BVL INTERNATIONAL SCIENTIFIC SYMPOSIUM ON LOGISTICS
June 19 – 20, 2023

# Development of an ETL-Pipeline for Automatic Sustainability Data Analysis

*Jettarat Janmontree, research associate,*
*Institut für Logistik und Materialflusstechnik, Otto von Guericke Universität, Magdeburg*

*Aditya Mehta, master student,*
*Systems Engineering for Manufacturing, Otto von Guericke Universität, Magdeburg*

*Hartmut Zadek, professor,*
*Institut für Logistik und Materialflusstechnik, Otto von Guericke Universität, Magdeburg*

**Extended Abstract**

**Summary.** As the scientific community and organizations increase their investments in sustainable development, the phrase is increasingly being used deceptively. To be sustainable, one must examine all three aspects, namely environmental, social, and economic. The release of sustainability reports has generated a vast amount of data regarding company sustainability practices. This data demands time and effort to evaluate and extract meaningful information. This research aims to create criteria that include a list of keywords for analyzing sustainability reports. Using these criteria, a proposed application based on the concepts of Extract, Transform, Load (ETL) was developed to automatize the process of data analysis. The results generated by the ETL tool can be used to conduct qualitative and quantitative assessments of the organization's sustainability practices as well as compare the transparency in sustainability reporting across different industries.

**Keywords.** accuracy of sustainability reporting, semi-automated text-mining, extract-transform-load-tool

## 1. Introduction

During the past years, sustainability has been a main focus of corporate practices. Many companies have introduced new goals to be more sustainable and reference corporate sustainability (CS). These sustainability strategies aim to fulfill both financial goals while addressing topics related to sustainable development including environmental, social, and governance (ESG) factors. Furthermore, pressure from stakeholders such as the government, non-governmental organizations (NGOs), customers, or suppliers influences corporations to implement sustainability practices (Visser 2022, 79-80). Therefore, sustainability reporting became an essential communication outlet for many organizations. Companies also use sustainability reports to publicly announce their achievements regarding their sustainability goals

and to address their performance on all sustainability aspects (Siano et al. 2016, 3). This led to an increasing number of sustainability reports published in recent years (Mutiha 2022, 1).

According to the research, the vast majority of investors as well as consultants, and financial researchers find sustainability reports to be helpful in evaluating corporate sustainability practices (Petrescu et al. 2020, 22; Al-Shaer and Hussainey 2022, 1-2). It is stated that the fundamental of sustainability reporting should be for an organization to be transparent about its impacts on all sustainability aspects. (Global Reporting Initiative 2002, 4). Even though sustainability reporting has been commonly used to evaluate and compare sustainability practices, there is unclear how much of the sustainability topic should be included in the sustainability report (Janmontree 2021, 49-51).

At present, several organizations have established regulations that set a standard for sustainability reporting. However, sustainability reporting is not yet considered mandatory for many companies. Corporations have been given the opportunity to freely decide on the content of the reports (Chen and Bouvain 2009, 302). This gives an opportunity for organizations to exploit the concept of sustainability reporting. Organizations can potentially decide to omit or minimize certain information that could be perceived negatively by the public (Du and Yu 2021, 256). Such practices can cause ambiguity in the outcome and reduce the effectiveness of communication. The term commonly used for this practice is Greenwashing. It is best defined as the attempt of corporates to present misleading information to the stakeholders (Moodaley and Telukdarie, 2023, 4; de Freitas Netto et al. 2020, 6-7).

Evaluating and extracting useful information from sustainability reports requires time and effort. In addition, with the increasing number of sustainability reports published each year, a large scale of data regarding corporates' sustainability practices was generated. The Extract-Transform-Load (ETL) systems are used to transfer data from one or more sources into a database. It has the potential to be utilized for the purpose of reporting, analysis, and generating business insights. This paper focuses on developing an ETL tool to analyze sustainability reports and to examine the sustainability practices of various industrial factors.

The paper is structured as follows. In section two, the methodology of this research is explained as well as the result of this research is presented. Finally, the conclusion and discussion of this study are described in the last section.

## 2. Research Methodology and Results

The research work was conducted in three phases as seen in Figure 1. In this research, we investigated the criteria for sustainability performance measurement. We used the method of summative content analysis to assess the scope of the keywords with the help of the developed ETL tool. Our analysis revealed several key findings by comparing the result of six industrial sectors. The following sections provide a detailed description of the research methodology and findings.

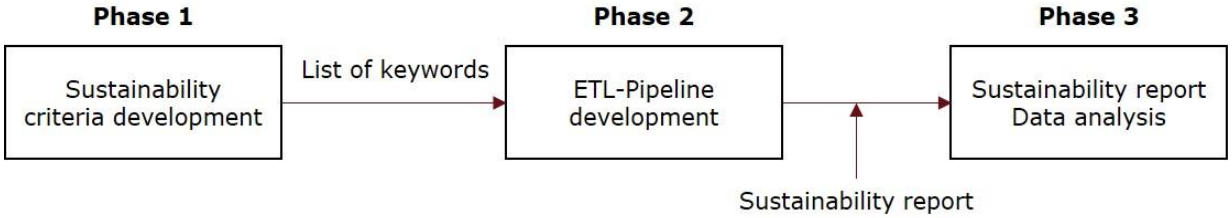


*Figure 1. Research methodology*

## 2.1. The Development of the Sustainability Keywords

The first phase is to develop a list of sustainability criteria (keywords). These keywords are based on the TBL framework which covers three dimensions of sustainability namely, social, environmental, and economic. Under the economic aspect, governance has also been included to broaden the scope of the keywords. A search for publications, journals, and articles was done using the Scopus, Elsevier, IEEE, and MDPI databases.

The literature search generated a list of keywords that were classified into three domains of sustainability. In each aspect of sustainability, two categories were established, including Functional Category, and Main Category. The latter is used to describe the end-point impact or final effect of the sustainability assessment process (Janmontree 2021, 66) as illustrated in Figure 2. To identify the crucial details of each functional category, a set of keywords is used. A total of 262 keywords were selected that cover relevant areas and criteria related to sustainability.

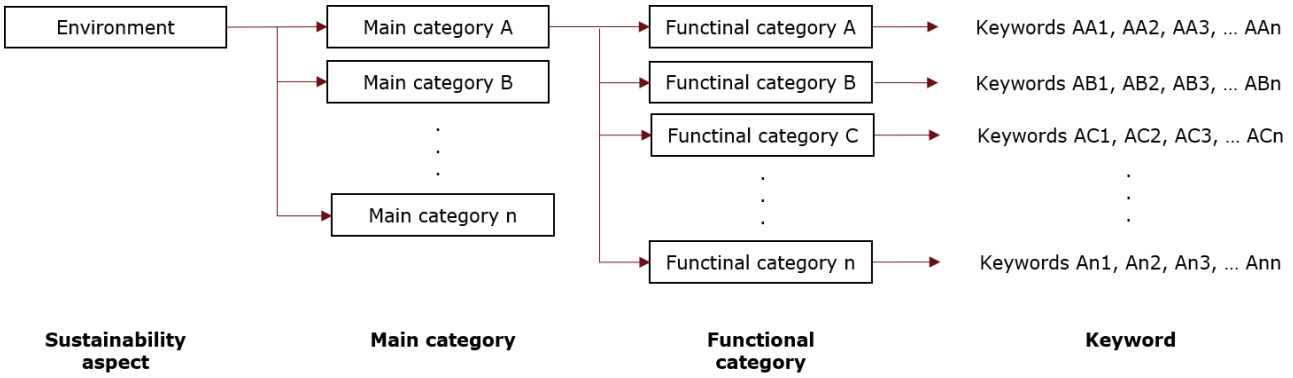


*Figure 2. Structure of the keywords*

## 2.2. Sustainability ETL Pipeline Architecture

The second phase is to develop an algorithm for a tool that could automate the assessment process. Along with the concept of ETL, a semi-automated text-mining technique is utilized in this study in order analyze sustainability reports and evaluate the level of coverage for each sustainability aspect. The ETL tool's algorithm combines both summative content analysis (quantitative) and sentence extraction (qualitative) techniques. By using a predefined list of sustainability keywords, the tool performs summative content analysis on the data retrieved from the document such as corporates' sustainability reports (Figure 3). The coverage of the sustainability keywords contributes towards the related main categories. Upon processing, the tool generates a PDF document as an output to summarize an overview of the result.

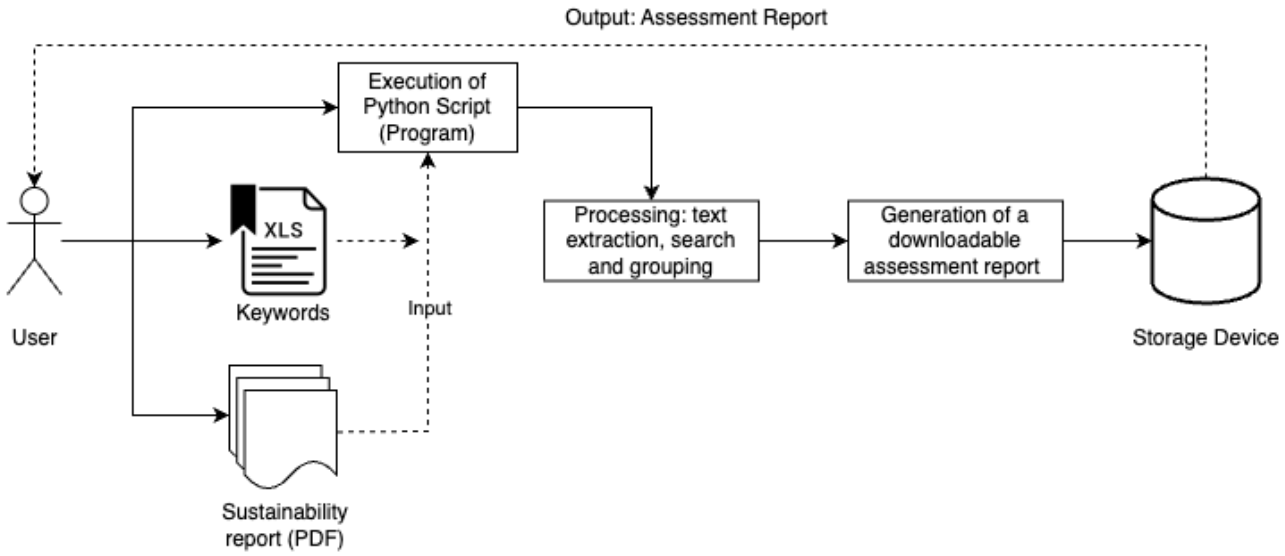


*Figure 3: ETL pipeline for sustainability performance assessment*

The tool also provides an additional Excel file that details the sentences, keywords, and corresponding areas of sustainability from the report. This alternative approach offers a more granular understanding of the report's sustainability-related content. The tool's robustness makes it suitable for use across various research areas for performance assessment purposes. Figure 3 shows an overview of the ETL tool framework.

### 2.3. Industry Benchmarks

The proposed ETL tool was used to evaluate the sustainability reports of six industrial sectors, including the agriculture, automotive, aviation, chemical, retail, and textiles industries. There are in total 20 companies from each sector were selected. The sustainability reports, along with the selected sustainability keywords file in .xlsx format, were used as inputs for the proposed ETL tool.

Figure 4 shows the benchmark result. Overall, sustainability reports from all industries contain less information regarding economic sustainability. In particular, the chemical, and textile industries have relatively high coverage of information related to environment and social sustainability when compare to other industries. On the other hand, sustainability reports of the agriculture industry contain less information on two sustainability aspects, including environmental and social aspects. However, to determine the possibility of greenwashing practices in the agriculture sector, a thorough evaluation of the result should be conducted.

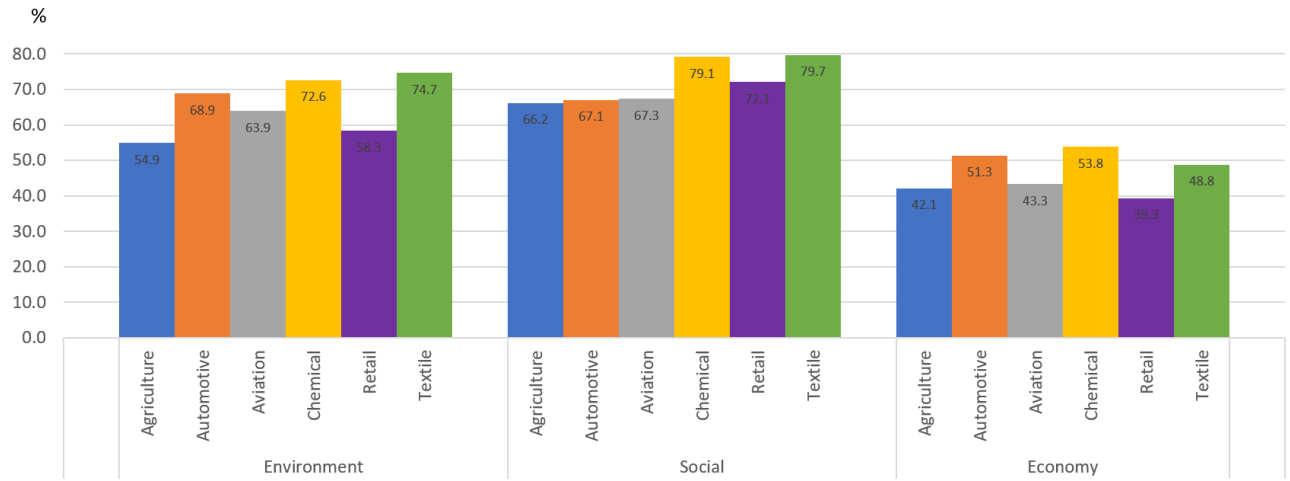


*Figure 4: Sustainability benchmark between each industry*

## 3. Conclusion and Discussion

Sustainability reports have been used to communicate and clarify corporates' practices. As the number of sustainability reports published by big corporates continues to increase, it is important to ensure that these reports are accurate and provide a complete picture of their sustainability performance. However, given a broad and complex nature of sustainability and a lack of specific regulations in sustainability reporting, this has led to the exploitation of the term sustainability.

To address this issue, we developed an Extract-Transform-Load (ETL) tool for analyzing sustainability reports. The tool demonstrated a high level of accuracy (84.70%) and the advantage of speed in assessing a sample sustainability report compared to manual evaluation. Furthermore, the tool is designed to be scalable, enabling it to process large sustainability reports quickly and efficiently. Custom keywords can be added to the tool to suit specific needs and changes.

Our analysis of sustainability reports from six different industries revealed that none of the selected industries has completely covered all aspects of sustainability. Overall, the economic aspect has the least coverage in keywords. This can be argued that, in general, sustainability reports tend to contribute more information about the environmental and social aspects. The agriculture industry has the worst performance in covering sustainability information, especially in the environmental aspect. This finding is concerning, as positive communication about environmental performance with a lack of actual practices can lead to the practice of greenwashing. Even though the proposed ETL tool demonstrates a fast and efficient approach to analyze the sustainability reports, further investigation should be made to ensure the accuracy of the result.

## References

Al-Shaer, Habiba, and Khaled Hussainey. 2022. "Sustainability reporting beyond the business case and its impact on sustainability performance: UK evidence." *Journal of Environmental Management* 311, 114883. https://doi.org/10.1016/j.jenvman.2022.114883.

Chen, Stephen, and Petra Bouvain. 2009. "Is corporate responsibility converging? A comparison of corporate responsibility reporting in the USA, UK, Australia, and Germany." *Journal of business ethics* 87, no. 1: 299–317.

de Freitas Netto, Sebastião Vieira, Marcos Felipe Falcão Petra, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. 2020. Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe* 32, no. 1: 1–12.

Du, Shuili, and Kun Yu. 2021. "Do Corporate Social Responsibility Reports Convey Value Relevant Information? Evidence from Report Readability and Tone." *Journal of Business Ethics* 172, no. 8: 253–274.

Global Reporting Initiative. 2022. A Short Introduction to the GRI Standards. Access on 24 March 2023). https://www.globalreporting.org/media/wtaf14tw/a-short-introduction-to-the-gri-standards.pdf

Janmontree, Jettarat. 2021. "Sustainability Performance Measurement Framework for the Product Life Cycle An Application for the Wind Turbine Industry." PhD diss., Otto-von-Guericke University.

Moodaley, Wayne, and Arnesh Telukdarie. 2023. "Greenwashing, Sustainability Reporting, and Artificial Intelligence: A Systematic Literature Review" *Sustainability* 15, no. 2: 1481. https://doi.org/10.3390/su15021481

Mutiha, Arthaingan H. 2022. "The Quality of Sustainability Report Disclosure and Firm Value: Further Evidence from Indonesia" *Proceedings* 83, no. 1: 26. https://doi.org/10.3390/proceedings2022083026

Petrescu, Anca Gabriela, Florentina Raluca Bîlcan, Marius Petrescu, Ionica Holban Oncioiu, Mirela Cătălina Türkeș, and Sorinel Căpuşneanu. 2020. "Assessing the Benefits of the Sustainability Reporting Practices in the Top Romanian Companies" *Sustainability* 12, no. 8: 3470. https://doi.org/10.3390/su12083470

Siano, Alfonso, Francesca Conte, Sara Amabile, Agostino Vollero, and Paolo Piciocchi. 2016. "Communicating Sustainability: An Operational Model for Evaluating Corporate Websites" *Sustainability* 8, no. 9: 950. https://doi.org/10.3390/su8090950

Visser, Wayne Africa Merlin-Tao. 2002. "Sustainability reporting in South Africa." *Corporate Environmental Strategy* 9, no. 1: 79–85.

This paper could be cited as:

Janmontree, Jettarat, Aditya Mehta, and Hartmut Zadek. 2023. "Development of an ETL-Pipeline for Automatic Sustainability Data Analysis." In *2023 International Scientific Symposium on Logistics: Conference Volume*, ed. by Thorsten Schmidt, Kai Furmans, Bernd Hellingrath, René de Koster, Anne Lange, and Hartmut Zadek, 63–68. Bremen: Bundesvereinigung Logistik. https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-855713. https://doi.org/10.25366/2023.125.