Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Extracting and Coding Digital Sustainability Data using Text Mining and AI: A Design Science Approach

Thomas Abraham

tabraham@kean.edu

Kean University

Viet Dao

Shippensburg University

Nesreen El-Rayes

Georgia State University

Research Article

Keywords:

Posted Date: August 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10538848/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Extracting and Coding Digital Sustainability Data using Text Mining and AI: A Design Science Approach

Abstract

This study develops artifacts to help researchers extract digital sustainability (DS) data effectively and efficiently from large textual resources such as sustainability reports, and code the data using established Information Systems literature frameworks addressing the challenges associated with manual data collection, processing, and coding. We adopt the design science research methodology to develop the artifacts. We identify the research problem and motivation and define the objective for the solution the artifact attempts to address. Then we describe the iterative process through which we design, develop, and demonstrate artifacts. Finally, we evaluate the artifacts by comparing data collected and coded via the newly designed semi-automatic process with data collected and coded manually by expert researchers. The results reveal that digital sustainability data can be extracted in greater quantity and in less time using text analytics supported by our DS dictionary. The results also reveal that our generative AI prompts coded DS data with a high degree of agreement with expert manual coded data. The study offers researchers a scalable process for building large digital sustainability datasets from textual sources. The artifacts help reduce manual effort in data extraction and coding while preserving the need for expert review and quality assurance. The study contributes to a novel design science approach that combines text mining and generative AI to support digital sustainability research. It provides validated process, dictionary, and prompt artifacts that enable scalable, theory-driven extraction and coding of digital sustainability initiatives from textual data sources.

1. Introduction

Information Systems (IS) scholars studying Digital Sustainability (DS) (Kotlarsky et al., 2023) focus on how IS helps companies holistically address both environmental and social sustainability internally and externally. Given the lack of centralized data sources for DS, researchers often must rely on extracting data from secondary sustainability data sources , such as sustainability reports, annual reports, etc. Without the support of an automated process, it is a time-consuming and labor-intensive process as researchers must first rely on manual screening of such reports to find and extract DS content, followed by a laborious process of manual content coding based on different coding schemes (Abraham & Dao, 2019; Carberry et al., 2019; Dao & Abraham, 2021; Zhang et al., 2025). The evolution of text mining, text analytics, and Artificial Intelligence (AI) provides researchers with opportunities to automate parts of the data collection and coding process. This should facilitate examining much larger datasets from secondary data sources (Zhang et al., 2025).

In this study, we adopt the design science research method, in which researchers start by addressing a specific problem through the development of an artifact and then extracting general knowledge from the design process (Livari, 2015). In particular, we develop a semi-automatic process for DS data collection and content coding using the WordStat text analytics tool (Provalis Research, 2025) and ChatGPT (OpenAI, 2023). The proposed process consists of two main phases. The first phase focuses on semi-automatic DS data collection by developing a DS dictionary (Deng et al., 2019) using WordStat as the supporting analytics tool. The second phase focuses on semi-automatic content coding of the collected DS content by developing and fine-tuning a coding scheme to guide ChatGPT's automated coding.

Our proposed approach makes four main contributions. First, it combines text analytics and generative AI tools to create a process artifact for DS data collection and coding. Second, it creates a DS data dictionary artifact that can be used to extract relevant DS content from secondary data sources. Third, it creates generative AI prompt artifacts based on established frameworks for IT Roles (Zuboff, 1988; Schein, 1992; Aral and Weill, 2007) and ISF Quadrants (Hart

& Millstein, 2003; Dao et al., 2011), to guide AI-supported content coding in DS research. Fourth, we extract general guidelines and principles from our design process.

Following design science research guidelines (vom Brocke et al., 2020), our process was carefully designed, developed iteratively, and evaluated by comparing the semi-automatically collected and coded data with previously manually collected and coded data done by expert researchers. By doing so, we aim to address a key research question:

(RQ) *How can text analytics and AI tools be effectively used to aid researchers with semi-automatic DS data collection and content coding?*

The rest of the paper is structured as follows. The next section provides a review of the literature. Section 3 discusses how we applied the design science research method. Section 4 identifies the problem and the objective of the study. Section 5 describes in detail the design, development, and demonstration of the artifacts, and Section 6 evaluates the artifacts by comparing the semi-automatically collected and coded data against data that were previously manually collected and coded by experts. The manuscript ends with our discussion of research contributions and the conclusion.

2. Research Background

2.1 Digital Sustainability Research and Empirical Data Collection

Although DS research has made significant progress over the last decade, the IS field has not kept pace with industry developments, resulting in a limited body of research on DS and its interaction with sustainability goals (Kotlarsky et al., 2023). Tackling this gap is crucial as sustainability has long been recognized as a key concern in IS research and IS has a strong potential to contribute to the achievement of sustainable development goals. Several papers call for emphasizing responsible digital transformation and positioning DS as a broader socio-technical agenda (Watson et al., 2010; Pappas et al., 2023; Schoormann et al., 2025). AI is increasingly viewed as a key enabler of DS. Pan & Nishant (2023) identified six dimensions through which AI supports DS: Supply Chain relationships, green creativity skills, metrics, strategies, tools improvement, and sensemaking. Florek-Paszkowska and Ujwary-Gil (2025) propose a DS ecosystem framework in which Industry 4.0 technologies drive sustainability transitions through real-time monitoring, predictive capabilities, and continuous improvement. Similarly, Corporate Environmental Intelligence links digital infrastructure, such as AI, IoT, and BCT with sustainability systems (Bressane et al., 2026; Balcioglu et al., 2024). Despite these conceptual advances, the challenge of collecting large datasets for DS initiatives remains significant because centralized data sources in IS remain scarce. As a result, scholars often need to rely on primary data collection or extract evidence from secondary sources such as web content, sustainability reports, and related organizational disclosures. This process requires extensive manual work (Zou et al., 2025). In addition, once the data have been collected, content coding and categorization in a systematic, theory-driven manner is also challenging and labor-intensive (Zhou et al., 2024). Consequently, future IS research needs to examine how technologies, including LLMs and text mining, can be responsibly deployed to advance sustainability goals (Brown, 2025), including supporting data extraction and coding.

2.2 Analytical Foundations for Extracting DS Data

Text analytics has been used for a variety of research purposes, including text classification, text summarization, and sentiment analysis (Ittoo et al., 2016). In the DS context, IS scholars have also suggested using text analytics to support environmental sustainability research, particularly through the development of dictionaries and using those dictionaries for subsequent text data collection and analysis (Deng et al., 2019). Dictionary-based text analytics offers a structured way to identify relevant sustainability content across large textual corpora.

Dictionary building can be accomplished through one of three approaches: manual, semi-automatic, or automatic approaches. Among these, the semi-automatic approach is the most widely used (Deng et al., 2019). The quality of a dictionary impacts the results of subsequent dictionary-based text analytics (Laver & Garry, 2000). However, developing and validating high-quality dictionaries is a time-consuming process, thus many researchers lean toward using predefined dictionaries (Boritz et al., 2013). For example, Guo et al. (2026) used an existing dictionary of AI terms (Mishra et al., 2022) to extract AI-related themes from the text of Management Science journal articles. Building such a dictionary of DS terms would support extracting data from sustainability and other reports.

2.3 Generative AI for Data Extraction and Content Coding

Recent research has explored the use of generative AI tools such as ChatGPT (OpenAI, 2025) to support automated data extraction and coding. Bu et al. (2025, p. 693) stated that AI is "a present-day reality driving a profound paradigm shift. Morgan (2023) used ChatGPT to reinvestigate datasets that were originally analyzed using traditional methods and found that ChatGPT performed reasonably well. Within the sustainability domain, AI tools have been primarily used for ESG ratings construction, disclosure classification, and sentiment evaluation (Alotaibi and Alwathnani, 2025).

Hamilton et al. (2023) compared AI and human coders in a study and found that humans identified certain themes that AI did not, while AI also identified patterns that humans overlooked. These differences have implications on users' cognitive trust in AI (Zhou & Lu, 2025) and vice versa. Zou et al. (2025) developed an LLM module for automated Environmental and Social Governance (ESG) data extraction, and benchmarked several LLMs, including ChatGPT 3.5, ChatGPT 4, Chat GLM, and QWEN, and found that ChatGPT 4 achieved the strongest data extraction performance reaching 77%. Generative AI has also been used in more structured qualitative analysis settings. Zhou et al. (2024) created a guided process using ChatGPT to analyze text data using a grounded theory methodology. Their findings suggest that generative AI can produce results comparable to those of human analysts, but only when the process is supported by proper guidance and rigorous evaluation. Specifically, they used prompt engineering patterns consisting of five elements: role setting, task instruction, context description, input data, and output indicators. We adopt their process-oriented approach and prompt engineering patterns.

Although generative AI offers clear potential for supporting analytical tasks, its use may introduce challenges. Chong et al. (2025) found that technostress associated with GenAI use reduces flow experience and continuance intention while increasing switching intention. These findings imply that AI-supported research tasks such as content coding should not be treated as fully automated. Instead, they should be designed as guided and cognitively manageable processes with clear researcher oversight.

Theory-driven content coding is important. We adopt two established frameworks to guide the coding process – IT Roles (Zuboff, 1988; Schein, 1992; Aral and Weill, 2007) and ISF Quadrants (Hart & Millstein, 2003; Dao et al., 2011). We describe these frameworks in much greater detail in section 4.

3. Design Science Research Method for DS Data Extraction and Coding

Design science research literature positions IS research as increasingly relevant to societal challenges, including sustainability transformation and socially significant problem solving. In IS, design science research frequently addresses problems related to the design of IS and associated artifacts. Such artifacts may take the form of system instantiations that automate or improve a process (Nunamaker et al., 1991; Hevner et al., 2004; Greger & Hevner, 2011). In this study, the primary design artifact is the semi-automated data collection and coding process. Supporting this central artifact are two additional artifacts: the data dictionary used for DS data extraction and the AI prompts used to guide content coding. These artifacts are intended to provide a more scalable, systematic, and theory-driven approach to collect and code DS initiatives from textual sources.

Livari et al. (2015) distinguish and contrast two design science research strategies in information systems. Our approach follows the second strategy where "a researcher attempts to solve a client's specific problem by building a concrete IT artifact in that specific context and distills from that experience prescriptive knowledge to be packaged into a general solution concept to address a class of problem." Markus et al. (2002) and Stein et al. (2011) are exemplars of this strategy. In our case, the client is the DS research community, and the class of problem is extracting and coding DS data from textual reports. Strobel et al. (2023) used AI and Peffers et al. (2007) design science research methods to address the lack of inclusive solutions for the deaf and to create artifacts that instantiate their solution. They then discuss the lessons learned from instantiating their artifacts. To the best of our knowledge, no studies have applied text mining and generative AI to create a process artifact for extracting and coding DS data.

We adopt the design science research methodology to develop our artifacts (Hevner et al., 2004; Peffers et al., 2007; Vom Brocke et al., 2020). The process is illustrated in Fig 1. First, we identify the research problem and define the objective for the solution the artifacts attempt to address. Then we describe the iterative process through which we

design, develop, and demonstrate artifacts. Finally, we evaluate the artifacts by comparing data collected and coded via the newly designed semi-automatic process with data collected and coded manually by expert researchers.

Using this design science research method, we develop a semi-automatic process artifact for DS data extraction and coding. In addition, we create two artifacts, one for the data extraction task and one for the data coding task. We divide the creation of the artifacts into two phases: Phase 1 develops a process to use text mining to extract DS data from secondary data sources using a data dictionary artifact; and Phase 2 develops a process to apply generative AI and prompt artifacts to code the extracted data using a literature-based coding scheme. Both phases require the process to go through the stages in Fig 1. We follow a semi-automatic dictionary-building process (Brier & Hopp, 2011; Deng et al., 2019) to develop our DS data dictionary.

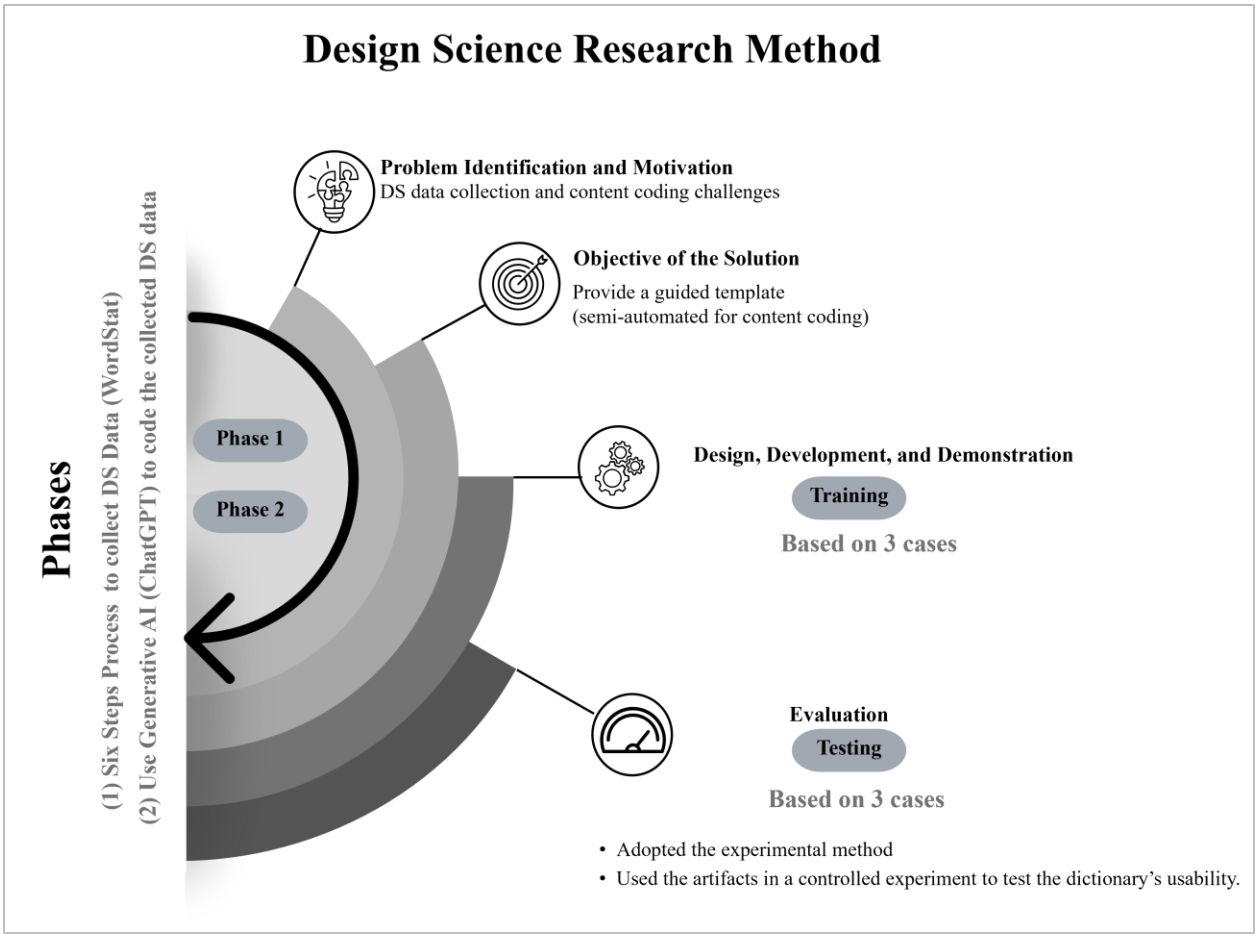


Fig. 1 Design Science Research Method

* DS: Digital Sustainability

4. Problem Identification and Objective of the Solution

4.1 Problem Identification

Through our review, we find limited empirical research on companies' use of IT/IS resources to address sustainability holistically. This could be explained by the challenge of collecting data on DS initiatives, given the lack of available centralized data sources on DS. To collect such data, scholars need to rely on primary data sources (which inherently are hard to collect) or extract such data from secondary data sources such as web content, sustainability reports, etc. Doing so usually requires extensive manual work.

Additionally, once the data have been collected, content coding and categorizing them in a systematic, theory-driven manner is also a challenging and labor-intensive process. This usually requires multiple experts/trained coders to review such content, and manually code them and compare their coding for inter-rater reliability. Together, the data collection and coding processes pose a significant challenge for researchers, especially when collecting a large dataset.

4.2 Objective of the Solution

In IS research, a design science research project aims to help enable the development and implementation of information systems or technology-based solutions - an artifact - to address unsolved business problems (Hevner et al. 2004). While such solutions/systems usually aim at addressing problems for IS practitioners, our research aims at developing an artifact that helps IS researchers solve the data collection and coding problem. It is also based on the utilization of technology, namely a text analytics tool (WordStat) and a generative AI tool (ChatGPT).

The main objective of the artifacts built in this research – a semi-automated data collection and coding process, a data dictionary, and coding prompts - aims to provide a guided template that helps researchers to more easily collect and content code a larger set of data on DS. The purpose of the data dictionary is to help identify DS within sustainability reports regardless of the industry being researched. The purpose of the prompts is to automatically code those DS initiatives into certain theory-driven categories, based on theoretically developed schemas. These content-coded DS initiatives can help researchers conduct further empirical analyses that will help shed more light on our collective understanding of the role of IS in enabling companies to be more sustainable.

Two key questions of design science research are "What utility does the new artifact provide?" and "What demonstrates that utility?" (Hevner et al. 2004, p.91). The artifacts of our research help provide a utility to researchers to simplify and improve two tasks: 1) data extraction with text mining and a data dictionary and 2) automatic data content-coding using generative AI prompts. We demonstrate this utility by comparing the results obtained by the artifacts with the results of a previous data extraction and coding exercise done manually.

5. Design, Development, and Demonstration

5.1 Phase 1: Data collection and dictionary-building

Once the objective of the dictionary is established, the next steps are corpus creation and pre-processing following an inductive consensus-building approach. The corpus consists of the documents used as the source material to help identify terms used in the dictionary. Instead of building a corpus from scratch, we used data from an earlier study (Abraham et al., 2026; El-Rayes et al., 2022), which extracted the abstracts, titles, and keywords from a huge collection of research articles on DS from the Scopus database. They used a text mining tool called VOS Viewer to extract over 20,000 terms from this extensive corpus and used a pre-processing technique called cutoff criteria to select 604 terms that met the criterion of at least ten occurrences. Our study continued the pre-processing step by extracting roughly 104 terms from that list based on their relationship to information systems or information technology (Fig 2).

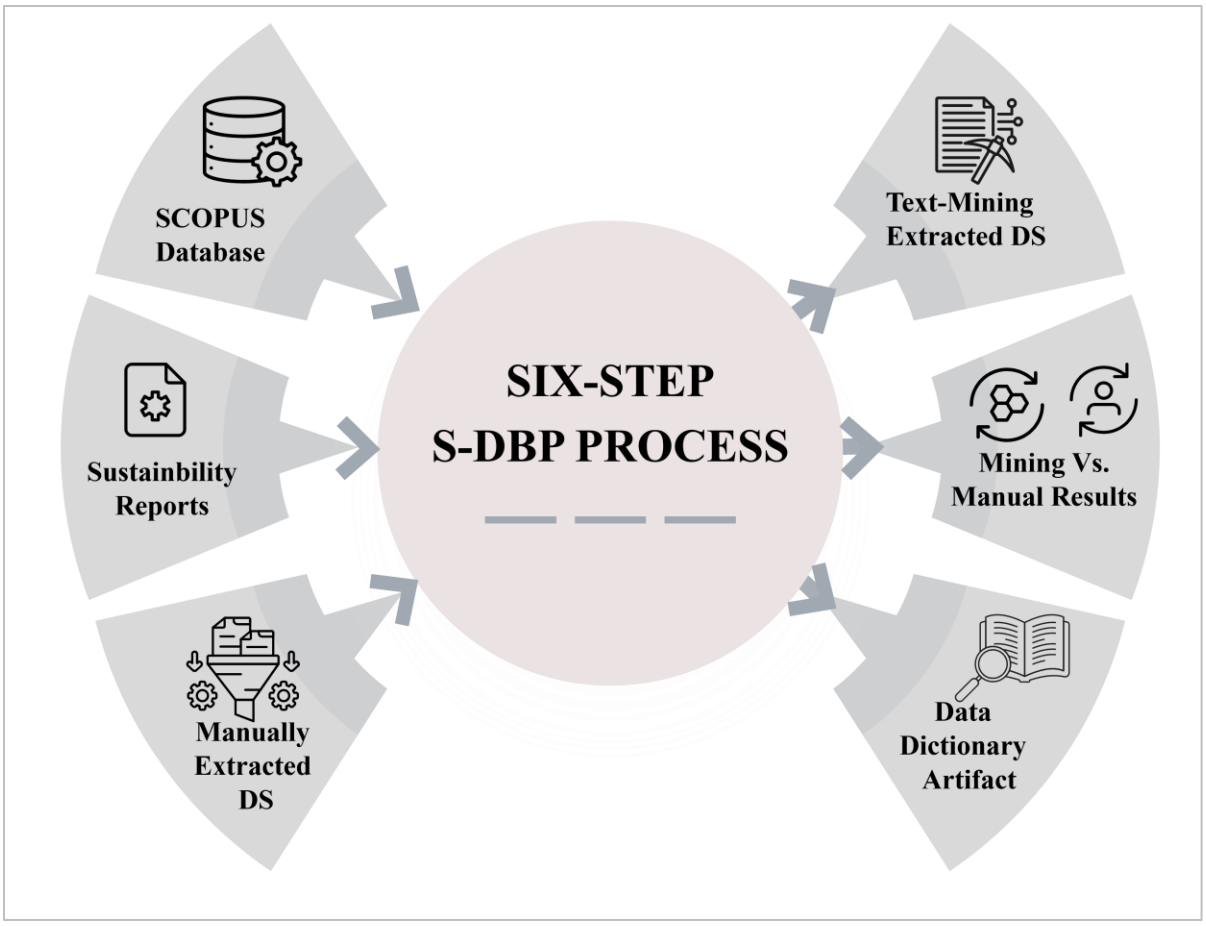


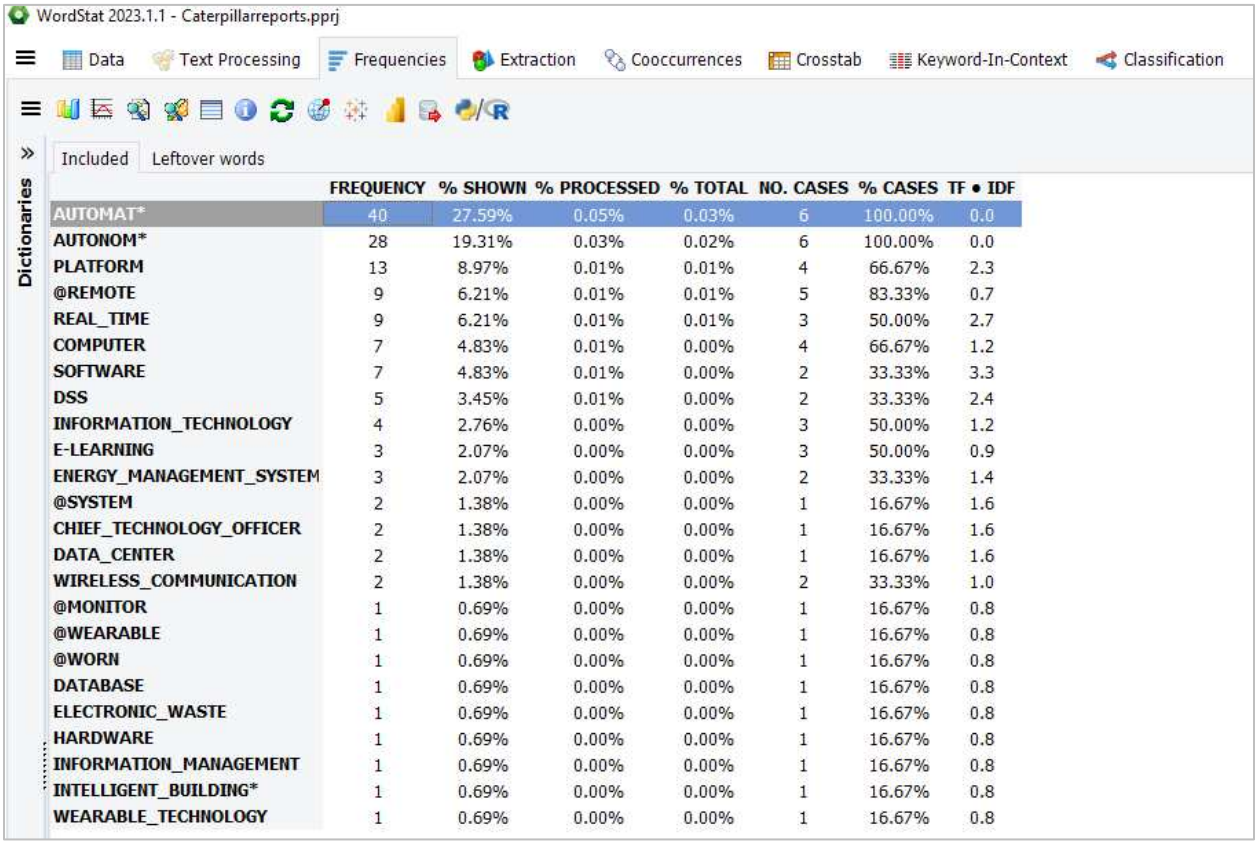
Fig. 2 Iterative dictionary-building and DS data collection

Following the pre-processing step is entry identification and categorization. We categorized dictionary entries as either Green Information Technology (Category 1) or Green Information Systems (Category 2). From our earlier review of research included in the corpus, IS sustainability research has mostly focused on environmental sustainability. Therefore, we found that the terms extracted might not sufficiently reflect social sustainability issues. To expand our scope to include social sustainability, we searched the Scopus database for additional terms using the AMCIS 2024 tracks as a source. We searched for articles that included terms such as digital equity, digital social entrepreneurship, social computing, health informatics, and so on, and categorized the terms as Social Information Systems (Category 3).

The fifth step of the dictionary-building process is extension and simplification. We added the early version of the data dictionary from the previous step (fourth) to a text-mining tool called WordStat. Using an earlier study in which we used a manual data collection process for six companies in three industries (pharmaceutical, food and beverages, and heavy equipment manufacturing) over 6 years from 2009 to 2015. We split the companies into training and testing data sets of 3 companies each. The training data set is used to develop our data extraction and coding processes, while the testing data set is used to evaluate the effectiveness of the developed processes. The training data set includes data from three companies: Biogen, Coca-Cola, and Caterpillar; while the testing data includes data from three companies: Johnson & Johnson, General Mills, and Volvo Construction.

Using the training set, we mined the companies' sustainability reports, searching for discussion of digital sustainability through several rounds. For each company, we created a WordStat project with the company's reports. We analyzed the documents by applying the DS data dictionary to them. Fig 3 shows an example word frequency table created by the Frequencies tool in WordStat for Caterpillar's data. Afterward, we used the Keyword Retrieval feature to extract and save the original paragraphs containing the terms. Each report was saved in a separate case in WordStat, corresponding to the year the report was published. This analysis took only a few minutes for each company's reports,

consisting of several hundred pages each. This process proved to be a significant improvement in terms of time and labor compared to the manual data collection process.



	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
AUTOMAT*	40	27.59%	0.05%	0.03%	6	100.00%	0.0
AUTONOM*	28	19.31%	0.03%	0.02%	6	100.00%	0.0
PLATFORM	13	8.97%	0.01%	0.01%	4	66.67%	2.3
@REMOTE	9	6.21%	0.01%	0.01%	5	83.33%	0.7
REAL_TIME	9	6.21%	0.01%	0.01%	3	50.00%	2.7
COMPUTER	7	4.83%	0.01%	0.00%	4	66.67%	1.2
SOFTWARE	7	4.83%	0.01%	0.00%	2	33.33%	3.3
DSS	5	3.45%	0.01%	0.00%	2	33.33%	2.4
INFORMATION_TECHNOLOGY	4	2.76%	0.00%	0.00%	3	50.00%	1.2
E-LEARNING	3	2.07%	0.00%	0.00%	3	50.00%	0.9
ENERGY_MANAGEMENT_SYSTEM	3	2.07%	0.00%	0.00%	2	33.33%	1.4
@SYSTEM	2	1.38%	0.00%	0.00%	1	16.67%	1.6
CHIEF_TECHNOLOGY_OFFICER	2	1.38%	0.00%	0.00%	1	16.67%	1.6
DATA_CENTER	2	1.38%	0.00%	0.00%	1	16.67%	1.6
WIRELESS_COMMUNICATION	2	1.38%	0.00%	0.00%	2	33.33%	1.0
@MONITOR	1	0.69%	0.00%	0.00%	1	16.67%	0.8
@WEARABLE	1	0.69%	0.00%	0.00%	1	16.67%	0.8
@WORN	1	0.69%	0.00%	0.00%	1	16.67%	0.8
DATABASE	1	0.69%	0.00%	0.00%	1	16.67%	0.8
ELECTRONIC_WASTE	1	0.69%	0.00%	0.00%	1	16.67%	0.8
HARDWARE	1	0.69%	0.00%	0.00%	1	16.67%	0.8
INFORMATION_MANAGEMENT	1	0.69%	0.00%	0.00%	1	16.67%	0.8
INTELLIGENT_BUILDING*	1	0.69%	0.00%	0.00%	1	16.67%	0.8
WEARABLE_TECHNOLOGY	1	0.69%	0.00%	0.00%	1	16.67%	0.8

Fig. 3 Dictionary terms found in 5 Caterpillar Sustainability Reports.

During each round of text mining using the training dataset, we first had to manually evaluate each paragraph returned by WordStat and select unique DS initiatives. For instance, WordStat returned 280 paragraphs, many of which included no clear DS initiatives and many duplicated initiatives. From that list, we identified 55 actual DS initiatives. This is why we opted for a semi-automated process. After selecting the DS initiatives, we compared the WordStat results with the prior manual search results. A researcher compared the DS paragraphs extracted by WordStat with those previously extracted manually and highlighted any that did not match. We then extended the dictionary's terms to improve its ability to capture DS initiatives that were missed in earlier iterations. Our extended and simplified dictionary automatically retrieved a dataset which included most paragraphs discussing DS initiatives that were manually retrieved. Table 1 provides a small sample of paragraphs extracted by WordStat versus a manual search from Biogen's sustainability reports

Table 1 Sample Biogen WordStat and Manual Search Results

WordStat extracted DS sample	Also extracted manually?	Paragraph from Biogen's reports
EMS	Yes	For years, we've developed innovative solutions to lessen our impact on the planet while operating our business efficiently. In 2009, we undertook a comprehensive review of our environmental practices and developed a long-term strategy to continue to improve our performance. We also began implementing an Environmental Management Information System to track our performance over time.

HVAC Optimization	No	• HVAC Optimization: Utilizing variable air volume and exhaust fan software to modulate static pressure and control air flow in campus buildings during unoccupied nights and weekends.
Green Data Center	Yes	Green Data Center - Designed with the installation of a high-density rack and self-contained cooling system, our green data center maximizes performance while reducing energy consumption by 35 to 45 percent.
Recycling Database	No	We created a database that helps employees determine which recycling or waste container to use for different waste streams and where they are located. The database also provides specific steps for the responsible disposal or treatment of each waste stream.
Health and Safety Management System	Yes	Biogen Idec is committed to maintaining a healthy and safe work environment. We have adopted a formal health and safety management system, which we update regularly to keep pace with the changing requirements of our business. We also require all employees to attend safety classes specifically tailored to their lines of work.
R&D IT Strategy	Yes	Another key development in 2013 was accelerating our R&D information technology (IT) strategy, with the aim of better integrating IT into the way we develop solutions. The outcome of this analysis will drive significant IT improvements in 2014 and beyond.

Since sustainability reports for each company use their own naming conventions for DS, we needed to create flexibility in the dictionary to capture this variation. We accomplished this by using a WordStat feature called proximity rules. An example of this is @MONITOR [MONITOR* BEFORE SYSTEM* /A /S4]. This rule captures any DS initiative that includes phrases such as "monitoring and tracking systems" as long as the variations of Monitor* precedes System* by up to 4 words, and they are in the same sentence. In addition, WordStat suggested extensions such as Portal and Server, which we included in the dictionary. We also added a new category (Category 4: IS Organization) to capture organizational titles such as CIO.

Table 2 shows the number of digital sustainability initiatives (DSI) semi-automatically extracted compared to the number of DSI manually extracted, in aggregate and broken down by company. We treat the first three companies' reports as training data. For each company, the "Overlap DSI" data column indicates the number of DSI that were included in both manually extracted and WordStat extracted datasets. "Total DSI" includes the number of distinct DSI extracted by both methods.

WordStat extracted many more DSI than the manual process. However, we see that there are still a few DSI found manually that were not picked up by WordStat: 3 for Biogen, 3 for Coca-Cola, and 1 for Caterpillar. Some of the terms we initially included, such as websites, generated a great number of results without identifying useful systems. We dropped website from our dictionary, which resulted in some missed systems. Nevertheless, in our training sample of three companies, WordStat successfully identified 54 out of 61 manually extracted DSI (89%). Further, of the total 114 DSI extracted, WordStat found 107 of them (94%).

Table 2 Comparing Manually Extracted vs. WordStat Extracted DSI

Data	Company	Manually extracted DSI (1)	Overlap DSI (2)	WordStat extracted DSI (3)	Total Distinct DSI (1 & 3)	% Manual	% Overlap	% WordStat
Training	Biogen	19	16	32	35	54%	46%	91%
	Coca-Cola	17	14	28	31	55%	45%	90%
	Caterpillar	25	24	47	48	52%	50%	98%
	Total	**61**	**54**	**107**	**114**	**54%**	**47%**	**94%**

Clearly, this iterative process will not result in a final, optimal data dictionary. For one thing, the terms that define information systems evolve constantly, and new terms will have to be continuously added. Our goal for the dictionary was to achieve some satisficing level of coverage. Table 3 lists our dictionary items and categories.

Table 3 DS Dictionary artifact

Category 1: Green Information Technology	Category 2: Green Information System	Category 3: Social Information Systems
Application Program*		Bioinformatics
Artificial Intelligence	Carbon Accounting System	Data privacy
Automat*	Communication System	Digital Divide
Autonom*	Data [Mining*/A/S]	Digital Equity
Big Data	Decision Support System	Digital Exclusion
Blockchain	DSS	Digital Government
Cloud Comput*	EHS+S Management System	Digital Health
Computer	Electronic Commerce	Digital Inclusion
Database	EMS	Digital Innovation
Data Center	Energy Informatics	Digital Literacy
Digital Storage	Energy Management System	Digital Rights
Digital Transformation	Enterprise Resource Management	Digital Skills
Discover [Data -Discover*/A/S]	Environmental Management System	E-Democracy
Distributed System	Geographic Information System	E-Government
Electronic Waste	GIS	E-Learning
Embedded System	Green IS	E-Participation
Emerging Technolog*	Green Manufacturing	E-Services
E-Waste	Human Computer Interaction	Education Computing
Green Computing	Information Management	E-Health
Green IT	Information Service	Electronic Health Record
Green Technology	Information System	Electronic Medical Record*
Grid Computing	Intelligent Building*	Health Information System
Hardware	Intelligent [System*/A/S3]	Hospital Information System
ICT	IT Governance	M-Government
Information and Communication Technologies	IT [System*/A/S3]	Medical Computing
Information Technology	Knowledge Management	Medical Informatics
Internet of Things	Mobile Telecommunication System	Medical Information Systems
IoT	Monitor [System*/A/S4]	Process Safety Management System
Learning System	Supply Chain Management System	Safety Management System
Machine Learning		Smart City
Mobile Computing		Smart Wearables
Network Architecture		Social Computing
Network Security	**Category 4: IS Organization**	Sustainab* Website
Platform		Telemedicine
Portal	Chief Information Security Officer	M-Health
Real Time	Chief Technology Officer	Wearable Computer
Remote [Monitor*/A/S]	CIO	Wearable [Wear* Device/A/S]
Security of Data	Chief Information Officer	Wearable Sensor
Server	CISO	Wearable Technology
Smart Device	CTO	Wearable System
Smart Grid	Data Scientist	Worn [Device/A/S]
Social Network	DBA	
Software	Database Administrator	
Technology Adoption	Systems Analyst	
Ubiquitous Computing		
Virtualization		
Virtual Reality		
Wireless Communication		

5.2. Phase 2: Coding the Data using Generative AI

A key challenge in DS research is not only extracting relevant initiatives from text, but also coding them in a systematic and theoretically meaningful way. Once DS initiatives data have been extracted, Phase 2 of our project involves using generative AI to aid in the content coding of the extracted initiatives based on certain theoretically grounded categorization frameworks. During this phase, we adopt two theoretically grounded categorization frameworks that have been widely used in prior DS research to develop AI prompt artifacts to design and evaluate the effectiveness of using generative AI in helping researchers with content coding of DS initiatives.

IT roles coding

The first categorization framework categorizes DS initiatives as Infrastructure, Automate, Informate, and Transform initiatives. It has been used to examine different types of IT assets, resources, and initiatives and their impacts on business performance (e.g., Aral and Weill, 2007; Dehning et al., 2003) as well as sustainability (Chen et al., 2008; Dao et al., 2011; Abraham and Dao, 2019).

- *Automate:* IS-enabled initiatives that help firms automate business processes to make them faster, more efficient, and more accurate.
- *Informate:* IS-enabled initiatives that make available data to employees and external entities to make better and faster decisions and carry out work processes more effectively and efficiently.
- *Transform:* IS-enabled initiatives that help firms fundamentally alter existing business processes and/or models that enable an organization to create new products/services that position firms more favorably in the product-markets.
- *Infrastructure*: IS-enabled initiatives that include both technical and human resources such as servers, networks, user devices, shared databases, help desks, etc., based on which automate, informate, and transform resources are developed and deployed.

At this stage, we iteratively designed, developed, and demonstrated the coding process using the same training dataset that includes data from Biogen, Coca-Cola, and Caterpillar. Fig 4 shows this process.

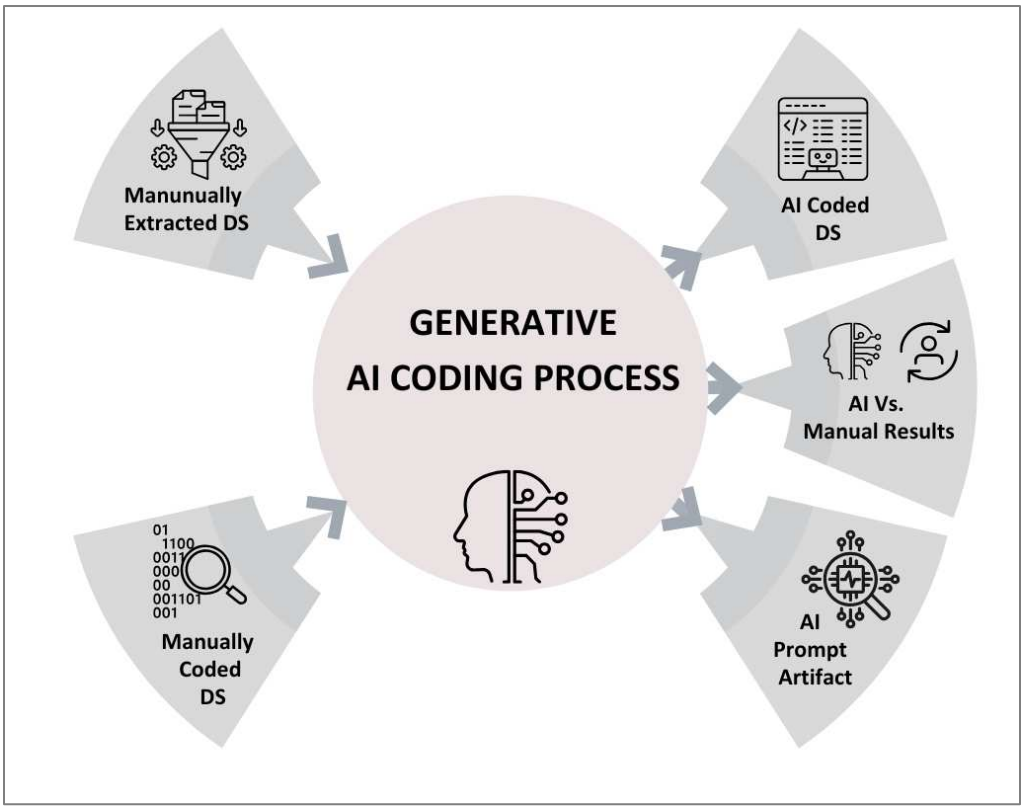


Fig. 4 Iterative prompt-building using a training dataset [Phase 2]

The coding and training process began by coding the IT Roles (AITInfra) with the most advanced model available at the time (*C*hatGPT 5.1), treating the model as a research associate and iteratively training it to understand both the coding prompt and the underlying rationale. We first developed a structured prompt (Fig 5), drawing on the coding scheme established in Dao and Abraham (2021).

I would like you to work as a researcher (***element 1 – role setting***) to conduct a content coding task as follows: We have a data set of discussions of IT-enabled business initiatives. You will need to read and analyze such discussions and classify the discussed IT-enabled business initiatives into one of the four following categories defined below (***element 2 – task instruction***):

- **Automate:** Initiatives that replace human labor by automating business processes.
- **Informate:** Initiatives that provide data/information to empower management, employees, or customers. OR initiatives that filter/target information for social impacts.
- **Transform:** Initiatives that fundamentally alter traditional ways of doing business by reconstituting business processes and relationships.
- **Infrastructure:** Initiatives that comprise of standardized IT services deployments (both technical and human, such as servers, networks, and help desk) that provide platforms through which standardized technical services are provisioned and from which automate, informate, and transform IT resources are utilized to deliver their intended business objectives. (***element 3 – context description***)

In the attached file, the third column includes discussions of the IT-enabled business initiatives. I want you to read through the discussions in the third column and conduct your content coding task (***element 4 – input data***). You should add your coding - Automate, Informate, Transform, or Infrastructure - to a fourth column titled "Coded role". You will include your rationale for assigning the coding to the discussion in the fifth column, titled "Explanation".

You will return the results as a spreadsheet file with the three original data columns and the fourth and fifth columns described above (***element 5 – context description***).

Fig 5 Original IT Roles Prompt artifact

The original scheme was revised and translated into a detailed prompt that specified the model's role, goal, and background, and organized each task into a description, expected output, and required output format. The prompt was used to code Biogen data, and AI-coded results were compared to manually coded results. After running the initial prompt on the Biogen dataset, we refined the prompt to improve the clarity and quality of the model's explanations, including asking the model to propose revisions and then testing the AI-revised prompt on the same data, which produced comparable results.

Afterward, we introduced a few-shot prompt to compare performance against the zero-shot prompt used earlier. In the few-shot prompt (Fig 6), a sample discussion for each of the initiative categories (Infra/A/I/T) is included in the prompt, together with a rationale on why such sample discussion is coded into the category. To compare results, we cleared chat history and created a new chat for each session to run the coding with zero-shot and few-shot prompts. Using this iterative approach, the model achieved an 84% match rate with the manually coded Biogen data (Table 4). The finalized zero-shot and few-shot prompts from this stage were then carried forward for use in the subsequent phase of the project.

During the next step, we use the refined prompt to code data of two additional companies, Coca-Cola and Caterpillar, for further prompt refinement and training. Using Coca-Cola data, the zero-shot prompt produced an initial match rate of 64%, while the few-shot version achieved 68.4%. To improve performance, we implemented a chain-of-thought fine-tuning procedure in which ChatGPT was provided with the manually coded Coca-Cola data and asked to compare its code to the manually coded data, examine coding discrepancies, and request clarification as needed. During this stage, the model identified several areas of divergence that prompted targeted clarification. For example, four initiatives related to energy management systems were coded by the model as "Automate" and "Informate," whereas expert researchers manually classified them as "Infrastructure". After explanation, the model adjusted its understanding to treat energy and plant-level systems as infrastructure rather than higher-level business activities. In another case, a manually coded "Transform" was coded as "Informate" by ChatGPT, which explained that it had erred on the conservative side for its coding rationale. After discussion, ChatGPT revised its logic accordingly. Through this iterative exchange, additional coding rules—mirroring the detailed decision criteria used in manual coding—were

incorporated into the prompt (Fig 6) to strengthen consistency and interpretive accuracy across datasets. We then applied the same iterative process to the Caterpillar dataset, where the zero-shot prompt initially matched 60% of the manual codes, and the few-shot prompt matched 72%.

Cohen's Kappa indexes were calculated to compare manually coded data and AI-coded data with few-shot prompts. The indexes were >0.7, indicating consistent high level of inter-rater reliability. To evaluate inter-rater reliability, Cohen's Kappa indexes were calculated to compare manually coded data and AI-coded data with the few-shots prompt. The indexes were >0.7, indicating consistent high level of inter-rater reliability.

Table 4 Comparing Manually Coded vs. ChatGPT-Coded IT Roles

Data	**Company**	**Manually extracted DSI**	**Matching Roles - zero-shot prompt**	**Matching Roles - with few-shot prompt**
Training	Biogen	19	16(84.2%)	16(84.2%)
	Coca-Cola	19	12(63.1%)	13(68.4)
	Caterpillar	25	15 (60.0%)	18(72.0%)
	Total	**61**	**43 (70.5%)**	**47 (77.5%)**

I would like you to work as a researcher (***role***) to conduct a content coding task as follows: We have a data set of discussions of IT-enabled business initiatives. You will need to read and analyze such discussions and classify the discussed IT-enabled business initiatives into one of the four following categories defined below (***task***):

- **Automate:** Initiatives that replace human labor by automating business processes.
- **Informate:** Initiatives that provide data/information to empower management, employees, or customers. OR initiatives that filter/target information for social impacts.
- **Transform:** Initiatives that fundamentally alter traditional ways of doing business by reconstituting business processes and relationships.
- **Infrastructure:** Initiatives that comprise of standardized IT services deployments (both technical and human, such as servers, networks, and help desk) that provide platforms through which standardized technical services are provisioned and from which automate, informate, and transform IT resources are utilized to deliver their intended business objectives. (***context description***)

In the attached file (***input data***), the fourth column includes discussions of the IT-enabled business initiatives. I want you to read through these discussions and conduct your content coding tasks. You should add your coding - A, I, T, or Infra for Automate, Informate, Transform, and Infrastructure, respectively - to a fifth column titled "Coded role". You will include your rationale for assigning the coding to the discussion in the sixth column, titled "Explanation".

You will return the results as a spreadsheet file with the four original data columns and the fifth and sixth columns described above (***output***).

Additional coding rules (***additional context***)**:**

- Code for highest Infrastructure à Automate à Informate àTransform when enough information available. Lowest when not enough information.
- Code for information about information technology that is embedded in industrial technology with enough detail.
- Initiatives creating new information flow: informate.
- Initiatives changing the way a marketplace operates: transform
- Teleconference system: infrastructure
- Energy infrastructure: Generally infrastructure, except when there is explicit information on IT-enabled automating energy management tasks or providing information to employees on energy management decision making.

To help you to better understand rationale on how to assign code to the different types of IT-enabled initiatives, I include below several initiatives that I have coded and the rationale on why I assigned A, I, T, or Infra code to the initiatives.

Automate IT-enabled Business Initiative (A)

Source: Coca-Cola's sustainability report:

Discussion: "We are enhancing our data collection processes all the time. A new database was developed by our Cooler Services team which has reduced manual data entry and helped to improve its quality. We have also rolled out our data systems for our manufacturing sites to our cold drinks centers and distribution sites, making data capture easier."

Rationale: Clearly, new IT has been implemented, but the discussion indicates that they have automated existing manual processes rather than created new information. Thus, automate is the appropriate code.

Informate IT-enabled Business Initiative (I)

Source: Johson & Johnson's sustainability report:

Discussion: "Meanwhile, we are creating a website to increase access to information about our responsibility approach and performance. We share progress internally through a variety of other mechanisms, including employee communications and intranet sites."

Rationale: Website enables the company to share its sustainability information with relevant stakeholders, helping them be more informed about the company's sustainability activities. Thus informate is the appropriate code.

Transform IT-enabled Business Initiative (T)

Source: Caterpillar's sustainability report:

Discussion: "One of the forward-looking projects under development is the creation of Semi-Autonomous Tractor Systems (SATS), which will enable one operator to control multiple dozers simultaneously."

Rationale: SATS initiative is revolutionary in terms of technology and market practices. Thus Transform is the appropriate code.

Infrastructure IT-enabled Business initiative (Infra)

Source: Biogen's sustainability report:

Discussion: "Green Data Center - Designed with the installation of a high-density rack and self-contained cooling system, our green data center maximizes performance while reducing energy consumption by 35 to 45 percent."

Rationale: Green data center provides core IT services that enable other types of IT-enabled initiatives. Thus, infrastructure is the appropriate code.

Fig 6 Final Few-shot IT Roles Prompt artifact

Quadrant Coding

To further evaluate the effectiveness of the AI-assisted content coding process artifact, we use the approach developed above and use a second DS initiatives categorization framework to conduct another round of content coding. Prior research on DS initiatives has also categorized them based on what areas of sustainability the initiatives address (Dao et al., 2011; Dao and Abraham, 2021). This categorization framework is based on the Integrated Sustainability Framework (ISF) that was theoretically developed in Dao et al. (2011). The framework includes four quadrants: internal today, external today, internal tomorrow, and external tomorrow, and IS-enabled initiatives are categorized as aiming to address sustainability issues within these quadrants:

- Quadrant 1: Initiatives addressing current internal sustainability performance (e.g., IS-enabled initiatives such as energy or safety monitors for internal use).
- Quadrant 2: Initiatives addressing current external sustainability performance (e.g., whole supply chain's sustainability performance).
- Quadrant 3: Initiatives addressing future internal sustainability performance. (e.g., IS-enabled initiatives such as data mining for internal new products).
- Quadrant 4: Initiatives addressing future external sustainability performance (e.g., IS-enabled initiatives such as wearable devices for underserved patients or self-driving trucks).

Building on the insights gained from the IT Roles coding process, we used a few-shot prompt approach for quadrant coding. The prompt was first developed and tested using Biogen data, yielding a 73.7% match with manually coded data by expert researchers. Subsequent fine-tuning was conducted with Coca-Cola and Caterpillar datasets, where match rates improved substantially to 94.7% and 84.0%, respectively (Table 5).

Table 5 Comparing manually vs. ChatGPT-coded ISF Quadrants

Data	Company	DSI	Matching Quadrant with few-shot prompt
Training	Biogen	19	14 (73.7%)
	Coca-Cola	19	18 (94.7%)
	Caterpillar	25	21 (84.0%)
	Total	**61**	**53(86.9%)**

The improved match percentages highlighted the need for more explicit contextual rules when coding DS initiatives. The few-shot prompt artifact for quadrant coding is shown in Fig 7. Cohen's Kappa indexes were calculated to compare manually coded data and AI-coded data with few-shot prompts. The indices were >0.7, indicating consistent high level of inter-rater reliability.

I would like you to work as a researcher (***role***) to conduct a content coding task as follows: We have a data set of discussions of IT-enabled business initiatives. You will need to read and analyze such discussions and classify the discussed IT-enabled business initiatives into one of the four following categories defined below (***task***):

- **Quadrant 1 (Internal-today):** Initiatives aimed at improving sustainability performance of current internal operation activities (e.g. Energy consumption reduction, inventory optimization, improving employee management practices.
- **Quadrant 2 (External-Today):** Initiatives aimed at current interactions with external stakeholders (Business partners, customers, supply chain partners, etc.) about current sustainability activities (e.g. Supplier evaluation, web portals for external stakeholders interaction, etc.)
- **Quadrant 3 (Internal-Tomorrow):** Initiatives aimed at developing (groundbreaking/radical) innovative technologies, products, and/or processes that disrupt industry standards practices that also help solve environmental and social concerns (e.g. New innovation platform, etc.)
- **Quadrant 4 (External-Tomorrow):** Initiatives aimed at exploring future opportunities with external stakeholders. For example, explore opportunities for unmet markets for sustainable new products/services that intersect with the companies' business competency.

In the attached file (***input data***), the fourth column includes discussions of the IT-enabled business initiatives. I want you to read through these discussions and conduct your content coding task. You should add your coding – Q1, Q2, Q3, or Q4 for Quadrant1, Quadrant 2, Quadrant 3, and Quadrant 4 respectively - to a fifth column titled "Coded Quadrant". You will include your rationale for assigning the coding to the discussion in the sixth column, titled "Explanation".

You will return the results as a spreadsheet file with the four original data columns and the fifth and sixth columns described above (***output***).

Additional coding rules (***additional context***)**:**

- If a paragraph does not clearly describe a specific IT system or application, still assign the quadrant based on its primary purpose (internal vs external, today vs tomorrow), assigning the lowest code when not enough information.
- Use Q3 or Q4 only when an initiative clearly involves new or experimental technologies, products, business models, or radical process redesigns aimed at future breakthroughs.
  If the initiative mainly improves the efficiency or performance of existing operations—even if benefits extend into the future—treat it as 'today' (Q1 or Q2)
- We treat sustainability broadly, including environmental (e.g., energy, CO_2, resource use) and social aspects (e.g., health access, safety, underserved communities, employee well-being.
- Websites or portals for internal communication with employees: code as Internal
- Initiatives focused on future social and economic benefits for underserved communities: code as external.
- Systems that ingest external data (customer/supplier) but are primarily used for internal decision-making and product or risk management: Code as internal.
- When tools are framed explicitly as CSR / dealership network / global operational reporting rather than plant-level operations: Code as external.
- Initiatives aimed primarily at Internal product development: Internal
- Initiatives aimed at future benefits for customers and society: coded as External tomorrow (Q4)
- Initiatives aimed at current or near-market benefits for customers and society: coded as external today (Q2)
- Internal decision-support with external data → Q1
  Examples: internal product-development databases, supplier risk-scoring systems, internal compliance dashboards, internal EHS risk models that use external benchmarks.
- External transparency / engagement tools → Q2 or Q4
  Examples:
  - Initiatives communicating or reporting sustainability-relevant information to customers, communities, suppliers, investors, regulators, etc.

- Initiatives providing ongoing services / tools directly to external stakeholders (e.g., customer portals, public websites with sustainability information, external telematics services, community-facing apps).

Examples of Coded initiatives and coding Rationale (*few shots*):
To help you to better understand rationale on how to assign code to the different types of IT-enabled initiatives, I include below several initiatives that I have coded and the rationale on why I assigned Q1, Q2, Q3, or Q4 code to the initiatives.

Q1 IT-enabled Business Initiative
Source: Biogen's Sustainability Report.

Discussion: "Biogen Idec is committed to maintaining a healthy and safe work environment. We have adopted a formal health and safety management system, which we update regularly to keep pace with the changing requirements of our business. We also require all employees to attend safety classes specifically tailored to their lines of work."

Rationale: The adoption of a formal health and safety management system helps maintain a good current internal working environment for the company's employees. Therefore, the focus of this initiative is on current and internal business operations, thus it is coded as a Quadrant 1 initiative.

Q2 IT-enabled Business Initiative

Source: Volvo Construction's Sustainability Report.

Discussion: "The Commute Greener! application has been developed by Volvo IT as an open innovation to welcome many partners. External target groups are corporations, cities, organizations and individuals. In addition to the application itself, Commute Greener! also provides an engaging community website where commuters can challenge each other to reduce their individual carbon dioxide footprint."

Rationale: The IT-enabled initiative enhances the company's communication and engagement with its current external stakeholders. Such enhanced engagement helps improve the company's current sustainability performance.

Q3 IT-enabled Business Initiative

Source: Coca-Cola, Sustainability Report.

Discussion: "Within our Information Technology (IT) department we have adapted the Lean Startup methodology to support our innovation process, enabling ideas to be generated and developed quickly and cost-effectively. Colleagues are able to access our Ideation Platform, either from their laptop or via a mobile phone app, where they can post an idea. If their idea receives enough 'likes', it's developed into a proof of concept and presented to senior management for a decision on future investment. We have learned equally from our successes and our failures and are building these insights into our processes. We have been trialing the use of a 3D printer within our business, as well as the use of beacon messaging. The idea generation process has also started to expand to other areas of our business such as with our Field Sales teams in Great Britain. I'm very excited to see what innovative ideas will be developed."

Rationale: New IT-enabled initiatives have been implemented to help with innovations, improving the company's future sustainability performance and success with new ideas.

Q4 IT-enabled Business Initiative

Source: Johnson & Johnson, Sustainability Report.

Discussion: In conjunction with other partners, launched the Mobile Alliance for Maternal Action, which uses mobile phones to deliver prenatal and post-birth health information to new and expectant mothers in the developing world. Over the next three years, the program will reach mothers in Bangladesh, India and South Africa.

Rationale: The initiative aims to reach out to under-served communities and external stakeholders. This helps improve services to these under-served communities and expand future opportunities for the company, improving future sustainability performance.

Fig 7 Final Few-shot Quadrants Prompt Artifact

6. Evaluation

The core activities of design science research are 'build' (construct an artifact for a specific purpose) and 'evaluate' (determine how well the artifact performs) (March & Smith, 1995). The process of extracting and coding DS data using text mining and generative AI, the DS data dictionary, and our coding prompts could all be considered the artifacts of this design science research project (Gregor & Hevener, 2011; Klessascheck et al., 2025). While the dictionary building and prompting process could be an ongoing iterative one, we found that the artifacts had reached a mature enough state to be evaluated. "The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods" (Hevner et al., 2004, p. 85). Hevner et al. (2004) list 5 methods for design evaluation (Table 2, p. 86). We adopted the experimental method and used the artifacts in a controlled experiment to test their outputs and compared them to manually collected and coded data.

6.1. Evaluating data collection and dictionary-building phase

To evaluate the usability of our dictionary, we tested it out on three other companies, J&J, General Mills, and Volvo. We compared the data semi-automatically retrieved using WordStat and our newly built dictionary with the data that had previously been manually extracted from these 3 companies.

The last row of Table 6 shows the results of the test. While the dictionary found only 34 manually extracted DSI out of 60, it found many more instances of DSI. Of the total 140 DSI extracted, WordStat found 114 (81%). We conclude that using text analytics with a DS data dictionary significantly outperforms manual data extraction.

Table 6 Comparing Manually Extracted vs. WordStat Extracted DSI

Data	**Company**	**Manually extracted DSI (1)**	**Overlap DSI (2)**	**WordStat extracted DSI (3)**	**Total Distinct DSI (1 & 3)**	**% Manual**	**% Overlap**	**% WordStat**
Testing	J&J	11	5	30	36	31%	14%	83%
	General Mills	21	12	29	38	55%	32%	76%
	Volvo	28	17	55	66	42%	26%	83%
	Total	**60**	**34**	**114**	**140**	**43%**	**24%**	**81%**

6.2 *Evaluating Data Coding using Generative AI*

To evaluate our coding artifacts, we compared our ChatGPT coded DSI with our previously manually coded initiatives. Again, we used the manually coded data for J&J, General Mills, and Volvo DSIs to evaluate the prompting artifacts. The last rows of Tables 7 and 8 show the results of the test. The ChatGPT codes matched 78% of the IT Roles and 85.3% of the Quadrants assigned by expert coders. We conclude that using generative AI is a viable option for coding DSI data. High inter-rater reliability was achieved with a calculated Cohen's Kappa consistently >0.7.

Table 7 Comparing manually vs. ChatGPT-coded IT Roles

Data	**Company**	**DSI**	**Matching Roles - with few-shot**
Testing	J&J	30	25(83.3%)
	General Mills	22	19(86.4%)
	Volvo	29	20 (69.0%)
	Total	**82**	**64 (78.0%)**

Table 8 Comparing manually vs. ChatGPT coded ISF Quadrants

Data	**Company**	**DSI**	**Matching Quadrants with few-shot**
Testing	J&J	30	28 (93.3%)
	General Mills	23	20(90.9%)
	Volvo	29	22 (75.9%)
	Total	**82**	**70 (85.3%)**

7. Discussion

In this paper, we developed a semi-automatic process model for DS data extraction and coding. The process is grounded in the existing DS and text analytics literature and can support future research focused on extracting and coding DSy initiatives from secondary textual data sources. Based on the design and evaluation of our semi-automatic

DS data collection and coding process artifacts, we can derive a set of lessons learned from our project and some general guidelines for researchers. Table 9 summarizes these lessons and guidance for the field as well as points to issues on which to reflect.

Table 9 Lessons Learned and Guidelines

Area	**Lessons Learned**	**General/Practical Guidelines**
Methodology	Our process artifact followed the design science stages, and the data dictionary and AI prompt artifacts were constructed through several iterations.	The design science approach of building artifacts in an iterative manner works very well with the context of unstructured text data. It could apply to other domains.
Theory-based	Our artifacts were based on prior, published DS research.	The quality of the artifacts depends on the use of established theories and frameworks and prior published research.
Tools	Text mining and generative AI tools were more efficient and effective than manual data extraction and coding.	The tools are rapidly improving through each version so results may continue to improve. Different AI models may prove more effective in different contexts.
Validity	The extracted data needed to be checked and validated manually. The coded data, after some training provides valid results, even helping us fine-tune our own coding rules.	The data extraction and coding is a semi-automatic process which needs human oversight. Generative AI can be an additional researcher to help validate the coding choices of human researchers.
Scalability	The semi-automated process reduces time and labor, enabling the analysis of large datasets.	Large-scale extraction and coding should be followed by validation procedures, such as manual sample review.
Limitations	The incompleteness of the dictionary artifact may reduce the accuracy if not adapted before use. AI coding is limited in understanding business context of initiatives and when coding depends on expert understanding of such business context nuances.	Dictionary building in general is never complete. Researchers need to iteratively update and tune the DS data dictionary artifact. It is important that human expert researchers involve in the early phase of the coding process to help train the AI model with the coding and develop detailed coding rules to be incorporated in coding prompts.

The rest of section 7 elaborates on lessons learned and general guidelines.

7. 1. Learning and Reflection

Our research has shown that using a design science approach and tools such as text mining and generative AI could be a very useful approach for researchers to extract and code larger DSI datasets from secondary data sources. Our process artifact followed the design science stages, and the data dictionary and AI prompt artifacts were constructed through several iterations. The semi-automated process artifact for extracting and coding text from reports should apply to any research context. However, the process will involve multiple iterations to refine the artifacts.

We grounded our artifacts in prior, published research. Following design science research requirements (Walls et al., 1992; Markus et al., 2002), our process artifact is based in theory, and it provides guidance to practitioners, in this case DS researchers. Our initial data dictionary emerged from keywords identified in a literature review paper (Abraham et al, 2026). Our AI coding prompts used the established IS frameworks of IT Roles and ISF Quadrants (Dao et al., 2011).

We found that extracting examples of DS from sustainability reports using a text mining tool like WordStat with our data dictionary artifact was far more effective and efficient than our prior manual approach. In our testing data, WordStat extracted 54 more examples of DSI than our manual search, even if it did miss 26 examples. This should give researchers a useful approach to extract a larger scale of DS data from secondary sources.

We also found that coding that data using a generative AI tool like ChatGPT and common DS frameworks was very accurate after fine-tuning our prompts through multiple training iterations. For the IT Roles coding, our testing data showed a 78% match between manual and ChatGPT codes. For quadrant coding, once we had gone through the process to develop a few-shot prompt with more clarification coding rules, our testing data showed an even higher 85% match between manual and ChatGPT codes.

In terms of validity, the data extracted by WordStat needed to be checked and validated manually. The tool returns many paragraphs that do not reflect distinct DS initiatives. We found that the generative AI model, after some training using few-shot prompts approach and chain of thought fine-tuning, provides valid content coding results. We also found that in some cases, the explanations provided by ChatGPT for its coding choices were compelling and helped us fine-tune our coding rules for coding consistency.

7.2: Generalized design principles

The design science approach of building artifacts in an iterative manner works very well with the context of unstructured text data. It could apply to other domains that require theory-based information extraction and coding from unstructured data by replacing the DS dictionary with domain-specific terms. However, the process will involve multiple iterations to refine the artifacts. Training and testing datasets should be kept separate, with the former used in development and demonstration stages while the latter used for evaluation.

The quality of the artifacts depends on the use of established theories and frameworks and prior published research. Research practitioners looking to adopt our process artifact will need to build their dictionary and prompt artifacts on context-specific theories and coding rules.

We expect that the text-mining and generative AI tools are likely to improve as these companies introduce newer versions and models. For instance, when we moved from ChatGPT 4 to 5, we experienced a marked improvement in coding performance. Nevertheless, human oversight by experts is necessary for valid results.

The coding results suggest that using generative AI for coding data can be used not just for efficiency but also as an additional researcher to help validate the coding choices for human researchers. For AI content coding, AI models should be treated like a collaborative research assistant that needs to be trained and supervised by expert researchers to help with content coding before being allowed to conduct the coding independently. In a typical content coding project, expert researchers would first develop a draft coding scheme, independently code an initial batch of data based on the coding scheme, then compare and discuss coding discrepancies to help them fine-tune the coding scheme and be more consistent with their coding to improve inter-rater reliability. Only when a satisfactory level of inter-rater reliability is reached (typically Cohen's Kappa indexes >0.7) then the "research assistant" would be able to independently help with coding the remaining part of a larger data set.

The same process should be applied for researchers who want to use AI models to help with content coding a large dataset. First, it is necessary that expert researchers develop a coding scheme and spend time working on coding an initial batch of data to help them understand better the nuances of the business context and to have a batch of human-coded data for AI training. The coding scheme should be carefully turned into a well-crafted prompt, which then is used to ask AI models to conduct the coding task. A few-shot approach should be adopted in crafting a prompt for content coding, with examples and coding rationales provided for coded categories, to help guide a generative AI's automated coding. Additionally, to enhance coding consistency, it is necessary for expert researchers to use an initial batch of human coded data set to feed to the AI model to train the model and add further clarification coding rules when needed. Only once a consistent high level of inter-rater reliability has been reached can the expert human researcher use the AI model to conduct the coding for the rest of the dataset.

7.3. Contributions and implications

This study makes several contributions to the literature on DS, text analytics, and the application of Generative AI in research practice. It creates a semi-automated process to address the persistent problem of collecting and coding data from dispersed secondary courses. The primary contribution is a reusable process artifact that integrates text mining

and generative AI for data extraction and theory-driven content coding. Moreover, the study provides a validated DS data dictionary spanning Green IT, Green IS, Social IS, and IS organization to support DS data extraction from large unstructured datasets. The generative AI prompt artifacts are anchored in the literature of IT Roles and ISF quadrants and were refined using techniques such as interactive dialogue between researcher and model and using zero-shot and few shot examples.

Our study also establishes the utility of technology tools such as text mining and generative AI in DS research with the important proviso that they are used with expert, human supervision. At a fraught time with threats to environment and society, including from AI and data centers, it will be incumbent on DS researchers to monitor the promises and perils of technology. In the field of Information Systems (IS), sitting at the intersection of business and technology, editorials from the top IS journals call for research on responsible technology (Benbya and Melville, 2025; Brown, 2025; Recker et al., 2025). To conduct empirical research in this area, tools such as the ones used in this study will prove very useful. The practical implication of our study is that it guides the research community through the process of embedding text mining and AI within the research process.

7.4. Limitations and Future Directions

Our research does have limitations. The first limitation is the incompleteness of the dictionary artifact. Dictionary building in general is never complete. As new terms appear in the field, the dictionary must be updated. We used data from 2009 to 2015; thus, the dictionary was not evaluated using newer reports.

Additionally, there is also a limitation in terms of the trade-off between keeping terms which produce a lot of noise in the results versus dropping them and missing important DS initiatives. Terms like 'Platform' produced a great deal of unhelpful data, but dropping terms like 'Website' caused us to miss some useful DS initiatives that we had extracted manually. As we saw in one of our testing cases, J&J, we missed 6 DS initiatives that we extracted using search terms like System and Information. These include Data Management System, Mobile Alliance for Maternal Action (MAMA), Employee Data Management System, Training and Grievance data, mMITRA, and Supplier Environmental Assessment. Researchers would need to iteratively update and tune the DS data dictionary artifact.

The second limitation is the need for contextual knowledge when using generative AI for coding. When conducting content coding, it is important that the coding process takes into consideration nuances and specific industry or business conditions in assigning code to initiatives. For example, initiatives that took multiple years to complete (e.g., General Mills' Global Safety Tracking and Environmental Management System) were mentioned in reports across multiple years. However, it would only make sense to code for the initiative during its first mentioning. In another example, information technology that is embedded in industrial technology for manufacturing should only be coded if enough detailed information is provided. Therefore, it is still important for expert researchers to be involved in the coding process to train AI models and to understand and incorporate more detailed coding rules into the prompt to account for such nuances and context-dependent coding situations.

8. Conclusion

In this study, we developed a semi-automatic process for data extraction and coding of DS data from textual sources. We adopted the design science research method to iteratively build and evaluate three design artifacts – the process, the DS data dictionary, and the generative AI coding prompts. The study follows two phases – phase one is data collection from textual sources and phase two is content coding using established frameworks. In each phase, we used the design science research method to design, develop, demonstrate, and evaluate the artifacts. We compared the results to data collected manually from the same sources in prior published research. We found that data collected via text mining was more efficient and effective than the manual method but still required substantial human supervision. When we compared codes generated by AI with manually generated codes, we found that the codes matched to a high degree but again needed tweaking by an expert researcher. ChatGPT also provided insights useful for identifying and tackling ambiguities. This suggests that AI can serve not only as a tool for efficiency, but also as an additional analytical aid for verifying and refining human coding judgments.

The study has several contributions to digital sustainability and IS literature. First, the artifacts provide a practical template for DS researchers seeking to create large datasets from textual sources and code extensive textual data. Second, the study offers general guidelines for researchers using tools such as text mining and generative AI. Third,

the study develops a digital sustainability dictionary that future researchers can adapt after doing minor tuning to align with their research objectives.

REFERENCES

Abraham, T and Dao, V (2019) A longitudinal exploratory investigation of innovation systems and sustainability maturity using case studies in three industries. Journal of Enterprise Information Management 32(4):668-687

Abraham, T, Dao, V, and El-Rayes, N (2026) Information Systems for Sustainability Research: A Literature Review Pacific Asia Journal of the Association for Information Systems. 18(2):75-100 https://doiorg/1017705/1pais18204

Alotaibi, E, and Alwathnani, A (2025) AI-Enabled ESG Compliance Audit for Stakeholders, Sustainability, 17(21)

Aral, S and Weill, P (2007) IT assets, organizational capabilities, and firm performance: how resource allocations and organizational differences explain performance variation. Organization Science, 18(5):763-780

Balcioglu, Y, Celik, A, Altindag, E (2024) Artificial Intelligence Integration in Sustainable Business Practices: A Text Mining Analysis of USA Firms. Sustainability, 16(15)

Benbya, H and Melville, N P (2025) Special Issue Editorial: Fueling the Potential of Artificial Intelligence for Societal Impact, MIS Quarterly Executive: 24(2), Article 2

Boritz, JE, Hayes, L and Lim, JH (2013) A content analysis of auditors' reports on IT internal control weaknesses: The comparative advantages of an automated approach to control weakness identification. International Journal of Accounting Information Systems, 14(2):138-163

Bressane, A, de Castro, MV, Nomura, LM and Ewbank, H (2026) Corporate Environmental Intelligence as an Emerging Governance Framework in Digital Sustainability: A Scoping Review. Environmental Development, 101467

Brier, A and Hopp, B (2011) Computer assisted text analysis in the social sciences. Quality & Quantity, 45(1):103-128

Brown, S (2025) Editor's Comments: Beyond Green IT: From Digital Sustainability to Responsible Technology, *MIS Quarterly*, 49(4):iii-viii

Bu, Y, Lee, CS, Wang, D and Morishima, A (2025) Editorial: Application of artificial intelligence in information services: part 1. The Electronic Library, 43(5):693-694

Carberry, EJ, Bharati, P, Levy, DL and Chaudhury, A (2019) Social movements as catalysts for corporate social innovation: Environmental activism and the adoption of green information systems. Business & Society, 58(5):1083-1127

Chen, A, Boudreau, M and Watson, R (2008) Information systems and ecological sustainability. Journal of Systems and Information Technology, 10(3):186-201

Chong, SE, Looi, KH and Shahudin, F (2025) Beyond the prompt: exploring technostress, flow and personality in generative AI behaviors. The Electronic Library, 43(5):733-756

Dao, VT and Abraham, T (2021) An Empirical Examination of the Use of IS-enabled Sustainability Initiatives Across the Integrated Sustainability Framework. Pacific Asia Journal of the Association for Information Systems, 13(3):57-85

Dao, V, Langella, I, and Carbo, J (2011) From green to sustainability: Information Technology and an integrated sustainability framework. The Journal of Strategic Information Systems, 20(1)):63-79

Dehning, B, Richardson, VJ and Zmud, RW (2003) The value relevance of announcements of transformational information technology investments. MIS Quarterly, 27(4):637-656

Deng, Q, Hine, J, Ji, S and Sur, S (2019) Inside the Black Box of Dictionary Building for Text Analytics: A Design Science Approach Journal of International Technology and Information Management, 27(3):119-159

El-Rayes, N, Abraham, T and Dao, V (2022) Bibliometric Analysis of Sustainability Information Systems. AMCIS 2022 Proceedings, 16

Florek-Paszkowska, A and Ujwary-Gil, A (2025) The Digital-Sustainability Ecosystem: A conceptual framework for digital transformation and sustainable innovation. Journal of Entrepreneurship, Management and Innovation, 21(2):116-137

Gregor, S, and Hevner, A R (2011) Introduction to the special issue on design science *Information Systems and e-Business Management*, 9(1):1-9

Guo, L, Wang, Y and Zhang, A (2026) Do AI-related papers in top management science journals contribute greater academic influence and online attention? The Electronic Library, 44(1):73-108

Hamilton, L, Elliott, D, Quick, A, Smith, S and Choplin, V (2023) Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. International Journal of Qualitative Methods, 22:1-13

Hart, S, Milstein, MB (2003) Creating sustainable value. Academy of Management Executive 17:56–67

Hevner, AR, March, ST, Park, J and Ram, S (2004) Design science in information systems research. MIS Quarterly, 28(1):75-105

Ittoo, A, Nguyen, LM and van den Bosch, A (2016) Text analytics in industry: Challenges, desiderata and trends. Computers in Industry, 78:96-107

Klessascheck, F, Weber, I, & Pufahl, L (2025) SOPA: a framework for sustainability-oriented process analysis and re-design in business process management. Information Systems and e-Business Management, 23(2):489-537

Kotlarsky, J, Oshri, I and Sekulic, N (2023) Digital sustainability in information systems research: Conceptual foundations and future directions. Journal of the Association for Information Systems, 24(4):936-952

Laver, M and Garry, J (2000) Estimating policy positions from political texts. American Journal of Political Science, 44(3):619-634

Livari, J (2015) Distinguishing and contrasting two strategies for design science research. European Journal of Information Systems, 24(1):107-115

March, ST and Smith, GF (1995) Design and natural science research on information technology. Decision Support Systems, 15(4):251-266

Markus, ML et al. (2002) A Design Theory For Systems That Support Emergent Knowledge Processes. MIS Quarterly, 26(3)

Mishra, S, Ewing, MT and Cooper, HB (2022) Artificial intelligence focus and firm performance. Journal of the Academy of Marketing Science, 50(6):1176-1197

Morgan, DL (2023) Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. International Journal of Qualitative Methods, 22:1-10

Nunamaker, J F, Chen, M, and Purdin, T D M (1991) Systems Development in Information Systems Research. Journal of Management Information Systems 7(3):89

OpenAI (2023) ChatGPT. https://chatopenaicom Accessed 30 September 2023

OpenAI (2025) GPT-5. https://openaicom/gpt-5/ Accessed 29 October 2025

Pan, SL and Nishant, R (2023) Artificial intelligence for digital sustainability: An insight into domain-specific research and future directions. International Journal of Information Management, 72:102668

Pappas, IO, Mikalef, P, Dwivedi, YK, Jaccheri, L and Krogstie, J (2023) Responsible digital transformation for a sustainable society. Information Systems Frontiers, 25(3):945-953

Peffers, K, Tuunanen, T, Rothenberger, MA and Chatterjee, S (2007) A design science research methodology for information systems research. Journal of Management Information Systems, 24(3):45-77

Peffers, K, Tuunanen, T and Niehaves, B (2018) Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research. European Journal of Information Systems, 27(2):129-139

Provalis Research (2025) Provalis Research. https://provalisresearchcom/ Accessed 1 June 2023

Recker, J Chatterjee, S Sundermeier, J and Tarafdar, M (2025) Digital Responsibility: Current Perspectives and Future Directions. Journal of the Association for Information Systems, 26(5):1222-1238

Schein, EH (1992) The role of the CEO in the management of change: The case of information technology, in Useem, TAKM (Ed). Transforming Organizations, Oxford University Press

Schoormann, T, Paeplow, J and Möller, F (2026) Digital Sustainability Definitions: The Same but Different. Proceedings of the 59th Hawaii International Conference on System Sciences, DOI: 1024251/HICSS2026109

Strobel, G, Schoormann, T, Banh, L and Möller, F (2023) Artificial Intelligence for Sign Language Translation – A Design Science Research Study. Communications of the Association for Information Systems, 53:42-64

vom Brocke, J, Hevner, A and Maedche, A (2020) Introduction to design science research. in Design Science Research Cases, Springer, Heidelberg1-13

Watson, RT, Boudreau, MC and Chen, AJ (2010) Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community. MIS Quarterly, 34(1):23-38

Zhang, M, Shen, Q, Zhao, Z, Wang, S and Huang, G (2025) Optimizing ESG reporting: Innovating with E-Bert models in nature language processing. Expert Systems with Applications, 265

Zhou, Y, Yuan, Y, Huang, K and Hu, X (2024) Can ChatGPT Perform a Grounded Theory Approach to do Risk Analysis? An Empirical Study. Journal of Management Information Systems, 41(4):982-1015

Zou, Y, Shi, M, Chen, Z, Deng, Z, Lei, Z, Zeng, Z, et al (2025) ESGReveal: An LLM-based approach for extracting structured data from ESG reports, *Journal of Cleaner Production*, 489, p 144572

Zuboff, S (1988) In the Age of the Smart Machine: The Future of Work and Power, Basic Books, New York