Providing Data on Financial Results of Public Companies Enriched with Provenance for OBInvest

GILBERTO PASSOS*, SAULO ALMEIDA*, VALQUIRE JESUS*, JORGE ZAVALETA*, and SÉRGIO SERRA*, Universidade Federal do Rio de Janeiro, Brazil

Financial Literacy (FL) initiatives, aimed at young people in formal or informal learning spaces, are defended and implemented in several countries, being encouraged since 2005 by the Organization for Economic Co-operation and Development (OECD). In Brazil, the teaching and learning process in several areas has been stimulated through Academic Competitions generally called Knowledge Olympics, which are essentially student contests that aim to encourage, find talent and awaken interest in the field knowledge presented in the competition. It was precisely for this purpose that the Brazilian Investment Olympics (OBInvest) was born, aiming to democratize access to education and promote reflections on economic and financial issues, through a FL perspective for high school students from all over the country. One of OBInvest's objectives is to help boosting the development of computational tools, aiming to provide easier access to fundamental data for decision-making in the field of finance. However, from the tools developed by OBInvest, it was noted that the creation of new educational tools would be enhanced through the use of datasets enriched with provenance and aligned with FAIR principles. This work aims to offer a computational strategy based on data science techniques, which is easy to use and also provides curated data series through a reproducible pipeline, using open data on financial reports from publicly listed Brazilian companies, provided by the Brazilian Security and Exchange Commission, called Comissão de Valores Mobiliarios (CMV).

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Gilberto Passos, Saulo Almeida, Valquire Jesus, Jorge Zavaleta, and Sérgio Serra. 2018. Providing Data on Financial Results of Public Companies Enriched with Provenance for OBInvest. *J. ACM* 37, 4, Article 111 (August 2018), 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Financial Literacy (FL) initiatives, including those aimed at young people, both in school spaces and in non-formal teaching environments, are advocated and implemented in several countries [1], has been encouraged since 2005 by actions of the Organization for Economic Co-operation and Development (OECD). In this scenario, addressing financial and economic notions in the Basic Education Mathematics curriculum was advocated in the guiding plan of national

*All authors contributed equally to this research.

Authors' address: Gilberto Passos, gilbertopassos@ufrj.br; Saulo Almeida, sauloandrade@ufrj.br; Valquire Jesus, valquire@ufrj.br; Jorge Zavaleta, zavaleta@pet-si.ufrrj.br; Sérgio Serra, serra@ppgi.ufrj.br, Universidade Federal do Rio de Janeiro, P.O. Box 68.530, Rio de Janeiro, Rio de Janeiro, Brazil, 21941-590.

Manuscript submitted to ACM

education, especially with the recent inclusion of FL as a transversal and integrative theme in the National Common Curricular Base [2].

In Brazil, the Brazilian Investment Olympiad (OBIinvest) [16], comes up in August 2020, as an Extension Project of Federal Center for Technological Education (CEFET) [4], with the aim of democratizing access and promoting reflections on economic and financial issues and situations through a FL perspective for high school students throughout Brazil.

From a multidisciplinary perspective and considering the teaching and methodological aspects, OBInvest tries to attract students to think about hypothetical situations and to make decisions. These types of activities contribute to the development of skills and competences necessary for the creation of critical thinking, emancipatory and inclusive of the individual, being a guide for the full exercise of citizenship and also enabling the insertion of these young people in a new job market.

OBInvest tries to guide the development of computational systems and tools in order to provide easy access to important and fundamental data for decision-making in financial field. However, it was noticed that the development of new educational tools could be enhanced with the use of datasets enriched with provenance [18] and aligned with the FAIR principles [10], which could add more reliability to decisions based on financial data and bring less risk. It is estimated that these two resources could also be used in banking and financial services.

This work aims to offer a computational strategy that is easy to use, that is reproducible and uses curated data series that can improve the development of skills and competences of young talents interested in Finance and Investments both in enriching the national activities of the Brazilian Investment Olympiad, such as professional preparation and enhancements for future financial certifications.

1.1 Work Contributions

As a contribution, we developed a tool called DRE-CVM capable of executing reproducible pipelines and using curated, fairified, and annotated data with the retrospective source metadata of the financial statements of publicly traded Brazilian companies. The artifact uses pipelines that can be reused by students and other interested parties in finance to study the behaviors of a company's time series results and thus introduce research on predicting future results.

1.1.1 Statement of Comprehensive Income (DRE). According to the Brazilian Securities and Exchange Commission (CVM) [7], the Statement of Comprehensive Income (DRE in Portuguese) is an accounting document that displays the computations of all a company's revenues and expenses over a period, generally at the end of a year or quarter. The DRE structure lists the company's gross revenue followed by the accounting deductions incurred on this value, until it informs, after all inconsistencies, what was the company's profit or loss in that period. All publicly traded companies are required to disclose the DRE and other financial statements on a quarterly basis and in the accumulated year-end period. The statements are submitted to the CVM following the International Financial Reporting Standards (IFRS), as expressed in CVM Instruction nº 485 [8].

It should be noted that the CVM makes individual and consolidated results available, differing from consolidated statements that consider the controlling companies and their subsidiaries [5]. In this work, the generated experiments are based on consolidated DRE's.

Among the accounts provided in the Statement of Income, we highlight the Gross Operating Revenue, the Cost of Goods and Services, whose difference is the Gross Profit (Gross Profit). We also have operating expenses, the difference from which toOBInvest Data Dictionary Gross Profit is the account "Result Before Financial Result and Taxes", also known as EBIT (Earnings Before Interest and Taxes) and is generally referred to as "operating profit" [21]. Another highlight is the important line called Net Result or Net Profit. We will show some important calculations in this article, one of them is the concept of Margin, which, "at different levels of the balance sheet, indicates what the gross, operational and net result of the company represents relative to its net revenue" [17].

1.1.2 Stakeholders Mapping. One of the contributions of this article is the definition of Stakeholders. These are interested in a company's financial statements (i.e., "any person, entity, or system that affects or is affected by the activities of an organization" [11]). Therefore, Stakeholders can be: the company's shareholders, employees, customers, suppliers, government, competitors, media, unions, among others.

The search for accounting information is necessary to make a professional analysis of investment companies and usually precedes the decision-making of those who intend to invest in a given company. Although publicly traded companies make these results available on investor relationship sites, data is often available by year and in distinct, standardized formats such as csv, xls or pdf. For the stakeholder who wants to follow the quarterly results of more than one company, this article will provide an experiment in capturing the DRE data submitted to the CVM by companies, making the historical series of accounts, line by line, available from 2011.

1.2 Work Structure

The rest of the work is organized as follows: Section 2 presents a list of the main works found that use CVM data, as well as some GitHub repositories. In section 3 the methodology used is presented. In section 4 the OBInvest dataset and information about FAIR, provenance, and reproducibility data is presented. A brief discussion is presented in section 5. In conclusion section 6 the conclusion of the work is made, and some suggests and improvements are presented for future works.

2 RELATED WORKS

There are few academic works that use CVM data with little expressive results, which motivated the present work. In [6], an analysis was proposed on the generation and distribution of wealth created by companies listed on the Securities and Exchange Commission using Value Added Statement (DVA) data from 2009 to 2017. In [3], a comparative analysis of the quality of profits of two accounting standards (COSIF and CPC) was proposed using information provided by CVM and BACEN, in the period from 2010 to 2018. [13] proposes some *valuation* models used by Emerging Companies Mutual Investment Funds (FMIEE), relying on data from the CVM and a survey based on questionnaires sent to the managers of the FMIEE funds. In [22] an attempt was made to identify differences and similarities in environmental liability disclosure practices between Brazil and the United States, using financial statements from 2002 to 2004.

The main projects found in Github repositories that use CVM data are [19] which contains a series of Python notebooks for obtaining various types of financial information. [9] is a project made in Google's GO Lang language, where you can directly obtain various information from CVM through command lines. The brFinance project [20] is a Python library that can be imported into Python projects and seeks financial data information directly from the CVM. The repository [23] contains a set of Python notebooks and one of them relies on the data from the Profit and Loss Account Statements (DREs), the same type of data analyzed in this work.

3 MATERIALS AND METHODS

The development of this work was based on the OSEMN framework [14], widely used in Data Science Processes when the objective is to make questions about the data after the end of the process, unlike what is done in CRISP-DM [25], where questions are made early. In this study, the Obtain, Scrub and Explore steps were developed, resulting in the OBIvenst dataset.

The execution of the steps was supported by computational tools, with a focus on the Python language v.3.9.12, the Pandas library v1.4.2, the KNIME workflow platform v.4.6.1, and the Jupyter integrated development environments (running on the Anaconda3 platform in version v2022.5 and Docker v20.10.14). The experiment was also executed in the Google Colabotory cloud environment.

3.1 Raw Data

Data available on the Brazilian Securities and Exchange Commission (CVM) open data portal was used, available at https://dados.cvm.gov.br/dataset/. The data is open source with an Open Data Commons Open Database License (ODbL) and is managed through the Comprehensive Knowledge Archive Network (CKAN) [15]. The information collected in this research is listed under the "Company" data group and three different sets were obtained, all in Comma Separated Values (CSV) format: ***(i)*** Company Open Data Cadastral (CAD-CIA) dataset; ***(ii)*** Quarterly Information Form (ITR) dataset; ***(iii)*** Standard Financial Statements Form (DFP) dataset.

3.2 Data Engineering

The data used in the work relates to publicly traded companies monitored by the CVM. To achieve the proposed objective, it was decided to analyze the historical and audited information for the period from 2011 to 2021. Since it is a historical, controlled data that will not be changed, it was understood that there would be no gain in downloading these information at the time of experiment execution, opting to store the datasets in a data directory in the project repository. The data was collected and stored on September 16, 2022. Out of the large amount of data made available by the CVM, only three datasets were used: **(i) Public Company Registration Data (CAD-CIA)** - The dataset with 2550 records contains registration information for publicly traded companies, allowing for grouped analyses by each company's area of operation; **(ii) Quarterly Information Form (ITR)** - The CVM provides a set of information regarding public companies in a Quarterly Information Form (ITR). From the various data made available by the CVM, the article chose to deal with the DRE information. The CVM provides two sets of information, the Individual DRE, which deals with the data of the companies and their subsidiaries and affiliates individually; and the Consolidated DRE, which deals with the information between holding and associated companies in a consolidated manner. For the investment analysis, the results data of companies in consolidated format is used. For the 11-year interval analyzed, the consolidated CSV files of Consolidated DREs totaled 1,654,787 records; and **(iii) Standardized Financial Statements Form (DFP)** - Information about a company's accounts and financial results each year. Also for the 11-year interval analyzed, the consolidated CSV files of annual DRE (DFP) type totaled 665,666 records.

3.3 Information deduplication

A significant amount of redundancy was observed in the raw DRE data. For all DRE account records, both in the quarterly and annual financial statements, two records are presented, where one contains the data of the current period, and a second record with the information of the same period, but from the previous year.

In the quarterly DRE (ITR), upon initial analysis, isolated records for quarters were observed, but there were also records for accumulations of the second and third quarters, which in a way creates redundancy. Another point noted about the ITR is that the record of the last quarter (the fourth quarter) was not presented, and to obtain this information, it would be necessary to obtain this data from the annual DRE dataset (DFP) and subtract from the accumulated third quarter existing in the ITR DRE dataset.

3.4 DATA PIPELINE

The first part of the exploration stage was carried out using the KNIME tool [12] creating a workflow to start the exploratory process of the working dataset, the workflow is available in the project repository. The exploration phase helps clarify which operations are necessary for creating the final work dataset and generating the provenance of these processes.

Next, the Pandas library was used as a tool to create a dataset that met the scope of the work and aggregated information that could be useful to OBInvest. A join of the CAD-CIA, ITR, and DFP datasets was performed with the necessary transformations, compositions, and aggregations deemed useful and essential. The main data preparation activities developed were: (i) Transformation between the CAD-CIA, ITR, and DFP datasets, aimed at filtering out companies with active registration and aggregating the "Activity Sector" column to the ITR and DFP, originating from CAD-CIA; (ii) Slicing of the ITR dataset, aimed at obtaining two dataframes. The first dataframe containing information from the first, second, and third quarters, called TRIM123. The second containing the accumulated value between the first and third quarters, relating to the year of the financial year, called ACM3; (iii) Slicing of the DFP dataset, with the purpose of filtering out duplicated data and acquiring a dataframe composed of the accumulated value between the first and fourth quarters, calling it ACM4; (iv) Subtraction between the values in the "Account Value" column of the ACM4 and TRIM123 dataframes, resulting in the dataframe containing the fourth quarter, called TRIM4; (v) Concatenation between the TRIM123 and TRIM4 dataframes, thus obtaining the OBInvest dataset.

During the preparation process, two more columns were added: "Year" and "Quarter." There was also the need to standardize the records contained in the "Account Value" column, keeping all values in multiples of a thousand. The steps of the developed workflow can be seen in Figure 1, available in the appendix seccion

3.5 Problems found in raw datasets

The following inconsistencies were observed in the records of all ITR and DFP datasets after the Obtainment, Exploration, and Workflow process: ***(i)*** Companies with irregular financial exercises, where the first quarter of the financial year does not coincide with the first quarter of the year; ***(ii)*** Companies that are exempt from filing their financial reports, making it impossible to calculate the financial performance; ***(iii)*** Companies only submitted their financial reports in the DFP dataset, making it impossible to determine the financial quarter.

3.6 Discussion about data, cleaning and record linking

The strategy adopted in this study was to disregard the data submitted by companies that fall into the problems listed in Subsection 3.5. The dataset after all treatments has a much lower number of records than the initial numbers. The main factor responsible for this decrease in records is the redundancy of information in the raw data, as explained in Section 3.3. In addition to redundancy, after treating all the problems found in the datasets, there was a reduction in both the number of companies and the total number of records, but not preventing the continuation of the study. The following reductions were observed in the OBInvest dataset: ***(i)*** The ITR and DFP datasets recorded 621 and 623 distinct companies, respectively. This amount declined to 425 companies; ***(ii)*** The sum of unique records found in the ITR and DFP datasets was 1,637,176, which was reduced to 413,641.

A possible solution to the problems found in the anomalous records of the companies would be to individually treat these records, so that they could be included in the OBInvest dataset, which is not part of the scope of this work. But they could be implemented in future works.

3.7 Visualization of Results

As an example of a data visualization, we created a visualization of the net margin of sectors. With an analogous procedure, we can determine the operating margin and gross margin by taking the accounts corresponding to gross profit and operational profit respectively instead of net profit. The historical quarterly net margin series of the sectors to which the companies are classified according to the CVM was obtained as the division of the sum of Net Profits by the sum of Net Revenues, by quarter, of the companies that make up a given sector. This is a result that can track investment opportunities in sectors that indicate increasing profitability, and this can be of great value for investors in general. The visualization of the sector margin is available in the project repository.

4 CURATED, REPRODUCIBLE, AND ENRICHED DATASETS WITH METADATA PROVENANCE

The OBInvest dataset containing the quarterly DREs of publicly listed companies on the CVM is obtained as a result of the experiment. Best practices related to the F.A.I.R. principles were sought. In addition, a more detailed description of the experiment was made, based on provenance, in an attempt to ensure both interoperability and reproducibility.

This section of the article presents details on the conduct of the construction of provenance and reproducibility. The last executable version of the experiment can be accessed through the *doi*: https://doi.org/10.5281/zenodo.7110653 and reproduced using the Docker container.

4.1 Data Dictionary

Data Dictionary for OBInvest dataset generated by the end of the data pipeline differs very little from the dictionary provided in the metadata information on the CVM website. A few fields from the CVM DRE dictionary have been suppressed, and the Year and Quarter fields have been added. The dictionary can be observed in more detail in table 1, available in the appendix seccion.

4.2 Data FAIRification

In [24], a series of relevant but obvious questions about the urgency of change in how research should be conducted to clearly and easily ensure the reuse of academic work. The concern is related to both the execution and the data used in research. The work leaves a series of reflections in the scientific community.

The principles, in English, Findable, Accessible, Interoperable and Reusable, creating the acronym F.A.I.R. (Recoverable, Accessible, Interoperable and Reusable), from 2017, began to be organized and promoted by the GO FAIR movement, in various countries around the world, including Brazil.

The experiment performed in this article tried to follow the principles defined in [10] during its creation. Table 2 , available in the appendix seccion, presents the understanding of which principles were achieved during the process. In the Status column, only the principles that the team considered to have been completely implemented were identified as "Implemented". For those in which there were still doubts about the completeness of the implementation, it was decided to mark them as "Not Implemented".

4.3 Provenance in the Curated Dataset

The generated provenance attempted to represent, as detailed as possible, the entire life cycle of the experiment: from the origin and location of the data used; all the agents involved; and a step-by-step execution of the experiment, where the metadata that allows interoperability and image of the provenance are created during the execution of the experiment.

The provenance was created using the python PROV v2.0.0 library, which respects the model established by the standard [18] maintained by W3C. There was great care in using namespaces that could assign semantics at each stage of the provenance creation process. The provenance can be viewed and analyzed in three main parts.

The first part tries to demonstrate the origin of the information, made available by the CVM agent, how it is made available and how it was stored in our repository.

In the second part, it is possible to visualize the hierarchy of agents involved in defining and executing the experiment. Finally, in the third and last part, a agent referring to the python notebook of the experiment, of the type "Software Agent", was created that details the entire execution of the experiment, including the timestamps of each stage during its execution.

Figure 2, available in the appendix seccion, shows a small excerpt of the generated provenance graph, and a complete image is accessible in the project repository.

4.4 Experiment Reproducibility

Another important point that was addressed during the development of the work was related to the idea of repeatability and reproducibility of the experiment.

When thinking about repeatability and reproducibility, the idea of executing the experiment in cloud environments such as Google Colaboratory or MyBinder always raises a lot of attention. However, some questions arose for the team members. The main ones were related to the lack of guarantee that this type of environment offers, about the evolution of language and library versions, and how this could impact repeatability in the future.

Thus, the decision was made to invest in repeatability and reproducibility actions using Docker containers and environments where the language and main library versions would be managed by Conda.

The idea was to create a Docker image in such a way that the exact version of Python that would be used, as well as the versions of the main libraries used in the experiment, could be controlled. All defined explicitly, thus guaranteeing a "freeze" both of the experiment and of the entire environment used.

To do this, during the creation of the Docker image, some of the steps focused on defining a Conda environment, tying the Python and main library versions used. Another verification created by the team was the creation of a method that is executed before the main code of the experiment, where the Conda environment and if the expected versions of the libraries are found in the Jupyter environment are verified.

Another step during the creation of the Docker image is the cloning of the OBInvest Github repository to the specific version that was defined (using a version control tag).

The directory where the repository is cloned serves as the path to the Jupyter instance that will be executed during the execution of this image and will be used during the creation of the Docker container, from where it is possible to reproduce the entire experiment (including the creation of the OBInvest dataset and the creation of provenance metadata at runtime).

The created Docker image was sent to the Docker Hub, a remote Docker image repository, which allows a container of an image to be downloaded and executed in a simplified way. To do this, just execute the command: *docker run -p 8888:8888 obinvest/drecvm*

5 DISCUSSION

The quarterly financial statement information offered as a result of creating the curated OBInvest Dataset was not found in any related works. Despite the exclusion of exceptional cases presented in section **??**, we understand that the offered set will assist a series of OBInvest activities.

Despite the lack of maturity with the GO FAIR principles and the lack of experience with issues related to reproducibility using containers, we understand that the work was reasonably successful in these aspects. It is possible, in a simple and clear way, to execute the procedure from any computer with Internet connection. Once this environment has been executed for the first time, the work can even be used in an "off-line" manner in subsequent executions.

The metadata generated during the development of the work was able to represent in detail the origin of the data, the main agents involved, and the detailed step-by-step of the experiment, where all this information that enables interoperability is created in execution time in the RDF Turtle, XML formats in addition to being presented in a png format image.

The sector margin analysis, although simple, clearly illustrates the potential for information that can be obtained from the dataset created in the experiment.

6 CONCLUSION

This article created a data enrichment strategy and made available a dataset with information of the quarterly financial statements of publicly traded companies, originally made available by CVM, so that it could be used in OBInvest activities. As a demonstration of the potential of OBInvest, a quarterly margin analysis of the sector was made, grouped by industrial sector, to simply illustrate how this information can be exploited.

During the development of the work, all activities tried to follow the GO FAIR principles and there was a great concern to ensure that the experiment could be simplified. To do this, both the experiment and the environment used were packaged through the use of Docker images and with the support of the Conda environment manager.

It is understood that various improvements can be incorporated into the presented work, the main suggestions for future work are: ***(i)*** Perform more substantial analyses on the created dataset, such as predicting results based on the history of demonstration results; ***(ii)*** Deal with each case, the exception situations of companies that presented DRE results outside the standard, as shown in section **??**, so that they can be added to the OBInvest Dataset; ***(iii)*** Obtain other types of information made available by CVM, to be used during the activities of the Brazilian Investment Olympics; ***(iv)*** Adapt the docker image so that it can be executed in the My Binder cloud environment, aiming to improve reproducibility issues. Since the My Binder environment already works on the container environment, this form of execution would provide a completely reproducible and executable environment directly from a cloud environment; ***(v)*** Adapt the component version verification step to a comparison of the complete list extracted from the Conda environment with the environment identified at the time of the experiment execution, instead of checking the version of the Python language and some important libraries, as is currently done.

REFERENCES

[1] Carmela Aprea, Eveline Wuttke, Klaus Breuer, nk Koh, Peter Davies, Bettina Greimel-Fuhrmann, and Jane Lopus. 2016. *International Handbook of Financial Literacy.* https://doi.org/10.1007/978-981-10-0360-8

[2] BNCC. 2018. Base Nacional Comum Curricular. Disponível em: http://basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_versaofinal_site.pdf, Acessado em: 20/09/2022.

[3] Anderson Catapan and Romualdo Douglas Colauto. 2020. Governança corporativa: uma análise de sua relação com o desempenho econômico-financeiro de empresas cotadas no Brasil nos anos de 2010–2012. *Contaduría y administración* 59, 3 (2020), 137–164.

[4] CEFET-RJ. 1917. Centro Federal de Educação Tecnológica Celso Suckow da Fonseca. Disponível em: http://www.cefet-rj.br, Acessado em: 20/09/2022.

[5] CFC. 2016. NBC TSP 6 - Demonstrações Consolidadas e Separadas. Disponível em: https://cfc.org.br/wp-content/uploads/2016/02/NBC_TSP_6_Demonstracoes_Consolidadas_e_Separadas.pdf, Acessado em: 20/09/2022.

[6] FELIPPE CHIELLA and FERNANDO RICHARTZ. 2019. MFC283-ANÁLISE DA GERAÇÃO E DISTRIBUIÇÃO DO VALOR ADICIONADO DAS EMPRESAS REGISTRADAS NA CVM NOS ANOS DE 2009 A 2017. (2019).

[7] CVM. 2017. *ANÁLISE DE INVESTIMENTOS - Histórico, Principais Ferramentas e Mudanças Conceituais para o Futuro.* Disponível em: https://www.investidor.gov.br/portaldoinvestidor/export/sites/portaldoinvestidor/publicacao/Livro/livro_TOP_analise_investimentos.pdf, Acessado em: 20/09/2022.

[8] CVM485. 2010. INSTRUÇÃO CVM Nº 485. Disponível em: https://conteudo.cvm.gov.br/export/sites/cvm/legislacao/instrucoes/anexos/400/inst485.pdf, Acessado em: 20/09/2022.

[9] dude333. 2021. Rapina. Disponível em: https://github.com/dude333/rapina, Acessado em: 12/10/2022.

[10] GO-FAIR. 2017. GO FAIR. Disponível em: https://www.go-fair.org/fair-principles/, Acessado em: 12/10/2022.

[11] IBGC. 2015. Código das Melhores Práticas de Governança Corporativa. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4382648/mod_resource/content/1/Livro_Codigo_Melhores_Praticas_GC.pdf, Acessado em: 20/09/2022.

[12] Knime. 2004. Knime Workflow. Disponível em: https://www.knime.com/, Acessado em: 12/10/2022.

[13] Vagner Antônio Marques, Jacqueline Veneroso Alves da Cunha, and Poueri do Carmo Mário. 2020. Modelos de Valuation Utilizados pelos Fundos Mútuos de Investimentos em Empresas Emergentes (FMIEE). (2020).

[14] H. Mason and C. Wiggins. 2010. A Taxonomy of Data Science. Disponível em: https://https://web.archive.org/web/20211219192027/http://www.dataists.com/2010/09/a-taxonomy-of-data-science/, Acessado em: 26/09/2022.

[15] Jennifer C Molloy. 2012. *The Open Knowledge Foundation: Open Data Means Better Science.* Working Papers id:4686. eSocialSciences. https://ideas.repec.org/p/ess/wpaper/id4686.html

[16] OBInvest. 2020. Olimpíada Brasileira de Investimentos. Disponível em: https://www.obinvest.org, Acessado em: 20/09/2022.

[17] Alexandre Povoa. 2012. *Valuation - Como Precificar Acoes.*

[18] PROV-Overview. 2013. An Overview of the PROV Family of Documents. Disponível em: https://www.w3.org/TR/prov-overview/, Acessado em: 12/10/2022.

[19] Código Quant. 2020. Python para Investimentos. Disponível em: https://github.com/codigoquant/python_para_investimentos, Acessado em: 12/10/2022.

[20] Eudes Rodrigo. 2022. brFinance. Disponível em: https://github.com/eudesrodrigo/brFinance, Acessado em: 12/10/2022.

[21] S.A. Ross, R.W. Westerfield, J. Jaffe, and R. Lamb. 2015. *Administração Financeira.*

[22] Suliani ROVER and José Alonso BORBA. 2006. A evidenciação das informações ambientais nas Demonstrações Contábeis das empresas que atuam no Brasil e que negociam ADR's na Bolsa de Valores dos Estados Unidos: uma análise das DFP's (CVM) e do relatório 20-F (SEC). In *VI CONGRESSO USP DE INICIAÇÃO CIENTÍFICA EM CONTABILIDADE*, Vol. 6.

[23] Lucas Vido. 2020. DRE CVM. Disponível em: https://gist.github.com/Vido/cbc33862dd27a22790df633f1d113ae6, Acessado em: 12/10/2022.

[24] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[25] Rüdiger Wirth. 2000. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*. 29–39.

A DEVELOPED WORKFLOW MODEL USED

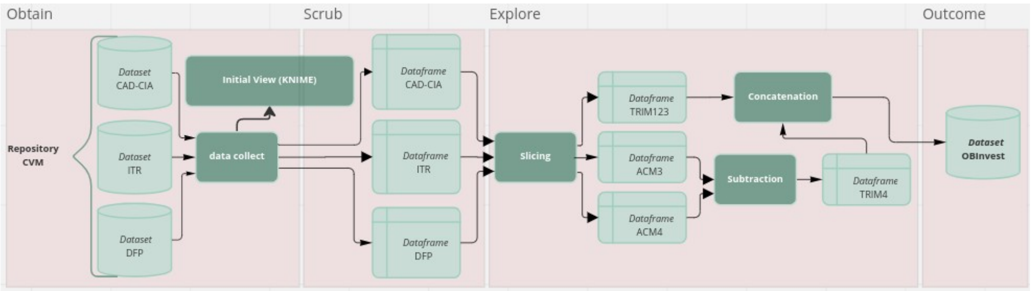

Fig. 1. Developed workflow model.

B OBINVEST DATA DICTIONARY

OBInvest Dataset Data Dictionary			
Field name	**Description**	**Type**	**Category**
CNPJ_CIA	Company CNPJ	varchar(20)	
DT_REFER	Reference date of the document	date(10)	YYYY-MM-DD
DENOM_CIA	Company business name	varchar(100)	
CD_CVM	CVM Code	char(6)	
GRUPO_DFP	Name and level of aggregation of the demonstration	varchar(206)	
MOEDA	Currency	varchar(100)	
ESCALA_MOEDA	Monetary scale	varchar(100)	
ORDEM_EXEC	Order of the social year	varchar(9)	
DT_INI_EXERC	Beginning date of the social year	date(10)	YYYY-MM-DD
DT_FIM_EXERC	End date of the social year	date(10)	YYYY-MM-DD
CD_CONTA	Account code	varchar(18)	
DS_CONTA	Account description	varchar(100)	
VL_CONTA	Account value	decimal(29,10)	
ST_CONTA_FIXA	Indicates if it is a fixed account or not	varchar(1)	'S':Yes; 'N':No
ANO	Year of the reference date of the document	decimal(4,0)	
SETOR_ATIV	Activity sector	varchar(100)	
TRIMESTRE	Quarter of the reference date of the document	decimal(1,0)	

Table 1. OBInvest quarterly DRE data dictionary representation

C FAIR LEVELS REACHED

Principles of the GO F.A.I.R. Initiative	
Principle	**Situation**
(F) – Findable	
(F1) Metadata must be assigned globally, persistently and identifiable identifiers	Implemented
(F2) Data is described using enriched metadata.	Implemented
(F3) The metadata clearly and explicitly includes the identifiers of the described data.	Implemented
(F4) The metadata is registered or indexed through a searchable resource.	Not implemented
(A) – Accessible	
(A1) Metadata can be retrieved using its identifiers through a standardized communication protocol.	Implemented
(A1.1) The protocol is free, open and universally implementable.	Implemented
(A1.2) The protocol allows for authentication and authorization procedures, when necessary.	Implemented
(A2) The metadata is accessible even when the data is no longer available.	Implemented
(I) – Interoperable	
(I1) Metadata uses a formal, accessible, shared and widely applicable language for knowledge representation.	Implemented
(I2) Metadata uses vocabularies that follow the FAIR principles.	Not implemented
(I3) Metadata includes qualified references to other metadata;	Implemented
(R) – Reusable	
(R1) Metadata is described with a plurality of accurate and relevant attributes.	Not implemented
(R1.1) Metadata is released with clear and accessible data usage licenses.	Implemented
(R1.2) Metadata is associated with detailed provenance.	Implemented
(R1.3) Metadata is part of domains with shared standards usage in communities	Implemented

Table 2. Implemented FAIR Principles

D PART OF PROVENANCE IMAGE GENERETED

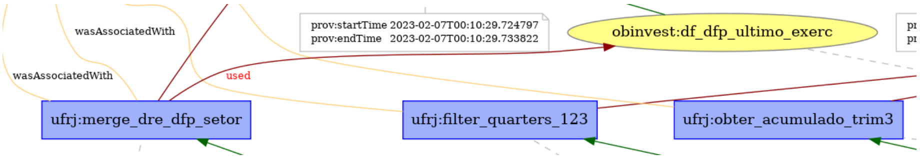
wasAssociatedWith
wasAssociatedWith
used
prov:startTime 2023-02-07T00:10:29.724797
prov:endTime 2023-02-07T00:10:29.733822
obinvest:df_dfp_ultimo_exerc
ufrj:merge_dre_dfp_setor
ufrj:filter_quarters_123
ufrj:obter_acumulado_trim3

Fig. 2. Detail of the experiment operations, created during execution.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009