IFC-Bank of Italy Workshop on "Data Science in Central Banking: Applications and tools"

14-17 February 2022

Creation of a structured sustainability database from company reports – a web application prototype for information retrieval and storage[1]

Eugenia Koblents and Alejandro Morales,
Bank of Spain

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Creation of a structured sustainability database from company reports

A web application prototype for information retrieval and storage

Alejandro Morales

Eugenia Koblents

Statistics Department, Banco de España

Abstract

A web application for semi-automated information extraction and storage has been developed at the Banco de España for retrieval of sustainability indicators from annual financial statements reported by Spanish non-financial corporations. The goal is to assist business users in the process of extracting sustainability indicators from large document databases and storing them in a structured format. The tool developed incorporates a set of pre-defined search terms for each indicator, which have been selected based on domain knowledge. For each company and indicator, the tool suggests the most relevant text snippets to the user, who identifies the correct indicator's value and stores it in the database via the user web interface.

This tool has been developed by two data scientists in three months, with the continuous support of a domain expert team for definition of requirements and refinements, input data collection and tool validation and testing.

The tool has been entirely implemented in Python and provides for interaction with the user by means of a user web interface. The web application has already been used by 20 people in the Banco de España's Statistics Department to create an initial version of the sustainability database, which currently contains more than 15,000 records, retrieved from a total of 800 reports submitted by 300 of the largest Spanish companies. To date, the new sustainability database contains information on 39 environmental, social and governance indicators, with plans for it to be extended in the near future.

This paper describes the technical approach adopted and the main modules of the prototype implemented, including text extraction, indexing and search, data storage and visualisation. It also presents an overview of the first version of the sustainability database created.

Keywords: sustainability, climate change, full-text search, OCR, databases, web application, Python.

JEL classification: Q5 (Environmental economics).

Index

1. Introduction

The Banco de España, the Spanish central bank, embarked on several avenues of research on sustainable finance in its 2021-2022 analytical programme. In this context, in March 2021 Statistics Department staff (data scientists and accounting experts), together with members of other departments, launched a project to create a structured database containing sustainability information reported by non-financial companies in their annual 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.

To allow for the creation of the new database in a short period of time (March to July 2021), an experimental prototype has been developed out of the Banco de España's regular IT environment. A web application for semi-automated information extraction and storage has been developed, which implements (digital and scanned) text extraction, indexing, search and storage 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 then stores 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 at the Central Balance Sheet Data Office (CBSO), a division of the Banco de España's Statistics Department. After the first phase of this project, the new sustainability database contains 39 environmental, social and governance (ESG) indicators (selected from a list of over 100), 77% of which are included in the Global Reporting Initiative (GRI) standard.

This paper describes the technical approach designed to create the new sustainability database and the results obtained during the first ingestion phase. Section 1 presents the motivation and goals of this project and describes the target information and the main challenges faced. Section 2 describes the main modules and technical details of this experimental prototype. Section 3 sets out the results obtained during the first data ingestion process, creating the first version of the sustainability database. Lastly, Section 4 summarises the conclusions and future avenues of research envisaged.

Target information

In the first phase of this project, completed during 2021, the goal was to retrieve the value of the 39 continuous and categorical sustainability indicators listed in Table 1.

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) (tC02e, etc.) ▪ Greenhouse gas emission reduction (tC02e, etc.) ▪ Circular economy (yes, no) ▪ Greenhouse gas emission intensity (ratio)

▪ Environmental policy (yes, no) ▪ Total waste generated (t) ▪ Waste not destined for disposal (t) ▪ Hazardous waste (t) ▪ Percentage of renewable energy (%) ▪ Company located in a stress area regarding water? (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 leave voluntarily) ▪ Employee training (hours) ▪ Work-life balance 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)

Table 1 ESG indicators collected during the first data ingestion phase.

The target information is presented in highly heterogeneous formats within the company reports, including plain text, tables, graphics and images. Figure 1 shows examples of the ways in which the information is presented. To date, no standard has been defined on how this information should be presented or how the metrics should be used. For this reason, a flexible tool capable of extracting the information in all possible formats had to be designed. European Union regulations are adapting and clear rules on how information should be reported are expected to be defined in the near future, making the information extraction process much easier.

Owing to the variety of formats in which information is reported, a semi-automated approach for information retrieval has been preferred, requiring user 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, which would make manual information extraction extremely costly. A fully automated information retrieval approach was ruled out owing to the

impossibility of handling such a complex and heterogeneous information retrieval problem without requiring human validation.

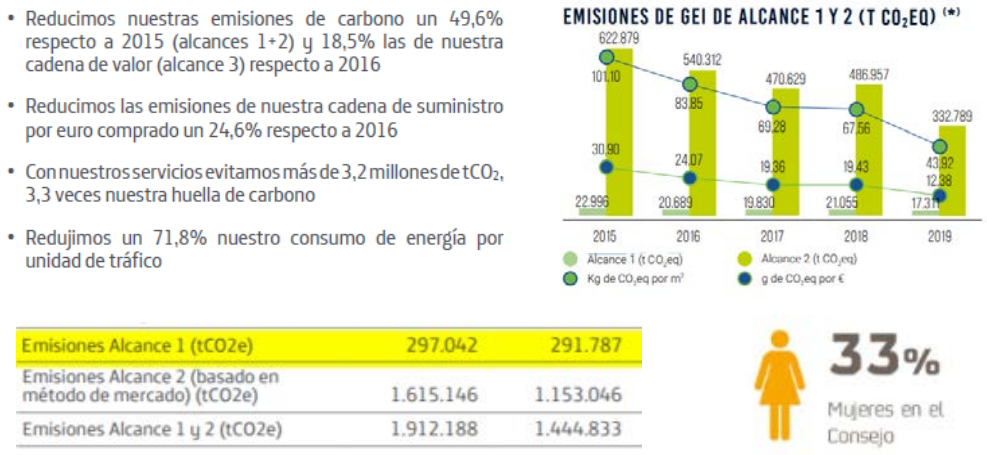


Figure 1 Examples of the ways information is presented in the reports.

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 [5] 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 [6] 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 one of 11 recommended disclosures related to the four recommendations [5]. In [1] and [8] 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 [2], [3], [4] and [7]. In [2] 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 [3] natural language processing (NLP) techniques are applied to pinpoint 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 [4] 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 [7] 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 a 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.

2. Technical approach

The technical goal of this project is to transform the non-structured information contained in large collections of documents into a structured relational database, in order to be able to combine it with additional information, to obtain meaningful insights into sustainability and climate change.

A semi-automated tool with full-text search and storage capabilities has been developed to fulfil this goal. The tool combines an offline and an online processing part and provides a user web interface to allow for easy interaction with the end user. Figure 2 summarises the main modules of the tool, which are described in the next subsections.

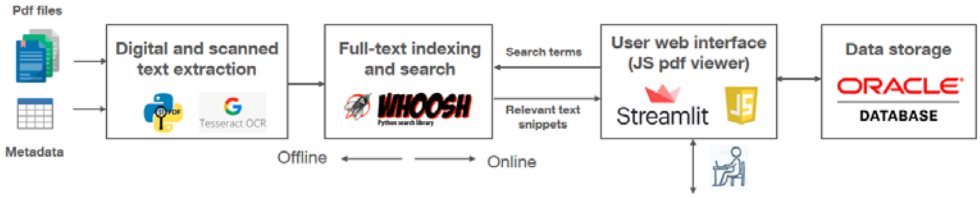


Figure 2 Main modules of the sustainability information retrieval and storage tool.

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

The offline processing part implements 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 through the interactive user web interface.

The user interface is a web application that allows end users to interact with all the tool functionalities in order to extract relevant information from a large collection of documents. Figure 3 shows the main view of the user web interface, which incorporates multiple filters and selectors, advanced search and storage capabilities and a control panel to track data ingestion.

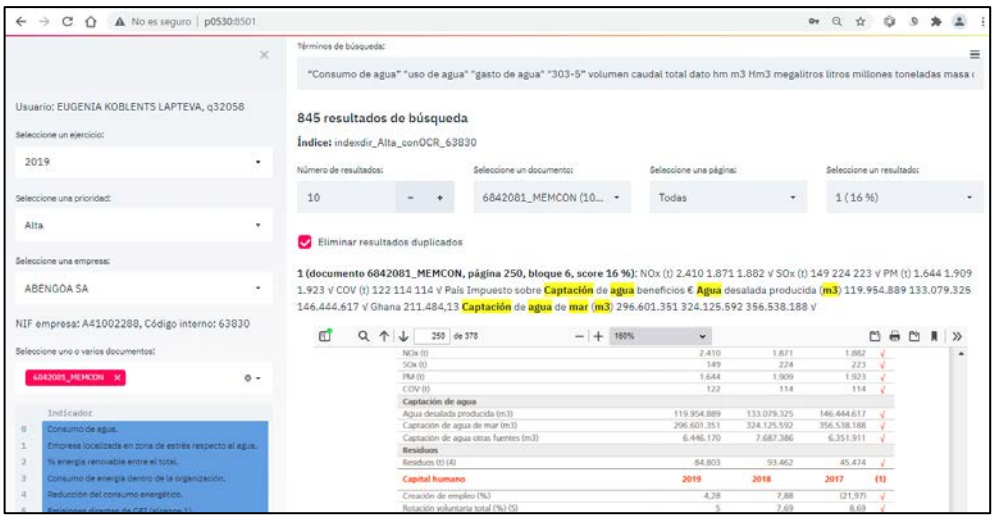


Figure 3 Main view of the user web interface.

Once all documents have been pre-processed and indexed in an offline processing phase, users can perform online searches and storage 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 those selections. In particular, an index associated to the company selected and a pre-defined list of search terms are loaded. 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 that which contains the indicator of interest. Lastly, they fill in a form with the information related to that indicator that needs to be stored. Additionally, the tool automatically stores complete context information on 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 next subsections.

Input data description

The main source of input data for the tool developed is a collection of pdf documents reported by Spanish non-financial companies. 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 Optical Character Recognition (OCR) process. Figure 4 shows an example of text fragments in digital format (left) and text contained in a scanned image (right). Over one thousand documents (6GB) have been processed to date. Only documents in the Spanish language have been considered so far; multilingual processing has not been addressed.

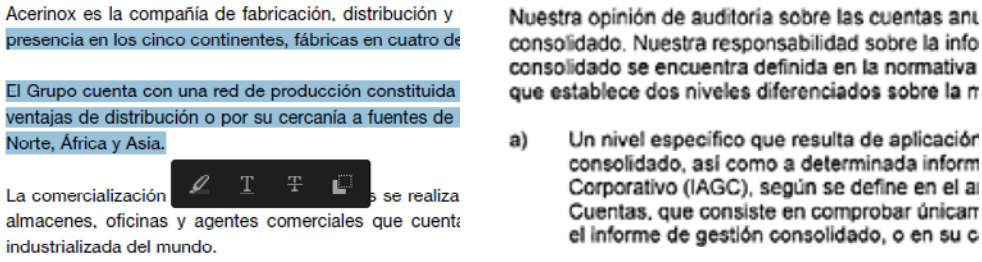


Figure 4 Examples of text in digital format (left) and text contained in scanned images (right).

The metadata related to the collection of documents and reporting companies are also required for the tool to operate. In particular, the metadata matrix contains a unique document identifier, type and date, as well as a company name and identifier, among other additional information. The metadata matrix is updated every time a new document is indexed and is currently stored on file together with the corresponding documents.

Using the input data sources described, the text extraction and indexing modules generate an index search for each company, which is also stored on file with a unique identifier. Lastly, a taxonomy of search terms needs to be defined for each sustainability indicator of interest; it 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 collection of documents, which will later be indexed by the indexing module. The module inputs are the collection of documents and the metadata file. Documents usually contain both text in digital format and text contained in scanned images, two text data sources that need to be processed in different ways.

Text in digital format can be readily extracted from documents and usually contains no errors. In this work, the Python library pymupdf was used to extract digital text from documents. This library enables extraction of text in multiple formats, including plain text, blocks, words, json, xml, and other formats. JSON format was preferred since it contains rich format information (text location, font and font size, etc.) additionally to the text content itself. Figure 5 shows an example of digital text extraction with pymupdf, where the output JSON structure contains rich information describing the extracted text, including the text bounding box coordinates, font and font size, colour, etc.

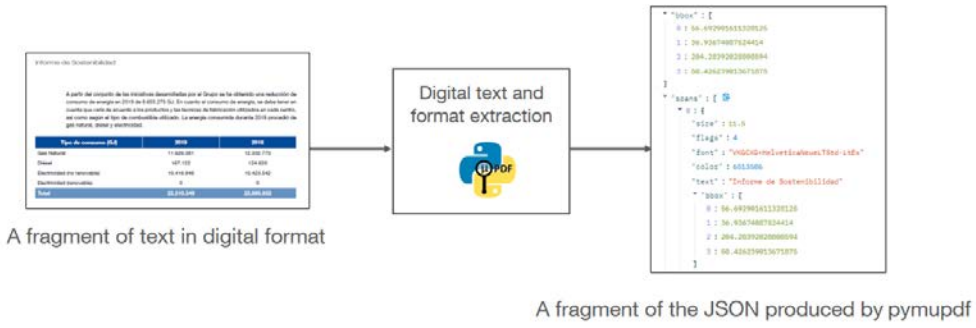


Figure 5 Digital text extraction with pymupdf.

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 before being processed with an OCR system. In this work, the Tesseract OCR engine, implemented in the Python library pytesseract, was applied to all images in order to extract scanned text present in the documents. Figure 6 shows the scanned text extraction workflow process.

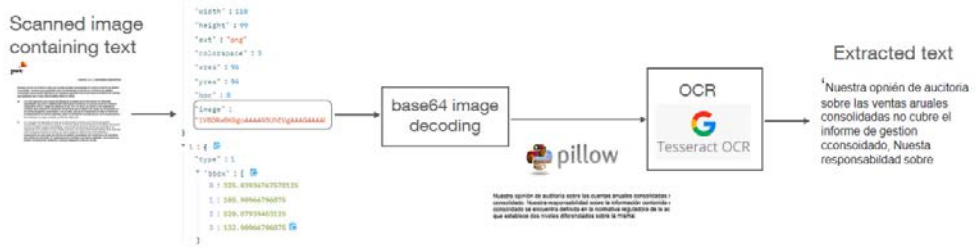


Figure 6 Scanned text extraction with pymupdf, base64 image decoding and Tesseract OCR.

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 implements fuzzy search, which may also be used as an alternative technique to account for errors in the extracted text. However, in this work, no error correction tools were used, since they can

potentially induce errors in correct but non-standard terms, such as company names and other words.

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

Full-text indexing and search

The indexing and search module has been developed using the Python library whoosh, which provides a fast, full-text indexing and search engine implementation in pure Python, suitable for moderate-sized document databases. Full-text search engines allow for quick searches of high 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 in which fragments each term appears.

In this work, digital and scanned text was indexed in blocks of more than 50 words, sorted by y-coordinate. An index schema and a language analyser including a tokenizer, a stemmer, a stop-word filter, etc., had to be defined. For scalability and robustness reasons, an index was 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 user web interface and the search indices generated.

To standardise the search terminology, a taxonomy of search terms was 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 7 shows the section of the user interface dedicated to the search of the sustainability indicators. The first selector allows the user to choose a sustainability indicator from a pre-defined list and a short description is loaded beside it in an expandable box. 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 them into the search bar using the tool as a standard full-text search engine. The pre-defined search query can be restored by clicking on the corresponding button.

The tool automatically searches the index associated with the selected company for the terms listed in the search bar. The total number of search results and the index name are shown below 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 computed by the whoosh library. In particular, the BM-25 scoring algorithm was used in this work. Each search result is listed together with the document name, page, text block and relevance score.

The tool automatically eliminates duplicated search results, when multiple copies of some 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.

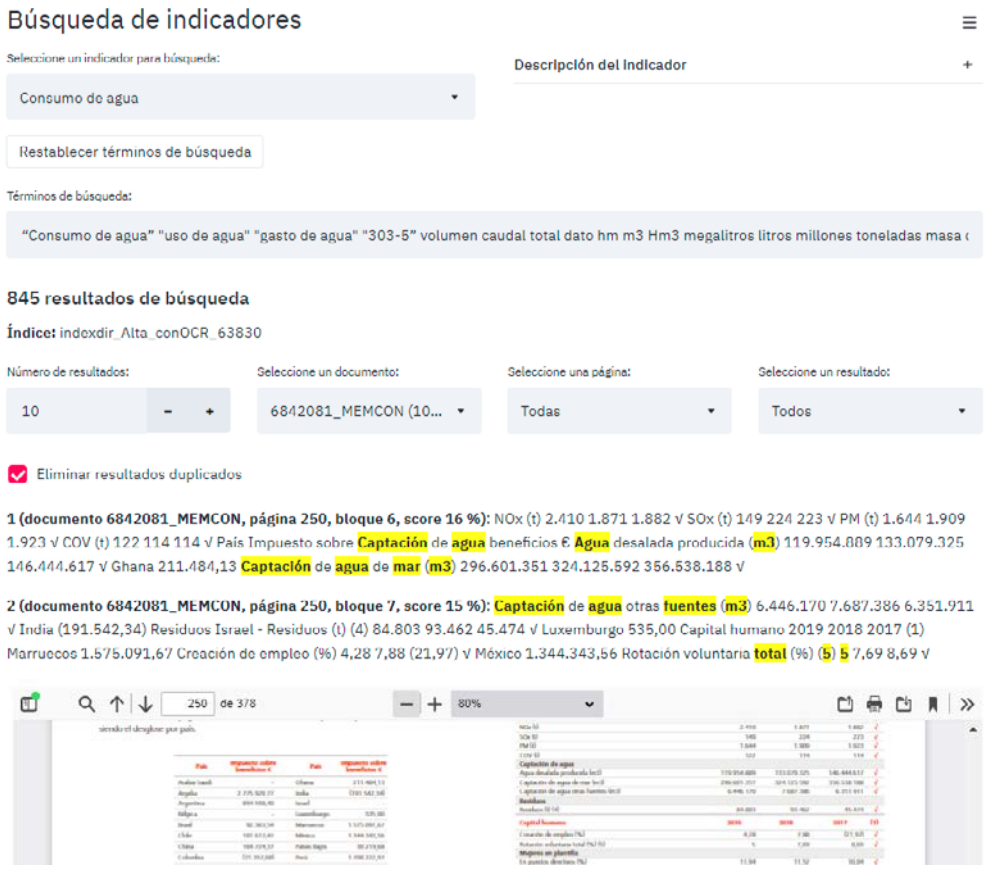


Figure 7 User interface section dedicated to search of sustainability indicators.

Data storage

A relational database was used to store the new database for two reasons. First, a relational database is the natural way to store structured information extracted from a collection of documents, allowing the new database to be merged with other relevant data, such as financial company data (activity, sales, employees, etc.) in order to generate meaningful and complete statistics.

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

The database should be as normalised as possible. In this case the primary key is formed by the variables "indicator", "year", "company", "country", "document year" and a Boolean feature indicating whether the data filled in is consolidated or not.

Access from the web app and to the database is by means of the python library cx_oracle, as can be seen in Figure 8. Thus, the operator fills in the data to be stored from a range of different values, in some cases, or directly by typing in some text, and the data are stored in the database via SQL commands through the 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 metrics, etc. Moreover, a backup table provides for traceability of insertion and deletion of indicators in the database and information on the user who made those modifications.

Provisionally, the authentication provided by the database has been used to authenticate in the web app. In the next steps other specific software such as Kerberos will be used for this purpose.

When the operator has found the correct results in the document and is ready to introduce the information into the database, there are a fixed range of parameters to choose from. Only a few of the fields are free text. The operator can also monitor the indicators that are left in the table in the left-hand margin (in red) and can check the results introduced in the table below and modify the data where necessary. These details can be seen in Figure 9.

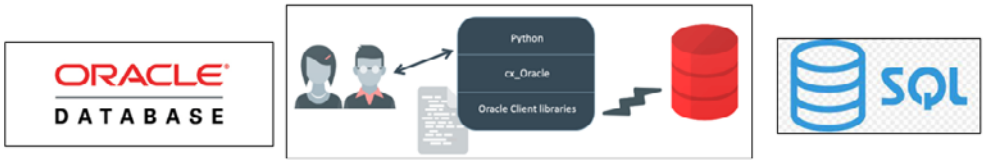


Figure 8 Communication from the web app to the database.

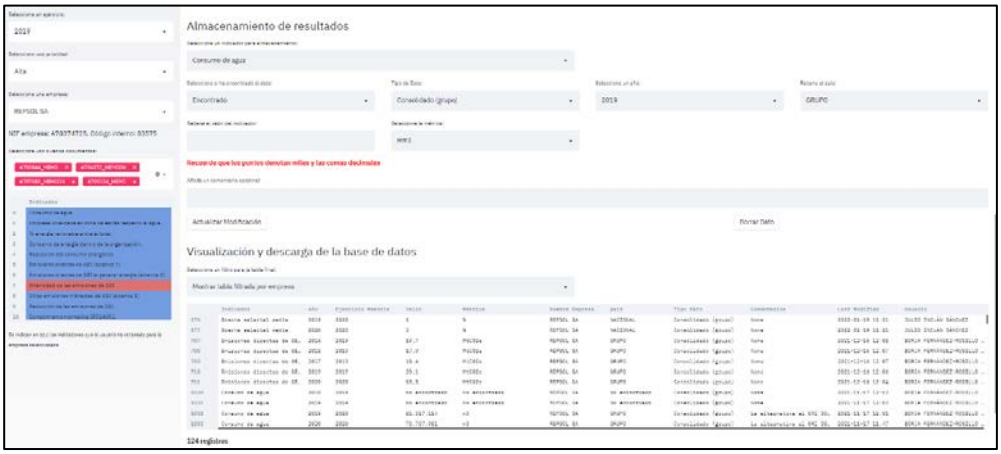


Figure 9 User interface section dedicated to storage.

User interface

The user web interface is one of the most important modules in this project. As previously mentioned, the process of transferring the information of indicators from pdf documents to a structured database is semi-automated, i.e. it needs a business user who validates the indicator candidates that the searcher finds and chooses the correct one (if any), giving it the appropriate format and also storing it in the database via the web app.

The web app incorporates a pdf viewer where the operator checks the raw data about to be stored. In some cases, the operator might find the indicator on a different page from the one the searcher reported. This may be because the values and descriptions are on different pages (e.g. the values may be on the following page, within an image or in an annex or table).

Other functionalities satisfied by the application are a control panel to track data insertion and detect which companies are pending completion by each worker, and other filters and selectors to navigate within the range of documents and companies. Also, the web app allows users to delete records directly, instead of having to write SQL sentences.

The user web interface has been temporarily implemented in Streamlit and deployed on local Banco de España Statistics Department machines. Streamlit is an easy-to-use open source app framework for data science visualisation and web development. This software will be migrated to the python framework Dash and deployed on a corporate server.

A deeper look into the previously mentioned modules in the app can be seen in Figure 10 (authentication), Figure 11 (search), Figure 12 (pdf viewer) and Figure 13 (storage).

Figure 10 Users have to introduce their credentials to log into the app.

Figure 11 For each indicator, users can customise the search terms. They can then navigate through the search results and choose the correct one.

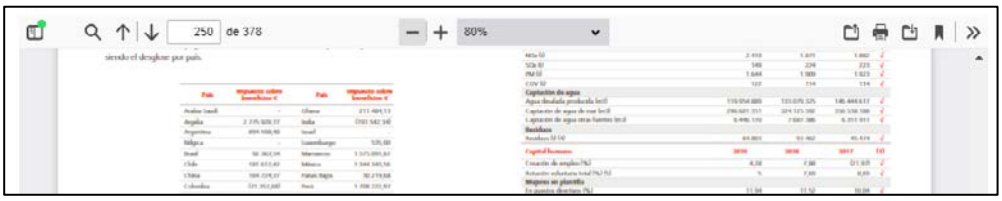


Figure 12 The pdf navigator is important to verify at source that the result found in the previous step is correct.

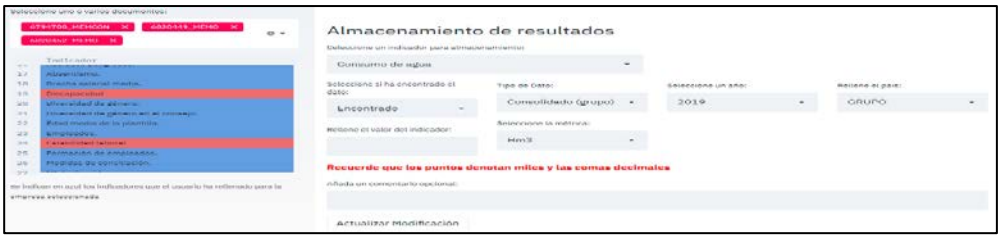


Figure 13 All the information is sent to the database in a fixed and structured format.

3. First data ingestion

Approximately 20 users worked part-time during two months in 2021 to make the first data ingestion. The 39 indicators were required to be filled in for 139 corporate groups. A total of 515 documents were available for these groups (making an average of approximately 2.9 documents per entity). Some 10,500 records were stored in this first data ingestion and 4,500 more in a second batch, making approximately 4,500 rows.

In 73% of cases the indicator was found for the given company, compared with 27% where it was not found. Listed groups provided more information than unlisted ones: 81% of the records belonging to listed companies were found. Social indicators were found for 80% of the data searched, whereas only 66% of the environmental indicators were found. 75% of government data was found. The documents were from 2019 and 2020, although some contain information from earlier years. This explains why historical series have been achieved for some indicators.

Some reports have been made and shown to the public. These reports are mainly dashboards made by PowerBI, as can be seen in Figure 14.

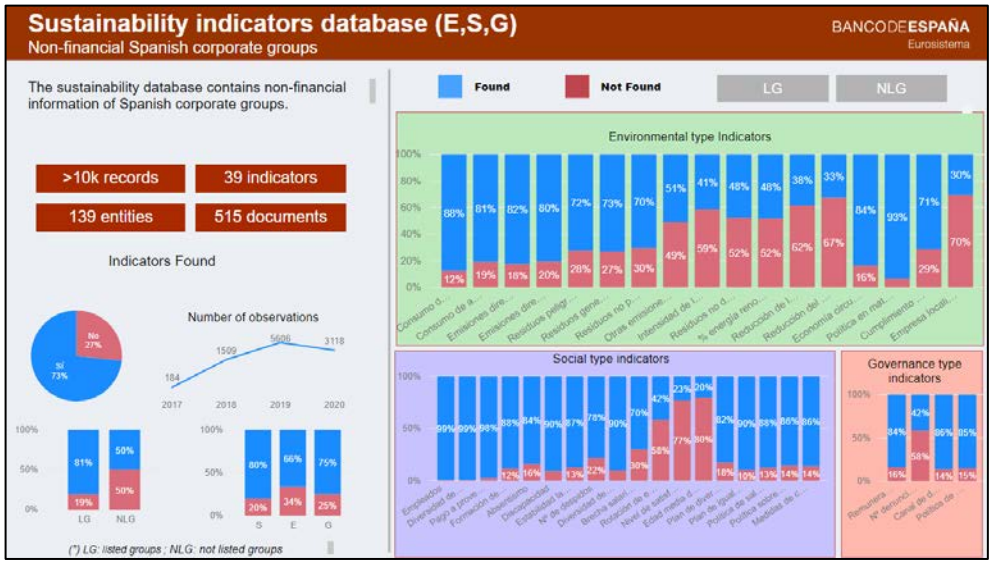


Figure 14 Several dashboards have been made to show the effectiveness of the product.

4. Conclusions

A web application has been developed and deployed for the creation of a new database of sustainability information from large collections of documents (Spanish corporate reports). This information has been obtained using a semi-automated approach that reduces the human effort required thanks to the search tools included.

The prototype was implemented by two data scientists in the Banco de España's Statistics Department, with the support of domain experts for requirements definition, testing, etc. The prototype is planned to be productionised with the help of the IT Department.

The tool incorporates authentication, data filters and selectors, full-text search in multiple documents, data storage, deletion and download and a control panel to track data insertion. The app allows the operators to customise the search terms (although default terms are given for each indicator).

Next steps

The project for storing companies' sustainability information is a long-term one as new regulations are arising and will arise in different legal frameworks. This means it is an ongoing project which will have to adapt to new documents, regulations, etc.

The work performed so far is a complete prototype prepared to be taken to production. Real data has already been introduced using the full life-cycle of the web app. Nevertheless, some lines of work are open for the short, medium and long term:

- **Database extension**: Currently, the database has been populated with information retrieved from documents of listed companies and consolidated groups. The plan is to expand this to other enterprises, and also to include financial entities.
- **Improvement of ontology using NLP**: Together with the above-mentioned 10,500 records, additional context information, such as the search terms, the location of the selected search result in the document, and its textual content has also been stored. NLP techniques could be used to automatically optimize the set of search terms for each indicator based on that context information.
- **Database publication within the Banco de España**: The database is accessible to the CBSO operators, but it has not been made available to Banco de España staff in general or to the public. The plan is for it to be released within the Banco de España environment and to external researchers within the Banco de España's data laboratory (BELab [9]) in the short term.
- **Migration to dash**: Dash is the most common Python tool for visualisation and is the official software for web development in Python at the Banco de España. For this reason, the current Streamlit prototype will be migrated to Dash in the near future.
- **Migration to production servers**: The app has been deployed on local servers in the Statistics Department but it will soon be migrated to official production servers.

- **Exploration of alternative OCR systems:** The IT Department is currently exploring the UIpath system, which provides different OCR engines (Tesseract, Omnipage, etc.). Currently, the library pytesseract is the package used in this prototype to extract text from image formats, but UIpath might be used in the future.

References

[1] Bingler, Julia Anna, Mathias Kraus and Markus Leippold. "Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures." Available at SSRN (2021).

[2] Friederich, David et al. "Automated Identification of Climate Risk Disclosures in Annual Corporate Reports." arXiv preprint arXiv:2108.01415 (2021).

[3] Luccioni, Alexandra and Hector Palacios. "Using Natural Language Processing to Analyze Financial Climate Disclosures." *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*. 2019.

[4] Luccioni, Alexandra, Emily Baylor and Nicolas Duchene. "Analyzing Sustainability Reports Using Natural Language Processing." arXiv preprint arXiv:2011.08073 (2020).

[5] Moreno, Ángel Iván and Teresa Caminero. "Application of Text Mining to the Analysis of Climate-Related Disclosures." (2020).

[6] Moreno, Ángel Iván and Teresa Caminero. "Analysis of ESG Disclosures in Pillar 3 Reports. A Text Mining Approach". Future Finance (forthcoming 2022).

[7] Tarquinio, Lara, Domenico Raucci and Roberto Benedetti. "An Investigation of Global Reporting Initiative Performance Indicators in Corporate Sustainability Reports: Greek, Italian and Spanish Evidence." Sustainability, 2018.

[8] Webersinke, Nicolas et al. "ClimateBert: A Pretrained Language Model for Climate-Related Text." arXiv preprint arXiv:2110.12010 (2021).

[9] https://www.bde.es/bde/en/areas/analisis-economi/otros/que-es-belab/.

BANCO DE ESPAÑA
Eurosistema

A WEB APPLICATION PROTOTYPE TO RETRIEVE AND STORE SUSTAINABILITY INFORMATION FROM UNSTRUCTURED TEXT

Alejandro Morales (alejandro.morales@bde.es)
Eugenia Koblents (eugenia.koblents@bde.es)

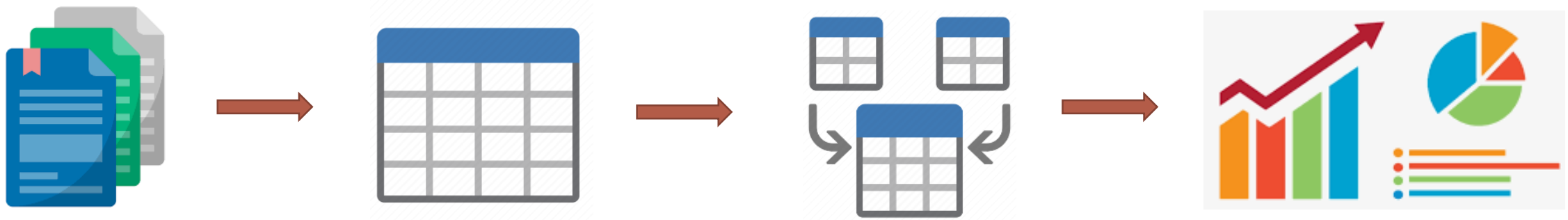
IFC-BANK OF ITALY WORKSHOP ON DATA SCIENCE IN CENTRAL BANKING
PART 2: DATA SCIENCE IN CENTRAL BANKING: APPLICATIONS AND TOOLS

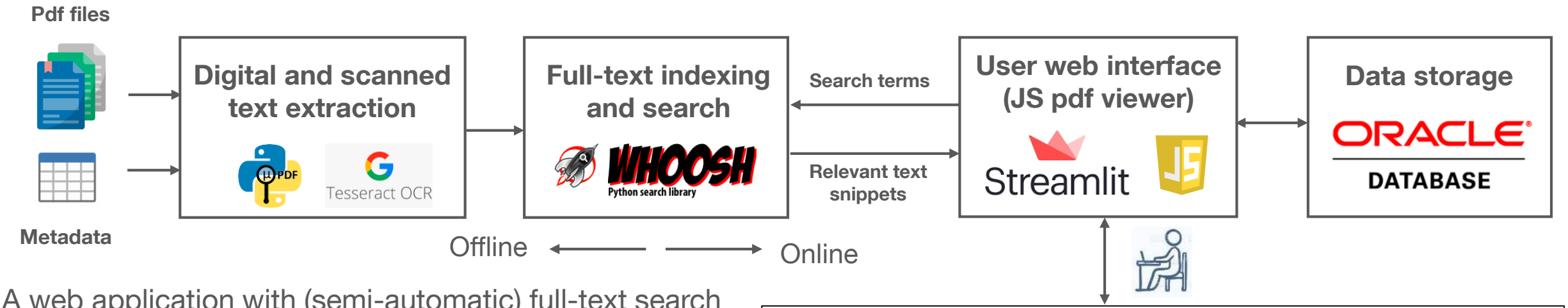
17/02/2022

STATISTICS DEPARTMENT

INDEX

- The goal of this project is the creation of a new database containing microdata information on sustainability extracted from **Spanish companies** reports in order to address **climate change**.
- The target information includes **39 environmental** (energy, water, emissions, residues, etc), **social** (employees, age, gender, etc) and **governance** (corruption, bribery, complaints, etc) **indicators**.
- To date, this information is only mandatory for **large corporate groups** but EU regulation is adapting.
- The reported information is presented in **highly heterogeneous formats**: plain text, tables, graphics and images. A broad variety of **metrics** are used for each indicator. **Manual extraction** is very costly.
- **Technical goal**: transform **non-structured information** (documents) into a **structured database** (table) to **merge** it with additional information and generate **statistics** on sustainability.
- **Project time span**: April-July 2021. Next phase planned for first semester of 2022.

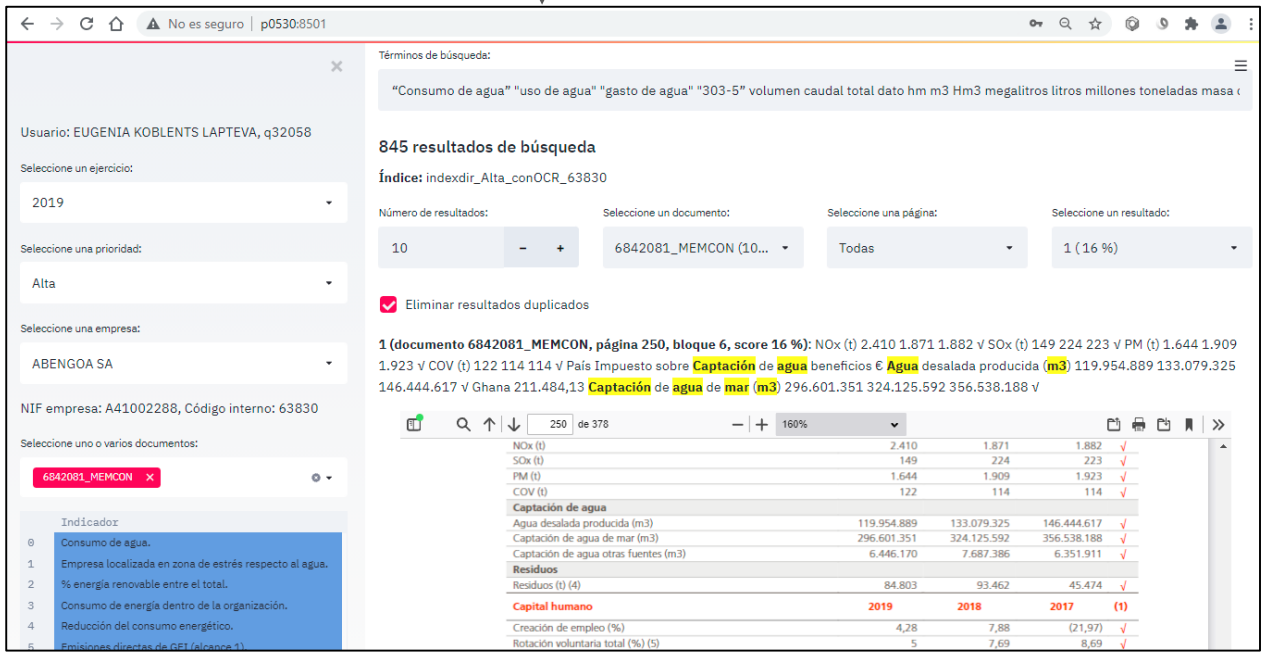




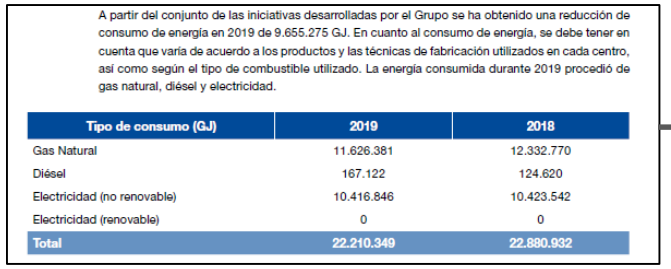
A web application with (semi-automatic) full-text search and storage capabilities has been developed in **Python**:

1. The user selects a year, company and indicator.
2. The **tool searches** for pre-defined terms and retrieves a sorted list of relevant text snippets.
3. The **user validates** the search results and **stores** the indicator value into the database.
4. Complete **context information** is also stored (location, surrounding text, etc.).

Streamlit tool temporarily deployed on a local machine, not supported by IT. Migration to **dash** and final **deployment** on a corporate web server planned for 2022.



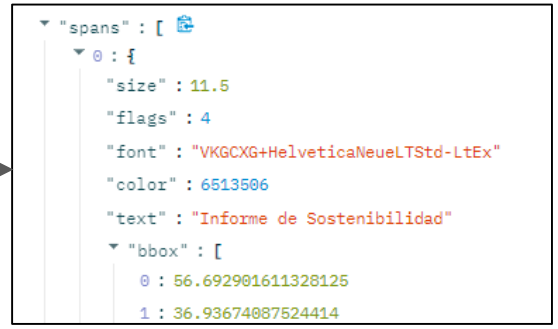
- **Input data:** Highly heterogeneous documents of variable length are available for each company, including text in **digital format** (readily extractable) and text contained in **scanned images** (requiring OCR).
- Text in **digital format** is readily extractable and (usually) contains **no errors**. The **pymupdf** library allows to extract digital text in **JSON** format, containing **rich format information** (text location, font, size, etc).



A partir del conjunto de las iniciativas desarrolladas por el Grupo se ha obtenido una reducción de consumo de energía en 2019 de 9.655.275 GJ. En cuanto al consumo de energía, se debe tener en cuenta que varía de acuerdo a los productos y las técnicas de fabricación utilizados en cada centro, así como según el tipo de combustible utilizado. La energía consumida durante 2019 procedió de gas natural, diésel y electricidad.

Tipo de consumo (GJ)	2019	2018
Gas Natural	11.626.381	12.332.770
Diésel	167.122	124.620
Electricidad (no renovable)	10.416.846	10.423.542
Electricidad (renovable)	0	0
Total	22.210.349	22.880.932

A fragment of text in digital format

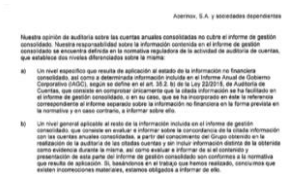


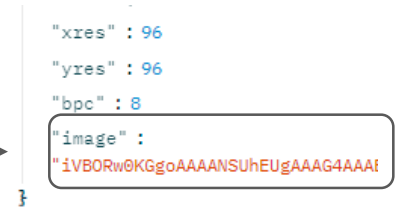
```
"spans" : [
  0 : {
    "size" : 11.5
    "flags" : 4
    "font" : "VKGCXG+HelveticaNeueLTStd-LtEx"
    "color" : 6513506
    "text" : "Informe de Sostenibilidad"
    "bbox" : [
      0 : 56.692901611328125
      1 : 36.93674087524414
```

A fragment of the pymupdf JSON

- **(Scanned) images** appear as base64 encoded strings in the pymupdf JSON and need to be decoded and converted into PIL images to be fed into pytesseract OCR. The quality of scanned images if often low, yielding OCR errors.

Scanned text





```
"xres" : 96
"yres" : 96
"bpc" : 8
"image" :
"iVBORw0KGgoAAAANSUhEUgAAAG4AAAA
}
```

Extracted text

‘Nuestra opnién de aucitoria sobre las ventas aruales consolidadas no cubre el informe de gestion

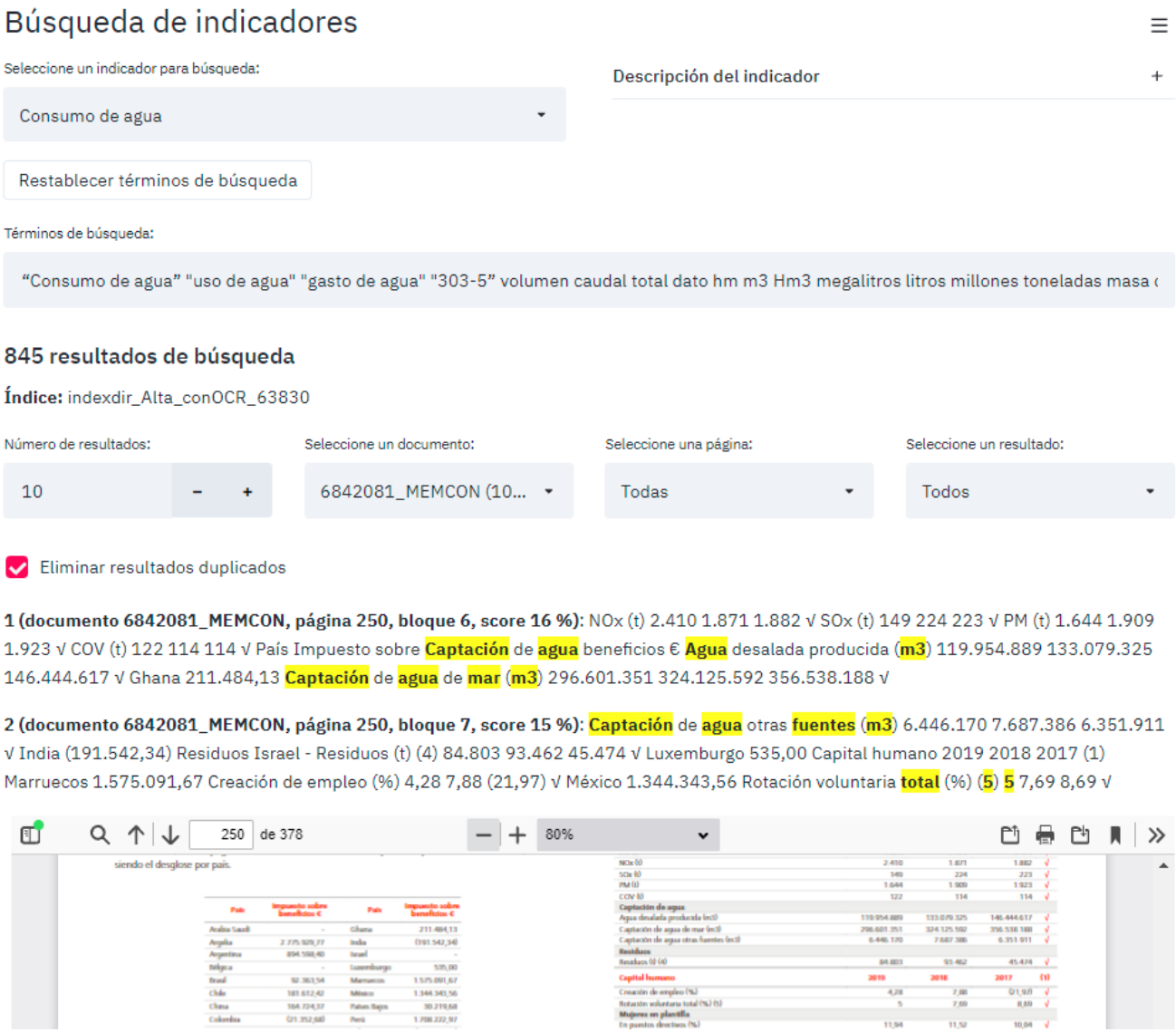
MAIN MODULES

Full-text indexing and search

- Digital and scanned text is indexed (offline) in **blocks of >50 words**. An **index** per company is created.
- A **language analyser** (tokenizer, stemmer, stop-word filter, etc) and an **index schema** need to be defined.
- A set of **search terms** has been defined for each indicator based on **expert knowledge**.
- The tool retrieves **relevant text snippets** in real time.
- The user can **filter results** by document and page.
- When a search result is selected, the corresponding text snippet is shown on a JavaScript **pdf viewer.**
- The tool can also be used as a standard **full-text search engine** by manually typing the search query.

WHOOSH Python search library

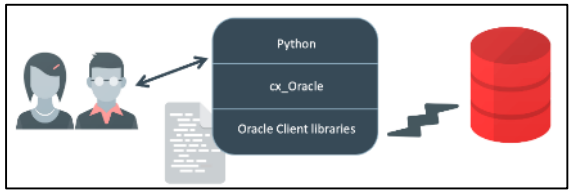
Streamlit



MAIN MODULES

Data storage and user interface

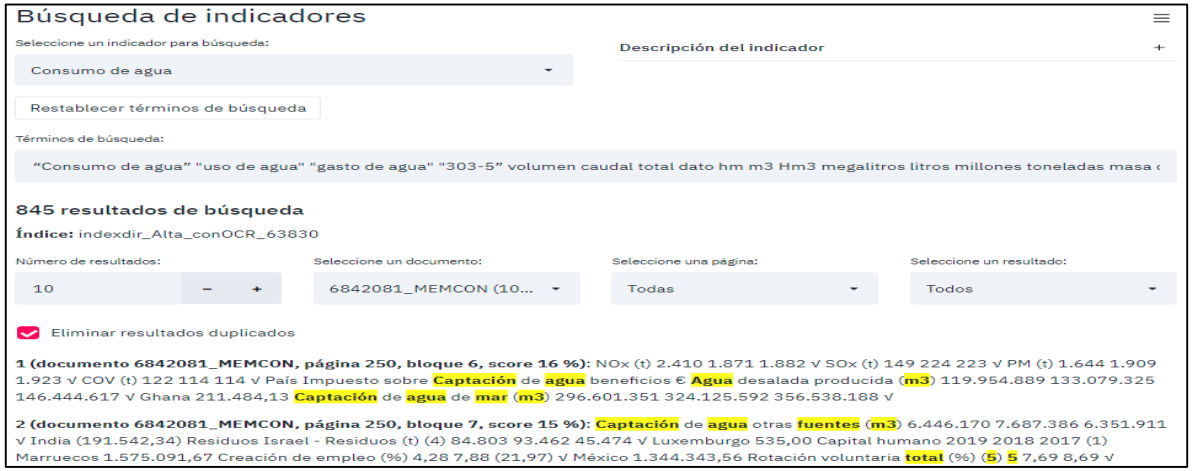
- The tool allows the user to **store** the value of the indicator in an **Oracle database**, together with additional information (metric, year, comments, etc). Connector: library **cx_Oracle** in python.
- Rich **context information** is also stored automatically (search terms, text snippets, user ID, date and time, etc).
- A **control panel** is automatically updated when new data is inserted into the database.
- The user interface allows to **download** the full as well as a filtered version of the dataset.
- More than 20 workers were required to work in this project and fill data into the database, good **simultaneous performance** has been a key requirement.
- The **authentication** in the database is used as registration in the interface.
- Backup and change **traceability** stored. Other intermediate tables are also kept in the database.



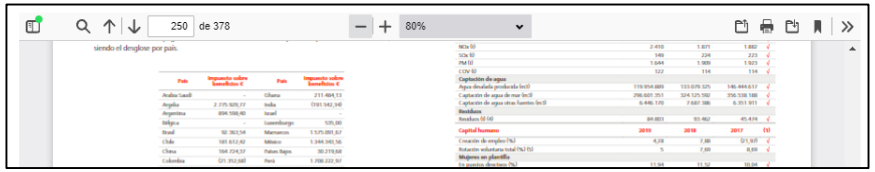
- The user web interface has been temporarily implemented in streamlit and deployed on a local machine.
- **Streamlit** is an easy to use open-source app framework for data science visualization and web development.
- The tool incorporates integrated authentication, multiple filters and selectors, **full-text search, data storage** and a control panel to track the insertion of data.
- The user interface will soon be migrated to **dash** and deployed on a corporate **web server**.

Authentication

Search

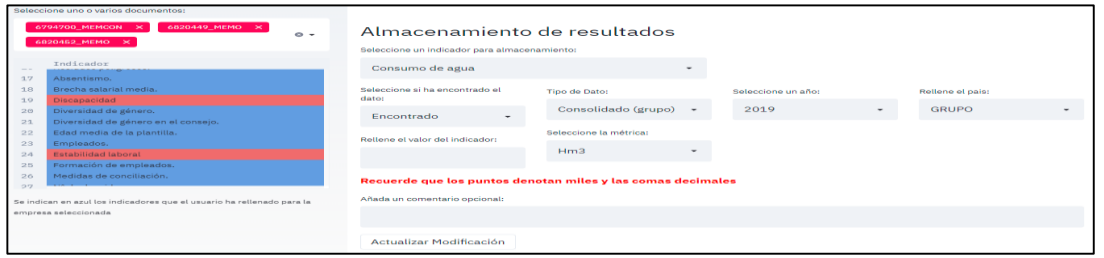


Document check



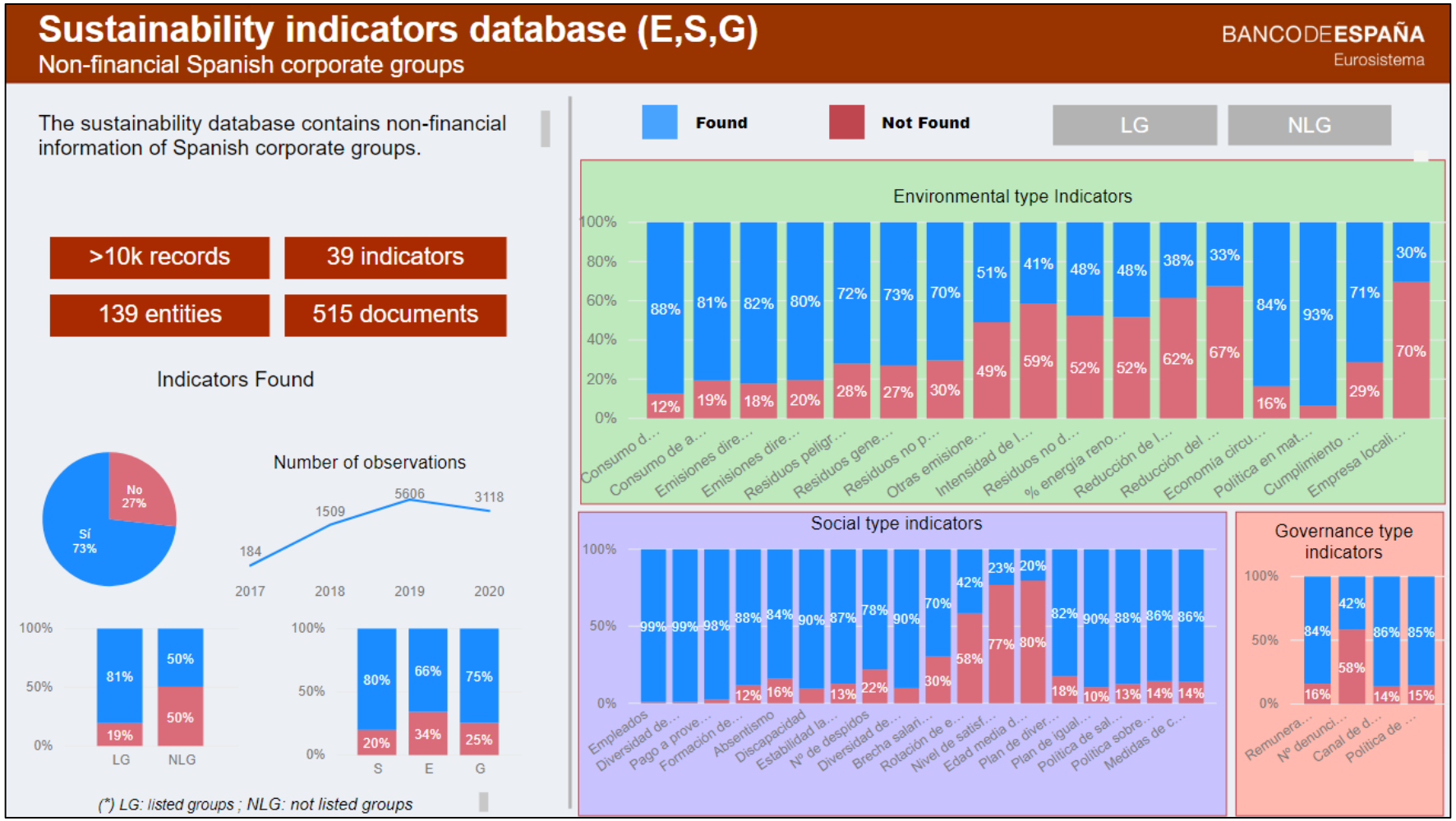
Streamlit

Storage



- 20 users were working part-time during 2 months to make the first data ingestion:
  - **> 10,000 records**
  - **39 indicators**
  - **139 companies**
  - **515 documents**

- Listed companies provided more percentage of successfully found indicators.
- Historical series have been obtained for some indicators.



CONCLUSIONS AND NEXT STEPS

Conclusions:

- A **web application** has been developed and deployed for creating a new database of sustainability information from large collections of documents (Spanish corporate groups reports).
- The tool incorporates authentication, **full-text search**, **data storage** and a control panel to track the insertion of data. Prototype implemented in 3 months by two **data scientists** with the support of **domain experts** for requirements definition, testing, etc.
- The current prototype is **not supported by the IT Department** but has been implemented and deployed using temporary tools.
- The tool allows to **standardize the search terms** (each user makes the same searches by default).
- **Context information stored** (search terms, search results, etc.) will allow to train ML methods to optimize the search process.
- A **semi-automatic approach** reduces the human effort in populating the new database.

Next steps:

- **Database extension** to include other enterprises and financial entities: Currently, only listed companies and consolidated groups.
- **Improvement of ontology using NLP and machine learning.**
- **Database publication within the Bank of Spain**: The database is only accessible by the operators of the CBSO at the moment.
- **Migration to dash**: dash is the most common Python tool for visualization and is the official software for web development in Python in the Bank of Spain. The user interface will soon be migrated to **dash** enhancing its functionality.
- **Migration to corporate production servers**: So far the app has been deployed on local machines in the Statistics Department.
- Exploration of **alternative OCR systems**.

BANCO DE ESPAÑA
Eurosistema

Thank you for your attention!

STATISTICS DEPARTMENT