Item 3 (Legal Proceedings) and Item 7 (Management´s Discussion and Analysis of Financial Condition and Results of Operations) contained in the 10-K section of the "Complete Submission Text File"

## 6. VALIDATION OF EXTRACTION ALGORITHMS

In order to validate the proposed extraction algorithms and to test their capabilities, I retrieve all Form 10-K filings listed in the SEC EDGAR form index files. Using the data gathering method as described in Section 3 in total 188,875 annual reports (167,599 on Form 10-K and 21,276 on Form 10-K405 ) filed between 1993 and 2016 are retrieved from the EDGAR database (SEC EDGAR Form 10-K types as used in Loughran and McDonald 2011a). The "Annual Report Algorithm" is applied to all submissions to derive different word counts for each filing made with the SEC. In addition to the overall word count of an annual report, for each core document (10-K section) and the exhibits embedded in a "Complete Submission Text File" an individual word count is retrieved in order to be compared (XBRL files declared as exhibits are deleted). Figure 3 illustrates how word counts for each filing and its components are obtained from the "Complete Submission Text File" for the document validation process of the "Annual Report Algorithm".

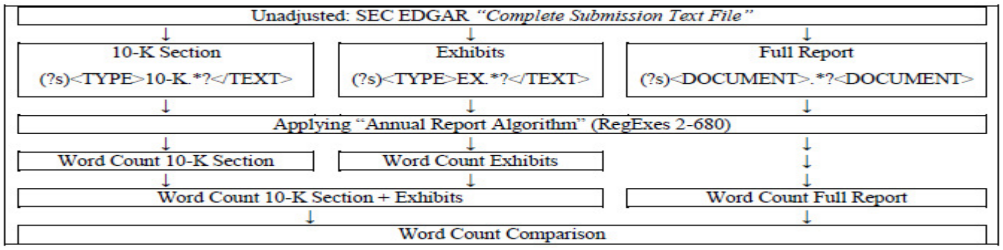

Unadjusted: SEC EDGAR *"Complete Submission Text File"*
10-K Section
(?s)<TYPE>10-K.*?</TEXT>
Exhibits
(?s)<TYPE>EX.*?</TEXT>
Full Report
(?s)<DOCUMENT>.*?<DOCUMENT>
Applying "Annual Report Algorithm" (RegExes 2-680)
Word Count 10-K Section
Word Count Exhibits
Word Count 10-K Section + Exhibits
Word Count Full Report
Word Count Comparison

Figure 3. Document validation process of the "Annual Report Algorithm".

Notes: The figure presents the document validation process of the "Annual Report Algorithm". The "Complete Submission Text File" of each financial statement as provided on the SEC server is used to extract all relevant components (documents). The "Annual Report Algorithm" is applied to each filing in order to retrieve word counts for all relevant documents embedded in the submission. The word count of all relevant documents is compared with the overall length of the submission. A mismatch between the word counts would indicate that the entire report contains nonrelevant document (file) types after applying the "Annual Report Algorithm".

This word count comparison between the overall report on full length and its different components cannot be a validation of the "Annual Report Algorithm" since the same algorithm is simply applied to different sets of textual information (10-K section, exhibits, full report). However, if the entire report would still contain document (file) types or elements which are not a part of the core 10-K section or a corresponding exhibit the word count of a certain financial statement would be artificially increased (Word Count Full Report). In fact, the ability to validate the entire extraction procedure by applying an alternative to the "Annual Report Algorithm" (e.g. HTML-Parser) is limited since to a certain extent the same regular expressions have to be used to create the input for both extraction methods in the first place (extracting core 10-K document and exhibits, deleting nonrelevant document (file) types etc.). Due to this disability in validating the entire extraction process from the beginning by applying an HTML-Parser one has to validate the input the proposed algorithm is creating and its extraction results separately, therefore validating the entire information extraction process. The validation of the textual input created by the "Annual Report Algorithm" is represented by the extraction algorithm itself since it uses only regular expressions combined with the electronic filing requirements introduced by the SEC (precisely not the SEC but Attain, LLC). According to the SEC, all documents embedded in a "Complete Submission Text File" must be equipped with a

"<TYPE>" tag representing the conformed document type of that particular submission part within the text version of the filing (<TYPE>10-K, <TYPE>10-Q, <TYPE>8-K, <TYPE>EX-1, <TYPE>EX-2 etc.) [45]. The "Annual Report Algorithm" (RegExes 1-29) uses these requirements in order to extract the core document and the corresponding exhibits from annual reports while deleting all documents associated with XBRL and other document (file) types. The search patterns of the "Annual Report Algorithm" which have been designed accordingly to the filing requirements of the SEC can be validated due to the general pattern notation of the regular expression language.

An output comparison between the "Annual Report Algorithm" and a common HTML-Parser shall serve as an additional validation for the remaining extraction procedure. Therefore, I modify the "Complete Submission Text Files" as provided by the SEC (unadjusted filings) and apply the first part of the "Annual Report Algorithm" (RegExes 1-29) in order to make the text version of the financial statements readable for the predefined HTML-Parser (adjusted filings). Since this part of the overall validation process focuses on how well the "Annual Report Algorithm" is capable of decoding escape sequences embedded in a "Complete Submission Text File" the aggregated text length of both procedures are compared rather than the word counts due to decimal character encodings (a simple word count comparison would not fully capture the disability of the "Annual Report Algorithm" in decoding these character references in relation to the HTML-Parser). Figure 4 illustrates the output validation process of the "Annual Report Algorithm".

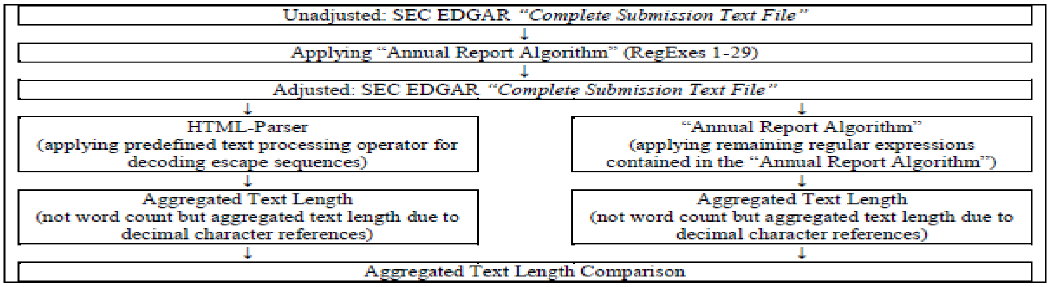

Unadjusted: SEC EDGAR *"Complete Submission Text File"*
Applying "Annual Report Algorithm" (RegExes 1-29)
Adjusted: SEC EDGAR *"Complete Submission Text File"*
HTML-Parser (applying predefined text processing operator for decoding escape sequences)
"Annual Report Algorithm" (applying remaining regular expressions contained in the "Annual Report Algorithm")
Aggregated Text Length (not word count but aggregated text length due to decimal character references)
Aggregated Text Length (not word count but aggregated text length due to decimal character references)
Aggregated Text Length Comparison

Figure 4. Output validation process of the „Annual Report Algorithm"

Notes: The figure presents the output validation process of the "Annual Report Algorithm". The "Complete Submission Text File" of each financial statement as provided on the SEC server is adjusted in order to compare the output of the algorithm with the output a common HTMLParser would produce. RegExes 1-29 modify the unadjusted document as provided on the EDGAR database before applying a predefined text processing operator (HTML-Parser). The aggregated text length for all filings of both procedures is compared in order to validate the capability of the „Annual Report Algorithm" in decoding escape sequences. The aggregated text length includes each individual element in an underlying text document (text, digits, spaces, special characters etc.).

In contrast to the "Annual Report Algorithm" the modified "Items Algorithm" is validated by its ability to distribute the extracted information to the individual items an annual report filed with the SEC is composed of. In order to test and validate the capabilities of the "Items Algorithm" I again use the "Complete Submission Text Files" as provided by the SEC and extract only the 10-K section of each filing. For each submission, I retrieve separate word counts for the 10-K section and for all individual items extracted by the "Items Algorithm". Despite textual information embedded in the 10-K section not contained in a particular item (introduction) a word count comparison between the overall 10-K section and all items represents an attempt to validate the capabilities of the "Items Algorithm" in extracting certain sections from the core document of an annual report filed with the SEC and its EDGAR system. Figure 5 illustrates the content validation process of the "Items Algorithm".

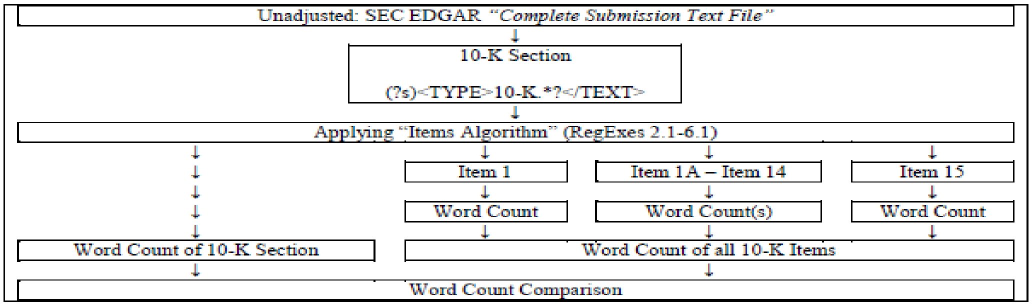

Unadjusted: SEC EDGAR "Complete Submission Text File"
10-K Section
(?s)<TYPE>10-K.*?</TEXT>
Applying "Items Algorithm" (RegExes 2.1-6.1)
Item 1
Item 1A – Item 14
Item 15
Word Count
Word Count(s)
Word Count
Word Count of 10-K Section
Word Count of all 10-K Items
Word Count Comparison

Figure 5. Content validation process of the "Items Algorithm"

Notes: The figure presents the content validation process of the "Items Algorithm". First, the entire 10-K section of each filing from the "Complete Submission Text File" as provided on the SEC server is extracted. Word counts for the entire 10-K section as well as for all individual items are retrieved by applying the "Items Algorithm" in order to be compared. Due to structural changes of the annual report on Form 10-K over time (different number of items) the relation of text length between the overall 10-K section and all individual items shall represent the ability of the algorithm to extract particular items from the 10-K section.

Table 9 presents the validation results for the "Items Algorithm".

| Year | Filings | | "Items Algorithm" | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Word Count Comparison | | Precision, Recall, and F-measure | | | | | | |
| | Number | % | ∑ of Items (%) | Rest/ Error (%) | Filings | Items | | | | | |
| | | | | | Tested | Exists | Extracted | Correct | Precision | Recall | F-measure |
| 2016 | 2,886 | 44.63 | 97.72 | 2.28 | 10 | 195 | 191 | 185 | 96.86 | 94.87 | 95.85 |
| 2015 | 3,714 | 46.51 | 97.60 | 2.40 | 10 | 200 | 200 | 195 | 97.50 | 97.50 | 97.50 |
| 2014 | 4,004 | 49.53 | 97.52 | 2.48 | 10 | 198 | 197 | 193 | 97.97 | 97.47 | 97.72 |
| 2013 | 3,962 | 48.88 | 97.53 | 2.47 | 10 | 192 | 188 | 188 | 100.00 | 97.92 | 98.95 |
| 2012 | 3,938 | 46.92 | 97.39 | 2.61 | 10 | 198 | 192 | 183 | 95.31 | 92.42 | 93.85 |
| 2011 | 4,104 | 46.43 | 97.43 | 2.57 | 10 | 193 | 191 | 189 | 98.95 | 97.93 | 98.44 |
| 2010 | 2,805 | 30.61 | 97.10 | 2.90 | 10 | 197 | 168 | 155 | 92.26 | 78.68 | 84.93 |
| 2009 | 2,719 | 27.63 | 97.15 | 2.85 | 10 | 196 | 181 | 164 | 90.61 | 83.67 | 87.00 |
| 2008 | 2,077 | 23.75 | 97.28 | 2.72 | 10 | 196 | 184 | 170 | 92.39 | 86.73 | 89.47 |
| 2007 | 2,065 | 24.08 | 97.38 | 2.62 | 10 | 195 | 184 | 173 | 94.02 | 88.72 | 91.29 |
| 2006 | 2,662 | 30.07 | 97.49 | 2.51 | 10 | 198 | 184 | 165 | 89.67 | 83.33 | 86.39 |
| 2005 | 3,122 | 34.62 | 97.70 | 2.30 | 10 | 181 | 175 | 163 | 93.14 | 90.06 | 91.57 |
| 2004 | 3,496 | 40.81 | 97.60 | 2.40 | 10 | 173 | 170 | 163 | 95.88 | 94.22 | 95.04 |
| 2003 | 3,903 | 46.09 | 97.33 | 2.67 | 10 | 161 | 161 | 154 | 95.65 | 95.65 | 95.65 |
| 2002 | 4,961 | 55.57 | 97.65 | 2.35 | 10 | 150 | 150 | 150 | 100.00 | 100.00 | 100.00 |
| 2001 | 5,799 | 62.71 | 97.61 | 2.39 | 10 | 146 | 144 | 135 | 93.75 | 92.47 | 93.10 |
| 2000 | 6,268 | 63.51 | 97.55 | 2.45 | 10 | 150 | 149 | 138 | 92.62 | 92.00 | 92.31 |
| 1999 | 6,302 | 62.26 | 97.55 | 2.45 | 10 | 146 | 145 | 143 | 98.62 | 97.95 | 98.28 |
| 1998 | 6,492 | 63.11 | 97.56 | 2.44 | 10 | 140 | 140 | 128 | 91.43 | 91.43 | 91.43 |
| 1997 | 6,397 | 64.62 | 97.43 | 2.57 | 10 | 132 | 129 | 125 | 96.90 | 94.70 | 95.79 |
| 1996 | 3,918 | 62.61 | 97.27 | 2.73 | 10 | 136 | 121 | 112 | 92.56 | 82.35 | 87.16 |
| 1995 | 1,907 | 58.93 | 97.07 | 2.93 | 10 | 135 | 135 | 127 | 94.07 | 94.07 | 94.07 |
| 1994 | 1,039 | 54.03 | 97.23 | 2.77 | 10 | 140 | 139 | 133 | 95.68 | 95.00 | 95.34 |
| 1993 | 1 | 25.00 | 98.49 | 1.51 | 1 | 14 | 14 | 14 | 100.00 | 100.00 | 100.00 |
| Total | 88,541 | 46.88 | 97.48 | 2.52 | 231 | 3,962 | 3,832 | 3,645 | 95.12 | 92.00 | 93.53 |

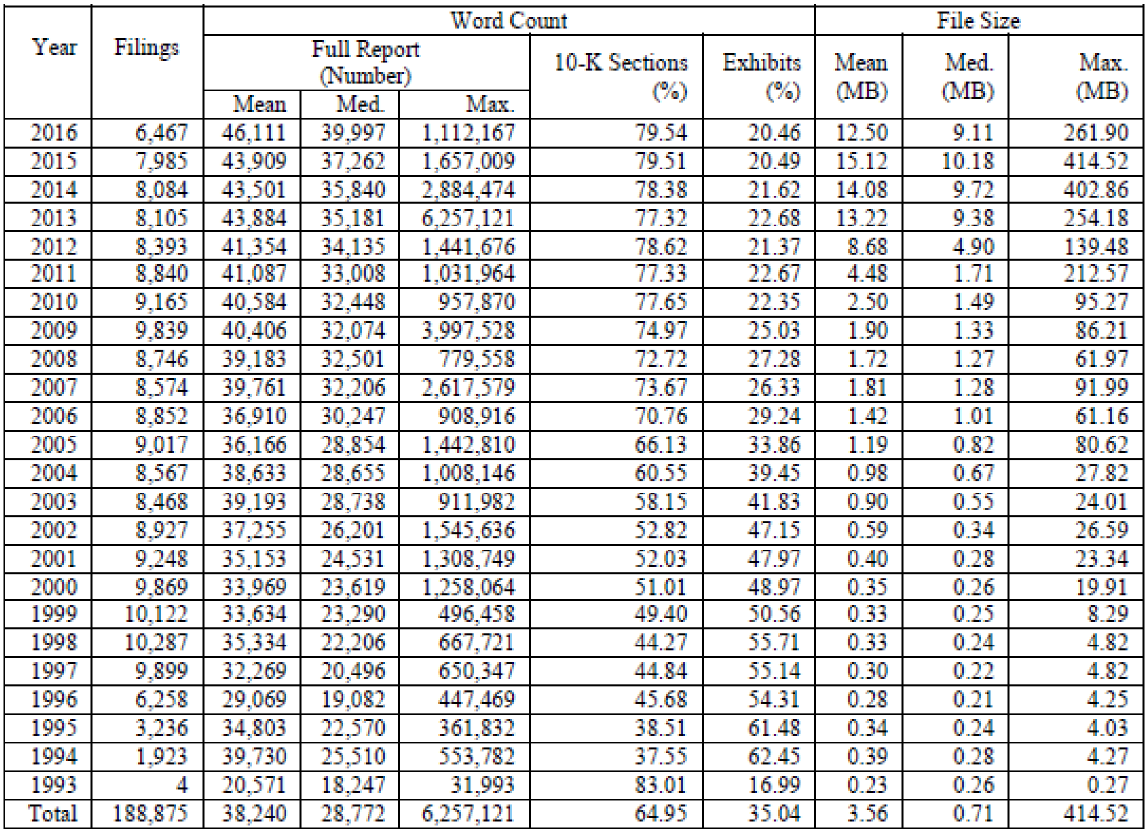
Table 9. Validation results of the "Items Algorithm"

Notes: The table presents the validation results of the "Items Algorithm". The second and third columns show the number of filings of which items could be extracted from by applying the "Items Algorithm" (filings were not machine-parsable due to lacks of content, inconsistent filing structure, table tags and HTML formatting inconsistencies). Only filings with extracted items length exceeding 90 percent of 10-K section are presented. The next two columns show the average amount of extracted information from each filing in a particular year since 1993. The next columns show the performance evaluation of the "Items Algorithm" using precision (=number of correct answers/number of total answers), recall (=number of correct answers/total possible correct answers), and F-measure (=2*precision*recall/precision+recall).

## 7. DESCRIPTIVE STATISTICS ON FORM 10-K CONTENTS

In total, I examine the textual composition of 188,875 annual reports filed with the SEC between 1993 and 2016. On average, an annual report on Form 10-K submitted to the EDGAR system during the sample period is composed of 38,240 words. The average word count of an annual submission increased from 39,730 in 1994 to 46,111 in 2016. The medians of the word counts increased accordingly. The majority of textual information embedded in an annual report on Form 10-K are contained in the core document (64.95 percent) whereas the disclosed exhibits represent only a minority of the overall textual elements stated in annual submissions (35.04 percent). By examining the EDGAR database and its Form 10-K filings in more detail, investors and researchers can see that the average file size (Megabyte) of an annual report made with the electronic disclosure system increased in recent years due to HTML formatting, ASCIIencodings and XBRL documents. Table 10 presents descriptive statistics of the text length and the file size of 188,875 annual reports on Form 10-K (Form 10-K405) filed with the SEC between 1993 and 2016.

Table 10. Descriptive statistics of SEC EDGAR Form 10-K reports



| Year | Filings | Word Count | | | | | File Size | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Full Report (Number) | | | 10-K Sections (%) | Exhibits (%) | Mean (MB) | Med. (MB) | Max. (MB) |
| | | Mean | Med. | Max. | | | | | |
| 2016 | 6,467 | 46,111 | 39,997 | 1,112,167 | 79.54 | 20.46 | 12.50 | 9.11 | 261.90 |
| 2015 | 7,985 | 43,909 | 37,262 | 1,657,009 | 79.51 | 20.49 | 15.12 | 10.18 | 414.52 |
| 2014 | 8,084 | 43,501 | 35,840 | 2,884,474 | 78.38 | 21.62 | 14.08 | 9.72 | 402.86 |
| 2013 | 8,105 | 43,884 | 35,181 | 6,257,121 | 77.32 | 22.68 | 13.22 | 9.38 | 254.18 |
| 2012 | 8,393 | 41,354 | 34,135 | 1,441,676 | 78.62 | 21.37 | 8.68 | 4.90 | 139.48 |
| 2011 | 8,840 | 41,087 | 33,008 | 1,031,964 | 77.33 | 22.67 | 4.48 | 1.71 | 212.57 |
| 2010 | 9,165 | 40,584 | 32,448 | 957,870 | 77.65 | 22.35 | 2.50 | 1.49 | 95.27 |
| 2009 | 9,839 | 40,406 | 32,074 | 3,997,528 | 74.97 | 25.03 | 1.90 | 1.33 | 86.21 |
| 2008 | 8,746 | 39,183 | 32,501 | 779,558 | 72.72 | 27.28 | 1.72 | 1.27 | 61.97 |
| 2007 | 8,574 | 39,761 | 32,206 | 2,617,579 | 73.67 | 26.33 | 1.81 | 1.28 | 91.99 |
| 2006 | 8,852 | 36,910 | 30,247 | 908,916 | 70.76 | 29.24 | 1.42 | 1.01 | 61.16 |
| 2005 | 9,017 | 36,166 | 28,854 | 1,442,810 | 66.13 | 33.86 | 1.19 | 0.82 | 80.62 |
| 2004 | 8,567 | 38,633 | 28,655 | 1,008,146 | 60.55 | 39.45 | 0.98 | 0.67 | 27.82 |
| 2003 | 8,468 | 39,193 | 28,738 | 911,982 | 58.15 | 41.83 | 0.90 | 0.55 | 24.01 |
| 2002 | 8,927 | 37,255 | 26,201 | 1,545,636 | 52.82 | 47.15 | 0.59 | 0.34 | 26.59 |
| 2001 | 9,248 | 35,153 | 24,531 | 1,308,749 | 52.03 | 47.97 | 0.40 | 0.28 | 23.34 |
| 2000 | 9,869 | 33,969 | 23,619 | 1,258,064 | 51.01 | 48.97 | 0.35 | 0.26 | 19.91 |
| 1999 | 10,122 | 33,634 | 23,290 | 496,458 | 49.40 | 50.56 | 0.33 | 0.25 | 8.29 |
| 1998 | 10,287 | 35,334 | 22,206 | 667,721 | 44.27 | 55.71 | 0.33 | 0.24 | 4.82 |
| 1997 | 9,899 | 32,269 | 20,496 | 650,347 | 44.84 | 55.14 | 0.30 | 0.22 | 4.82 |
| 1996 | 6,258 | 29,069 | 19,082 | 447,469 | 45.68 | 54.31 | 0.28 | 0.21 | 4.25 |
| 1995 | 3,236 | 34,803 | 22,570 | 361,832 | 38.51 | 61.48 | 0.34 | 0.24 | 4.03 |
| 1994 | 1,923 | 39,730 | 25,510 | 553,782 | 37.55 | 62.45 | 0.39 | 0.28 | 4.27 |
| 1993 | 4 | 20,571 | 18,247 | 31,993 | 83.01 | 16.99 | 0.23 | 0.26 | 0.27 |
| Total | 188,875 | 38,240 | 28,772 | 6,257,121 | 64.95 | 35.04 | 3.56 | 0.71 | 414.52 |

Notes: The table presents descriptive statistics of the text lengths, document compositions and file sizes for all annual reports filed with the SEC since 1993. Columns 3-5 show the means, medians and maxima of word counts of Form 10-K filings made on EDGAR. The average distribution of textual information between the 10-K sections and exhibits contained in the “Complete Submission Text Files” is presented in column 6 and 7.

The distribution of textual elements among the various 10-K items is unequal. On average 22.65 percent of all textual information are contained in Item 1 (“Business”). Describing a company´s business as well as its main products and services, the item may also include information about the competition, regulations and other issues a particular company is faced with [46] [47]. Item 7 (“Management´s Discussion and Analysis of Financial Condition and Results of Operations – MD&A”) represents 18.58 percent of the given information within Form 10-K filings made with the SEC. The item states information about a company´s operations and financial results in addition to its liquidity and capital resources. The section may include off-balance sheet arrangements and contractual obligations alongside key business risks [46] [47]. Item 8 (“Financial Statements and Supplementary Data”) requires a company to disclose audited financial statements [46] [48] [47]. Additional information explaining the financial statements in more detail (“Notes to Consolidated Financial Statements”, “Report of Management”, “Report of Independent Registered Accounting Firm” etc.) represent 15.96 percent of all given information in the 10-K section of an annual report. Item 1A (“Risk Factors”) describes significant factors that may adversely affect a filer´s business, financial condition or future financial performance [46] [47]. Since electronic filings became available on average 8.42 percent of all textual information disclosed in annual submissions are contained in this section. Each of the remaining items only represent a fraction of the overall textual information embedded in Form 10-K filings. While the length for most sections in annual reports remained constant over time the amount of textual information contained in Item 1A (“Risk Factors”) increased from 12.56 percent in 2006 to 20.10 percent in 2016 indicating that SEC EDGAR filers disclose more information about risks in recent years.

## 8. SUMMARY

This paper displays the huge amount and variety of publicly available corporate information filed with the SEC and distributed by its EDGAR database. It shows how massive data can be retrieved from the SEC server in a fast and efficient way using simple and easy accessible software. The second main purpose of this paper is to create standardized procedures (“Annual Report Algorithm” and “Items Algorithm”) investors and researchers can use to extract any kind of textual information from financial statements filed with the SEC. This is achieved by providing regular expressions for multiple steps of data cleaning and filtering. Using these dynamic and platform-independent extraction algorithms the paper analyses the textual composition of more than 180,000 annual reports filed with the SEC via the EDGAR system between 1993 and 2016. The algorithms are tested for validity in several ways. The tools and algorithms intend to reduce costs and lower technical boundaries for researchers in the field of finance and accounting to engage in textual analysis.

## ACKNOWLEDGEMENTS

I appreciate the valuable comments by Klaus Henselmann and Daniel Büchs

## REFERENCES

[1] Grant G H, Conlon S J (2006) EDGAR Extraction System: An Automated Approach to Analyze Employee Stock Option Disclosures, in: Journal of Information Systems (JIS), 20(2)/2006, 119-142

[2] Wilks Y (1997) Information Extraction as a Core Language Technology, in: M T Pazienza, Ed.Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer-Verlag, Berlin Heidelberg, Germany

[3] Mooney R J, Bunescu R (2005) Mining Knowledge from Text Using Information Extraction, in: SIGKDD Explorations (SIGKDD), 7(1)/2005, 3-1

[4] Gaizauskas R, Humphreys K, Azzam S, Wilks Y (1997) Concepticons vs. Lexicons: an Architecture for Multilingual Information Extraction, in: M T Pazienza, Ed. Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer- Verlag, Berlin Heidelberg, Germany

[5] Garcia D, Norli O (2012) Crawling EDGAR, in: The Spanish Review of Financial Economics (SRFE), 10/2012, 1-10

[6] Stümpert T (2008) Extracting Financial Data from SEC Filings for US GAAP Accountants, in: D Seese, C Weinhardt, F Schlottmann, Eds. Handbook on Information Technology in Finance. Springer-Verlag, Berlin Heidelberg, Germany

[7] Bovee M, Kogan A, Nelson K, Srivastava R P, Vasarhelyi M A, (2005) Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL), in: Journal of Information Systems (JIS), 19(1)/2005, 19-41

[8] O´Riain S (2012) Semantic Paths in Business Filings Analysis. Ph.D. thesis, National University of Ireland, Galway, Ireland

[9] Loughran T, McDonald B (2014) Measuring Readability in Financial Disclosures, in: The Journal of Finance (JoF), 69(4)/2014, 1643-1671

[10] Gerdes J Jr (2003) EDGAR-Analyzer: automating the analysis of corporate data contained in the SECs EDGAR database, in: Decision Support Systems 35/2003, 7-29

[11] Kambil A, Ginsburg M (1998) Public Access Web Information Systems: Lessons from the Internet EDGAR Project, in: Communications of the ACM (CACM), 41(7)/1998, 91-97

[12] Davis A K, Tama-Sweet I (2012) Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A, in: Contemporary Accounting Research (CAR), 29(3)/2012, 804-837

[13] Loughran T, McDonald B (2016) Textual Analysis in Accounting and Finance: A Survey, in: Journal of Accounting Research (JAR), 54(4)/2016, 1187-1230

[14] Tetlock P C (2007) Giving Content to Investor Sentiment: The Role of the Media in the Stock Market, in: The Journal of Finance (Jof), 62(3)/2007, 1139-1168

[15] Loughran T, McDonald B (2011a) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, in: The Journal of Finance (JoF), 66(1)/2011, 35-65

[16] Jegadeesh N, Wu D (2013) Word power: A new approach for content analysis, in: Journal of Financial Economics (JFE), 110(3)/2013, 712-729

[17] Stümpert T, Seese D, Centinkaya Ö, Spöth R (2004) EASE – a software agent that extracts financial data from the SEC´s EDGAR database, in: Proceedings of the 4th International ICSC Symposium on Engineering of Intelligent Systems (EIS 2004). Funchal, Portugal

[18] Engelberg J, Sankaraguruswamy S (2007) How to Gather Data Using a Web Crawler: An Application Using SAS to Search Edgar. Working Paper, SSRN

[19] Cong Y, Kogan A, Vasarhelyi M A (2007) Extraction of Structure and Content from the Edgar Database: A Template-Based Approach, in: Journal of Emerging Technologies in Accounting (JETA) 4(1)/2007, 69-86

[20] Thai V, Davis B, O´Riain S, O´Sullivan D, Handschuh S (2008) Semantically Enhanced Passage Retrieval for Business Analysis Activity, in: Proceedings of the 16th European Conference on Information Systems (ECIS 2008). Galway, Ireland

[21] Chakraborty V, Vasarhelyi M A (2010) Automating the process of taxonomy creation and comparison of taxonomy structures, in: 19th Annual Research Workshop on Strategic and Emerging Technologies, American Accounting Association. San Francisco, California, USA

[22] Hernandez M A, Ho H, Koutrika G, Krishnamurthy R, Popa L, Stanoi I R, Vaithyanathan S, Das S (2010) Unleashing the Power of Public Data for Financial Risk Measurement, Regulation, and Governance. IBM Technical Report #RJ10475

[23] Srivastava R P (2016) Textual Analysis and Business Intelligence in Big Data Environment: Search Engine versus XBRL, in: Indian Accounting Review (IAR), 20(1)/2016, 1-20

[24] SEC (2013) What We Do, Available online on URL: https://www.sec.gov/about/whatwedo.shtml

[25] SEC (2010) Important Information about EDGAR, Available online on URL:

https://www.sec.gov/edgar/aboutedgar.htm

[26] SEC (2006) Electronic Filing and the EDGAR System: A Regulatory Overview, Available online on URL https://www.sec.gov/info/edgar/regoverview.htm

[27] Pagell R A (1995) EDGAR: Electronic Data Gathering and Receiving, in: Business Information Review (BIR), 11(3)/1995, 56-68

[28] SEC Release 34-36997 (1996) EDGAR Phase-in Complete on May 6, 1996, Available online on URL: https://www.sec.gov/info/edgar/ednews/34-36997.htm

[29] SEC Regulation S-T (2016) General Rules and Regulations for electronic Filings, Available online on URL:
http://www.ecfr.gov/cgi-bin/textidx? node=17:3.0.1.1.14&rgn=div5#se17.3.232_1100

[30] SEC Release 33-8099 (2002) Mandated EDGAR Filing for Foreign Issuers, Available online on URL: https://www.sec.gov/rules/final/33-8099.htm

[31] SEC (2015) Information for FTP Users, Available online on URL:
https://www.sec.gov/edgar/searchedgar/ftpusers.htm

[32] SEC Index Files (2016) Full Index Files, Available online on URL: ftp://ftp.sec.gov/edgar/fullindex/

[33] W3C Recommendation (1999) HTML 4.01 Specification, Available online on URL:
https://www.w3.org/TR/html401/cover.html

[34] Filer Manual (2016) Filer Manual – Volume II EDGAR Filing, Available online on URL:
https://www.sec.gov/info/edgar/edgarfm-vol2-v37.pdf

[35] Ditter D, Henselmann K, Scherr E (2011) Using XBRL Technology to Extract Competitive Information from Financial Statements, in: Journal of Intelligence Studies in Business (JISIB), 1/2011, 19-28

[36] W3C Strict DTD (1999) HTML 4.01 Strict DTD, Available online on URL:
https://www.w3.org/TR/html4/strict.dtd

[37] SEC (2000) HTML Specifications for EDGAR Rel. 7.0, Available online on URL: https://www.sec.gov/info/edgar/ednews/edhtml.htm [38] Bodnaruk A, Loughran T, McDonald B (2015) Using 10-K Text to Gauge Financial Constraints, in: Journal of Financial and Quantitative Analysis (JFQA), 50(4)/2015, 623-646

[39] Loughran T, McDonald B (2011b) Internet Appendix for “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks”, Available online on URL:
http://www.afajof.org/SpringboardWebApp/userfiles/afa/file/Supplements%20and%20Data%20
Sets/Internet%20Appendix%20for%20When%20Is%20a%20Liability%20Not%20a%20Liability%20Text
ual%20Analysis,%20Dictionaries,%20and%2010-Ks%206989-IA-Feb-2011.pdf

[40] SEC EDGAR Archives (2016) Coca Cola Company´s Financial Statement Submissions on 2016-02-25, Available online on URL:
https://www.sec.gov/Archives/edgar/data/21344/000002134416000050/0000021344-16-000050-index.htm

[41] Friedl J E F (2006) Mastering Regular Expressions, Third Edition, O´Reilly Media, Inc., Sebastopol, California, USA

[42] Thompson K (1968) Regular Expression Search Algorithm, in: Communications of the ACM (CACM), 11(6)/1968, 419-422

[43] Goyvaerts J, Levithan S (2012) Regular Expressions Cookbook, Second Edition, O´Reilly Media, Inc., Sebastopol, California, USA

[45] SEC EDGAR (2015) Public Dissemination Service (PDS) Technical Specification, Available online on URL:
https://www.sec.gov/info/edgar/specifications/pds_dissemination_spec.pdf

[46] SEC (2011) Fast Answers – How to Read a 10-K, Available online on URL:
https://www.sec.gov/answers/reada10k.htm

[47] SEC Regulation S-K (2016) Standard Instructions for filing Forms under Securities Act of 1933, Securities Exchange Act of 1934 and Energy Policy and Conservation Act of 1975-Regulation S-K, Available online on
URL:http://www.ecfr.gov/cgi-bin/textidx?
SID=8e0ed509ccc65e983f9eca72ceb26753&node=17:3.0.1.1.11&rgn=div5#se17.3.229_1101

[48] SEC Regulation S-X (2016) Form and Content of and Requirements for Financial Statements; Securities Act of 1933, Securities Exchange Act of 1934, Investment Company Act of 1940, Investments Advisers Act of 1940, and Energy Policy and Conservation Act of 1975- Regulation S-X, Available online on URL:
http://www.ecfr.gov/cgi-bin/textidx?
SID=8e0ed509ccc65e983f9eca72ceb26753&node=17:3.0.1.1.8&rgn=div5#se17.3.210_11_601