Large-Scale Open Corporate Data Collection and Analysis as an Enabler of Corporate Social Responsibility Research

Vasiliki Gkatziaki(✉), Symeon Papadopoulos, Sotiris Diplaris,
and Ioannis Kompatsiaris

Information Technologies Institute (ITI), CERTH -ITI, Thessaloniki, Greece
vasgat@iti.gr

Abstract. During the last years, citizens and transparency initiatives put increasing pressure on governments, organizations, and companies to be more transparent and to publicize information pertaining to their operations. Although several organizations have started engaging in open data practices, data quality, structure and availability is still highly inconsistent across organizations, which makes it challenging and effort-intensive to obtain and analyze large-scale high-quality datasets. To this end, this paper examines how publicly available financial and corporate data can be leveraged to extract useful inferences regarding the financial and social performance of companies. Numerous reports have been collected from the Securities Exchange Commission (SEC) and analyzed to study hypotheses regarding the corporate practices and social responsibility of companies.

Keywords: Open data · Information retrieval · Corporate Social Performance · eXtensible Business Reporting Language (XBRL)

1 Introduction

During the last decade, important efforts have been expended towards ensuring transparent reporting regarding the Environmental, Social and Governance performance of companies. Governments, shareholders and other stakeholders require companies to become more transparent regarding their financial, social and environmental activities. For instance, the Securities and Exchange Commission (SEC), a government commission created by the US Congress with the goal of protecting investors from deceitful and manipulative market practices, requires US public companies to file statements regarding their financial performance in order to protect investors. In 2008, SEC adopted a new rule [21] that enforces public companies to disclose their financial statements using the eXtensible Business Reporting Language (XBRL) format[1] in an attempt to increase financial transparency.

[1] XBRL (https://www.xbrl.org/) is a global standard for business reporting that allows the digitization of financial statements.

I. Kompatsiaris et al. (Eds.): INSCI 2017, LNCS 10673, pp. 266–280, 2017.
https://doi.org/10.1007/978-3-319-70284-1_21

WikiRate[2] is a platform for collecting and analyzing information about companies' Environmental, Social and Governance (ESG) performance and it aims to make corporations more transparent, reactive and ethical by making data about their ESG performance available to everyone [17]. Numerous metric points, coming from different sources, have been collected and integrated into the WikiRate platform. Moreover, a plethora of financial metric points have been extracted from 10-K filings[3] and integrated into the platform.

In this paper, we investigate how open financial data can be leveraged to extract useful inferences regarding the social performance of companies. We collected 39,029 10-K reports in XBRL format from SEC and after processing these reports, we managed to extract more than 529,000 financial facts related to different aspects of the financial performance of companies in different reporting years. Additionally, we collected about 3,442 Conflict Minerals reports. Then, we performed data analysis on two datasets to study research hypotheses regarding specific aspects of Corporate Social Responsibility (CSR). The first dataset, we studied, comprised 25,500 observations regarding 7,700 companies, while the second comprised 465 observations regarding 465 companies and contained financial facts extracted from XBRL reports as well as green scores defined by Newsweek in 2016[4].

The main contributions of this work can be summarized as follows:

- *Collecting a large amount of data from the Web in relation to the financial performance of companies*: A large database of over 500,000 financial facts about more than 50,000 companies was extracted from 10-K filings (available by SEC). Additionally, a REST API was developed to make the collected data available to third parties for further research.
- *Demonstrating the value of open data for CSR research:* We argue that financial open data can lead to useful inferences regarding the social performance of companies. We support our claim by testing four hypotheses on two datasets that integrate different types of data regarding the social and environmental performance of companies.

Figure 1 presents an overview of the methodology adopted in this paper. This comprises three key steps: *Data Collection*, *Data Integration* and *Data Analysis*. *Data Collection* is responsible for constructing appropriate wrappers for the given Web sources, executing the constructed wrappers and handling the extracted data. *Data Integration* results in the construction of appropriate datasets by integrating data coming from different sources. Finally, *Data Analysis* on the created datasets performs *descriptive statistics* and *statistical inference* to verify selected research hypotheses.

[2] http://wikirate.org/.

[3] A 10-K is a comprehensive summary report of a company's performance that must be submitted annually to the SEC.

[4] Newsweek, an American news magazine, publishes yearly rankings for the 500 largest publicly-traded US companies based on their overall environmental performance. For 2016, those are available on http://www.newsweek.com/green-2016.

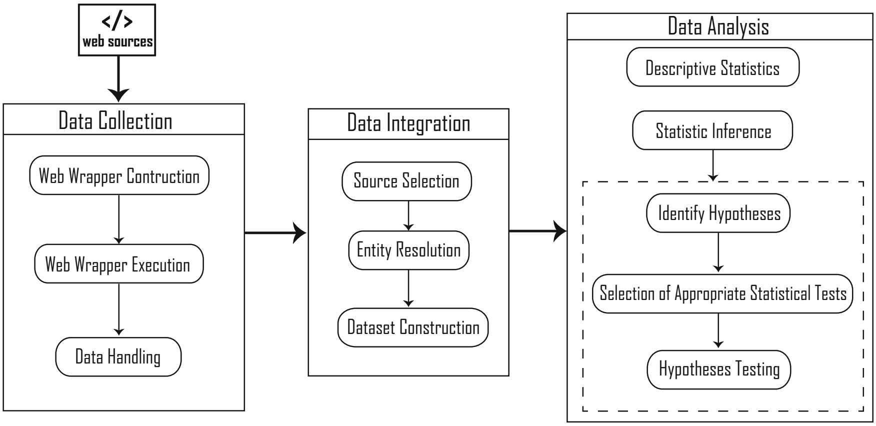

Fig. 1. Research methodology.

2 Related Work

The concept of Corporate Social Responsibility (CSR) emerged in the United States almost 60 years ago [2]. Over the years, numerous researchers tried to provide different definitions of CSR, to achieve better understanding and to explore different dimensions of the concept [5,7,11,18,22]. For instance, Wanderley et al. [24] investigated 127 corporations and how their CSR communication on the Web is influenced by the country or industry. They concluded that CSR disclosure on the Web is significantly influenced by the country where the corporation operates but less by the industry sector. Cai et al. [4] researched 475 companies from controversial industries in terms of producing dangerous products for humans, environment or society to study the importance of CSR engagement. Their results indicated that CSR engagement even for those industries is important and helped them increase their value. Bauman and Skitka [1] studied how CSR affects employees and developed a framework that identifies the correlation between employees' needs and companies' actions. Parguel et al. [20] addressed the question of how sustainable ratings affect consumers and discovered that such ratings significantly influence consumers to make more effective assessments on companies. Namkung and Jang [19] conducted a survey to discover which characteristics play an important role on consumers to pay more for green practices in restaurants. Finally, Loughran et al. [13] studied the use of ethics-related terms in 10-K reports obtained from SEC and detected an inconsistency between companies' social performance and the use of such terms.

The relationship between Corporate Social Performance (CSP) and Corporate Financial Performance (CFP) has received considerable attention over the years. Numerous studies have been performed to investigate this relationship. Bragton and Marlin [3] were among the first to study the connection between CSP and CFP. They investigated the assumption that more strict pollution controls translate to less profit. Their outcomes negate the hypothesis that environmental virtue and profitability are incompatible goals. Bromiley and Marcus [14] studied the assumption that stock market reaction to problematic behaviors can function as an instrument of social control to discourage such behaviors. Their findings indicated that stock market as a medium of social control is inadequate. Waddock and Graves [23] tried to use a more representative measurement approach for CSP by taking into consideration different aspects and discovered a positive connection between CSP and CFP. Margolis et al. [15] performed a meta-analysis on 167 studies to define which CSR aspects are more effective over the financial performance of a firm and they discovered that there is a small positive effect of CSP over CFP. Cheng et al. [6] researched how CSP affects a firm's ability to access capital. They concluded that companies with superior social performance have lower capital limitations. Eckbo et al. [9] studied the effect of gender-balancing in the board of directors in the case of Norwegian companies. The results showed that gender-quota has neutral effect on a firm's value.

With the emergence of Web 2.0 and the prevalence of the Web as a global communication medium, governments are increasingly demanding companies to disclose data regarding their financial and social performance in order to protect citizens and initiate an open dialogue. As a result, large amounts of information about companies' financial and social performance are available online by governments, NGOs and other initiatives. In this paper, we study how open financial data disclosed by US companies to the SEC can lead to useful inferences regarding specific aspects of their social performance. Additionally, an integrated dataset is studied that combines data from two different sources and investigates the relationship of CSP and CFP.

3 Data Collection, Integration and Publishing

3.1 Company Data Model and Metrics

Figure 2 illustrates the developed data model, comprising the following entities:

- `companies`: US companies, each associated with a name, a Central Index Key[5] (CIK) and a Standard Industrial Classification (SIC) code.
- `metrics`: pieces of information related to companies; each metric is defined by its name, value, source, the year to which it refers and the CIK referring to the associated company.

The underlying database, mongoDB, is a schema-free document database, which offers the required flexibility allowing to add more fields based on the needs of each extracted metric. The serialization format is BSON (binary JSON), which offers higher storage efficiency.

[5] https://www.sec.gov/edgar/searchedgar/cik.htm.

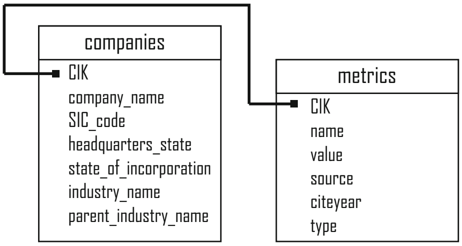

Fig. 2. Data model used for storing the collected data.

3.2 Data Collection and Extraction from SEC

This study has been based on data collected from the Securities and Exchange Commission (SEC). SEC supports transparency and requires public companies to file registration statements, periodic reports and other securities forms electronically. Most of these high-value data are received, processed and disseminated through the Electronic Data Gathering, Analysis and Retrieval (EDGAR) system. We obtained from the public database of EDGAR two types of report:

- *Conflict Minerals Report:* companies are obliged to disclose in these documents their use of conflict minerals considered necessary to production.
- *Form 10-K:* these documents make available audited financial statements regarding the annual financial performance of the company.

Collecting the aforementioned reports in large scale was a challenging task, which necessitated the development of a custom Web wrapper, i.e. a highly automated piece of software that iterates through a number of Web pages of interest and extracts structured information from each page [10]. Algorithm 1 specifies the operation of the developed SEC Wrapper. Given as input the set of all available Standard Industry Codes (SIC) and the report types of interest, the wrapper searches for companies, and for each company it extracts relevant information, such as the company name, the Central Index Code, the state of incorporation, etc. from the appropriately selected DOM[6] elements. In addition to basic company information, all reports of the selected type (e.g., Form 10-K and Conflict Minerals) are extracted from the respective company page. To specify the target DOM elements, i.e. the placeholders of interesting information within the fetched HTML pages, we composed extraction rules in the form of *CSS selectors*. The implementation of the wrapper was based on the JSoup[7] Java library. Using the wrapper, we collected data for 50,219 US companies, 3,442 Conflict Minerals Reports and 162,349 Form 10-K filings.

SEC is also supporting efforts to make data easy to process in a programmatic way. Numerous reports in EDGAR are available in this format. Due to

[6] The Document Object Model (DOM) is a programming interface for HTML and XML documents.

[7] http://jsoup.org/.

Algorithm 1. SEC Wrapper

```
Input: set of standard industry codes $S_{SIC}$; report types $S_{RT}$

foreach $sic \in S_{SIC}$ do
    $companies \leftarrow$ lookup($sic$)
    foreach $company\ c \in companies$ do
        url$_c$ $\leftarrow$ create_company_url($c$)
        extract_company_info(url$_c$)
            (i) extract content from defined DOM elements
            (ii) post-processing of the extract content if needed
            (iii) store company data
        foreach $r \in S_{RT}$ do
            extract_reports(url$_c$, $r$)
                (i) extract content from defined DOM elements
                (ii) store report
```

SEC, companies have to use the Generally Accepted Accounting Principles (usgaap) for financial reporting, i.e. a standard framework of guidelines for financial accounting, including standards, conventions and rules to be followed by accountants.

XBRL supports the representation of complex financial statements. In total, we managed to collect 39,029 Form 10-K filings in XBRL format. For extracting financial facts related to a company from the collected reports, we leveraged the DOM tree of the instance document of the Form 10-K. Overall, we managed to extract a plethora of financial data points related to 27 metrics (Table 1). Out of those, our data analysis focused on the four metrics of Table 2.

Table 1. Metrics extracted from XBRL reports.

Metrics
Advertising Expense, Assets, Comprehensive Income Net of Tax, Cost of Goods Sold, Cost of Goods and Services Sold, Current Foreign Tax Expense Benefit, Current Income Tax Expense Benefit, Deferred Income Tax Expense Benefit, Deferred Tax Liabilities Undistributed Foreign Earnings, Effective Income Tax Rate Continuing Operations, Effective Income Tax Rate Reconciliation at Federal Statutory Income Tax Rate, Good Will, Gross Profit, Income Loss From Continuing Operations Before Income Taxes Foreign, Income Tax Expense Benefit, Income before income taxes, Profit, Net Income Attributable to Non Controlling Interest, Net Income Attributable to the company, Research and Development Expense, Revenues, Sales Revenue Goods Net, Sales Revenue Net, Selling and Marketing Expense, Undistributed Earnings Of Foreign Subsidiaries, Unrecognized Tax Benefits Reductions Resulting from Lapse of Applicable statute of Limitations, Unrecognized Tax Benefits That would impact Effective Tax Rate

Table 2. Metrics extracted from XBRL reports and used in our data analysis.

Metric	Details
Profits	The consolidated profit or loss of the company for the reported period
Good Will	Good Will is the value of a company's brand name and it depends on the customer base, customer relations, employee relations, its patents and technology
Research & Development (R&D) Expense	The costs related to R&D for the reporting period which aims on discovering new knowledge with the ambition to exploit such knowledge on developing new products or services
Undistributed earnings of foreign subsidiaries	Undistributed earnings of foreign subsidiaries that are intended to be indefinitely reinvested for the reported period

3.3 Data Availability

We developed a RESTful API on top of the collected data in order to make them available to third parties for further study. The API, which is available on the endpoint http://easie.iti.gr/sec_dataset/, returns data in JavaScript Object Notation (JSON) and offers two basic methods are available:

- *GET companies*[8], which is responsible for returning a collection of companies by querying their names with a search term. If no query is defined then the method returns all available companies.
- *GET metrics*[9], which is responsible for returning a collection of metrics data points by querying their name with a search term. If no query is defined then the method returns all available metrics.

There are several parameters that users can specify to limit the results on both methods. These are specified in Tables 3 and 4.

Table 3. Available options of *GET companies* method.

Parameter	Description
`q`	Query companies given a search term
`CIK`	Search for a company given its central index key
`SIC`	Search for companies based on industry code
`headquarters_state`	Search for companies based on their headquarters location
`state_of_inc`	Search for companies based on their state of incorporation
`page`	Pagination of the results (100 results per page)

[8] http://easie.iti.gr/sec_dataset/companies.

[9] http://easie.iti.gr/sec_dataset/metrics.

Table 4. Available options of *GET metrics* method.

Parameter	Description
`q`	Query metrics given a search term
`CIK`	Search for metrics related to a company based on its central index key
`citeyear`	Search metrics related to a specific year
`page`	Pagination of the results (100 results per page)

4 Data Analysis

We carried out our data analysis on two datasets. We first provide some descriptive statistics regarding our primary dataset obtained from SEC and then perform statistical inference on both datasets to study selected research hypotheses regarding the CSR performance of companies.

4.1 Descriptive Statistics

The dataset obtained from SEC comprises 50,219 US companies. A large number of the collected companies (43%) belong to Finance, Insurance and Real Estate Industry, 20% to Manufacturing, 12% to Services and 7% to Transportation, Communications, Electric, Gas and Sanitary Service. It is common practice among companies to have a different location for their headquarters compared to the location of incorporation. Even though most of the collected companies are located in California (16%), New York (16%) and Texas (8%), a large percentage of these companies are incorporated in Delaware (52%) and Nevada (22%), most likely due to their highly attractive tax rates. Delaware is considered as a domestic tax haven and its paradigm has been extensively studied [8,25].

We also collected about 3,500 Conflict Minerals Reports related to 1,265 companies. The main industry the companies report in those is Manufacturing (76%), followed by Retail Trade (7%) and Services (7%) (Fig. 3). Figure 4 depicts

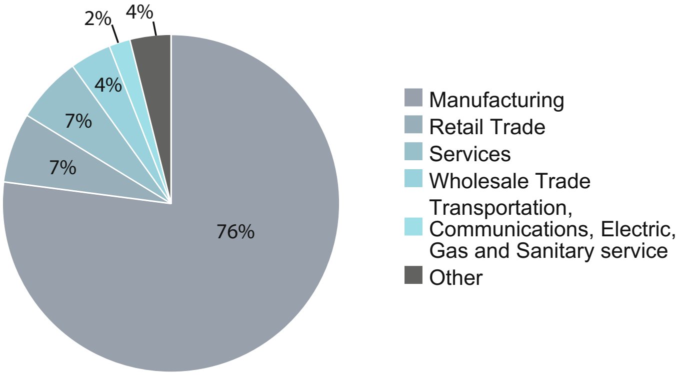

Fig. 3. Companies industry reporting conflict minerals.

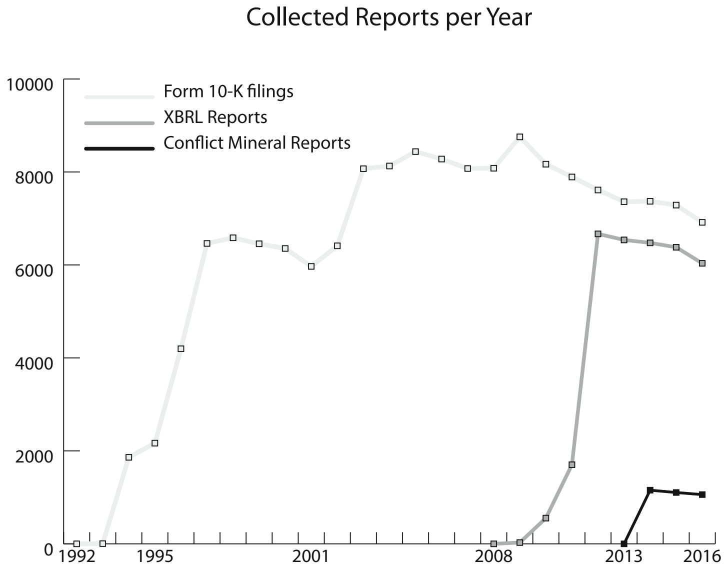

Fig. 4. Overview of the different types of collected reports per year.

the collected reports per year. It is noteworthy that companies started publishing financial annual reports in 1993 and over the years an increasing number of companies publish financial statements. In 2009, 27 companies started using XBRL in Form 10-K and 10-Q reporting. Companies started using XBRL more extensively in 2013 (6,668 companies published Form 10-K in XBRL format).

4.2 Statistical Inferences on CSR Performance

In this section, we are going to demonstrate the usefulness of the available financial open data for inference extraction regarding the social performance of companies. In particular, we perform data analysis on two datasets. The first contains more than 25,500 financial facts (from XBRL reports) for about 7,700 companies, and the second combines data about the environmental performance of 465 companies based on their green score in Newsweek 2016 Green List.

Financial facts dataset. Over the years, large corporations developed tax avoidance strategies to reduce their effective tax rate. These strategies are legitimate and companies adopting such strategies are not associated with inappropriate actions. For instance, the US government is actually incentivizing companies and individuals to increase their spending on charities by providing tax deductions in relation to such spending. In the following, we are going to study a different aspect of tax avoidance. Undistributed earnings of foreign subsidiaries (also known as permanently reinvested earnings) have been increasing over the

Table 5. Non-parametric correlation of the undistributed earnings to foreign subsidiaries to profits and R&D expenses. Each variable is perfectly correlated with itself (r = 1) and to this end the significance value is not calculated and is annotated in the table with ".".

		v1	*v2*	*v3*
Undistributed earnings of foreign subsidiaries (*v1*)	Correlation coefficient	1.000	0.637^{**}	0.582^{**}
	Sig. (2-tailed)	.	0.000	0.000
	N	3643	3512	2039
Profit (*v2*)	Correlation coefficient	0.637^{**}	1.000	0.135^{**}
	Sig. (2-tailed)	0.000	.	0.000
	N	3512	24278	7786
Research and development expense (*v3*)	Correlation coefficient	0.582^{**}	0.135^{**}	1.000
	Sig. (2-tailed)	0.000	0.000	.
	N	2039	7786	8040

**Correlation is significant at the 0.01 level (2-tailed).

last few years[10]. Corporations re-invest these earnings to their subsidiaries to avoid taxation. These practices can eventually lead to economic contraction and a shift of jobs overseas [16].

Laplante and Nesbitt [12] investigated the relationship between the trapped cash in foreign subsidiaries, the permanently reinvested earnings and the foreign cash. They found that the three terms are intersecting and also discovered that companies which have trapped cash are more likely to invest heavily in Research & Development. To this end, we are going to study the relationship of the undistributed earnings of foreign subsidiaries with not only profits but also with R&D Expenses. In this context, we are studying the following two hypotheses:

- **Hypothesis 1.** There is no correlation between the reported undistributed earnings of foreign subsidiaries, the profits and the R&D expenses.
- **Hypothesis 2.** There is no difference between the means of reported undistributed earnings of foreign subsidiaries between the years 2013, 2014, 2015 and 2016.

To study the first hypothesis, we used correlation tests and more specifically the Spearman's correlation coefficient that is appropriate for non-normal data. Both Kolmogorov-Smirnov and Shapiro-Wilk tests for normality showed that the undistributed earnings of foreign subsidiaries, profits and R&D expenses are significantly non-normal with $p < .001$. The Spearman's rho correlation coefficient showed a significant positive association between all three variables and resulted in the rejection of the null Hypothesis 1. The corresponding correlation coefficients are presented in Table 5. It is noteworthy that the relationship

[10] https://blogs.wsj.com/cfo/2015/04/24/companies-leaving-trillions-in-cash-overseas/.

Table 6. Top 10 companies ranked by the amount of profits in 2016. UEFS stands for Undistributed Earnings of Foreign Subsidiaries, and FIRE stands for Finance, Insurance and Real Estate.

Company	Industry	State of Inc.	Profits (in $M)	UEFS (in $M)
Apple Inc.	Manufacturing	California	$ 45,687	$ 109,800
JP Morgan Chase & Co.	FIRE	Delaware	$ 24,733	$ 38,400
Berkshire Hathaway Inc.	FIRE	Delaware	$ 24,427	$ 12,400
Wells Fargo & Co.	FIRE	Delaware	$ 22,045	$ 2,400
Alphabet Inc.	Services	Delaware	$ 19,478	$ 60,700
Bank of America Corp.	FIRE	Delaware	$ 17,906	$ 17,800
Microsoft Corp.	Services	Washington	$ 16,798	$ 124,000
Wal Mart Stores Inc.	Retail Trade	Delaware	$ 15,080	$ 26,100
Citigroup Inc.	FIRE	Delaware	$ 14,975	$ 47,000
Gilead Sciences Inc.	Manufacturing	Delaware	$ 13,488	$ 37,600

of R&D expenses appears to be stronger to undistributed earnings of foreign subsidiaries than to profits in terms of the coefficient value. This indicates that companies that permanently reinvest earnings of foreign subsidiaries are likely to invest heavily on R&D. Strong positive association between the undistributed earnings of foreign subsidiaries and company profits are also observed. Hence, we could claim that companies with high profits tend to adopt tax avoidance practices in terms of higher undistributed earnings of foreign subsidiaries.

Table 6 presents the top 10 companies in terms of profits in 2016 along with their undistributed earnings of foreign subsidiaries. The highest profits were reported by Apple Inc. followed by JP Morgan Chase & Co. We notice that in many cases the undistributed earnings of foreign subsidiaries exceed the amount of profits. The amount of the undistributed earnings of foreign subsidiaries for Apple Inc. was almost 2.5 times greater than the reported profits while for Microsoft was 7 times.

Continuing, the one-way analysis of variance (one-way ANOVA) test is commonly used to research the existence of statistically significant differences among the means of two or more groups and it would be a good candidate for researching the second hypothesis. However, our data failed the assumptions of normality for each category of the independent variable (year) as well as the homogeneity of variances assumption. To test normality within the groups (group by year), we used Kolmogorov-Smirnov's test and the undistributed earnings of foreign subsidiaries in reporting years 2013 ($D(859) = .401, p < .001$), 2014 ($D(911) = .407, p < .001$), 2015 ($D(939) = .403, p < .001$) and 2016 ($D(934) = .414, p < .001$) were found significantly different from the normal distribution. To test variance homogeneity, we used Lavene's test and the variances of the undistributed earnings of foreign subsidiaries in the four groups (grouped by year) were found significantly different, $F(3, 3639) = 5.32, p < .01$ and thus the homogeneity of variance could not be assumed.

In order to detect differences between the groups we used the non-parametric test of Kruskal-Wallis, equivalent to the one-way ANOVA. The results showed that the undistributed earnings of foreign subsidiaries were significantly different over the years, $H(3) = 14.867, p < .01$ and that led to the rejection of the null hypothesis 2. Mann-Whitney tests were used to reveal between which years there was a significant difference in the undistributed earnings of foreign subsidiaries. A Bonferroni correction was applied and thus all effects are reported at a .0083 level of significance. Significant differences in the undistributed earnings of foreign subsidiaries were detected only regarding years 2013 to 2015 ($U = 372062.5, r = -.067$) and 2016 ($U = 361298.0, r = -.086$). We can conclude that the undistributed earnings of foreign subsidiaries in 2015 and 2016 were significantly higher compared to 2013. Finally, Jonckheere's test revealed a significant trend in the data: the median of the undistributed earnings of foreign subsidiaries are significantly increased over the years ($J = 2621883.0, z = 3.790, r = .063$).

To sum up, the above analysis provides evidence that over the years profitable companies try to find ways to avoid taxation.

Newsweek green rankings dataset. Newsweek publishes rankings based on the environmental performance of the 500 largest publicly-traded Global and US companies since 2009. They score companies based on their performance on eight specific environmental aspects. In this section, we are going to study if and how we can extract useful inferences by combining data residing at different sources. More specifically, we combined the 2016 Newsweek green scores of 465 US companies with several financial statements extracted from 10-K filings. We are going to study the following two hypotheses:

- **Hypothesis 3.** There is no correlation between the Newsweek green score with Profits, Good Will and R&D Expenses.
- **Hypothesis 4.** There is no difference in the environmental performance (in terms of Newsweek green score) of companies reporting conflict minerals with those that do not.

To study Hypothesis 3, we used correlation tests for discovering relations between the four variables of interest (Newsweek green score, Profits, Good Will and R&D Expenses). The selection of appropriate correlation tests required to check for normality inside our sample with Shapiro-Wilk's test of normality. The distributions of all four variables were found to be significantly non-normal with $p < .001$ in all cases. Thus, we select Spearman's correlation coefficient to discover correlations and the strength of the existing relationships between them. Spearman's correlation coefficient in Table 7 indicates that there is strong positive correlation between all of them. This suggests that the higher the Newsweek green score is the higher the Profits, the Good Will and the R&D Expenses. Note that the highest coefficient occurs in the case of Profits with $r = .401$. Consequently, we could claim that big companies, in terms of financial performance, tend to adopt more sustainable practices.

Table 7. Non-parametric correlation of the Newsweek green score to Profits, Good Will and R&D Expenses. Each variable is perfectly correlated with itself (r = 1) and to this end the significance value is not calculated and is annotated in the table with ".".

		v1	*v2*	*v3*	*v4*
Newsweek green score (*v1*)	Correlation coefficient	1.000	0.401**	0.197**	0.232**
	Sig. (2-tailed)	.	0.000	0.000	0.005
	N	465	416	383	145
Profit (*v2*)	Correlation coefficient	0.401**	1.000	0.478**	0.437**
	Sig. (2-tailed)	0.000	.	0.000	0.000
	N	416	416	371	141
Good Will (*v3*)	Correlation coefficient	0.197**	0.478	1000	0.459**
	Sig. (2-tailed)	0.000	0.000	.	0.000
	N	383	371	383	136
Research and development expense (*v4*)	Correlation coefficient	0.232**	0.437**	0.459**	1.000
	Sig. (2-tailed)	0.005	0.000	0.000	.
	N	145	141	136	145

**Correlation is significant at the 0.01 level (2-tailed).

To test Hypothesis 4, we performed Mann-Whitney's non-parametric test since our data violated the normality assumption. Results indicate that the environmental performance of companies that report conflict minerals is significantly higher than those companies that do not ($U = 20358, r = -.15, p < .01$).

5 Conclusions

Over the years, more data are available online regarding the financial and social performance of companies. In this work, we examined how financial open data can be harnessed to extract useful inferences regarding specific aspects of Corporate Social Responsibility. We managed to collect numerous financial and conflict minerals reports from the SEC database and to extract a plethora of financial facts regarding the financial performance of companies. We performed data analysis on two datasets and tested four hypotheses regarding corporate tax avoidance and environmental performance. Integrating data from different sources is a challenging task but it could lead to important insights. To this end, we make the collected data available to third parties though a Web API for further use and investigation.

Future work could focus on researching more complex hypotheses as well as on integrating data from different sources related not only to the financial but also to the social performance of companies.

Acknowledgements. This work has been supported by the WikiRate and ChainReact projects, partially funded by the European Commission under contract numbers FP7-609897 and H2020-687967 respectively.

References

1. Bauman, C.W., Skitka, L.J.: Corporate social responsibility as a source of employee satisfaction. Res. Organ. Behav. **32**, 63–86 (2012)
2. Bowen, H.R., Johnson, F.E.: Social Responsibility of the Businessman. Harper, New York City (1953)
3. Bragdon, J., Marlin, J.: Is pollution profitable? Risk Manag. **19**(4), 9–18 (1972)
4. Cai, Y., Jo, H., Pan, C.: Doing well while doing bad? CSR in controversial industry sectors. J. Bus. Ethics **108**(4), 467–480 (2012)
5. Carroll, A.B.: Corporate social responsibility: evolution of a definitional construct. Bus. Soc. **38**(3), 268–295 (1999)
6. Cheng, B., Ioannou, I., Serafeim, G.: Corporate social responsibility and access to finance. Strateg. Manag. J. **35**(1), 1–23 (2014)
7. Dahlsrud, A.: How corporate social responsibility is defined: an analysis of 37 definitions. Corp. Soc. Responsib. Environ. Manag. **15**(1), 1–13 (2008)
8. Dyreng, S.D., Lindsey, B.P., Thornock, J.R.: Exploring the role delaware plays as a domestic tax haven. J. Financ. Econ. **108**(3), 751–772 (2013)
9. Eckbo, B.E., Nygaard, K., Thorburn, K.S.: Does gender-balancing the board reduce firm value? (2016)
10. Ferrara, E., De Meo, P., Fiumara, G., Baumgartner, R.: Web data extraction, applications and techniques: a survey. Knowl.-Based Syst. **70**, 301–323 (2014)
11. Johnston, K.A., Beatson, A.T.: Managerial conc1eptualisation of corporate social responsibility: an exploratory study (2005)
12. Laplante, S.K., Nesbitt, W.L.: The relation among trapped cash, permanently reinvested earnings, and foreign cash. J. Corp. Financ. **44**, 126–148 (2017)
13. Loughran, T., McDonald, B., Yun, H.: A wolf in sheeps clothing: the use of ethics-related terms in 10-k reports. J. Bus. Ethics **89**, 39–49 (2009)
14. Bromiley, P., Marcus, A.: The deterrent to dubious corporate behavior: profitability, probability and safety recalls. Strateg. Manag. J. **10**(3), 233–250 (1989)
15. Margolis, J.D., Elfenbein, H.A., Walsh, J.P.: Does it pay to be good? A meta-analysis and redirection of research on the relationship between corporate social and financial performance. Ann Arbor **1001**, 48109–1234 (2007)
16. Marr, C., Highsmith, B.: Tax holiday for overseas corporate profits would increase deficits, fail to boost the economy, and ultimately shift more investment and jobs overseas. Center on Budget and Policy Priorities, Washington, D.C. (2011)
17. Mills, R., et al.: WikiRate.org – leveraging collective awareness to understand companies' environmental, social and governance performance. In: Bagnoli, F., Satsiou, A., Stavrakakis, I., Nesi, P., Pacini, G., Welp, Y., Tiropanis, T., DiFranzo, D. (eds.) INSCI 2016. LNCS, vol. 9934, pp. 74–88. Springer, Cham (2016). doi:10.1007/978-3-319-45982-0_7
18. Moir, L.: What do we mean by corporate social responsibility? Corp. Gov.: Int. J. Bus. Soc. **1**(2), 16–22 (2001)
19. Namkung, Y., Jang, S.: Are consumers willing to pay more for green practices at restaurants? J. Hosp. Tour. Res. **41**(3), 329–356 (2017)
20. Parguel, B., Benoît-Moreau, F., Larceneux, F.: How sustainability ratings might deter greenwashing: a closer look at ethical corporate communication. J. Bus. Ethics **102**(1), 15 (2011)
21. SEC: Interactive data to improve financial reporting, June 2017. https://www.sec.gov/rules/final/2009/33-9002.pdf

22. Sethi, S.P.: A conceptual framework for environmental analysis of social issues and evaluation of business response patterns. Acad. Manag. Rev. **4**(1), 63–74 (1979)
23. Waddock, S.A., Graves, S.B.: The corporate social performance-financial performance link. Strateg. Manag. J. **18**(4), 303–319 (1997)
24. Wanderley, L.S.O., Lucian, R., Farache, F., de Sousa Filho, J.M.: CSR information disclosure on the web: a context-based approach analysing the influence of country of origin and industry sector. J. Bus. Ethics **82**(2), 369–378 (2008)
25. Wayne, L.: How delaware thrives as a corporate tax haven. N.Y. Times **30** (2012). http://www.nytimes.com/2012/07/01/business/how-delaware-thrives-as-a-corporate-tax-haven.html