GARBAGE IN, GARBAGE OUT (GIGO): ENFRENTANDO ESTA MÁXIMA NOS CONJUNTOS DE DADOS ASSOCIADOS AO PROGRAMA DINHEIRO DIRETO NA ESCOLA (PDDE)

Garbage In, Garbage Out (GIGO): Confronting this Maxim in the Datasets Associated with the Money Direct to School Program (MDSP)

Garbage In, Garbage Out (GIGO): Enfrentando este máximo en los conjuntos de datos asociados al Programa Dinero Directo en la Escuela (PDDE)

Guilherme Ataíde Dias
Doutor, Universidade de São Paulo (USP), São Paulo, SP, Brasil.
Professor Titular, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
URL:Lattes: http://lattes.cnpq.br/9553707435669429
ORCID: https://orcid.org/0000-0001-6576-0017

Wagner Junqueira de Araújo
Doutor, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Professor Associado III, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Lattes: http://lattes.cnpq.br/6762905361803183
ORCID: https://orcid.org/0000-0002-2301-4996

Adriana Valéria Santos Diniz
Doutora, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Professora Associado II, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Lattes: http://lattes.cnpq.br/7196551398849603
ORCID: https://orcid.org/0000-0002-2720-2433

Flavio Ribeiro Córdula
Doutor, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Analista de TI, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Lattes: http://lattes.cnpq.br/7466802181232338
ORCID: https://orcid.org/0000-0001-8680-9190

Paulo Roberto Santos Costa
Mestre, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Cargo ocupado, Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), João Pessoa, PB, Brasil.
Lattes: http://lattes.cnpq.br/7257600761427884
ORCID: https://orcid.org/0000-0001-5833-0676

Resumo

Introdução: A investigação apresenta e discute a relevância da qualidade dos dados na Ciência de Dados, trazendo o conceito de "Garbage In, Garbage Out" (GIGO) ao Programa Dinheiro Direto na Escola (PDDE). **Objetivos**: Teve como objetivo descrever o processo de extração, transformação e carga de dados para a geração de painéis de informação pelo Centro Colaborador de Apoio ao Monitoramento e à Gestão de Programas Educacionais (CECAMPE/NE). **Metodologia**: A metodologia empregada configura-se como quantitativa e utilizou ferramentas típicas da Ciência dos Dados. Os dados relacionados com o PDDE foram coletados de sistemas transacionais e de enquetes com as populações envolvidas. **Resultados**: Os resultados mostraram uma redução satisfatória do GIGO, embora tenham sido identificados desafios como divergências de atualização de dados, falta de documentação adequada das tabelas de dados e problemas com campos de digitação livre nos formulários das enquetes. **Conclusão**: A análise reitera que o conceito de GIGO é um desafio significativo para a utilização eficaz dos dados do PDDE e destaca a necessidade de elaboração de melhores práticas de gestão de dados no contexto do programa e que todos os profissionais que usam dados como a matéria prima para a realização de suas atividades profissionais devem estar conscientes desses desafios e trabalhar em prol de soluções eficazes.

Palavras-chave: Programa Dinheiro Direto na Escola; Garbage in Garbage out; qualidade de dados; painéis de informação; centro colaborador de apoio ao monitoramento e à gestão de programas educacionais.

Abstract

***Introduction**: The investigation presents and discusses the relevance of data quality in Data Science, introducing the concept of "Garbage In, Garbage Out" (GIGO) to the Money Direct to School Program (PDDE). **Objectives**: The objective was to describe the process of data extraction, transformation, and loading for the generation of information panels by the Collaborative Support Center for Monitoring and Management of Educational Programs (CECAMPE/NE). **Methodology**:The employed methodology was quantitative and used typical Data Science tools. The data related to the PDDE were collected from transactional systems and surveys with the involved populations. **Results**: The results showed a satisfactory reduction of GIGO, although challenges were identified such as data update discrepancies, lack of adequate documentation of data tables, and issues with free typing fields in survey forms. **Conclusion**: The analysis reiterates that the GIGO concept is a significant challenge for the effective use of PDDE data and emphasizes the need to develop better data management practices in the context of the program. It suggests that all professionals who use data as raw material for their professional activities should be aware of these challenges and work towards effective solutions.*

***Keywords**: Money Direct to School Program; Garbage in Garbage out; data quality; dashboards; collaborative center for support in monitoring and management of educational programs.*

Resumen

***Introducción**: La investigación presenta y discute la relevancia de la calidad de los datos en Ciencia de Datos, introduciendo el concepto de "Garbage In, Garbage Out" (GIGO) en el Programa Dinero Directo en la Escuela (PDDE). **Objetivos**: El objetivo era describir el proceso de extracción, transformación y carga de datos para la generación de paneles de información por el Centro Colaborador de Apoyo al Monitoreo y Gestión de Programas Educativos (CECAMPE/NE). **Metodologia**: La metodología empleada fue cuantitativa y utilizó herramientas típicas de Ciencia de Datos. Los datos relacionados con el PDDE se recogieron de sistemas transaccionales y encuestas con las poblaciones involucradas. **Resultados**: Los resultados mostraron una reducción satisfactoria de GIGO, aunque se identificaron desafíos como discrepancias en la actualización de datos, falta de documentación adecuada de las tablas de datos y problemas con campos de digitación libre en los formularios de las encuestas. **Conclusión**: El análisis reitera que el concepto de GIGO es un desafío significativo para el uso efectivo de los datos del PDDE y enfatiza la necesidad de desarrollar mejores prácticas de gestión de datos en el contexto del programa. Sugiere que todos los profesionales que utilizan datos como materia prima para sus actividades profesionales deben ser conscientes de estos desafíos y trabajar hacia soluciones efectivas.*

***Palabras clave**: Programa Dinero Directo en la Escuela; Garbage in Garbage out; calidad de datos; paneles de información; centro colaborador de apoyo al monitoreo y gestión de programas educativos.*

> "Garbage in, garbage out" provavelmente é a primeira lição que os aspirantes a cientistas de dados aprendem sobre suas futuras empreitadas analíticas." (OZMINKOWSKI, 2021, p. 1, tradução nossa1).

1 Texto no original: "*Garbage in, garbage out* is probably the first lesson budding data scientists learn about their forthcoming analytic endeavors."

DOI 10.22477/vi.widat.32

1. Introdução

Esse trabalho aborda sobre uma questão importante que muitos cientistas de dados podem enfrentar nas etapas iniciais de suas investigações, a garantia da qualidade dos dados obtidos. Qualidade esta que poderá ter repercussões positivas ou negativas nas análises a serem realizadas, bem como em todos os entregáveis resultantes dos conjuntos de dados utilizados. A questão do impacto da qualidade dos dados em processos computacionais e de informação são bastante reconhecidos desde meados do século XX, visto que, a partir do desenvolvimento e popularização dos primeiros computadores digitais com uma maior capacidade de processamento (*mainframes*) a quantidade de geração e processamento de dados vem crescendo, literalmente explodindo a partir dos anos 1990 com a massificação dos microcomputadores e uma pletora de outros dispositivos das Tecnologias Digitais da Informação e Comunicação (TDICs).

Associado ao processo de tratamento dados, mencionamos o conceito de *Garbage In, Garbage Out* (GIGO). Literalmente, essa expressão idiomática em Língua Inglesa, significaria em Língua Portuguesa algo como "Lixo entra, Lixo sai". Stenson (2016, *online*, tradução nossa2) explica que: "A ideia por trás de GIGO remonta ao próprio amanhecer da computação, no início do século XIX, quando Charles Babbage apresentou o projeto de sua 'máquina diferencial' ao Parlamento da Inglaterra.". O mesmo autor, explica ainda, que possivelmente a expressão foi criada por um funcionário da IBM no ano de 1958 ou 1959 durante um treinamento no computador 305 RAMAC para clientes da empresa em Nova Iorque.

O conceito de GIGO é mais impactante hoje do que em qualquer época, devido à nossa dependência que temos dos dados em todas as instâncias das atividades humanas. Mesmo como os avanços nos produtos de *software* que possibilitam a consistência nos processos de tratamento de dados, os desafios persistem. O impacto da possibilidade de entrada de "lixo" afeta todos os tipos de sistema, desde os sistemas tradicionais de folha de pagamento até os mais sofisticados sistemas de Inteligência Artificial (AI), baseados em modelos de aprendizado profundo, que são absolutamente dependentes de vastos volumes de dados.

Nesse contexto, a pesquisa tem como por objetivo descrever o processo de extração, transformação e carga de dados para a geração de painéis de informação pelo Grupo de Monitoramento do Centro Colaborador de Apoio ao Monitoramento e à Gestão de Programas Educacionais (CECAMPE/NE). Os dados obtidos são relativos ao Programa Dinheiro Direto na Escola (PPDE), uma iniciativa do Fundo Nacional de Desenvolvimento da Educação (FNDE). Diniz *et al.* (2022, p. 7-8) explicam que:

> O PDDE é um programa federal de transferência suplementar de recursos direto às instituições de ensino da educação básica pública. Ao lado de outros programas federais, a exemplo do Programa Nacional de Transporte Escolar (PNATE), Programa Caminho da Escola, Programa Nacional de Ali-

2 Texto no original: "The idea behind GIGO actually dates to the very dawn of computation, the early 19th century, when Charles Babbage presented the design for his "difference engine" to England's Parliament."

> mentação Escolar (PNAE), Programa Nacional de Livro Didático (PNLD), promove a melhoria física e pedagógica das instituições de ensino, contribuindo com o fortalecimento da gestão escolar, uma vez que abre espaço para a comunidade participar da tomada de decisão, tanto no que se refere aos aspectos administrativo-financeiros e didático-pedagógicos. (DINIZ *et al*., 2022, p. 7-8)

Córdula *et al*. (2022, p. 68) explicam o que são os Centros Colaboradores de Apoio ao Monitoramento e à Gestão de Programas Educacionais:

> Os Centros Colaboradores de Apoio ao Monitoramento e à Gestão de Programas Educacionais (CECAMPEs) são universidades vinculadas do Fundo Nacional de Desenvolvimento da Educação (FNDE) que realizam atividades de assistência técnica e monitoramento a estados, municípios e escolas, dando suporte para que estes possam aprimorar a execução e o desempenho do Programa Dinheiro Direto na Escola (PDDE) e suas Ações Integradas, do Programa Caminho da Escola e do Programa Nacional de Apoio ao Transporte do Escolar (PNATE).

No Nordeste o CECAMPE está representado pela Universidade Federal da Paraíba, doravante referido como CECAMPE/NE.

Os dados gerados no âmbito do PDDE após os devidos tratamentos (eliminar GIGO) foram utilizados para a elaboração de painéis de informações pela equipe monitoramento do CECAMPE/NE. Esses painéis fornecem subsídios necessários para que o FNDE, o CECAMPE/NE e toda a sociedade possam acompanhar os mais diversos indicadores acerca do PDDE. Destaca-se a possibilidade das equipes de campo do CECAMPE/NE realizarem intervenções quando em atividades nas escolas, mediante a consulta aos painéis, seja em uma escola da capital ou em uma escola situada em uma área rural no interior do Nordeste.[3]

3 Os painéis disponibilizados pelo CECAMPE/NE podem ser acessados através do seguinte URL: https://www.cecampe.ufpb.br/paineisdeinformacoes.

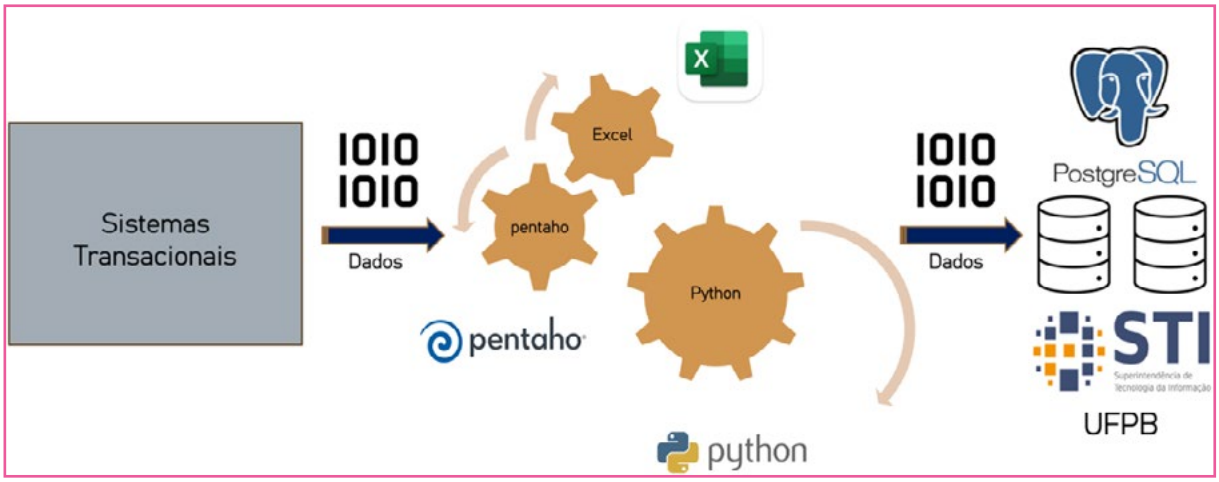
2. Procedimentos Metodológicos

Os dados obtidos para a elaboração dos painéis vieram de quatro fontes: 1) do FNDE, que aconteceu, primeiramente, a partir da Plataforma Ágil de Serviços de Dados do Banco Central do Brasil, conhecida como Olinda, e, posteriormente, por meio da plataforma de aplicação *Web* da Microsoft conhecida com *Sharepoint*; 2. Dados obtidos através de formulários gerados a partir de enquetes realizadas com o *Google Forms* com as populações atendidas pelo PDDE; 3. Microdados do Censo Escolar da educação básica, extraídos do *site* oficial do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP); e 4. da divisão territorial brasileira, que teve como fonte o *site* do Instituto Brasileiro de Geografia e Estatística (IBGE).

O período de coleta de dados compreendeu os anos de 2021 e 2023.

Quanto aos dados obtidos através da Plataforma Olinda e *Sharepoint*, eles foram submetidos às atividades de ETL (***E****xtract*, ***T****ransform* and ***L****oad*) (Vide Figura 1). Esse processo envolveu a extração de dados com *scripts* específicos da plataforma e sua subsequente transformação na estação de trabalho do cientista de dados, foram utilizando os seguintes produtos de *software*: *pentaho*, *Python, R* e *Microsoft Excel*. O formato final dos dados, após as devidas transformações, foi o *Comma-Separated Values* (CSV). Esses dados no formato CSV foram carregados em uma base de dados do gerenciador de banco de dados *PostgreSQL* da Superintendência de Tecnologia da Informação da UFPB (STI). O acesso a esse banco de dados é concedido apenas aos usuários devidos credenciados através do IP 150.165.130.10, sendo que, atualmente, essa comunidade é constituída majoritariamente pelos desenvolvedores de painéis.

Figura 1 – Captura de dados dos sistemas transacionais



Fonte: Desenvolvimento dos autores (2023)

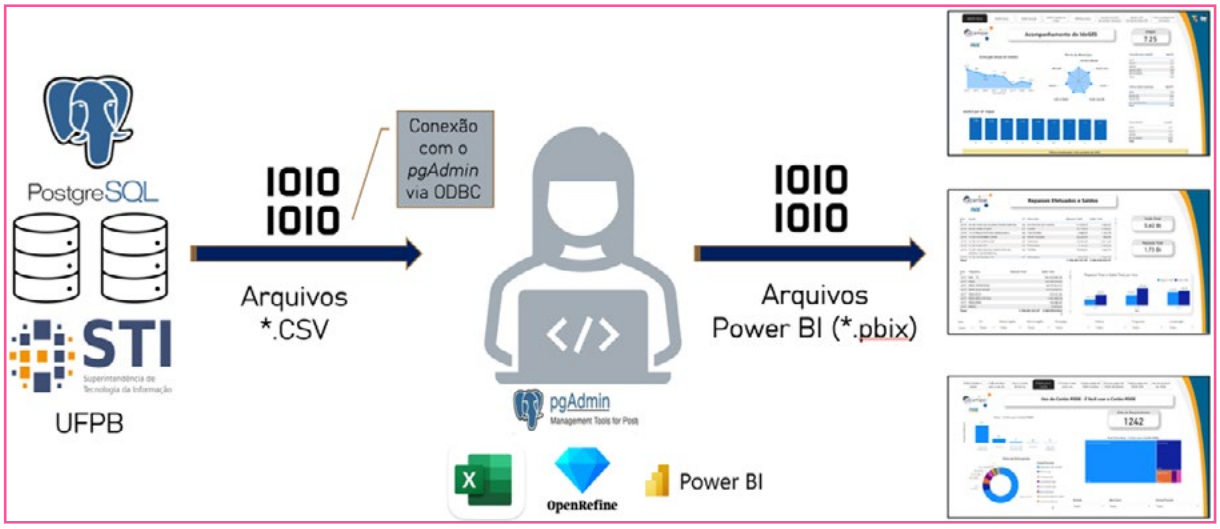
É importante mencionar que os dados extraídos da Plataforma Olinda, durante o período inicial dos desenvolvimentos, apresentavam divergências. Para sanar estas divergências, conforme descrito no relatório apresen-

tado pelo CECAMPE/NE (BRASIL, 2022), foi necessário a utilização de cinco bases distintas, providas pelo FNDE, conforme disponibilidade dos dados.

Os dados obtidos por meio de formulários gerados a partir de enquetes realizadas com o *Google Forms* foram exportados no formato CSV para o programa de planilhas eletrônicas *Microsoft Excel*, após passarem por uma etapa de transformações com o *software OpenRefine*. Este é um produto *Open Source* que facilita a limpeza e a transformação de dados, disponibilizando diversos algoritmos para verificar a consistência deles. Os dados obtidos através de enquetes realizadas com o *Google Forms* não foram disponibilizados no gerenciador de banco de dados *PostgreSQL* da STI/UFPB, mas armazenados no sistema de arquivos das estações de trabalho dos Cientistas de Dados.

A Figura 2 ilustra o processo da importação de dados a partir do gerenciador de banco de dados *PostgreSQL* da STI/UFPB pelos cientistas de dados e desenvolvedores de painéis.

Figura 2 – Produção de painéis a partir dos Bancos de SQL do STI/UFPB



Fonte: Desenvolvimento dos autores (2023)

Após a importação dos dados e a realização de eventuais ajustes que sejam necessários com o *Open Refine* ou outra ferramenta, esses dados são transferidos para o programa de *Business Analytics* da Microsoft, conhecido como *Power BI*. A aplicação gerada, é então disponibilizada na *Web* para consumo por toda a comunidade.

3. Resultados

Após a apresentação dos procedimentos metodológicos indicamos que a redução do GIGO no processo extração, transformação e carga dos dados do PDDE foram consideradas satisfatórias pelos cientistas de dados do CECAMPE/NE. Ficou evidenciado que as anomalias identificadas nos processos iniciais de extração, contribuíram para minorar resultados inconsistentes nos painéis de dados

Dentre os desafios encontrados no tratamento dos dados obtidos a partir da Plataforma Olinda mencionamos: 1. as divergências no intervalo temporal entre a atualização dos dados disponibilizados pelas bases do FNDE, SIGEF e SAE. Isso pode causar divergências em resultados calculados por aplicações que usem uma ou outra fonte de dados. O FNDE orientou que os dados usados para a elaboração de painéis passassem a ser obtidos através de sua plataforma interna Microsoft *Sharepoint* (DIAS, 2022); 2. a ausência de documentação associada às tabelas e a muitos dos seus respectivos atributos. Frequentemente, os identificadores utilizados para nomear as tabelas não eram suficientes para possibilitar uma compreensão efetiva da semântica desse objeto. Uma situação análoga também foi detectada nos nomes dados a alguns atributos das tabelas (DIAS, 2022). Por exemplo, uma tabela que modela os fornecedores de uma empresa devia ser chamada de "fornecedores" e não algo como "abend" ou "xpto". A mesma lógica se aplica aos atributos da tabela; 3. O uso de valores fora do padrão, como, por exemplo, o símbolo usado como separador de casas decimais ou formatos de datas diferentes do utilizado por padrão no Brasil.

Em relação aos desafios encontrados no tratamento dos dados obtidos por meio dos formulários gerados a partir das enquetes realizadas com o *Google Forms*, consideramos importantes mencionar um problema associado ao preenchimento de dados pelos respondentes dos questionários em campos de livre digitação. Por permitir a inserção de qualquer valor, um campo de livre digitação possibilita a captura de um mesmo conceito, representado por sintaxes diversas, o que demanda uma etapa adicional no tratamento dos dados. O Quadro 1 ilustra algumas das possibilidades associadas ao cargo de "Diretor(a)".

Quadro 1 - Problemas terminológicos em dados provenientes de enquetes.

Diretor	Diretor	Diretora Adjunta	Diretor adjunto
Diretor Adjunto	Diretor administrativo	Diretor Administrativo da Educação	Diretor de Cultura e Eventos
Diretor de Departa-mento de Gestão	Diretor de ensino	Diretor de Ensino	Diretor de escola Adjunto
Diretor de escola(a)	Diretor de projetos especiais	Diretor Escolar	Diretor Escolar Adjunto
Diretor Financeiro (Tesoureiro)	Diretor geral da Se-cretaria	Diretor na SEDUC	Diretora
Diretora da Direc	Diretora adjunta	Diretora Adjunta	Diretora Adjunta escolar
Diretora administrativa e financeira	Diretora da Semed	Diretora da área de prestação de contas	Diretora de Ensino
Diretora Financeira	Diretora NTE		

Fonte: Dias *et al*. (2022, p. 84).

O problema mencionado ocorreu na realização das primeiras enquetes, posteriormente os campos de livre digitação foram substituídos por opções de resposta com valores fixos (vide Figura 3).

Figura 3 - Produção de painéis a partir dos Bancos de SQL do STI/UFPB.

Fonte: Dias *et al*. (2022, p. 84)

O aprendizado alcançado pela equipe de cientistas de dados do CECAMPE/NE na redução do GIGO possibilitou a elaboração de produtos de informação mais fidedignos, com potencial para atender de maneira mais assertiva toda a comunidade de usuários.

4. Considerações Finais

Após o término do período de dois anos de coleta dos dados do PDDE, foi possível constatar a existência de questões importantes relacionadas à qualidade desses dados, conforme apresentado na seção anterior. Essas questões têm o potencial para afetar negativamente as análises e as conclusões que emergem desses dados, materializadas na forma de painéis, caso não tivessem sido efetivamente abordadas.

O conceito de GIGO, popular nos centros de processamento de dados na segunda metade do século XX, volta a ser proeminente em nossa contemporaneidade, caracterizada pela "datificação" de todos os aspectos da atividade humana. A partir desta investigação podemos concluir que a máxima GIGO representa um desafio real para a utilização efetiva dos dados do PDDE. De modo a perpassar este desafio, sugerimos a implementação de melhores práticas na gestão de dados no contexto do FNDE/PDDE, atentando para a padronização de todas as etapas relacionadas com um ciclo de vida de dados no contexto de atuação dos gestores de bancos de dados, cientistas de dados, desenvolvedores de aplicações e de todos(s) aqueles(as) que usam os dados como a matéria prima para a realização de suas atividades profissionais.

Referências

CORDULA, Flávio Ribeiro; DIAS, Guilherme Ataíde; COSTA, Paulo Roberto Santos; ARAÚJO, Wagner Junqueira; DINIZ, Adriana Valéria Santos. Desenvolvimento da Integração de Dados para o Projeto Cecampe - NE no Contexto do "Programa Dinheiro Direto Na Escola". *In*: SEMINÁRIO DO CECAMPE NORDESTE, 1., 2022, João Pessoa. **Anais [...]**. João Pessoa, PB: Editora do CCTA, 2022. v. 1, p. 67-68. Tema: Os novos gerenciamentos de ações para o fortalecimento do Programa Dinheiro Direto na Escola.

DIAS, Guilherme Ataíde; COSTA, Paulo Roberto Santos; ARAÚJO, Wagner Junqueira; CORDULA, Flávio Ribeiro; DINIZ, Adriana Valéria. Santos. Dados e painéis informacionais: insumos e tecnologias habilitadoras para a disseminação do conhecimento no contexto do "Programa Dinheiro Direto na Escola". *In*: DINIZ, Adriana Valéria Santos; PRESTES, Emília Maria da Trindade; SANTOS, José Lucas Batista dos; FITTIPALDI, Ítalo; PEREIRA, Maria Aparecida Nunes; ARAÚJO, Wagner Junqueira de. (org.). **Os novos gerenciamentos de ações para o fortalecimento do programa dinheiro direto na escola**. João Pessoa: Editora do CCTA, 2022, v. 1, p. 79-87.

DINIZ, Adriana Valéria Santos; PEREIRA, Maria Aparecida Nunes; ARAÚJO, Wagner Junqueira; FITTIPALDI, Ítalo. O fortalecimento do Programa Dinheiro Direto na Escola na Região Nordeste como estratégia para a gestão democrática e para a qualidade da educação. *In*: SEMINÁRIO DO CECAMPE NORDESTE, 1., 2022, João Pessoa. **Anais [...].** João Pessoa, PB: Editora do CCTA, 2022. v. 1, p. 6-23. Tema: Os novos gerenciamentos de ações para o fortalecimento do Programa Dinheiro Direto na Escola.

BRASIL. MINISTÉRIO DA EDUCAÇÃO. **O fortalecimento do Programa Dinheiro Direto na Escola na região Nordeste como estratégia para a gestão democrática e para a qualidade da educação CECAMPE--Região Nordeste**: relatório técnico: análise do IDEGes e sua evolução. João Pessoa, PB, 2022. Disponível em: https://bit.ly/42MiEP8. Acesso em: 21 maio 2023.

OZMINKOWSKY, Ron. Garbage In, Garbage Out. **Towards Data Science**, nov. 2021. Disponível em: https://bit.ly/46bbreK. Acesso em: 18 maio 2023.

STENSON, Rob. Is This the First Time Anyone Printed, 'Garbage In, Garbage Out'? **Atlas Obscura**, mar. 14, 2016. Disponível em: https://www.atlasobscura.com/articles/is-this-the-first-time-anyone-printed-garbage-in--garbage-out. Acesso em: 18 maio 2023.