

MICRO-DATABASE FOR SUSTAINABILITY  
(ESG) INDICATORS DEVELOPED  
AT THE BANCO DE ESPAÑA

2022

BANCO DE **ESPAÑA**  
Eurosistema

Notas Estadísticas  
N.º 17

Borja Fernández-Rosillo San Isidro,  
Eugenia Koblents Lapteva  
and Alejandro Morales Fernández

## CONTENTS

<b>Abstract</b>	4	<b>4 Conclusions</b>	43
<b>Resumen</b>	5	Next steps	43
<b>ANALYSING CLIMATE CHANGE DATA GAPS</b>	7	<b>References</b>	45
<b>1 Introduction</b>	7	Statistical notes published	46
<b>2 Researching and establishing the relevant indicators</b>	8		
<b>3 Extracting ESG information – transformation from unstructured to structured form</b>	10		
<b>4 Data gaps and limitations in current ESG reporting</b>	10		
<b>5 Improving data quality: quality control of sustainability indicators</b>	18		
<b>6 Conclusions</b>	20		
<b>Next steps</b>	21		
<b>References</b>	22		
<b>ANNEX 1 Selection of indicators</b>	23		
<b>ANNEX 2 Dictionary of words associated with each indicator</b>	24		
<b>A WEB APPLICATION PROTOTYPE FOR INFORMATION RETRIEVAL AND STORAGE</b>	27		
<b>1 Introduction</b>	27		
Target information	28		
<b>2 Technical approach</b>	32		
Description of input data	34		
Digital and scanned text extraction	35		
Full-text indexing and search	37		
Data storage	38		
User interface	40		
<b>3 First data ingestion</b>	42		

**MICRO-DATABASE FOR SUSTAINABILITY (ESG) INDICATORS  
DEVELOPED AT THE BANCO DE ESPAÑA**

## ABSTRACT

In recent years, awareness of social and environmental issues has been on the rise. As a result, the demand for sustainability data has grown exponentially. This has driven the Banco de España's Statistics Department to develop a micro-database for sustainability indicators (ESG).

This document presents two papers which analyse the process developed to gather this information and the numerous limitations and difficulties found along the way when dealing with sustainability microdata. Specifically, the two topics covered by these papers are:

- 1 **“Analysing climate change data gaps”** (presented at the 11th Biennial IFC (Irving Fisher Committee) Conference on “Post-pandemic landscape for central bank statistics” held on 25-27 August 2022, Session 3.B “Environmental statistics”)
- 2 **“Creation of a structured sustainability database from company reports: A web application prototype for information retrieval and storage”** (presented at the IFC Bank of Italy workshop on “Data science in central banking” held on 14-17 February 2022, Session 4.3 “Text Mining and ML utilized in Economic Research”) (Koblents and Morales (2022))

The first paper focuses on the various limitations encountered and achievements made in the process of developing a micro-database for sustainability indicators for non-financial companies. After carefully researching current ESG standards, consulting ESG experts, analysing regulatory obligations and conducting practical research, a list of the 39 most relevant ESG indicators was selected from those normally reported by companies. Currently more than 15,000 data samples have been gathered for the period 2019-2020 using a semi-automatic search application developed in-house (presented in detail in the second paper). Numerous challenges were identified during the process, such as the use of different metrics for reporting sustainability information, a lack of information and of a downloadable digital format, comparability difficulties and regulatory restrictions.

The second paper focuses on the tool developed to create the micro-database presented in the first paper. This web application aims, through semi-automatic extraction and storage, to retrieve sustainability indicators from annual non-financial statements reported by Spanish non-financial corporations. This application aims to make it easier for users to search for sustainability indicators in large document databases and store them in a structured database. The tool developed incorporates a set of pre-defined search terms for each indicator which have been selected based on domain knowledge and artificial intelligence in subsequent developments. For each company and indicator, the tool suggests the most relevant text snippets to the user, who then identifies the correct value of the indicator and stores it in the database using the web's user interface. This tool was created by two data scientists in three months, with the continuous support of a team of experts that helped to define the system specifications, propose refinements, collect input data and validate and test the tool. This paper describes the technical approach and the main modules of the implemented prototype, which include text extraction, indexing and search, data storage and visualization.

**Keywords:** ESG (Environmental, Social, Governance), data gaps, standards, climate change, sustainability, databases, metrics, digitised, regulation, full-text search, OCR, databases, web application, Python.

**JEL classification:** Q5 (environmental economics).

## RESUMEN

En los últimos años, la preocupación por los temas sociales y medioambientales ha ido en aumento y, en consecuencia, la demanda de datos sobre sostenibilidad se ha incrementado exponencialmente. Por esta razón, se ha desarrollado en el Departamento de Estadística del Banco de España una base de microdatos sobre indicadores de sostenibilidad (ESG).

Este documento presenta dos artículos que analizan el proceso desarrollado para capturar esta información, así como las numerosas limitaciones y dificultades encontradas a lo largo del camino de búsqueda de microdatos sobre sostenibilidad. Concretamente, los dos temas que tratan los artículos son:

- 1 **“Analysing climate change data gaps”** (presentado en la 11th Biennial IFC (Irving Fisher Committee) Conference on “Post-pandemic landscape for central bank statistics” durante los días 25-27 de agosto de 2022 en la sesión 3.B “Environmental statistics”)
- 2 **“Creation of a structured sustainability database from company reports: A web application prototype for information retrieval and storage”** (presentado en el IFC Bank of Italy workshop on “Data science in central banking” los días 14-17 de febrero de 2022 en la sesión 4.3 “Text Mining and ML utilized in Economic Research”) (Koblents and Morales (2022))

El primer artículo se centra en las numerosas limitaciones encontradas y logros conseguidos en el proceso de desarrollo de la base de microdatos sobre indicadores de sostenibilidad para sociedades no financieras. Tras analizar detalladamente los estándares actuales de información ESG, consultar a expertos en la materia, analizar las obligaciones regulatorias y llevar a cabo un ejercicio práctico de búsqueda de esta información, se seleccionó una lista de los 39 indicadores más relevantes para comenzar la búsqueda. Actualmente se han recopilado más de 15.000 datos correspondientes al período 2019-2020 utilizando una herramienta semiautomática de búsqueda de información desarrollada internamente (presentado en detalle en el segundo artículo). Durante el proyecto se identificaron numerosas dificultades tales como el uso de diferentes métricas al reportar los indicadores, falta de información y de soporte digital para la descarga, así como dificultades de comparabilidad y restricciones regulatorias.

El segundo artículo se centra en la herramienta desarrollada para crear la base de microdatos presentada en el primer artículo. Esta aplicación web tiene como objetivo, mediante la extracción y almacenamiento semiautomático, obtener los indicadores de sostenibilidad de los estados no financieros anuales presentados por las sociedades no financieras españolas. El objetivo de la aplicación es facilitar a los usuarios el trabajo de búsqueda de indicadores de sostenibilidad en múltiples documentos y su almacenamiento en una base de datos estructurada. La herramienta desarrollada incorpora un conjunto de términos de búsqueda predefinidos para cada indicador que han sido seleccionados en base a conocimiento experto e inteligencia artificial en desarrollos posteriores. Para cada empresa e indicador, la herramienta sugiere los fragmentos de texto más relevantes al usuario, quien a su vez identifica el valor correcto del indicador y lo almacena en la base de datos utilizando la interfaz web de usuario. Esta herramienta ha sido creada por dos científicos de datos en tres meses, con el apoyo continuo de un equipo de expertos que ha contribuido a la definición de requisitos y propuestas de mejora, la recopilación de datos, así como la validación y prueba de la herramienta. A lo largo del artículo, se realiza una descripción del enfoque técnico y los principales módulos del prototipo implementado, incluyendo la extracción de texto, indexación y búsqueda, almacenamiento de datos y visualización.

**Palabras clave:** ASG (Ambientales, Sociales y Gobernanza), data gaps, estándares, cambio climático, sostenibilidad, bases de datos, métricas, digitalizado, regulación, búsqueda de texto completo, OCR, aplicación web, Python.

**Códigos JEL:** Q5 (economía ambiental).



## ANALYSING CLIMATE CHANGE DATA GAPS

### 1 Introduction

In recent years, awareness of social and environmental issues has been on the rise. As a result, the demand for sustainability data has grown exponentially. Although progress has been made in regulation and ESG reporting, and despite today's numerous ESG requirements for companies, there is still a long way to go in terms of data availability, homogeneity, robustness and reliability. Although rich, the granular information available is still insufficient since:

**It does not cover the entire population** of companies, mainly because (i) current regulations exclude small and medium-sized enterprises from mandatory reporting requirements and (ii) companies belonging to a group are not required to report ESG indicators if the parent company of the group reports this information.

**It is not homogeneous** as there is no single definition of the indicators to be reported and their units. Each company bases its reporting on a series of different standards with numerous ways of calculating and measuring the ESG indicators, making comparison for the purposes of analysis difficult. Moreover, as a result of the first limitation, there is a mix of consolidated data (i.e. from companies that present their emissions for the entire group) and individual data (i.e. from companies that report their emissions directly). However, the problem of the lack of homogeneity of the indicators and metrics being reported will mostly be solved by the current regulatory developments – namely the Corporate Sustainability Reporting Directive (CSRD) and the European Sustainability Reporting Standards (ESRS) that are being developed at European level by the European Financial Reporting Advisory Group (EFRAG) and the IFRS Sustainability Disclosure Standard that is being developed by the International Sustainability Standards Board (ISSB) and the Value Reporting Foundation's Sustainability Accounting Standards Board (SASB).

**It is not digitised**, making it difficult to download data for processing. The ESG information is still not reported in a digital reporting language such as XBRL (Extensible Business Reporting Language). However, in the near term it is expected to be digitised like the financial statements.

Taking into account the limitations mentioned above and bearing in mind the need to assess the impact of economic policy measures on climate change, as well as the

role of the financial system in channelling investments towards environmentally sustainable activities, the Banco de España is interested in using the information available to generate ESG statistics. Consequently, and in line with the institution's objectives, the Statistics Department is working on the development of a micro-database for sustainability indicators for non-financial companies, the technical part of which has already been presented at the Irving Fisher Committee (IFC).<sup>1</sup> It is also actively participating in three working groups relating to climate change.<sup>2</sup>

The aim of this paper is to present the process developed at the Central Balance Sheet Data Office for obtaining, processing and analysing the ESG indicators of non-financial corporations.

## 2 Researching and establishing the relevant indicators

Before determining what indicators to look for, a rigorous analysis of the different aspects that needed to be taken into consideration was conducted, as follows:

- 1 The current regulatory obligations were analysed.** Here it was detected that, under the Non-Financial Reporting Directive (Directive 2014/95/EU, transposed into Spanish law by Law 11/2018), companies were required to present a report containing non-financial information. However, the precise indicators and the format for reporting them were not specified. Additionally, they were not reported digitally and they did not follow a structured form.
- 2 Research was carried out on the national and international ESG standards.** In order to better understand the current indicators reported by companies and the technical aspects of their preparation different ESG standards were analysed. We focused on the technical papers of the Global Reporting Initiative (GRI) since most Spanish companies report in line with this standard.
- 3 A preliminary list of 120 indicators was established.** First, lists of indicators were consulted, such as the one compiled by the Spanish Association of Accounting and Business Administration (AECA) and the one compiled by the private company INFORMA. In conjunction with the indicators reflected in the GRI standards, these consultations enabled a

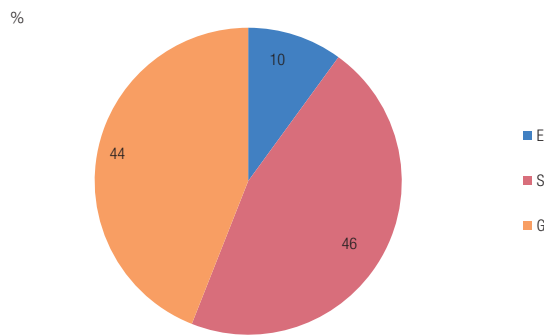
---

1 For further information regarding this web application prototype, see Koblents and Morales (2022), presented at the IFC.

2 Working groups: (1) Workstream on exposure of financial institutions to climate-related physical risks from the STC Expert Group on Climate Change and Statistics / (2) Workstream on climate-related data gaps from the FSC Task Force on Climate Impact Assessment Analytics / (3) Workstream on ESG non-financial information reporting from the XBRL Spanish association.

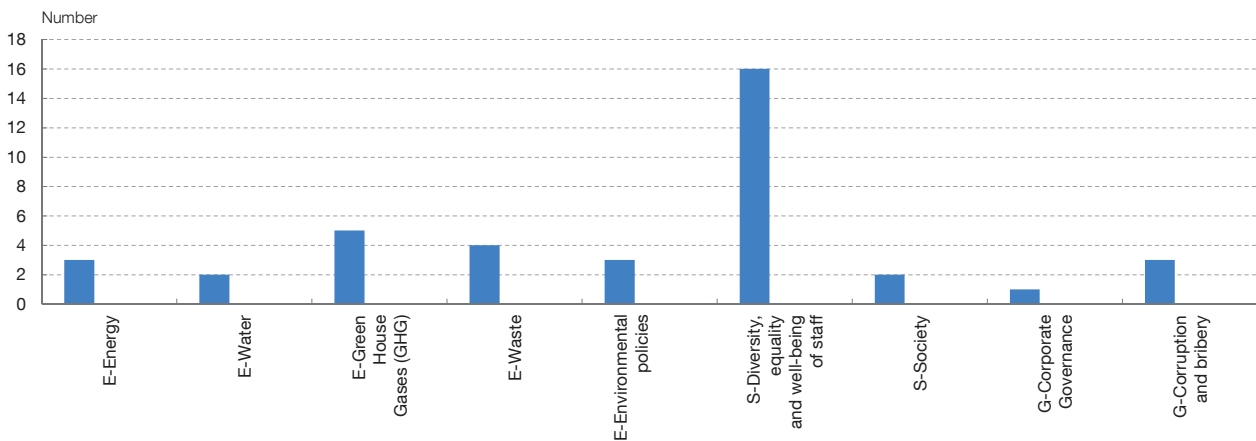


Chart 1  
**DISTRIBUTION OF THE 39 SELECTED INDICATORS BY TYPE (ESG)**



SOURCE: Devised by authors.

Chart 2  
**DISTRIBUTION OF THE 39 SELECTED INDICATORS BY SUBTYPE (ENERGY, WATER, ETC.)**



SOURCE: Devised by authors.

preliminary list of 120 indicators to be drawn up from which to select the most feasible and interesting ones.

- 4 A practical research exercise involving six listed companies was conducted.** In order to select the indicators with the highest probability of being found and of greatest interest a practical exercise was performed by searching for the 120 indicators on the preliminary list in the non-financial reports of six listed companies published in the year 2019. Additionally, a survey of three experts was conducted (one from the rating and risk sustainability field, another from the sustainable finance field and the last one from the accounting and ESG reporting field) who categorised each indicator into one of three categories (low, medium and high),

according to its potential interest. The result of this exercise, taking into consideration various aspects of each indicator (difficulty, interest, feasibility, etc.), was a list of 39 ESG indicators to search for (see Table A1.1 in Annex 1). Charts 1 and 2 depict the distribution of the 39 selected indicators by type (environmental, social or governance) and subtype (energy, water, etc.).

### 3 Extracting ESG information – transformation from unstructured to structured form

Once the indicators to look for had been established, it was necessary to address the process of retrieving them from non-financial reports. The main issue was the lack of homogeneity in the indicators reported and the fact that they were not digitised. To solve this issue, and in close collaboration with data scientists of the Statistics Department, a semi-automatic search application was developed in order to obtain these data and transform them, going from an unstructured format to a structured database.<sup>3</sup>

To make it easier to gather this information, the search process was based on a set of pre-defined terms for each sustainability indicator. The dictionary of ESG terms associated with each indicator was prepared on the basis of the above-mentioned practical exercise involving six listed companies and the information supplied by the indicator descriptions obtained from the various standards consulted. The result was a preliminary list of words (see Table A2.1 in Annex 2) which would help the web application locate this information.

When the first list of words was compiled, there was no previous experience of this process of extracting ESG indicators from non-financial reports. Having now gathered roughly 15,000 indicators and saved the search terms for each indicator, a success rate has been determined for each word list, making it possible to improve the current ontology. Natural Language Processing (NLP) and Artificial Intelligence (AI) tools enable the application to learn from current searches and automatically improve the ontology in order to maximise the success rate.

### 4 Data gaps and limitations in current ESG reporting

From the experience of this project there is enough evidence to confirm that most companies in the sample analysed are aware of ESG risks and reporting. However, a wide range of limitations were found during the process.

---

<sup>3</sup> For further information regarding this web application prototype, see Koblents, Eugenia, and Alejandro Morales (2022). "Creation of a structured sustainability database from company reports: A web application prototype for information retrieval and storage" presented at the IFC workshop on "Data science in central banking."

Table 1

**“ENERGY CONSUMPTION WITHIN THE ORGANISATION” INDICATOR FOR TELEFONICA (SPANISH TELECOMMUNICATIONS COMPANY) IN MWh**

ENERGÍA	2015	2016	2017	2018	2019
Consumo total de Energía (MWh)	7.031.436	6.865.919	6.901.216	6.991.253	6.958.516
Electricidad (Mwh)	6.612.778	6.391.248	6.461.695	6.543.895	6.574.002

ENERGY	2015	2016	2017	2018	2019
Total energy consumption (MWh)	7,031,436	6,865,919	6,901,216	6,991,253	6,958,516
Electricity (Mwh)	6,612,778	6,391,248	6,461,695	6,543,895	6,574,002

SOURCE: Telefonica sustainability report (2019).

Table 2

**“ENERGY CONSUMPTION WITHIN THE ORGANISATION” INDICATOR FOR ENDESA (SPANISH ENERGY COMPANY) IN TJ**

INTERNAL ENERGY CONSUMPTION BY PRIMARY SOURCE (TJ)*			
Fuel type	2017	2018	2019
Coal	244,764	221,079	81,527
Fuel oil	58,205	53,313	47,755
Diesel oil	33,357	34,590	34,457
Natural gas	6,768	51.160,000	64,932
Uranium	280,139	254,926	279,042
<b>Total ENDESA consumption</b>	<b>684,142</b>	<b>615,336</b>	<b>507,614</b>

CONSUMO ENERGÉTICO INTERNO POR FUENTE PRIMARIA (TJ)*			
Tipo de combustible	2017	2018	2019
Carbón	244.764	221.079	81.527
Fuel óleo	58.205	53.313	47.755
Gasóleo	33.357	34.59	34.457
Gas natural	67.676	51.160	64.932
Uranio	280.139	254.926	279.042
<b>Total consumo ENDESA</b>	<b>684.142</b>	<b>615.336</b>	<b>507.614</b>

SOURCE: Endesa sustainability report (2019).

**Limitation 1 – Different metrics and units.** During the process of extracting ESG information we found a wide variety of units for some indicators which were, at first, a clear barrier to direct comparison. However, in most cases it was possible to perform a simple transformation of the indicator into the homogeneous unit defined. Some examples for specific indicators are shown in Tables 1, 2 and 3.

Table 3

**“ENERGY CONSUMPTION WITHIN THE ORGANISATION” INDICATOR FOR FERROVIAL (SPANISH CONSTRUCTION AND TRANSPORT COMPANY) IN GJ**

<b>302-1 CONSUMO ENERGÉTICO DENTRO DE LA ORGANIZACIÓN</b>				
		<b>2017*</b>	<b>2018*</b>	<b>2019</b>
Combustibles utilizados en fuentes Estacionarias y Móviles (total) (GJ)	Diésel	6.058.020	5.167.428	4.532.451
	Fuel	78.994	98.703	157.533
	Gasolina	472.599	289.117	586.315
	Gas Natural	3.039.568	260.542	304.364
	Carbón	390.225	570.558	361.701
	Queroseno	21.189	20.221	24.938
	Propano	18.467	27.732	22.793
	LPG	11.540	6.600	6.856
	<b>TOTAL</b>	<b>10.090.602</b>	<b>6.440.901</b>	<b>5.996.951</b>

<b>302-1 ENERGY CONSUMPTION WITHIN THE ORGANISATION</b>				
		<b>2017*</b>	<b>2018*</b>	<b>2019</b>
Fuel used in Stationary and Mobile sources (total) (GJ)	Diesel	6,058,020	5,167,428	4,532,451
	Oil	78,994	98,703	157,533
	Gasoline	472,599	289,117	586,315
	Natural gas	3,039,568	260,542	304,364
	Coal	390,225	570,558	361,701
	Kerosene	21,189	20,221	24,938
	Propane	18,467	27,732	22,793
	LPG	11,540	6,600	6,856
	<b>TOTAL</b>	<b>10,090,602</b>	<b>6,440,901</b>	<b>5,996,951</b>

SOURCE: Ferrovial sustainability report (2019).

It is clear that companies use different units to present this information. However, in these specific cases, where it was possible to convert all the indicators to the same unit, a preliminary analysis was carried out to see which unit was the most common (MWh in the case of “energy consumption within the organisation”). All the data stored in the database were then converted into that unit in order to facilitate comparison and economic analysis. Hence, the different units do not prevent analysis of this type of information, but they do make the process longer and more difficult. Hopefully, future regulatory standards will specify a common unit in order to minimise this limitation.

In this second list of examples relating to the “GHG emissions intensity” indicator (see Tables 4 and 5 and Chart 3), companies presented this ratio in many different ways, as the current definition allows companies to use the business parameters they consider most accurate (for example, companies use as the denominator of the ratio their total output, total sales, total employees, square metres of production plants, total km travelled, audio-visual production hours, etc.). Although some specific cases may be transformed (e.g. to use total employees or total sales), the numerator posed further difficulties as various options could also be used (e.g. Scope 1, Scope 2 or Scope 1+2, among others). Consequently, in cases like this, the

Table 4

“GHG EMISSIONS INTENSITY” INDICATOR FOR EUSKALTEL (SPANISH TELECOMMUNICATIONS COMPANY) IN KgCO<sub>2</sub>E/Output

Intensidad de las emisiones GEI	CO <sub>2</sub>	7,76	7,15	Kg CO <sub>2</sub> e./prod.
GHG emissions intensity	CO <sub>2</sub>	7.76	7.16	Kg CO <sub>2</sub> e/output

SOURCE: Euskaltel sustainability report (2019).

Table 5

“GHG EMISSIONS INTENSITY” INDICATOR FOR GRIFOLS (SPANISH INDUSTRIAL COMPANY) IN TCO<sub>2</sub>/€M

INTENSIDAD DE EMISIONES DE CO <sub>2</sub> e			
T/CO <sub>2</sub> e/millones de euros	2019	2018	2017
Total Grifols	64,8	66,6	69,3

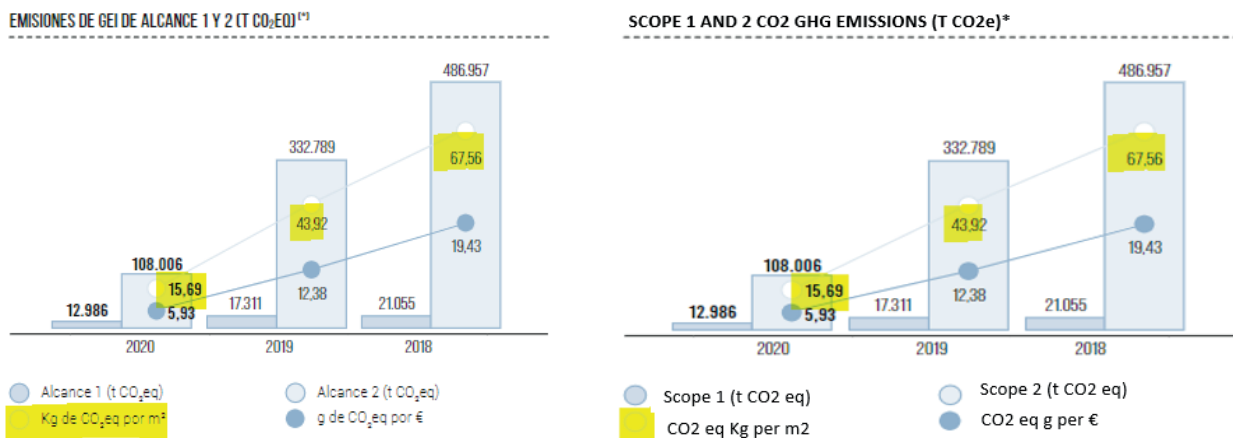
  

ENERGY INTENSITY IN CO <sub>2</sub> e			
TCO <sub>2</sub> e/million euros	2019	2018	2017
Total Grifols	64.8	66.6	69.3

SOURCE: Grifols sustainability report (2019).

Chart 3

“GHG EMISSIONS INTENSITY” INDICATOR FOR INDITEX (SPANISH RETAIL COMPANY) IN KgCO<sub>2</sub>/m<sup>2</sup>



SOURCE: Inditex sustainability report (2019).

Table 6

**“SCOPE 1, 2 AND 3 GHG EMISSIONS” INDICATOR FOR GRUPO EZENTIS SA  
(SPANISH TELECOMMUNICATIONS AND INDUSTRIAL COMPANY) IN TCO<sub>2</sub>E**

Distribución de las emisiones según alcances:

Emisiones <sup>1</sup> (TCO <sub>2eq</sub> )	2018*	2018 (comparable con 2019)	2019
Alcance 1	7.477	36.223	32.761
Alcance 2	949	949	1.159
Alcance 3	29.471	725	532
TOTAL	37.897	37.897	34.453

\*Desglose reportado en EINF 2018

Distribution of emissions according to scope:

Emissions (TCO <sub>2eq</sub> )	2018*	2018 (comparable with 2019)	2019
Scope 1	7,477	36,223	32,761
Scope 2	949	949	1,159
Scope 3	29,471	725	532
TOTAL	37,897	37,897	34,453

\* Reported breakdown in 2018 non-financial statement

SOURCE: Grupo Ezentis sustainability report (2019).

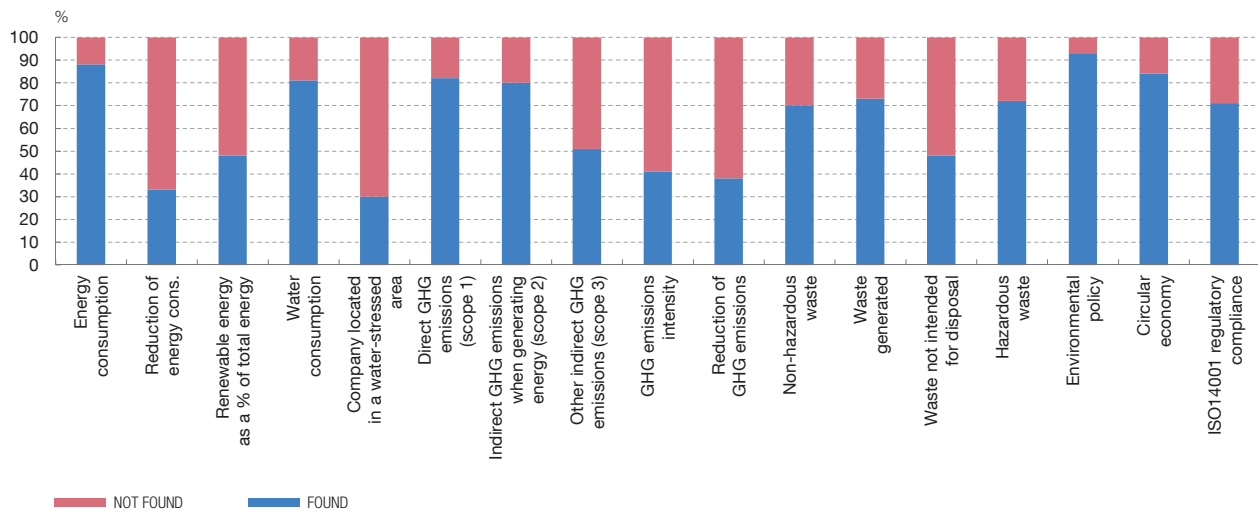
data should only be taken into consideration for qualitative analysis, without attempting to make direct comparisons. In the case of this specific indicator it may be more beneficial to internally devise a ratio that can be calculated for all companies using the same parameters.

**Limitation 2 – Changes in data over time.** During the process it was not uncommon to see data change from one year to the next. There were a number of companies in the sample that changed data compilation criteria and, consequently, the number they had given for a specific indicator one year did not coincide with the number for the same indicator in the subsequent year’s non-financial report. When an explanation was given in the non-financial report, it usually cited a change in the calculation method. These changes are probably due to the novelty of this information and the limited number of years of experience in collecting and preparing it. In these cases, the most up-to-date data were taken to be the relevant ones. Our intention is to closely follow the evolution of this information in order to see if it stabilises over time and if the number of changes falls significantly.

In the example reflected in Table 6 the company presents in its 2019 non-financial report 2018 data that have been recalculated using a different method. This leads to a substantial difference between years for the same indicator. During this initial period of adaptation to ESG reporting, which companies are currently in, such changes in criteria and data will be common. However, after a reasonable time of

Chart 4

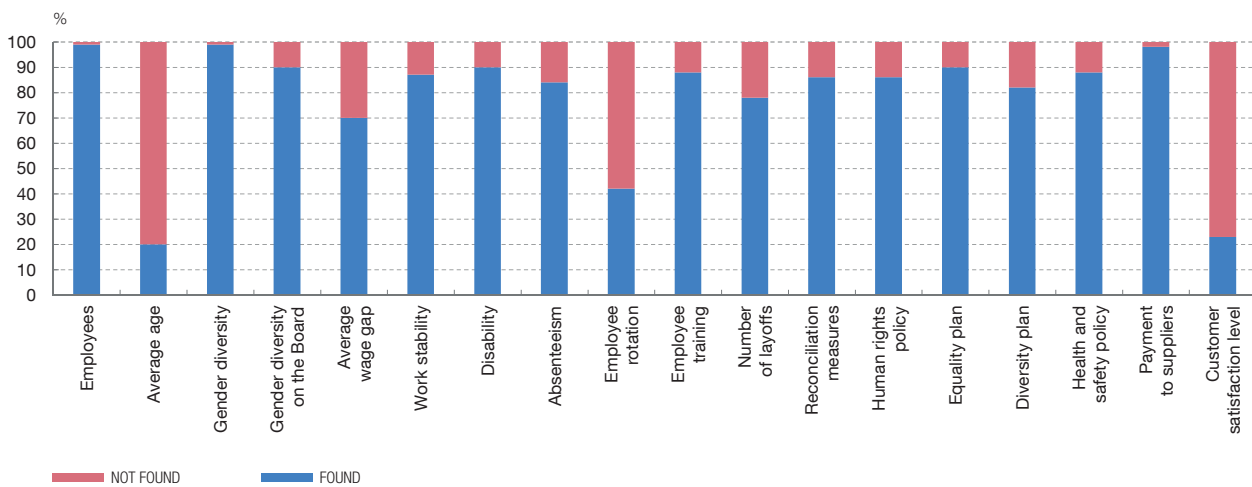
**DISTRIBUTION OF FOUND VS NOT FOUND FOR ENVIRONMENTAL INDICATORS (2019)**



SOURCE: Devised by authors.

Chart 5

**DISTRIBUTION OF FOUND VS NOT FOUND FOR SOCIAL INDICATORS (2019)**



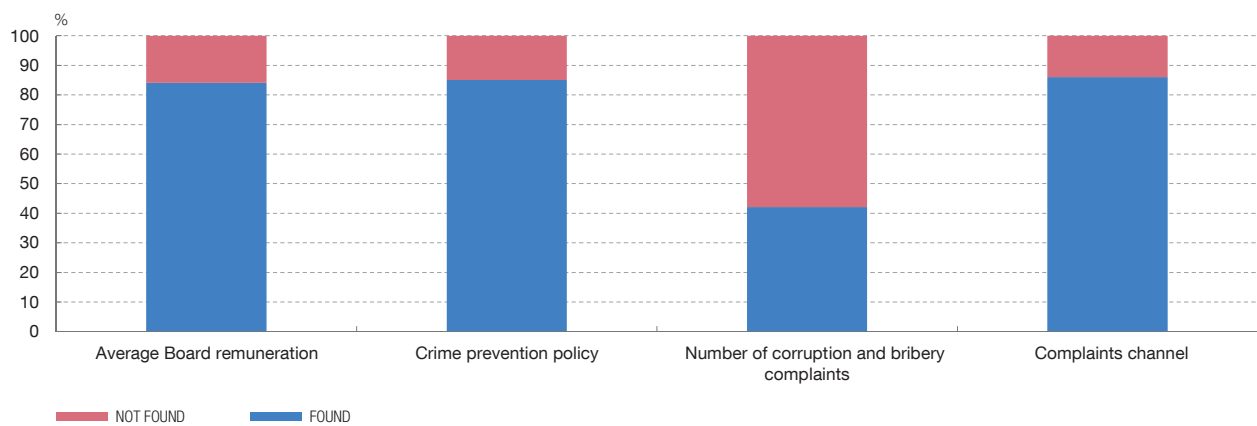
SOURCE: Devised by authors.

compiling and processing non-financial information, these changes in criteria and data can be expected to be fewer and to have less of an impact.

**Limitation 3 – Lack of information.** Some indicators were not found for some companies in the search process. Moreover, it is important to highlight that searches for some of the indicators from the preliminary selection list had a higher success rate than for others (see Charts 4, 5 and 6).

Chart 6

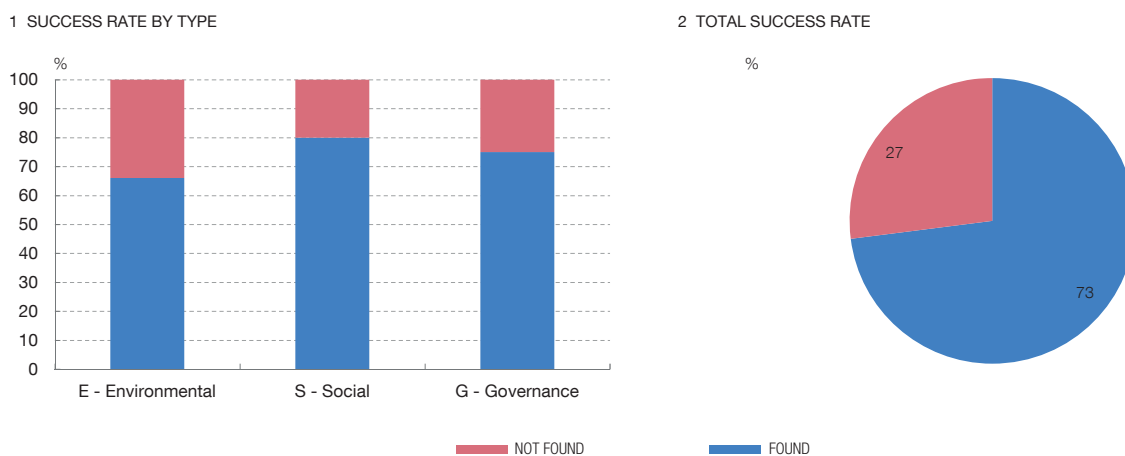
**DISTRIBUTION OF FOUND VS NOT FOUND FOR GOVERNANCE INDICATORS (2019)**



SOURCE: Devised by authors.

Chart 7

**DISTRIBUTION OF FOUND VS NOT FOUND BY TYPE OF INDICATOR (ESG) AND TOTAL (2019)**



SOURCE: Devised by authors.

The following conclusions may be drawn from the analysis presented above:

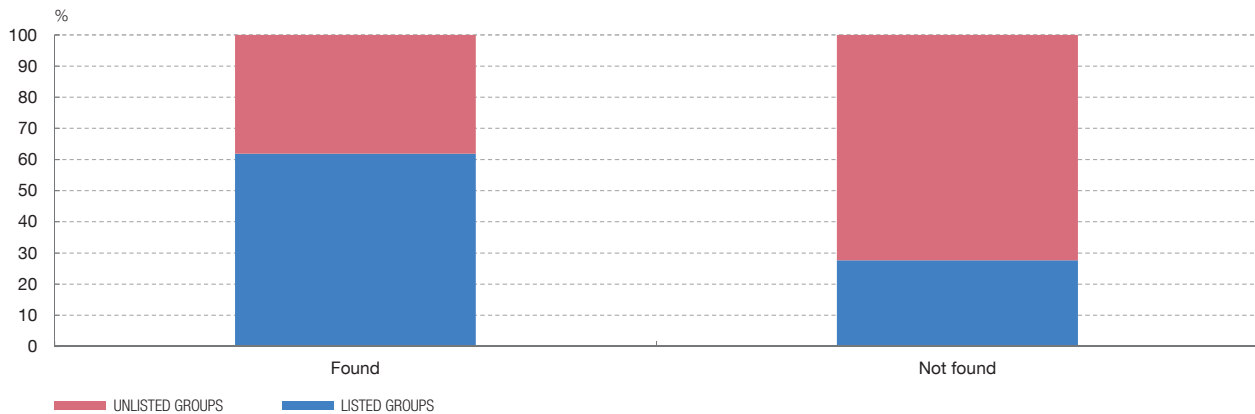
- Social indicators stand out as the type of indicator with the highest success rate (80%), followed by governance (75%) and environmental (66%) (see Chart 7). This result was to be expected since social information (employees, diversity data, disability...) was being reported long before the current boom in environmental information.
- The overall success rate stands at 73%.



Chart 8

**DISTRIBUTION OF FOUND VS NOT FOUND BY TYPE OF COMPANY (LISTED VS UNLISTED) (2019)**

SUCCESS RATE BY TYPE OF GROUP (LISTED VS UNLISTED)



SOURCE: Devised by authors.

- A substantial difference was observed in the success rates for listed companies (81%) vs unlisted companies (50%) (see Chart 8). This is because listed companies have greater public exposure and stakeholder demand than unlisted companies.

**Limitation 4 – Comparability difficulties.** Although the main comparability issue relates to the use of different units and can be solved relatively easily there are still a number of issues that make it difficult to compare this type of data, such as:

**Individual vs consolidated data.** The current regulations exempt individual companies from presenting the non-financial report when the company reports the information at a consolidated level. Consequently, an added difficulty is having to gather data which in some cases are to be found at a consolidated level and in others (less often) at an individual level.

**Global data from international companies.** Currently, many companies (particularly the largest) have an international presence all over the world. This affects the ESG information reported in their non-financial reports as it refers to the performance of the company or group worldwide, hindering the task of measuring national impacts (e.g. Scope 1 emissions in Spain). However, for a very small percentage of companies and indicators, a breakdown is presented by country or economic region.

**Different calculation methodologies.** Calculation methods vary across some indicators and companies. Therefore, although theoretically data should be comparable, there is not enough evidence that such data have been obtained using the same method or based on similar criteria. A case in point are the Scope 2

emissions, which may be reported using a market approach or a location-based approach. This poses no problem when the company reports the method used, but in the majority of the sample analysed the calculation method was not specified in the non-financial report.

**Limitation 5 – Lack of digitisation of the information presented.** The information presented in the non-financial reports is still not digitalised and is reported in a wide variety of formats (e.g. charts, tables, etc.) which makes the process of locating it more difficult. However, the upcoming CSRD (Corporate Sustainability Reporting Directive) provides for the digitisation and standardisation of ESG reporting.

**Limitation 6 – Lack of official ESG verification.** Although some progress has been made in terms of ESG verification there is still no rigorous technical supervision of the information presented. Additionally, even if the information presented can be checked, verification should be stepped up by introducing objective and technical measures that guarantee the quality and veracity of the data presented (i.e. that the emissions a company reports are correctly calculated and real).

## 5 Improving data quality: quality control of sustainability indicators

The heterogeneity of the information and the novelty of having to analyse new types of information (e.g. greenhouse gas emissions and energy consumption) made it extremely important to conduct a rigorous and robust quality control assessment of the data. Consequently, each individual datum was reviewed, compared and contextualised, in terms of activity, in order to guarantee consistency in the database.

Moreover, during the process of analysis of each individual indicator the most common unit was identified to help define the standardised unit to which to convert all the data to facilitate comparison. Additionally, and as we learnt from this new information, an interval analysis was conducted in order to establish the possible maximum and minimum values for each indicator. The possible values of an indicator were thus narrowed down, which improved subsequent processes of searching for this information.

To increase the accuracy and quality of future ESG data uploads, individual guides were prepared for each of the indicators with the following fields:

This first part of the guides (*1-Information of interest on the indicator*) relates to general information on the indicator to improve the process of locating such information. The specific aspects section contains information related to alternative ways of calculating its value or concrete aspects relevant to its interpretation. Additionally, the units sections help to identify whether data being uploaded is consistent in terms of value (see Figure 1).

Figure 1

**EXAMPLE OF THE ASPECTS COVERED IN THE FIRST PART OF THE GUIDES PREPARED FOR EACH INDICATOR (1-INFORMATION OF INTEREST ON THE INDICATOR)**

**1 Information of interest on the indicator**

Guía Indicadores BBDD SOST: Emisiones directas de GEI (alcance 1)  
 Proyecto creación BBDD INF  
 Departamento de Estadística (División de Central de Balances)

Emisiones directas de GEI (alcance 1)	
<b>Indicador</b>	<b>Medioambiente</b>
<b>Tipo</b>	Gases de Efecto Invernadero (GEI)
<b>Subtipo</b>	Tomadas de emisiones de CO2 (TCO2e)
<b>Métrica establecida como homogénea para los datos</b>	Toneladas de emisiones de CO2 (TCO2e)
<b>Posibles métricas a encontrar en la búsqueda de este indicador</b>	1. TnCO2e (Toneladas de CO2 equivalente) 2. Miles TnCO2e 3. MCO2e 4. KgCO2e (Kilogramos de CO2 equivalentes) 5. Kt CO2e
<b>Estándares asociados al indicador</b>	GRI 305-1
<b>Descripción del indicador</b>	
Buscar las emisiones directas de gases de efecto invernadero (GEI o GHG) (Scope 1 o alcance 1) desde fuentes en propiedad o gestionadas por la compañía. La organización informante debe presentar el valor bruto de emisiones directas de GEI (alcance 1) en toneladas métricas de CO2 equivalente, sea tres de las emisiones directas producidas por quema de combustibles por parte del emisor.	
<b>Aspectos específicos a tener en cuenta</b>	
1. Existe la posibilidad de encontrar información por regiones, concretamente distinguiendo Nacional, Unión Europea y Terceros Países. La suma de estas tres categorías nunca debería superar el valor incluido en la categoría Grupo. 2. Existen cambios en los métodos de cálculo o impones que varían de un ejercicio a otro, siempre dejar en todo el dato más actualizado prescindiendo de los valores anteriores (p.e. si tenemos un dato de año 2019 de la memoria de 2019 pero en la memoria de 2020 el dato de 2019 es distinto se debe consignar este último y eliminar el del ejercicio anterior para evitar duplicidades). 3. Consignar siempre los valores con un máximo de dos decimales (ejemplo correcto: 34,64 vs ejemplo incorrecto: 34,646). 4. Siempre consignar un dato que se esté especificando el tipo de emisiones (si nos encontramos "emisiones scope1" en más explicación no se autoriza información como para poder consignar el dato va que desconocemos si hace referencia a Alcance 1, 2 o 3). 5. En los casos en los que se aporte información desglosada por línea de negocio simplemente habrá que sumarlos para obtener el total. 6. A la hora de introducir un dato utilizar como referencia la información histórica, si la hubiera, los rangos posibles del indicador y la métrica de sector. Si se va a meter un dato en una métrica distinta a la acordada como "homogénea" para el indicador tener siempre en mente la coherencia de valor una vez transformado a la métrica homogeneizada.	
<b>Conversiones a métrica homogeneizada</b>	
1 tCO2e = 1.000.000 TnCO2e 1 KgCO2e = 0,001 TnCO2e 1 Mt TnCO2e = 1.000 TnCO2e 1 Kt CO2e = 1.000 TnCO2e	
<b>Ontología asociada al indicador</b>	
- Emisiones directas - scope 1 - Alcance 1 - Toneladas de CO2	- TCO eq - toneladas de CO2 equivalente - GRI 305-1 - TCO

- 1 Name of the indicator
- 2 Type of indicator
- 3 Standardised unit after review (based on that most commonly reported)
- 4 Possible alternative units that can be found for the selected indicator
- 5 Associated ESG standard
- 6 Indicator description
- 7 Specific aspects and precautions that must be taken into account when uploading this information to the database
- 8 Different unit conversions and equivalences to convert to the standardised unit above
- 9 Revision of the ontology, including new words that are considered important and removing those that are irrelevant or that distort the search

SOURCE: Devised by authors.

Figure 2

**EXAMPLE OF THE ASPECTS COVERED IN THE SECOND PART OF THE GUIDES PREPARED FOR EACH INDICATOR (2-ANALYSIS OF THE INDICATOR BY SECTOR)**

**2 Analysis of the indicator by sector**

**Análisis del indicador**

*Disclaimer: debemos tener en cuenta que el indicador no está siempre disponible para la misma empresa lo largo de la serie histórica y, además, para determinados ejercicios todavía no se ha procedido a buscar dicha información. No obstante, estos datos y resultados deben servir como fuente de apoyo y contraste a la hora de meter un valor en la bbdd y poder hacernos una idea de si el valor que vamos a introducir tiene cierta idiosia o no. Hay que destacar que los valores se presentan para los datos de "Relevante país: GRUPO" y con la métrica ya homogeneizada a "TnCO2e".*

A continuación, se presenta la muestra por sectores utilizada para el análisis:

Nº empresas					
Nº Sector	2020	2019	2018	2017	2016
1 ENERGÍA	6	6	6	7	9
2 INDUSTRIA	24	31	25	5	9
3 COMERCIO Y HOSTELERÍA	6	11	7	1	1
4 INFORMACIÓN Y TELECOM.	9	10	9	2	1
5 CONSTRUCCIÓN E INMOB.	6	7	7	1	0
6 TRANSPORTE Y ALMAC.	4	5	2	2	2
7 RESTO	10	14	10	10	0
8 TOTAL	69	87	69	16	6

A continuación, se presentan los valores medios del indicador para la muestra analizada:

Media (TnCO2e)					
Nº Sector	2020	2019	2018	2017	2016
1 ENERGÍA	6.600.928	6.824.465	11.390.897	11.158.234	n.d.
2 INDUSTRIA	1.283.204	1.021.864	1.201.962	3.634.407	9.607.200
3 COMERCIO Y HOSTELERÍA	125.045	52.269	23.898	48.110	47.619
4 INFORMACIÓN Y TELECOM.	25.596	25.790	28.436	148.790	291.797
5 CONSTRUCCIÓN E INMOB.	514.008	536.714	551.968	150.000	n.d.
6 TRANSPORTE Y ALMAC.	3.019.308	5.420.719	6.189.620	14.513.179	14.261.770
7 RESTO	37.891	863.166	1.097.862	601.525	n.d.
8 TOTAL	1.338.571	1.557.027	2.095.310	4.547.820	8.112.824



- 1 A description of the type of information used and its current limitations, to be used as supporting information or as a comparison when new indicators are loaded
- 2 Distribution of the sample used by year and sector (in the case of numerical indicators)
- 3 The average values (in the standardised unit) that will serve as reference. Another table has been included with the maximum values so that the usual ranges can be established for each indicator (in the case of the numerical ones)

SOURCE: Devised by authors.

The second part of the guides (*2-Analysis of the indicator by sector*) relates to sample data for the indicator in question as a reference to determine whether the new data being introduced into the database are consistent and realistic. These tables allow the possible intervals of the indicator to be narrowed down and give the average values by sector and year in order to minimise the potential number of future errors that can be entered into the database (see Figure 2).

The third and last part of the guides (*3-Real examples in non-financial reports*) provides real examples of how this information can appear in the non-financial reports in order to facilitate the location of this information (see Figure 3).

Figure 3

**EXAMPLE OF THE ASPECTS COVERED IN THE THIRD PART OF THE GUIDES PREPARED FOR EACH INDICATOR (3-REAL EXAMPLES FROM NON-FINANCIAL REPORTS)**

1

Various examples of the way this information can be found in the non-financial reports, which serve as reference in the search for this type of information

**3 Real examples from non-financial reports**

**Ejemplos de memorias (Scope 1):**

A continuación, se muestran una serie de ejemplos en referencia al Scope 1 y la manera en la que las empresas lo están mostrando en sus memorias.

**Emissiones CO2e (Toneladas)**

	2018			2019		
	Total	Intensidad	100%	Total	Intensidad	100%
Alcance 1	474.237	1.007 T/m²	49.962	394.974	1.017 T/m²	40.141
Alcance 2	3.909.287	1.073 T/m²	79.974	3.191.284	1.171 T/m²	64.141
Total Alcance 1+2	4.383.524	1.038 T/m²	129.936	3.586.258	1.198 T/m²	104.282

\*Intensidad en base de 2018 de acuerdo a 100% ventas

Emissiones de CO2e (Toneladas)		2018	2019
Alcance 1 - Emisiones directas	474.237	474.237	394.974
Alcance 2 - Emisiones indirectas	3.909.287	3.909.287	3.191.284
Total	4.383.524	4.383.524	3.586.258

En el año 2020, el total de emisiones de alcance 1 (emisiones directamente vinculadas con el core business de la compañía) ha sido de 633 toneladas equivalentes de CO<sub>2</sub> e incluye las emisiones asociadas al gas natural. En el caso de las emisiones de alcance 2 (emisiones indirectas) el resultado ha sido de 24.46 toneladas equivalentes de CO<sub>2</sub> e, incluyendo las emisiones asociadas a la electricidad.

**EMISIONES DE GASES DE EFECTO INVERNADERO (CO2e)**

	2018	2019
Alcance 1	633 t eq CO <sub>2</sub> e	633 t eq CO <sub>2</sub> e
Alcance 2	21.83 t eq CO <sub>2</sub> e	24.46 t eq CO <sub>2</sub> e
Total emisiones Alcance 1+2 en base al beneficio obtenido	654.83 t eq CO <sub>2</sub> e	657.46 t eq CO <sub>2</sub> e

\* Incluye las emisiones de gases de efecto invernadero

**5.2 Cambio climático y energía**

Total de emisiones de CO<sub>2</sub>e (base 1) de la actividad de producción de cemento por país (toneladas)

País	2018	2019	2020	Variación 2020
Francia	1.047.477,0	892.276,3	555.000,0	-49%
Argentina	1.632.000,0	1.308.700,0	1.290.900,0	-1%
Italia	4.345.833,0	3.943.740,0	4.400.000,0	12,4%
Emisiones	386.000,0	372.300,0	320.000,0	-16,4%
Burkina Faso	103.000,0	90.000,0	80.000,0	-22,2%
Brasil	3.348.000,0	346.500,0	3.000.000,0	-10,2%
Italia	446.000,0	466.000,0	400.000,0	-10,3%
Guatemala	90.000,0	90.000,0	90.000,0	0%
Total	9.770.000,0	8.091.200,0	6.776.000,0	-30,8%

**Emisiones**

CO <sub>2</sub> e	Indicador	Definición	2018	2019	2020
Europa*	CR 205-1	Emisiones directas	62.277,93	43.124,42	37.424,42
	CR 205-2	Emisiones indirectas	79.277,33	62.719,05	58.422,00
	CR 205	Total	141.555,26	105.843,47	95.846,42
Norteamérica	CR 205-1	Emisiones directas	22.424,82	27.242,34	22.342,49
	CR 205-2	Emisiones indirectas	79.842,08	76.247,48	82.282,01
	CR 205	Total	102.266,90	103.489,82	104.624,50
Brasil	CR 205-1	Emisiones directas	10.121,20	14.287,18	8.911,82
	CR 205-2	Emisiones indirectas	1.706,83	1.438,41	477,23
	CR 205	Total	11.828,03	15.725,59	9.389,05
Asia (India/China)	CR 205-1	Emisiones directas	10.714,47	10.294,47	10.912,44
	CR 205-2	Emisiones indirectas	206.151,11	214.289,24	200.124,89
	CR 205	Total	216.865,58	224.583,71	211.037,33
TOTAL	CR 205-1	Emisiones directas	61.702,22	68.094,39	60.639,13
	CR 205-2	Emisiones indirectas	365.179,87	354.848,08	324.414,97
	CR 205	Total	426.882,09	422.942,47	385.054,10

SOURCE: Devised by authors.

**6 Conclusions**

Sustainability reporting is becoming increasingly relevant to measure companies impact on climate change. The numerous demands from the economy's stakeholders and the new green investment and financing decisions based on this information are

key to the exponential growth in ESG reporting. However, despite the current progress there is still a long way to go as there are still data gaps, heterogeneity in reporting, a lack of digitisation and verification and regulatory limitations.

The experience gained in gathering ESG indicators from Spanish non-financial corporations' reports shows that listed companies present more and richer information than unlisted companies. Moreover, the ability to measure national ESG impacts is still a huge issue as multinational companies present information for the whole group and local companies either do not present ESG information, as they are exempted from doing so by the current regulations, or they belong to a group which presents consolidated data.

All in all, this project, in combination with the regulations currently in progress (the CSRD) and the increasing awareness of the importance of reporting by companies, means that statisticians can be optimistic about ESG data availability in the near future.

### Next steps

The achievements made and the experience obtained so far have provided the necessary catalyst to continue working on the compilation of a robust ESG micro-database that can meet the various requirements of stakeholders and help minimise current ESG data gaps. Consequently, and bearing in mind the long-term goal of this project, the following lines of work are in the pipeline:

- Increasing the sample of companies from which information is currently extracted.
- Developing robust and solid quality control tests to minimise inconsistencies and errors in the data (automatic quality checks for future information uploads).
- Adapting the micro-database to the upcoming new ESG regulation.
- Analysing the possibility of increasing the number of indicators that can be searched for.
- Refining the search process to increase the rate of success in terms of location and automation.

## REFERENCES

ESG standards developed internally at INFORMA [<https://www.informa.es/>]

GRI (Global Reporting Initiative) standards [<https://www.globalreporting.org/>]

Koblents, Eugenia, and Alejandro Morales. (2022). "Creation of a structured sustainability database from company reports: A web application prototype for information retrieval and storage". IFC workshop on "Data science in central banking".

Moreno, Ángel Iván, and Teresa Caminero. (2020). "Application of Text Mining to the Analysis of Climate-Related Disclosures". International Conference on "Statistics for Sustainable Finance". Irving Fisher Committee on Central Bank Statistics, 14-15 September 2021.

Spanish Association of Accounting and Business Administration (AECA) "Modelo AECA de información integrada para la elaboración del Estado de Información no Financiera - Cuadro integrado de indicadores (CII - FESG)" [<https://is.aeca.es/en/integrated-scoreboard-documents/>]

## ANNEX 1 Selection of indicators

### Selection of indicators for collecting sustainability information (39)

Table A1.1

#### SUSTAINABILITY INFORMATION IN NON-FINANCIAL CORPORATIONS

Type (E,S,G)	Subtype	Indicator	
E	Energy	Energy consumption within the organization	
		Reduction of energy consumption	
		Renewable energy as a % of total energy	
	Water	Water consumption	
		Company located in a water-stressed area	
	Green House Gases (GHG)	Direct GHG emissions (scope 1)	
		Indirect GHG emissions when generating energy (scope 2)	
		Other indirect GHG emissions (scope 3)	
		GHG emissions intensity	
		Reduction of GHG emissions	
	Waste	Non-hazardous waste	
		Waste generated	
		Waste not intended for disposal	
		Hazardous waste	
	Environmental policies	Environmental policy	
		Circular economy	
		ISO14001 regulatory compliance	
	S	Diversity, equality and well-being of staff	Employees
			Average age of employees
Gender diversity			
Gender diversity on the Board			
Average wage gap			
Work stability			
Disability			
Absenteeism			
Employee rotation			
Employee training			
Number of layoffs			
Reconciliation measures			
Human rights policy			
Equality plan			
Diversity plan			
Health and safety policy			
Society		Payments to suppliers	
		Customer satisfaction level	
G		Corporate Governance	Average Board remuneration
	Corruption and bribery	Crime prevention policy	
		Number of corruption and bribery complaints	
		Complaints channel	

SOURCE: Banco de España.

## ANNEX 2 Dictionary of words associated with each indicator

Table A2.1

### DICTIONARY OF WORDS ASSOCIATED WITH EACH INDICATOR

Type (E,S,G)	Indicator	Dictionary of words (original queries in Spanish)	Dictionary of words (English)
E	Energy consumption within the organization	Consumo de energía energético eléctrico electricidad dentro de la organización KWh, MWh, GWh, TWh, kJ, MJ, GJ, TJ, PJ «302-1» dato total vatios julios	Consumption of energy electrical energy electricity within the organization KWh, MWh, GWh, TWh, kJ, MJ, GJ, TJ, PJ “302-1” total figure watts joules
E	Reduction of energy consumption	Reducción del consumo energético Energía KWh, MWh, GWh, TWh, kJ, MJ, GJ, TJ, PJ «302-4» dato total evitar descender julios vatios	Reduction of energy consumption Energy KWh, MWh, GWh, TWh, kJ, MJ, GJ, TJ, PJ “302-4” total figure avoid falling joules watts
E	renewable energy as a % of total energy	porcentaje energía renovable total % ratio fuentes TJ consumo energético combustible fósil verde KWh GWh origen «menores emisiones» alternativa eólica solar 302 «no convencionales» KWh, MWh, GWh, TWh, kJ, MJ, GJ, TJ, PJ	percentage of total renewable energy % ratio sources TJ energy consumption green fossil fuel KWh GWh source “lower emissions” solar wind alternative 302 “non-conventional” kWh, MWh, GWh, TWh, kJ, MJ, GJ, TJ, PJ
E	Water consumption	«Consumo de agua» «uso de agua» «gasto de agua» «303-5» volumen caudal total dato hm m3 Hm3 megalitros litros millones toneladas masa captación dulce	“Water consumption” “water usage” “use of water” “303-5” volume total flow figure hm m3 hm3 megalitres litres million tonnes fresh catchment mass
E	Company located in a water-stressed area	Estrés hídrico localización zona área agua escasez recursos hídricos H20 303	Water stress location area water scarcity water resources H20 303
E	Direct GHG emissions (scope 1)	Emisiones directas de GEI «alcance 1» «scope 1» «ambito 1» «Gases de Efecto Invernadero» Mt Tn Tm CO2* «305-1» miles millones toneladas dato total kt calcul*	Direct GHG emissions “scope 1” “scope 1” “scope 1” “Greenhouse Gases” Mt Tn Tm CO2* “305-1” billion tonnes total data kt calcul*
E	Indirect GHG emissions when generating energy (scope 2)	Emisiones indirectas de GEI genera* energía energético «alcance 2» «scope 2» «ambito 2» «Gases de Efecto Invernadero» Mt Tn Tm CO2* «305-2» miles millones	Indirect GHG emissions generat* energy “scope 2” “scope 2” “scope 2” “Greenhouse Gases” Mt Tn Tm CO2* “305-2” billions
E	Other indirect GHG emissions (scope 3)	Otras emisiones indirectas de GEI genera* energía energético «alcance 3» «scope 3» «ambito 3» «Gases de Efecto Invernadero» Mt Tn Tm CO2* «305-3» miles millones toneladas eléctrico electricidad dato total kt calcul* TCO eq equivalente tCO* dióxido de carbono GHG	Other indirect GHG emissions generat* energy “scope 3” “scope 3” “scope 3” “Greenhouse Gases” Mt Tn Tm CO2* “305-3” billion tonnes electric electricity total figure kt calcul* TCO eq equivalent tCO* carbon dioxide GHG
E	GHG emissions intensity	Intensidad de las emisiones de GEI «Gases de Efecto Invernadero» Ratio «305-4» dato tCO* CO2* kg kilogramo carbono TJ intensidad energética MtCO2e GHG	Intensity of GHG emissions “Greenhouse Gases” Ratio “305-4” data tCO* CO2* kg kilogram carbon TJ energy intensity MtCO2e GHG
E	Reduction of GHG emissions	Reducción de las emisiones de GEI objetivo evitar «Gases de Efecto Invernadero» Mt Tn CO2* «305-5» miles millones toneladas dato TCO* eq equivalente GHG	Reduction of GHG emissions objective to avoid “Greenhouse Gases” Mt Tn CO2* “305-5” billions of tonnes TCO data* GHG equivalent eq
E	Non-hazardous waste	Residuos generados «no peligrosos» reutilización reciclaje recuperación incineración compostaje toneladas Tn «306-2»	Waste generated “non-hazardous” reuse recycling recovery incineration composting tonnes Tn “306-2”
E	Waste generated	Residuos generados peso total toneladas Tn «306-3»	Waste generated total weight tonnes Tn “306-3”
E	Waste not intended for disposal	Residuos generados no destinados a eliminación Tn Toneladas composición «306-4»	Waste generated not intended for disposal Tn Tonnes composition “306-4”
E	Hazardous waste	Residuos generados peligrosos Tn Toneladas 306	Hazardous waste generated Tn Tonnes 306
E	Environmental policy	Medio Ambiente Política medioambiental Futuro limpio de emisiones Estrategia climática menor impacto productos bajos 30X	Environment Environmental policy Clean future of emissions Climate strategy less impact low products 30X

SOURCE: Banco de España.



Table A2.1

**SUSTAINABILITY INFORMATION IN NON-FINANCIAL CORPORATIONS (cont'd.)**

Type (E,S,G)	Indicator	Dictionary of words (original queries in Spanish)	Dictionary of words (English)
E	Circular economy	Proyectos de "economía circular" Reducción del uso de "materias primas" Ecología Circularidad Ecodesign alternativas Eficiencia en procesos Innovación	"Circular economy" projects Reduction in the use of "raw materials" Ecology Circularity Ecodesign alternatives Efficiency in processes Innovation
E	ISO14001 regulatory compliance	ISO 14001 cumplimiento normativa Estándar en gestión ambiental Responsabilidades medioambientales reducir impacto corporativa Requerimientos regulatorios	ISO 14001 regulatory compliance Environmental management standard Environmental responsibilities reduce corporate impact Regulatory requirements
S	Employees	Número de empleados Mujeres Hombres Distribución plantilla Contrato laboral Fijos Temporales "102-8"	Number of employees Women Men Workforce distribution Employment contract Permanent Temporary "102-8"
S	Average age of employees	edad media de la plantilla años empleo empleados	average age of the workforce years of employment employees
S	Gender diversity	Nº de hombre* mujer* total empleadas Empleo Distribución por sexo género igualdad laboral integración "102-8"	No. of m*n wom*n total employees Employment Distribution by sex gender labour equality integration "102-8"
S	Gender diversity on the Board	distribución por sexo "102-8" Consejo de administración mujer* Composición Compromiso igualdad Consejeros	distribution by sex "102-8" Board of Directors wom*n Composition Commitment to equality Directors
S	Average wage gap	Brecha salarial Sueldos por sexo Igualdad de oportunidades Retribución Paridad Políticas de género a favor de hombre* mujer* equidad remuneración equitativa	Salary gap Salaries by sex Equal opportunities Remuneration Parity Gender policies in favour of m*n wom*n equity equal pay
S	Work stability	Empleados por contrato laboral "102-8" Permanente Temporal Indefinidos Personal fijo tipo Tipología laborales	Employees by employment contract "102-8" Permanent Temporary Indefinite Permanent staff type Type of work
S	Disability	Discapacidad "405-1" Empleados Personal discapacitado grado Grupos vulnerables integración social inserción laboral igualdad de oportunidades accesibilidad exclusión diversidad discriminación respetuoso	Disability "405-1" Employees Disabled staff degree Vulnerable groups social integration job placement equal opportunities accessibility exclusion diversity discrimination respectful
S	Absenteeism	absentismo días puesto de trabajo ausencia ausentismo perdidos laborables laborales total horas jornadas	absenteeism days job absence absenteeism lost working days total hours working hours
S	Employee rotation	rotación de empleados abandono de puesto de trabajo fin relación laboral número personas abandonan "404-1" plantilla cese voluntario	employee turnover job abandonment termination of employment relationship number of people leaving "404-1" template voluntary resignation
S	Employee training	total horas de formación cursos empleado trabajador % año	total hours of training courses employee worker % year
S	Number of layoffs	Número de despidos empleados despedidos extinción rescisiones fin relación laboral cese indemnización terminación rescindir contrato trabajo	Number of layoffs dismissed employees termination terminations end of employment relationship cessation compensation termination terminate employment contract
S	Reconciliation measures	Conciliación bienestar empleados medidas de laboral políticas sociales reducir estrés emocional mental vida privada familiar equilibrio hijos mayores discapacitados compaginar conciliar	Reconciliation well-being employees labour measures social policies reduce mental emotional stress private family life balance older children disabled reconcile
S	Human rights policy	Compromiso social política de derechos humanos 412 vulneración abusos vulnerable responsabilidad	Social commitment human rights policy 412 violation abuses vulnerable responsibility
S	Equality plan	Compromiso social plan política de igualdad oportunidades "102-8" "406-1" "404-2" inclusión discriminación raza sexo color religión diversidad género mujer*	Social commitment equal opportunities policy plan "102-8" "406-1" "404-2" inclusion discrimination race sex colour religion diversity gender wom*n

SOURCE: Banco de España.

Table A2.1

**SUSTAINABILITY INFORMATION IN NON-FINANCIAL CORPORATIONS (cont'd.)**

Type (E,S,G)	Indicator	Dictionary of words (original queries in Spanish)	Dictionary of words (English)
S	Diversity plan	Políticas de inclusión Igualdad de derechos LGTBI LGBTI Plan de diversidad plantilla 405 origen étnico orientación identidad sexual sin prejuicios	Inclusion policies Equal rights LGTBI LGBTI Diversity plan 405 template ethnic origin orientation sexual identity without prejudice
S	Health and safety policy	Salud y seguridad política de asistencia sanitaria seguros médicos beneficios sociales planes de ayuda beneficios sociales 403 bienestar de los empleados	Health and safety health care policy medical insurance social benefits assistance plans social benefits 403 employee welfare
S	Payments to suppliers	Periodo medio pago a proveedores días PMP cadena de suministro grupos de interés	Average payment period to suppliers days PMP supply chain stakeholders
S	Customer satisfaction level	Nivel de satisfacción cliente servicio postventa trato experiencia compra usuario grado "102-43" GRI "102- 44" insatisfacción grupos interés	Level of customer satisfaction after-sales service treatment user purchase experience grade "102-43" GRI "102-44" dissatisfaction with interest groups
G	Average Board remuneration	Remuneración media percibida salario del consejo retribución del consejo administración 405 miles anual euros variable especie monetaria fija a largo plazo	Average remuneration received salary of the board remuneration of the board of directors 405 thousand euros per year variable long-term fixed monetary species
G	Crime prevention policy	Modelo de Prevención de Delitos Certificado AENOR Sistema de gestión de compliance penal antisoborno 205 Anticorrupción UNE-ISO 37001 UNE 19601 soborno	Crime Prevention Model AENOR Certificate Anti-bribery criminal compliance management system 205 Anti- corruption UNE-ISO 37001 UNE 19601 bribery
G	Number of corruption and bribery complaints	Casos de corrupción soborno políticas antisoborno nº "205-3" fraude prevenir luchar sanción indebido denuncia	Corruption cases bribery anti-bribery policies number "205-3" fraud prevent fight sanction improper complaint
G	Complaints channel	Canal de denuncias Ética y Cumplimiento sistemas de comunicación comunicar fraud* empleados transparencia ilegal irregular* conducta Anonimato Confidencialidad 205 de forma confidencial y anónima	Ethics and Compliance Whistleblowing Channel Communication Systems Communicate Fraud* Employees Transparency Illegal Irregular* Conduct Anonymity Confidentiality 205 Confidentially and Anonymously

SOURCE: Banco de España.

# A WEB APPLICATION PROTOTYPE FOR INFORMATION RETRIEVAL AND STORAGE

## 1 Introduction

Banco de España opened several lines of research on sustainable finance in its 2021-2022 analytical program. Within this context, members of the Statistics Department (data scientists and accounting experts), together with other departments, began in March 2021 a project to create a structured database containing sustainability information reported by non-financial corporations in their annual non-financial reports.

Although some sustainability indicators have become mandatory in the non-financial reports annually submitted by Spanish non-financial corporations to the Mercantile Registers, they are often reported in an unstructured format, such as tables or images contained in the annexes of their annual non-financial statements.

The creation of the new database in a short period of time (March to July 2021) was made possible by an experimental prototype developed outside of the regular IT environment of Banco de España. A web application for the semi-automatic extraction and storage of information was developed, for extracting (digital and scanned) text, and for indexing, searching and storing data in a relational database. The tool suggests the most relevant text fragments to the user, who needs to validate the search results by selecting the correct value for each indicator and storing it in the database.

The tool developed has recently been used by 20 business experts to create the first version of the new sustainability database, which is hosted in the Central Balance Sheet Data Office. After the first phase of this project, the new sustainability database contains 39 environmental, social and governance indicators (selected from a list of over 100), 77% of which are included in the GRI international standard.

This paper describes the technical approach used to create the new sustainability database and the results obtained during the first ingestion phase. Section 1 outlines the motivation and goals of this project and describes the target information and the main challenges faced. Section 2 sets out the main modules and technical details of this experimental prototype. Section 3 describes the results obtained during the first data ingestion process whereby the first version of the sustainability database was created. Finally, Section 4 summarizes the conclusions and future lines of research.

Previous attempts have been made to extract sustainability information from unstructured company reports. However, to the best of our knowledge, none of

these have produced a database of sustainability indicators, which is the main goal of the present work. In Moreno and Caminero (2020), the authors apply text mining techniques to analyse the Task Force on Climate-related Financial Disclosures (TCFD) recommendations on climate-related disclosures of the 12 Spanish significant financial institutions, using publicly available corporate reports from 2014 to 2019. In Moreno and Caminero (2022), the authors present an extension of this work to Pillar 3 reports.

In its 2018 and 2019 Status Reports, the TCFD also used supervised machine learning techniques to identify areas of the corporate reports potentially containing information related to each of 11 recommended disclosures related to the four recommendations (Moreno and Caminero (2020)). In Bingler, Kraus and Leippold (2021) and Webersinke (2021), the authors train ClimateBert, a deep neural language model, on thousands of sentences related to climate-risk disclosures aligned with the TCFD recommendations. This model can be used for various climate-related downstream tasks like text classification, sentiment analysis and fact-checking.

Other attempts to analyse climate-related documents using machine learning include Friederich (2021), Luccioni and Palacios (2019), Luccioni, Baylor and Duchene (2020) and Tarquinio, Rauchi and Benedetti (2018). In Friederich (2021), the authors use machine learning to automatically identify disclosures of five different types of climate-related risks, creating a dataset of over 120 manually-annotated annual reports by European firms. In Luccioni and Palacios (2019), natural language processing (NLP) techniques are applied to determine the companies that divulge their climate risks and those that do not, identify the types of vulnerabilities that are disclosed and follow the evolution of these risks over time. In Luccioni, Baylor and Duchene (2020), a custom NLP model named ClimateQA is proposed, which enables analysis of financial reports in order to identify climate-relevant sections based on a question answering approach. Finally, in Tarquinio, Rauchi and Benedetti (2018), the authors explore the performance indicators disclosed in the GRI-based Sustainability Reports (SRs) produced by the companies of three different countries: Italy, Spain and Greece. They use regression trees to describe how the companies' variables explain the different use of the indicators. Their findings show that Spanish companies, on average, disclose the greatest number of indicators. Labour-related social indicators are those most frequently reported in the SRs of the three countries.

## Target information

The first phase of this project, completed during 2021, sought to retrieve the values of the 39 continuous and categorical sustainability indicators of interest listed in Table 7.

Table 7

**ENVIRONMENTAL, SOCIAL AND GOVERNANCE INDICATORS COLLECTED DURING THE FIRST DATA INGESTION PHASE**

## Environmental indicators (17):

Energy consumption and reduction (MWh, GJ, etc.)
Total water consumption (m3, Hm3, mega litres, etc.)
Greenhouse gas emissions scope 1, 2 and 3 (tCO2e, etc.)
Greenhouse gas emissions reduction (tCO2e, etc.)
Circular economy (yes or no)
Greenhouse gas emissions intensity (ratio)
Environmental policy (yes or no)
Total waste generated (t)
Waste not intended for disposal (t)
Hazardous waste (t)
Renewable energy as a % of total energy (%)
Company located in a water-stressed area (yes, no)
ISO 14001 certification (yes, no)
Non-hazardous waste (t)

## Social indicators (18):

Diversity plan (yes, no)
Number of employees
Number of employees with disabilities
Gender diversity (number of women employed)
Gender diversity on the board (% of women)
Permanent employees (%)
Equality plan (yes, no)
Health and safety policy (yes, no)
Human rights policy (yes, no)
Average age of the workforce (years)
Number of dismissals
Average pay gap (%)
Employee absenteeism (days, hours)
Employee turnover (number of employees who abandon voluntarily)
Employee training (hours)
Reconciliation measures (yes, no)
Payments to suppliers (days)
Customer satisfaction level (%)

## Governance indicators (4):

Number of corruption and bribery complaints
Channel for complaints (yes, no)
Average Board remuneration (€)
Crime prevention policy (yes, no)

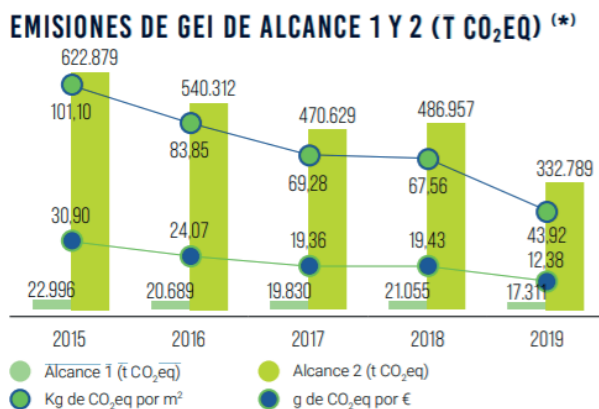
**SOURCE:** Banco de España.

The target information is presented in highly heterogeneous formats within the company reports, including plain text, tables, graphics and images. Figure 4 shows examples of the ways in which information is presented. To date, no standard has been defined on how this information should be presented nor on the units to be used. For this reason, a flexible tool had to be designed in order to extract the information in all possible formats. EU regulation is adapting and it is expected that clear rules on how information should be reported will be defined in the near future, making the information extraction process much easier.

Figure 4

**EXAMPLES OF THE WAYS INFORMATION IS PRESENTED IN THE REPORTS**

- Reducimos nuestras emisiones de carbono un 49,6% respecto a 2015 (alcances 1+2) y 18,5% las de nuestra cadena de valor (alcance 3) respecto a 2016
- Reducimos las emisiones de nuestra cadena de suministro por euro comprado un 24,6% respecto a 2016
- Con nuestros servicios evitamos más de 3,2 millones de tCO<sub>2</sub>, 3,3 veces nuestra huella de carbono
- Redujimos un 71,8% nuestro consumo de energía por unidad de tráfico
- Hemos sido reconocidos con la máxima clasificación "A" en el CDP Climate Change
- Reciclamos el 98,4% de nuestros residuos



SOURCE: Inditex sustainability report (2019).

Due to the range of formats in which this information is reported, a semi-automatic approach for information retrieval was preferred, which requires the user’s validation in order to guarantee that only high quality data populate the new database. Documents usually contain both text in digital format and (often low-quality) scanned images, making manual information extraction extremely costly. However, a fully automatic information retrieval approach was discarded due to the impossibility of handling the problem of retrieving such complex and heterogeneous information in a fully automatic way.

Previous attempts have been made to extract sustainability information from unstructured company reports. In Moreno and Caminero (2020) the authors applied text mining techniques to analyse TCFD recommendations on climate-

Figure 4

**EXAMPLES OF THE WAYS INFORMATION IS PRESENTED IN THE REPORTS (cont'd.)**

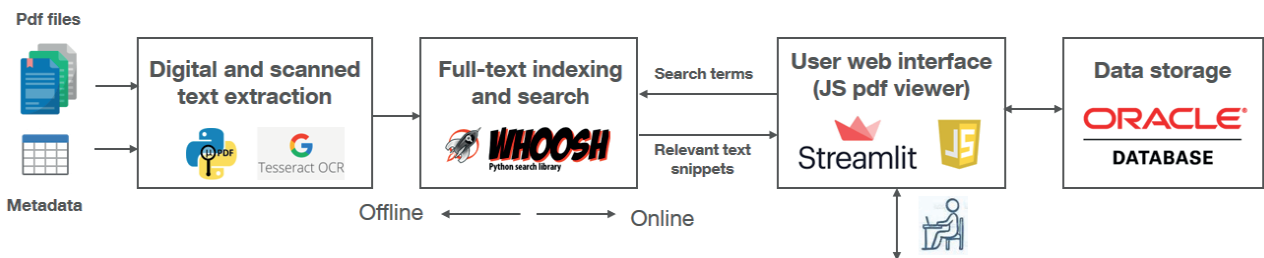
Emisiones Alcance 1 (tCO2e)	297.042	291.787	295.622	252.937	237.620	-20,0 %
Emisiones Alcance 2 (basado en método de mercado) (tCO2e)	1.615.146	1.153.046	1.059.796	923.719	725.326	-55,1 %
Emisiones Alcance 1 y 2 (tCO2e)	1.912.188	1.444.833	1.355.418	1.176.656	962.946	-49,6 %



SOURCE: Telefonica and Repsol sustainability report (2019).

Figure 5

**MAIN MODULES OF THE SUSTAINABILITY INFORMATION RETRIEVAL AND STORAGE TOOL**

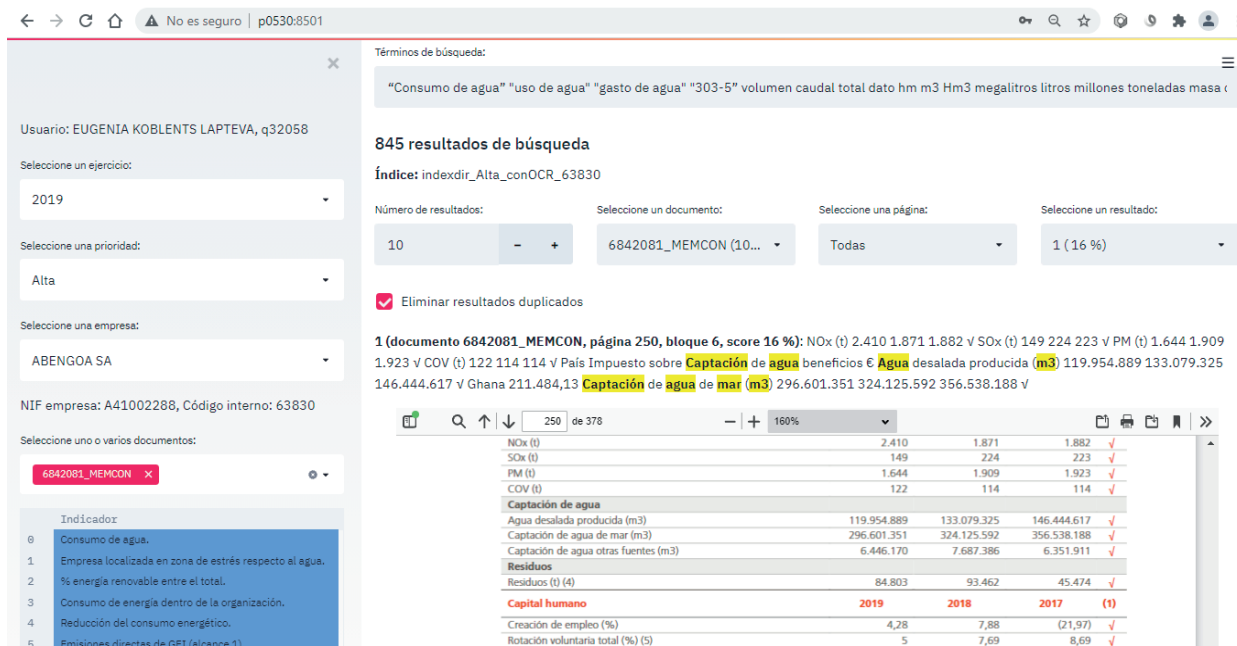


SOURCE: Devised by authors.

related disclosures of the 12 significant Spanish financial institutions using publicly available corporate reports from 2014 until 2019.

Figure 6

MAIN VIEW OF THE WEB USER INTERFACE



SOURCE: Devised by authors.

In its 2018 and 2019 Status Reports the TCFD also made use of supervised machine learning techniques to identify sections of the corporate reports potentially containing information on each of the 11 recommended disclosures related to the four recommendations (see Moreno and Caminero (2020)).

## 2 Technical approach

The technical goal of this project is the transformation of unstructured information contained in large sets of documents into a structured relational database, so that it can be combined with additional information to get meaningful insights on sustainability and climate change.

A semi-automatic tool with full-text search and storage capabilities has been developed to fulfil this goal. The tool consists of an offline and an online processing part with a web user interface to allow for an easy interaction with the end user. Figure 5 summarizes the main modules of the tool, which are described in the following sub-sections.

The input data to the system consist of a set of pdf documents, the metadata related to those documents and the pre-defined taxonomy of search terms.



EXAMPLES OF TEXT IN DIGITAL FORMAT (BOTTOM) AND TEXT CONTAINED IN SCANNED IMAGES (TOP)

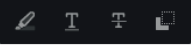
Nuestra opinión de auditoría sobre las cuentas anuales consolidadas no cubre el informe de gestión consolidado. Nuestra responsabilidad sobre la información contenida en el informe de gestión consolidado se encuentra definida en la normativa reguladora de la actividad de auditoría de cuentas, que establece dos niveles diferenciados sobre la misma:

- a) Un nivel específico que resulta de aplicación al estado de la información no financiera consolidado, así como a determinada información incluida en el Informe Anual de Gobierno Corporativo (IAGC), según se define en el art. 35.2. b) de la Ley 22/2015, de Auditoría de Cuentas, que consiste en comprobar únicamente que la citada información se ha facilitado en el informe de gestión consolidado, o en su caso, que se ha incorporado en éste la referencia correspondiente al informe separado sobre la información no financiera en la forma prevista en la normativa y en caso contrario, a informar sobre ello.

## 1. Quiénes somos

Acerinox es la compañía de fabricación, distribución y venta de acero inoxidable más global del mundo con presencia en los cinco continentes, fábricas en cuatro de ellos y suministro a clientes de 81 países.

El Grupo cuenta con una red de producción constituida por seis fábricas – ubicadas estratégicamente por sus ventajas de distribución o por su cercanía a fuentes de materias primas – distribuidas en Europa, América del Norte, África y Asia.

La comercialización  se realiza a través de una red formada por centros de servicio, almacenes, oficinas y agentes comerciales que cuenta con capacidad para suministrar a cualquier región industrializada del mundo.

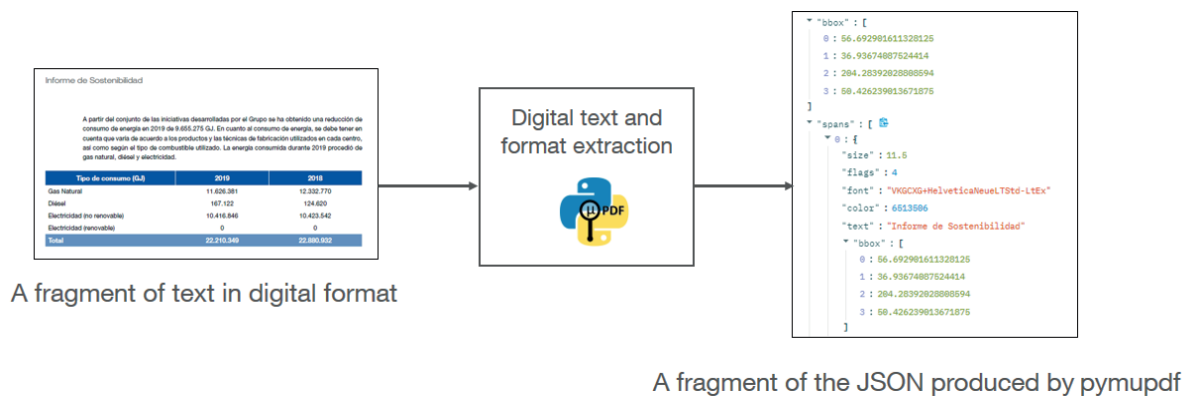
SOURCE: Acerinox sustainability report (2019).

The offline processing part carries out the pre-processing, text extraction and indexing, and generates a search index for each company as a result. Then, search and data storage are performed online by the end-user via the interactive web user interface.

The user interface is a web application that allows end users to interact with all of the tool’s functionalities to extract relevant information from a large set of documents. Figure 6 shows the main view of the web user interface, which incorporates multiple filters and selectors, advanced search and storage capabilities and a control panel to track data ingestion.

Once all the documents have been pre-processed and indexed during an offline processing phase, users can conduct online searches and store the data using the interactive web interface. Users log into the web application using their personal credentials. They then select a year, a company name and a sustainability indicator and the tool automatically loads all the information related to that selection (in particular, an index associated to that company and a pre-defined list of search terms). The tool then searches for all the pre-defined search terms in the corresponding index and retrieves a list of relevant text snippets sorted by a relevance score. Users can filter the results and select the result that contains the

Figure 8  
DIGITAL TEXT EXTRACTION WITH PYMUPDF



SOURCE: Devised by authors.

indicator of interest. They finally fill in a form with the information related to that indicator that needs to be stored. Additionally, the tool automatically stores the full context information of the selected search result (location in the document, surrounding text, etc.). All the modules involved in this process are described in more detail in the following sub-sections.

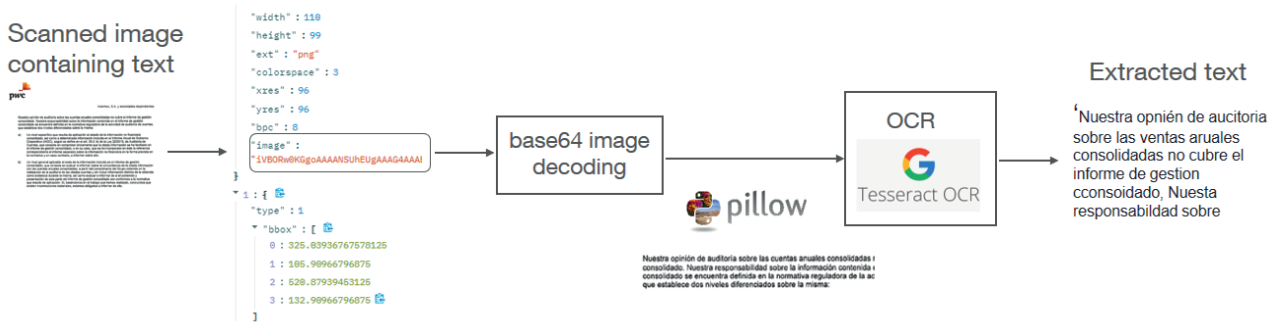
### Description of input data

The main source of input data for the developed tool is a set of pdf documents reported by Spanish non-financial corporations. Highly heterogeneous documents of variable length are available for each company. These documents contain text both in digital format, which is readily extractable, and text contained in (often low-quality) scanned images, which usually contains errors due to the OCR (Optical Character Recognition) process. Figure 7 shows an example of text fragments in digital format (bottom) and text contained in a scanned image (top). Over one thousand documents (6GB) have been processed to date. Only documents in Spanish have been considered so far and multilingual processing has thus not been addressed.

The metadata related to the set of documents and reporting companies are also required for the tool to operate. In particular, the metadata matrix includes a unique document identifier, the document type and date, and a company name and identifier, among other additional fields. This metadata matrix is updated every time a new document is indexed and is currently stored together with the corresponding documents.

Figure 9

SCANNED TEXT EXTRACTION WITH PYMUPDF, BASE64 IMAGE DECODING AND TESSERACT OCR



SOURCE: Devised by authors.

Using the input data sources described above the text extraction and indexing modules generate an index search for each company, which is also stored with a unique identifier. Finally, a taxonomy of search terms needs to be defined for each sustainability indicator of interest, which is stored in the database and is fed into the search module.

### Digital and scanned text extraction

This module extracts the textual content of the set of documents, which will later be indexed by the indexing module. It takes as input the set of documents and the metadata file. Documents usually contain both text in digital format and text contained in scanned images. These two sources of text data need to be processed in different ways.

Text in digital format can be readily extracted from documents and it usually does not contain any errors. In this case the digital text is extracted from the documents using the Python library pymupdf. This library makes it possible to extract text to multiple formats, including plain text, blocks, words, JSON, xml, and other formats. The JSON format was preferred, since it contains rich format information (text location, font, size, etc.) in addition to the text content itself. Figure 8 shows an example of digital text extraction with pymupdf. The output JSON structure contains rich information describing the extracted text, including the coordinates of the text bounding box font, size, color, etc.

Images, possibly containing scanned text, appear as base64 encoded strings in the JSON structure provided by pymupdf. These strings need to be decoded and converted into PIL images prior to being processed with an OCR system. The Tesseract OCR engine, implemented in the Python library pytesseract, is then

Figure 10

SUSTAINABILITY INDICATOR SEARCH USER INTERFACE

Búsqueda de indicadores

Seleccione un indicador para búsqueda: Descripción del indicador +

Consumo de agua

Restablecer términos de búsqueda

Términos de búsqueda:

agua consumo m3 303-5 fuentes reutilizada dato hídrico uso total subterránea m² captación toneladas litros mar volumen hm río dulce marina lago millones caudal gasto hm3

938 resultados de búsqueda

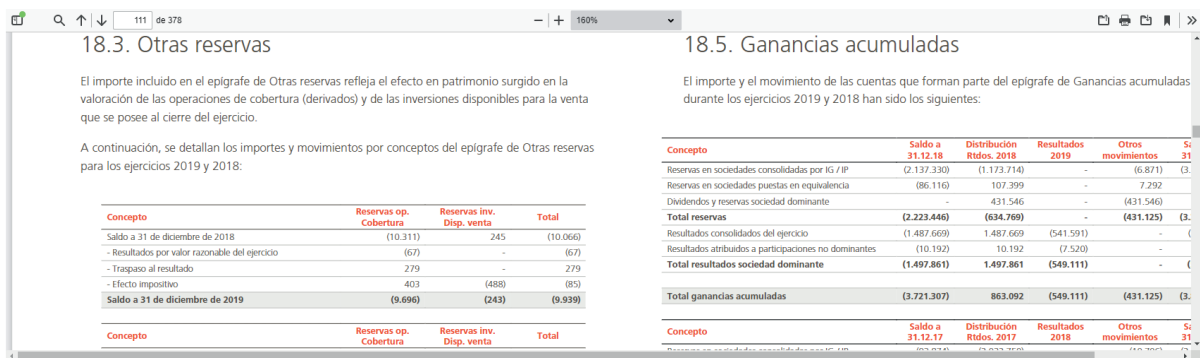
Número de resultados: Seleccione un documento: Seleccione una página: Seleccione un resultado:

10 - + 6842081\_MEMCON (100%) Todas Todos

Eliminar resultados duplicados

1 (documento 6842081\_MEMCON, página 250, bloque 6, score 15 %): NOx (t) 2.410 1.871 1.882 v SOx (t) 149 224 223 v PM (t) 1.644 1.909 1.923 v COV (t) 122 114 114 v País Impuesto sobre **Captación de agua** beneficios E **Agua** desalada producida (m3) 119.954.889 133.079.325 146.444.617 v Ghana 211.484,13 **Captación de agua de mar** (m3) 296.601.351 324.125.592 356.538.188 v

2 (documento 6842081\_MEMCON, página 251, bloque 6, score 14 %): **Consumo de agua** y el suministro de **agua** de acuerdo con las limitaciones locales 76 **303-1 Consumo** de materias primas y las medidas adoptadas para mejorar la eficiencia de su **uso** 76 301-1 **Consumo**, directo e indirecto, de energía, 76 302-1 Medidas tomadas para mejorar la eficiencia energética 88 302-4



SOURCE: Devised by authors.

applied to all images in order to extract the scanned text contained in the documents. Figure 9 shows the scanned text extraction workflow implemented in this project.

The quality of scanned images is often low, yielding OCR errors that can sometimes be corrected using available error correction tools. In particular, the pyspellchecker and hunspell libraries are available in Python. The indexing and search engine implemented in the Python library whoosh is able to conduct fuzzy searches, which can also be used as an alternative technique to mitigate errors in the extracted text. However, no error correction tools have been used here, since they can potentially give rise to errors in correct but non-standard terms, such as the names of entities and other words.

Both the text extraction and indexing modules are executed offline every time a new document becomes available and are not accessible from the web user interface.

## Full-text indexing and search

The indexing and search module has been developed using the Python library whoosh, which provides fast, featureful, full-text indexing and searches in pure Python, suitable for moderate-sized document databases. Full-text search engines make it possible to quickly search large volumes of text for a custom textual query, as opposed to metadata-based or exact-match search engines. Full-text search systems rely on an inverted index that indicates the fragments where each term appears.

In this case, the digital and scanned text is indexed in blocks of more than 50 words, sorted by their y-coordinate. An index schema and a language analyser including a tokenizer, a stemmer, a stop-word filter, etc., need to be defined. For scalability and robustness reasons, an index is created for each company. This indexing process is performed offline by the tool developers.

Once the textual content has been extracted from the documents and indexed, end-users can search for (sustainability) information in real-time using the web user interface and the generated search indices.

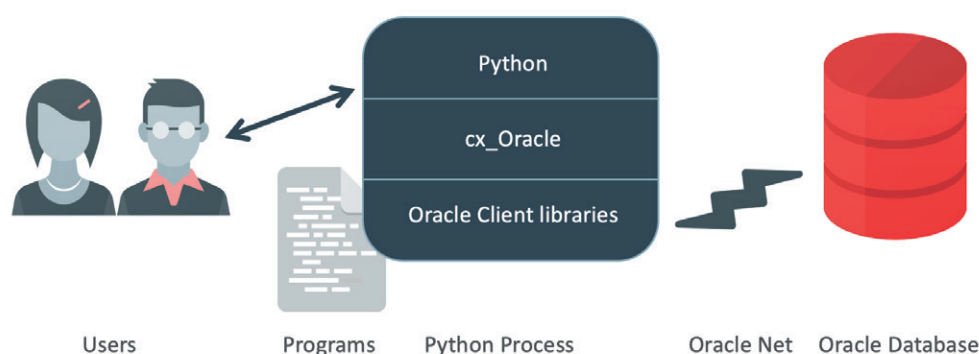
To standardise the search terminology, a taxonomy of search terms has been pre-defined for each sustainability indicator of interest based on expert knowledge. This taxonomy is stored in an Oracle database and is accessible from the user web interface. When the user selects a sustainability indicator of interest in the user interface, the pre-defined list of search terms is automatically loaded and a search query is built based on those terms. By default, OR operators have been used for the query creation but other logical operators can be used to build custom search queries.

Figure 10 shows the part of the user interface dedicated to searching for the sustainability indicators. The first selector allows the user to choose a sustainability indicator from a pre-defined list. A short description is loaded alongside it in an expandable container. The list of pre-defined search terms is automatically loaded into the search bar. The user can also modify the search terms in real time by typing into the search bar using the tool as a standard full-text search engine. The pre-defined search query can be restored by pressing the corresponding button.

The index associated with the selected company is automatically searched for the terms listed in the search bar. The tool allows the user to select the number of search results to be shown, and filter them by document and page. The filtered list of the most relevant text snippets is shown below, sorted by the scoring metric calculated by the whoosh library. In particular, the BM-25 scoring algorithm has been used in this case. Each search result is listed together with the document name, page, text block and relevance score.

Figure 11

## WEB APP-DATABASE COMMUNICATION



SOURCE: Devised by authors.

The tool automatically eliminates duplicated search results, when multiple copies of certain fragments of text appear within the documents or when multiple OCR systems have been applied to the images contained in the documents.

When a search result is selected, the corresponding text snippet is shown on an integrated JavaScript pdf viewer, which incorporates additional pdf handling capabilities, such as exact match search, page navigation, zoom, etc.

### Data storage

A relational database was chosen to store the new database for two reasons. First, a relational database is the natural way to store structured information extracted from a set of documents, making it possible to merge the new database with other relevant data, such as the firms' financial data (activity, sales, employees, etc.), to generate meaningful and complete statistics.

Second, a relational database allows for simultaneous read-write access by a large number of users, which was an important implementation requirement. In particular, an Oracle database is used in this project because it is the standard software used by the Banco de España's IT team for this type of applications. Alternative solutions have not been explored. The implemented storage solution based on an Oracle relational database has performed very well in terms of speed, simultaneous access, data protection and the possibility to recover past versions of the database, etc.

The database should be as standardised as possible. In this case, the primary database comprises the variables indicator, year, company, country, document year and a Boolean feature indicating whether the data is consolidated or not.

Figure 12  
**STORAGE USER INTERFACE**

The screenshot displays the 'STORAGE USER INTERFACE' for a user named ALEJANDRO MORALES FERNÁNDEZ. The interface is divided into several sections:

- User Information:** Shows the user's name and a list of selected documents (e.g., 6794700\_MEMCON, 6500449\_MEMO).
- Form Fields:** Includes dropdowns for 'Selección de ejercicio' (2019), 'Selección de prioridad' (Alta), 'Selección de una empresa' (CEMENTOS MOLINS SA), and 'Selección de un año' (2019). There are also fields for 'Selección de un indicador' (Consumo de agua), 'Selección de un tipo de dato' (Consolidado (grupo)), and 'Selección de una métrica' (Hm3).
- Table of Indicators:** A list of indicators with checkboxes. Indicators 17-23 are highlighted in blue, while indicators 24-27 are highlighted in red, indicating they are not yet included in the table.
- Data Entry Form:** A section for entering values, including a 'Recuerde que los puntos denotan miles y las comas decimales' warning, a 'Rellene el valor del indicador' field, and a 'Selección de una métrica' dropdown.
- Visualization and Download:** A section titled 'Visualización y descarga de la base de datos' with a filter dropdown set to 'Mostrar tabla filtrada por empresa'.
- Table of Results:** A table with columns: Indicador, año, Ejercicio Memoria, Valor, Métrica, Nombre Empresa, país, Tipo Dato, and Comentarios. The table contains 10 rows of data for various indicators.

SOURCE: Devised by authors.

The web app accesses the database by means of the python library `cx_oracle`, as can be seen in Figure 11. Thus, the user enters the data that will be stored by choosing from a range of different values, in some cases, or by simply typing them in directly, and the data is stored in the database via SQL commands using the above-mentioned python library.

There are other input and intermediate tables which are also stored in this database, such as the ontology table, a table containing the units, etc. Moreover, the traceability of the insertions and deletions and of who made those changes is also kept in a backup table.

Provisionally, the authentication provided by the database is used for authentication in the web app. In the future other specific software such as Kerberos will be used for this purpose.

Once users have found the correct results in the document and are ready to enter the information into the database, they have a fixed range of parameters to choose from. Only a small number of fields are free text. They can also monitor which indicators are not yet included in the table in the left margin (in red). Finally, they can check the results entered into the table at the bottom and modify the data if needed. These details can be seen in Figure 12.

Figure 13

### USERS MUST ENTER THEIR CREDENTIALS TO LOG INTO THE APP

Introduzca su q de usuario: Ejemplo q32057

Introduzca su password a la bbdd

SOURCE: Devised by authors.

Figure 14

### USERS CAN CUSTOMISE THE SEARCH TERMS FOR EACH INDICATOR. THEY CAN THEN NAVIGATE THE SEARCH RESULTS AND CHOOSE THE CORRECT ONE

Búsqueda de indicadores

Seleccione un indicador para búsqueda: Consumo de agua

Restablecer términos de búsqueda

Términos de búsqueda: agua consumo m3 303-5 fuentes reutilizada dato hídrico uso total subterránea m³ captación toneladas litros mar volumen hm río dulce marina la

938 resultados de búsqueda

Número de resultados: 10

Seleccione un documento: 6842081\_MEMCON (100...)

Seleccione una página: Todas

Seleccione un resultado: Todos

Eliminar resultados duplicados

1 (documento 6842081\_MEMCON, página 250, bloque 6, score 15 %): NOx (t) 2.410 1.871 1.882 √ SOx (t) 149 224 223 √ PM (t) 1.644 1.909 1.923 √ COV (t) 122 114 114 √ País Impuesto sobre Captación de agua beneficios € Agua desalada producida (m3) 119.954.889 133.079.325 146.444.617 √ Ghana 211.484,13 Captación de agua de mar (m3) 296.601.351 324.125.592 356.538.188 √

2 (documento 6842081\_MEMCON, página 251, bloque 6, score 14 %): Consumo de agua y el suministro de agua de acuerdo con las limitaciones locales 76 303-1 Consumo de materias primas y las medidas adoptadas para mejorar la eficiencia de su uso 76 301-1 Consumo, directo e indirecto, de energía, 76 302-1 Medidas tomadas para mejorar la eficiencia energética 58 302-4

SOURCE: Devised by authors.

## User interface

The web user interface is one of the most important modules in this project. As previously mentioned, the process of transferring indicator information from the pdf documents to a structured database is semi-automatic, i.e. it needs a user to validate the possible indicators found by the search engine and choose the correct one (if any). The user also gives it the appropriate format and stores it in the database via the web app.

The web app incorporates a pdf viewer where the user can check the raw data about to be stored. In some particular cases, the user might find the indicator is on a different page than the one the indicated by the search engine. This could be due to the fact the indicator's value and description may be on different pages



Figure 15

THE PDF NAVIGATOR IS IMPORTANT TO VERIFY THE RESULTS OBTAINED IN THE PREVIOUS STEP IN THE ORIGINAL DOCUMENT

El detalle de las Reservas en sociedades consolidadas por integración global/proporcional y por el método de participación es el siguiente:

Actividad de negocio	Saldo a 31.12.19		Saldo a 31.12.18	
	IG / IP	MP	IG / IP	MP
Ingeniería y construcción (*)	(1.220.777)	(1.993)	(282.892)	(99.313)
Infraestructura tipo concesional	(2.097.138)	30.568	(1.854.438)	13.197
<b>Total</b>	<b>(3.317.915)</b>	<b>28.575</b>	<b>(2.137.330)</b>	<b>(86.116)</b>

(\*) Incluye la actividad discontinuada correspondiente a bioenergía.

### 18.6. Participaciones no dominantes

En este epígrafe se recoge la parte proporcional del Patrimonio neto de las sociedades del Grupo consolidadas por integración global y en las que participan otros accionistas distintos al mismo.

El movimiento del epígrafe de Participaciones no dominantes para el ejercicio 2019 es el siguiente:

Sociedad	Saldo a 31.12.18	Cambios en el perímetro	Variaciones (1)	Imputación Rdo 2019	Saldo a 31.12.19
Solar Power Plant One	21.043	-	(1.514)	4.633	24.162
Société d'Eau Dessalée d'Agadir	11.860	-	3.856	2.195	17.911
Khi Solar One	3.617	-	80	(5.409)	(1.712)
Tenes Lyimyah	64.260	-	(6.605)	5.850	63.505
Zona Norte Engenharia	22.129	(24.228)	-	2.099	-
Abengoa Abenewco 1, S.A.U.	-	-	104.625	-	104.625
Otros menores	4.704	-	4.819	(1.849)	7.674
<b>Total</b>	<b>127.613</b>	<b>(24.228)</b>	<b>105.261</b>	<b>7.519</b>	<b>216.165</b>

(1) Variaciones producidas principalmente por ampliaciones/disminuciones de capital, reparto de dividendos, diferencias de conversión y cambios en el método de consolidación.

El movimiento del epígrafe de Participaciones no dominantes para el ejercicio 2018 es el siguiente:

Sociedad	Saldo a 31.12.17	Cambios en el perímetro	Variaciones (1)	Imputación Rdo 2018	Saldo a 31.12.18
LAT Brasil en operación	347.964	(347.964)	-	-	-
Solar Power Plant One	20.185	-	(2.752)	3.610	21.043
Société d'Eau Dessalée d'Agadir	9.113	-	1.744	1.003	11.860
Khi Solar One	9.786	-	(639)	(5.530)	3.617
Tenes Lyimyah	52.646	-	1.174	10.440	64.260
Zona Norte Engenharia	22.796	-	(3.268)	2.601	22.129
Otros menores	(417)	418	6.635	(1.932)	4.704
<b>Total</b>	<b>462.073</b>	<b>(347.546)</b>	<b>2.894</b>	<b>10.192</b>	<b>127.613</b>

(1) Variaciones producidas por ampliaciones/disminuciones de capital, diferencias de conversión principalmente y cambios en el método de consolidación.

Al cierre del ejercicio 2018, la disminución del epígrafe de participaciones no dominantes se corresponde con la salida del perímetro de consolidación de las líneas de transmisión en operación de Brasil por venta (ATE XI, Manaus Transmissora de Energia, S.A. y ATE XIII, Norte Brasil Transmissora de Energia, S.A.) (véase Nota 6.2.b)).

La relación de sociedades ajenas al Grupo que poseen una participación igual o superior al 10% del capital de alguna sociedad consolidada por el método de integración global del perímetro de consolidación a 31 de diciembre de 2019 se muestra en el Anexo VIII.

En la mayoría de los casos las participaciones no dominantes cuentan con los habituales derechos de protección, fundamentalmente en cuanto a restricciones de inversión, desinversión y financiación.

SOURCE: Devised by authors.

Figure 16

ALL THE INFORMATION IS SENT TO THE DATABASE IN A STANDARDISED AND STRUCTURED FORMAT

Selección uno o varios documentos:

6957113\_MEMCON X 6957192\_MEMCON X  
6957193\_EINF X

### Almacenamiento de resultados

Selección un indicador para almacenamiento:

Consumo de agua

Selección si ha encontrado el dato: Encontrado Tipo de Dato: Consolidado (grupo) Selección un año: 2021 Rellene el país: GRUPO

Rellene el valor del indicador: Selección la métrica: megalitros

**Recuerde que los puntos denotan miles y las comas decimales**

Añada un comentario opcional:

Actualizar Modificación Borrar Dato

Indicador:

- Consumo de agua \*
- Empresa localizada en zona de estrés respecto al agua \*
- % energía renovable entre el total \*
- Consumo de energía dentro de la organización.
- Reducción del consumo energético.
- Emissiones directas de GEI (alcance 1).
- Emissiones directas de GEI al generar energía (alcance 2).
- Intensidad de las emisiones de GEI \*
- Otras emisiones indirectas de GEI (alcance 3) \*
- Reducción de las emisiones de GEI \*
- Cumplimiento normativa ISO14001.

Se indican en verde los indicadores que han sido rellenados con éxito para la empresa seleccionada, en amarillo los que han sido buscados, pero no encontrados y en rojo los que todavía no han sido tratados

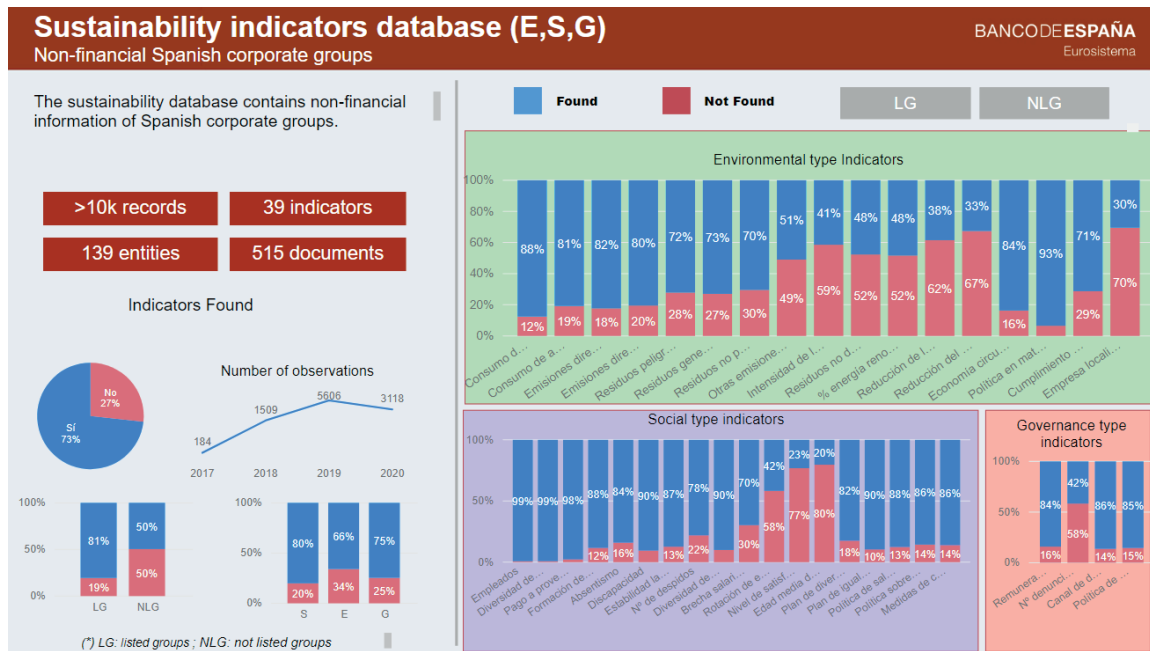
SOURCE: Devised by authors.

(e.g. the values could be on the following page, within an image or in the annex, stored in a table).

The application also has a control panel to track the insertion of data and the companies whose completion is pending for each user, along with other filters and

Figure 17

**SEVERAL DASHBOARDS WERE DRAWN UP TO SHOW THE EFFECTIVENESS OF THE APPLICATION**



SOURCE: Devised by authors.

selectors to navigate the range of documents and companies. Also, the web app allows users to delete records directly, instead of having to write SQL commands.

The web user interface has been temporarily implemented in Streamlit and deployed on local machines of the Banco de España’s Statistics Department. Streamlit is an easy-to-use open-source app framework for data visualization and web development.

This software will be migrated to the python framework Dash and run on a corporate server.

A detailed view of the above-mentioned modules of the app (authentication, search, pdf viewer and storage) can be seen in Figures 13, 14, 15 and 16 respectively.

### 3 First data ingestion

Approximately 20 users worked part-time for two months in 2020 to make the first data ingestion.

The 39 indicators had to be entered for 139 corporate groups. 515 documents were available for these groups (i.e. an average of approximately 2.9 documents per entity).

Around 10,500 records were stored in this first data ingestion and a further 4,500 were stored in a second batch, which is equivalent to some 4,500 rows.

The indicator was found for the given company 73% of the time. It was not found in 27% of the cases.

Listed groups provided more information than unlisted ones (81% of the records belonging to listed companies were found).

Social indicators were found in 80% of the searches compared to only 66% of environmental indicators. The proportion of governance data found was 75%.

The documents were from 2019 and 2020, although some of them contain information of past years, making it possible to create historical series for some indicators.

Several reports have been drawn up and made public. These reports are mainly dashboards made using PowerBI as can be seen in Figure 17.

## 4 Conclusions

A web application has been developed and deployed for creating a new database of sustainability information from large sets of documents (Spanish corporate groups' reports). This information has been obtained using a semi-automatic approach that reduces human effort, thanks to the search tools implemented.

This prototype was implemented in three months by two data scientists at the Banco de España's Statistics Department, with the support of domain experts for the definition of specifications, testing, etc. The prototype is expected to be put into production with the help of the IT department.

The tool includes the following functionalities: authentication, data filters and selectors, full-text search in multiple documents, data storage, deletion and download and a control panel to track the insertion of data.

The app allows users to customise the search terms (although default terms are given for each indicator).

## Next steps

The long-term focus of this project, which aims to store companies' sustainability information, means it will have to adapt as new regulations arise in different legal frameworks.

The work carried out so far has essentially consisted in the development of a production-ready prototype. Real data have already been extracted using the full life-cycle of the web app. Nevertheless, some work streams remain for the medium and long term:

- **Extension of the database:** The current database has been populated with information retrieved from documents of listed companies and consolidated groups. The plan is to expand this to other enterprises. Moreover, further steps will be taken to include financial entities, which will involve changes to the entire architecture.
- **Improvement of the ontology using NLP:** important information, such as the search terms, the snippet of text found or the location of the indicator, is stored with the above-mentioned 10,500 records. By using NLP, a better ontology of terms for each indicator can be established.
- **Publication of the database by the Banco de España:** The database is accessible to Central Balance Sheet Data Office users and to other departments in the Bank. This information will also be made available to external researchers in the BELAb data laboratory in the near future.
- **Migration to Dash:** Dash is the most common Python visualisation tool and is the Banco de España's official web development software in Python.
- **Migration to production servers:** The app has been deployed on local servers in the Statistics Department.
- **Exploration of alternative OCR systems:** The IT department has additional tools such as Uipath, which provides different OCR engines (Tesseract, Omnipage, etc.). Currently, the package used in this prototype to extract text from image formats is the pytesseract library.

## REFERENCES

- Bingler, Julia Anna, Mathias Kraus and Markus Leippold. (2021). "Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures", available at SSRN.
- Friederich, David et al. (2021). "Automated Identification of Climate Risk Disclosures in Annual Corporate Reports", arXiv preprint arXiv:2108.01415.
- Luccioni, Alexandra, and Hector Palacios. (2019). "Using Natural Language Processing to Analyze Financial Climate Disclosures", *proceedings of the 36th International Conference on Machine Learning, Long Beach, California*.
- Luccioni, Alexandra, Emily Baylor and Nicolas Duchene. (2020). "Analyzing Sustainability Reports Using Natural Language Processing", arXiv preprint arXiv:2011.08073.
- Moreno, Ángel Iván, and Teresa Caminero. (2020). "Application of Text Mining to the Analysis of Climate-Related Disclosures". International Conference on "Statistics for Sustainable Finance". Irving Fisher Committee on Central Bank Statistics, 14-15 September 2021.
- Moreno, Ángel Iván and Teresa Caminero. "Analysis of ESG Disclosures in Pillar 3 Reports. A Text Mining Approach". SUERF Policy Brief No 332, May 2022.
- Tarquino, Lara, Domenico Raucci and Roberto Benedetti. (2018). "An Investigation of Global Reporting Initiative Performance Indicators in Corporate Sustainability Reports: Greek, Italian and Spanish Evidence", *Sustainability*.
- Webersinke, Nicolas et al. (2021). "ClimateBert: A Pretrained Language Model for Climate-Related Text", arXiv preprint arXiv:2110.12010.

<https://www.bde.es/bde/en/areas/analisis-economi/otros/que-es-belab/>.

- 1 DEPARTAMENTO DE ESTADÍSTICA Y CENTRAL DE BALANCES: Registro de los Servicios de Intermediación Financiera en Contabilidad Nacional a partir de 2005. (Publicada una edición en inglés con el mismo número.)
- 2 DEPARTAMENTO DE ESTADÍSTICA Y CENTRAL DE BALANCES: Valoración de las acciones y otras participaciones en las *Cuentas Financieras de la Economía Española*. (Publicada una edición en inglés con el mismo número.)
- 3 DEPARTAMENTO DE ESTADÍSTICA Y CENTRAL DE BALANCES: Registro de los Servicios de Intermediación Financiera en Contabilidad Nacional a partir de 2005. Adendum. (Publicada una edición en inglés con el mismo número.)
- 4 LUIS GORDO MORA Y JOÃO NOGUEIRA MARTINS: How reliable are the statistics for the stability and growth pact?
- 5 DEPARTAMENTO DE ESTADÍSTICA: Nota metodológica de las *Cuentas Financieras de la Economía Española*.
- 6 DEPARTAMENTO DE ESTADÍSTICA: Nota metodológica de las *Cuentas Financieras de la Economía Española*. SEC-2010.
- 7 DEPARTAMENTO DE ESTADÍSTICA: *Holdings* y sedes centrales en el marco del SNA 2008/ SEC 2010.
- 8 DEPARTAMENTO DE ESTADÍSTICA: Presentación de los resultados de la encuesta de satisfacción de los usuarios de las estadísticas del Banco de España.
- 9 DEPARTAMENTO DE ESTADÍSTICA: Los cambios en la Balanza de Pagos y en la Posición de Inversión Internacional en 2014.
- 10 DEPARTAMENTO DE ESTADÍSTICA: Impacto de la revisión *benchmark* 2019 sobre la capacidad/necesidad de financiación y la Posición de Inversión Internacional de la economía española.
- 11 DEPARTAMENTO DE ESTADÍSTICA: La estimación de los ingresos por turismo en la Balanza de Pagos.
- 12 DEPARTAMENTO DE ESTADÍSTICA: Revisión extraordinaria de las *Cuentas Financieras de la Economía Española* (2019).
- 13 DANIEL SÁNCHEZ MENESES: The advantages of data-sharing: the use of mirror data and administrative data to improve the estimation of household external assets/liabilities (2020).
- 14 Ana Esteban e Ignacio González: Efectos de la aplicación de la NIIF16 sobre arrendamientos en los grupos cotizados españoles no financieros (2020).
- 15 ROBERTO BADÁS ARANGÜENA: La inversión exterior directa en España: ¿cuáles son los países inversores inmediatos y cuáles los últimos? (2021).
- 16 JAVIER JAREÑO MORAGO: Notas estadísticas relativas a las series históricas de los tipos de interés del Banco de España 1938-1998 (2022).

- 17 BORJA FERNÁNDEZ-ROSILLO SAN ISIDRO, EUGENIA KOBLENTS LAPTEVA Y ALEJANDRO MORALES FERNÁNDEZ: Micro-database for sustainability (ESG) indicators developed at the Banco de España (2022).

Se permite la reproducción para fines docentes o sin ánimo de lucro, siempre que se cite la fuente.

© Banco de España, Madrid, 2022  
ISSN 2530-7495 (edición electrónica)