easIE: Easy-to-Use Information Extraction for Constructing CSR Databases From the Web

VASILIKI GKATZIAKI and SYMEON PAPADOPOULOS, Information Technologies Institute (ITI), CERTH-ITI
RICHARD MILLS, University of Cambridge
SOTIRIS DIPLARIS, IOANNIS TSAMPOULATIDIS, and IOANNIS KOMPATSIARIS, Information Technologies Institute (ITI), CERTH-ITI

Public awareness of and concerns about companies' social and environmental impacts have seen a marked increase over recent decades. In parallel, the quantity of relevant information has increased, as states pass laws requiring certain forms of reporting, researchers investigate companies' performance, and companies themselves seek to gain a competitive advantage by being seen to operate fairly and transparently. However, this information is typically dispersed and non-standardized, making it complicated to collect and analyze. To address this challenge, the WikiRate platform aims to collect this information and store it in a standardized format within a centralized public repository, making it much more amenable to analysis. In the context of WikiRate, this article introduces easIE, an easy-to-use information extraction (IE) framework that leverages general Web IE principles for building datasets with environmental, social, and governance information from the Web. To demonstrate the flexibility and value of easIE, we built a large-scale corporate social responsibility database comprising 654,491 metrics related to 49,009 companies spending less than 16 hours for data engineering, collection, and indexing. Finally, a data collection exercise involving 12 subjects was performed to showcase the ease of use of the developed framework.

CCS Concepts: • **Information systems** → **Information retrieval**; *Web mining*; • **Mathematics of computing** → Exploratory data analysis;

Additional Key Words and Phrases: Information extraction, Web wrapper, corporate social responsibility (CSR), environmental, social, and governance (ESG)

ACM Reference format:
Vasiliki Gkatziaki, Symeon Papadopoulos, Richard Mills, Sotiris Diplaris, Ioannis Tsampoulatidis, and Ioannis Kompatsiaris. 2018. easIE: Easy-to-Use Information Extraction for Constructing CSR Databases From the Web. *ACM Trans. Internet Technol.* 18, 4, Article 45 (April 2018), 21 pages.
https://doi.org/10.1145/3155807

1 INTRODUCTION

Companies are increasingly expected to, and in some cases legally required to, report on their environmental, social, and governance (ESG) performance. The voluntary production of corporate

This work was supported by the WikiRate and ChainReact projects, partially funded by the EC under contracts FP7-609897 and H2020-687967.
Authors' addresses: V. Gkatziaki, S. Papadopoulos, R. Mills, S. Diplaris, I. Tsampoulatidis, and I. Kompatsiaris; emails: {vasgat, papadop}@iti.gr, rm747@cam.ac.uk, {diplaris, itsam, ikom}@iti.gr.

https://doi.org/10.1145/3155807

social responsibility (CSR) reports is now commonplace among large companies. For instance, in the past few years, legislation has come into effect that requires (1) companies that trade in the United States and file with the U.S. Securities and Exchange Commission (SEC) to produce conflict minerals reports [27], (2) companies that trade in the UK to publish statements about the steps they take to avoid slavery in their supply chains [17], and (3) Indian companies over a certain size to spend 2% of their profits on CSR activities [24]. Driving this trend is increased public awareness of the scale of corporate impacts and the dubious nature of some of these impacts. This increased awareness stems from the work of researchers and investigators who have exposed unethical and unsustainable behavior within companies and their supply chains.

WikiRate[1] is a platform for collecting and analyzing information about companies' ESG impacts in a transparent manner, with the aim of making that information accessible to all and using it to push for improved ESG performance on the part of companies. Underpinning WikiRate's approach is the concept of a "metric," which asks the same question about many different companies, and "metric values," which represent the answers to those questions for a particular company in a particular year. WikiRate metrics have been designed to accommodate many different types of information in a standardized format, including (1) low-level numerical indicators such as quantity of water use or greenhouse gas emissions; (2) binary answers to questions such as whether a company engages in a particular CSR-related practice; and (3) ratings of company performance as produced by external research, advocacy, and media organizations.

WikiRate's metric system [21] is open and transparent. Users have access to the methodology of the calculated metrics, as well as to the low-level data that are leveraged to create these metrics. All metric values have a source, and a reader can easily follow links to see these sources. Everything on the platform can be discussed and edited to facilitate discourse about how information should be interpreted and which questions are most important. Importantly, any user can create new metrics, posing new questions of companies' performance, allowing the voices of stakeholders that are marginalized in the CSR space to be heard—to ask the questions that matter to them and seek answers to those questions.

The subject of corporate ESG performance raises some issues related to the trustworthiness and sourcing of information. Much available data ultimately comes from companies themselves, either directly from their sustainability reporting or having been analyzed or processed by an intermediary [21]. Data that depicts a company as (un)ethical or (un)sustainable can have an impact on how that company is perceived and valued, and so the what and how of assessing performance are contested. Companies engage in this reporting for their own reasons and have control over their reporting output. Intermediaries that provide analyses of or assurances about performance may also be influenced by factors such as wanting to maintain a good relationship with covered companies to build influence, or in the other direction by a biased preconception of poor performance against one company or companies in general. WikiRate's design eschews the common Wiki model of relying on trusted sources to provide "objective facts" as the basis of the resource, favoring a model of representing who is saying what about companies, based on which evidence. All of WikiRate's metrics have a "designer" who formulated the question and methodology for determining the correct answer, and each answer to that question must be backed by citing a particular source. The WikiRate team does not control this system to decide which ways of assessing ESG performance or sources of evidence are legitimate; rather, any member of the community can become a metric designer or add metrics designed by some external entity. Within this open system, the visibility of metrics and their data is determined by user preference/voting, how often the metric is used in analyses, and the amount of data it has available.

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

There are three ways in which WikiRate is producing this comprehensive resource: (1) organizations or individuals who produce company-relevant data can create metrics to represent their data and import it in bulk; (2) any user can add a new metric value for a company, and this method is being used to extract information from semistructured sources such as CSR report documents; and (3) there are many existing publicly accessible and semistructured sources of company-relevant data scattered across the Web sites of different organizations. Without any technical support, the collection and integration of these data into the WikiRate platform would be a challenging and time-consuming task, as it requires one to build custom information extraction (IE) logic (also known as a wrapper) for each Web source. Web pages are semistructured documents, and different Web pages follow different structures. This article addresses the problem of extracting data from the Web into a structured format without the need to implement from scratch a separate wrapper for each source.

To this end, an easy-to-use open-source framework is proposed, named *easIE*, for extracting data from external sources related to companies. The framework supports easy generation of Web wrappers to perform IE from user-designated Web sources. As a result, users with little or no programming skills can contribute to the process of data gathering more actively by extracting data from both static and dynamic HTML pages simply by defining a configuration file. This is an important development for WikiRate, because the capacity to bring data to the platform at scale affords one some influence over what the resource has to say about ESG performance. By enabling additional users, without necessarily having technical expertise, to extract data from public sources and import this to WikiRate at scale, easIE helps to democratise this process, allowing the depiction of ESG performance that emerges from the data on the platform to be a better reflection of how the community wishes to assess this performance.

In a more direct and practical sense, the requirement to represent the source of every data point on WikiRate means that when data is extracted using easIE, the record must be augmented with details of where it came from so that this can be cited as the source on WikiRate and a user can follow this link to see where the data came from. Having companies as the subject raises an additional issue in that the precise name of the same company may vary between sources, and many legal entities may bear similar names. This introduces a step whereby as the data is imported to WikiRate, each entity represented must be matched to the appropriate WikiRate company and these matches verified. Otherwise, the process could distort the representation of what the source says about company performance by associating the data with an incorrect company.

Using this framework, we have extracted financial and CSR data from several external sources about several companies and integrated this within the WikiRate database. The proposed framework can be used to create large CSR databases, but it is possible to be adapted for extraction of data from Web sources related to different disciplines. The easIE data model is developed around the notion of companies. The remaining entities of the model are tied on companies. Adapting easIE to different domains should be straightforward, as the approach behind metrics extraction from Web sources remains the same and the only change pertains to the central entity of the model. For instance, if one is interested in environmental data about cities, the main model entity would be the city. The contributions of our work are not referred with regard to the easIE data model but to the overall work. To this end, we suggest to present the contributions as follows:

Overall, the contributions of this work can be summarized in the following points:

— easIE, an easy-to-use open-source[2] framework that extracts data about companies from heterogeneous Web sources in a semi-automatic manner.

[2]The source code is publicly available at https://github.com/MKLab-ITI/easIE.

— a large database of over 650,000 metric values about more than 49,000 companies, generated with only minimal configuration effort for easIE in very short time.
— a unified data model for representing company and CSR information and a Web API that makes the collected data available to third parties.

To the best of our knowledge, easIE is the first framework that addresses the problem of extracting CSR data from Web pages in a well-structured format. Users with limited programming skills can easily create large CSR databases, as demonstrated in Sections 4.1 and 4.2. Finally, in Section 4.3, we conduct a few data exploration exercises by visualizing parts of the collected data to demonstrate the potential of the framework for assisting CSR research.

2 RELATED WORK

This article touches upon two areas that typically have been studied separately in the past: CSR reporting and data extraction from the Web.

2.1 CSR Reporting

Adam and Shavit [1] proposed a theoretical model that evaluates the conditions under which corporations increase their CSR efforts to be included in a CSR ranking. Belkaoui and Karpik [3] researched which factors affect corporate decisions to communicate CSR information to stakeholders, and they developed a model that assesses the correlation between social information disclosure and corporate social and financial performance. Wanderley et al. [32] assessed whether CSR communication on the Web is influenced by the country or industry of corporations. Their analysis concluded that corporate communication on the Web is indeed influenced by country and less by the industry sector. Finally, Scalet and Kelly [26] investigated the impact of CSR rankings on corporate behavior. They discovered that corporations usually do not try to improve their rankings as reported by several CSR rating agencies, but they are mostly focusing on demonstrating their CSR plans and actions. The WikiRate approach posits that the effectiveness of ratings as a means of incentivizing companies to improve their performance can be significantly increased by producing these ratings in a more transparent and open manner.

A popular method for quantifying CSR performance is content analysis of corporate communication reports. For instance, Widiarto Sutantoputra [34] performed content analysis on reports that were written using the global reporting initiatives (GRI) framework and used standardized measures to allow comparison of CSR performance between different companies. Tate et al. [30] performed content analysis on CSR reports to discover the factors that CSR-aware companies address to improve their ESG performance relatively to their internal functionality and supply chain management.

There are some existing services that ingest company-related ESG information from many sources and use this to produce ratings of company behavior following certain themes (e.g., CSRhub.com, ethicalconsumer.org). WikiRate differs from these services in two ways. First, WikiRate is a public platform and all information is available at no cost, whereas most similar services are subscription based. Second, other services do not show the precise workings of how data of different types is turned into a rating. In contrast, ratings on WikiRate are entirely transparent, as any reader can understand how a particular rating works, see the raw data the rating for on which any company is based, and links to the sources that this raw data comes from.

To return to the example of Sutantoputra's framework for ratings based on GRI indicators [34], WikiRate's approach is to define metrics that capture this indicator-level data and populate these metrics with data so that one can see the data a company reported for every indicator. WikiRate's rating system can be used to reproduce Sutantoputra's scoring methodology, and once in place,

that rating will be dynamically calculated and displayed for every company where the relevant data is available, allowing for easy comparisons between companies' performance through this framework. A reader who opens this rating will see how the rating is calculated, what the company reported for each indicator, and how that affected their score. This rating could then be incorporated into a higher-level rating alongside other methods of rating companies' ESG performance, each weighted according to the level of importance afforded by that metric's designer.

2.2 Web Data Extraction

The problem of extracting data from the Web is also known as Web data extraction (WDE) [10]. WDE techniques allow collecting large amounts of data from the Web and are used in a wide range of research fields. To extract data from a particular Web source, a Web wrapper needs to be created (i.e. a procedure that seeks a set of data of interest and automatically extracts it from the page into a structured format) [10]. The Web wrapper life cycle consists of three stages: (1) wrapper generation, in which the wrapper is fully defined for a source of interest; (2) wrapper execution, in which the wrapper is executed on a set of target pages and the underlying data are extracted and stored; (3) wrapper maintenance, where the wrapper is updated due to changes in the HTML structure of the Web source.

Building hand-coded Web wrappers requires effort and is time consuming. A plethora of research studies have been conducted in the field of WDE and specifically in the automation of wrapper generation. Wrapper induction is the task of generating wrappers automatically by learning extraction rules from a set of manually labeled HTML documents. Kushmerick [16] was the first to introduce a wrapper induction algorithm for automatic data extraction. His algorithm requires a number of labeled instances based on which a wrapper is constructed, and to ensure the quality of the produced wrapper, probably approximately correct (PAC) analysis is used. After Kushmerick, numerous researchers studied the problem of supervised wrapper induction [2, 8, 14, 22, 35].

One of the main downsides of supervised wrapper induction is the amount of human effort required in the process of labeling. Several studies tried to eliminate human effort by proposing unsupervised methods. Crescenzi et al. [7] proposed RoadRunner, a system for automatic wrapper generation that seeks patterns that are repeated between different but similar-structured pages to extract the data of interest. Wang and Lochovsky [33] tried to uncover part of the Hidden Web by proposing DeLa, a system that sends queries through HTML forms to obtain parts of back-end databases and extract important part of data into well-structured tables. DeLa generates wrappers automatically based on HTML-tag structures and by discovering continuous repeated patterns. Liu et al. [18] also tried to uncover part of the Hidden Web by utilizing the visual features of the hidden Web pages. Gentile et al. [12] proposed an unsupervised data extraction methodology that generates dictionaries from Linked Data repositories based on the type of information to extract. Then, these dictionaries are used to annotate given Web pages and discover repeated patterns. Wong and Lam [35] proposed a framework based on Bayesian learning and expectation-maximization for addressing the problem of wrapper adaptation, which leverages previously learned wrappers to extract data from unseen Web sources.

The implementation of WDE tools comes with several challenges. For instance, there are numerous cases that the aforementioned methodologies fail or are unable to extract important pieces of information and therefore require ad hoc post-processing logic, which makes their use complicated for users with few programming skills. Our goal is to create a generally applicable WDE tool, but the design process of such a tool needs to take into consideration our application domain (collecting data about companies' ESG performance). In particular, an important challenge that we needed to address has been the maximum possible reduction of user effort by providing a high degree of automation. The objective of this tool has been to enable users with few programming

skills, to contribute more actively to the process of data collection. The design of easIE was inspired by Ducky [15], a system where data extraction is implemented by means of a set of rules that are serialized in a configuration file. Compared to Ducky, easIE also supports the IE from dynamic HTML pages.

3 FRAMEWORK DESCRIPTION

The easIE framework[3] enables users with limited programming skills to contribute more actively to the process of data collection by only defining a configuration file for each Web source of interest. The framework then generates and executes a wrapper for each source and achieves high accuracy and adaptivity to any Web page structure.

3.1 Data Model

We model CSR data around three main entities: companies, metrics, and articles:

- *Companies* represent corporations and are associated with a name, a set of aliases, country, and Web site.
- *Metrics* are pieces of information (quantitative or qualitative) related to companies; metrics are primarily defined by their name and value, the year to which they refer, and information related to the source that was used to derive them.
- *Articles* refer to online resources (posts) that refer to CSR activities and issues in relation to a company or a set of companies.

3.2 Architecture

To tackle the diversity of Web sources where CSR information resides, easIE employs a semi-automatic WDE approach that generates a custom wrapper for a particular Web source based on a set of handcrafted data extraction rules. Users only need to define these rules and store them in a configuration file for easIE to generate and execute a custom wrapper to collect the target data from the selected Web source. Defining these rules is a simple process and makes the collection of CSR data possible from both static and dynamic HTML pages.

Figure 1 provides an illustration of the easIE architecture. A configuration file is provided as input by the user, and the *Configuration File Validator* checks for inconsistencies. In case such inconsistencies are detected, appropriate messages are generated to guide users to necessary corrections. Once the validation is successful, the *Configuration Parser* translates the configuration file and provides input (settings and extraction rules) to the *Web Wrapper Generator*, which creates a custom wrapper for the target Web source. If a navigation step is defined in the configuration file (Table 1), then the *Web Wrapper Generator* produces a new wrapper for the new source. This wrapper (or wrappers in case navigation is required) is then applied on the page by the *Wrapper Executor*. As a final step, the *Data Handler* performs data integration by matching records referring to the same company and updating the CSR database. The *Data Handler* is also responsible for handling the different units of measure in case of numeric values. Note that easIE supports the extraction of text in UTF-8 encoding, and thus it supports multiple languages in all types of metric.

Two types of wrapper are generated depending on whether the HTML content is static or dynamic. Each of them executes the following processing flow: page fetching, data extraction, data postprocessing, and start-over—that is, identifying the next page to fetch from within the currently processed Web page. In case of static pages, the executor applies the wrapper on a sequence of Web pages that is defined either through a pagination operation (*Pagination Iterator*) or through

[3]Documentation and demonstration of the framework is available at http://easie.iti.gr.

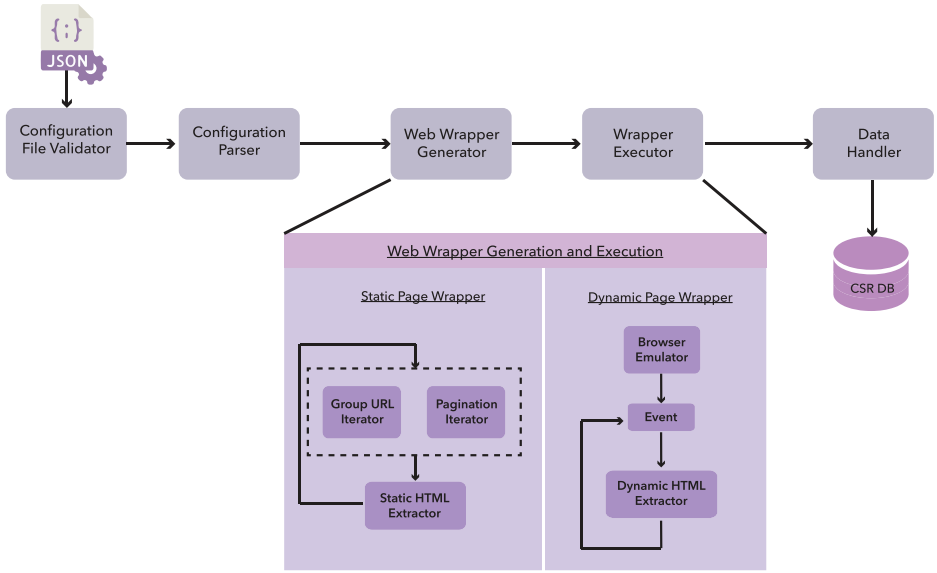

Fig. 1. Overview of the easIE architecture.

Table 1. Main Elements of the easIE Configuration File

Field	Type	Description
`metadata`	Object	Optional field based on the Dublin Core schema for adding annotations to the configuration file
`url`	Object	URL of target page comprised of two subfields: `base_url` and `relative_url`
`source_name`	String	User-selected source name
`company_info`	Array	Array of JSON elements further specified in Table 2
`metrics`	Array	Array of JSON elements further specified in Table 2
`dynamic_page`	Boolean	Specifies whether the page is dynamic or static
`events`	Array	Array of JSON elements further specified in Table 3
`list_selector`	String	CSS selector that points to the DOM subtree where the data of interest reside
`next_page_selector`	String	CSS selector that points to the next page element in the HTML document
`group_of_urls`	Array	Group of similarly structured URLs from which to extract data
`crawl`	Object	Nested JSON with configuration settings that specifies the next Web page to be fetched
`store`	Object	Specification of storage

a group of URLs (*Bunch URL Iterator*). At the moment, easIE supports search from HTML forms indirectly by defining the search values in the relative URL. In case of dynamic pages, the executor launches a browser emulator and executes a set of user events (as specified in the configuration file), namely click and scroll events, to fetch and extract the target data.

Wrapper maintenance is an important stage in the Web wrapper's life cycle. When changes occur in the HTML structure of a Web page, this may require that users redefine the extraction rules in the configuration file. To this end, proper messages are displayed and prompt users to check the validity of the extraction rules or for possible changes in the Web page in cases where it is not possible to detect the defined elements or to extract content from them.

The framework was developed in Java. JSoup[4] and Selenium[5] were used to automatically build the wrappers capable of extracting data from both static and dynamic HTML pages, respectively. More specifically, JSoup is a Java HTML parser and offers an API for manipulating data using document object model (DOM)[6] and CSS queries. Selenium is a Java library for browser automation, and we used it to automate the construction of wrappers for dynamic HTML pages and automate click and scroll down events. Additionally, the underlying database for storing the collected data is mongoDB,[7] a schema-free document-oriented database using BSON (binary JSON) as the serialization format. This offers the required flexibility allowing us to easily extend the framework in the future and add or remove more fields or nested fields where needed. Finally, we make the collected data available through a RESTful API that was developed based on the Java Servlets framework [6].

3.3 Configuration File

To extract and collect data from a Web source, a configuration file needs to be prepared by the user. This is defined in JSON format and consists of a set of elements (listed in Table 1) that specify (1) the type of data that is to be extracted and (2) details about the steps of the wrapper processing workflow (fetching, data extraction, post-processing, start-over).

The entry point for a wrapper is the `url` field that determines the seed Web page that the wrapper will first fetch to start the data collection process for the source under `source_name` (i.e., a label referring to the particular source). Note that users need to define the base and relative URL of the seed page on the `url` field of the configuration file for all relative URLs of the page to be easily converted to absolute ones based on the `base_url`. Then, the configuration file should specify the (1) attributes that can be extracted for each company (Web site, country, sector, etc.) through the `company_info` elements and (b) the `metrics` that will be extracted and associated with each company (these will be further described in the following).

Another essential configuration element is a Boolean flag (`dynamic_page`) indicating whether the target page should be handled by the wrapper as static or dynamic. In case of static pages, the target data are available upon page load; instead, for dynamic pages, the configuration file should also specify the set of user `events` (click or scroll) that are required to load the necessary HTML snippets in the browser emulator. Algorithm 1 describes the wrapper processing workflow for static Web pages, and Algorithm 2 describes the corresponding one for dynamic Web pages.

The actual data extraction process is initiated by pointing the wrapper to the DOM subtree specified by the `list_selector` field. This is a CSS selector that points the wrapper to the set of HTML elements that contain the target information about company attributes and metrics. The wrapper iteratively visits all elements of the selected subtree and applies the data extraction rules specified by `metrics` and `company_info` (Table 2) that map parts of these elements to the respective structured fields. For instance, to select each row from a table with id `list-table-body` we should declare `'list_selector' :'#list-table-body > tr'` and to map

[4] https://jsoup.org/.
[5] http://www.seleniumhq.org/.
[6] DOM is a programming interface that handles HTML and XML documents as trees.
[7] http://mongodb.com/.

ALGORITHM 1: Static Web Page Wrapper

```
Data: configuration
Result: extracted_company_info, extracted_metrics
page = fetch_page(configuration.url);
extract_company_info(page, configuration.company_info);
(I) extract content from defined elements
(II) post-processing of the extracted content if the replace field is defined
extract_metrics(page, configuration.metrics);
go-to step 1 if:
(I) configuration.group_of_urls is defined
    configuration.url = configuration.group_of_urls.next
(II) configuration.next_page_selector is defined
    configuration.url = next_page
(III) configuration.crawl is defined
    configuration = configuration.crawl
return extracted_company_info, extracted_metrics;
```

ALGORITHM 2: Dynamic Web Page Wrapper

```
Data: configuration
Result: extracted_company_info, extracted_metrics
start_browser_emulator(configuration.url)
execute_events(configuration.events)
page = fetch_page(configuration.url);
extract_company_info(page, configuration.company_info);
(I) extract content from defined elements
(II) post-processing of the extracted content if replace field is defined
extract_metrics(page, configuration.metrics);
go-to step 1 if:
(I) AFTER_EACH_EVENT type of events occur
(II) configuration.crawl is defined
    configuration = configuration.crawl
return extracted_company_info, extracted_metrics;
```

the third column to a metric with the name "Newsweek Global Rank 2014" the respective CSS selector would be `{'label':'Newsweek Global Rank 2014', 'selector':'td:nth-child(3)', 'type':'text'}`. In addition to `text`, the `type` field can also take one of the following values:

- `boolean`: Transforming the corresponding value of the selected element to Boolean
- `numerical`: Transforming the corresponding value of the selected element to numerical and assigning the proper units of measure if available
- `categorical`: Assigning the corresponding value of the selected element as is
- `link`: Corresponding to the `href` attribute of the selected element
- `image`: Corresponding to the `src` attribute of the selected element
- `list`: Referring to a list of text elements
- `other`: A string specifying the name of the attribute of interest (e.g., `'type': 'src'`).

In several cases, it is necessary to extract parts of the selected text element. To address this requirement, easIE supports processing the selected pieces of text by a simple declarative

Table 2. Specification of Metrics and Company_info

label	User-defined metric name or company info name
selector	CSS selector that expresses the extraction rule for extracting the target piece of information
type	Type of extracted data (text, Boolean, numerical, categorical, link, image, etc.)
replace	Pair of regex and with fields that define a post-processing step for the extracted data
regex	array of regex expressions specifying the patterns to be replaced
with	array of strings to replace the patterns detected by regex

Note: For Company_info, users need to define at least one field with label equal to company_name to specify to which company the extracted metric(s) refer.

Table 3. Specification of Events for Dynamic Pages

type	Event type (CLICK or SCROLL_DOWN)
selector	CSS selector of the element where the event is going to be executed
times_to_repeat	Times to repeat the specified event
extraction_type	Specifies when to invoke the data extraction, one after each event (AFTER_EACH_EVENT) or once at the end (AFTER_ALL_EVENTS)

specification (replace) comprising a set of regex patterns accompanied by a set of string values (with) to replace the detected patterns. For instance, the specification 'replace': {'regex':[':.*'],'with':['']} removes the part of a string after and including the colon. Note that the find and replace operations are performed sequentially starting from left to right, and the number of the defined regular expressions should match the number of the defined replaces (in the 'with' field).

In numerous cases, a source of interest contains multiple pages with data. In such cases, instead of defining different configuration files and only changing the url field, the user needs to define an array of URLs (group_of_urls) to be collected by the same wrapper. Navigation within the page is enabled if the crawl field is defined. The user needs to declare a nested configuration file that takes as input the URL(s) obtained by the field named crawl_to. Another multi-page data collection scenario pertains to paginated data (i.e., entries spanning multiple pages with paging controls available to browse from one to the next). To enable paginated access to such data, one needs to specify the next_page_selector field by a CSS selector pointing to the element where the "next" button is located in the page.

A final option of page processing concerns the fetching and processing of HTML elements from dynamic Web pages. In such cases, the fetching and processing workflow is specified by the events field (Table 3). There are two types of event that can be executed by the browser emulator:

- CLICK: In this case, users have to also define the selector of the element to which the event will apply, as well as the number of times that the event will be executed (times_to_repeat).
- SCROLL_DOWN: In this case, the wrapper will execute the event until no more data can be further loaded from the page. Users may also specify the times to repeat the scroll down event, as there are pages that are based on an "infinite scroll" design.

If the `times_to_repeat` field is specified, then one also needs to specify the `extraction_type` field declaring whether data extraction should be performed after the execution of each event (`AFTER_EACH_EVENT`) or after the execution of all events (`AFTER_ALL_EVENTS`).

Finally, users can add annotations to the configuration file in the `metadata` field based on the Dublin Core schema.[8] More specifically, they can provide a small description about the configuration file, the date, creator, contributor, and so on. Additional details about the framework, several tools to facilitate the understanding of easIE, and several examples of configuration files can be found online in the Web page of the framework.[9]

4 EVALUATION

The two primary performance aspects of easIE that we evaluate include (1) flexibility in terms of different Web sources that can be handled and (2) ease of use, quantified with respect to the required effort by non-experts to ingest new sources. To this end, the framework was tested on the real-world task of collecting CSR metrics to contribute as input to the WikiRate platform. As a starting point, an extensive list of Web sources for CSR data was compiled by the WikiRate consortium and was made available to us. Potential sources were identified (1) from previous experience of researching companies' ESG performance, (2) through Google searches, and (3) by browsing the source lists of established aggregators of CSR information (e.g., CSRHub). Out of this list, several sources (Table 4) were selected as targets for data collection based on several criteria: sources needed to have company-level data that was immediately interesting to the audience of WikiRate users or lower-level data that could later be productively used in the rating system. Sources were prioritized based on the ease of extraction, the number of companies they covered, and the relevance of the information they contained.

We conducted two experiments: (1) *collection effort*, a controlled experiment on a set of 16 sources to measure the ease of use and speed of data collection, and (2) *collection scale*, a larger-scale data collection exercise to assess the flexibility of easIE. In addition, we conducted a few data exploration exercises by visualizing parts of the collected data to showcase their potential for additional insights regarding the CSR performance of companies. Finally, a data-driven hypothesis testing task was performed on three samples of data extracted by easIE, and three hypotheses were studied.

4.1 Collection Effort

In the first experiment, we assessed the effort required for learning and using the proposed framework, as well as the time required for the data collection and indexing. To this end, we asked 12 researchers from our research lab (nine male and three female, between ages 25 and 37 years) to perform a data collection exercise involving 16 CSR sources using easIE. All 12 researchers are computer scientists and familiar with the concepts of HTML, CSS, and DOM. They have also had some limited previous experience with extracting information from Web pages, but only by means of custom scripts and code. None of them was involved in any way in the development of easIE, nor was any familiar with the WikiRate project.

The subjects were given a short (15-minute) introduction to the concept of easIE, they were presented with a few example configuration files, and they were familiarized with the concepts of CSS selectors, as well as with the possibility of quickly forming appropriate selector expressions with the use of the browsers' built-in inspection tools. Overall, this learning phase lasted approximately half an hour. Then, they were handed the list of 16 selected sources and were asked to prepare one

[8] http://dublincore.org/documents/dcmi-terms/.
[9] http://easie.iti.gr/Documentation.html#config_examples.

Table 4. List of CSR Sources Along With Basic Statistics and Experimental Results

	Conf. Lines (#)	**Rules (#)**	**Type**	**Metrics (#)**	**et** (sec)	**net** (sec)	**acc** (%)	**ext** (sec)
	Collection Effort Experiment							
WBCSD	27	2	Static	171	68	839	83.33	5
Sustainability Asia Ranking	45	5	Static	100	121	703	98.33	4
RSPO	51	6	Satic	7,689	143	891	98.61	54
PETA: No Tests on Animals	32	3	Dynamic	3,700	142	403	72.22	76
PETA: Tests on Animals	32	3	Dynamic	258	22	278	72.22	59
PERI	70	8	Static	594	163	523	100	4
Newsweek Global Green List (2011, 2012, 2014, 2015)	**130**	**15**	Dynamic	16,156	**251**	**901**	98.33	145
Newsweek U.S. Green List (2011, 2012, 2014, 2015)	90	11	Dynamic	7,461	120	587	95.45	134
Access to Medicine Index	37	4	Static	140	97	321	100	7
Access to Nutrition Index	37	4	Static	636	48	297	97.92	10
Measure Up	101	12	Static	550	235	556	100	2
Good Company Index (2011, 2012, 2014)	77	9	Static	1,813	117	397	100	5
Forbes	74	8	Dynamic	10,000	120	279	100	2,616
Ethisphere's Most Ethical Companies (2007–2016)	42	4	Static	1,166	73	580	91.67	14
EPEAT	26	2	Static	59	50	326	100	3
BSR	26	2	Static	294	81	442	83.33	5
EICC	26	2	Static	105	43	392	83.33	4
2020 WOB	30	4	Static	1,796	102	510	91.67	44
	Collection Scale Experiment							
Forbes	83	9	Both	92,560	224	–	–	2,676
UN Global Compact	101	14	Static	64,308	491	–	–	2,580
Climate Counts	45	5	Static	634	179	–	–	48
CorporateKnights (2010–2016)	45	5	Static	1,892	179	–	–	684
Electronic Frontier Foundation (2011–2014)	591	29	Static	502	714	–	–	70
Environmental Investment Organization (2011, 2013)	340	26	Static	10,524	394	–	–	26
FCPA	1,771	22	Static	995	441	–	–	69
Influence Map	72	8	Static	480	266	–	–	151
Rankabrand	70	3	Static	347	545	–	–	46
ReportWatch	159	15	Static	2,400	233	–	–	25
EPA	94	10	Static	8,269	444	–	–	159
EU Transparency	182	20	Both	104,409	1,104	–	–	3,002
BCorporations	95	13	Static	18,621	421	–	–	994
Fortune	873	87	Static	2,980	1,995	–	–	544
Wikipedia	–	–	Static	102,219	–	–	–	10,800
Ethical Guide	–	–	Static	10,650	–	–	–	1,800
Security and Exchange Commission (SEC)	–	–	Static	154,442	–	–	–	18,185

Note: *et* and *net* stand for expert and nonexpert time (in seconds), respectively; *acc* stands for accuracy (percentage of correct extraction rules); and *ext* stands for execution time (in seconds).

configuration file for each of them while recording the required time for creating each file. In the end, they returned the prepared configuration files so that we could test their validity and correctness. It is noteworthy that four of the subjects managed to create all configuration files without any help, whereas the rest required some help with a couple of cases.

Table 4 (upper part) presents the results of this experiment. For each of the 16 tested sources, the table presents the number of lines, extraction rules, and extracted metrics, along with the time (in seconds) required by the subjects on average to create the respective configuration files and its correctness measured as the percentage of correct extraction rules. For comparison reasons, we also report the time that an expert (the lead easIE developer) spent on the creation of the corresponding configuration files. Finally, we report the execution time for extracting and storing

the desired data into the database. In total, the expert spent approximately 35 minutes to build the configuration files, whereas the nonexpert subjects spent on average a total of approximately 155 minutes for the same task and managed to compose correctly 92.6% of the extraction rules. Given the amount of collected information (52,688 metrics), this is certainly a reasonable amount of manual effort, which attests to the fact that easIE is a straightforward framework to use with an easy learning curve.

The most complex configuration file in terms of number of extraction rules and time required by an expert is the one associated with the Newsweek 2015 Global Green List. Furthermore, looking into the individual performance of the 12 subjects, we noted that there was limited intersubject variation. The majority of subjects required between 100 and 200 minutes to complete the task (one took only 50 minutes, whereas two took a bit more than 200 minutes); six (half) of the subjects achieved an accuracy (percentage of correct rules) of about 96% or higher, and the lowest accuracy among the subjects was 74%.

4.2 Collection Scale

In the second experiment, we assessed the scalability of the data collection mechanism by targeting additional sources, most of which were of higher complexity compared to the ones tested in the first experiment. Here, 17 new Web sources were targeted, eight of which required navigation within the Web page to obtain the data of interest. In particular, 6 Web sources had one or more dedicated company pages that included more information about the companies; Forbes, UN Global Compact, Climate Counts, Influence Map, BCorporations, and EU Transparency contained such dedicated Web pages. The aforementioned sources required one navigation step, except for UN Global Compact, which required two. As already mentioned, easIE allows users to navigate within Web pages and obtain additional information by declaring nested configuration files in the field `crawl` of the configuration file. Forbes was also included in the previous collection effort experiment, but in this case we also extracted the data that exists in the dedicated company pages.

Some additional sources were addressed that required important modifications in the framework. These sources were included in our study because of the rich CSR data they contained. More specifically, extracting data from Wikipedia, Ethical Guide, and the SEC required major adaptations in the data extraction logic. For Wikipedia, this was due to the fact that there is no uniform way of accessing lists of companies, and hence some custom crawling logic was built to extract a comprehensive list of Wikipedia URLs pointing to company articles. Then, for each article, structured information was extracted from the company infobox; however, the linking to the sources where the data came from (i.e., citations) was not consistent, and hence it required developing a custom technique to address this inconsistency. Extracting CSR data from the Ethical Guide Web page required extensive adaptations to the WDE logic mainly due to the fact that the HTML content was often not consistently structured. Last, obtaining financial and conflict minerals reports from the SEC required to search for the reports by company name.

The second experiment was performed by the lead developer of easIE. Combined, the data collection tasks conducted within the first and second experiments resulted in a database of 654,491 metrics related to 49,009 companies and required a total of less than 16 hours of data engineering (configuration file creation, tweaks in the WDE approach by the expert) and collection. It is noteworthy to point out that only extracting data from the SEC required more than 5 hours due to the rate limits of their Web site. The next slowest collection time is observed in case of Wikipedia, which took 3 hours. In total, 47% of the collected metric data points are numerical, 5% categorical, 39% textual, and 9% Boolean.

The main goal of WikiRate is to make the information accessible to everyone and use it to push for improved ESG performance by the companies. Thus, to integrate the extracted data to the

WikiRate platform and expose them to third parties, we made all collected data available through a RESTful API[10] that consists of two endpoints: companies[11] and metrics.[12] It is noteworthy that this database constituted 92% of the total number of metrics that were available in the WikiRate platform (the rest were manually contributed by members of the community) according to measurements that were carried out at the end of October 2015.

4.3 Data Exploration

Exploratory data analysis and visualization are becoming increasingly important means of summarizing and gaining insights from data. The CSR metrics collected with the help of easIE are leveraged by the WikiRate platform and community to produce useful insights regarding the CSR performance of companies.

4.3.1 Data Visualizations. A first type of valuable insights that can be gleaned from the large database of collected metrics relate to the positioning of a company's CSR practices with respect to the rest. This can be achieved by visualizing the distribution of a selected metric across the full set of companies in our collection and highlighting the value for the company of interest.

This provides a quick and easily digestible view on how the company performs in terms of the given metric relative to other companies. For instance, Figure 2 depicts the performance of Apple Inc. in terms of eight different metrics. In the case of the seven numeric metrics, larger values correspond to better performance. The last bar chart (Figure 2(h)) depicts the participation of Apple Inc. in several non-governmental organizations related to sustainable activities; out of the four such organizations, Apple Inc. participates in the highlighted three, and the height of each bar corresponds to the number of companies that participate in the respective organization. Overall, Apple Inc. seems to have moderate to good CSR performance in most of the depicted metrics. Nevertheless, it could improve its performance in terms of the number of women on the board of directors and the number of persons involved in the EU transparency activities. Such insights can be very useful both for the company management and for investigators and analysts who are interested in rating the company's practices.

Moreover, several interesting visualizations and insights can be gleaned from the associations between companies and topics that can be exported from CSR articles (see the data model in Section 3.1). As a proof of concept, we conducted a collection of 8,541 articles from the Business and Human Rights Resource Centre[13] based on easIE.[14] Since in many cases multiple companies may be discussed in the same article, it is possible to generate a graph among companies that depicts this type of co-occurrence. To help explore such relations, we developed a simple Web application (snapshot depicted in Figure 3). By selecting a node (company), the connections between this and other companies are presented along with a list of articles from where these relationships were extracted. Given that such articles typically discuss the CSR practices of multiple companies as an ensemble, this graph could help discover similarities among companies in terms of CSR practices. In addition, one may leverage the power of social network analysis methods, such as centrality analysis [5] and community detection [23], to obtain a more nuanced image of the interplay and relative position of companies with respect to CSR practices.

A second type of visualization that can be generated from CSR articles depicts the associations among companies, Web sources, topics, and geographical regions. An interactive Web application

[10] http://easie.iti.gr/csr_api.html.
[11] http://easie.iti.gr/csr_api/companies.
[12] http://easie.iti.gr/csr_api/metrics.
[13] http://business-humanrights.org/.
[14] A slightly modified data collection process compared to the one in Figure 1 was used for that purpose.

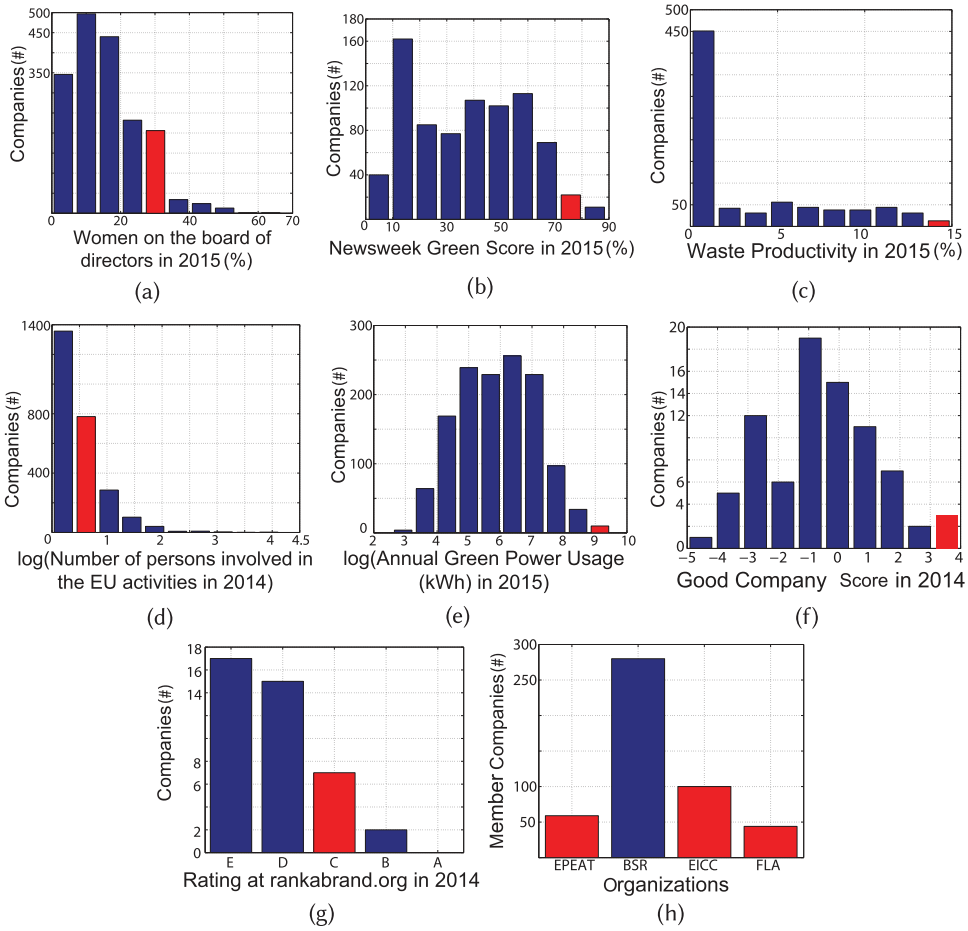

Fig. 2. Overview of Apple Inc. CSR performance in terms of eight metrics: percentage of women on the board of directors (a), Newsweek Green score (b), waste productivity (%) (c), number of persons involved in EU transparency activities (d), annual green power usage (kWh) (e), Good Company score (f), and rating at rankabrand.org (g); (h) depicts the membership of the company in several non-profit organizations. Values referring to the company are highlighted in red.

that visualizes such associations is available online.[15] For instance, Figure 4 depicts the focused graph around Google. It appears that Google is often discussed in the same articles with Hewlett Packard, Microsoft, General Electric, and Oracle, and the most popular tags of the articles that refer to Google include Environment, Health, Poverty, Labour, Security, Privacy, and Access to information. The most popular sources that publish articles about Google include Reuters, New York Times, Washington Post, Financial Times, Associated Press, BBC, and Huffington Post. Finally, geographical regions that are mentioned in articles about Google include USA, China, India, United Kingdom, Myanmar, Australia, and Vietnam.

[15] http://easie.iti.gr/HR-Graph/.

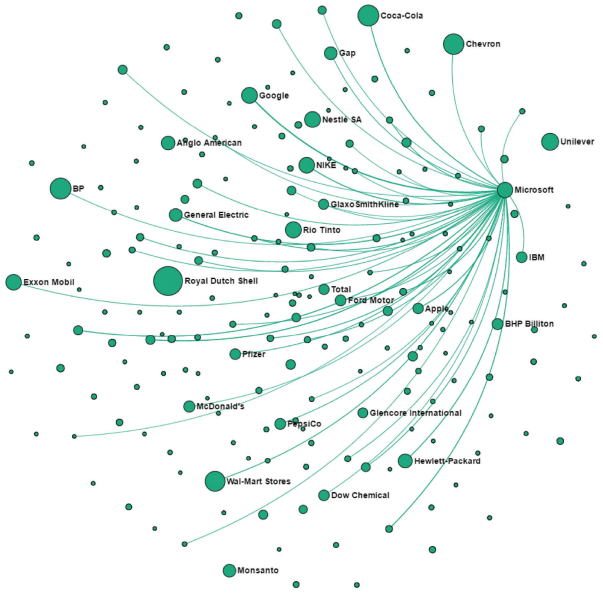

Fig. 3. Company relations graph: edges mean that the two companies are mentioned together in the same article. Available online: http://easie.iti.gr/HR-Companies-Relationships/.

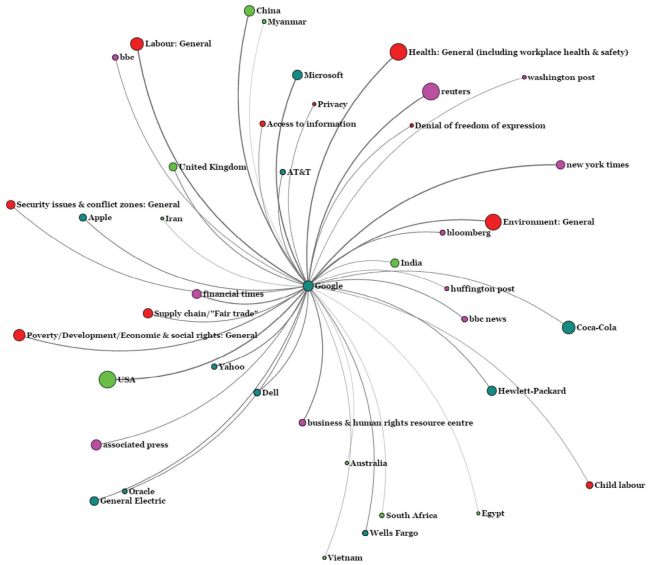

Fig. 4. Graph depicting associations between selected company and other entities (topics, sources, regions).

4.3.2 Data Analysis. Except for the useful visualizations, the collected data could be leveraged to extract useful data analytics, inferences, and prediction models related to companies' CSR behavior. In the past 40 years, a plethora of studies have been conducted that examine the relationship between corporate social performance (CSP) and corporate social profitability (CFP) [4, 9, 20, 29]. In the early 1970s, Bragdon and Marlin [4] were among the first to study this relationship from

Table 5. Communication Status: Company Type Cross Tabulation

Communication Status		Type of Company					
		Large	Medium	Employees (#)	Small	Small-medium	Total
Active	Count	875	1,522	28	3,082	2,247	7,754
	Within type of company	94.2%	83.9%	13.7%	62%	75.6%	71.3%
	Total	8.1%	14.0%	0.3%	28%	20.7%	71.3%
Noncommunicating	Count	54	291	176	1,867	726	3,114
	Within type of company	5.8%	16.1%	86.3%	38%	24.4%	28.7%
	Total	0.5%	2.7%	1.6%	17%	6.7%	28.7%
Total	Count	929	1,813	204	4,949	2,973	10,868
	Within type of company	100%	100%	100%	100%	100%	100%
	Total	8.5%	16.7%	1.9%	46%	27.4%	100%

both a manager's and an investor's perspective. To date, the relationship between CSP and CFP is still under investigation. Margolis et al. [19] conducted a meta-analysis on 167 studies to define which aspects of CSR are more effective in CFP, and they discovered a small but positive effect of CSP on CFP but a more important effect of specific aspects of CSP on CFP.

Numerous studies have also investigated the relationship between social responsibility disclosure and company size [25, 31]. Bigger companies tend to disclose more information about their social responsibility than smaller ones. Stanwick and Stanwick [28] studied the relationship between CSP and organizational size, financial, and environmental performance. To demonstrate the usefulness of ESG data analysis for conducting such inferences, we performed data analysis on three datasets collected with easIE to investigate three hypotheses. Note that the size of the samples that we study is highly dependent on the available data in our database regarding the tested variables.

The first dataset that we analyzed comprised collected data about companies that have been active participants in the UN Global Compact initiative or were in the past. In particular, we studied a set of 10,868 companies, being interested in the following hypothesis.

Hypothesis 1. *There is no difference in the achieved communication status of companies (Active, Non-communicating) between the different company types (small, small-medium, medium, etc.).*

The cross tabulation (Table 5) shows that among the different types of companies, larger companies tend to have higher percentage of active communication with the UN Global Compact initiative than smaller ones. More specifically, large companies have the highest percentage of active communication status at 94.2%, followed by medium companies at 83.9%. Consequently, smaller companies tend to be less consistent in their communication process. Companies with no employees have the highest percentage of non-communicating status at 86.3%, followed by small companies at 37.7%. The chi-square tests led to $\chi^2(4, n = 10{,}868) = 934.53$, $p < .001$, and the result indicates that there is a strong association between company type and communication status. Therefore, the result does not support null hypothesis 1, and there is a significant difference in the achieved communication status of companies between the different company types.

Next, a sample of 1,094 companies located in the United States was studied. More particularly, we wanted to study the following hypothesis.

Hypothesis 2. *There is no correlation between the percentage of women on the board of directors and the net income attributable to the company in 2015.*

Table 6. Results of the Games-Howell Post Hoc Test

Multiple Comparisons

Dependent Variable: green_score

Games-Howell

(I) country	(J) country	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
china	france	–0.374049*	0.027890	0.000	–0.44895	–0.29914
	germany	–0.320929*	0.033226	0.000	–0.41161	–0.23025
	japan	–0.257089*	0.030264	0.000	–0.33751	–0.17667
france	china	0.374049*	0.027890	0.000	0.29914	0.44895
	germany	0.053120	0.038237	0.514	–0.04964	0.15588
	japan	0.116960*	0.035694	0.010	0.02232	0.21160
germany	china	0.320929*	0.033226	0.000	0.23025	0.41161
	france	–0.053120	0.038237	0.514	–0.15588	0.04964
	japan	0.063841	0.040001	0.391	–0.04292	0.17061
japan	china	0.257089*	0.030264	0.000	0.17667	0.33751
	france	–0.116960*	0.035694	0.010	–0.21160	–0.02232
	germany	–0.063841	0.040001	0.391	–0.17061	0.04292

*The mean difference is significant at the 0.05 level.

The Spearman's correlation coefficient was used to check hypothesis 2 and discover the direction and strength of the relationship between the two variables (if such relationship exists). We preferred Spearman's Rho correlation coefficient since the distribution of the variables are not normal. Spearman's correlation coefficient is the non-parametric equivalent of the Pearson correlation, and as Field [11] stated, "Spearman's correlation coefficient is a non-parametric statistic and so can be used when the data have violated parametric assumptions such as non-normally distributed data." The Spearman's correlation coefficient indicated that there is a strong relationship between the percentage of women on the board of directors and the net income of the company ($r_s = 0.238, p < .001$). This suggests that the higher the net income of the company, the higher the percentage of women on the board of directors. Consequently, we could claim that companies with higher net income tend to employ more women on the board of directors.

Finally, a sample of 115 companies that have been scored in the Newsweek Top Green Companies List in 2015 was studied. The headquarters of the selected companies are located in one of these four countries: Germany, France, China, or Japan. We chose only these four countries for demonstration purposes. The last hypothesis we wanted to investigate was the following.

Hypothesis 3. *There is no difference between the means of the CSR performance of companies according to Newsweek in different countries.*

To investigate the aforementioned hypothesis, we performed one-way analysis of variance (ANOVA), which is used to compare means of three or more groups by using the F distribution, which is a right-skewed probability distribution. The ANOVA results indicated a significant difference between the means of the CSR performance and the rejection of null hypothesis 3. As our data did not meet the homogeneity of variances assumption, we performed the Games-Howell post hoc test to reveal between which groups the differences in the mean of CSR performance are significant. The results of the post hoc test are presented in Table 6. The mean of the CSR performance of companies located in China presents a significant negative mean difference from the means of the other three countries. However, France presents a positive significant mean difference of the CSR performance with China and Japan. The mean of the CSR performance of companies located in Germany displays a significant positive difference only with that of companies from China. Finally, Japan presents a significant positive difference with China and a negative one with

France. The results showed that companies from China tend to environmentally perform worse compared to companies from all studied countries (France, Germany, and Japan) in 2015, whereas companies from France outperformed companies from Japan and China but performed similarly with companies from Germany.

This is a preliminary work on data analysis of the collected data by easIE. The proposed framework enables users to collect a large amount of company-related data and increases the confidence in the analysis results due to the amount of the available information. The aggregated information from multiple sources allows us to investigate more complex hypotheses. Our follow-up work in this area is presented in Gkatziaki et al. [13].

5 OPEN ISSUES AND FUTURE WORK

The development of easIE is a work in progress and leaves several issues open for future work. Our main goal is to extend easIE so that users, WikiRate community members, with no programming skills will be able to contribute more actively in the process of data collection. Consequently, we are planning to implement a graphical interface that will allow users with no programming skills to interactively build Web wrappers, thus further reducing the barriers of entry to the framework. Such extensions would enable crowd-sourcing data collection and adding value to collective awareness platforms for sustainability and social innovation.[16] It is noteworthy that easIE is already used within the WikiRate,[17] ChainReact,[18] and hackAIR[19] projects.

In the future, we are also going to extend easIE to provide domain-specific IE for supporting the automatic extraction of data from Web sources by leveraging Linked Open Data related to corporate financial data and CSR performance. Another area of interest concerns the task of *company mapping*, a data integration task where company-related data residing at different sources need to be matched. Our goal is to exploit state-of-the-art relation extraction methodologies to achieve high precision and recall on data integration.

Finally, we plan to extract more useful analytics and inferences similar to the ones presented in Section 4.3. Our goal is to create regression models by exploiting the collected data to predict companies' behavior regarding certain aspects of CSR performance.

6 CONCLUSIONS

This article addressed the problem of gathering corporate ESG performance data from diverse Web sources and integrating it into an open CSR database. To this end, easIE, an easy-to-use framework, was proposed that enables users to extract information of interest from selected Web sources by creating appropriate configuration files using simple sets of rules. To demonstrate the ease of use of the framework, a user study was carried out on 12 non-experts and showed that using easIE can be learned in very short time and that the majority of users could create almost 100% accurate configuration files for performing data extraction. In addition, to demonstrate the scale of data collection that is possible with the help of the framework, a selected set of Web sources were processed to collect approximately 650,000 metrics related to more than 49,000 companies, spending only a few hours of data engineering, collection, and indexing. A first set of statistics and visualizations was also presented on top of the collected data, and a first set of analysis results were presented on the collected data to showcase the potential of extracting useful inferences regarding corporate ESG performance.

[16] https://ec.europa.eu/digital-single-market/en/collective-awareness.
[17] http://wikirate.org/.
[18] http://chainreact.org/.
[19] http://hackair.eu/.

REFERENCES

[1] Avshalom Madhala Adam and Tal Shavit. 2008. How can a ratings-based method for assessing corporate social responsibility (CSR) provide an incentive to firms excluded from socially responsible investment indices to invest in CSR? *Journal of Business Ethics* 82, 4, 899–905.

[2] Tobias Anton. 2005. XPath-wrapper induction by generalizing tree traversal patterns. In *Proceedings of Lernen, Wissensentdeckung and Adaptivitt (LWA'05)*. 126–133.

[3] Ahmed Belkaoui and Philip G. Karpik. 1989. Determinants of the corporate decision to disclose social information. *Accounting, Auditing and Accountability Journal* 2, 1, 1–16.

[4] J. H. Bragdon Jr. and J. A. Marlin. 1972. ls pollution profitable? *Risk Management* 19, 9–18.

[5] Peter J. Carrington, John Scott, and Stanley Wasserman. 2005. *Models and Methods in Social Network Analysis*. Vol. 28. Cambridge University Press, Cambridge, MA.

[6] Danny Coward and Yutaka Yoshida. 2003. *Java Servlet Specification Version 2.3*. Sun Microsystems.

[7] Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. 2001. RoadRunner: Towards automatic data extraction from large Web sites. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB'01)*, Vol. 1. 109–118.

[8] Nilesh Dalvi, Ravi Kumar, and Mohamed Soliman. 2011. Automatic wrappers for large scale Web extraction. *Proceedings of the VLDB Endowment* 4, 4, 219–230.

[9] Islam Elshahat, Clark Wheatley, and Ahmed Elshahat. 2015. Is pollution profitable? A cross-sectional study. *Academy of Accounting and Financial Studies Journal* 19, 2, 59.

[10] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. 2014. Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems* 70, 301–323.

[11] Andy Field. 2013. *Discovering Statistics Using IBM SPSS Statistics*. Sage.

[12] Anna Lisa Gentile, Ziqi Zhang, Isabelle Augenstein, and Fabio Ciravegna. 2013. Unsupervised wrapper induction using linked data. In *Proceedings of the 7th International Conference on Knowledge Capture*. ACM, New York, NY, 41–48.

[13] Vasiliki Gkatziaki, Symeon Papadopoulos, Sotiris Diplaris, and Yiannis Kompatsiaris. 2017. Large-scale open corporate data collection and analysis as an enabler of corporate social responsibility research. In *Proceedings of the International Conference on Internet Science*.

[14] Chun-Nan Hsu and Ming-Tzung Dung. 1998. Generating finite-state transducers for semi-structured data extraction from the Web. *Information Systems* 23, 8, 521–538.

[15] Kei Kanaoka, Yotaro Fujii, and Motomichi Toyama. 2014. Ducky: A data extraction system for various structured Web documents. In *Proceedings of the 18th International Database Engineering and Applications Symposium*. ACM, New York, NY, 342–347.

[16] Nicholas Kushmerick. 1997. *Wrapper Induction for Information Extraction*. Ph.D. Dissertation. University of Washington.

[17] Legislation.gov.uk. 2015. Modern Slavery Act 2015. Retrieved March 15, 2018, from http://www.legislation.gov.uk/ukpga/2015/30/section/54/enacted.

[18] Wei Liu, Xiaofeng Meng, and Weiyi Meng. 2010. Vide: A vision-based approach for deep Web data extraction. *IEEE Transactions on Knowledge and Data Engineering* 22, 3, 447–460.

[19] Joshua D. Margolis, Hillary Anger Elfenbein, and James P. Walsh. 2007. 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.

[20] Jean B. McGuire, Alison Sundgren, and Thomas Schneeweis. 1988. Corporate social responsibility and firm financial performance. *Academy of Management Journal* 31, 4, 854–872.

[21] Richard Mills, Stefano De Paoli, Sotiris Diplaris, Vasiliki Gkatziaki, Symeon Papadopoulos, Srivigneshwar R. Prasad, Ethan McCutchen, Vishal Kapadia, and Philipp Hirche. 2016. WikiRate.org—leveraging collective awareness to understand companies' environmental, social and governance performance. In *Proceedings of the International Conference on Internet Science*. 74–88.

[22] Ion Muslea, Steve Minton, and Craig Knoblock. 1998. Stalker: Learning extraction rules for semistructured, Web-based information sources. In *Proceedings of the AAAI-98 Workshop on Artificial Intelligence and Information Integration*. 74–81.

[23] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos. 2012. Community detection in social media. *Data Mining and Knowledge Discovery* 24, 3, 515–554.

[24] Ashok Prasad. 2014. India's new CSR law sparks debate among NGOs and businesses. Retrieved March 15, 2018, from http://www.theguardian.com/sustainable-business/india-csr-law-debate-business-ngo.

[25] M. Purushothaman, G. Tower, R. Hancock, and R. Taplin. 2000. Determinants of corporate social reporting practices of listed Singapore companies. *Pacific Accounting Review* 12, 2, 101–133.

[26] Steven Scalet and Thomas F. Kelly. 2010. CSR rating agencies: What is their global impact? *Journal of Business Ethics* 94, 1, 69–88.
[27] SEC. 2014. Fact Sheet: Disclosing the Use of Conflict Minerals. Retrieved March 15, 2018, from https://www.sec.gov/News/Article/Detail/Article/1365171562058.
[28] Peter A. Stanwick and Sarah D. Stanwick. 1998. The relationship between corporate social performance, and organizational size, financial performance, and environmental performance: An empirical examination. *Journal of Business Ethics* 17, 2, 195–204.
[29] Ambec Stefan and Lanoie Paul. 2008. Does it pay to be green? A systematic overview. *Academy of Management Perspectives* 22, 4, 45–62.
[30] Wendy L. Tate, Lisa M. Ellram, and Jon F. Kirchoff. 2010. Corporate social responsibility reports: A thematic analysis related to supply chain management. *Journal of Supply Chain Management* 46, 1, 19–44.
[31] Ken T. Trotman and Graham W. Bradley. 1981. Associations between social responsibility disclosure and characteristics of companies. *Accounting, Organizations and Society* 6, 4, 355–362.
[32] Lilian Soares Outtes Wanderley, Rafael Lucian, Francisca Farache, and José Milton de Sousa Filho. 2008. CSR information disclosure on the Web: A context-based approach analysing the influence of country of origin and industry sector. *Journal of Business Ethics* 82, 2, 369–378.
[33] Jiying Wang and Fred H. Lochovsky. 2003. Data extraction and label assignment for Web databases. In *Proceedings of the 12th International Conference on World Wide Web*. ACM, New York, NY, 187–196.
[34] Aries Widiarto Sutantoputra. 2009. Social disclosure rating system for assessing firms' CSR reports. *Corporate Communications: An International Journal* 14, 1, 34–48.
[35] Tak-Lam Wong and Wai Lam. 2010. Learning to adapt Web information extraction knowledge and discovering new attributes via a Bayesian approach. *IEEE Transactions on Knowledge and Data Engineering* 22, 4, 523–536.

Received December 2016; revised September 2017; accepted October 2017