An Exploratory Analysis of the Use of ETL in Transparency Portals*

Marcus Vinicius Santana Poletti[a], Methanias Colaço Junior[b] and
André Vinicius Rodrigues P. Nascimento[c]
Federal University of Sergipe, Avenida Marechal Rondon S/N Caixa Postal 49,100-000 São Cristovão, SE, Brazil

Keywords: Open Data, e-Government, Public Transparency, ETL, Data Pipeline.

Abstract: Context: Government transparency portals are built based on ETL (Extract, Transform and Load) processes, which increase the quality and interoperability of data, making a critical subsystem for these applications, subject to evaluative research for improvements. Objective: To analyze publications on the use of ETL in transparency portals, in order to characterize them in relation to their scenarios, impacts, empirical methods and general bibliometric data. Method: Using the PICO strategy (Population, Intervention, Comparison and Outcome), a systematic mapping of the literature was performed. Summary of Results: In a total of 204 publications researched, 25 works were selected, of which 40% present, as the main impact for the portals, the availability of support for the construction of loads through a graphical interface, followed by the possibility of connectivity between bases of heterogeneous data (27%) and the ability to monitor loads (22%). Regarding the real automation of loads and their quality control, respectively, only 8% and 3% of the works discussed the impacts of these characteristics. Conclusion: The research showed that the use of ETLs in transparency portals still lacks comparative and feasibility studies. In this sense, an existing challenge is the lack of research that carries out replications to consolidate and validate the works already published, evidenced by the scarcity of controlled experiments in the area. Finally, analyzes on the quality control of loads was an important gap identified.

1 INTRODUCTION

Although Brazil has adopted an open data policy since 2011, in accordance with the provisions of Law No. 12,527 - Law on Access to Information (LAI) (Brazil, 2011), access to transparency portals in the 26 Brazilian states, and Federal, shows that theory and practice are at different stages. LAI establishes that every municipality with more than ten thousand inhabitants must make government data available on a transparency portal (Brazil, 2011). For greater effectiveness, data must be made available following the principles of the Open Government Data (OGD) standard, proposed by the World Wide Web Consortium (W3C) (Transparency International, 2020).

The predominance in transparency portals of Brazilian capitals and states is the data availability in formats that do not follow the OGD principles, resulting in a lack of interoperability (Eberhardte & Silveira, 2018; Oliveira & Silveira, 2018; Bachtiar, Suhardi & Muhamad, 2020; Cenci, Fillotrani & Ardenghi, 2017). There are also inadequacies regarding the completeness, primacy, opportunity, and accessibility of the data available on these portals (Oliveira & Silveira, 2018). This is a problem that deserves attention, as the reuse of this data, in initiatives that promote population engagement, becomes a major technological challenge (Bachtiar, Suhardi & Muhamad, 2020; Muller, Gil-Garcia & Tirelli, 2018; Dahbi, Lamharhar & Chiadmi, 2018). In December 2020, the Federal Government's Anti-Corruption Plan was published, bringing the implementation of 142 actions distributed among prevention, detection, and accountability mechanisms (Tian et al., 2021). These actions also aim to improve public transparency and promote population

[a] https://orcid.org/0000-0002-9047-5995
[b] https://orcid.org/0000-0002-4811-1477
[c] https://orcid.org/0000-0003-4219-3704
Postgraduate Program in Computer Science – PROCC

engagement. However, it is a necessary technological structure that makes possible their executions.

In this context, one of the critical points for these transparency portals creation is the data loading, in large institutions, is usually spread over several heterogeneous systems with the most varied types of databases (Pan, Zhang & Qin, 2018). In other words, the data need to be Extracted (E) from these source systems, Transformed (T), and Loaded (Load – L) in integrated databases that will be accessed by the portals (Sun & Lan, 2012). These three tasks gave names to the systems that perform them, they are the ETL Systems (Extract, Transform, and Load) (Sun & Lan, 2012).

Portals need to build and/or acquire these systems, considering impacts such as connectivity to all types of bases, ease and agility of use through a graphical interface, control and data quality assurance, automation, and monitoring of loads (Sreemathy, Infant Joseph, Nisha, Chaaru & Gokula, 2020). In addition, ETL systems are also directly involved with creating metadata, publishing data and metadata on the web, as well as creating appropriate catalog records (Saranya et al., 2021).

Given this need, in this article, we present a systematic mapping to characterize articles in relation to the performance of ETL tools in transparency portals, identifying impacts, approaches, scenarios, research methods, and general bibliometric data. The rest of this article is organized as follows: Section 2 describes how the systematic mapping was designed. Section 3 presents how the systematic mapping was conducted. In section 4, the results obtained during the study are presented. Section 5 presents a discussion of the results. Section 6 presents the threats to validity found and, finally, in section 7, the conclusion is presented.

2 SYSTEMATIC MAPPING PLANNING

2.1 Objective

This mapping had as its principal objective to identify and characterize the impacts caused by the ETL Systems in transparency portals use.

2.2 Research Questions

The research questions were developed to present an area overview, highlighting fundamental aspects of primary studies (Kitchenham, 2004; Petersen, Vakkalanka & Kuzniarz, 2015). They were prepared based on the PICO model (Bergin & Wraight, 2006;

Santos, Pimenta & Nobre, 2007), to highlight the effects of an intervention in a given population and structure the research into four fundamental elements: Population, Intervention, Control, and "Outcomes" (Results). These elements, according to (Santos, Pimenta & Nobre, 2007), can be used to build research questions of different natures. Table 1 illustrates the PICO model used in this work.

Table 1: PICO model for research question compliance.

Acronym	Category	Description
P	Population	Publications that directly address government transparency portals, with or without open data.
I	Intervention	Development and/or application of ETL tools to optimize the process of Extraction, Transformation, and Loading of portals.
C	Control	Articles on Transparency Portals that do not use tools developed specifically for ETL processes. **Articles that fit the intervention:** LinkedPipes *ETL* in use: Practical publication and consumption of Linked Data (Klímek & Skoda, 2017); A Content-Driven *ETL* Processes for Open Data (Berro & Teste, 2015). **Control article:** Analysis of the energy service in non-interconnected zones of Colombia using business intelligence (Colmenares-Quintero, 2021).
O	Result	Automation, scheduling, orchestration, and monitoring of ETL processes, prioritizing quality through clean and accurate data, as well as ease through graphical interfaces.

Thus, from the PICO model definition, research questions were elaborated, based on the guidelines of the Systematic Literature Mapping protocol observed in (Kitchenham, 2004; Petersen, Vakkalanka & Kuzniarz, 2015). They are: Q1: What methods are used in research on ETLs that deal with transparency portals?; Q2: What are the approaches used for ETL?; Q3: What are the most used data extraction, transformation, and loading tools within the transparency portals scope?; Q4: What are the portal ownership scenarios (see table 6)?; Q5: What publication types or forums have addressed the ETL issue in the public transparency context?; Q6: Which

countries have the most researchers who have published in this area?; Q7: Which years had the most publications in this area?; Q8: What are the impacts of using ETL on Transparency Portals?

2.3 Search and Selection Strategy

To search for articles, the databases responsible for publishing the main journals in the Computer Science field were selected, namely, ACM Digital Library (ACM), IEEE Xplore (IEEE), SCOPUS, and Science WEB. The searches were carried out using the filtering tools available in each database, considering the searches: title, abstract, and keywords. Access to the databases was performed through the CAPES journal portal (Capes, 2021) using an institutional subscription without any article restrictions.

For digital databases search, a search string was defined and composed of English terms and synonyms, associated with transparency portals and advantages of the applying ETL. The identified terms from the papers of the PICO model and the control articles are defined in Table 1, and later refined and adapted for better string use. Table 2 presents the terms before refinement.

Table 2: Terms before refinement.

Category	Description
Population	Open government data portals, open government, e-government, public sector information, OGD, OGD portals, open government data, digital government, e-services, government data, transparency, government accountability, government transparency, open government data ecosystem, open data platforms, open data
Intervention	Extraction, transformation, and loading; *ETL*; Extract, transform, load; Load Procedures; Extraction Procedures; Load Software; Extraction Software; Load Program; Extraction Program.
Control	-
Result	Automation; Orchestration; Scheduling; Quality; Connectivity; Graphic Interface; Graphical UserInterface; User Interface; Ease; Facilitate; Monitoring; Alerts; Broadcasting.

After refinement, the adjusted terms were used to build the search string, which is described in Table 3.

Table 3: Strings chosen after refinement.

Search string terms		
Population	Intervention	Result
open data, open government, e-government, digital government, public transparency, government transparent, electronic governmeny, government accountability	extraction, transformation and loading, etl, extract, transform and load	Orchestr*, gui, graphical user interface, eas*, facilit*, monitor*, alert*, warn*, connectivity, schedul*, quality

From the terms highlighted above, the following search string was created:
TITLE-ABS-KEY((“open government” OR “open data” OR “e-government” OR “digital government” OR “public transparency” OR “government transparency” OR “electronic government” OR “government accountability”) AND (“extraction, transformation and loading” OR “etl” OR “extract, transform and load”) AND (“orchestr” OR “gui” OR “graphical user interface” OR “eas*” OR “facilit*” OR “monitor*” OR “alert*” OR “warn*” OR “connectivity”OR “schedul*” OR “quality”))*

2.4 Source Selection Criteria

Inclusion and exclusion criteria were established to filter articles relevant to systematic mapping. Only the studies selected for evaluation, which passed the inclusion and exclusion criteria, were counted.

The inclusion criteria are presented below:
1. Articles that are available online and in digital libraries;
2. Articles that contain the search String in the title, abstract, or keywords;
3. Articles that identify and characterize existing studies on the use of ETL tools in transparency portals;
4. The articles must have a publication date between the years 2011 and 2022. This period, significant and above 10 years, was chosen based on Law No. 12,527 - Law on Access to Brazilian Information (LAI), of 2011;
5. Articles written in English.

The exclusion criteria were:
1. Duplicate studies;
2. Surveys*;*
3. Systematic review;
4. Preliminary studies;

5.   Short Papers.

2.5   Information Extraction Strategy

To assess the work quality and answer the research questions set out in section 2.2, tables 4, 5, and 6 were designed, which served as a guide for researchers to have no doubts about the data to be extracted for each article read in the whole. According to (Kitchenham, 2004), data extraction forms should be designed to collect all the information necessary to address the questions and quality criteria of the study.

Table 4: Impacts identified and extracted from the articles.

Impacts	Definition
Connectivity	Providing connectivity to unstructured data and cloud data sources.
Ease of use through Graphical Interface	Resource savings with developers through an intuitive, code-free environment for extracting, transforming, and loading data.
Quality	Quality control, to determine the data consistency, accuracy, and control.
Automation	Automation capabilities with task scheduling and process orchestration.
Monitoring	Process metadata storage, such as, for example, the load time and the number of records loaded, issuing alerts and warnings about deviations and the failure or success of loads to stakeholders, as well as providing a monitoring panel.

Table 5: Classification of Extracted Approaches.

Approach	Description
Market ETL	Ready and commercially available open-source or closed-source ETL tools.
Developed ETL	ETL tool developed especially for the project.
Not Specified	The article did not provide details about the Tool.

3   SYSTEMATIC MAPPING CONDUCTION

Responsible for publishing the major journals in the Computer Science field, the Scopus database (Yu et al., 2019) was chosen as the basis for defining and refining the search string. This base also includes articles from distinct scientific databases, such as IEEE, ACM, and Web of Science. After being

defined, refined, and judged adequate, the string was translated to the other search engines used in this work, which were: IEEE. ACM and Web of Science. In total, 204 works were returned, 24 (12%) from Scopus, 15 (7%) from Web of Science, 22 (11%) from IEEE, and 143 (70%) from ACM. The data are represented in Figure 1.

After searching the articles in the databases, the filtering process began based on the selection criteria defined in section 2.4. Each paper was classified as accepted or rejected. Of the 204 analyzed publications, 149 (73.04% of the total) were in accordance with the exclusion criteria. After removing these articles, the remaining works were read completely, 55 publications, 26.96% in relation to the total of publications. After being analyzed, following the inclusion criteria, 25 publications were selected, that is, 12.25% of the total will be analyzed. Figure 1 shows the summary of this step.
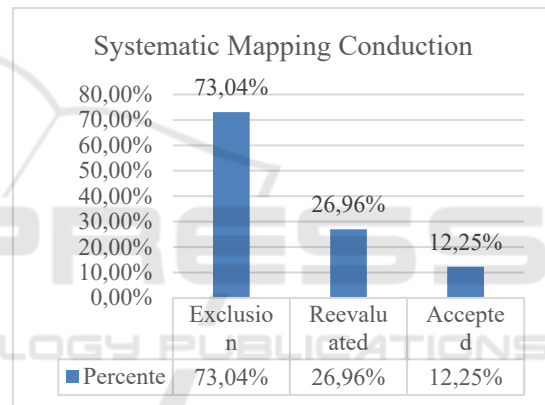

Figure 1: Systematic Mapping Conduction.

4   DATA SYNTHESIS AND RESULT PRESENTATION

In this section, the systematic mapping result will be presented according to the extracting process of the obtained articles and answering the research questions according to the extracted data.

4.1   What Methods Are Used in Research on ETLs Dealing with Transfer Portals?

Figure 2 presents the methods used in research on ETLs dealing with transparency portals. The highlight goes to Exploratory Studies, with 56% of the works.
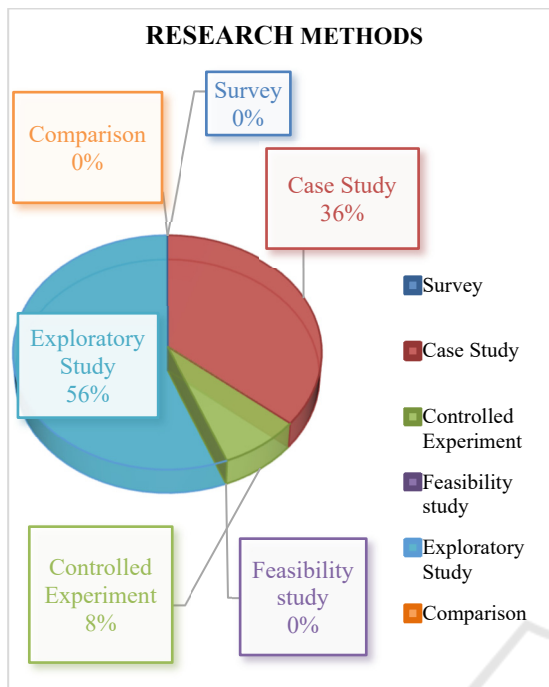
Figure 2: Works selected by Research Methods.

4.2 What Are the Approaches Used for ETL?

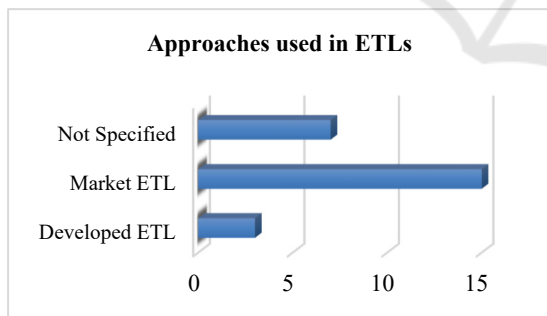Figure 3 presents the approaches used. We can see that 60% of the works use the ETL tools on the market.

Figure 3: Approaches used in ETLs.

4.3 What Are the Most Used Data Extraction, Transformation, and Loading Tools within the Scope of Transparency Portals?

Among the works analyzed, 28% did not specify the tool used. The instrument most cited in the works was Kettle, from the open-source software Pentaho, used in 32% of the analyzed papers. The second most

quoted tool was Talend Open Studio, used in 16% of the analyzed works. The third most cited tool, with 12%, was Linked Pipes ETL. The other most representative were ODET (4%), Cylon (4%), and SpatialETL (4%).

4.4 What Are the Portal Ownership Scenarios?

Of the scenarios found in the works, 52% of them are independent portals, that is, they can be portals for universities and NGOs (Non-Governmental Organizations) but use government data. We present the scenarios found in the works in percentages in figure 4.
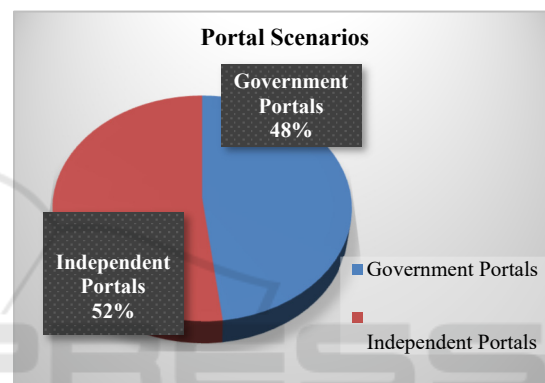

Figure 4: Scenarios found in the analyzed works.

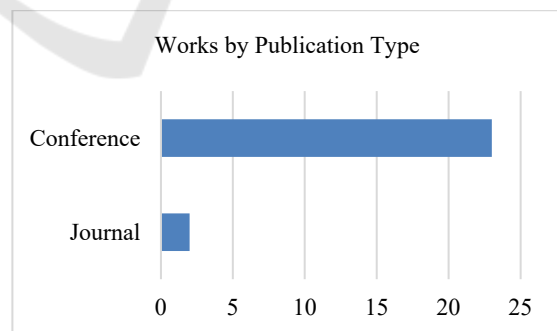## 4.5 What Publication Types or Forums Have Addressed the ETL Issue in the Context of Public Transparency?

Figure 5: Works selected by Publication Type.

4.6 Which Countries Have Published the Most in This Area?

Figure 6 shows the countries where the analyzed works were published. The United States appears with 48% of published works, being the country with

the most publications in the area, followed by Spain, with 12%, and Macedonia, with 8%.
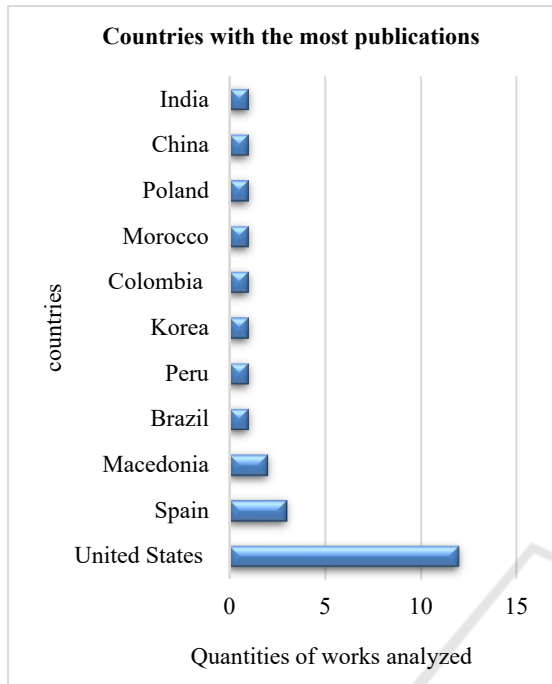
Countries with the most publications

Figure 6: Countries with more published works.

4.7 Which Years Had the Most Publications in This Area?

As shown in Figure 7, the year in which more works were published was 2019, with six publications.
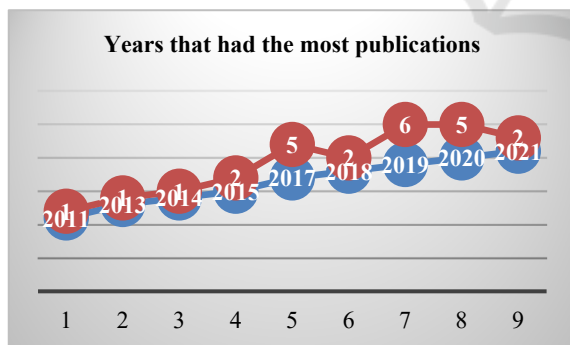
Years that had the most publications

Figure 7: The years with the most publications.

4.8 What Are the Impacts of Using ETL on Transparency Portals?

Figure 8 shows the most evident impacts in the works. Ease of use through a Graphical Interface was the most obvious impact, representing 40% of the analyzed papers.
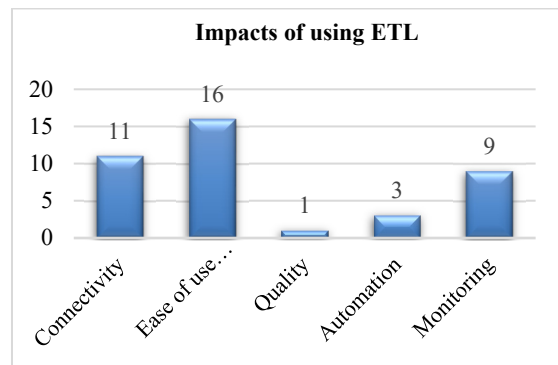
Impacts of using ETL

Figure 8: Impacts of using ETL.

5 NARRATIVE SYNTHESIS

In this section, the main aspects found and lessons learned in the analyzed articles will be discussed. Most articles published on the topic occurred in 2019, which indicates that the research area is still young. The results also show that publications related to the matter are present in several countries, which shows that the search for solutions for ETL usage in transparency portals is expanding.

The evidence on ETL quality control and assurance was addressed in only 3% of the studies studied. It denotes the need for research on how data transformations are carried out, the data amount that is rejected and not validated in approval processes, source codes generated, and how these transformed value validations are carried out, especially before the portal starts production.

The analysis of the publications also highlighted the absence of evaluations and experiments with closed source tools, such as Microsoft's Integration Services and Oracle Cloud Infrastructure Data Integration, both among themselves and against open-source tools

Finally, Controlled Experiments represent only 8% of the works, in addition, the comparison methods and feasibility study were not identified in the papers. It demonstrates a lack of research that performs replications for work consolidation and validation and the experimental study scarcity with more rigorous protocols that allow these replications. A knowledge base on ETLs and effective transparency portals will depend on directing research toward the effective application of the aforementioned scientific method.

6 THREATS TO VALIDITY

Threats to validity may limit the ability to interpret and/or describe data results obtained in a study (Chapetta, 2006). Therefore, some threats should be taken into account:

- **Construct Validity:** The search string and elaborated research questions span the entire related studies area. To mitigate this threat, an attempt was made to develop a string based on identified and refined terms with the help of control articles guided by the PICO model.
- **Internal Validity:** To mitigate data extraction and characterization problems, forms were designed to be filled in by the publication analysis and each analysis and extraction were reviewed by two researchers.

7 CONCLUSION

This work carried out a systematic mapping with the objective of identifying the ETLs usage in transparency portals, analyzing and highlighting the most relevant works in the area. Of the 204 studies retrieved from the scientific databases, 25 were accepted by the inclusion criteria, of which 24% were published only in 2019. This fact reveals the trend in the area, with recent attention from more researchers to the problem. Among the principal means of publication, conferences stood out, with 23 (92%) works, while journals were represented by only 2 articles (8%).

As identified gaps, the need for more studies on the quality process of ETLs was highlighted, as well as the absence of benchmarks with closed source tools widely used in the market. From a methodological point of view,

An existing challenge is the lack of research that carries out replications to consolidate and validate the works already published, as well as the scarcity of well-formatted experiments, which allow these replications. On the other hand, the results obtained in this research demonstrate that the application of ETLs in transparency portals is researched in several countries, allowing adaptations to the different contexts of governance and transparency. Finally, it is believed that this work presents relevant results to the academy, providing a source of consultation of the main gaps and trends in the use of ETLs in transparency portals.

REFERENCES

Brazil (2011). Lei Federal N° 12.527 de 18 de novembro de 2011. Disponível em: http://www.planalto.gov.br/ ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: junho 2021.

Open Knowledge Foundation. Manual dos dados abertos: governo traduzido e adaptado de opendatamanual.org. (2011). Disponível em: https://www.w3c.br/pub /Materiais/PublicacoesW3C/Manual_Dados_Abertos_ WEB.pdf. Acesso em: junho de 2021.

André Eberhardt and Milene Selbach Silveira. (2018). Show me the Data! A Systematic Mapping on Open Government Data Visualization. In dg.o '18: dg.o 2018: Proceedings of the 19th Annual International Conference on Digital Government Research, May 30-June 1, 2018, Delft, Netherlands. ACM, New York, NY, USA, 10 pages. https://doi.org/10. 1145/3209281.3209337.

Emilio Feliciano de Oliveira and Milene Selbach Silveira. (2018). Open Government Data in Brazil A Systematic Review of its Uses and Issues. In dg.o '18: dg.o 2018: Proceedings of the 19th Annual International Conference on Digital Government Research, May 30-June 1, 2018, Delft, Netherlands. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3 209281.3209339.

Arief Bachtiar, Suhardi, Wardani Muhamad. (2020). Literature Review of Open Government Data. International Conference on Information Technology Systems and Innovation (ICITSI) Bandung - Padang, October 19 - 23, 2020 ISBN: 978-1-7281-8196-7 ©2020 IEEE.

Karina Cenci, Pablo Fillottrani, and Jorge Ardenghi. (2017). Government Data Interoperability: a Case Study from Academia. In Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance (ICEGOV '17). Association for Computing Machinery, New York, NY, USA, 625–628. DOI: https://doi.org/10.1145/3047273.3047382.

P. Müller, J. Gil-Garcia, C. Tirelli. (2018). The Impact of Political, Technological and Social Variables on the Development of Local EGovernment: Lessons from Brazil. In Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance, Galway, Ireland, April 2018 (ICEGOV'18), 10 pages. DOI: 10.1145/32094 15.3209440.

K.Younsi Dahbi, H. Lamharhar, D. Chiadmi (2018). Exploring dimensions influencing the usage of Open Government Data Portals. In Proceedings of 12th International Conference on Intelligent Systems: theories and applications, Rabat Morocco, October 2018 (SITA'18), 6 pages. DOI: 10.1145/ 3289402.3289526.

Tian, Q., Han, Z., Yu, P. et al. (2021). Application of openEHR archetypes to automate data quality rules for electronic health records: a case study. BMC Med Inform Decis Mak 21, 113 (2021). https://doi. org/10.1186/s12911-021-01481-2

Bin Pan, Guomin Zhang, Xuepei Qin. (2018). Design and Realization of an ETL Method in Business Intelligence Project. 3rd IEEE International Conference on Cloud Computing and Big Data Analysis.

Kunjian Sun, Yuqing Lan. (2012). SETL: A Scalable and High Performance ETL System. 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization.

Saranya N, Brindha R, Aishwariya N, Kokila R, Matheswaran P, Poongavi P. (2021). Data Migration using ETL Workflow.7th International Conference on Advanced Computing & Communication Systems (ICACCS).

Sreemathy J, Infant Joseph V, Nisha S, Chaaru Prabha I, Gokula Priya RM. (2020). Integração de dados em ETL usando TALEND. 6th International Conference on Advanced Computing and Communication Systems (ICACCS).

Pristi Sukmasetya, Harry Budi Santoso, Dana Indra Sensuse. (2018). Current E-Government Public Service on User Experience Perspective in Indonesia. 2018 International Conference on Information Technology Systems and Innovation (ICITSI), 2018, pp. 159-164, doi: 10.1109/ICITSI.2018.8695962. Acesso em: junho de 2021.

W. A. Chapetta. (2006). Uma Infra-estrutura para Planejamento, Execução e Empacotamento de Estudos Experimentais em Engenharia de Software. Ph. D. Dissertation. Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ, Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

Jianyang Yu, Yuanyuan Qiao, Nanfei Shu, Kewu Sun, Shenshen Zhou, and Jie Yang. (2019). Neural Network Based Transaction Classification System for Chinese Transaction Behavior Analysis. In 2019 IEEE International Congress on Big Data (BigDataCongress). 64–71. https://doi.org/10.1109/BigDataCongress.2019.00021

Transparency International (2020). Índice de Percepção da Corrupção 2020. Disponível em: <https://comunidade.transparenciainternacional.org.br/ipc-indice-de-percepcao-da-corrupcao-2020>. Acesso em junho: de 2021.

Shan Bergin and Paul Wraight. (2006). Silver based wound dressings and topical agents for treating diabetic foot ulcers (Review). Cochrane database of systematic reviews (Online) 2006 (01 2006), CD005082. https://doi.org/10.1002/14651858. CD005082.pub2

Cristina Mamédio da Costa Santos, Cibele Andrucioli de Mattos Pimenta, and Moacyr Roberto Cuce Nobre. (2007). A estratégia PICO para a construção da pergunta de pesquisa e busca de evidências. Revista Latino-Americana de Enfermagem 15, 3 (jun. 2007), 508–511. https://doi.org/10.1590/S0104-11692007000300023

Barbara Kitchenham. (2004). Procedures for Performing Systematic Reviews. Keele, UK, Keele Univ. 33 (08 2004).

Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. Information

and Software Technology 64 (08 2015). https://doi.org/10.1016/j.infsof. 2015.03.007

Capes. (2021). Portal de periódicos CAPES/MEC [Journal Portal CAPES/MEC]. Disponível em: https://www.periodicos.capes.gov.br. Acesso em: 01 dezembro 2021.

Jakub Klímek and Petr Škoda. (2017). LinkedPipes ETL in Use: Practical Publication and Consumption of Linked Data. In Proceedings of The 19th International Conference on Information Integration and Web-based Applications &Services, Salzburg, Austria, December 4–6, 2017 (iiWAS '17), 5 pages. https://doi.org/10.1145/3151759.3151809.

Berro, A., Megdiche, I., Teste, O. (2015). A Content-Driven ETL Processes for Open Data. New Trends in Database and Information Systems II. Advances in Intelligent Systems and Computing, vol 312. Springer, Cham. https://doi.org/10.1007/978-3-319-10518-5_3

Ramón Fernando Colmenares-Quintero, Gina Paola Maestre-Gongora, Lina Judith Pacheco-Moreno, Natalia Rojas, Kim E. Stansfield & Juan Carlos Colmenares-Quintero (2021) Analysis of the energy service in non-interconnected zones of Colombia using business intelligence, Cogent Engineering, 8:1, 1907970, DOI: 10.1080/23311916.2021.1907970. : https://doi.org/10.1080/23311916.2021.1907970