Content Analysis of Non-Financial Reporting of Sustainable Companies

Fumiharu Otsubo[a]　　Huang Haixiang[b]

Abstract

This study analyzed the characteristics of internationally evaluated corporate non-financial reporting (NFR) based on the International Integrated Reporting Council (IIRC) framework and Global Reporting Initiative (GRI) standards. The analysis target was 400 NFRs of Global 100 companies published in four years (2018-2021). Content analysis using text mining was used. We showed the extent to which the companies being evaluated are disclosing reports that match international disclosure standards and identified the position of NFRs by year and their overall trends. Understanding overall trends is useful for designing individual company reports and considering what information should be disclosed.

Keywords: information disclosure, NFR, NFI, integrated report, IIRC framework, GRI standards, text mining, corporate sustainability

1. Introduction-International trends in Non-Financial Information (NFI) disclosure

Since early 2000, there has been greater market interest in NFI. This information has been expanded regarding decision usefulness for investors and investment responsibility. Non-financial pressures are increasing internationally, as evidenced by non-financial initiatives, guidelines, changes in corporate valuation standards, and investment selection trends. These include Science Based Targets initiative (SBTi), United Nations Global Compact, Paris Agreement Targets, GRI standards, IIRC framework, Value Reporting Foundation (VRF), Task Force on Climate-related Financial Disclosures (TCFD), Sustainability Accounting Standards Board (SASB), Climate Disclosure Standards Board (CDSB), and Carbon Disclosure Project (CDP). Companies are under pressure to respond.

Corporate non-financial disclosure is traced back to the environmental reports of the 1980s. As environmental management became more widespread, companies attempted to reduce environmental costs and risks while improving their environmental performance. They disclosed the results to society through environmental reports. Subsequently, through GRI's international efforts, NFI extended to social and economic performance, and a comprehensive sustainability performance based on the triple bottom line (see Elkington 1997) was proposed. GRI guideline (2000, 2002, 2013, 2016) is primarily intended to provide information to multi-stakeholders. Around 2000, sustainability and CSR reports were published, and the information has been expanded and reorganized.

a　Faculty of Economics, Dokkyo University
b　Institute of Informatics, Dokkyo University

Therefore, to enrich or enlarge the information, there is an international trend toward materiality, key performance indicator (KPI) approaches to information selection, and the current integrated reporting, aiming to reorganize financial information and NFI. Therefore, although there are differences between countries and regions, the IIRC proposes an international discussion on how financial and non-information should be combined and disclosed for integrated reporting.

The IIRC was established by the Prince's Accounting for Sustainability Project (A4S) and GRI in 2010 to develop an international framework for integrated reporting. The draft and final International Integrated Reporting Framework were published in 2011 and 2013, respectively. This significantly influenced corporate disclosure practices. IIRC Integration has an impact on investor relations information because "an integrated report should be prepared primarily for providers of financial capital" (IIRC 2013, p8). As a result, integrated reporting entered a new phase with the release of the revised International Integrated Reporting Framework in 2021 and the establishment of the VRF through the merger of the IIRC and SASB.

2. Research question and related work

The global spread of integrated reporting has had a profound impact on traditional NFR practices, and increasingly diversified NFR. However, not all organizations have made progress toward integrated reporting, and traditional disclosure practices that suit the organization's characteristics prevail. Much of the content is freely described because most NFRs are voluntary disclosures. Although there is some structure commonality according to the guidelines, the individuality of the companies is demonstrated in detail.

Assuming diversified disclosure practices depend on the future size of the company, industry, etc., individual organizations have to consider the most appropriate reporting model for their organization. Additionally, they should consider the kind of effective information to be disclosed. Therefore, it is necessary to understand the position of one's organization among others. The organization can determine the reports to refer to in the future and the difference between the reports being evaluated and our own by understanding the organization's position.

It is necessary to ensure a sufficient number of analyses and individually scrutinize each report to analyze the overall trends and characteristics of NFR. However, since most report information is narrative, the scope of analysis is limited, and the number of analyses that can be handled is also difficult.

Peripheral devices and language processing technology have made processing many analyses possible in recent years. As a result, in accounting research, there have been many studies on using language processing technology to replace narrative information with quantitative information and derive empirical evidence (Türegün 2019). It also enables big data analysis through machine learning and attracts significant interest in analyzing trending words.

There are few studies on content analysis of NFI. Therefore, it is important to use language processing technology to quantify and analyze NFI because it is often difficult to recognize, measure, and understand NFI quantitatively (see also Goloshchapova et al. 2019, Barkemeyer et al. 2009, Teuteberg 2013, Modapothala et al. 2010, Bala et al. 2015, Rivera et al. 2014). These related studies differ in their target samples and scales, the parts of speech and terms to be extracted, and their analytical methods and processes (PN analysis, see also Teuteberg 2013, Shahi et al. 2014, Harymawan et al. 2020, word cloud analysis see also Fiandrino et al. 2021). However, content analysis focusing on narrative information has provided new insights.

This study identifies the characteristics of corporate NFRs by setting axes. The axes use the IIRC framework and GRI standards, significantly impacting corporate NFR.

3. Methodology

3-1. Sample selection

The study sample consisted of NFRs published by companies selected for the Global 100 index. The analysis covered four years from 2018 to 2021, with a sample size of 400 reports. The Global 100 ranks the world's most sustainable companies, published annually at the World Economic Forum. The sample for each year is based on the most recently disclosed reports, considering that the selection process starts several months before the previous year.

The Global 100 ranking is for publicly listed companies with $ PPP currency sales of $1 billion or more. The evaluation method begins with screening companies based on their financial health, product categories, and misconduct. Second, it scores the selected companies from the environmental, social, governance, and economic matrices and finally ranks them based on the results. Each year, the evaluation method was slightly modified, and the selected companies were highly replaceable.

Additionally, the NFR published by Global 100 companies is diverse because of voluntary disclosure, such as annual, integrated, CSR, NFR sustainability, and ESG reports (see Table 1).

Table 1 Sample type and numbers (2018-2021)

Type	Number of reports
Annual report	66
Integrated report	32
Sustainability report	171
CSR report	63
NFR	4
ESG report	7
Sustainability and CSR	5
Annual and Sustainability	5
Annual and CSR	1
Integrated and Sustainability	2
other	44
Total	400

Source: Companies in the Global 100 for 2018-2020.

3-2. Analytical model

This study aimed to understand the positional relationship by clustering each report content. First, PDF files were converted into text format, text preprocessing, and text clustering.

Currently, many tools are available for converting PDFs into text. Each has its advantages, but this study used the Python programming language (PyPDF2).

For text pre-processing, the main tasks are morphological analysis of words, conversion to normal form, recognition of named entity recognition (NER), removal of noise, accounting for word density, and calculation of the importance of each word.

Tokenizing words (morphological analysis) is easy using the Natural Language Toolkit (NLTK) since English documents have spaces between words. In order to convert the extracted words into normal form, the words need to be stemmed and lemmatized. For example, the plural form of a noun or the conjugated form

of a verb needs to be returned to its original form. The former ignores the context, while the latter takes the context into account when processing. Regarding the recognition of NER, this analysis is only based on the recognition by the NLTK dictionary. As an improvement, a corresponding technical term dictionary should be constructed. Besides, for noise elimination, we eliminate punctuation and stop words that are unnecessary for the analysis. In order to account for word density, the frequency at which a word appears in each sample is noted, and the importance of each word is calculated according to the term frequency-inverse document frequency (TF-IDF) formula (tfidf (t,d) =tf (t,d) × idf (t)), and a matrix table of each report and word is created (see Table 2).

Table 2 Matrix of terms used and their frequency of occurrence in each sample.

file_name	w_1	w_2	w_n
Criterion			
NFR a			
NFR b			
NFR n			

The TF-IDF method is often used to measure the importance of a sentence term in which a word appears (see also Rivera et al. 2014, Liew et al. 2014). The words used in the analysis were based on their appearance in IIRC and GRI. Therefore, the characteristics of each report are clarified based on these.

JMP statistical software was used for cluster analysis of each preprocessed sample's word lists and TF-IDF information. Although many cluster analysis methods exist, this study used the frequently used K-means method for analysis. However, as shown in Shahi et al. (2014), general documents and corporate reports classification is still different. Therefore, we should consider suitable methods for classifying corporate reports as a future task.

4. Results

4-1. Distribution and typology of companies based on GRI standards (2016)

Figure 1 shows the position of NFRs published by 100 selected companies by year based on the GRI standards. Specifically, NFRs were categorized by comparing the words used in each company's NFR and their frequency of occurrence against the 3,534 words used in the GRI standards and their frequency of occurrence (a total of 65,168 occurrences of 3,534 words). In other words, it is a typology based on the degree of mention of the GRI standards.

The principal components (PC) should represent the object of analysis and have a certain directionality. However, this study targeted words appearing in the report and did not arbitrarily select words as principal components. Therefore, analysis was not performed in the direction of the principal components. However, using the words appearing in the GRI standards as an axis, we can observe the distance between each company's reports.

Looking at the NFRs of the 2018 Global 100 companies, 62 reports (cluster 3) were vertically aligned with the GRI standards. This group is close to the GRI standards. It can be assumed that many reports contain

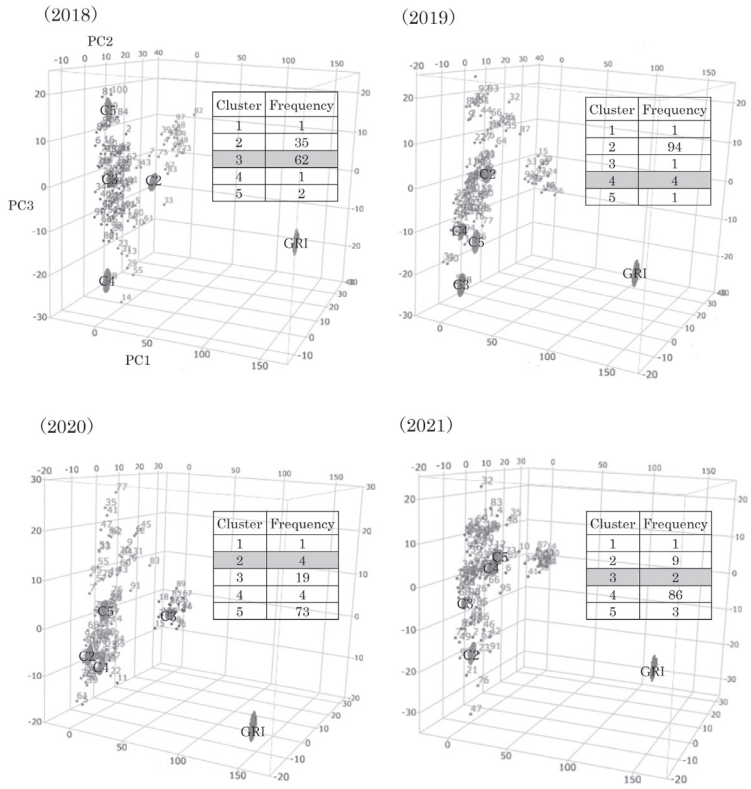

Figure 1 Distribution based on GRI standards (2018-2021)
*The area shaded grey is the cluster closest to the axis.

content that conforms to the GRI standards. However, from 2019 to 2021, there were fewer reports in the cluster in proximity to the GRI standards, even though there has not been much change in the types of reports (CSR, sustainability, integrated, annual, etc.) compared with 2018. Many of the NFRs of Global 100 companies continue to trend away from the GRI standards.

Figure 2 shows the PC 2 and PC 3 planes in 2021. The 2021 NFRs of cluster groups and companies closest to the GRI standards are sustainability reports, closely following the GRI standards. Reports that fall furthest from the GRI standards are the groups that commonly publish annual reports and integrated reports, which contain more financial information (see areas with strong financial information in Figure 2). The countries and industries of the companies in these reports vary.

A closer analysis of this distribution shows that five Japanese companies ranking in the Global 100 belong to Cluster 4, the largest group. Furthermore, four of these companies are located remarkably close to each other. The same trend can be observed in other years, and the regional characteristics of information disclosure are of interest.

The change in the evaluation method of the Global 100 may have had some impact. As a result, the trend of disclosing as much as possible on items required by the GRI standards in the past (e.g., attaching the GRI standard control sheet) may have led to streamlining NFRs. Therefore, much of the information may be shifted to disclosure in a more flexible, decentralized, or timely manner.

In each year, there was no correlation between the group composition and the Global 100 ranking results.

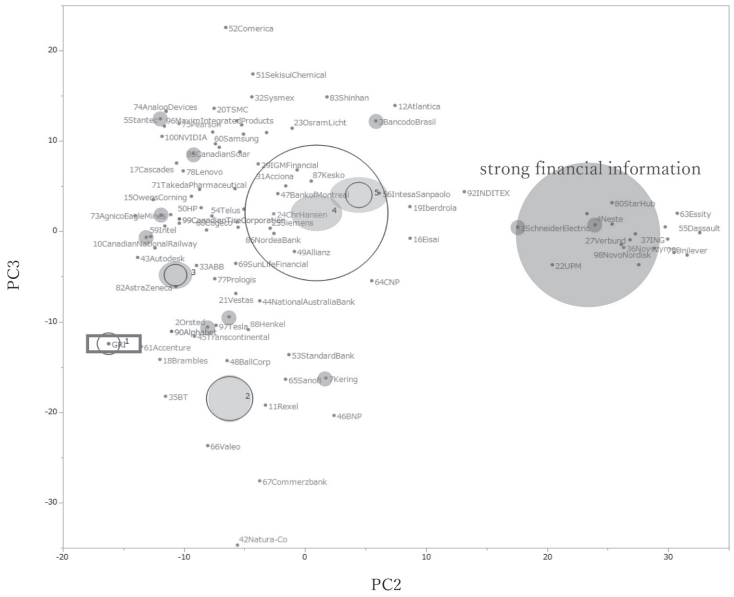

Figure 2 Distribution of NFRs (2021) in view of PC 2 and 3.
＊The small circles indicate the top 10 companies in the Global 100

In addition, the formation of groups of companies in the same industry was not confirmed, indicating that there is no bias in the disclosure of information specific to the industry.

4-2. Distribution and typology of companies based on IIRC framework (2013)

Figure 3 shows the position of NFRs published by the 100 selected companies based on the IIRC framework. The number of words used in the IIRC framework were 1,706, and the occurrences were 9,533. For this, the degree of adaptation of the words used and their frequency of occurrence in each NFR was calculated, and the NFRs were typified.

Figure 3 shows that 64 reports (cluster 3) in 2018 and 51 (cluster 2) in 2019 were close to the IIRC framework. Many reports in 2018 and 2019 made extensive use of words used in the IIRC framework. However, in 2020 and 2021, the number of reports in clusters close to the IIRC framework is small. In other words, most of the NFRs of Global 100 companies tend to be far from the IIRC framework, and the analysis results are based on the GRI standards.

By 2021, more reports will be grouped closer to the IIRC framework than to the GRI case (2021). Of these, the group closest to the IIRC framework was Cluster 5. However, of the 12 reports in Cluster 5, only one report was published under the title Integrated Report, and not all reports titled Integrated Report are close to the IIRC framework. This is due to the diverse approaches to integrated reporting. For example, some reports comply with the IIRC framework, and others are company-specific. Furthermore, there are many titles for NFRs, which do not always match the report's content. In other words, some integrated reports may deviate from the IIRC framework, and some sustainability reports may be strongly compliant. However, many of the integrated reports of the 2021 Global 100 companies are gradually moving toward their own disclosure practices in terms of report structure, content, and methods.

The IIRC case also confirms the similarity between Japanese companies (see Figure 4). Four Japanese companies belong to Cluster 3, which is the largest group. They are also located remarkably close, except for

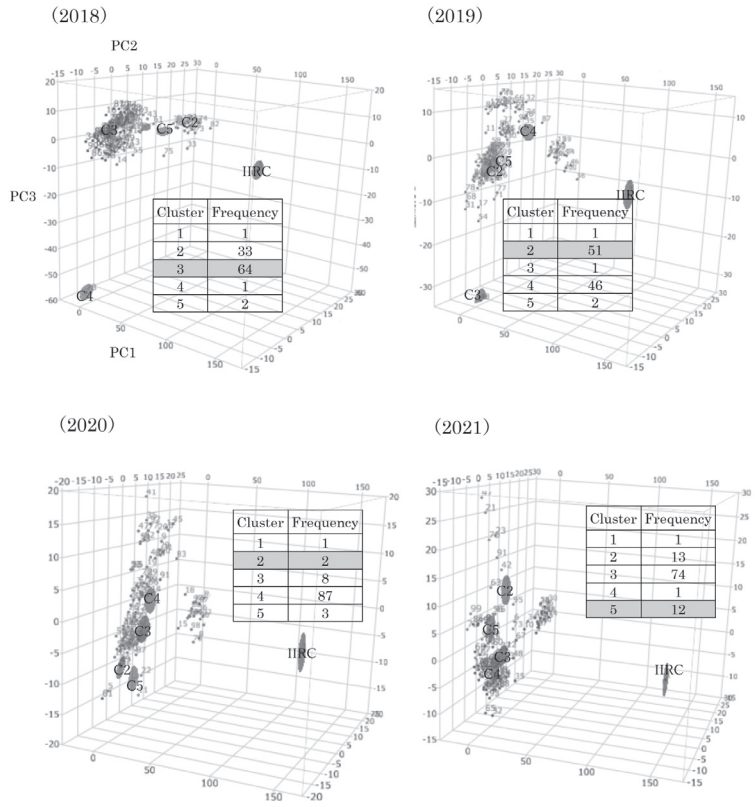

Figure 3 Distribution based on IIRC framework (2018-2021)
*The area shaded grey is the cluster closest to the axis.

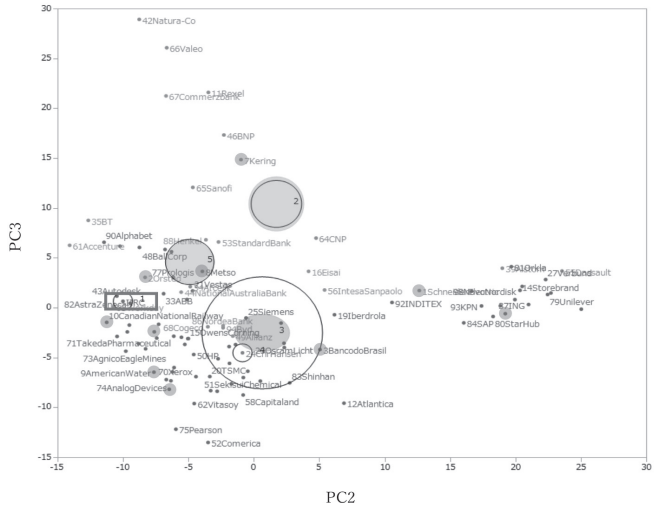

Figure 4 Distribution of NFRs (2021) in view of PC 2 and 3.
* The small circles indicate the top 10 companies in the Global 100

one company that publishes an integrated report ranked in the Global 100. The same trend was observed in other years. This suggests the possibility of resemblance through mutual influence within a domain, as in the case of GRI. However, the case of the IIRC framework also shows no correlation between cluster composition and the Global 100 rankings or industries in each year.

5. Conclusion

This study identified the characteristics of NFRs published by companies ranked in the Global 100 from 2018 to 2021 through content analysis using Natural language processing technology. The analysis was based on internationally influential GRI standards and the IIRC framework to enhance the features of the report.

Although the characteristics of GRI and IIRC cases vary from year to year, the recent trend is to move away from the GRI standards and the IIRC framework and practice their own information disclosure; that is, the words used in the GRI standards and the IIRC framework are the basic elements, but much information is incorporated into the NFR content. Alternatively, it indicates that the information within NFRs may be reduced in size and disseminated in various ways. Therefore, information disclosure is expected to progress from NFRs to the use of all other means. However, as the means of information disclosure becomes more diverse, we need to consider ways to ensure the reliability of the information.

In addition, the finding that national and regional characteristics exist in the context of the increasing globalization of information disclosure is a new insight. In this study, we focused only on Japanese companies. However, future studies could explore similarities in other countries and regions.

However, if it is assumed that non-financial pressures will become stronger and information disclosure will become more uniform internationally, then NFRs that demonstrate the individuality of companies will be sought and developed. Therefore, to achieve this, it is necessary to understand the position of the company among others. The Global 100 company platform presented in this study can be used, for example, to include a sample of a company's NFRs to understand their position within the overall group, set a target group, and determine what information to focus on in order to reach the target.

As mentioned above, the study of classification methods remains an issue for improving report classification accuracy. The selection of indicator words should also be considered to clarify the direction of principal components.

Reference

Bala, G., Bartel, H., Hawley, J. P., Lee, Y. J. (2015) "Tracking "real-time" corporate sustainability signals using cognitive computing" *Journal of Applied Corporate Finance*, Vol.27, pp95-102.

Barkemeyer, R., Figge, F., Holt, D., Hahn, T. (2009) What the Papers Say: Trends in Sustainability. A Comparative Analysis of 115 Leading National Newspapers Worldwide, *Journal of Corporate Citizenship*, pp69-86.

Elkington, J (1997), *Cannibals wit forks*, new society publishers.

Fiandrino, S., Tonelli, A. (2021) "A Text-Mining Analysis on the Review of the Non-Financial Reporting Directive: Bringing Value Creation for Stakeholders into Accounting" *Sustainability*, vol. 13(2), pp1-18.

Global Reporting Initiative (2016), *GRI standers*.

Goloshchapova, I., Pooon, S-H., Pritchard, M., Reed, P. (2019) "Corporate Social Responsibility Reports: Topic

Analysis and Big Data Approach" *European Journal of Finance*, Volume 25, Issue 17, pp 1637-1654.

Harymawan, I. et al (2020) "Text mining on sustainability reporting: A case study", *Journal of Security and Sustainability Issues*, Vol.9, pp48-55.

International Integrated Reporting Committee (2013), *the international <IR> Framework*, IIRC Paper. Corporate Knights, Global 100 Ranking (http://www.corporateknights.com)

Liew, W., Adhitya, A., Srinivasan, R. (2014) "Sustainability trends in the process industries: A text mining-based analysis" *Computers in Industry*, 65, pp393-400.

Modapothala, J. R., Issac, B., Jayamani, E. (2010) "Appraising the Corporate Sustainability Reports - Text Mining and Multi-Discriminatory Analysis" *Innovations in Computing Sciences and Software Engineering*, pp489-494.

Rivera, S. J., Minsker, B. S., Work, D. B., & Roth, D. (2014) "A text mining framework for advancing sustainability indicators" *Environmental Modelling & Software*, Volume 62, pp 128-138.

Shahi, A. M., Issac, B., Modapothala, J. R. (2014) "Automatic Analysis of Corporate Sustainability Reports and Intelligent Scoring" *International Journal of Computational Intelligence and Applications*, Vol.13, No.1, pp1-27.

Teuteberg, F. (2013) "Corporate social responsibility reporting - A transnational analysis of online corporate social responsibility reports by market-listed companies: Contents and their evolution" *International Journal of Innovation and Sustainable Development*, Vol.7 No.1, pp1-26.

Türegün, N. (2019)"Text Mining in Financial Information" *Current Analysis on Economics & Finance*, pp18-26.