Kalpa Publications in Computing
Volume 15, 2023, Pages 1–14
Proceedings of The 2022 International Conference on
Digital Transformation and Innovation Technology

Data use as a tool for regulation and decision-making in communications services: An academy-state innovation experience

Jhon Garcia[1], Jose Suárez-Jurado[2], David Niño-Torres[3], Omar Roa[4], Ana Barbosa[5], Juan Plazas-Hernández[6], Catalina García-Acevedo[7], Julieth López-Castiblanco[8] and Ingrid Picón[9]
[1] Lab101, Universidad National de Colombia, Bogotá, Colombia
[2] Comisión de regulación de comunicaciones, Bogotá, Colombia
jhagarciaca@unal.edu.co, jdsuarezj@unal.edu.co, jdninot@unal.edu.co,
oeroaq@unal.edu.co, ambarbosac@unal.edu.co, juan.plazas@crcom.gov.co,
acgarciaa@unal.edu.co, julalopezcas@unal.edu.co, ingrid.picon@crcom.gov.co

Abstract

Colombia is a country with significant development, but with a great distance to go regarding Digital Transformation (DT) in companies and in the public sector, especially if one examines the deep lag that it has compared to other Latin American countries. This article digs into the illustrative case and searches for DT in the Colombian public and government sector of the *Comisión de Regulación de Comunicaciones* -CRC- (Communications Regulation Commission). During 2020 and 2021, CRC led the creation of two tools: the first, a rate and plan comparator (services offered by Internet providers, mobile and fixed telephony in Colombia) that was created in partnership with the *Laboratorio de innovación, creatividad y nuevas tecnologías* -LAB101- (Laboratory

[1] LAB101 leader

[2] Back-end developer of the plan comparator project

[3] Created the first draft of this document

[4] Back-end and front-end developer coordinator of the plan comparator project

[5] Statistics LAB101 leader

[6] CRC executive director

[7] Product leader of the plan comparator project

[8] Created the first draft of this document and made the adjustments to the template

[9] CRC developer of AnalIsa project

of Innovation, Creativity and new technologies) of the National University of Colombia (*Universidad Nacional de Colombia* -UNAL-). The comparator provided CRC an agile and automated deep understanding of the market for communication services in the country. Although the comparator was initially conceived to meet the operational needs of CRC, Data Analytics and Web Development, through Web Scraping and Machine Learning techniques, it allowed the construction of an interactive platform that was released and, nowadays, facilitates the decision-making of Colombian citizens. The second tool was a legal comparator called AnalIsa -created by lawyers, economists and engineers from CRC- which seeks to compile current regulations and administrative acts to promote consultation and legal decision-making by CRC. AnalIsa was created following Artificial Intelligence guidelines and Agile Methodologies within the SCRUM framework. Its target audience are CRC employees, suppliers and consumers who are impacted by the legal decisions of the Commission.

The two comparators -developed under the guidance of CRC- supported the idea that one of the innovation principles of DT is the use and appropriation of contemporary inputs for the solution of problems of different interest groups and stakeholders. In this case, the public sector, citizens, suppliers and academia.

Keywords: Innovation, Creativity, Regulation, Web Scraping, Machine Learning, Data Analysis, Digital Transformation, Artificial Intelligence, document comparison, Automation, Optimization

1 Introduction

1.1 Colombia and the digital transformation: potentialities, limitations, and needs

The Colombian government and the business sector -in tune with global dynamics and cutting-edge technology- project a progressive annual growth of digital transformation (DT) scenarios in the country and a deep penetration in various sectors of the national economy [1]. However, there are multiple barriers to make this possible, such as an inadequate or heterogeneous business structure or culture, a lack of DT strategies and return on investment and visibility, and even the perception of a "cannibalization" or disappearance of existing businesses due to negative impacts of DT [2]; in addition to external and social barriers, such as lack of recognition on how DT can improve the society, insufficient infrastructure, inadequate regulations and the lack of resources, especially for small and medium enterprises [2].

Colombia is a country that is still at an early stage of DT; compared to other countries in the region such as Chile or Perú, small, medium, and large Colombian companies are relegated in terms of digital maturity and only startups manage to have a high score, an advanced state in the digital maturity index proposed by the consulting firm Virtus Partners and its report published in 2021 [3]. Furthermore, in this same report, the consulting firm constructs a score from 1 to 100 to evaluate 6 dimensions of DT in Colombia, with 1 being the lowest possible score and 100, the highest possible score. Dimensions such as Data and Analytics have a score of 29.4 and Processes, technology, and digital operations 35.8, concluding that there is a precarious level of digitization (analog maturity level).

According to the data collected by the report, the use of data analytics for business management is less than half of the organizations that took part in the study, while the use of advanced data analytics for segmentation, prediction, optimization, and recommendation is distributed as follows in the business sector: 49% for large companies, 35% for SMEs and 29% for startups [3]. Here, a drastic decrease is evident between large, consolidated sectors, and entrepreneurs.

Under a critical context of constant review and assessment, and following the guidelines of the OCDE in 2019 [1] about a digital government (DGI), the Colombian State has promoted the DT of its entities for some years, initially issuing public policies such as CONPES 3975 of 2019 [4], as a framework that directs the digital transformation and modernization, thus seeking to move to a useful and efficient State, in tune with the citizenship, responding to the dynamics of contemporary times. After all, the Colombian State is positioned under the position of two possible options in relation to industries, innovation and technology; in the first scenario, if companies keep a standard model, they tend to maintain a traditional, closed and operational business in a single market; in a second option, if companies adopt a disruptive model, they tend to open additional opportunities to capitalize on new markets and innovate in technologies, processes, products or services. This second scenario advocates the need for DT for innovation and points out that one of the main objectives of DT is to improve the accessibility and quality of digital services offered to the population [1].

This article seeks to address a work experience located in the research and efforts of the Colombian State in order to generate transformation in its institutions, overcome stages of analog maturity or digitization, to reach a stage of DT, which benefits both the entity internally, as well as the growth of the institution itself. Thus, the following section starts by addressing the concept of DT, highlighting that to achieve it, it is necessary the conjunction of different profiles and tools, and conclusively, wrups up with a brief description of two very specific initiatives that benefited different groups, both based on developments from data analytics.

1.2 The case of the CRC: An experience of joint work and contributions for the DT in the Colombian public sector

The principle of innovation in DT processes is the use and appropriation of contemporary inputs for the solution of problems of different stakeholders [5]. By “contemporary inputs" we refer to a set of tools such as emerging technologies [6], contemporary research methodologies, information capture and analysis [7], and the meeting of different professionals from different areas of knowledge. When we mention the stakeholder category, we mean that DT is, above all, a human process [8] whose essential beneficiaries are precisely the people belonging to different social groups, such as institutions belonging to the public or private sector, non-governmental entities, consolidated companies, commercial enterprises, communities, collectives, and cultural groups, among many other examples that are part of a universe of possibilities.

This article focuses on DT conceived for meeting the particular needs of a Colombian public entity: the *Comisión de Regulación de Comunicaciones* -CRC- (Communications Regulation Commission). For this purpose, it describes two specific tools. The first one is the Internet, telephony and Pay television services comparator, which grants CRC a quick reading on the behavior of the market providers of these services in Colombia, while allowing users to compare and choose the best fit services according to their needs. This comparator was developed thanks to the articulation between academia and the State, i.e. the collaborative partnership between the CRC and the *Laboratorio de innovación, creatividad y nuevas tecnologías* -LAB101- (Laboratory of Innovation, Creativity and new technologies) of the National University of Colombia (*Universidad Nacional de Colombia* -UNAL-).

The second tool is the legal data comparator called *AnalIsA*, developed by CRC. This comparator was conceived and created thanks to the articulated work of two coordinations of CRC: the coordination of Legal Counsel and Dispute Resolution and the coordination of Information Technologies and Systems. AnalIsA comparator performs text analysis, by comparing documents based on CRC administrative acts and regulations in force. The main purpose is to deliver a result that allows members of CRC to make decisions based on existing documents and thus, avoid legal damage, inaccuracy, contradiction or lack of completeness in the issuance of a new regulation or decisional acts issued by the entity. This tool facilitates the investigative work and decision-making of CRC's legal advisors.

By mentioning both tools and exploring the evolution that allowed their creation, we seek to emphasize the importance of interdisciplinary processes that enhance digital transformation for public and private sector companies. While it is important to highlight the presence of fourth-generation technologies, techniques, and tools or emerging technologies such as Web Scraping, Machine Learning, Artificial Intelligence, specific web development, all addressed in this text, the biggest richness of the production of both comparators mentioned is the articulation of different fields of knowledge, including law, economics, statistics, graphic design and finally web development. All of them foster structural changes and provide tools for the consolidation of a culture of innovation within a company or entity, such as CRC.

The following sections detail the methodology in terms of data analysis, software development and design that allowed the creation of the two tools. All this work was carried out under the leadership of CRC in 2020 and 2021. Finally, the description of the concrete results will be addressed: the functionalities, modules and filters that ensure the usability and interactivity features within the AnalIsa Comparator and the Internet, telephony and Pay television services comparator. In addition, it concludes with a series of analyses of the impact of these tools and also proposes recommendations oriented to future processes of similar purpose and scope.

2 Methodology

2.1 CRC needs: Problem Statement

As Colombia's regulator of communication services, CRC seeks to have detailed and updated information on the behavior of this specific market: the services offered by providers and their relationship with users. In addition, CRC analyzes this information in order to ensure the protection and welfare of users, in terms of a fair relationship between the quality of services, diversity of plans and prices. In this regard, in 2016, CRC requested operators to fill out Format 1.2 [9]-a Microsoft Excel database, delivered to the entity every three months-. It has 65 fields that indicate, in a thorough and complete manner, all the characteristics of the services offered. This information, although detailed, is overwhelming both for the receiving entity, taking into account the number of operators that report (national coverage operators and local operators); and for the providers who must build these records manually, filling out this information box by box at least 4 times a year. The communication reports between CRC and operators state: "the agents of the sector have been submitting observations regarding the difficulties that arise [sic] in the uploading and reporting of the information in Format 1.2, a situation that hinders the possibility of carrying out a proper monitoring of the provision of fixed and mobile telephony and internet services"[10].

The CRC's need to read and understand the market is clear and the 65 fields of the Format 1.2 are crucial, however, due to the amount of information generated, it is imperative to transform the way in which the data is captured and the mechanisms that organize it and facilitate its analysis. It is precisely at this point where CRC is committed to create innovative and automated strategies that even reduce the periodicity of data collection (from quarterly to daily data collection), allowing a more detailed historical reading without the need to generate administrative and operational burdens in any institution or company.

A similar need is the origin and significance of the AnalIsa legal comparator. The Digital Government policy regulated in Decree 1008 of 2018 [11] establishes the general guidelines for the use and exploitation of information and communication technologies within State entities. For this reason, CRC in its Regulatory Agenda for 2020-2021 [12] incorporated in the Innovation axis, encourage the use of emerging technologies that provide support in problem-solving, contribute to risk mitigation in the management of its processes through the proper management of information, data quality and thus can have greater confidence in strategic decision making in the Entity; proposals aligned to the National Development Plan 2018-2022 issued by Law 1955 of 2019 in its Article 147 "Public Digital Transformation" and to what is established in CONPES 3975 [4] "National Policy for Digital Transformation and Artificial Intelligence" to increase the generation of social and economic value from the use of digital technologies in the public and private sector, which considers the following actions:

- The reduction of the barriers that prevent the incorporation of digital technologies in the public and private sectors is related to the lack of culture and lack of knowledge of these technologies.
- The grant of enabling conditions through international partnerships for innovation, design, and implementation of initiatives to promote entrepreneurship and digital transformation.
- The strengthening of the competencies of human capital to face the 4RI (4th Industrial Revolution) by favoring the development of digital competencies during the educational trajectory and the configuration of innovation ecosystems through international partnerships for the training of talent with priority in AI.
- The increase of the enabling conditions to prepare Colombia for the economic and social changes brought about by AI.
- The promotion of other 4RI technologies by fostering the development of digital technologies, the creation of regulatory testing environments, funding for AI research and technological development, among others.

With respect to the subject previously mentioned, the use of intelligent tools provides solutions aimed at improving the provision of services and delivery of products to users and citizens, redesigning processes or creating new solutions that avoid repetitive tasks, improve the availability of quality data, encourage the use of information technology systems such as artificial intelligence and Machine Learning, so that they can copy human behavior and through algorithms, the machine learns through training and solves problems in an agile and effective way to improve strategic business processes.

Moreover, the use of mechanisms supported by artificial intelligence reduces costs and times in the management of processes, minimizing errors and with the possibility of reaching a higher degree of accuracy, as well as providing high-speed information analysis, and among others, allows finding knowledge in the data, predicting situations in the medium and long term.

2.2 Development of the Internet, telephony and Pay television plans comparator

2020, Data analysis, definition, and testing of comparable data - key guidelines for the comparator

The plan comparator was developed between 2020 and 2021 through the joint work of CRC and LAB101 UNAL. Its initial phase, in 2020, focused on the analysis of data and information held by CRC at that time on the internet, mobile, and fixed telephony service operators. All this information was collected between 2016 and 2020 in CRC’s Format 1.2. The initial phase also focused on reviewing the mechanisms and ways in which this information was obtained, evaluating its efficiency, its timeliness, and above all, proposing mechanisms for the automation of information collection.

In the first instance, this was achieved, thanks to data analytics and a descriptive statistical perspective on the information provided by CRC; and secondly, data collection by the LAB101 UNAL team using the Web Scraping technique. This technique was used to navigate the documentation containing relevant information but finding and accuracy required deep searches or extensive reading and organization work.

The work carried out in 2020 can be summarized as follows: first, data analysis through the collection of two samples associated with Format 1.2. These samples were used to identify relevant and necessary variables for use in subsequent processes. After this identification, the analytical team proceeded to perform data cleaning under a data mining perspective, that is: eliminating duplicate records, standardizing and normalizing the base values, reconstructing variables with new values, and data imputation. Data mining involves evaluating, correcting, and unifying values, monitoring their consistency and concordance with a data dictionary, in this case, provided by CRC. For the data debugging process, the Python [13] programming language was used in a Jupyter [14] environment, making use of the different libraries to manipulate the DataFrame.

The second step consisted of the development of a particular scraper to complement the data analysis through web scraping and thus feed the specialized PostgreSQL database [15] and facilitate the statistical analysis. Finally, the team proceeded to the creation of Grafana dashboards [16] that allowed us to concisely visualize the achieved analyses. Within Grafana, summary statistics and time series are implemented with the different segregations necessary and relevant to consult.

While the 2020 results were illustrative and a step forward in the understanding and potential of the data, Grafana graphs could be used as a reporting document that needed to be periodically and manually updated as well as CRC data collection forms. In other words, while the data, graphs, and analysis were quite robust, the level of automation and interactivity of the tool was low. In 2020, all the data analyzed were taken from offline documentary sources not amenable to automated updating.

2021, potentialities of Web development guided by the implemented Data Analytics model

Thus, in 2021 the project focused on the use of the information obtained, a new proposal for information sources, and new use of the web scraping technique within the project so that searches could be carried out online within the pages of the communications service providers. This required a tripartite consultation process between CRC, LAB101 team, and the providers so that, through a standardized model of tags in HTML language [17], they could standardize their web pages and thus enable the process of collecting and indexing information [18], it should be noted that what we wanted to standardize was the development language of each site and never the information or the appearance of the websites for the users.

In this phase of the project, the use of web scraping focused on the programming of a scraper -a bot-whose main objective is to collect, from the providers' web page, information related to characterization (e.g. prices, number of minutes, number of GB, coverage, among others), without the need for them to report this information to CRC in excel formats or documents that could only be consulted through downloads. The scraper was programmed with the ability to feed a precise database, where it organizes, in a matter of milliseconds, all the necessary information to make the comparison of plans and rates, which only a millisecond later would be visible on a web platform and would build the necessary reports for the exercise of regulation and understanding of the services market by CRC. The scraper based its programming on the guidelines of the data analysis model created in 2020, its main difference lies in the ability to focus on online data that can be updated every 24 hours and thus detect promotions, special events, or any behavior by an operator that would indicate a sudden change in the conditions of the services it offers.

In technical terms, the scraper was programmed as follows: the team used the Python programming language to manipulate the data and the BeautifulSoup [19] and Selenium [20] libraries were used as Frameworks that guided the programming and navigation of the scraper within the HTML language and, to give greater flexibility to the capture model, the information was stored in the unstructured database MongoBD [21].

It is at this point that the platform, organization, analysis, interactivity, and usability were conceived as a tool that should not only be in the hands of CRC, internally but could be published and open to the public so that all Colombians could know detailed and updated information on the services available in the country and thus facilitate the choice of a provider.

The structure of the comparator was designed in such a way that it allows the user -as a citizen or as a company- to make optimal filtering to obtain customized results in which he can compare the plans he wants by supplying the details and specifications. As a basis for the construction of this component, NestJS [22] and NuxtJS [23] were used as NodeJS Framework [24] for building web pages based on the model-view-controller architectural pattern, making use of user-side HTML rendering. In like manner, TypeORM [25] and Mongoose [26] were used as the data manipulation tools between the base Framework and its respective database engine.

Finally, regarding the deployment of the framework, the comparator was based on the Docker container system [27] within a host operating system, which provides the flexibility to use different technologies for the implementation of the various components. This system was designed to have an ecosystem that takes into account DDoS security parameters.

Deep learning models

In the last stage of the project, LAB101 team proposed to develop a Machine Learning (ML) model to achieve forecasting within the communication services market by applying emerging technologies. The potential of this model lies in the fact that CRC, in addition to having historical and descriptive information, would have prospective information on the behavior of operators. For this purpose, the data organized in the analysis made in the first stage of the project was used and a major cleaning and one hot coding were carried out, taking into account only the variables related to impact and performance. Subsequently, it analyzed the data to observe trends and thus build the forecasting models, taking into account descriptive statistics and ML parameters.

In the case of ML, the team took advantage of the fact that the data obtained are time series and applied recurrent network models associated with neural models, to perform sequence analysis to extract contextual information by defining the dependencies between timestamps [28]. The models used were RNN and its variations: LTSM, long-term memory units, which allow greater accuracy in handling large data sets and varying the predictive model according to the nature of the data; and GRU, closed recurrent units, with fewer training parameters, lower memory usage and therefore shorter execution time [29]. These variations were used because they can maintain long-term interrelationships and also nonlinear dynamics, which are frequent in economic data [30].

In implementing these models: two types of errors were considered to evaluate the accuracy of the models and to guide a readjustment of these: the first, MAE, which measures the average magnitude of the errors in the set of predictions, without considering their direction; this was used to validate the accuracy of the implemented models. The second, MSE, is the average of the squared differences between the prediction and the actual observation. It was found that this error gave more importance to the most significant errors of the set.

2.3 AnalÍsa Comparator Development

To maintain a common thread of legality and technical-legal coherence in the resolutions issued by CRC, traditionally a manual consultation work must be done through the use of standard Microsoft Office tools such as Microsoft Word and summary tables of concepts in Microsoft Excel, which is an arduous search and study work for the employees of the entity responsible for such work. Besides, it was also necessary to call on the memory of officials who had been familiar with these issues in the past.

Well aware of the benefits offered by Artificial Intelligence (AI) and machine learning (ML) technologies, mentioned above, in 2020 the Entity developed a proof of concept and a minimum viable product with a third party, which applies these technologies through free-to-use algorithms, This allowed CRC to materialize the implementation of the software tool called "AnalIsA", which performs text analytics through the comparison of documents, focused on compliance with the Prevention of Antilegal Damage Policy proposed by the entity and approved by the National Agency for the Legal Defense of the State -ANDJE, by its Spanish initials-.

Likewise, in 2021 the functionalities of AnalIsA were increased, linking other mission processes such as Regulatory Design and Audiovisual Content and some features to facilitate the monitoring and incorporation of new documents through batch processing to the document base, thus obtaining a complete, robust, and highly capable tool to meet the text analytics CRC needs.

At the functional level, the Coordinator of Legal Counsel and Dispute Resolution and the Functional Leader with the specific knowledge to implement the Anti-legal Damage Prevention Policy of the Communications Regulation Commission applicable for the 2020-2022 period, before the National Agency of Legal Defense of the State -ANDJE-. Likewise, for the construction of the data dictionaries, work tables were held with an interdisciplinary team composed of lawyers, engineers, and economists.

At the technical level, the Coordinator of Information Technologies and Systems and the Manager and technical leader of the Project on behalf of CRC for the direction and definition of the phases and execution of the project. Likewise, for the implementation of the AnalIsA tool, a third party was hired to implement the project, with expertise and technical knowledge for the development of emerging technology projects in the Colombian State.

This tool was conceived and developed taking into account the best practices and guidelines of Digital Government for the incorporation of emerging technologies in Colombian State entities, as well as the transversality of the Digital and Information Security component.

It is a web solution implemented under the framework of agile software development -SCRUM [31]- and emerging technologies such as Artificial Intelligence and Machine Learning, whose Frontend was developed in React [32], on an Apache Web server and Java technology [33] for the user interface and a Backend in Wildfly [34] with code in Python [13] and R [35] for analytics and learning algorithms.

3 Results

3.1 The Case of the Plan Comaparator

Data Analytics. Information debugging

About 65 variables with 8122003 unduplicated records were identified for the first sample, while 66 variables with 4629475 unduplicated records were found for the second sample. For both samples, about 40% of the fields were found to be empty, with less than 11% of erroneous data (data associated with out-of-range numeric fields). Samples contained numeric, categorical, and geographic location data. Regarding the visualizations, the different dashboards created present a common structure that includes a dashboard description: dashboard summary, filters with required and expected segregations, dashboard update button, and drop-down sections with detailed graphs depending on the type of service consulted.

The Plan Comparator

There were 175 operator web pages exposed to consumer users of which 28 were viable for implementing the web scraping system with adequate data collection. These web pages presented two types of base specifications: one is implemented with sample information on a single page, and the other presents interactive components where an interactivity bot must be used to access the information. A proposal for each type of page was presented in an implementation guide that was given to the suppliers to standardize their web pages. After this, the web platform was created and scaled by the regulatory entity for its use in subscription TV services, so that the comparator contains plans for mobile telephony, mobile internet, fixed internet, fixed telephony, and subscription TV. The web portal was launched to the public on March 23, 2022. It is available at: https://comparador.crcom.gov.co/

The page has a single-use option for companies and another for users, for the latter it allows different filters depending on the services that the user wishes to review to obtain more personalized results and compare the plans of interest. In the case of selecting more than one service, the duo, triple or multiple options will be marked depending on the user's choice. The comparator has a coverage filter that is used in fixed services where the user can indicate the location and stratum, otherwise, the user will skip this part and go directly to the results.

Lastly, in each tab the user has two options: the user can see more details of each plan of interest or can select different plans and use the comparator table to see their characteristics in parallel, here it is possible to choose up to 4 plans. The page -in its web and mobile version- is useful for citizens or companies to find the plan that best suits their preferences within the different service operators.

Deep learning models

The data used for the predictive modeling of demand in the telecommunications market in Colombia were those extracted from the reports of Format 1.2 with quarterly frequency and those captured with Web Scraping:

- F1_2 (7.5 Million records, Size: 13.8 GB): contains only Active plans from 2017 through Quarter 1, 2021, of Format 1.2.
- WS (84 Thousand records, Size: 23.6 MB): contains the sample execution of the Web Scraping algorithm, without information on the number of subscribers (demand), since it has not yet been possible to capture this variable.

The data model for Format 1.2 and WS shares the same structure where the number of subscribers is the dependent variable to be modeled and represents the market demand for communication plans and where some of the independent variables were: start date since the product was offered, price, provider, type and modality of the plan, targeted stratum and download and upload speed of the fixed Internet.

Two types of errors were examined, MAE and MSE, and determined that MSE gave much more importance to the most significant errors, while MAE, by giving equal importance to each error, allowed a more intuitive perception for the evaluations. Of the different models evaluated through iterative training, the most effective hyperparameter configuration (with the lowest loss) was chosen according to the validation set.

It can be observed in the results that the predictions obtained from the autonomous learning, present loss errors lower or very close to 1, so that the models manage to provide reliable predictions of the demand of the plans according to the variability of the behavior of the historical supply/demand. In addition, it was found that the market price of each type of service was the most influential variable in explaining the behavior of telecommunications demand in Colombia. This statement supports the theoretical relationship between prices and demand for each good or service in the economy.

3.2 Results AnalIsa tool

Taking as a reference for innovation in public management, and using emerging technology tools, such as Machine Learning algorithms and data science, to facilitate the work of the Entity's collaborators and data-based decision making. The result is the construction of this software tool for text analytics that consists of several modules, which meet the needs of users, among which are:

- “Buscar"/Search module: This module is the heart of the tool and allows searching with the following parameters:
- Text search: Search for a text within the system's document base, to find out which documents in that document base match the indicated text.
- Document Search: Compare a document with those already in the document base to establish which of those documents the requested document is related to.
- Category Search: Search documents within the document base according to topics previously defined in the data dictionaries.
- “Cargar”/Upload module: It allows sending documents and indexing them to the tool's document base. The incorporation of new documents can be done in the following way:
- Through the "Upload" option: Users can include documents one by one to the document base.
- Through the "Bulk Upload" option: Documents previously uploaded to a shared folder in the Entity's file server by the AnalIsA system can be automatically included.
- “OCR" module: This allows the user to take a document (in PDF or possibly scanned image) and generate a plain text version, by means of an OCR (Optical Character Recognition) tool. This option is valuable to have greater control over the incorporation of new documents in terms of quality and completeness.
- “Dictionary" module: This allows administrator users to upload a Microsoft Excel file containing the data dictionary through an established template so that the topics or categories required for text analytics can be incorporated or modified in AnalIsA.
- “History" module: Allows to review and consult the searches, indexing of new documents, and OCR executed in the tool by users.
- “Administration" Module: Allows performing administrative review tasks such as: - Modify the topics that appear to the user to make searches.
  - Review indexing results in Backend.
  - Review the tasks performed by all users.

4 Conclusions

4.1 CRC needs and potential benefits to multiple stakeholders

In the case of the Plan Comparator, it is important to highlight how data analytics and web development made it possible to build solutions that provided benefits to various stakeholders at different levels. First, it contributed to the solution of an internal operational need of CRC, which wanted to simplify and automate the capture of information from each of the service operators regarding the characteristics of the internet, mobile, and fixed telephony plans. CRC sought to facilitate its reading exercise and its role as regulator and protector of the users of communications services.

Secondly, it reduced the administrative burden on operators who had to report changes in their services to CRC on a quarterly basis and almost manually through an Excel format. The comparator made it possible to digitally transform and automate the delivery of this data for CRC's clarity on a daily basis, freeing them from the obligation to create the Excel format or the periodicity of delivery, but generating new responsibilities around the development of web pages, specifically in HTML language to facilitate the capture of information, encouraging a DT process also in the operators.

Thirdly, the comparator benefited users of communications services, as they could make comparisons between operators and even consult regional companies with less experience compared to traditional service providers who have greater access to the media to advertise. This tool facilitates purchasing and consumption decisions by users of the Internet, mobile telephony, and pay television services throughout the country.

A fourth agent benefited in the project that allowed the generation of the plan comparator was the academia, represented in the team of LAB101, who through research exercises, tests and comparisons managed not only to theorize but also to put into practice and service the knowledge achieved, giving an applied character to everything discussed and consulted.

4.2 Methodological inputs for DT and sustainability of comparators

In the case of the Plan Comparator, it was found that the vast majority of operators have their information grouped in card-type elements, which makes the maintainability of the scripts developed viable in the medium term.

Despite the above, the monitoring of mobile and fixed services is useful for CRC professionals as it allows the visualization of many of the variables of Format 1.2, the temporal analysis, and the variation of the offer. In addition, it is of social importance since it facilitates information on mobile and fixed services to end consumers.

In the case of AnalIsa comparator, this tool has allowed the development of a cognitive ecosystem that leverages emerging technologies, strengthening the digital transformation process of the Entity and making public management more efficient and compliance with the Policy for the Prevention of Antilegal Damage of the Communications Regulation Commission (CRC) applicable before the National Agency of Legal Defense of the State (ANDJE), being a national reference in the incorporation of emerging technologies and digital transformation in the Legal context of Public Entities. The main points that facilitate the work of the collaborators are:

- Strengthens knowledge management by providing a reliable and easy-to-consult information base.
- Avoids the risk and improves the induction and knowledge transfer process due to personnel rotation and the hiring of new collaborators.
- Establishes an agile and reliable way to perform data analysis tasks on a knowledge bank, strengthening the Regulatory Design, Audiovisual Content, and Dispute Resolution

mission processes, relieving the operational burden, and allowing employees to focus on more strategic activities in the mission.

The creation of a complete and appropriate document base allowed the tool to make a reliable comparison of the documents under study, without the need for human knowledge or expertise. However, it was a challenge to achieve the use and appropriation of the tool by the collaborators of the Coordination of Legal Counsel and Dispute Resolution by formalizing it in the dynamics of their daily work within the process.

References

[1] «Índice de Gobierno Digital OCDE 2019. Resultados y mensajes clave», Organización para la cooperación y el desarrollo económico OCDE, 2019. [En línea]. Disponible en: https://www.oecd.org/gov/digital-government/digital-government-index-2019-highlights-es.pdf [2] C. Ebert y C. H. C. Duarte, «Digital Transformation», IEEE Softw., vol. 35, n.º 4, pp. 16-21, jul. 2018, doi: 10.1109/MS.2018.2801537.

[3] J. J. De la Torre et al., «La transformación digital en Colombia y su índice de madurez», Consultora Virtus Partners, Estudio, 2021. Accedido: 13 de abril de 2022. [En línea]. Disponible en: https://www.cesa.edu.co/news/transformacion-digital-en-colombia-indice-de-madurez-digital-delas-empresas/

[4] Departamento de Planeación, Ministerio de Tecnologías de la Información y Comunicaciones, y Departamento Administrativo de la Presidencia de la República, POLÍTICA NACIONAL PARA

LA TRANSFORMACIÓN DIGITAL E INTELIGENCIA ARTIFICIAL - (Documento CONPES 3975). 2019. [En línea]. Disponible en: https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3975.pdf

[5] M. A. G. Castrillón y A. I. Mares, «Innovación empresarial, difusión, definiciones y tipología: una revisión de literatura», Dimens. Empres., vol. 11, n.º 1, pp. 45-60, 2013.

[6] J. I. Aguaded Gómez y J. Cabero Almenara, «Avances y retos en la promoción de la innovación didáctica con las tecnologías emergentes e interactivas», Educar, 2014, doi: 10.5565/rev/educar.691.

[7] E. B. Ramírez, C. W. G. Estrella, y S. K. S. Gárate, «La inteligencia de negocios y la analítica de datos en los procesos empresariales», Rev. Científica Sist. E Informática, vol. 1, n.º 2, Art. n.º 2, jul. 2021, doi: 10.51252/rcsi.v1i2.167.

[8] M. S. Ramírez-Montoya, «Innovación abierta, interdisciplinaria y colaborativa para formar en sustentabilidad energética a través de MOOCs e investigación educativa», Open, interdisciplinary and collaborative innovation to train in Energy Sustainability through MOOCs and educational research, dic. 2018, Accedido: 19 de abril de 2022. [En línea]. Disponible en: https://repositorio.grial.eu/handle/grial/1467

[9] «RESOLUCIÓN No. 5050 DE 2016», Comisión de regulación de comunicaciones, 2016. Accedido:

30 de marzo de 2021. [En línea]. Disponible en: https://www.crcom.gov.co/sites/default/files/normatividad/Compilada_2016_11_11.pdf

[10] «REVISIÓN DEL FORMATO 1.2. “TARIFAS Y SUSCRIPTORES DE PLANES INDIVIDUALES Y EMPAQUETADOS” DEL TÍTULO DE REPORTES DE INFORMACIÓN DE LA RESOLUCIÓN CRC 5050 DE 2016». agosto de 2020. [En línea]. Disponible en:

https://crcom.gov.co/system/files/Proyectos%20Comentarios/2000-71-15/Propuestas/documento_soporte%281%29.pdf

[11] «Decreto 1008 de 2018», MINISTERIO DE TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES, DEPARTAMENTO NACIONAL DE PLANEACIÓN, DEPARTAMENTO ADMINISTRATIVO DE LA FUNCIÓN PÚBLICA, jun. 2018. Accedido: 17 de abril de 2022. [En línea]. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=86902

[12] Z. C. Vargas Mesa, C. E. Lugo Silva, S. Martínez Medina, y J. P. Hernández Marcenaro, «Agenda Regulatoria 2020 - 2021 [Proyecto 5000-2020-1]», Comisión de Regulación de Comunicaciones, p. 36, dic. 2019.

[13] G. van Rossum, «Python reference manual», Art. n.º R 9525, ene. 1995, Accedido: 4 de abril de 2022. [En línea]. Disponible en: https://ir.cwi.nl/pub/5008

[14] T. Kluyver et al., «Jupyter Notebooks – a publishing format for reproducible computational workflows», 2016, pp. 87-90. doi: 10.3233/978-1-61499-649-1-87.

[15] P. G. D. Group, PostgreSQL 9.0 - Internals and Appendixes, vol. 5. Fultus Corporation, 2011.

[16] M. Chakraborty y A. P. Kundan, «Grafana», en Monitoring Cloud-Native Applications: Lead Agile Operations Confidently Using Open Source Software, M. Chakraborty y A. P. Kundan, Eds. Berkeley, CA: Apress, 2021, pp. 187-240. doi: 10.1007/978-1-4842-6888-9_6.

[17] K. Patel, «Incremental Journey for World Wide Web: Introduced with Web 1.0 to Recent Web 5.0 – A Survey Paper», ijarcsse, vol. Volume 3, p. 10, oct. 2013.

[18] S. Raj Mohan, «Scrape data from the web using Python and Watson Studio», IBM Developer, 5 de marzo de 2019. https://developer.ibm.com/tutorials/scrape-data-from-the-web-using-watsonstudio/ (accedido 6 de abril de 2022).

[19] L. Richardson, «Beautiful Soup Documentation [Release 4.4.0]». Beautiful-soup-4, 24 de diciembre de 2019. [En línea]. Disponible en: https://beautiful-soup-4.readthedocs.io/en/latest/#

[20] Harvard, «Selenium Documentation [Release 1.0]», 26 de julio de 2012. Accedido: 6 de abril de 2022. [En línea]. Disponible en: https://pdf4pro.com/docs/selenium-webdriver-1da98d.html

[21] N. O'Higgins, MongoDB and Python: Patterns and Processes for the Popular Documentoriented Database. O'Reilly Media, Inc., 2011.

[22] M. A. Alvares, «Manual de NestJS». desarrolloweb. Accedido: 6 de abril de 2022. [En línea]. Disponible en: https://desarrolloweb.com/manuales/manuales-nestjs

[23] NuxtJS, «The Intuitive Vue Framework», Nuxt. https://nuxtjs.org/ (accedido 6 de abril de 2022).

[24] L. M. Surhone, M. T. Tennoe, y S. F. Henssonow, Node.js. Beau Bassin, MUS: Betascript Publishing, 2010.

[25] Typeorm, «TypeORM - Amazing ORM for TypeScript and JavaScript». https://typeorm.io/ (accedido 6 de abril de 2022).

[26] A. Kurniawan, Object-Relational Mapping (ORM): MongoDB, Mongoosejs and Node.js By Example. PE Press.

[27] D. Merkel, «Docker: lightweight Linux containers for consistent development and deployment», Linux J., vol. 2014, n.º 239, p. 2:2, mar. 2014.

[28] B. Kumaraswamy, «6 - Neural networks for data classification», en Artificial Intelligence in Data Mining, D. Binu y B. R. Rajakumar, Eds. Academic Press, 2021, pp. 109-131. doi: 10.1016/B978-0-12-820601-0.00011-2.

[29] T. M. Navamani, «Chapter 7 - Efficient Deep Learning Approaches for Health Informatics», en Deep Learning and Parallel Computing Environment for Bioengineering Systems, A. K.

Sangaiah, Ed. Academic Press, 2019, pp. 123-137. doi: 10.1016/B978-0-12-816718-2.00014-2.

[30] O. Campesato, Artificial Intelligence, Machine Learning, and Deep Learning. Mercury Learning and Information, 2020.

[31] K. Schwaber, «SCRUM Development Process», en Business Object Design and Implementation, London, 1997, pp. 117-134. doi: 10.1007/978-1-4471-0947-1_11.

[32] F. Copes, The React Handbook. flaviocopes.

[33] D. Holmes y J. Gosling, THE Java^{TM} Programming Language, 4.ª ed. Addison-Wesley Professional, 2005. Accedido: 19 de abril de 2022. [En línea]. Disponible en: https://www.oreilly.com/library/view/the-javatm-programming/0321349806/

[34] L. Fugaro, WildFly Cookbook. Packt Publishing Ltd, 2015.

[35] R Core Team, «R: The R Project for Statistical Computing», vol. 3, p. 201, 1 de mayo de 2013.