Straive™

Connecting the dots

Why data extraction is the game changer in ESG reporting

WHITEPAPER

Driving Value from Data

Companies increasingly face many environmental, social, and governance (ESG)requirements from sustainability, supply chain compliance, and stakeholder activism perspectives. But nearly 80%-90% of the data that a company has access to isunstructured. This data is mainly in text form spread across base documents likeinvoices, utility bills, policy documents, etc.

The process of manual data extraction is tedious and time-consuming. Moreover, some of the data is in the form of scanned PDFs and images, adding another layer of complexity. Hence, the text intelligence process is required to analyze valuable insights from this data.

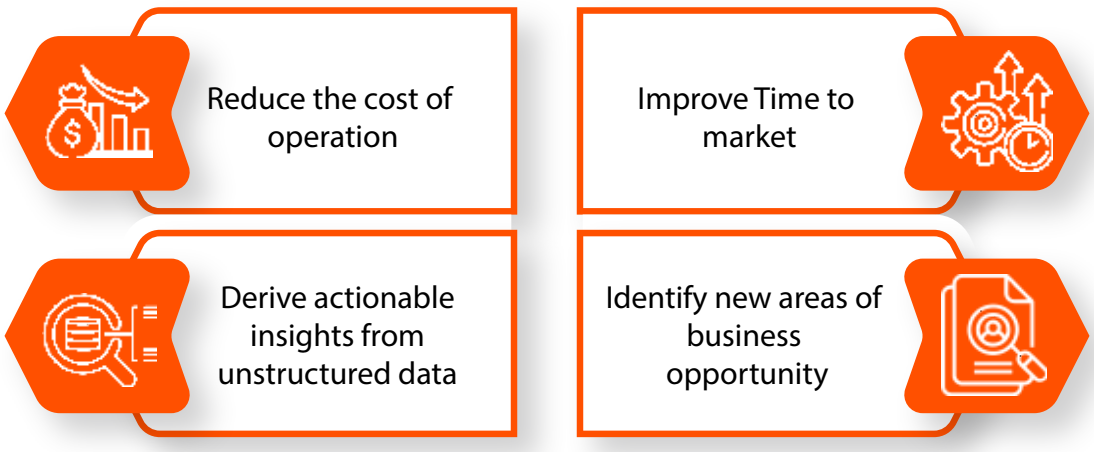
Not accessing and using this data means not unlocking potential insights for an organization, and missing out on a significant opportunity to accelerate innovation and boost revenue. Data extraction tools help unlock substantial insights that fasten business growth.

Text Intelligence involves retrieving information from unstructured data to uncover trends and patterns and derive insights from the output data. For enterprises, it has been a key enabler in tracking the digital footprints of customers.

Recently, there has been a shift to a global analysis of textual data for more varied use cases. Nevertheless, garnering intelligence from unstructured data in the form of text – in emails, social media conversations, chats, annual reports, press releases, scientific publications, blogs, mobile transactions, customer transcripts, title deeds, and more – is easier said than done.

It is demanding to build automated tools to analyze textual data from scanned PDFs/images, as it requires a blend of artificial intelligence (AI), machine learning (ML), and natural language processing (NLP).

Exhibit 1: Advantages of text intelligence

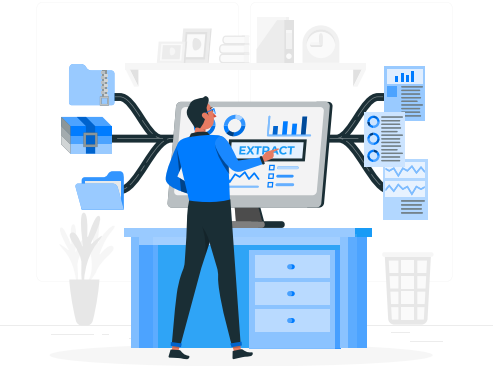


Source: Straive.

There are many challenges in leveraging public data for gaining insights regarding ESG use cases too. For instance, such data is unstructured and raw. Most of the content in the public space is not meant for structured consumption and analysis. The content has been developed as research-based and informative sources. The publication frequency is also highly variable, ranging from regular to annual, half-yearly, quarterly, to sometimes daily and hourly frequencies. Ordinarily, there is no specific cadence, standard, or format to enable data processing efficiently.

Thus, effective ESG integration is made possible by an automated data extraction tool that can automatically identify, extract, and summarize meaningful information for better insights and faster decision-making. For instance, the Straive Data Platform (SDP) is a data extraction tool built on a microservices architecture leveraging ML/NLP algorithms to enable enterprises to get data faster, with better quality and scale.

Data extraction effectiveness is a competitive advantage

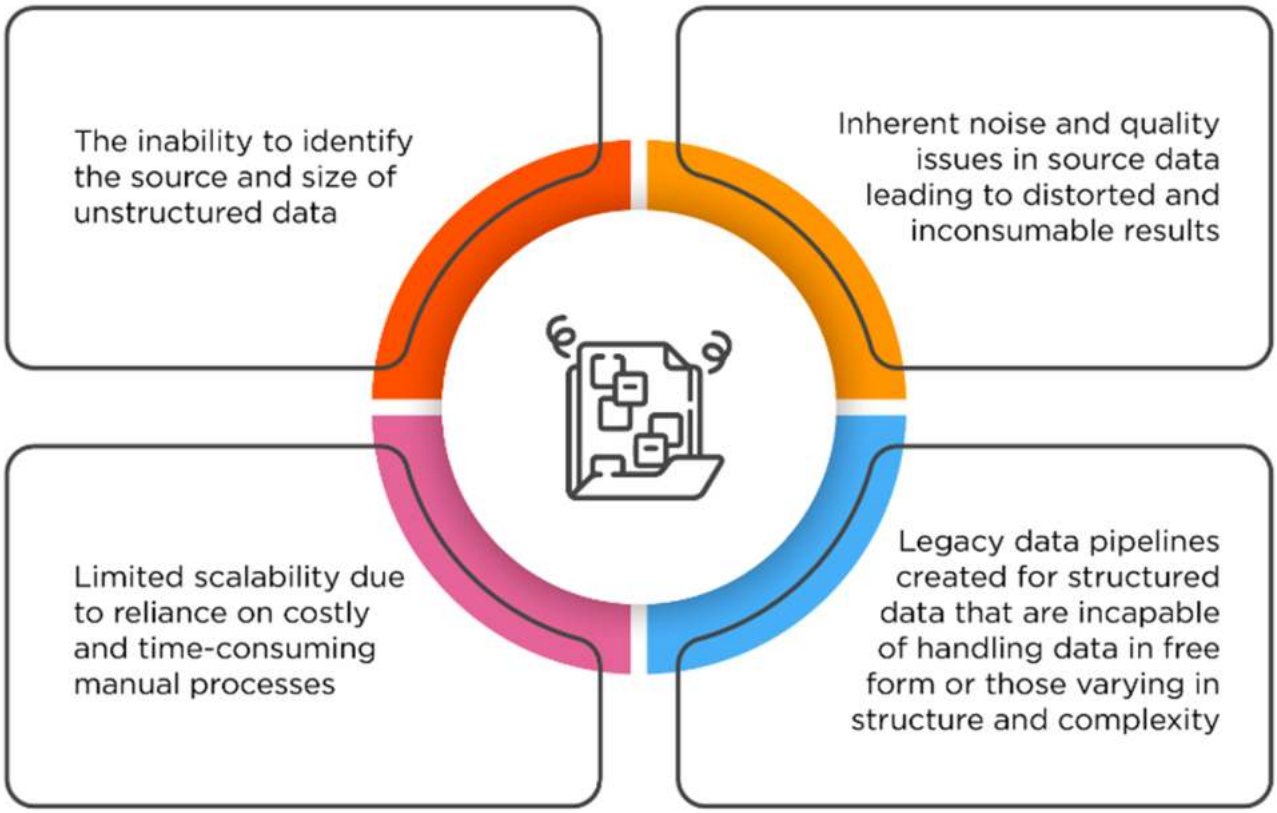


The lack of standardized formats for capturing and processing ESG data is an industry-wide problem. The lack of comparable ESG metrics and associated ratings often makes understanding materially significant ESG allocation difficult for asset managers. They need support when gathering, processing, and interpreting ESG data reliably at scale.

Vast ESG data sets, composed mainly of unstructured data, require advanced data analytics and AI to process and report ESG data quickly. Prompt reporting, in turn, will enable investors to verify progress reports with reliable data. ESG data solution activities begin at the collection stage and intensify during the analysis and verification stages.

However, many companies lack the technology and the tools to perform sophisticated data analytics and AI tasks on a large scale. Access to tools like ML/NLP is necessary. Many ESG data and insights are trapped in documents, digital pictures ,and videos, audio files, web content, social media chatter, climate sensors, etc. Leveraging unstructured data comes with its share of challenges.

Exhibit 2: Challenges in ESG unstructured data analysis



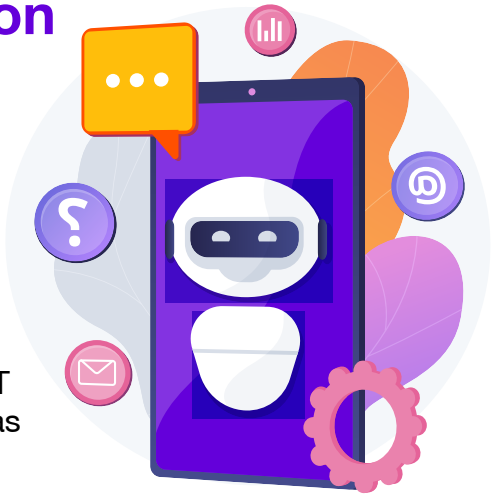
Source: Straive.

ESG data are usually updated only once a year, following annual company reports. However, daily tracking and a weekly update of critical events may be essential for monitoring a company's performance across metrics. System alerts can be used for controversy tracking to quickly communicate material ESG changes to those supervising developments, for example, by using customized email alerts. To that end, ML/NLP offers enormous potential for extracting information and mining insights from unstructured data.

Because traditional self-reported ESG data, such as corporate social responsibility (CSR) reports, are released at particular times, they lack a real-time connection. But alternative data offer a quasi-real-time view of market sentiments using various routes, for example, social media chatter, news articles, satellite images, independent non-governmental organizations' reports, and others. Often associated with financial analysis, alternative data is a term that typically refers to externally sourced information about a particular company to gain additional business insights.

Companies are better placed for ESG integration when data can be extracted efficiently from numerous unstructured documents and curated in a customized format for data analytics tools to consume. Having unstructured data in a straightforward, easy-to-understand format, a user can detect the patterns and trends without manually trawling for those data

ChatGPT and ESG Data Extraction



The advent of ChatGPT has upended and disrupted many workflows across industries and made multiple workflows easier. Does this have a bearing on ESG data extraction workflows too?

As an AI language model, ChatGPT can help provide information and extract data related to ESG topics. It can explain key concepts related to ESG, like environmental impact metrics, social responsibility indicators, etc. ChatGPT can also provide information on sources of ESG data, such as research firms, sustainability reports, databases, etc.

To those in the thick of ESG action, ChatGPT can assist in identifying specific ESG metrics and indicators commonly used to assess sustainability and ethical performance, like carbon emissions, employee diversity, board composition, etc. Analytics too gets easier since it can help analyze and compare the ESG performance of various companies based on available data, including examining company reports, financial statements, ESG ratings from third-party sources, etc.

ChatGPT can even provide information about various ESG reporting frameworks and standards, like the Global Reporting Initiative (GRI), Sustainability Accounting Standards Board (SASB), and Task Force on Climate-related Financial Disclosures (TCFD). But it cannot directly access real-time or proprietary data since it cannot directly web scrape or use APIs of data providers. ChatGPT, when deployed without SQL support, depends solely on NLP techniques to understand and generate responses when dealing with tabular data.[https://www.pragnakalp.com/data-extraction-from-tabular-data-with-chatgpt/] This can often result in unreliable or inaccurate responses. Though results improve when combined with SQL, reliability can remain a significant issue for sensitive data operations.

Analysis of real-time data with ChatGPT is prone to risks and not advisable at all. The biases in the training data can significantly influence outcomes deleteriously. ChatGPT has shown glimpses of its incapacity to perform aggregations (like summing or averaging tabular entries), rendering it a less than reliable ally to deal with name-value pairs across structured data tables and historical time series data. This also means that separate training models may be needed to suit ESG customizations when deploying ChatGPT, which is an additional effort.

An effective data extraction process

An effective data extraction process should be able to identify and extract data from any number of raw data sources. Further, data annotation is critical in labeling data found in images, scanned PDFs, and videos. It helps AI models identify specific data types and deliver relevant output.

The process of labeling data sets to make them machine-readable is data annotation. Data annotation or labeling is regarded as an indispensable adjutant to machine learning. They help develop and enhance the capability of machines to identify patterns from previous experience or data.

Consequently, if data is incorrectly annotated, the algorithms will deliver results that do not meet the business objective for which they were generated. The algorithms will learn the wrong lessons from the incorrectly annotated data, execute wrong calculations, and deliver misleading results. For specific and complex business objectives, Straive believes it is advisable to engage subject matter experts to ensure the quality and relevance of the annotated data.

Data annotation and labeling solutions fulfill the critical and specific role of providing enterprises with quality training data for their AI models. Straive's Straive Data Platform (SDP) uses in-built NLP and ML-based engines to enrich data by automating—

- Named Entity Recognition (NER)
- Concept Extraction using domain taxonomies or controlled vocabularies
- Summarization
- Classification to Taxonomy
- Business relevancy terms
- Creation of Domain Taxonomies or Ontologies

Exhibit 3: Data annotation types

Type	Brief
Visual data annotation	Visual data annotation is labeling images by identifying key points (or pixels) with precise tags in a format the system grasps. A reference frame called bounding boxes is placed in specific sections of an image to recognize a particular trait or object characteristic.
Audio data annotation	Audio data labeling is used in NLP, transcription, etc. Use cases include Alexa or Siri responding to verbal cues in real-time. The underpinning AI/ML models are trained on large, labeled vocal commands datasets, which helps generate suitable responses.
Text data annotation	Unstructured text data can be parsed by AI/ML systems trained with suitable datasets to interpret written language. This enables the machines to classify text residing in images, videos, PDFs, etc., and can surface context within the text. Examples include chatbots and virtual assistants.

Source: Analytics India Magazine, Straive.

Straive provides user interfaces (UIs) to review the outcomes by Subject Matter Experts and Data Stewards to ensure accuracy and validity. In addition, SDP offers a configurable workflow to manage the content from acquisition and ingestion to enrichment and delivery. In areas where we observe high data accuracy and consistency, the platform can also be configured as a robotic process automation (RPA) process. In this case, only data with missing/invalid points (exceptions) are presented to the analyst to enrich for human augmentation if mandated.

We develop ML models for project-specific requirements while using generic ML models like NERs and Summarizers to quickly onboard projects, as needed. The platform delivers accuracy at scale, with classification, detection, segmentation, and annotation tools that enable quick and accurate data labeling for any use case. We design and tailor the required ontology and instructions as per the ESG use case.

High-quality labeled data make smooth operations of AI/ML models possible. Thus, a secure and cost-effective data labeling approach is highly sought. Structuring and classifying data so that AL/ML models can distinguish between the human and the background, the roads and vehicles, etc., provide critical ground truth data to drive reliable ESG predictions.

Legacy information systems or existing web extraction solutions struggle to keep up with the more complex websites and security restrictions such as captcha, making collection sub-optimal. Most information services providers depend on manual data extraction processes to extract relevant data points from text-based documents. These processes are time-consuming and labor-intensive.

Also, continuous monitoring is required to check the currency of data. If an information services company lacks proper data governance and management strategy, it will lead to data silos.

How Straive adds value

Our proprietary Straive Data Platform (SDP) enables data extraction from multiple sources. By leveraging web crawlers and native data connectors, data extraction can be performed on websites, PDFs, blogs, RSS feeds, APIs, etc. Furthermore, SDP's intelligent engines clean, de-duplicate, and normalize the collected data. To create structured datasets, NLP and ML models are used to extract relevant data entities. Finally, Straive's SME team of 500+ associates curates and validates the data to deliver high-quality datasets in structured or semi-structured formats such as CSV, Excel, Flat files, JSON, XML, and API for seamless integration.

Straive's data extraction capabilities rests on seven pillars:

Data Extraction and Enrichment:

Our data extraction and enrichment modules have prebuilt AI and ML algorithms trained for data excerption. Information services providers can leverage these modules to extract data from annual reports, scientific literature, regulatory documents, and more.

Curation:

Straive enables an information services company to deliver enriched and differentiated data sets by leveraging an AI-based auto-curation solution layered with deep domain expertise.

Knowledge Management:

Straive enables information services providers to drive the discoverability and reusability of data by creating and managing taxonomies, ontologies, and graph technologies for knowledge management.

Product Development and Management:

Straive empowers providers to identify and accelerate the market launch of new data products through end-to-end data identification, extraction, enrichment, and cleansing services to deliver differentiated datasets.

Research and Reports:

Straive provides secondary research and report writing services such as industry and researcl reports, people profiles, competitive research, and market trends analysis.

Data Processing Audit:

Straive allows deep forensic analysis of the data processing supply chain – people, process, and technology -- to identify gaps and recommend improvements.

Our Solutions:

Straive provides information providers, across industry segments like Research, Education, BFSI, CPG, Retail, Logistics, Information Services, and Emerging Markets, with scalable data solutions across the data lifecycle to help uncover data intelligence and analytics from their data assets. Straive's end-to-end data solution enables enterprises to perform data extraction and transform the extracted data into insights. It involves implementing data governance practices and strategies for extracting, enriching, storing, transforming, processing, retrieving, using, or making information available.

Straive helps organizations with highly scalable data solutions across the data lifecycle, allowing them to uncover data intelligence and analytics from unstructured and structured data assets. This end-to-end data solution aids an organization's journey from data to intelligence. It involves implementing data governance practices and strategies right from data acquisition, extraction, enrichment, and transformation to ingestion and consumption.

About Straive

Straive is a market-leading content technology enterprise that provides data services, subject matter expertise (SME), and technology solutions to multiple domains, such as research content, eLearning/EdTech, and data/information providers. With a client base scoping 30 countries worldwide, Straive's multi-geographical resource pool is strategically located in seven countries - the Philippines, India, the United States, Nicaragua, Vietnam, the United Kingdom, and Singapore, where the company is headquartered.

www.straive.com
straiveteam@straive.com

https://www.pragnakalp.com/data-extraction-from-tabular-data-with-chatgpt/