Universidade de Brasília
Faculdade De Ciência Da Informação
Programa De Pós-Graduação Em Ciência Da Informação - PPGCINF

**GABRIEL SALDANHA OSTWALD CORBAL**

# ARQUITETURA DA INFORMAÇÃO: MODELO DE ANÁLISE DE DADOS ABERTOS DE UNIVERSIDADES FEDERAIS BRASILEIRAS

Brasília
2023

**GABRIEL SALDANHA OSTWALD CORBAL**

# ARQUITETURA DA INFORMAÇÃO: MODELO DE ANÁLISE DE DADOS ABERTOS DE UNIVERSIDADES FEDERAIS BRASILEIRAS

Dissertação apresentada como requisito para qualificação do Mestrado do Programa de Pós-Graduação em Ciência da Informação.

Orientador: Dr. Márcio de Carvalho Victorino.

Brasília

2023

Corbal, Gabriel Saldanha Ostwald.
Arquitetura da informação: modelo de análise de dados abertos de universidades federais brasileiras/ Gabriel Saldanha Ostwald Corbal. – Brasília, 2023.
76 f.

Orientador: Dr. Márcio de Carvalho Victorino.

Dissertação (Mestrado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília – UnB, Brasília, DF, 2023.

1. Dados abertos. 2. Dados públicos. 3. Administração pública. 4. Arquitetura da informação. 5. Ciência de dados. 6. Análise exploratória. 7. Tomada de decisão. 8. Business intelligence. I. Título.

CDU 004.62

Catalogação na publicação: Bibliotecário Alessandro Meneses da Silva – CRB1/2777

**UNIVERSIDADE DE BRASÍLIA**

**PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO**

**Ata Nº: 34**

Aos onze dias do mês de dezembro do ano de dois mil e vinte e três, instalou-se a banca examinadora de Dissertação de Mestrado do aluno Gabriel Saldanha Ostwald Corbal, matrícula 210007095. A banca examinadora foi composta pelos professores Dr. João de Melo Maricato / membro interno / PPGCINF/UnB, Dr. Felipe Lopes da Cruz / Membro externo / IDP, Dr. Dalton Lopes Martins / PPGCINF/UnB, Suplente e Dr. Marcio de Carvalho Victorino / orientador/presidente / PPGCINF/UnB. O discente apresentou o trabalho intitulado "ARQUITETURA DA INFORMAÇÃO: MODELO DE ANÁLISE DE DADOS ABERTOS DE UNIVERSIDADES FEDERAIS BRASILEIRAS".

Concluída a exposição, procedeu-se a arguição do(a) candidato(a), e após as considerações dos examinadores o resultado da avaliação do trabalho foi:

( X ) Pela aprovação do trabalho;

( ) Pela aprovação do trabalho, com revisão de forma, indicando o prazo de até 30 dias para apresentação definitiva do trabalho revisado;

( ) Pela reformulação do trabalho, indicando o prazo de **(Nº DE MESES)** para nova versão;

( ) Pela reprovação do trabalho, conforme as normas vigentes na Universidade de Brasília.

Conforme os Artigos 34, 39 e 40 da Resolução 0080/2021 - CEPE, o(a) candidato(a) não terá o título se não cumprir as exigências acima.

Dr. Marcio de Carvalho Victorino, PPGCINF/UnB

(Presidente/orientador)

Dr. João de Melo Maricato, PPGCINF/UnB

(Membro interno)

Dr.Felipe Lopes da Cruz, IDP

(Membro externo)

Dr. Dalton Lopes Martins, PPGCINF/UnB

(Suplente)

Gabriel Saldanha Ostwald Corbal

(Mestrando)

Documento assinado eletronicamente por **Marcio de Carvalho Victorino**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 27/12/2023, às 17:41, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Joao de Melo Maricato**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 02/01/2024, às 16:34, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Gabriel Saldanha Ostwald Corbal**, **Usuário Externo**, em 10/01/2024, às 12:13, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **FELIPE LOPES DA CRUZ**, **Usuário Externo**, em 18/01/2024, às 16:51, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Clovis Carvalho Britto**, **Coordenador(a) da Pós-Graduação da Faculdade de Ciência da Informação**, em 31/01/2024, às 10:42, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **10487498** e o código CRC **2BFB6095**.

**Referência:** Processo nº 23106.125673/2023-93 SEI nº 10487498

## RESUMO

Este trabalho aborda a Arquitetura da Informação (AI) aplicada à análise de dados abertos de universidades federais brasileiras, um tema relevante considerando o Brasil como um dos principais disseminadores de dados abertos no mundo. Apesar da disponibilidade desses dados, observa-se uma lacuna no aproveitamento efetivo das informações devido à falta de interesse e compreensão adequada do conteúdo disponibilizado. Visando superar essa barreira, a pesquisa propõe uma Arquitetura da Informação estruturada para facilitar a análise e interpretação desses dados.

A metodologia empregada envolve a Teoria do Enfoque Meta Analítico Consolidado (TEMAC) como forma de posicionar a pesquisa em ambito global, além do processo CRISP-DM, focando na preparação, inter-relação e detalhamento integrador dos dados. A fundamentação teórica abrange desde dados abertos e transparência pública até técnicas de análise de dados para geração de informação e conhecimento.

O cerne desta dissertação é a construção e validação de uma Arquitetura da Informação, tanto como processo quanto como estrutura, que foi submetida a uma prova de conceito. Esta consistiu na análise do orçamento da educação superior em algumas instituições de ensino superior brasileiras, fornecendo insights valiosos sobre a gestão de recursos e a tomada de decisões nessas universidades.

Através de uma abordagem prática, o estudo demonstra a eficácia da AI proposta, destacando sua aplicabilidade em facilitar o entendimento e a utilização dos dados abertos por parte dos gestores e da sociedade em geral. As considerações finais refletem sobre o impacto e as possibilidades futuras dessa abordagem no contexto da educação superior e da gestão de dados no Brasil.



**Palavras-chave:** dados abertos; dados públicos; administração pública; arquitetura da informação; análise de dados; transparência; *business intelligence.*

## ABSTRACT

This work addresses the application of Information Architecture (IA) to the analysis of open data from Brazilian federal universities, a significant topic considering Brazil as one of the main disseminators of open data globally. Despite the availability of these data, there is a gap in effectively leveraging the information due to a lack of interest and proper understanding of the content provided. Aiming to overcome this barrier, the research proposes a structured Information Architecture to facilitate the analysis and interpretation of these data.

The employed methodology involves the Consolidated Meta-Analytical Approach Theory (TEMAC) to position the research in a global context, along with the CRISP-DM process, focusing on the preparation, interrelation, and integrated detailing of the data. The theoretical foundation covers everything from open data and public transparency to data analysis techniques for information and knowledge generation.

The core of this dissertation is the construction and validation of an Information Architecture, both as a process and a structure, which was subjected to a proof of concept. This consisted of analyzing the higher education budget in some Brazilian higher education institutions, providing valuable insights into resource management and decision-making in these universities.

Through a practical approach, the study demonstrates the effectiveness of the proposed IA, highlighting its applicability in facilitating the understanding and use of open data by managers and society in general. The final considerations reflect on the impact and future possibilities of this approach in the context of higher education and data management in Brazil.



**Keywords:** open data; public data; public administration; information architecture; data analysis; transparency; business intelligence.

## LISTA DE FIGURAS

SUMÁRIO

## 1 INTRODUÇÃO

O dia 18 de novembro de 2011 é um marco da transparência no Brasil, quando foi sancionada a Lei 12.527/2011, conhecida como Lei de Acesso à Informação (LAI), que regulamentou o direito de acesso a várias informações públicas por meio da política de publicação de dados abertos estabelecida no Decreto nº 8.777 de 11 de maio de 2016.

Esses dados são publicados no Portal Brasileiro de Dados Abertos (Brasil, [2023a]) e no Portal da Transparência (Brasil, [2023b]), com o objetivo de disponibilizar informações sobre uma ampla gama de tópicos relacionados à administração pública. No entanto, esses portais fornecem apenas dados abertos, os arquivos ou documentos com algum tipo de sigilo não são publicados no portal.

Embora os dados estejam disponíveis no portal, o acesso e a compreensão desses dados ainda são restritos a profissionais técnicos, que às vezes têm dificuldade em navegar e encontrar as informações que precisam. Esse efeito foi apontado por Wersig (1993) e está relacionado à credibilidade do conhecimento por meio da tecnologia que permite a observação. Com a sofisticação das teorias, metodologias e técnicas de coleta e processamento de dados, o conhecimento produzido fica restrito a certos grupos ou comunidades (Nhacuongue; Ferneda, 2015).

Compreendendo o papel da tecnologia como um grande instrumento político, educacional e social, é necessário encontrar maneiras de organizar e democratizar o acesso às informações disponíveis pelo governo federal para uso da população.

Uma das informações mais significativas deste portal é a publicação dos dados relacionados a Lei Orçamentária Anual (LOA), que estabelece os orçamentos da União e cuja aprovação e confecção cabe ao Congresso Nacional (Câmara dos Deputados, 2022). Dentro da LOA encontramos os orçamentos de praticamente todos os órgãos governamentais, sendo possível compara-los e entender para onde está indo o dinheiro do contribuinte.

Com uma quantidade avassaladora de dados presentes na LOA, o recorte proposto do estudo se dá especificamente na área da educação superior do país, tema constantemente em discussão e sendo parte significativa da vida e da formação

de diversos brasileiros. O debate em relação ao ensino superior gratuito do país, deve seguir com lentes justas em relação ao que estamos julgando. Partindo desse pressuposto, as verbas que o Ministério da Educação fornece servem como base de análise.

## 1.1 Justificativa

O Brasil é um dos países mais transparentes do mundo, atualmente no top 30 de transparência pública de acordo com o Ranking da instituição "*Corruption Risk*" (Corruption [...], [2023]). Mesmo com essa massiva divulgação de dados abertos, o entendimento desses dados ainda não se transformou em informação prática para o cidadão.

A ciência, em sua essência, busca a compreensão e a explicação dos fenômenos, visando alcançar um consenso racional sobre diversas áreas do conhecimento, conforme destacado por Ziman (1968). Nesse contexto, a interdisciplinaridade emerge como uma abordagem poderosa, permitindo que diferentes campos do saber se complementem e enriqueçam mutuamente. A Ciência da Informação e a Ciência da Computação, embora distintas em suas origens e objetivos primários, apresentam uma confluência natural quando se trata de gerenciar, processar e disseminar informações em um mundo cada vez mais digitalizado.

A Ciência da Informação, por sua vez, foca na organização, representação, recuperação e disseminação da informação. Ela se preocupa com a estruturação e categorização da informação, garantindo que ela seja acessível e compreensível. Por outro lado, a Ciência da Computação oferece as ferramentas e técnicas necessárias para implementar, de forma prática, as teorias e estratégias propostas pela Ciência da Informação. Em outras palavras, enquanto a Ciência da Informação define "o quê" e "por que", a Ciência da Computação responde "como".

Unir essas duas disciplinas, portanto, não é apenas uma questão de conveniência, mas uma necessidade para enfrentar os desafios contemporâneos da gestão da informação. Ao combinar a perspectiva organizacional da Ciência da Informação com as capacidades técnicas da Ciência da Computação, é possível desenvolver soluções mais robustas, eficientes e adaptadas às demandas atuais. Esta sinergia entre os campos permite uma abordagem holística, onde a teoria e a prática

se entrelaçam de maneira harmoniosa, potencializando os resultados e contribuições para a sociedade.

O Portal da Transparência, desde sua implementação, tornou-se uma ferramenta essencial para o exercício da cidadania no Brasil. Além de servir como um mecanismo de controle e fiscalização do uso dos recursos públicos, o portal proporcionou aos cidadãos a capacidade de monitorar gastos, contratos, salários de servidores e diversas outras informações relacionadas à gestão pública.

Muitos cidadãos, jornalistas e organizações da sociedade civil utilizam o portal para realizar investigações, reportagens e auditorias cidadãs, contribuindo para a identificação e denúncia de irregularidades e malversações. Além disso, a disponibilidade desses dados em formato aberto tem permitido o desenvolvimento de aplicativos e plataformas de terceiros que facilitam a visualização e análise das informações, ampliando ainda mais o alcance e a utilidade do portal para a população. Este portal tem sido fundamental para fortalecer a democracia, promover a responsabilidade governamental e incentivar a participação ativa dos cidadãos na gestão pública.

Buscando incentivar outras pesquisas e facilitar futuras análises exploratórias dos dados públicos brasileiros por outros pesquisadores, serve como motivação desta pesquisa, documentar e mapear todo o processo de entendimento das fontes de dados que serão analisadas neste trabalho. Com diversos dados abertos presentes na plataforma, este estudo busca organizar e entender como cruzar diferentes fontes de dados presentes em portais públicos de dados abertos. Os gastos relacionados ao investimento na educação superior serão o tema central deste trabalho, através de uma perspectiva da ciência da informação e sua relação com a tecnologia.

Iniciativas públicas de facilitação da divulgação de dados abertos podem ser vistas em diferentes formas ao redor do mundo, tendo como inspiração o projeto *NYC OPEN DATA*, que é um projeto desenvolvido pelo Departamento de Análise de Dados do Governo dos Estados Unidos e pelo Departamento de Informação, Tecnologia e Telecomunicações de Nova Iorque. A missão deste portal é engajar os nova-iorquinos a utilizarem e conhecerem as informações produzidas e utilizadas pelo governo da cidade. Este é um portal centrado no usuário que publica bancos de dados já limpos

e normalizados para análise, além de possuir funcionalidades de *dashboards* dentro da própria plataforma (NYC Open Data, 2021).

Diferente da proposta do portal da NYC OPEN DATA, que parte do governo entregar as informações tratadas e resumidas, está sendo proposto municiar o cidadão através de técnicas e ferramental necessário para impulsionar pesquisas e descobertas próprias do interesse individual de cada cidadão.

## 1.2 Objetivos

A proposta de estudo tem por objetivo geral desenvolver uma arquitetura da informação para a análise de conjuntos de dados públicos originários de diferentes fontes relacionadas à educação superior.

Para alcançar o objetivo geral seguiremos os seguintes objetivos específicos:

- Identificar os conjuntos de dados abertos relacionados a educação superior e dados correlacionados;
- Fazer uma análise exploratória dos dados verificando possibilidades de análise e sua qualidade em ambiente local;
- Propor um modelo relacional para dados de diferentes fontes;
- Propor uma arquitetura da informação para análise dos dados relacionados;
- Implementar uma prova de conceito desta arquitetura da informação e gerar *Dashboards* que apresentem a análise dos dados tratados.

## 2 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa é definida como aplicada, uma vez que se direciona à resolução de um problema específico e tangível. Adota uma abordagem quantitativa, baseando-se em dados numéricos e estatísticos para suas análises. Em termos de objetivos, ela é tanto exploratória quanto descritiva, o que significa que busca aprofundar a compreensão sobre um fenômeno e, simultaneamente, descrevê-lo detalhadamente.

O modelo experimental é empregado, caracterizando-se pela execução de uma intervenção e subsequente mensuração de seus efeitos. Em relação aos métodos adotados, a pesquisa se vale de técnicas avançadas de análise de dados e de uma prova de conceito, que se manifesta na proposta de desenvolvimento de uma arquitetura integral de informação, com o intuito de disponibilizar os dados coletados e analisados.

Os procedimentos metodológicos adotados estão estruturados em duas partes principais, a primeira consiste em uma análise fundamentada na Teoria do Enfoque Meta Analítico Consolidado (TEMAC), enquanto a segunda parte se concentra em uma prova de conceito baseada no processo Cross-Industry Standard Process for Data Mining (CRISP-DM).

A revisão sistemática da literatura, mesmo não sendo um objetivo deste trabalho, visa auxiliar esta pesquisa a encontrar os principais estudos e autores para fornecer um embasamento teórico sólido, além de posicionar a presente pesquisa no panorama de estudos similares conduzidos globalmente. Esta revisão sistemática é orientada pela TEMAC, conforme proposto por Mariano e Santos (2017). Para enriquecer a revisão de literatura, foram incorporados textos e autores advindos das disciplinas da Faculdade de Ciência da Informação.

Como base metodológica será utilizado o *framework* CRISP-DM. Um *framework* que aborda de maneira integradora projetos que envolvem dados. É composto por seis fases, a compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação.

## 2.1 Teoria do Enfoque Meta Analítico Consolidado (TEMAC)

### *2.1.1 Preparação da pesquisa*

De acordo com a literatura, na primeira etapa define-se os termos de pesquisa a serem realizados nas bases científicas. Duas principais bases foram definidas para a pesquisa, a Web of Science (WOS) e Scopus tanto pelo seu prestígio, quanto pela estrutura otimizada de filtros e possibilidades de exploração de documentos relacionados à pesquisa.

Para realizar a busca de artigos relacionados nas plataformas internacionais de pesquisa, o idioma padrão utilizado para a criação do conjunto de palavras-chave escolhido foi o inglês, com o objetivo de encontrar referências internacionais e entender o debate mundial sobre o tema.

Os termos utilizados para filtro em cada uma das plataformas foram:

1. Web of Science: 137 resultados para:

   (ALL=("open data" OR "public data" OR "public datasets") AND ALL=("government" OR "public administration" OR "public sector") AND ALL=("data analysis" OR "EDA" OR "ADE" OR "exploratory analysys" OR "decision making")) AND (PY=(2007-2022) AND TASCA==("COMPUTER SCIENCE INFORMATION SYSTEMS" OR "INFORMATION SCIENCE LIBRARY SCIENCE" OR "COMPUTER SCIENCE THEORY METHODS" OR "COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS" OR "PUBLIC ADMINISTRATION"))

2. Scoups:

   (ALL("open data" OR "public data" OR "public datasets") AND ALL("government" OR "public administration" OR "public sector") AND ALL("data analysis" OR "EDA" OR "ADE" OR "exploratory analysys" OR "decision making")) AND ( LIMIT-TO ( SUBJAREA,"SOCI" ) OR LIMIT-TO ( SUBJAREA,"COMP" ) OR LIMIT-TO ( SUBJAREA,"DECI" ) ) AND

( LIMIT-TO ( PUBYEAR,2022) OR LIMIT-TO ( PUBYEAR,2021) OR LIMIT-TO ( PUBYEAR,2020) OR LIMIT-TO ( PUBYEAR,2019) OR LIMIT-TO ( PUBYEAR,2018) OR LIMIT-TO ( PUBYEAR,2017) OR LIMIT-TO ( PUBYEAR,2016) OR LIMIT-TO ( PUBYEAR,2015) OR LIMIT-TO ( PUBYEAR,2014) OR LIMIT-TO ( PUBYEAR,2013) OR LIMIT-TO ( PUBYEAR,2012) OR LIMIT-TO ( PUBYEAR,2011) OR LIMIT-TO ( PUBYEAR,2010) OR LIMIT-TO ( PUBYEAR,2009) OR LIMIT-TO ( PUBYEAR,2008) OR LIMIT-TO ( PUBYEAR,2006) ) AND ( LIMIT-TO ( EXACTKEYWORD,"Open Data" ) OR LIMIT-TO ( EXACTKEYWORD,"Decision Making" ) )

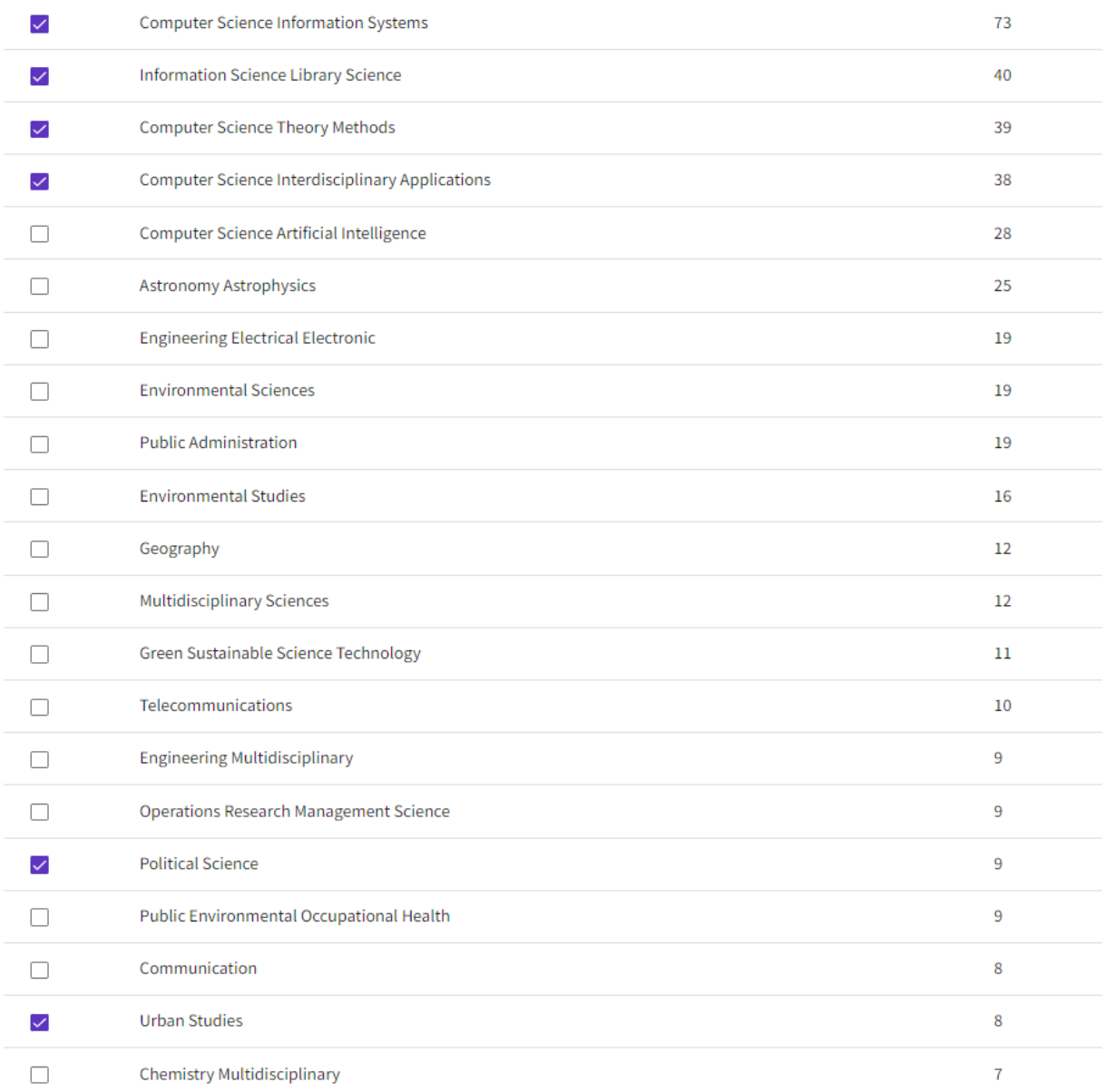
O ponto inicial da pesquisa começou a partir do termo *“Open Data”*. Com os resultados em mãos, foi-se adicionando sinônimos relacionados ao tema e afunilando os resultados para se aproximar do objeto de estudo. Primeiro afunilando para resultados relacionados a repositórios abertos de dados, que vem de Governos.

Por ter como objetivo de estudo análises exploratórias de dados e empoderar o cidadão através de informação, outros termos de pesquisa complementares foram adicionados, relacionados a Análise Exploratória de Dados e sinônimos acoplados com a tomada de decisão.

Esse procedimento foi repetido algumas vezes em ambas as plataformas até praticamente todos os objetos de pesquisa terem algo relacionado ao tema e objetivos de pesquisa.

Na plataforma WOS ainda foi realizado um filtro de acordo com as categorias pré-disponibilizadas na plataforma, a fim de direcionar ainda mais o referencial teórico proposto. Antes dessa filtragem, as duas maiores categorias de artigos publicados no filtro são Ciência da Computação e Sistemas da Informação com 73 publicações e em seguida Ciência da Informação e Biblioteconomia com 40 publicações. Ambas disciplinas servem de base para este trabalho.

**Figura 1** - Quantidade de artigos por categorias no Web of Science



| | | |
|---|---|---|
| ☑ | Computer Science Information Systems | 73 |
| ☑ | Information Science Library Science | 40 |
| ☑ | Computer Science Theory Methods | 39 |
| ☑ | Computer Science Interdisciplinary Applications | 38 |
| ☐ | Computer Science Artificial Intelligence | 28 |
| ☐ | Astronomy Astrophysics | 25 |
| ☐ | Engineering Electrical Electronic | 19 |
| ☐ | Environmental Sciences | 19 |
| ☐ | Public Administration | 19 |
| ☐ | Environmental Studies | 16 |
| ☐ | Geography | 12 |
| ☐ | Multidisciplinary Sciences | 12 |
| ☐ | Green Sustainable Science Technology | 11 |
| ☐ | Telecommunications | 10 |
| ☐ | Engineering Multidisciplinary | 9 |
| ☐ | Operations Research Management Science | 9 |
| ☑ | Political Science | 9 |
| ☐ | Public Environmental Occupational Health | 9 |
| ☐ | Communication | 8 |
| ☑ | Urban Studies | 8 |
| ☐ | Chemistry Multidisciplinary | 7 |

Fonte: Web Of Science, 2023.

Na busca do Scopus também foi aplicado um filtro relacionado a categorias nativas do próprio site, limitando os estudos nas categorias de Ciências Sociais Aplicadas, Ciência da Computação e Ciências de Decisão.

Também dentro da plataforma Scopus, para reduzir e afunilar a quantidade de artigos relacionados às buscas foram limitadas para trabalhos que contenham as palavras exatas Dados Abertos e Tomada de Decisão.

Do ponto de vista temporal, para diminuir a quantidade de trabalhos fora do contexto atual de disponibilidade e processamento de dados foi-se aplicado um filtro

contendo somente as publicações que ocorreram entre 2007 e 2022 nas duas plataformas.

### *2.1.2 Apresentação de interrelação dos dados*

Com duas bases de dados diferentes, a análise sobre os artigos foi feita separadamente de acordo com cada base, sempre buscando similaridades.

No Web of Science, após a filtragem final, restaram 137 artigos. Os principais autores que publicaram sobre o assunto foram:

- Janssen M - 8 artigos
- Bernardini F - 4 artigos
- Viterbo J - 4 artigos
- Choi J - 3 artigos
- Ishikawa E - 3 artigos
- Kleiman F - 3 artigos

No Scopus após a filtragem final, restaram 1035 documentos. Os principais autores que publicaram sobre o assunto foram:

- Janssen M - 38 artigos
- Zuiderwijk A - 14 artigos
- Luthfi A - 11 artigos
- Tarabanis K - 9 artigos
- Kleiman F - 8 artigos
- Leung C.K - 8 artigos

Janssen M aparece na primeira posição de quantidade de publicações relacionadas ao tema nas duas plataformas, podendo ser considerado como o pesquisador com mais volume e disponibilidade de informações sobre o tema e será ponto chave do referencial teórico deste trabalho. Outro autor que aparece no ranking de ambas plataformas é o Kleiman F., que também é citado e sua obra explicada no decorrer do trabalho.

Os países com maior publicações relacionadas ao tema com pelo menos 8 artigos publicados na WOS foram:

- Estados Unidos da América - 20 artigos
- Espanha - 17 artigos
- Holanda - 15 artigos
- Alemanha - 14 artigos
- Brasil - 12 artigos
- Inglaterra - 9 artigos
- China - 8 artigos
- Coréia do Sul - 8 artigos

Os países com maior publicações relacionadas ao tema com pelo menos 8 artigos publicados na plataforma Scopus foram:

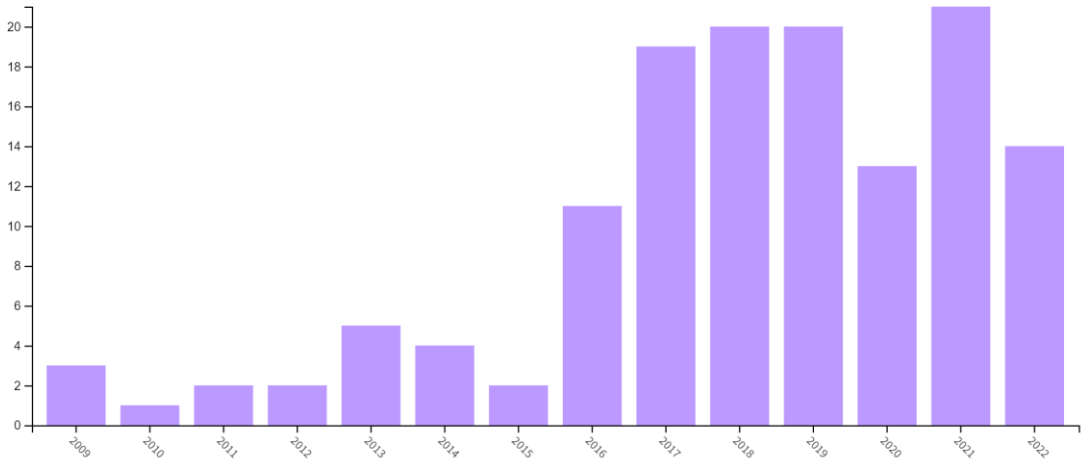
- Estados Unidos da América - 143 artigos
- Reino Unido - 110 artigos
- Holanda - 92 artigos
- Espanha - 76 artigos
- Alemanha - 69 artigos
- Brasil - 56 artigos
- Austrália - 52 artigos

- China - 52 artigos

O Brasil é um grande publicador de temas relacionados a dados abertos públicos associados à tomada de decisão e aparece sempre nas primeiras colocações da lista de acordo com a quantidade de artigos publicados.

Uma análise dos anos de publicação, sendo que o primeiro artigo é de 2009 e percebe-se uma crescente da quantidade de artigos principalmente depois de 2015. O gráfico da Figura 2 foi exportado da Web Of Science e mostra que o tema está aquecido porém ainda há espaços de aprofundamento da discussão.

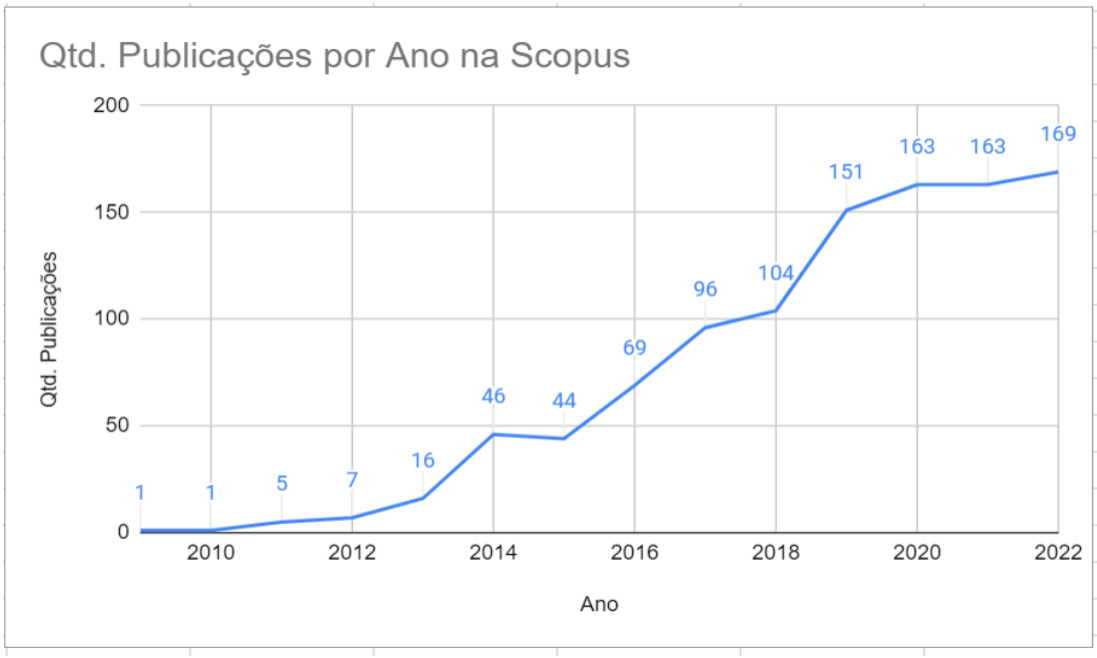
**Figura 2** - Quantidade de artigos por ano no Web of Science




Fonte: Web Of Science, 2023.

Na base da Scopus a crescente da quantidade de artigos relacionados ao tema foi ainda mais atenuada, contendo um volume baixo de artigos pré 2015 e aumentando de maneira muito significativa nos anos seguintes, em constante crescente.

**Figura 3** - Quantidade de artigos por ano no Scopus




Fonte: Adaptação da extração dos dados da Scopus, 2023.

Os cinco artigos mais citados na Web of Science são demonstrados na Figura 4:

**Figura 4** - Artigos mais citados no Web of Science

| Nome do Artigo | Data | Autor | Citações |
|---|---|---|---|
| Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities | 03/06/2020 | Ricardo Matheus, Marijn Janssen | 88 |
| When transparency and collaboration collide: The USA Open Data program | 17/09/2011 | Alon Peled | 79 |
| Comparison of metadata quality in open data portals using the Analytic Hierarchy Process | 01/01/2018 | Sylvain Kubler | 57 |
| A longitudinal cross-sector analysis of open data portal service capability: The case of Australian local governments | 02/04/2017 | Akemi Takeoka Chatfield | 54 |
| Using government websites to enhance democratic E-governance: A conceptual model for evaluation | 02/04/2019 | Seulki Lee-Geiller | 52 |

Fonte: Extração dos dados de Web Of Science, 2023, adaptado pelo autor.

Nos dados da plataforma Scopus, os cinco artigos com mais citados são demonstrados na Figura 5:

**Figura 5** - Artigos mais citados no Scopus

| Nome do Artigo | Data | Autor | Citações |
|---|---|---|---|
| A systematic review of open government data initiatives | 04/10/2015 | Judie Attard | 475 |
| Big Data for Development: A Review of Promises and Challenges | 13/12/2015 | Martin Hilbert | 326 |
| Civic open data at a crossroads: Dominant models and current challenges | 03/07/2015 | Renee E. Sieber | 175 |
| Accidental, open and everywhere: Emerging data sources for the understanding of cities | 01/05/2014 | Daniel Arribas-Bel | 147 |
| Barriers to open data release: A view from the top | 12/06/2014 | Emilya Barry, Frank Bannister | 139 |

Fonte: Extração dos dados de Scopus, 2023, adaptado pelo autor.

### *2.1.3 Detalhamento do modelo integrador e validação por evidências*

Com os dados coletados, filtrados e comparados entre si, agora segue a fase de análise de indicadores bibliométricos. Para este trabalho, foram selecionadas as análises de cocitação,acoplamento bibliográfico (coupling) e frequência de palavras-chave.

A análise de cocitação é utilizada para encontrar artigos comumente citados juntos em um mesmo artigo, o que pode ser considerado um indicativo de possuírem informações sobre o mesmo tema. O acoplamento bibliográfico é útil para analisar quais artigos têm citações em comum, para identificar possíveis frentes de pesquisa dentro de um mesmo assunto.

Com a análise dos dados contidos na Web Of Science, utilizando a ferramenta VoS viewer, para auxílio no cruzamento de informações, a primeira análise a ser feita, é a de cocitação, em que a unidade de análise é “referências citadas”. Levantando em consideração um limite mínimo de 14 citações por referências, somente 4 referências preencheram esses requisitos.

**Figura 6** - Análise de cocitação no Web of Science

| Nome do Artigo | Data | Autor | Co-Citações |
|---|---|---|---|
| Benefits, Adoption Barriers and Myths of Open Data and Open Government | 05/10/2012 | Marijn Janssen | 33 |
| A systematic review of open government data initiatives | 04/10/2015 | Judie Attard | 16 |
| On the barriers for local government releasing open data | 01/06/2014 | Peter Conradie | 14 |
| Open data policies, their implementation and impact: A framework for comparison | 01/01/2014 | Anneke Zuiderwijk | 14 |

Fonte: Extração dos dados de Web Of Science, 2023, adaptado pelo autor.

Percebe-se logo a importância do artigo de Marijn Janssen como um dos precursores a debater sobre o tema, sendo um autor relevante até os dias atuais.

Com a análise dos dados contidos na Scopus, também utilizando a ferramenta VoS Viewer, seguimos para entendimento das métricas de cocitação, em que a unidade de análise é “referências citadas”. O filtro utilizado foi um limite mínimo de 20 citações por referências, e aqui também,somente 4 referências preencheram esses requisitos nas mais de 50.000 referências contidas nos dados.

Ao investigar as referências, percebe-se uma duplicidade do mesmo artigo de Janssen, o tratamento para correção da computação de citações foi atualizado após a extração dos resultados.

**Figura 7** - Análise de cocitação no Scopus

| Nome do Artigo | Data | Autor | Co-Citações |
|---|---|---|---|
| Benefits, Adoption Barriers and Myths of Open Data and Open Government | 05/10/2012 | Marijn Janssen | 100 |
| Open data policies, their implementation and impact: A framework for comparison | 01/01/2014 | Anneke Zuiderwijk | 29 |
| A systematic review of open government data initiatives | 04/10/2015 | Judie Attard | 24 |

Fonte: Extração dos dados de Scopus, 2023, adaptado pelo autor.

Comparando as análises de cocitação tanto da plataforma Web Of Science, quanto da Scopus, fica evidente a relevância do artigo "*Benefits, Adoption Barriers and Myths of Open Data and Open Government*" de 2012. Sendo o artigo mais cocitado em ambas as bases, seguido logo atrás de outro artigo de aparente relevância de Anneke Zuiderwijk, "*Open data policies, their implementation and impact: A framework for comparison*" de 2014. Os textos são providenciais para o aprofundamento de pesquisa e entendimento do tema.

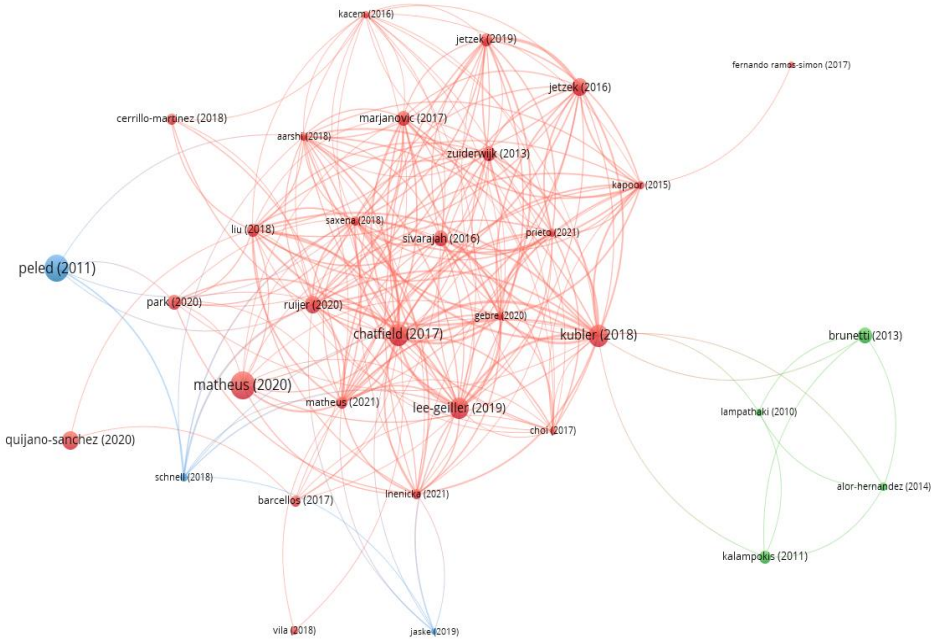
Segundo Grácio (2015), que nos ajuda a entender o acoplamento bibliográfico como, "a relação que ocorre entre dois artigos quando estes referenciam pelo menos uma publicação em comum". Complementando seu raciocínio explicando que "acoplamento bibliográfico estabelece uma conexão entre dois artigos ao utilizarem as mesmas referências."

Na análise de Coupling na plataforma WoS, utilizando como parâmetro os documentos em si, e como métrica artigos contendo pelo menos 5 citações, encontramos 33 principais documentos relacionados entre si. De acordo com a

8, fica evidente que 2 clusters menores se formam (verde e azul), e 1 cluster enorme (vermelho). Com uma análise detalhada dos trabalhos em si, percebe-se que em azul estão alguns artigos relacionados pelas palavras "democrático" e "transparência". Publicadas em algumas revistas de temas diferentes de ciência da informação e ciência da computação.

O maior cluster nesta análise é formado por publicações mescladas de ciência da informação com ciência da computação, algo completamente relacionado a linha de pesquisa deste projeto, aplicar técnicas de ciência da informação, com ferramentas da ciência da computação.

**Figura 8** - Análise de coupling no Web of Science

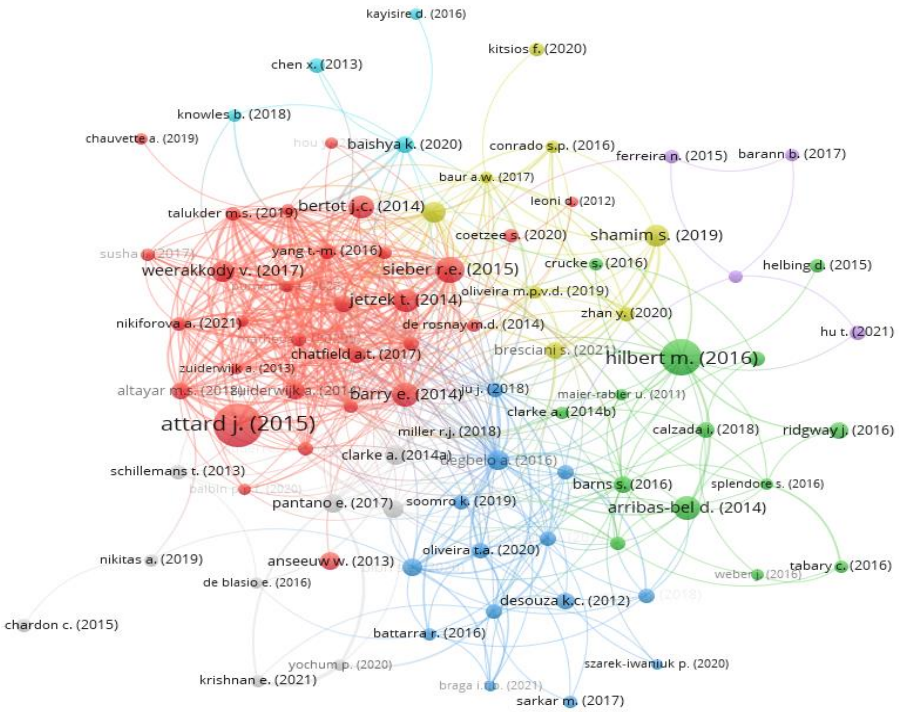



Fonte: Extração dos dados de Web Of Science utilizando VoSViewer, 2023, adaptado pelo autor.

Para a análise de acoplamento bibliográfico na plataforma Scopus, usando como unidade de análise todos os documentos extraídos, delimitando para um mínimo de 30 citações, dos 1035 documentos, 102 estão dentro do filtro mas somente 85 se relacionam entre si, foi optado por excluir os que não se relacionam.

Nesta análise, 6 clusters são formados. O cluster azul claro aborda temas relacionados à adoção de tecnologias por algumas perspectivas. O cluster em amarelo aborda a temática relacionada à tomada de decisão. O cluster verde se formou através da similaridade com o tema de Big Data. O cluster em azul é relacionado a “Cidades Inteligentes”. O cluster vermelho, o maior de todos e com relação com todos os outros clusters, foi criado a partir das palavras “Open Data”.

**Figura 9** - Análise de coupling no Scopus

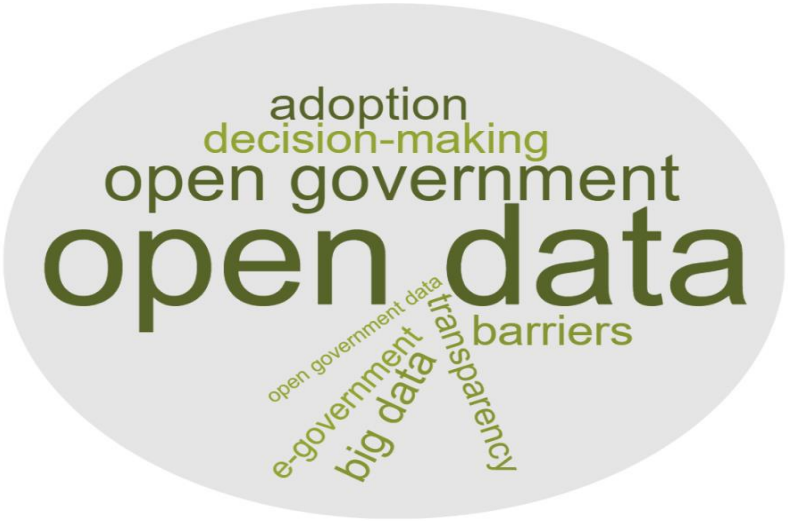



Fonte: Extração dos dados de Scopus, 2023, adaptado pelo autor.

Para a análise de frequência de palavras-chave também foi utilizada a plataforma VoSviewer, porém auxiliada pelo site Word Clouds (Wordclouds, [2023]) Segundo Mariano e Santos (2017) a “co-ocorrência de palavras-chave [...] evidencia palavras-chaves citadas juntas e a frequência de palavras-chaves que mapeiam as principais linhas de pesquisa.”

Começando pela análise das palavras da plataforma da WoS, o critério para a busca dessas palavras foi uma ocorrência de pelo menos 10 vezes entre os trabalhos. Das 657 palavras disponíveis, 9 entraram nesse filtro e podem ser visualizadas abaixo:

**Figura 10** - Análise de palavras-chave no Web of Science




Fonte: Extração dos dados de Web Of Science utilizando VoSViewer, 2023, adaptado pelo autor.

O critério para a busca dessas palavras na plataforma Scopus, foi uma ocorrência de pelo menos 50 vezes entre os trabalhos. Das 6775 palavras disponíveis, 20 entraram nesse filtro e podem ser visualizadas na Figura 11:

**Figura 11** - Análise de palavras-chave no Scopus

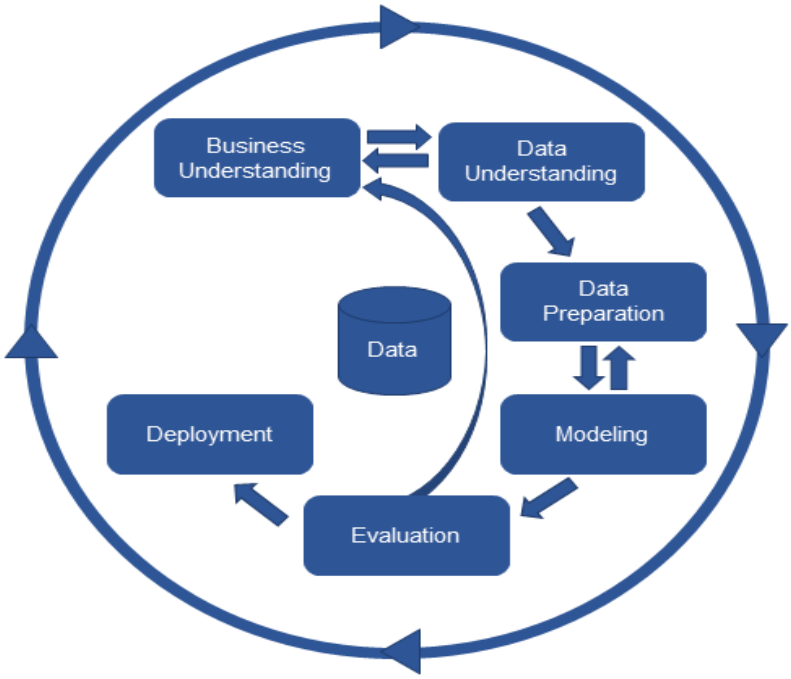

Fonte: Extração dos dados de Web Of Science utilizando VoSViewer, 2023, adaptado pelo autor.

Em ambas plataformas, percebe-se um peso similar para as palavras “Open Data” e “*decision making*”. Destaque também para “*big data*” como sendo um tema frequente nos estudos relacionados ao tema desta pesquisa

## 2.2 CRISP-DM

O Cross-Industry Standard Process for Data Mining (CRISP-DM) é um processo amplamente reconhecido para a condução de projetos de mineração de dados. Este processo é composto por seis etapas principais, que orientam o desenvolvimento de projetos desde a compreensão inicial do problema até a implementação de soluções. A seguir são detalhadas cada uma dessas etapas:

**Figura 12** - Framework CRISP-DM




Fonte: DSPA, 2013.

Compreensão do Negócio: Esta é a fase inicial, onde os objetivos e requisitos do projeto são definidos a partir de uma perspectiva de negócio. É essencial entender o problema a ser resolvido e traduzi-lo em um problema de mineração de dados. Além disso, é nesta etapa que se determinam os critérios de sucesso do projeto.

Compreensão dos Dados: Uma vez definido o problema, a próxima etapa envolve coletar, descrever e explorar os dados disponíveis. Isso inclui identificar a qualidade dos dados, suas características e entender as estruturas de dados disponíveis.

Preparação dos Dados: Esta é frequentemente a etapa mais demorada do processo. Envolve limpar, transformar e enriquecer os dados para que estejam prontos para a modelagem. As tarefas podem incluir tratamento de valores faltantes, transformação de variáveis e criação de novos atributos.

Modelagem: Com os dados preparados, a fase de modelagem envolve a seleção e aplicação de técnicas de mineração de dados para criar modelos. Isso pode envolver a escolha de algoritmos, a definição de parâmetros e a utilização de técnicas de validação, como a validação cruzada.

Avaliação: Após a modelagem, é essencial avaliar a qualidade e eficácia dos modelos gerados. Isso é feito comparando os resultados dos modelos com os critérios de sucesso definidos na primeira etapa. Além disso, é importante considerar todos os aspectos do projeto, incluindo a preparação dos dados e os objetivos de negócio, para garantir que o modelo atenda às necessidades do projeto.

Implantação: Na fase final, os modelos são colocados em prática e os resultados são aplicados ao ambiente de negócios. Isso pode envolver a integração do modelo em sistemas operacionais, a criação de relatórios e a monitorização do desempenho do modelo ao longo do tempo.

O CRISP-DM é um processo iterativo, o que significa que é comum retornar a etapas anteriores à medida que se avança no projeto. Por exemplo, durante a modelagem, pode-se descobrir que é necessário voltar à etapa de preparação dos dados para fazer ajustes adicionais.

Em resumo, o CRISP-DM oferece uma estrutura robusta e flexível para conduzir projetos de mineração de dados, garantindo que todas as etapas essenciais sejam abordadas e que os resultados sejam alinhados com os objetivos de negócio.

# 3 FUNDAMENTAÇÃO TEÓRICA

## 3.1 Dados abertos

A definição de dados, segundo o dicionário Oxford, afirma que dados são fatos e estatísticas reunidos para referência ou análise de algo. Semidão (2014) complementa que dados também são o ponto base de um processo cognitivo.

No contexto atual, com o crescente e constante avanço das tecnologias digitais, o conceito de "dados" refere-se a uma ampla variedade de elementos que se originam em ambientes digitais, incluindo textos, números, imagens, vídeos, áudios, entre outros (Koltay, 2017).

Dos dados que existem no ambiente digital, alguns são classificados como dados abertos que, levando em consideração o conceito da Open Knowledge Foundation (Tribunal de Contas da União, 2015), “dados são abertos quando qualquer pessoa pode livremente usá-los, reutilizá-los e distribuí-los, estando sujeito, no máximo, à exigência de creditar a sua autoria e compartilhar pela mesma licença”. Em sua maioria, os maiores divulgadores de dados são instituições governamentais e de pesquisa ao redor do mundo.

A definição mais aceita de dados abertos governamentais pode ser descrita por Eaves (2009) como a publicação e disseminação das informações do setor público na Web, compartilhadas em formato bruto aberto. Esse conceito foi endossado pela World Wide Web Consortium (W3C), um consórcio internacional com a missão de conduzir a web ao máximo do seu potencial através de padrões e diretrizes que garantam sua evolução e segurança.

Janssen (2012) define dados abertos como:

> [...] definimos dados abertos como dados não restritos à privacidade e não confidenciais, produzidos com dinheiro público e disponibilizados sem quaisquer restrições quanto ao seu uso ou distribuição. Dados privados, confidenciais e classificados são excluídos, pois esse tipo de dado não é apropriado para divulgação pública.

Para garantir a transparência e veracidade dos dados abertos públicos, existem diversas organizações ao redor do globo que seguem protocolos e critérios para avaliação dos dados disponibilizados. Um desses modelos é descrito por Berners-Lee, Hendler e Lassila (2006) que leva em consideração uma avaliação de

uma a cinco estrelas para avaliar a qualidade dos dados divididos nas seguintes categorias:

- Uma Estrela: disponível na internet em qualquer formato desde que com licença aberta, para que seja considerado dado aberto;
- Duas Estrelas: disponível na internet de modo estruturado (por exemplo, em uma planilha MS-Excel);
- Três Estrelas: disponível na internet de modo estruturado e em formato não proprietário (em uma planilha OpenOffice.org ou Comma Separated Values – CSV em vez de MS-Excel);
- Quatro Estrelas: seguindo todas as regras anteriores, mas dentro dos padrões estabelecidos pelo W3C (Resource Description Framework - RDF e SPARQL Protocol and RDF Query Language - SPARQL), uso de Uniform Resource Locator – URL para a identificação de coisas e propriedades, de modo que todos possam direcionar para suas publicações; e
- Cinco Estrelas: todas as regras anteriores e mais a conexão de seus dados a outros dados, fornecendo um contexto.

Dados abertos e transparência pública é um conceito e prática,

Mesmo com esse modelo descrito a tanto tempo, faltam estudos que avaliem e investiguem de maneira empírica os dados disponibilizados pelos portais, para avaliar o real valor dos arquivos ali presentes e suas reais capacidades. (Weerakkody, 2016)

Um estudo realizado classifica do ponto de vista da disponibilidade dos dados, 24 portais de dados abertos e identificou que a grande maioria desses portais estagnam na avaliação de três estrelas, somente sendo possível realizar o download dos dados, sem disponibilizar outras maneiras de se conectar e facilitar a extração (Berners-Lee; Hendler; Lassila, 2006; Kalampokis; Hausenblas; Tarabanis, 2011; Kubler, 2018).

Estudar e compreender o uso dos dados abertos públicos é um dos assuntos mais relevantes e discutidos na área de Ciência da Informação, em resposta aos problemas apontados por Wersig (1993, p. 232):

> Se mais conhecimento se torna despersonalizado, por outro lado, mais conhecimento tem de ser acreditado [...]. A situação se tornará mais complicada com as novas tecnologias [...]. Por isso, cada vez mais temos que ter cuidado com os dados de observação em dois aspectos: em primeiro lugar, temos que aceitar a tecnologia que originou os dados e, depois, levar em conta o que poderia ter acontecido com os dados brutos em processo de transformação. Para aceitar o conhecimento, temos que ser muito críticos em relação às tecnologias de coleta e manipulação.

Entendendo este problema, a área de ciência da informação procura soluções por meio de políticas de acesso a dados abertos e um enfoque na importância da descrição detalhada de todo o processo envolvido na coleta e tratamento desses dados.

Dados são a base de qualquer pesquisa e a importância do seu gerenciamento e organização através da ótica da biblioteconomia pode ser descrito por Tenopir *et al*. (2017, p. 25):

> [...] Os dados de pesquisa são uma parte essencial do registro acadêmico, e o gerenciamento de dados de pesquisa é cada vez mais visto como uma tarefa importante para as bibliotecas acadêmicas. [...] o bom gerenciamento de dados de pesquisa é essencial para garantir a transparência da pesquisa científica, preservar os dados e permitir a reutilização e reanálise dos dados e o avanço do conhecimento. [...] os dados da pesquisa são cada vez mais vistos como parte essencial do registro acadêmico. Como as bibliotecas acadêmicas tradicionalmente têm um papel no fornecimento de acesso ao registro acadêmico de várias formas, não é de surpreender que o gerenciamento de dados de pesquisa seja um problema global para as bibliotecas acadêmicas.

Podemos definir dados abertos governamentais como conjuntos de *datasets* abertos ao público em geral. Alguns exemplos são: despesas públicas, votações parlamentares, dados de transporte público, dados sobre educação e saúde, dentre outros. Esses dados são disponibilizados em portais públicos de dados, sendo uma das primeiras iniciativas na adoção de movimentos sobre dados abertos. (Attard, Orlandi; Scerri; Auer, 2015; Gebre; Morales, 2019)

## 3.2 Dados e transparência Pública

A transparência governamental, como conceito e prática, tem uma rica tapeçaria histórica que se estende por séculos. Uma análise profunda da transparência governamental ao longo dos últimos 250 anos, particularmente focada nos Países Baixos, revela padrões e tendências que têm implicações mais amplas

para outros países ocidentais (Meijer, 2015). Originalmente, a transparência foi concebida como um pilar fundamental da democracia representativa. Nesse contexto, servia como uma ferramenta que permitia ao povo monitorar e avaliar as ações de seus representantes, garantindo que os interesses públicos fossem priorizados (Meijer, 2015).

Com o passar do tempo, a noção de transparência evoluiu e adaptou-se às mudanças nas estruturas políticas e sociais. Em vez de ser apenas uma ferramenta de monitoramento, a transparência começou a ser vista como um meio de promover a participação ativa dos cidadãos no domínio público (Meijer, 2015). Isso marcou uma transição da transparência como um pilar da democracia representativa para um fundamento da democracia participativa.

No entanto, a era digital trouxe consigo novos desafios e oportunidades para a transparência pública. A ascensão dos dados abertos, em particular, tem sido vista como uma extensão lógica da transparência na era digital. No entanto, não está isenta de controvérsias. Há preocupações crescentes de que, sob o pretexto de promover a transparência, os dados abertos possam ser usados para fins políticos ocultos (Levy; Johns, 2016).

Por exemplo, a linguagem da transparência de dados, embora carregada de apelo positivo, pode ser usada estrategicamente para promover agendas políticas específicas (Levy; Johns, 2016). Iniciativas que enfatizam o acesso aberto aos dados, especialmente em contextos políticos atuais, podem ser usadas para impulsionar objetivos que vão além da simples promoção da transparência (Levy; Johns, 2016).

É essencial, portanto, abordar a questão da transparência e dos dados abertos com uma compreensão crítica e nuance. Enquanto a transparência tem o potencial de promover a responsabilidade, a participação pública e a confiança nas instituições, também é suscetível a ser cooptada para fins políticos. Como tal, é imperativo garantir que a transparência e a abertura dos dados sejam usadas de maneira responsável e ética.

Em conclusão, a transparência pública, desde suas origens históricas até sua manifestação na era digital, é um conceito dinâmico. Para garantir que continue a servir ao bem público, é crucial estar ciente de seus potenciais desafios e oportunidades na era dos dados abertos.

## 3.3 Informação e conhecimento

Uma discussão recorrente no âmbito da Ciência da Informação está no entendimento do seu principal objeto de estudo, a informação. Não existe consenso sobre o significado dessa palavra, e para sua definição, algumas premissas devem ser assumidas. Porém, uma premissa amplamente aceita por diversos cientistas é o princípio aristotélico, de que todas as ciências, para avançar no seu desenvolvimento, precisam entender "o que é" (Correia, 2017).

Existem diversos conceitos no universo da Ciência da Informação buscando sintetizar o significado da palavra informação. Segundo Fogl (1979), esse termo pode ser descrito como "a forma material da existência do conhecimento chamamos de informação". Além disso, ele complementa que a informação pode ser expressada através da linguagem natural ou de outros sistemas de signos.

Para se compreender o conceito de informação, podem-se observar diversos prismas e abordagens. No âmbito epistemológico da informação, origina-se à luz de quem a estuda, além de sua aplicação conceitual, que depende, por vezes, do contexto específico (Souza; Dias, 2011). Para Ademais, Ribeiro (2012), "a informação é um conjunto estruturado de representações mentais e emocionais codificadas - sinais e símbolos - que são desenvolvidos pela interação social". Capurro *et al*. (2007) complementa que sua definição ainda pode ser encontrada na origem da palavra *informatio* (dar forma).

No contexto atual, Capurro *et al*. (2007) afirma que a natureza digital da informação é o que a torna tão importante nos dias de hoje. Com o avanço da web, o volume de dados disponíveis vem aumentando consideravelmente a cada ano que passa, e entender a melhor forma de organizar e extrair o conhecimento necessário com informação torna-se essencial.

O objetivo da organização da informação é dar suporte ao tratamento, estudo e recuperação de objetos informacionais, sejam eles de forma estruturada, semi-estruturada ou não-estruturada. Segundo Medeiros e Café (2008), esse tema pode ser dividido em Organização da Informação (OI) e Organização do Conhecimento. O primeiro "visa a construção de modelos de mundo que se constituem em abstrações

da realidade", enquanto a organização do conhecimento "constitui-se em uma estrutura conceitual que representa modelos de mundo".

Svenonius (2000), além de ser uma grande pesquisadora da área de ciência da informação, apresenta em seus trabalhos uma análise extensa dos conceitos e teorias fundamentais da área. Além disso, ela traz um prisma sobre como essa ciência se intercala com os problemas contemporâneos gerados pela internet e como as técnicas e práticas tradicionais precisam se adaptar a uma nova realidade.

Um outro termo que se destaca como forma de síntese de informação, segundo a ciência da computação, são os Indicadores-chave de Desempenho, ou do inglês Key Performance Indicators (KPIs). Um dos principais autores que aborda a temática de KPIs é David Parmenter. Em sua obra *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs* (Parmenter, 2010), ele apresenta um guia completo para o desenvolvimento, implementação e utilização de KPIs efetivos. Ele destaca a importância da definição de objetivos claros e mensuráveis, bem como da escolha adequada dos indicadores a serem utilizados.

Outra referência importante é a obra *The Balanced Scorecard: Translating Strategy into Action* (Kaplan; Norton, 1996), de Robert Kaplan e David Norton. Os autores introduzem o conceito de *Balanced Scorecard*, que utiliza KPIs para medir o desempenho de uma organização em quatro perspectivas diferentes: financeira, cliente, processos internos e aprendizado e crescimento. Essa abordagem holística permite que os gestores avaliem o desempenho de uma empresa de forma integrada e equilibrada.

De acordo com Guimarães (2007), a organização da informação tem como objetivo criar repositórios organizados de informação e desenvolver técnicas que auxiliam na prevenção da criação de redes confusas de conceitos, que podem levar os usuários a desperdiçar muito tempo navegando sem encontrar o que procuram.

A falta de conhecimento acerca das técnicas e fundamentos relacionados a dados abertos, tanto da perspectiva dos publicadores, quanto da população em geral ainda é uma grande barreira para esses portais atingirem seu real potencial. Sua exploração e entendimento ainda é baixo entre a população em geral e se reduz a profissionais especialistas em ciências de dados do setor privado e pesquisadores (Gebre; Morales, 2019).

## 3.4 Técnicas de análises de dados para gerar informação e conhecimento

### *3.4.1 Análise exploratória de dados*

A análise exploratória de dados (AED) é uma abordagem de análise de dados, cujo objetivo principal é compreender características, padrões e tendências dos dados coletados. Segundo Tukey (1977), a AED é um processo de exame visual e quantitativo de dados, visando descobrir padrões, identificar possíveis problemas de qualidade dos dados e gerar hipóteses para análises mais avançadas. Essa é uma etapa crucial para o analista compreender os tipos de dados presentes e como prosseguir suas análises.

Segundo Michael (2020), uma boa prática comum relacionada à análise exploratória de dados é gastar a maior parte do tempo de um trabalho de um analista ou cientista de dados na análise exploratória, buscando direcionar e entender os principais conceitos que existem nos dados dispostos.

Cleveland (1993) propõe que a criação de gráficos não deve ser apenas uma ferramenta de comunicação, mas também um meio para explorar e entender os dados. O autor defende que a escolha do tipo de gráfico a ser utilizado deve estar diretamente ligada às características dos dados e às perguntas que se quer responder.

O autor também propõe diversas técnicas para a criação de gráficos desde os mais simples, como o gráfico de barras e o histograma, até os mais complexos, como o gráfico de múltiplas variáveis e o gráfico de séries temporais. O autor sempre enfatiza que os gráficos têm que ser claros e objetivos, transmitindo a informação de maneira eficiente e sem distorções (Cleveland, 1993).

Outro autor que se destaca no quesito exploratório dos dados é Chambers (1983), que reforça a importância dos gráficos para a compreensão dos dados, defendendo que eles devem ser utilizados tanto para explorar as informações quanto para comunicá-las de maneira eficiente.

Chambers (1983) também destaca a importância da organização e da gestão dos dados e a necessidade de se utilizar técnicas e ferramentas específicas para sua melhor organização.

Sivarajah *et al.* (2016) diz que dados abertos públicos brutos, enquanto podem ser um facilitador e um potencializador para os cidadãos monitorarem o governo e as contas públicas, estes mesmos dados precisam de intermediação e interpretação. Técnicas de análises exploratórias são necessárias para entender o potencial de cada dado aberto público.

### *3.4.2 Extração transformação e carga*

O processo de ETL é uma técnica de integração de dados em três etapas (extração, transformação e carga), utilizada para combinar informações provenientes de diversas fontes, para criar um repositório de dados (SAS, 2021). Segundo o Gartner (2023), o ETL é fundamental para a implementação de soluções de Business Intelligence (BI) e análise de dados em empresas, pois a integração de informações de diversas fontes é fundamental para a tomada de decisões estratégicas.

Rainer, Prince e Watson (2018) destacam que o processo de ETL pode ser complexo e requer a utilização de ferramentas e técnicas específicas para garantir a integridade e a qualidade dos dados. Além disso, eles ressaltam a importância da fase de transformação do ETL, que envolve a limpeza, a validação e a padronização dos dados a fim de possibilitar sua utilização de maneira efetiva.

Diversos trabalhos destacam que a qualidade dos dados é fundamental para a obtenção de insights precisos e que o processo de ETL é uma etapa crítica para garantir a confiabilidade das informações. Nessa fase, as principais regras de limpeza, padronização e tratamento são estabelecidas. (Chamber, 1983; Heise, 2012)

Janssen (2020) explica que “os dados podem ter qualidades variadas e serem coletados de maneiras diferentes. [...] eles precisam ser processados para sua visualização se tornar possível através de *dashboards* para facilitar o entendimento”.

### *3.4.3 Business intelligence*

O conceito de *Business Intelligence*, também conhecido como Inteligência de Negócios ou Inteligência Empresarial, é composto por um conjunto de metodologias de gestão que são aplicadas por meio de ferramentas de *software*. O objetivo principal dessa abordagem é otimizar os processos decisórios gerenciais e de alta

administração nas organizações, utilizando a capacidade analítica das ferramentas de software para integrar todas as informações necessárias em um único lugar (Angeloni; Reis, 2006, p. 3).

Ao transformar dados em conhecimento, o *Business Intelligence* tem como objetivo gerar vantagens competitivas para a empresa. Isso é possível por meio da análise das informações coletadas, que fornecem insights valiosos para a tomada de decisões estratégicas. O *Business Intelligence* é uma abordagem essencial para o sucesso empresarial, permitindo que as empresas tomem decisões informadas e efetivas com base em dados precisos e confiáveis (Angeloni; Reis, 2006, p. 3).

Outro conceito muito utilizado para descrever BI é o de Barbieri (2001 *apud* Angeloni; Reis, 2006, p. 5):

> [...] um guarda-chuva conceitual, visto que se dedica à captura de dados, informações e conhecimentos que permitam às empresas competirem com maior eficiência em uma abordagem evolutiva de modelagem de dados, capazes de promover a estruturação de informações em depósitos retrospectivos e históricos, permitindo sua modelagem por ferramentas analíticas. Seu conceito é abrangente e envolve todos os recursos necessários para o processamento e disponibilização da informação ao usuário.

Uma das formas de disponibilização de informações utilizando técnicas de BI é por meio de *dashboards*, que são painéis contendo gráficos e estatísticas organizadas (Vila; Estevez; Fillottrani, 2018).

Também é importante destacar que, para obter painéis de controle significativos e representativos, os indicadores-chave de desempenho (KPI) devem ser cuidadosamente projetados (Vila; Estevez; Fillottrani, 2018).

O uso dos *dashboards* devem resultar em transparência e responsabilidade e, finalmente, em mais confiança no governo (Harrison; Sayogo, 2014; Villeneuve, 2014). Além disso, os *dashboards* podem estimular o engajamento do cidadão. Os governos também podem desenvolver painéis de controle para sua própria tomada de decisão com base na contribuição do engajamento cidadão (Janssem, 2020).

### *3.4.4 Arquitetura da informação*

Não existe uma definição exata do que é ou quais etapas constituem uma ‘Arquitetura da Informação’ (AI), porém, este termo foi utilizado pela primeira vez por um arquiteto, Richard Saul Wurman, que em 1976, definiu o temo como “a ciência e a arte de criar instruções para espaços organizados” (Macedo, 2005; Victorino, 2011).

De acordo com Brancheau e Wetherbe (1986), a ‘Arquitetura da Informação’ envolve a elaboração de um plano para modelar os requisitos de informações de uma organização. Essa abordagem tem como propósito mapear de forma eficiente as informações necessárias para a organização, levando em consideração os processos de negócio e a documentação dos inter-relacionamentos.

Rosenfeld e Morville (2002) defendem que a AI como uma disciplina, que está na interseção entre o ‘contexto’, que pode ser descrito como os objetivos de negócio, o ‘conteúdo’, que são as informações e dados ali presentes e por fim os ‘usuários’, que são parte fundamental de qualquer solução, entendendo as necessidades e vontades do público-alvo.

Mais uma vez trazendo à tona as necessidades dos usuários, também podemos definir o termo como o processo de planejamento, organização e estruturação de conteúdos e funcionalidades de sistemas digitais, que leva em consideração as necessidades dos usuários e os objetivos do negócio (Kalbach, 2016).

Em alguns autores podemos perceber o conceito de AI, sendo confundidos com conceitos relacionados a Usabilidade. Mesmo que a Usabilidade, é apenas uma das várias disciplinas que podemos abordar, através da ótica de construção de soluções de arquiteturas da informação. O conceito é mais abrangente do que somente apresentações de informações em sítios da *web*. (Rosenfeld; Morville, 2002; Garrett, 2003; Victorino, 2011; Kalbach, 2016)

Sua relação com a tecnologia pode ser apontada por Victorino (2011):

> A tecnologia desempenha um papel importante em uma Arquitetura da Informação, mas o objetivo da AI é a organização e armazenagem dos objetos informacionais estruturados, semi-estruturados e não-estruturados em repositórios informacionais (bancos de dados, sistemas de arquivos, etc) providos de consistência, compartilhamento, documentação, privacidade e

> recuperação eficaz de seus conteúdos, sem se prender a técnicas específicas de modelagem de dados ou arquitetura de sistemas de informação.

A carreira de Arquiteto da Informação foi descrita como um profissional com duas funções claras. Primeiro, alguém que torna o complexo claro, através da organização dos dados por padrões inerentes. E também, aquele que cria a estrutura ou o mapa da informação, ajudando outros a encontrarem seu próprio caminho para o conhecimento (Wurman, 1996).

Então, ‘Arquitetura da Informação’ é um conjunto de processos e técnicas, que visam organizar e disponibilizar informações relevantes para o usuário final. Atualmente, as soluções estão em sua grande maioria, em ambientes digitais. Todos os processos, etapas e instruções precisam ser documentados e organizados, possibilitando a reutilização e análise das informações divulgadas, por qualquer outro usuário. (Brancheau; Wetherbe, 1986; Wurman, 1996; Rosenfeld; Morville, 2002; Macedo, 2005; Victorino, 2011; Kalbach, 2016).

Para fins deste projeto, o entendimento final sobre arquitetura da informação se assemelha mais com o de Macedo (2005), que afirma que Arquitetura da Informação é uma metodologia de ‘desenho’ que se aplica a qualquer ‘ambiente informacional’, sendo este compreendido como um espaço localizado em um ‘contexto’, constituído por conteúdos em fluxo, que serve a uma comunidade de ‘usuários’. A finalidade da Arquitetura da Informação é, portanto, viabilizar o fluxo efetivo de informações por meio do desenho de ‘ambientes informacionais’. (Victorino, 2011)

# 4 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados alguns trabalhos relacionados à tomada de decisão por meio da análise de dados públicos, de diferentes perspectivas. Durante a pesquisa dos trabalhos relacionados, percebe-se o cruzamento das disciplinas de ciência da computação e ciência da informação.

Os trabalhos selecionados neste capítulo passaram por alguns filtros estipulados em análises descritivas nas bases de dados extraídas e explicadas no capítulo da metodologia. Além de levar em consideração os indicadores bibliométricos selecionados (cocitação e coupling). Essa seleção e revisão de literatura apontam as semelhanças dos desafios que todos os países do mundo enfrentam atualmente do ponto de vista de dados abertos.

Mesmo com todos os filtros anteriormente expostos, não seria possível navegar e explorar todos os trabalhos extraídos das plataformas. Os critérios de seleção utilizados foram trabalhos mais citados, trabalhos mais relevantes, relevância de autores na área, e somente foram selecionados artigos que tratavam de dados abertos do setor público. Além disso, foram levados em consideração os indicadores bibliométricos anteriormente citados. Assuntos relacionados à qualidade, dificuldade e desafios de dados abertos também foram incorporados.

A Figura 13 representa a tabela com o top 10, de todos os trabalhos filtrados em ambas as plataformas e os motivos que levaram à sua escolha e curadoria.

**Figura 13** - Curadoria de trabalhos

| Nome do Artigo | Data | Autor | Critério | tema |
|---|---|---|---|---|
| Benefits, Adoption Barriers and Myths of Open Data and Open Government | 05/10/2012 | Marijn Janssen | - artigo relevante na análise de co-citação tanto do WoS, quanto Scopus<br>- autor que mais publica sobre o assunto tanto do WoS, quanto Scopus | 1.Geral - benefícios e problemas |
| A systematic review of open government data initiatives | 04/10/2015 | Judie Attard | - artigo mais citado na plataforma Scopus<br>- artigo relevante na análise de co-citação tanto do WoS, quanto Scopus | 1.Geral - revisão sistemática |
| Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities | 03/06/2020 | Ricardo Matheus, Marijn Janssen | - artigo mais citado na plataforma WoS<br>- similaridade com o tema de dashboards e tomada de decisão<br>- top 10 artigo mais citado na Scopus | 2.Tomada de decisão |
| The design and use of dashboards for driving decision-making | 04/04/2018 | Roman A. Vila, Elsa Estevez, Pablo R. Fillottrani | - similaridade com o tema de dashboards e tomada de decisão<br>- top 6 relevancia com as palavras-chave na plataforma WoS | 2.Tomada de decisão |
| How accessible is open data | 12/12/2019 | Engida H. Gebre | - artigo relevante do ponto de vista do usuário de dados abertos em si<br>- curadoria própria | Qualidade dos dados - Acessibilidade de dados |
| The role of e-participation and open data in evidence-based policydecision making in local government | 07/03/2016 | U. Sivarajah, V. Weerakkody, P. Waller, H. Lee, Z. Irani, Y. Choi, R. Morgan & Y. | - artigo relevante do ponto de vista da discussão da participação do cidadão na tomada de decisão política<br>- curadoria própria na plataforma Scopus | 2.Tomada de decisão |
| Open data and its usability: an empirical view from the Citizen's perspective | 23/07/2016 | Vishanth Weerakkody1 & Zahir Irani 1 & Kawal Kapoor1 & Uthayasankar Sivarajah1 & Yogesh K. Dwivedi | - artigo relevante do ponto de vista da discussão da participação do cidadão na tomada de decisão política<br>- curadoria própria na plataforma Scopus | 1.Geral - perspectiva do cidadão |
| Comparison of metadata quality in open data portals using the Analytic Hierarchy Process | 01/01/2018 | Sylvain Kubler | - artigo relevante de citação na plataforma WoS<br>- artigo relevante do ponto de vista de qualidade dos dados públicos | 3.Qualidade dos dados |
| Combining Social and Government Open Data for Participatory Decision-Making | | Evangelos Kalampokis1,2, Michael Hausenblas1, and Konstantinos Tarabanis2 | - um dos autores que mais publica sobre o assunto pela scopus (tarabanis)<br>- artigo relevante mescalando tomanda de decisão e a perspectiva do cidadão | 1.Geral - perspectiva do cidadão<br>Tomada de decisão |
| Integrating open government data with stratosphere for more transparency | 02/04/2012 | Arvid Heise, Felix Naumann | - artigo relevante sobre o tema de cruzamento de dados abertos públicos<br>- curadoria pessoal | 3.Qualidade dos dados DataLinking |

Fonte: Elaborado pelo autor, 2023.

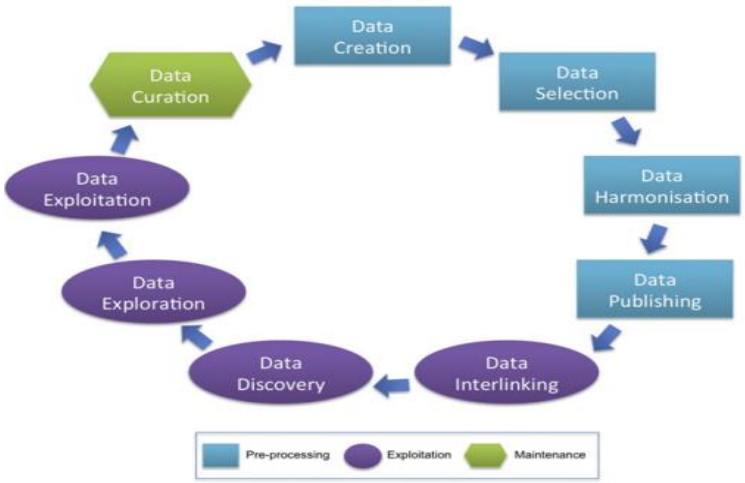
Organizando e revisando o material, os artigos foram classificados em 3 principais temas. Uma revisão geral sobre tópicos relacionados a dados abertos, envolvendo benefícios, problemáticas, revisões e perspectivas do cidadão.

O segundo tema trata dos dados abertos públicos com um viés direcionado a tomada de decisão, envolvendo entregas reais de valor ao cidadão. Por fim, trabalhos relacionados à qualidade dos dados e em como juntar diferentes fontes de dados públicos.

### 4.1 Trabalhos relacionados: perspectivas gerais sobre o tema

O artigo mais citado na plataforma Scopus é uma revisão sistemática, assim como parte da metodologia bibliográfica base deste trabalho. Diferente do propósito deste estudo de caso, que se concentra em uma análise específica através de um estudo de caso, o estudo em questão adota uma abordagem mais ampla para mapear as iniciativas e revisar os principais conceitos relacionados aos dados governamentais. A discussão inclui uma análise da terminologia e do ciclo de vida dos dados públicos, dividido em três etapas principais: pré-processamento, exploração e manutenção (Figura 14) (Attard; Orlandi; Scerri; Auer, 2015).

**Figura 14** - Ciclo de vida dos dados públicos proposto




Fonte: Attard; Orlandi; Scerri; Auer, 2015.

O artigo continua a discussão sobre as iniciativas de dados abertos governamentais em todo o mundo e seus principais desafios, seguido por uma análise da publicação e análise desses dados e conclui que isso gera impacto nas pessoas interessadas, afirmando que "as iniciativas de dados abertos do governo são

baseadas em transparência, participação cidadã e colaboração para fortalecer a democracia" (Attard; Orlandi; Scerri; Auer, 2015).

Complementando essa discussão, outro artigo cita que após diversas pesquisas, pode-se perceber três principais grupos de vantagens relacionadas a dados abertos, “benefícios políticos e sociais”, “benefícios econômicos” e "benefícios operacionais e técnicos". Sendo o primeiro visto como a categoria mais importante. (Janssem; Charalabidis; Zuiderwijk, 2012)

Após leitura do artigo, foi realizada uma curadoria das principais vantagens de cada um dos grupos, conforme Figura 15.

**Figura 15** - Benefícios de dados abertos

| Benefícios Políticos e Sociais | Benefícios Econômicos | Benefícios Operacionais e Técnicos |
|---|---|---|
| Maior transparência | Crescimento econômico e estímulo à competitividade | A capacidade de reutilizar dados / não ter que coletar os mesmos dados novamente e evitar a duplicação desnecessária e os custos associados (também por outras instituições públicas) |
| Maior participação e autonomia dos cidadãos (usuários) | Estímulo à inovação | Otimização de processos administrativos |
| Criação de confiança no governo | Desenvolvimento de novos produtos e serviços | Melhoria das políticas públicas |
| Engajamento público | Criação de um novo setor agregando valor à economia | Tomada de decisão justa possibilitada pela comparação |
| Novos serviços governamentais para os cidadãos | Disponibilidade de informações para investidores e empresas. | Capacidade de mesclar, integrar e combinar dados públicos e privados. |

Fonte: Janssem; Charalabidis; Zuiderwijk, 2012.

E as barreiras mais comuns, podem ser divididas entre os grupos de: “barreiras institucionais”, “barreiras das complexidades conceituais das tarefas em si”, “barreiras de uso e participação”, “barreiras de legislação”, “barreiras de qualidade da informação” e “barreiras técnicas”. (Janssem; Charalabidis; Zuiderwijk, 2012)

Do mesmo modo que com os benefícios, foi realizado um filtro com as principais barreiras de cada categoria, relacionados com o presente trabalho.

**Figura 16** - Barreiras de dados abertos

| Barreiras Institucionais | Barreiras das Complexidades Conceituais | Barreiras de Uso e Participação |
|---|---|---|
| Ênfase em barreiras e negligência de oportunidades | Falta de habilidade para descobrir os dados apropriados | Nenhum incentivo para os usuários |
| Compensação incerta entre valores públicos (transparência vs. valores de privacidade) | Sem acesso aos dados originais (apenas dados processados) | Organizações públicas não reagem ao feedback dos usuários |
| Ausência de política uniforme para divulgação de dados | Sem explicação do significado dos dados | Falta de conhecimento para utilizar ou compreender os dados |
| Falta de recursos para divulgação de dados (especialmente em pequenas agências) | Nenhuma informação sobre a qualidade dos dados abertos (ver categoria "Qualidade da Informação") | Falta da capacidade necessária para utilizar as informações |
| Promoção dos interesses de organizações locais em detrimento dos interesses dos cidadãos | O foco está em fazer uso de conjuntos de dados individuais, enquanto o valor real pode vir da combinação de vários conjuntos de dados | Nenhum conhecimento estatístico ou entendimento do potencial e limitações das estatísticas |
| **Barreiras de Legislação** | **Barreiras de Qualidade da Informação** | **Barreiras Técnicas** |
| Violação de privacidade | Falta de informação | Ausência de padrões |
| Segurança | Falta de precisão da informação | Falta de metapadrões |
| Nenhuma licença para usar dados | Informação incompleta, apenas parte da imagem total mostrada ou apenas uma certa faixa | Sem software padrão para processamento de dados abertos |
| Condições limitadas para o uso de dados | Dados obsoletos e não válidos | Fragmentação de software e aplicativos |
| É necessário obter permissão prévia por escrito para acessar e reproduzir dados | Dados similares armazenados em diferentes sistemas produzem resultados diferentes. | Sistemas legados que complicam a publicação de dados. |

Fonte: Janssem; Charalabidis; Zuiderwijk, 2012.

Através da óptica do cidadão, em um questionário realizado no Reino Unido a respeito de como pessoas viam os dados abertos e suas políticas. Foi constatado três principais perspectivas. Uma dessa é que falta clareza relacionado a disponibilidade dos dados públicos (36,6% das respostas). Outra perspectiva diz que o governo está abrindo dados muito significativos e isso pode ser perigoso (20,7% das respostas). Por fim, a perspectiva predominante no questionário da pesquisa, é de que os dados abertos devem oferecer percepções inovadoras para mudanças políticas e sociais (42,6% das respostas). (Weerakkody, 2016)

Existe também, uma visão relacionada a tomada de decisão política, através de ferramentas de votos online, disponíveis a toda população. Com isso, seria possível empoderar ainda mais o cidadão através das votações públicas. Os defensores dessa tese, acreditam, que isso poderia redistribuir a relação de poder entre o estado e a população, com isso, teríamos uma sociedade mais equilibrada, que está mais próxima de conseguir mudanças significativas no país e em suas comunidades (Kalampokis; Hausenblas; Tarabanis, 2011; Gebre; Morales, 2019).

## 4.2 Trabalhos relacionados: tomada de decisão

Nesta revisão, pode-se observar que na Web Of Science o artigo com o maior número de citações aborda o tema através de uma perspectiva de transparência e uso prático de ciência de dados como forma de empoderamento do cidadão. Janssen (2020) explana que:

> Na ciência de dados, o compartilhamento, uso e interpretação de dados são aspectos-chave na redução da lacuna entre o governo e o público. [...]. *Dashboards* podem ser usados para fornecer informações aos tomadores de decisão governamentais, mas também para que o público possa examinar as ações do governo, participar dos processos de tomada de decisão e melhorar a tomada de decisão. Os *Dashboards* devem ajudar a facilitar a transparência, a governança, a confiabilidade e permitir que os cidadãos participem da tomada de decisão em cidades inteligentes.

Além disso, Janssen expande em seu trabalho a análise de dois estudos de caso relacionados ao tráfego urbano. Um ponto importante em seu trabalho é que, em ambos os estudos, ocorre uma parceria entre uma empresa pública (SmartCity - Rio) e uma empresa privada (Waze e Moovit) a fim de cruzar informações e enriquecer os dados em conjunto.

O auxílio à tomada de decisão por meio de *dashboards* com dados públicos, tanto do ponto de vista dos gestores públicos quanto da população, também é explorado no texto de Vila, Estevez e Fillottrani (2018). Um ponto interessante abordado no artigo é que, na geração dos dados públicos, os mesmos sempre são gerados com um viés e não com uma análise técnica neutra. Diversas vezes, os dados são disponibilizados para atender critérios de transparência, mas seu cruzamento e análise é dificultado por meio de publicações separadas sem colunas identificadoras ou múltiplas camadas de colunas identificadoras.

Vila, Estevez e Fillottrani (2018) conclui seu texto com a proposta de um dashboard público geral para cada cidade, que contenha as principais métricas e acompanhamento de resultados do setor público. Tanto Vila, Estevez e Fillottrani (2018) quanto Jansen (2020) abordam a ideia de *dashboards* públicos como forma de democratizar e aproximar o público da tomada de decisão, munindo-os de mais informações. Ambos os trabalhos analisam essa ótica por meio de um estudo de caso sobre como um dashboard impactou ou pode impactar a sociedade como um todo.

O uso de técnicas de visualização para facilitar o entendimento humano de dados é amplamente estudado e validado. Assim que dados são compartilhados de maneira inteligente, eles abrem um leque de possibilidades e novas análises a serem realizadas por outros usuários (Sivarajah *et al*., 2016).

### 4.3 Trabalhos relacionados: qualidade dos dados

Gebre e Morales (2020) escreve sobre o quão "realmente acessível" são as ferramentas de Open Data e faz uma análise sobre as informações do ponto de vista dos comentários dos usuários. Suas conclusões apontam que as descrições dos bancos de dados são limitadas e os usuários têm que encontrar os próprios meios para entender e descobrir o que há nos *datasets*. Além disso, aponta que atualmente os quatro principais desafios dos portais públicos de dados abertos giram em torno de: organização e acessibilidade dos dados, clareza e completude, utilidade e precisão, e linguagem (ortografia e gramática).

Uma ótica sobre qualidade dos dados classifica e compara o real valor dos dados abertos com base em quatro principais indicadores. Primeiro, um indicador básico do conjunto de dados: determina a presença de um conjunto pré-definido de dados abertos de alto valor com base em nove categorias: Finanças e Economia, Meio Ambiente, Saúde, Energia, Educação, Transporte, Emprego, Infraestrutura, População (Kubler, 2018).

Um indicador de transparência, que consiste em dois indicadores (i) Transparência Governamental, que é observada como medida de conhecimento sobre as tarefas, processos e operações governamentais; e (ii) Transparência de Dados, que é calculada como uma média dos valores de Autenticidade, Compreensibilidade e Reutilização de Dados (Kubler, 2018).

E por fim, um ponto de vista prático de Indicadores de Participação e Colaboração da população, que o envolvimento do usuário é usado como fonte para os indicadores de participação e colaboração (Kubler, 2018).

Um outro estudo relacionado a ótica de despesas públicas e repositórios de dados abertos analisou como essas fontes poderiam ser relacionadas com outros repositórios, cruzando as informações e gerando mais valor à população em geral (Heise, 2012).

# 5 ARQUITETURA DA INFORMAÇÃO PROPOSTA

A arquitetura da informação proposta nesse estudo se dá no ambiente informacional de dados abertos, para a comunidade acadêmica, jornalistas especializados em educação ou outros profissionais das áreas da ciência e tecnologia. A viabilização desta arquitetura é composta por uma versão adaptada do *framework* CRISP-DM tendo como guia alguns alicerces básicos de segurança da informação (integridade, disponibilidade e autenticidade).

A integridade, refere-se à garantia de que a informação não foi alterada de forma não autorizada, seja intencionalmente ou acidentalmente. A integridade assegura que a informação mantenha sua forma original durante seu armazenamento, transmissão ou processamento. (Zanon, 2015)

A disponibilidade é um conceito que está relacionado à garantia de que a informação ou recurso estará acessível e utilizável quando necessário. Em outras palavras, a informação deve estar disponível para os usuários autorizados sempre que eles precisarem dela (Silva, 2022). Já a autenticidade, refere-se à garantia de que a informação é genuína e pode ser confiável, garantindo que a informação provém de uma fonte confiável e que não foi falsificada ou alterada de forma maliciosa. (Silva, 2022).

Todos os alicerces básicos escolhidos de segurança da informação permeiam todas as etapas adaptadas do *framework* CRISP-DM. O objetivo da arquitetura da informação proposta é servir como uma base processual e ferramental para a análise de dados oriundos de portais públicos. Todos os arquivos e código gerados, são disponibilizados em repositório abertos, garantindo a integridade e autenticidade da informação apresentada.

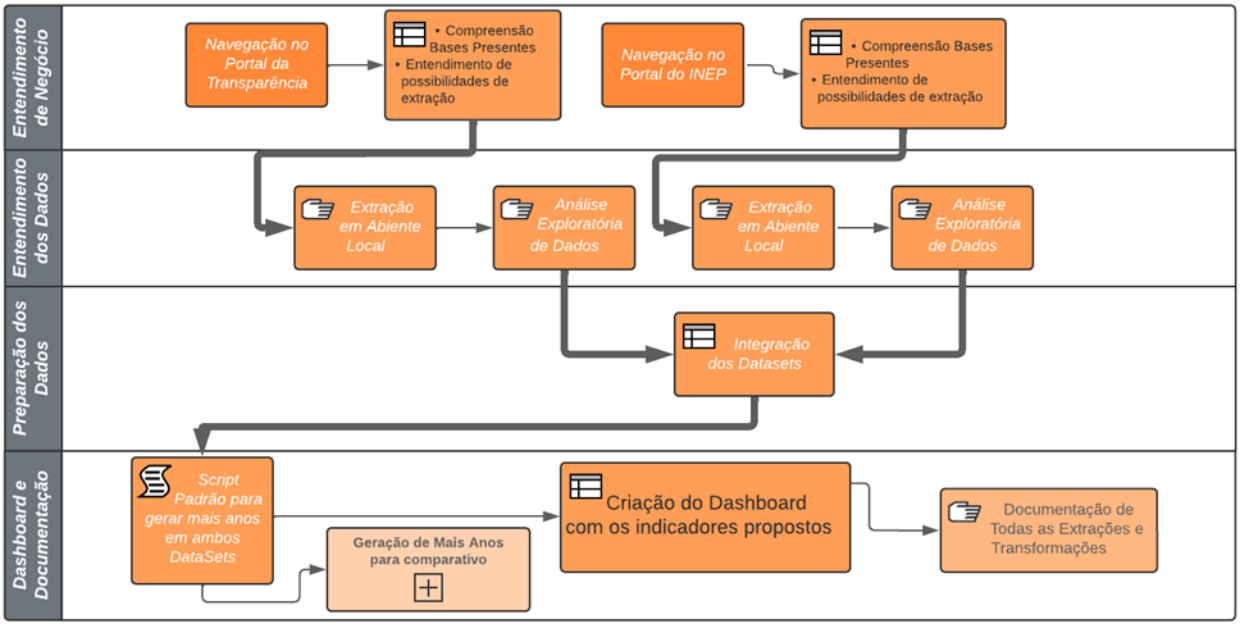
A arquitetura da informação proposta pode ser vista através de dois prismas principais que se relacionam entrei si, a AI como um processo e a AI como uma estrutura, que no caso deste estudo é uma arquitetura de software.

## 5.1 Arquitetura da Informação como processo

O conjunto de processos que compõem a arquitetura da informação proposta pode ser organizado em 4 principais etapas como pode ser visualizado na Figura 17,

sendo as 3 primeiras praticamente idênticas ao proposto pelo CRISP-DM, a adaptação em relação ao *framework* original encontra-se em sua parte final, compreendendo as etapas de Modelagem, Avaliação e Publicação. Essas três partes são agrupadas em um subconjunto denominado "*Dashboard*", uma maneira de disponibilização das informações geradas, através de um *dashboard* público.

**Figura 17 –** Arquitetura como Processo




Fonte: elaborado pelo autor.

A primeira etapa é uma investigação manual nos portais de dados buscados para o propósito deste estudo, nesta navegação busca-se responder algumas principais questões como por exemplo, quais dados estão disponíveis, quais são os formatos de extração e quais são os limites de extrações.

A partir disso, o próximo passo é realizar o *download* de alguns arquivos de amostra para ambos *datasets*, na etapa denominada "Entendimento dos Dados", o objetivo dessa etapa é explorar as amostras o máximo possível com o objetivo de a partir disso encontrar maneiras de integra-los para ser possível cruzar informações entre os arquivos. Através dessa integração, encontra-se também o formato final desejado para os arquivos e assim termina a fase de "Preparação dos Dados"

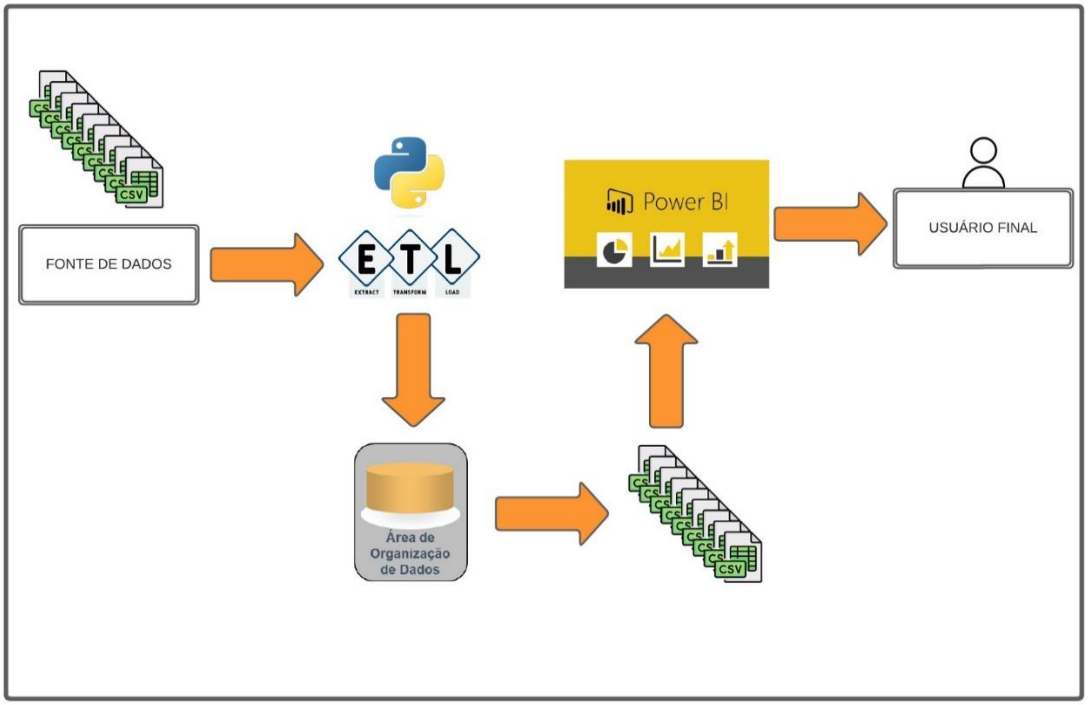
A última etapa desta arquitetura é construir uma forma de extração de dados para gerar mais arquivos e aplicar o mesmo tipo de tratamento dos arquivos de

amostra que foram manipulados localmente e transformar esses arquivos em um painel de visualização com gráficos para responder algumas perguntas principais. O projeto se encerra, após toda a documentação ser divulgada e catalogada.

## 5.2 Arquitetura da Informação como estrutura

A partir do entendimento dos dados ocorrido na fase da AI como um processo, passa-se para a fase seguinte da formatação dessa AI, que é a fase da arquitetura como uma estrutura. No caso desse estudo, uma arquitetura de software que pode ser visualizada de acordo com a Figura 18:

**Figura 18** – Arquitetura da Informação como estrutura




Fonte: Elaborado pelo autor.

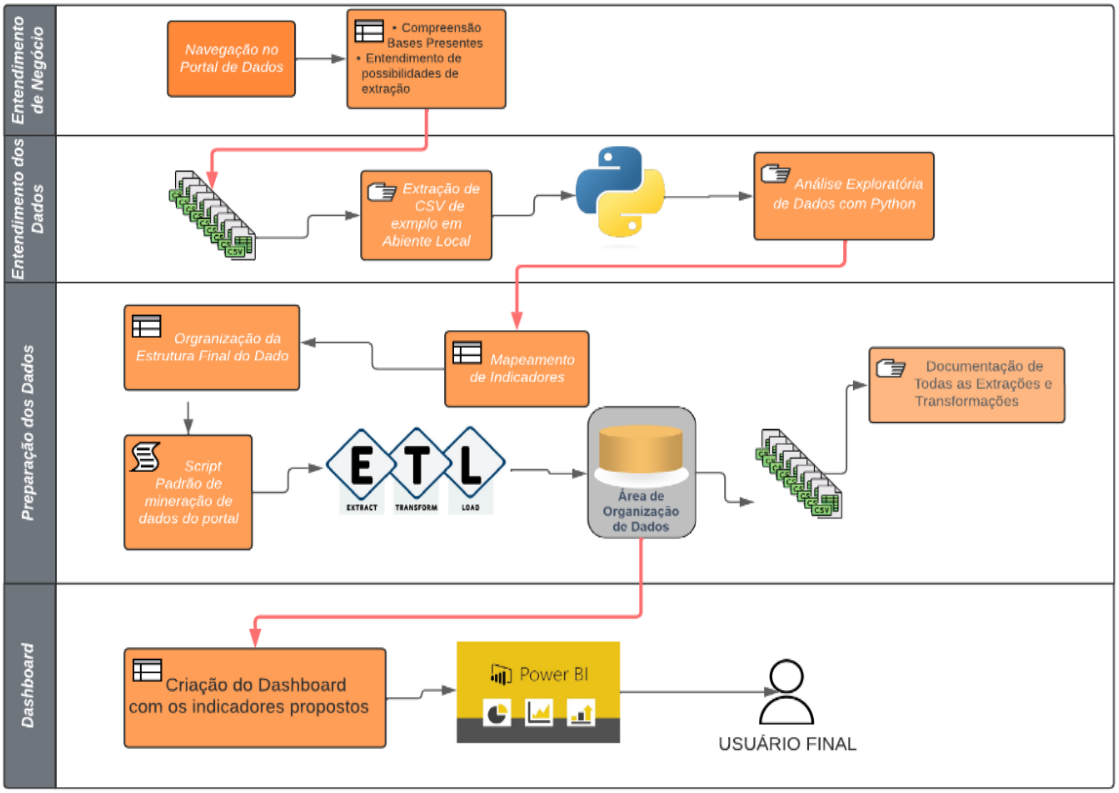
A estrutura de software é composta a partir dos arquivos de bases identificadas nas etapas da arquitetura como processo. O objetivo após determinar quais são os formatos das fontes de dados, utilizando a linguagem de programação python aplicamos a etapa de extração, transformação e carga. A carga desses dados é realizada em um repositório local com os dados organizados.

A partir dos dados organizados em uma área específica, carrega-se esses dados no Power BI, *software* de desenvolvimento de *dashboards* e ali é desenvolvida de fato a estrutura final de publicação e compartilhamento dos resultados, o painel de livre acesso aos usuários finais.

### 5.3 Arquitetura da Informação Processual e Estrutural

Unindo ambas as perspectivas da AI foi formatado o desenho final desta arquitetura, organizando tanto os processos como a estrutura dentro de uma mesma visualização. A Figura 19 é a representação dessa união.

**Figura 19** - Arquitetura da informação proposta




Fonte: elaborado pelo autor.

Esta arquitetura é também subdivida em 4 principais etapas como pode ser visualizado na Figura 19, sendo as 3 primeiras praticamente idênticas ao proposto pelo CRISP-DM, nesta adaptação somente foi adicionado o ferramental necessário e utilizado em cada uma dessas etapas. A adaptação em relação ao *framework* original

encontra-se em sua parte final, que no que seriam as etapas de Modelagem, Avaliação e Publicação. Essas três partes são agrupadas em um subconjunto denominado “*Dashboard*”, uma maneira de disponibilização das informações geradas, através de um *dashboard* público.

Na etapa inicial de “Entendimento de Negócio” o objetivo é navegar pelo portal de dados abertos da sua escolha e buscar compreender quais são as informações apresentadas, as principais respostas que devem ser buscadas são: qual a disponibilidade desses dados? Qual a forma de extração desses dados? Qual é a granularidade do dado? Como esses dados atualizados? Quais são as tabelas necessárias?

Com o entendimento de negócio aplicado e o entendimento exato de quais tabelas serão utilizadas e quais são suas formas de extração, segue-se para a etapa de “Entendimento dos Dados”, nesta etapa realizamos a extração de uma base CSV de amostra em ambiente local. Lê-se essa informação utilizando a linguagem de programação Python e é realizada uma primeira análise exploratória de como estão os dados baixados no arquivo de amostra. Essa etapa é de suma importância para a definição do escopo e das limitações do projeto.

Finalizada a análise de dados exploratória de todas as tabelas selecionadas, segue para a etapa de “Preparação dos Dados”, aqui já é possível mapear alguns indicadores principais que são possíveis com os arquivos de amostra e alguns comparativos entre essas informações. Com o entendimento completo de quais são os indicadores e como os dados se comportam, finalmente desenha-se a estrutura tabelar final do dado.

A próxima parte da fase de preparação é a criação de um script para minerar as informações no portal de dados abertos, com o objetivo de extraí-lo no formato tabelar final estipulado. Essa é a etapa de ETL, a extração do dado em ambiente local, a transformação estipulada na fase anterior como padrão e a carga desses dados em um diretório organizado. Os códigos que compõe o ETL e os dados minerados em si, são então publicados em um repositório público.

Através dos dados organizados em um diretório, diversas são as possibilidades de próximos passos, no caso desta arquitetura em específico, o objetivo é disponibilizar as análises através de *dashboards* públicos utilizando a ferramenta Power BI, liberando o acesso para o usuário final.

Outro ponto importante da arquitetura da informação proposta e ilustrada através da Figura 19 é que o recorte no estudo proposto ocorre na área de orçamento e censo do ensino superior, porém, o método, ou a sequência de processos para chegar as conclusões, pode ser replicado em qualquer outra área de interesse.

# 6 PROVA DE CONCEITO

A prova de conceito tem por objetivo validar e explicar de maneira prática a arquitetura da informação para a extração de bases de dados públicas proposta no capítulo anterior. A mesma estrutura apresentada na Figura 19 será o guia desta prova de conceito.

Os indicadores base selecionados para a construção da prova de conceito são:

● Investimento ao Longo do Tempo: comparativo de queda ou aumento do valor absoluto Grupo de Despesa, investimentos.

● Investimento por Aluno: quanto a instituição dentro de seu Grupo de Despesa de investimentos, investe no seu aluno.

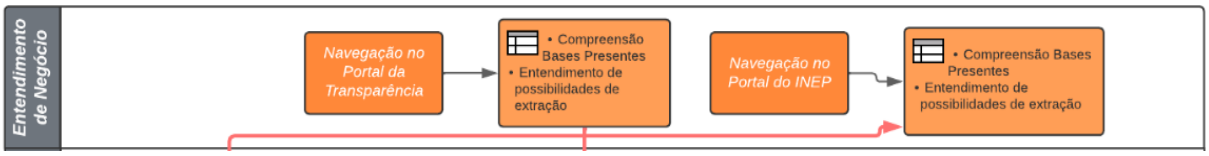
● Custo por Servidor: quanto cada instituição gasta em média no Grupo de Despesa - Pessoal e Encargos com sua folha de funcionários.

● Custo de outras despesas por Pessoa: quanto custa a infraestrutura, que recebe os alunos e professores do ponto de vista tanto dos Alunos que ali frequentam, quanto dos professores que têm as instituições como seus locais de trabalho.

## 6.1 Entendimento de negócio

O Portal da Transparência é um portal público de dados abertos com um acervo gigantesco de informação. Como forma de facilitar a busca e o entendimento dos dados para o cidadão, a plataforma além de disponibilizar dados em diversos formatos, também possui algumas ferramentas de busca e junções de tabelas.

**Figura 20** – Etapa Entendimento de Negócio




Fonte: Lucidchart, [2023], elaborado pelo autor.

Uma dessas facilidades é a ferramenta de detalhamento de despesas públicas (Brasil, 2023). Nesta aba é possível identificar e filtrar por período, por órgão

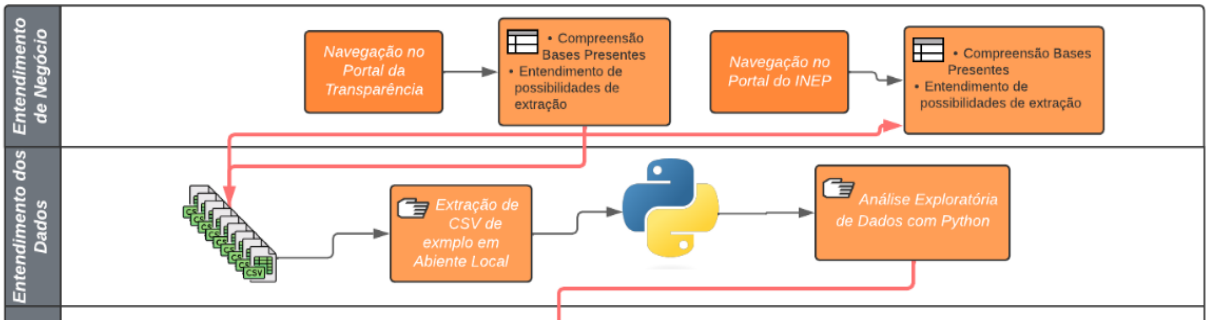
e navegar até o nível da despesa por documento. A plataforma também disponibiliza o botão para *download* do *dataset* gerado a partir dos filtros. Como essa pesquisa se propõe a entender contas relacionadas à educação, podemos facilmente limitar a busca por dados relacionados ao Ministério da Educação, que corresponde ao número identificador: 26000.

Encontrando como a despesa é distribuída fez-se necessário encontrar parâmetros de proporção para relacionar os custos entre instituições. Ao realizar buscas em outros portais de dados abertos públicos destacou-se o Censo da Educação Superior, produzido pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais). Uma coletânea de dados demográficos relacionados à educação superior no país que data desde 1959.

## 6.2 Entendimento dos dados

Foram utilizadas duas bases para o desenvolvimento desta prova de conceito, a base de despesas extraídas pelo portal da transparência e os dados do censo da educação superior pelo INEP.

**Figura 21** – Entendimento dos Dados




Fonte: Lucidchart, [2023], elaborado pelo autor.

Através de uma primeira navegação nos dados das despesas, filtrando-os somente para dados relacionados ao ministério da educação, fica claro que cada linha do resultado representa uma despesa específica, dividida por colunas classificadoras. As colunas são:

'Mês Ano','Órgão/Entidade Vinculada','Unidade Gestora', 'Área de atuação (Função)', 'Subfunção', 'Programa Orçamentário', 'Ação Orçamentária', 'Programa de Governo', 'Autor Emenda', 'Plano Orçamentário', 'Grupo de Despesa', 'Elemento de Despesa', 'Modalidade de Despesa', 'Valor Empenhado','Valor Liquidado', 'Valor Pago', 'Valor Restos a Pagar Pagos'.

**Figura 22** - Dados extraídos do portal da transparência

| DETALHAR | MÊS ANO | ÓRGÃO SUPERIOR | ÓRGÃO/ENTIDADE VINCULADA | UNIDADE GESTORA | ÁREA DE ATUAÇÃO (FUNÇÃO) | SUBFUNÇÃO | PROGRAMA ORÇAMENTÁRIO | AÇÃO ORÇAMENTÁRIA | PROGRAMA DE GOVERNO | NOME DO AUTOR DA EMENDA | PLANO ORÇAMENTÁRIO | GRUPO DE DESPESA | ELEMENTO DE DESPESA | MODALIDADE DE APLICAÇÃO | VALOR EMPENHADO | VALOR LIQUIDADO | VALOR PAGO | VALOR RESTOS A PAGAR PAGOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Detalhar | 12/2022 | 26000 - Ministério da Educação | 26244 - Universidade Federal do Rio Grande do Sul | 153114 - UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL | 12 - Educação | 364 - Ensino superior | 0032 - PROGRAMA DE GESTAO E MANUTENCAO DO PODER EXECUTIVO | 20TP - ATIVOS CIVIS DA UNIAO | 00 - NAO ATRIBUIDO | 0000 | 0000 - ATIVOS CIVIS DA UNIAO | 1 - Pessoal e Encargos Sociais | 07 - Contribuição a Entidades Fechadas de Previdência | 90 - Reserva de Contingência | - 50.422,60 | 321.391,98 | 593.718,45 | 0,00 |
| Detalhar | 12/2022 | 26000 - Ministério da Educação | 26244 - Universidade Federal do Rio Grande do Sul | 153114 - UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL | 12 - Educação | 364 - Ensino superior | 5013 - EDUCACAO SUPERIOR - GRADUACAO, POS-GRADUACAO, ENSINO, PESQUISA E EXTENSAO | 20RK - FUNCIONAMENTO DE INSTITUICOES FEDERAIS DE ENSINO SUPERIOR | 00 - NAO ATRIBUIDO | 0000 | 0000 - FUNCIONAMENTO DE INSTITUICOES FEDERAIS DE ENSINO SUPERIOR | 3 - Outras Despesas Correntes | 36 - Outros Serviços de Terceiros - Pessoa Física | 90 - Reserva de Contingência | 14.184,29 | 1.802,02 | 1.802,02 | 0,00 |
| Detalhar | 12/2022 | 26000 - Ministério da Educação | 26244 - Universidade Federal do Rio Grande do Sul | 153114 - UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL | 12 - Educação | 846 - Outros encargos especiais | 0032 - PROGRAMA DE GESTAO E MANUTENCAO DO PODER EXECUTIVO | 09HB - CONTRIBUICAO DA UNIAO, DE SUAS AUTARQUIAS E FUNDACOES PARA O CUSTEIO DO REGIME DE PREVIDENCIA DOS SERVIDORES PUBLICOS FEDERAIS | 00 - NAO ATRIBUIDO | 0000 | 0000 - CONTRIBUICAO DA UNIAO, DE SUAS AUTARQUIAS E FUNDACOES PARA O CUSTEIO DO REGIME DE PREVIDENCIA DOS SERVIDORES PUBLICOS FEDERAIS | 1 - Pessoal e Encargos Sociais | 13 - Obrigações Patronais | 91 - APLICACOES DIRETAS - OPER.INTRA-ORCAMENTARIAS | - 11.144.612,42 | 14.086.700,84 | 14.086.700,84 | 0,00 |

Fonte: Elaborado pelo autor, 2023.

Pode-se notar que existe uma coluna identificadora chamada “ORGÃO/ENTIDADE VINCULADA” que contém qual o nome da instituição em que a verba foi relacionada. O Grupo da Despesa classifica o tipo das despesas presentes. Somente 3 tipos, pessoal e encargos, outras despesas correntes e investimentos. Cada tipo de despesa possui uma “Ação Orçamentária” que também a classifica, e pode ser ainda mais destrinchado até o “Elemento de Despesa” específico.

Ao fazer o download dos dados do censo da educação, percebe-se sua estrutura muito organizada. Possui pastas separando o material explicativo do repositório dos dados brutos em si. O repositório possui 2 anexos principais, os dicionários dos dados presentes e o questionário do censo em si. Para este estudo, foi utilizado principalmente o dicionário de dados para entendimento.

Existem 2 tabelas de dados presentes, um censo relacionado aos profissionais da instituição superior em si (IES), e outro relacionado aos cursos e quantidades de alunos (CURSOS). Na tabela do IES percebe-se que possui diversas informações referentes a localização e quantidade detalhada de cada tipo de profissional que está nessa instituição (sua formação, nível de especialização entre outros), além do detalhamento de quanto histórico dessa informação existe e grande parte dos dados principais são catalogados desde 2009. Os dados possuem mais de 80 colunas classificadoras

Na tabela de cursos, consta todas as informações detalhadas envolvendo estudantes, e seus respectivos cursos (ingressantes, vagas disponíveis, formandos, alunos matriculados dentre outros). São mais de 200 colunas classificadoras somente neste relatório.

**Figura 23** - Dados extraídos do INEP

| N | Nome da Variável | Descrição da Variável | Tipo | Tam. | Categoria | Coleta por ano ("s"=sim;"-"=não ) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 1 | NU_ANO_CENSO | Ano de referência do Censo da Educação Superior | Num | 4 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| | DADOS DA INSTITUIÇÃO DE ENSINO SUPERIOR (IES) - SEDE ADMINISTRATIVA/REITORIA | | | | | | | | | | | | | | | | | |
| 2 | NO_REGIAO_IES | Nome da região geográfica da sede administrativa ou reitoria da IES | Char | 20 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 3 | CO_REGIAO_IES | Código da região geográfica da sede administrativa ou reitoria da IES | Num | 2 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 4 | NO_UF_IES | Nome da Unidade da Federação da sede administrativa ou reitoria da IES | Char | 50 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 5 | SG_UF_IES | Sigla da Unidade da Federação da sede administrativa ou reitoria da IES | Char | 2 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 6 | CO_UF_IES | Código da Unidade da Federação da sede administrativa ou reitoria da IES | Num | 2 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 7 | NO_MUNICIPIO_IES | Nome do Município da sede administrativa ou reitoria da IES | Char | 150 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 8 | CO_MUNICIPIO_IES | Código do Município da sede administrativa ou reitoria da IES | Num | 7 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 9 | IN_CAPITAL_IES | Informa se a sede administrativa ou reitoria da IES está localizada na capital da Unidade da Federação | Num | 2 | 0. Não<br>1. Sim | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 10 | NO_MESORREGIAO_IES | Nome da Mesorregião da sede administrativa ou reitoria da IES | Char | 100 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 11 | CO_MESORREGIAO_IES | Código da Mesorregião da sede administrativa ou reitoria da IES | Num | 4 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 12 | NO_MICRORREGIAO_IES | Nome da Microrregião da sede administrativa ou reitoria da IES | Char | 100 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 13 | CO_MICRORREGIAO_IES | Código da Microrregião da sede administrativa ou reitoria da IES | Num | 5 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 14 | TP_ORGANIZACAO_ACADEMICA | Tipo de Organização Acadêmica da IES | Num | 1 | 1. Universidade<br>2. Centro Universitário<br>3. Faculdade<br>4. Instituto Federal de Educação, Ciência e Tecnologia<br>5. Centro Federal de Educação Tecnológica | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 15 | TP_CATEGORIA_ADMINISTRATIVA | Tipo de Categoria Administrativa da IES | Num | 1 | 1. Pública Federal<br>2. Pública Estadual<br>3. Pública Municipal<br>4. Privada com fins lucrativos<br>5. Privada sem fins lucrativos<br>6. Privada - Particular em sentido estrito<br>7. Especial<br>8. Privada comunitária<br>9. Privada confessional | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 16 | NO_MANTENEDORA | Nome da mantenedora da IES | Char | 100 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 17 | CO_MANTENEDORA | Código único de identificação da mantenedora da IES | Num | 8 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 18 | CO_IES | Código único de identificação da IES | Num | 8 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 19 | NO_IES | Nome da IES | Char | 200 | | s | s | s | s | s | s | s | s | s | s | s | s | s |
| 20 | SG_IES | Sigla da IES | Char | 20 | | s | s | s | s | s | s | s | s | s | s | s | s | s |

Fonte: INEP, 2022.

Com o entendimento base dos dados em mãos, a preparação dos dados começou a partir da base do portal da transparência. Foi extraído através da plataforma do detalhamento das despesas, um resumo de todas as despesas de janeiro de 2021 relacionadas ao Ministério da Educação - 26000.

Foi gerado uma url de pesquisa (https://portaldatransparencia.gov.br/despesas/orgao/consulta?paginacaoSimples=true&tamanhoPagina=&offset=&direcaoOrdenacao=asc&de=01%2F01%2F2021&ate=28%2F02%2F2021&orgaos=OS26000&colunasSelecionadas=linkDetalhamento%2CmesAno%2CorgaoSuperior%2CorgaoVinculado%2CunidadeGestora%2Cfuncao%2CsubFuncao%2Cprograma%2Cacao%2CprogramaGoverno%2Cautor%2CplanoOrcamentario%2CgrupoDespesa%2CelementoDespesa%2CmodalidadeDespesa%2CvalorDespesaEmpenhada%2CvalorDespesaLiquidada%2CvalorDespesaPaga%2CvalorRestoPago) .

Após o download desses dados foi feita uma contagem de valores únicos da coluna Órgão/Entidade Vinculada para busca manual de algumas principais instituições de ensino, divididas por região.

**Figura 24** - Análise das entidades vinculadas

```
In [4]: dfgeral['Órgão/Entidade Vinculada'].unique()

Out[4]: array(['26256 - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca',
               '26257 - Centro Federal de Educação Tecnológica de Minas Gerais',
               '26201 - Colégio Pedro II',
               '26443 - Empresa Brasileira de Serviços Hospitalares',
               '26291 - Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior',
               '26292 - Fundação Joaquim Nabuco',
               '26271 - Fundação Universidade de Brasília',
               '26270 - Fundação Universidade do Amazonas',
               '26272 - Fundação Universidade do Maranhão',
               '26350 - Fundação Universidade Federal da Grande Dourados',
               '26284 - Fundação Universidade Federal de Ciências da Saúde de Porto Alegre',
               '26278 - Fundação Universidade Federal de Pelotas',
               '26268 - Fundação Universidade Federal de Rondônia',
               '26281 - Fundação Universidade Federal de Sergipe',
```

Fonte: Elaborado pelo autor, 2023.

Para realizar um filtro mais adequado dos dados, as instituições selecionadas foram:

- UNIFESP - '26262 - Universidade Federal de São Paulo',
- UFRGS '26244 - Universidade Federal do Rio Grande do Sul',
- UFMG '26238 - Universidade Federal de Minas Gerais',

- UNB - '26271 - Fundação Universidade de Brasília',
- UFAM - '26270 - Fundação Universidade do Amazonas',
- UFRJ - '26245 - Universidade Federal do Rio de Janeiro',
- UFBA - '26232 - Universidade Federal da Bahia',

Após a localização do identificador único de cada instituição buscada, foi identificado qual era o texto padrão de busca da url para cada instituição, e com o resultado final de: https://portaldatransparencia.gov.br/despesas/orgao/consulta?de=01/01/2021&ate=31/12/2021&orgaos=OR-INSIRA-AQUI-O-ORGAO-&ordenarPor=mesAno&direcao=desc

Foi realizado o download individual de cada instituição do ano de 2021, um tratamento base de tipos de dados foi aplicado em cada dataset e no final foi realizada a junção em um dataset único de despesas do ano de 2021 das instituições filtradas.

Com relação aos dados publicados pelo INEP, após uma detalhada lida na documentação, e formas de criptografia dos dados, foi feita a leitura dos csv's, começando pelo IES. Foi identificado uma coluna chamada SG_IES, que representa a sigla da instituição de ensino, após uma rápida criação de um filtro contendo as siglas das mesmas instituições utilizadas no portal da transparência.

Uma filtragem de colunas foi realizada, para a seguinte formatação: coluns = ['NU_ANO_CENSO', 'NO_REGIAO_IES', 'NO_UF_IES', 'SG_UF_IES','NO_MUNICIPIO_IES','CO_IES', 'NO_IES', 'SG_IES','QT_TEC_TOTAL', 'QT_DOC_TOTAL']. Destacando somente as informações descritivas da instituição, a quantidade de profissionais técnicos totais na instituição e a quantidade total de docentes.

Na investigação da base de CURSOS não foi possível a identificação da mesma coluna SG_IES presente no estudo anterior para filtragem das instituições de ensino. Porém, foi identificado uma coluna compartilhada entre ambas as bases, que é a CO_IES, que ao invés de possuir o nome escrito em si, possui um código identificador único de cada instituição de ensino. Foi então, realizado um filtro para resgatar os mesmos códigos presentes no tratamento final dos dados IES.

Primeiramente filtramos as colunas necessárias para o projeto:

coluns = ['NU_ANO_CENSO','CO_IES','QT_VG_TOTAL', 'QT_INSCRITO_TOTAL', 'QT_ING', 'QT_MAT', 'QT_CONC' ]. Dando destaque para a quantidade de vagas, quantidade de inscritos, quantidade de ingressantes, quantidade de matriculados e quantidade de concluintes. Com essas colunas filtradas, foi realizado um agrupamento somando a quantidade de todas essas colunas para todos os cursos por instituição de ensino e ano.

Com ambos DataSets, foi realizado uma união entre as duas bases para se criar uma base comparativo anual do censo geral das instituições de ensino escolhidas, conforme Figura 23:

**Figura 25** - Exemplo dos dados interligados

| SG_IES | QT_TEC_TOTAL | QT_DOC_TOTAL | QT_VG_TOTAL | QT_INSCRITO_TOTAL | QT_ING | QT_MAT | QT_CONC |
|---|---|---|---|---|---|---|---|
| UNB | 3081 | 2942 | 11131 | 53764 | 6393 | 40145 | 3729 |
| UFAM | 1862 | 1981 | 5852 | 29453 | 2925 | 30467 | 2408 |
| UFMG | 4176 | 3322 | 10361 | 119156 | 7686 | 31655 | 3998 |
| UFBA | 3036 | 3069 | 9463 | 126628 | 7229 | 30567 | 3211 |
| UFRGS | 2573 | 2905 | 8747 | 36251 | 6333 | 34203 | 3117 |
| UFRJ | 4783 | 5181 | 13464 | 150906 | 10725 | 46569 | 3935 |
| UNIFESP | 1873 | 1757 | 5613 | 60238 | 3348 | 12895 | 1908 |
| UNIRIO | 2483 | 938 | 4534 | 65407 | 2545 | 13830 | 1634 |

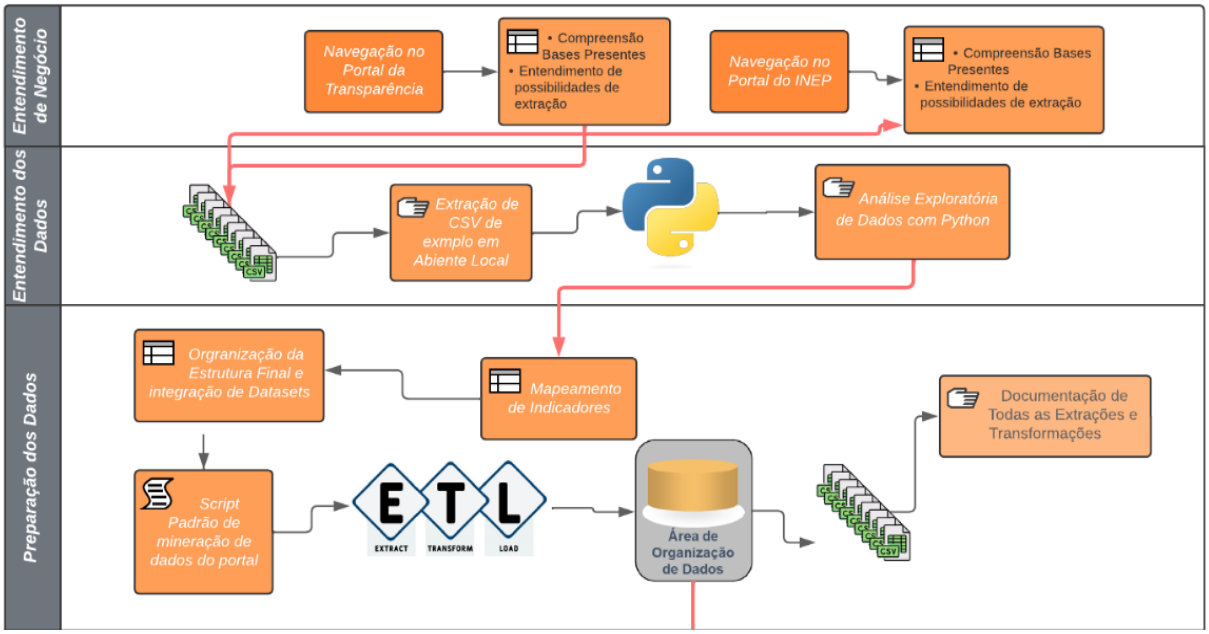
Fonte: Elaborado pelo autor, 2023.

Através da união de ambos os *datasets* do INEP, conseguimos demonstrar algumas informações principais relacionadas a quantitativo de técnicos, professores, alunos e concluintes das instituições superiores selecionadas.

## 6.3 Preparação dos dados

Após a análise exploratória dos dados é possível identificar alguns indicadores para os fins da análise dos dados relacionados a investimento em educação da lei orçamentária e como esse número se comporta relacionado com sua população através do censo do INEP.

**Figura 26** – Preparação dos Dados




Fonte: Lucidchart, [2023], elaborado pelo autor.

### *6.3.1 Mapeamento dos indicadores*

Os principais conjuntos de dados utilizados neste trabalho foram os bancos de dados do senso de educação superior do INEP e a Lei Orçamentaria Anual, o primeiro como forma de encontrar indicadores relacionados a quantitativo de professores, alunos e terceirizados, já a LOA serviu como base comparativa de valores financeiros destinados a cada instituição de ensino. Com ambas tabelas carregadas e relacionadas na plataforma de análise do Power BI, alguns indicadores foram selecionados para comparar as instituições de ensino.

Em relação a LOA, os indicadores selecionados foram: valor total alocado em cada instituição ano a ano; divisão das categorias de custo de cada instituição ano a

ano. Os indicadores relacionados a fonte de dados do INEP escolhidos foram: razão de alunos e professores, razão de alunos e terceiros e razão de professores e terceiros.

Relacionando ambos as tabelas, alguns outros indicadores se tornam possíveis, como por exemplo: o valor alocado da instituição por quantidade de alunos e o valor alocado da instituição em folha por quantidade de funcionários (terceiros e professores). Além dos indicadores inicias propostos no início da validação da hipótese que é quanto cada instituição investe por cada aluno e cada professor e quanto a instituição gasta com custos relacionados a infraestrutura (outras despesas).

### *6.3.2 Organização da estrutura final e integração de datasets*

Os dados extraídos do portal da transparência já vêm em um formato organizado e em uma única tabela, não sendo necessário realizar grandes transformações em seus dados para uma estrutura final de análise.

Já os dados do portal do INEP que originalmente eram formados por 2 tabelas, foram transformados em uma só tabela como apresentado na seção 6.2 de entendimento dos dados.

### *6.3.3 Script padrão de mineração de dados*

A partir das informações catalogadas no decorrer dos capítulos anteriores, se faz necessário extrair uma massa maior de dados para enriquecer as bases comparativas. Fazer isso de maneira manual, poderia gerar erros e problemas no manuseio dos dados. Para garantir a integridade da informação, criou-se uma forma de extrair e tratar os dados utilizando-se de técnicas de Extração, Transformação e Carga dos dados, o ETL citado na fundamentação teórica.

Para gerar o histórico de dados, de todas as instituições escolhidas na validação da hipótese, foi utilizado uma ferramenta de mineração de dados, de *webscrapping* técnica que utiliza a navegação web para automatizar navegação e movimentos na internet dentro da linguagem de programação Python.

**Figura 27** - Código base parte 1

```
import time
import os
import pandas as pd
from selenium import webdriver



PATH = r"C:\Program Files (x86)\chromedriver.exe" # Substitua pelo diretório do chromedriver correto
download_dir = r"C:\Users\Gabriel\Downloads"  # Substitua pelo diretório de downloads correto

anos = list(range(2014, 2021))
faculdades = ['26262', '26244', '26238', '26271', '26270', '26245', '26232']

# Inicializa o driver do Chrome
driver = webdriver.Chrome(executable_path=PATH)

# Configura as opções de download
options = webdriver.ChromeOptions()
prefs = {"download.default_directory": download_dir}
options.add_experimental_option("prefs", prefs)
```

Fonte: Elaborado pelo autor, 2023.

O programa funciona de maneira simples, fazemos a importação das bibliotecas principais utilizadas (algo rotineiro de qualquer código python), incluindo o *Selenium*, nossa ferramenta de navegação *web*. As variáveis principais do código são o “PATH” que é o caminho onde está instalado o programa de acesso à *web*, uma lista de anos para a geração histórica, que no caso desta análise é a partir de 2014 até 2022, esse limite foi definido através de uma sequência de tentativas e erros para buscar o maior número de anos de todas as instituições selecionadas durante a análise exploratória da base do LOA na seção 6.2.

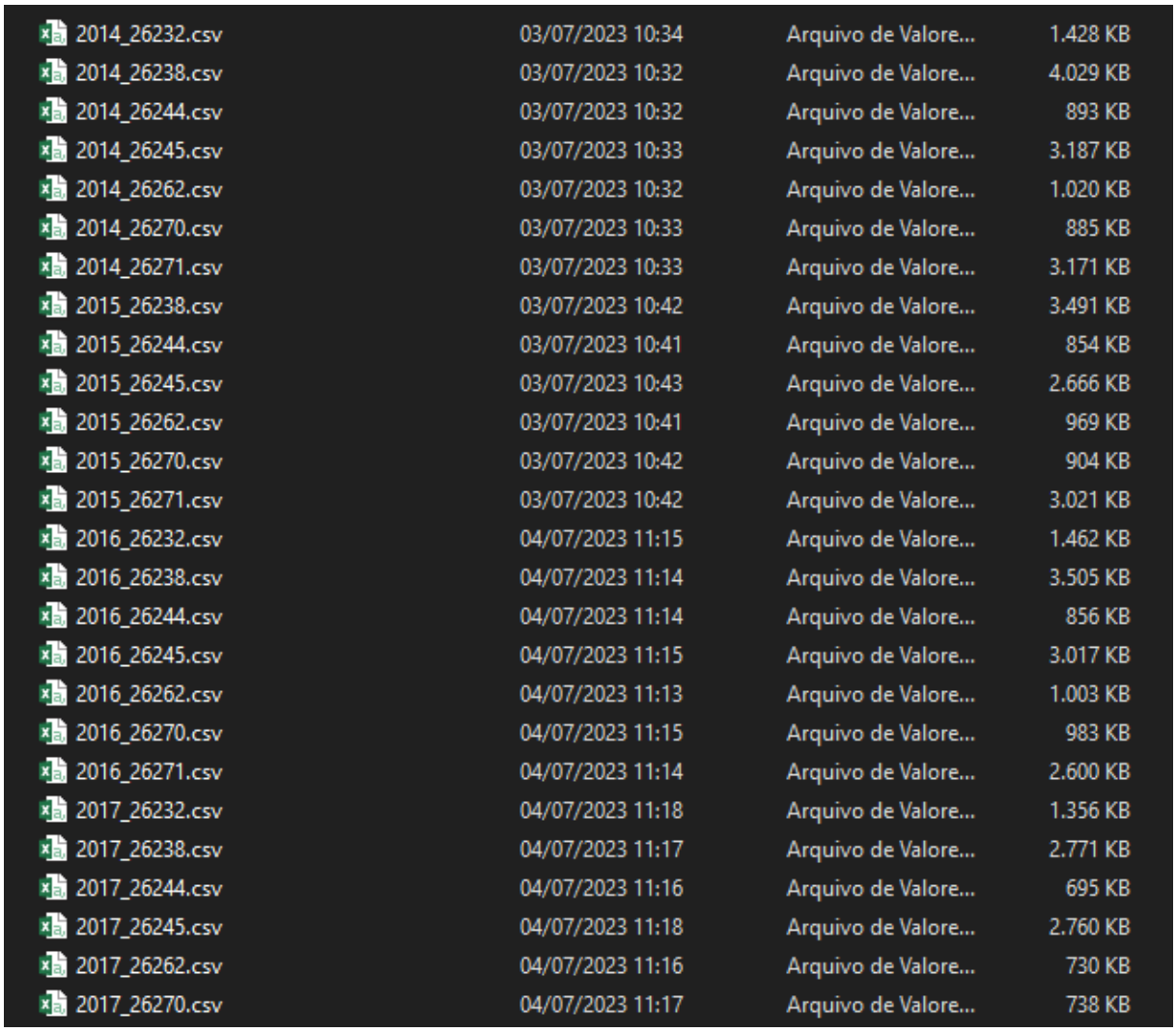
**Figura 28** - Código base parte 2

```
for ano in anos:
    for faculdade in faculdades:
        # Define a URL do link
        url = f"""https://portaldatransparencia.gov.br/despesas/orgao/consulta?de=01/01/{ano}&
        ate=31/12/{ano}&orgaos=OR{faculdade}&ordenarPor=mesAno&direcao=desc"""

        # Acessa a URL
        driver.get(url)

        # Localiza o botão "Baixar" pelo ID e clica nele
        botao_baixar = driver.find_element_by_id("btnBaixar")
        botao_baixar.click()

        # Aguarda 10 segundos para o download ser concluído
        time.sleep(10)

        # Obtém o nome do arquivo baixado

        arquivo_baixado = r"C:\Users\Gabriel\Downloads\despesas.csv"  # Substitua pelo caminho correto
        #dados_despesas = pd.read_csv(arquivo_baixado,encoding='utf8', sep = ';')
        time.sleep(10)


        # Renomeia o arquivo para o padrão "ano_faculdade"
        novo_nome_arquivo = f"{ano}_{faculdade}.csv"
        os.rename(arquivo_baixado, novo_nome_arquivo)

        # Exibe os dados lidos
        print(f"{ano} e {faculdade} finalizados, indo para o próximo")
        time.sleep(3)


# Encerra o driver do Chrome
driver.quit()
```

Fonte: Elaborado pelo autor, 2023.

Fazendo o uso de 2 “*loopings*” criamos uma sequência para navegar por ano e por instituição no site do portal da transparência, baixar o arquivo em ambiente local e salva-lo em uma pasta com a descrição correta de qual instituição e de qual período é o arquivo. O formato final dos downloads é um repositório local, contendo a nomeclatura de “ano_faculdade” salvo em formato CSV.

**Figura 29** - Pasta com arquivos LOA




Fonte: Elaborado pelo autor, 2023.

Os arquivos possuem data de geração e seguem a mesma lógica de *download* e tratamento, garantindo a integridade total da geração e dos dados obtidos. Esses dados também serão disponibilizados no repositório final do projeto. O tratamento dos dados relacionados ao INEP segue uma lógica similar, porém o download de cada uma das bases foi de forma manual. Os arquivos baixados vem no formato .zip, e foram extraídos em uma mesma pasta local.

**Figura 30** - Pasta com arquivos INEP

| | | | |
|---|---|---|---|
| dadosfim | 19/10/2023 10:54 | Pasta de arquivos | |
| Microdados do Censo da Educação Superior 2014 | 31/10/2022 09:48 | Pasta de arquivos | |
| Microdados do Censo da Educação Superior 2015 | 31/10/2022 09:48 | Pasta de arquivos | |
| Microdados do Censo da Educação Superior 2016 | 31/10/2022 09:48 | Pasta de arquivos | |
| Microdados do Censo da Educação Superior 2017 | 31/10/2022 09:48 | Pasta de arquivos | |
| Microdados do Censo da Educação Superior 2018 | 31/10/2022 09:48 | Pasta de arquivos | |
| Microdados do Censo da Educação Superior 2019 | 31/10/2022 09:48 | Pasta de arquivos | |
| Microdados do Censo da Educação Superior 2020 | 31/10/2022 09:48 | Pasta de arquivos | |
| Microdados do Censo da Educação Superior 2021 | 31/10/2022 09:47 | Pasta de arquivos | |
| microdados_censo_da_educacao_superior_2014.zip | 03/07/2023 10:29 | WinRAR ZIP archive | 8.335 KB |
| microdados_censo_da_educacao_superior_2015.zip | 03/07/2023 10:29 | WinRAR ZIP archive | 8.204 KB |
| microdados_censo_da_educacao_superior_2016.zip | 03/07/2023 10:29 | WinRAR ZIP archive | 8.887 KB |
| microdados_censo_da_educacao_superior_2017.zip | 03/07/2023 10:29 | WinRAR ZIP archive | 10.419 KB |
| microdados_censo_da_educacao_superior_2018.zip | 03/07/2023 10:29 | WinRAR ZIP archive | 12.908 KB |
| microdados_censo_da_educacao_superior_2019.zip | 03/07/2023 10:29 | WinRAR ZIP archive | 16.510 KB |
| microdados_censo_da_educacao_superior_2020.zip | 03/07/2023 10:29 | WinRAR ZIP archive | 21.742 KB |
| microdados_censo_da_educacao_superior_2021.zip | 30/06/2023 17:00 | WinRAR ZIP archive | 25.933 KB |

Fonte: Elaborado pelo autor, 2023.

Com cada uma das pastas geradas, utilizando como base o tratamento feito na análise exploratória de dados no decorrer do capítulo anterior foi realizado um *looping* ano a ano dos dados do INEP com o objetivo de ler cada um dos arquivos, aplicar o tratamento necessário de transformação dos dados e de seleção de colunas detalhados no decorrer da seção 6.2.

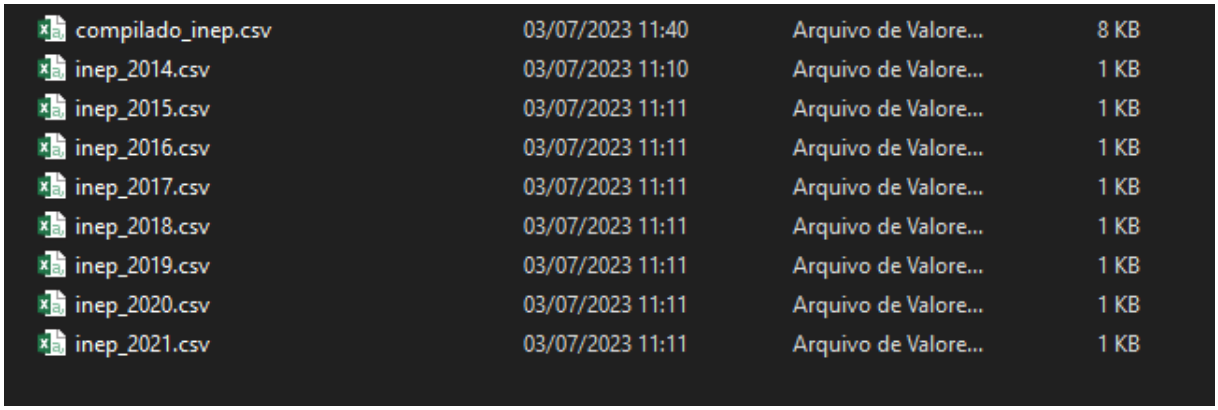
**Figura 31** - Código Base INEP

```
import pandas as pd

listafaculdades = ['UNB', 'UNIFESP', 'UFRGS', 'UFMG', 'UFAM', 'UFRJ', 'UFBA']
caminhobase = (r"C:\Users\Gabriel\Documents\Mestrado\dados\INEP")

for year in range(2014, 2022):
    #caminhobase = "caminho_da_base"  # substitua pelo caminho adequado
    filename_ies = f"\Microdados do Censo da Educação Superior {year}\dados\MICRODADOS_CADASTRO_IES_{year}.CSV"
    filename_alunos = f"\Microdados do Censo da Educação Superior {year}\dados\MICRODADOS_CADASTRO_CURSOS_{year}.csv"

    dfies = pd.read_csv(caminhobase + filename_ies, sep=";", encoding='iso-8859-1', dtype=str)
    dfies = dfies[dfies['TP_CATEGORIA_ADMINISTRATIVA'] == '1']
    dfies = dfies[dfies['SG_IES'].isin(listafaculdades)]
    colunsdfies = ['NU_ANO_CENSO', 'NO_REGIAO_IES', 'NO_UF_IES', 'SG_UF_IES','NO_MUNICIPIO_IES','CO_IES', 'NO_IES', 'SG_IES','QT
    dfies = dfies[colunsdfies]

    dfalunos = pd.read_csv(caminhobase + filename_alunos, sep=";", encoding='iso-8859-1', dtype=str)
    dfalunos = dfalunos[dfalunos['CO_IES'].isin(dfies.CO_IES)]
    colunsdfalunos = ['NU_ANO_CENSO','CO_IES','QT_VG_TOTAL', 'QT_INSCRITO_TOTAL', 'QT_ING', 'QT_MAT', 'QT_CONC' ]

    dfalunos = dfalunos[colunsdfalunos]

    dfalunos = dfalunos.groupby(['NU_ANO_CENSO', 'CO_IES'])["QT_VG_TOTAL",
                                                            'QT_INSCRITO_TOTAL',
                                                            'QT_ING', 'QT_MAT',
                                                            'QT_CONC'].apply(lambda x: x.fillna(0).astype(int).sum()).reset_inde

    dfjoin = pd.merge(dfies, dfalunos, how='left', left_on=["NU_ANO_CENSO", "CO_IES"], right_on=["NU_ANO_CENSO", "CO_IES"])
    dfjoin.to_csv()
    # Restante do código usando o dfjoin gerado

    filename_output = f"inep_{year}.csv"
    dfjoin.to_csv(filename_output, index=False)

    print(f"Loop concluído para o ano {year}.")
```

Fonte: Elaborado pelo autor, 2023.

Existem duas páginas principais nos arquivos dos dados extraídos do censo de educação, uma tabela das instituições de ensino e uma tabela relacionado a quantitativo de alunos por curso da instituição de ensino. O tratamento dos dados, primeiro foca os dados da Instituição de Ensino, e depois realiza um agrupamento junto da tabela com dados relacionados a alunos de cada instituição de ensino.

**Figura 32** - Pasta com arquivos tratados INEP




Fonte: Elaborado pelo autor, 2023.

Os dados finais do INEP são tratados individualmente ano a ano, salvos em ambiente local, sempre em formato CSV e por fim agrupados em um arquivo final, chamado “compilado_inep.csv”.

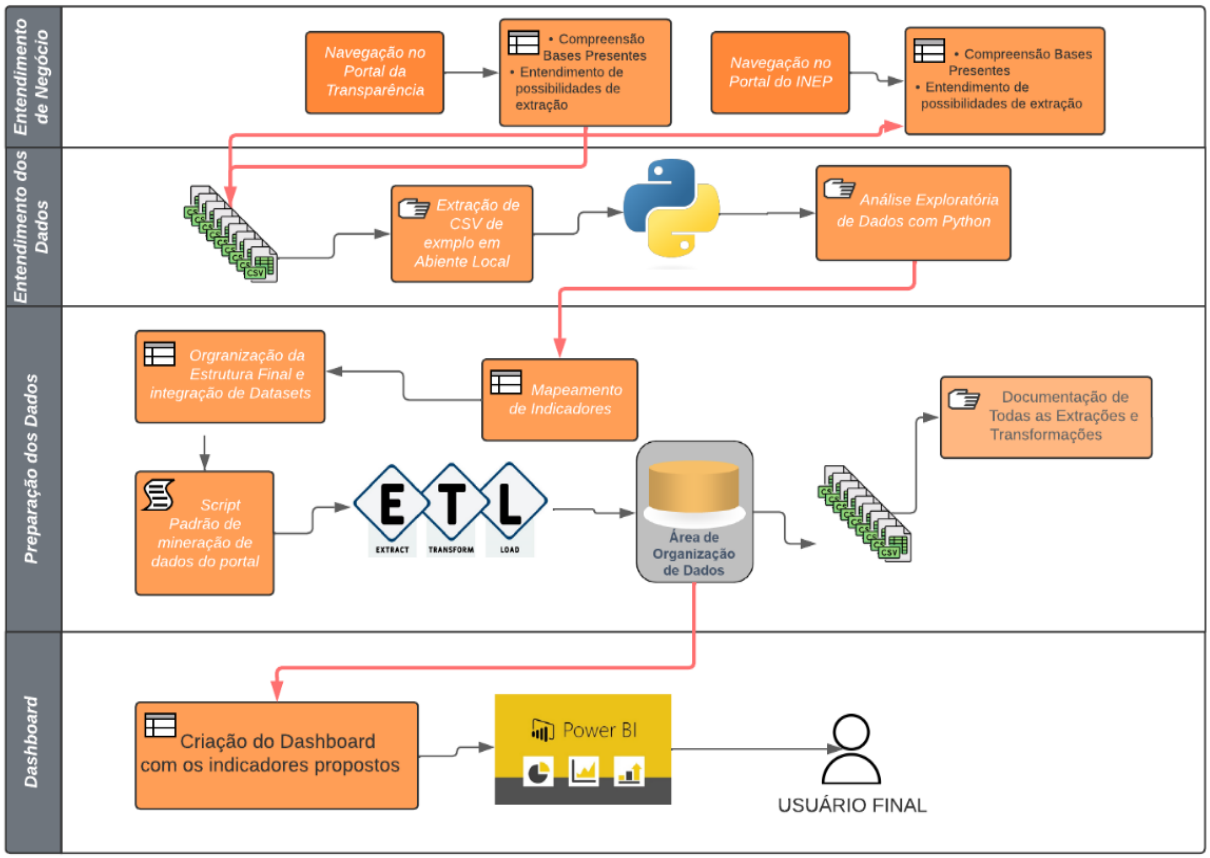
### *6.3.4 Área de Organização de Dados e Documentação*

Com os dados organizados e catalogados em ambiente local foi organizado um repositório online com todos os scripts e dados utilizados nesse projeto disponibilizados dentro de uma pasta aberta no Google Drive, uma ferramenta de armazenamento de arquivos que mantem os metadados intactos desde sua geração. Os arquivos contendo os códigos fontes podem ser encontrados através do link: https://drive.google.com/drive/folders/1I7FG1WoeBKNsCIdfGpFXsvt-R36hOrcN?usp=sharing . O painel público que foi ponto chave para as análises pode ser acessado através do link: https://shorturl.at/lrxAB .

## 6.4 Dashboard

Para a criação do *dashboard* como objetivo deste trabalho além do compartilhamento dos dados gerados, a ferramenta escolhida foi o Power BI pela sua facilidade de divulgação e publicação das informações geradas. A ferramenta usa como fontes de dados ambos os arquivos compilados gerados.

**Figura 33 -** *Dashboard*

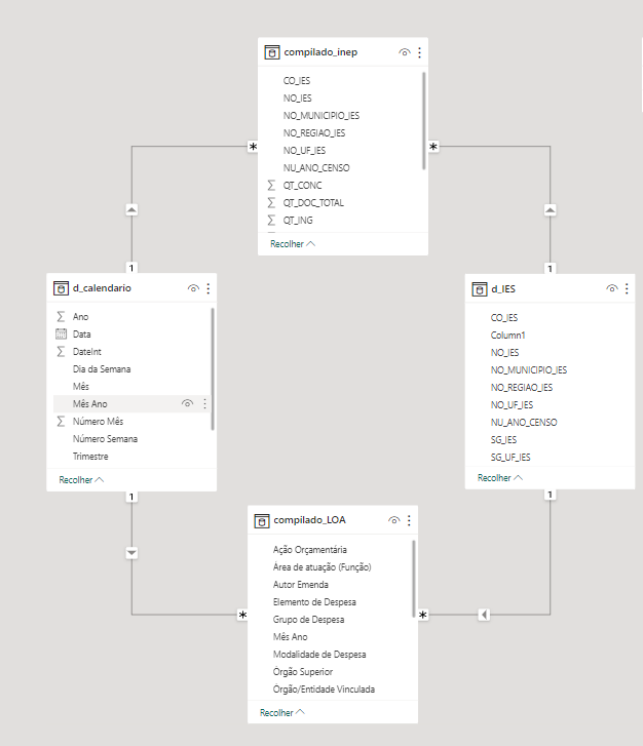



Fonte: Lucidchart, [2023], elaborado pelo autor.

A estrutura final da arquitetura da informação proposta para o desenvolver da prova de conceito pode ser visualizada na figura 33. A partir dos dados presentes no repositório de organização dos dados a estrutura base do *dashboard* foi desenvolvida para os fins de facilitação na disponibilização das informações geradas para o usuário fim.

Através da análise exploratória de dados desenvolvida e exemplificada no decorrer Capítulo 6 foi possível propor um modelo relacional para dados com diferentes fontes. O modelo relacional utilizado para disponibilizar a informação pode ser visto de acordo com a Figura 34:

**Figura 34** - Modelo Relacional

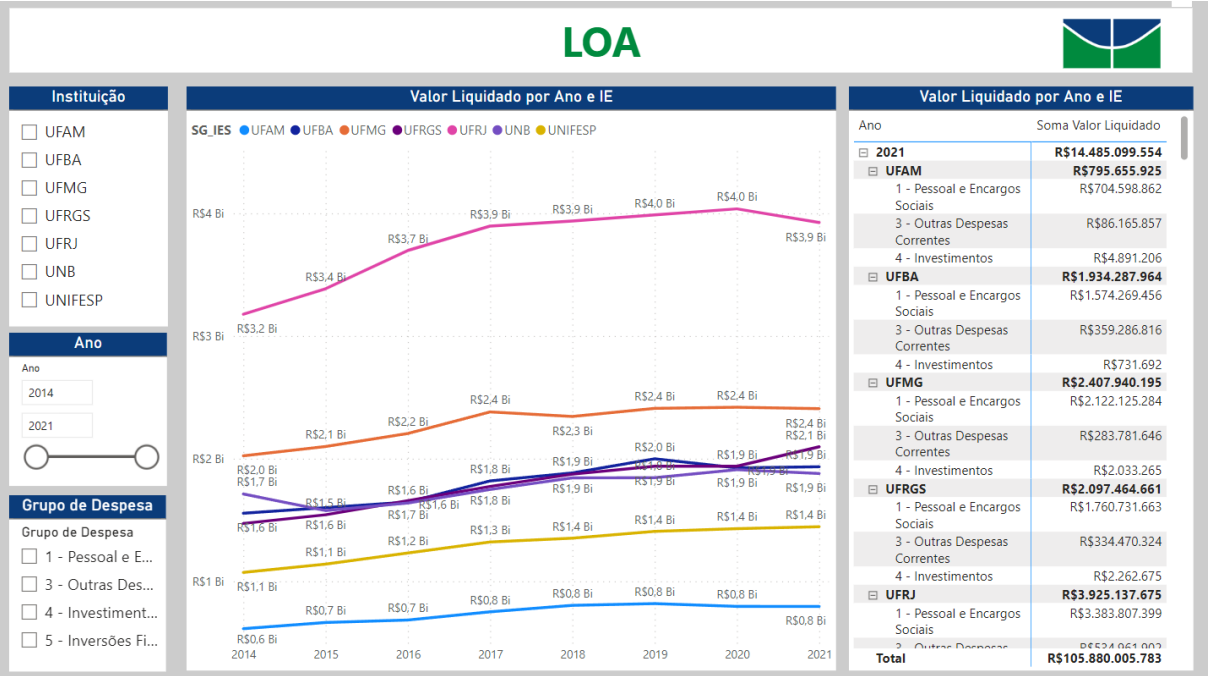



Fonte: Elaborado pelo autor, 2023.

As tabelas de base comparativa que relacionam ambos os *datasets* são uma tabela dimensão com as informações de características das instituições de ensino, que se relacionam através da coluna “Sigla Instituição de Ensino”, presente em ambas tabelas. Para comparar valores do ponto de vista temporal, se faz necessário uma outra tabela auxiliar de “Calendário” que também relaciona ambas as tabelas e torna possível realizar filtros relacionados a “datas”.

Tomando como base os indicadores relacionados a LOA, primeiramente o valor total alocado em cada instituição ano a ano, essa informação pode ser vista através do *Dashboard* LOA presente na Figura 35, no gráfico “Valor Liquidado por Ano e IE” em formato gráfico, e no formato tabelar através da tabela “Valor Liquidado por Ano e IE”.

**Figura 35** - *Dashboard* LOA

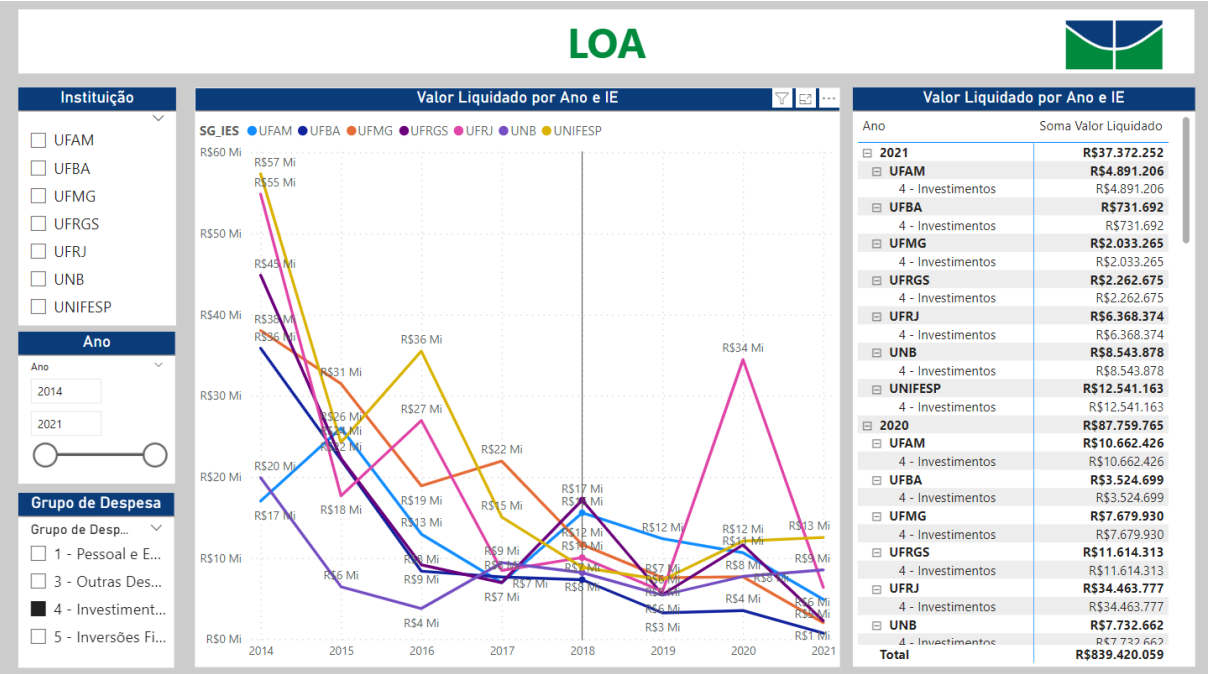



Fonte: Elaborado pelo autor, 2023.

Do ponto de vista de análise das categorias de verba disponíveis na LOA, podemos tanto filtrar essas informações através do filtro “Grupo de Despesa” como também já o visualizar no gráfico tabelar “Valor Liquidado por Ano e IE”. O objetivo desta página é tornar fácil e livre a navegação e comparativo entre instituições, por isso a página também conta com um filtro de multi-seleção de “Instituição” e de “Ano”.

Com a visualização do valor gasto ao longo do tempo, é notável que a UFRJ é a instituição que mais recebeu verba em comparação com todas as outras instituições escolhidas. UNB, UFRGS e UFBA apresentam um investimento total similar entre as instituições. A universidade que menos recebe verba neste comparativo é a UFAM.

**Figura 36** - *Dashboard* Investimento ao longo do tempo

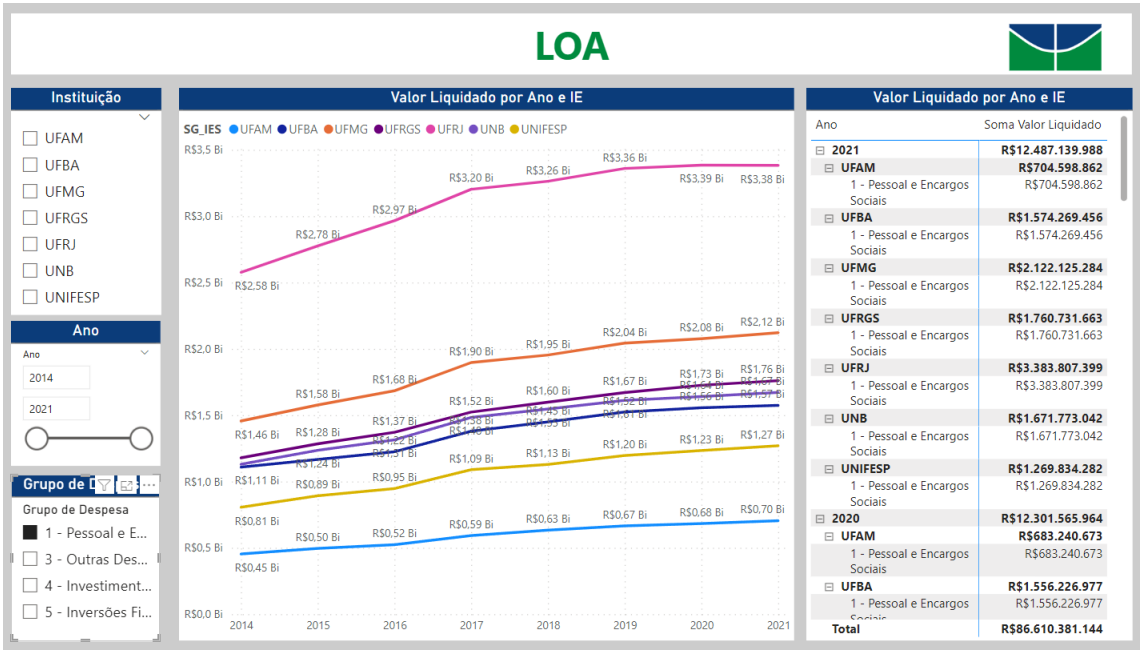



Fonte: Elaborado pelo autor, 2023.

Ao selecionar no filtro “Grupo de Despesa” a linha de “ 4 – Investimentos” pode ser visualizado na figura 36 como a linha de despesa de “Investimentos” é uma constante para baixo a partir de 2015 mesmo o valor total alocado em cada instituição sendo um crescente positivo ano após ano. Para investigar com precisão o que aconteceu com o dinheiro alocado em cada linha da lei orçamentária anual, seria necessário uma investigação mais detalhada linha a linha nas instituições levantadas para entender onde ocorreram as maiores variações.

Ao filtrar o “Grupo de Despesa” na linha “1 – Pessoal e encargos” fica evidente na Figura 37 que este grupo de Despesa é uma constante crescente em cada ano que passa. Os dados relacionados a população do quadro de funcionários da instituição podem complementar a análise, caso essa população também tenha aumentado ao longo do tempo, poderia justificar também o aumento constante desta linha de despesa.

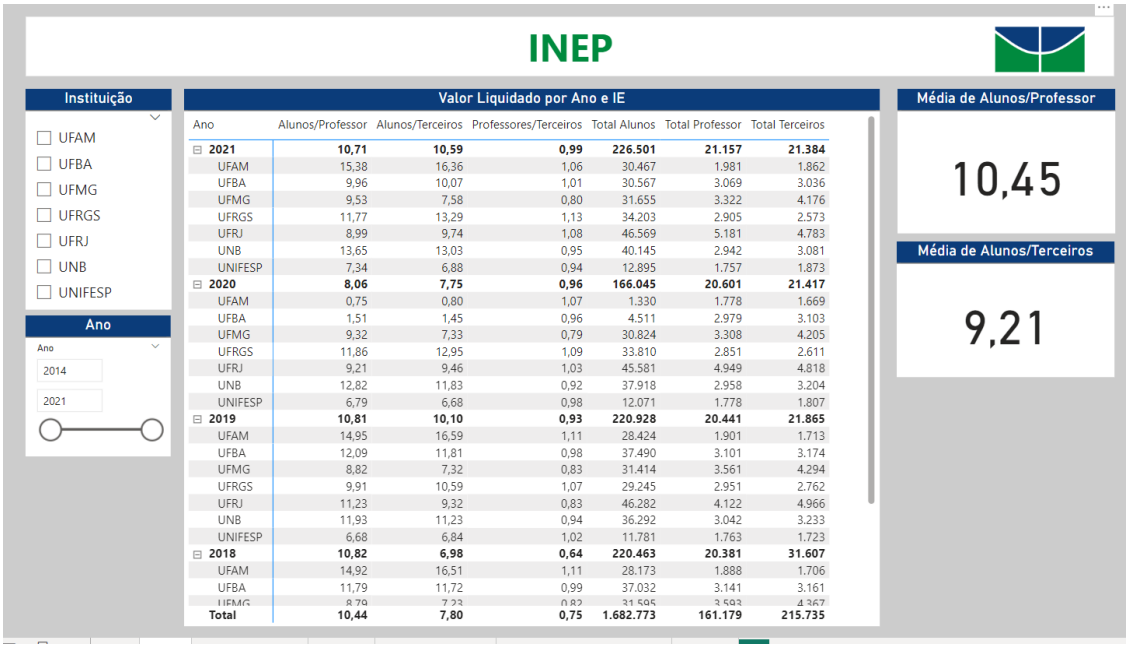
**Figura 37** - *Dashboard* Pessoal e Encargos ao longo do tempo




Fonte: Elaborado pelo autor, 2023.

Para os indicadores relacionados a fonte de dados oriundos do INEP, são três principais indicadores, e todos são razões comparativas entre populações. Sendo elas: razão de alunos e professores, razão de alunos e terceiros e razão de professores e terceiros.

**Figura 38** - *Dashboard* INEP



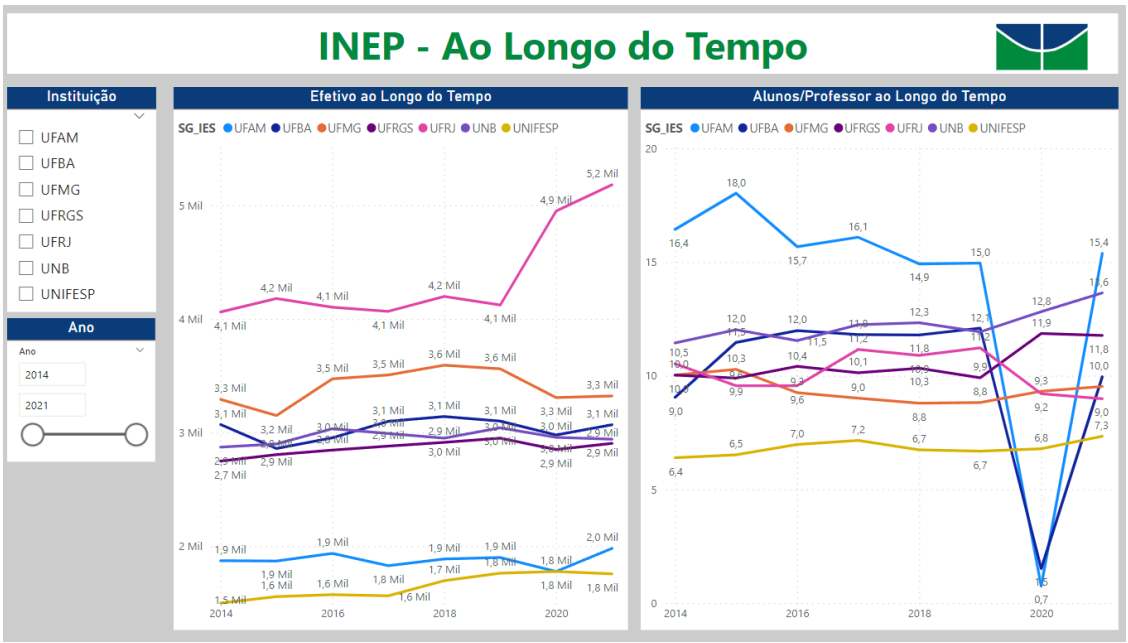

Fonte: Elaborado pelo autor, 2023.

Todos os indicadores citados podem ser visualizados na Figura 38, na tabela “Valor Liquidado por Ano e IE”. Esses valores são agrupados por ano e instituição, e também possuem os mesmos filtros da página anterior, tanto de “Instituição”, quanto de “Ano”. O filtro de grupo de despesa não faz sentido nessa visualização visto que não estamos abordando valores financeiros nesta parte da análise. Além das principais razões vistas adicionadas, também uma parte contendo o total de cada uma dessas populações (alunos, professores e terceiros).

A média de Alunos/Professores segue uma média de 10,45 entre as instituições buscadas, tomando como base o ano de 2021, UFAM e UNB se destacam por uma quantidade maior de alunos para cada professor. O grande *outlier* dessa equação é a UNIFESP que apresenta uma média de 7,34 alunos por professor, faz-se necessário mais informações sobre a instituição para entender o porquê dessa média ser tão baixa.

Através das considerações da linha de despesa de custo de pessoal e encargos no gráfico anterior, é necessário também uma visualização ao longo do tempo de como esses números se comportam. Esses números de população podem vistos na Figura 39.

**Figura 39** - *Dashboard* INEP Ano a Ano




Fonte: Elaborado pelo autor, 2023.

No gráfico de “Efetivo ao longo do Tempo”, soma-se o quantitativo de professores e terceirizados presentes nos dados do INEP. Fica evidente o crescente no aumento de efetivo da UFRJ a partir de 2018 e nas outras instituições uma grande constância. No gráfico da razão entre alunos e professores ao longo do tempo fica evidente dados faltantes para a instituição da UFAM e da UFBA no ano de 2020. Melhora de quantitativo de professores por alunos para a UNB e UFGRS que vem em uma constante crescente nesse indicador.

Após a análise de cada fonte de dados de maneira individual, por fim a parte final dos indicadores que cruzam ambas as tabelas, tornando possível a realização de novas análises mesmo com bancos não previamente relacionados. Os indicadores selecionados foram: valor alocado da instituição por quantidade de alunos e o valor alocado da instituição em folha por quantidade de funcionários (terceiros e professores). Esses indicadores podem ser visualizados na Figura 40.

**Figura 40** - *Dashboard* Indicadores Cruzados


| Ano | Valor Liquidado por Aluno Anual | Valor Liquidado por Efetivo Anual | Valor Liquidado por Aluno Mensal | Valor Liquidado por Efetivo Mensal |
|---|---|---|---|---|
| **2021** | **R$63.952** | **R$340.497** | **R$5.329** | **R$28.375** |
| UFAM | R$26.115 | R$207.040 | R$2.176 | R$17.253 |
| UFBA | R$63.280 | R$316.837 | R$5.273 | R$26.403 |
| UFMG | R$76.068 | R$321.144 | R$6.339 | R$26.762 |
| UFRGS | R$61.324 | R$382.889 | R$5.110 | R$31.907 |
| UFRJ | R$84.286 | R$393.932 | R$7.024 | R$32.828 |
| UNB | R$46.810 | R$312.000 | R$3.901 | R$26.000 |
| UNIFESP | R$112.093 | R$398.193 | R$9.341 | R$33.183 |
| **2020** | **R$87.060** | **R$344.041** | **R$7.255** | **R$28.670** |
| UFAM | R$598.528 | R$230.938 | R$49.877 | R$19.245 |
| UFBA | R$426.794 | R$316.552 | R$35.566 | R$26.379 |
| UFMG | R$78.491 | R$322.029 | R$6.541 | R$26.836 |
| UFRGS | R$57.333 | R$354.892 | R$4.778 | R$29.574 |
| UFRJ | R$88.563 | R$413.309 | R$7.380 | R$34.442 |
| UNB | R$50.383 | R$310.034 | R$4.199 | R$25.836 |
| UNIFESP | R$118.429 | R$398.761 | R$9.869 | R$33.230 |
| **2019** | **R$65.213** | **R$340.549** | **R$5.434** | **R$28.379** |
| UFAM | R$28.802 | R$226.528 | R$2.400 | R$18.877 |
| UFBA | R$53.325 | R$318.591 | R$4.444 | R$26.549 |
| UFMG | R$76.725 | R$306.842 | R$6.394 | R$25.570 |
| UFRGS | R$66.297 | R$339.375 | R$5.525 | R$28.281 |
| UFRJ | R$86.162 | R$438.794 | R$7.180 | R$36.566 |
| UNB | R$50.841 | R$294.041 | R$4.237 | R$24.503 |
| UNIFESP | R$119.471 | R$403.754 | R$9.956 | R$33.646 |
| **2018** | **R$63.683** | **R$270.058** | **R$5.307** | **R$22.505** |
| UFAM | R$28.555 | R$223.838 | R$2.380 | R$18.653 |
| UFBA | R$50.873 | R$298.941 | R$4.239 | R$24.912 |
| UFMG | R$74.196 | R$294.501 | R$6.183 | R$24.542 |
| **Total** | **R$62.920** | **R$280.913** | **R$5.243** | **R$23.409** |

Fonte: Elaborado pelo autor, 2023.

Na tabela “Valor Liquidado por Ano e IE” possuímos os principais indicadores comentados. Um valor alocado por aluno e o valor alocado pelo efetivo de funcionários (professores mais terceirizados), além disso uma perspectiva do valor mensal alocado

foi adicionada também, para entendermos quanto a instituição tem de orçamento por aluno mensalmente.

O número que mais sobressai, é o da UNIFESP com uma média de Valor Anual Liquidado por aluno de mais de 110 mil reais. Transformando um custo mensal por cada aluno na instituição em uma mensalidade de mais de R$ 9.000 reais. A UFAM é a que menos tem de custo proporcional por alunos. Faz sentido uma investigação a fundo para buscar compreender se existem relação entre o tempo de vida da instituição e o passivo dos custos de folha.

**Figura 41** - *Dashboard* Indicadores Cruzados Filtrado


Fonte: Elaborado pelo autor, 2023.

Nesta página existem os filtros de “Instituição”, “Ano” e “Grupo de Despesa”, através deste último, se torna possível observamos alguns indicadores a mais, como o ultimo indicador citado, o custo do efetivo cruzando com os valores do “Grupo de Despesa” sendo igual a “ 1 – Pessoal e Encargos Sociais”. Como apresentado na Figura 41

A Figura 42 detalha que ao filtrar o Grupo de Despesa “3 – Outras Despesas Correntes” fica evidente que o valor alocado por pessoa nessa linha de despesa é

correspondente a um valor bem abaixo do ponto de vista de alunos do que de professores.

**Figura 42** - *Dashboard* Indicadores Cruzados Filtrado 2


Valor Liquidado por Ano e IE

| Ano | Valor Liquidado por Aluno Anual | Valor Liquidado por Efetivo Anual | Valor Liquidado por Aluno Mensal | Valor Liquidado por Efetivo Mensal |
|---|---|---|---|---|
| **2021** | **R$8.656** | **R$46.087** | **R$721** | **R$3.841** |
| UFAM | R$2.828 | R$22.422 | R$236 | R$1.868 |
| UFBA | R$11.754 | R$58.851 | R$980 | R$4.904 |
| UFMG | R$8.965 | R$37.848 | R$747 | R$3.154 |
| UFRGS | R$9.779 | R$61.057 | R$815 | R$5.088 |
| UFRJ | R$11.488 | R$53.689 | R$957 | R$4.474 |
| UNB | R$4.953 | R$33.016 | R$413 | R$2.751 |
| UNIFESP | R$12.646 | R$44.921 | R$1.054 | R$3.743 |
| **2020** | **R$12.446** | **R$49.183** | **R$1.037** | **R$4.099** |
| UFAM | R$76.796 | R$29.631 | R$6.400 | R$2.469 |
| UFBA | R$81.028 | R$60.098 | R$6.752 | R$5.008 |
| UFMG | R$10.861 | R$44.560 | R$905 | R$3.713 |
| UFRGS | R$5.952 | R$36.844 | R$496 | R$3.070 |
| UFRJ | R$13.543 | R$63.203 | R$1.129 | R$5.267 |
| UNB | R$6.910 | R$42.522 | R$576 | R$3.544 |
| UNIFESP | R$15.208 | R$51.207 | R$1.267 | R$4.267 |
| **2019** | **R$10.354** | **R$54.071** | **R$863** | **R$4.506** |
| UFAM | R$4.941 | R$38.862 | R$412 | R$3.239 |
| UFBA | R$12.580 | R$75.162 | R$1.048 | R$6.263 |
| UFMG | R$11.444 | R$45.768 | R$954 | R$3.814 |
| UFRGS | R$8.973 | R$45.934 | R$748 | R$3.828 |
| UFRJ | R$13.434 | R$68.413 | R$1.119 | R$5.701 |
| UNB | R$6.279 | R$36.313 | R$523 | R$3.026 |
| UNIFESP | R$17.308 | R$58.493 | R$1.442 | R$4.874 |
| **2018** | **R$10.826** | **R$45.908** | **R$902** | **R$3.826** |
| UFAM | R$5.507 | R$43.171 | R$459 | R$3.598 |
| UFBA | R$11.533 | R$67.771 | R$961 | R$5.648 |
| UFMG | R$11.994 | R$47.608 | R$1.000 | R$3.967 |
| **Total** | **R$10.944** | **R$48.860** | **R$912** | **R$4.072** |

Fonte: Elaborado pelo autor, 2023.

Ao filtrar o campo para o grupo de despesa “4 – Investimentos” fica evidente o valor bem baixo por aluno de todas as instituições. Sendo um valor mensal por aluno próximo a 30 reais em média nos últimos 4 anos.

**Figura 43** - *Dashboard* Indicadores Cruzados Filtrado 3

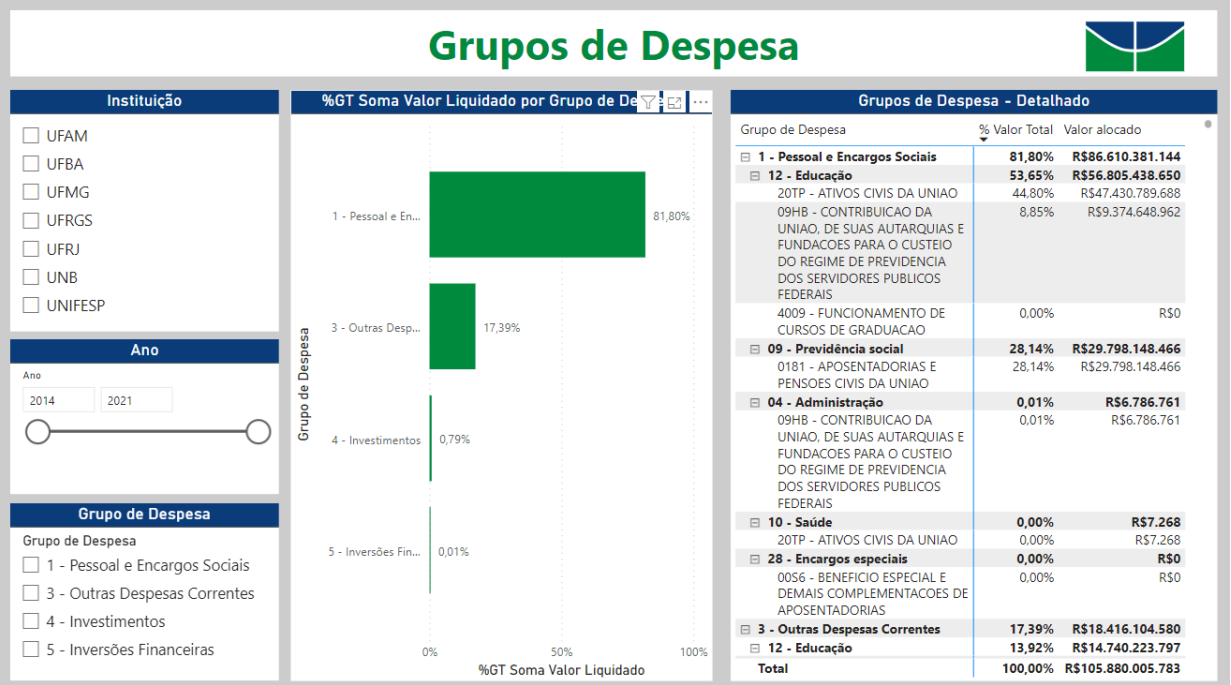

Valor Liquidado por Ano e IE

| Ano | Valor Liquidado por Aluno Anual | Valor Liquidado por Efetivo Anual | Valor Liquidado por Aluno Mensal | Valor Liquidado por Efetivo Mensal |
|---|---|---|---|---|
| **2021** | **R$165** | **R$878** | **R$14** | **R$73** |
| UFAM | R$161 | R$1.273 | R$13 | R$106 |
| UFBA | R$24 | R$120 | R$2 | R$10 |
| UFMG | R$64 | R$271 | R$5 | R$23 |
| UFRGS | R$66 | R$413 | R$6 | R$34 |
| UFRJ | R$137 | R$639 | R$11 | R$53 |
| UNB | R$213 | R$1.419 | R$18 | R$118 |
| UNIFESP | R$973 | R$3.455 | R$81 | R$288 |
| **2020** | **R$529** | **R$2.089** | **R$44** | **R$174** |
| UFAM | R$8.017 | R$3.093 | R$668 | R$258 |
| UFBA | R$781 | R$580 | R$65 | R$48 |
| UFMG | R$249 | R$1.022 | R$21 | R$85 |
| UFRGS | R$344 | R$2.126 | R$29 | R$177 |
| UFRJ | R$756 | R$3.529 | R$63 | R$294 |
| UNB | R$204 | R$1.255 | R$17 | R$105 |
| UNIFESP | R$1.001 | R$3.370 | R$83 | R$281 |
| **2019** | **R$214** | **R$1.118** | **R$18** | **R$93** |
| UFAM | R$436 | R$3.425 | R$36 | R$285 |
| UFBA | R$86 | R$515 | R$7 | R$43 |
| UFMG | R$241 | R$964 | R$20 | R$80 |
| UFRGS | R$188 | R$964 | R$16 | R$80 |
| UFRJ | R$127 | R$649 | R$11 | R$54 |
| UNB | R$150 | R$865 | R$12 | R$72 |
| UNIFESP | R$619 | R$2.092 | R$52 | R$174 |
| **2018** | **R$357** | **R$1.516** | **R$30** | **R$126** |
| UFAM | R$553 | R$4.335 | R$46 | R$361 |
| UFBA | R$198 | R$1.163 | R$16 | R$97 |
| UFMG | R$369 | R$1.465 | R$31 | R$122 |
| **Total** | **R$499** | **R$2.227** | **R$42** | **R$186** |

Fonte: Elaborado pelo autor, 2023.

A Figura 44 mostra que de todas as 7 instituições selecionadas para a validação da arquitetura da informação quase 82% dos valores destinados são para a linha de “Pessoal e Encargos” e praticamente um terço dessa despesa dentro da “Ação Orçamentária” de previdência social.

**Figura 44** - *Dashboard* Grupos de Despesa




Fonte: Elaborado pelo autor, 2023.

Os custos de “3 - Outras Despesas” representa o custo de manutenção e uso da estrutura das instituições, é perceptível que esse valor fica reduzido em comparação ao resto dos custos, sendo menos de 18% da linha da despesa. Para entender mais a fundo como esse gasto é correlacionado com a área e estrutura de cada instituição se faz necessário um outro estudo. A linha de investimentos também fica comprometida com menos de 1% de todo o valor alocado depositado nessa linha, limitando as possibilidades de melhora e expansão de oportunidades para os alunos.

# 7 CONSIDERAÇÕES FINAIS

Através da arquitetura da informação proposta no desenvolvimento do estudo foi possível analisar e compreender conjuntos de dados abertos públicos originários de diferentes fontes, sendo as escolhidas o portal da transparência, através dos dados da LOA e o INEP com o censo de educação superior.

Os dados foram identificados e correlacionados entre si através da prova de conceito exposta no Capítulo 6. A Análise exploratória dos dados foi amparada através do processo CRISP-DM detalhado na seção 2.2 e posto à prova na seção 6.2. O modelo relacional dos dados foi desenvolvido através de técnicas de tratamento de dados e viabilizada através da ferramenta do Power BI.

O produto final da Arquitetura da Informação pôde ser visualizado no Capítulo 5 e detalhado através da prova de conceito apresentada no decorrer do Capítulo 6. A prova de conceito é por fim um *dashboard* que pode ser acessado através do link público (https://shorturl.at/lrxAB).

Os indicadores que serviram como base para a prova de conceito foram explorados no decorrer da seção 6.4. Ao reparar o grupo de despesa “investimentos” nos dados presentes na LOA é evidente esses valores nas instituições estão reduzidos e em constante queda. Faz necessário também entender qual é o número ótimo da razão entre alunos/professores para utilizar como base comparativa.

Essa é uma primeira parte de vários outros possíveis estudos envolvendo dados abertos relacionados a investimento em educação, um tema que nunca diminui sua importância e é peça crucial para o desenvolvimento da nação. Alguns outros exemplos de estudos futuros propostos são: a relação de investimento das instituições, e as notas recebidas pelos órgãos de medições de resultados do ensino superior; outra opção seria extrair uma amostra maior de ensino superior público no país e encontrar as médias reais, levando em consideração todas as IEs; por fim, uma outra proposta seria entender também, como as instituições privadas de ensino no Brasil, mantém essa média em relação a Alunos e professor e Alunos e terceirizados.

# REFERÊNCIAS


ANGELONI, M. T.; REIS, M. F. R. A. **Introdução à metodologia do trabalho científico**. 3. ed. São Paulo: Atlas, 2006.

ANTONELLI, R. A. **Conhecendo o Business Intelligence (BI)**. Curitiba: TECAP, 2009.

ATTARD, J.; ORLANDI, F.; SCERRI, S.; AUER, S. A systematic review of open government data initiatives. **Government Information Quarterly**, v. 32, n. 4, p. 399-418, 2015.

BARBIERI, C. **BI**: Business Intelligence: modelagem e tecnologia. Rio de Janeiro: Axcel Books, 2001

BATISTA, E. O. **Sistemas de informação**. São Paulo: Saraiva, 2004

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific american**, v. 284, n. 5, p. 28-37, 2001.

BRANCHEAU, J. C.; WETHERBE, J. C. Information Architectures: Methods and Pratice. **Information Processing & Management**, v. 22, n. 6, p. 453-463, 1986.

BRASIL. Controladoria Geral da União. **Portal da transparência**. [2023b]. Disponível em: https://portaldatransparencia.gov.br. Acesso em: 07 jun. 2023.

BRASIL. Controladoria Geral da União. Portal da transparência: detalhamento da despesa pública. **Gov.br**, 2021a. Disponível em: https://portaldatransparencia.gov.br/despesas/orgao/consulta?paginacaoSimples=true&tamanhoPagina=&offset=&direcaoOrdenacao=asc&de=01%2F01%2F2021&ate=28%2F02%2F2021&orgaos=OS26000&colunasSelecionadas=linkDetalhamento%2CmesAno%2CorgaoSuperior%2CorgaoVinculado%2CunidadeGestora%2Cfuncao%2CsubFuncao%2Cprograma%2Cacao%2CprogramaGoverno%2Cautor%2CplanoOrcamentario%2CgrupoDespesa%2CelementoDespesa%2CmodalidadeDespesa%2CvalorDespesaEmpenhada%2CvalorDespesaLiquidada%2CvalorDespesaPaga%2CvalorRestoPago&ordenarPor=mesAno&direcao=desc. Acesso em: 07 jun. 2023.

BRASIL. Controladoria Geral da União. Portal da transparência: detalhamento da despesa pública. **Gov.br**, 2021b. Disponível em: https://portaldatransparencia.gov.br/despesas/orgao/consulta?de=01/01/2021&ate=31/12/2021&orgaos=OR-INSIRA-AQUI-O-ORGAO-&ordenarPor=mesAno&direcao=desc. Acesso em: 07 jun. 2023.

BRASIL. Controladoria Geral da União. Portal da transparência: execução da despesa por órgão. **Gov.br**, 2023. Disponível em: https://portaldatransparencia.gov.br/despesas/orgao. Acesso em: 07 jun. 2023.

BRASIL. Decreto no 8.638 de 15 de janeiro de 2016. Institui a Política de Governança Digital no âmbito dos órgãos e das entidades da administração pública

federal direta, autárquica e fundacional. **Diário Oficial da União**, 18 jan. 2016. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8638.htm. Acesso em: 07 jun. 2023.

BRASIL. Governo Federal. Dados abertos. **Gov.br**, [2023a]. Disponível em: https://dados.gov.br. Acesso em: 07 jun. 2023.

BRASIL. Lei n. 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º , no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. **Diário Oficial da União**, 18 nov. 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 07 jun. 2023.

CÂMARA DOS DEPUTADOS. **Orçamento da União**: Lei Orçamentária Anual (LOA). 2022. Disponível em: https://www2.camara.leg.br/orcamento-da-uniao/leis-orcamentarias/loa/lei-orcamentaria-anual-loa. Acesso em: 07 jun. 2023.

CAPURRO, R.; HJORLAND, B.; CARDOSO, A. M. P.; TRAD., M. G. A. F.; AZEVEDO, M. A.; (TRAD.), A. M. P. C.; (TRAD.), M. G. A. F.; (TRAD.), M. A. A. O conceito de informação. **Perspectivas em Ciência da Informação**, v. 12, n. 1, 2007. Disponível em: http://hdl.handle.net/20.500.11959/brapci/33134. Acesso em: 7 jun. 2022.

CHAMBERS, J. M. **Graphical methods for data analysis**. Belmont, CA: Wadsworth, 1983.

CLEVELAND, W. S. **Visualizing data**. Oxfordshire, UK: Hobart Press, 1993.

CORREIA, M. E. R. O uso de indicadores de desempenho na gestão da informação em empresas de tecnologia da informação e comunicação. **Revista de Informação Contemporânea**, v. 5, n. 1, p. 25-40, 2017.

CORRUPTION Risk Forecast. [2023]. Disponível em: https://corruptionrisk.org/. Acesso em: 07 jun. 2023.

COSTA, Alexandre de Souza *et al*. O uso do estudo de caso na ciência da informação no Brasil. **InCID**: R. Ci. Inf. e Doc., Ribeirão Preto, v. 4, n. 1, p. 49-69, jan./jun. 2013. DOI: 10.11606/issn.2178-2075.v4i1p49-69. Disponível em: https://www.revistas.usp.br/incid/article/view/59101/62099. Acesso em: 07 jun. 2023.

DSPA. **What is CRISP DM?** jan 2013. Disponível em: https://www.datascience-pm.com/crisp-dm-2/. Acesso em: 06 maio 2023.

EAVES, D. Public data and open data. **Eaves.ca**, 2009. Disponível em: http://eaves.ca/2009/06/04/public-data-and-open-data/. Acesso em: 30 jan. 2023.

FOGL, J. Relation of the concepts ‘Information’ and ‘Knowledge’. **International Forum for Information and Documentation**, The Hague, v. 4, n. 1, p. 21-24, 1979.

GARRETT, J. J. **The Elements of User Experience**. New York: New Riders Publishing, 2003.

GARTNER. **Magic quadrant for data integration tools**. 2023. Disponível em: https://www.gartner.com/document/3993653. Acesso em: 04 abr. 2023.

GEBRE, E. H.; MORALES, E. How “accessible” is open data? Analysis of context-related information and users’ comments in open datasets. **Government Information Quarterly**, v. 37, n. 4, 2020.

GRÁCIO, Maria Cláudia Cabrini. Acoplamento bibliográfico e análise de cocitação: revisão teórico-conceitual. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, [*S. l.*], v. 21, n. 47, p. 82-99, 2016. DOI: 10.5007/1518-2924.2016v21n47p82. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2016v21n47p82. Acesso em: 7 jun. 2023.

GUIMARÃES, J. A. C.; PINHO, F. A. Desafios da representação do conhecimento: abordagem ética. **Informação & Informação**, v. 12, n. 1, p. 19-39, 2007. DOI: 10.5433/1981-8920.2007v12n1p19 Acesso em: 22 mai. 2023.

HARRISON, T.; SAYOGO, D. S. Information transparency in disaster management: an evaluation of information policy in Brazil using the Transparency Portal. **Government Information Quarterly**, v. 31, n. 3, p. 399-408, 2014.

HEISE, Arvid. Integrating open government data with stratosphere for more transparency. **Journal of Web Semantics**, v. 14, p. 45-46, jul., 2012.

JANSSEM, M. Data science empowering the public: data-driven dashboards for transparent and accountable decision-making in smart cities. **Journal of Urban Technology**, v. 27, n. 3, p. 59-80, 2020.

JANSSEM, Marijn; CHARALABIDIS, Yannis, ZUIDERWIJK, Anneke. Benefits, adoption barriers and myths of open data and open government. **Information Systems Management**, v. 29, p. 258-268, 2012. DOI: 10.1080/10580530.2012.716740. Disponível em: https://www.tandfonline.com/doi/epdf/10.1080/10580530.2012.716740?needAccess=true&role=button. Acesso em: 07 jun. 2023.

KALAMPOKIS, Evangelos; HAUSENBLAS, Michael; TARABANIS, Konstantinos. Combining social and government open data for participatory decision-making. **Electronic Participation**, p 36-47, 2011.

KALBACH, J. **Mapping Experiences**: A Complete Guide to Creating Value through Journeys, Blueprints, and Diagrams. Sebastopol, CA: O'Reilly Media, 2016.

KAPLAN, Robert S.; NORTON, David P. **The balanced scorecard**: translating strategy into action. Massachusetts: Harvard Business Review Press, 1996.

KOLTAY, T. Data literacy for researchers and data librarians. **Ciência da Informação**, v. 49, n. 1, 2016. Disponível em: https://doi.org/10.1177/0961000615616450. Acesso em: 07 jun. 2023.

KUBLER, Sylvain; ROBERT, Jérérmy; NEUMAIER, Sebastian; UMBRICH, Jürgen, LE TRAON, Yves. Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. **Government Information Quarterly**, v. 35, n. 1, p. 13-29, jan. 2018. DOI: https://doi.org/10.1016/j.giq.2017.11.003. Disponível em: https://www.sciencedirect.com/science/article/pii/S0740624X16301319. Acesso em: 07 jun. 2023.

LEVY, K.; JOHNS, D. **When open data is a Trojan Horse**: The weaponization of transparency in science and governance. 2016. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/2053951715621568. Acesso em: 25 set. 2023.

LUCIDCHART. **Lucid**, [2023]. Disponível em: https://lucid.app/lucidchart. Acesso em: 07 jun. 2023.

MACEDO, F. L. O. **Arquitetura da informação**: aspectos epistemológicos, científicos e práticos. 2005. 190 p. Dissertação (Mestrado). Departamento de Ciência da Informação e Documentação, Universidade de Brasília. Brasília. 2005.

MARIANO, Januário Albino; FERNEDA, Edberto. O campo da ciência da informação: contribuições, desafios e perspectivas. **Perspectivas em Ciência da Informação**, v. 20, n. 2, p. 3-18, abr./jun. 2015. DOI: http://dx.doi.org/10.1590/1981-5344/1932. Disponível em: https://www.scielo.br/j/pci/a/j68g9dXT7SxHFL4nMtdwJKz/?format=pdf&lang=pt. Acesso em: 27 out. 2023.

MEDEIROS, M. B. B.; CAFÉ, L. M. A. Organização da informação ou organização do conhecimento? *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo. **Anais** [...]. São Paulo: USP, 2008. p. 1-14. Disponível em: http://hdl.handle.net/20.500.11959/brapci/176535. Acesso em: 5 jan. 2023.

MEIJER, Albert. **Government Transparency in Historical Perspective**: From the Ancient Regime to Open Data in The Netherlands. 2015. Disponível em: https://dspace.library.uu.nl/bitstream/handle/1874/329767/Government.pdf?sequence=1&isAllowed=y . Acesso em: 15 set. 2023.

MICHAEL, P. Notter. **Advanced exploratory data analysis (EDA)**. 1 fev. 2020. Disponível em: https://miykael.github.io/blog/2022/advanced_eda/. Acesso em: 21 set. 2022.

NHACUONGUE, Ari Melo; SANTOS, Maíra Rocha. Revisão da literatura: apresentação de uma abordagem integradora. *In*: CONGRESSO INTERNACIONAL AEDEM, 26., 2017; AEDEM International Conference -Economy, Business and Uncertainty: ideas for a European and Mediterranean industrial policy?, 26. Calabria, Italy, 2017. **Anais** [...]. Calabria, 2017.

NYC OPEN DATA. **Open data for all New Yorkers**. 2021. Disponível em: https://opendata.cityofnewyork.us/. Acesso em: 07 jun. 2023.

PARK, S.; GIL-GARCIA, J. R. Open data innovation: Visualizations and process redesign as a way to bridge the transparency-accountability gap. **Government Information Quarterly**, v. 39, n. 1, 2022.

PARMENTER, David. **Key performance indicators**: developing, implementing, and using winning KPIs. Hoboken, NJ: Wiley, 2010.

PARMENTER, David. **Key performance indicators**: developing, implementing, and using winning KPIs. New Jersey: Wiley, 2010.

RAINER, R. Kelly; PRINCE, Brad; WATSON, Hugh J. **Management information systems**: moving business forward. Hoboken, NJ: Wiley, 2018.

RIBEIRO, F. Organização e uso da informação: conhecer bem para bem representar. **IRIS**: Revista de Informação, Memória e Tecnologia, v. 1, n. 1, p. 7-16, 2012. Disponível em: http://hdl.handle.net/20.500.11959/brapci/93394. Acesso em: 14 abr. 2023.

ROSENFELD, L.; MORVILLE. **Information Architecture for the World Wide Web**. 2. ed. Cambridge: O'Reilly, 2002. 461 p.

SAS. **ETL (Extract, Transform, Load)**: O que é ETL? 2021. Disponível em: https://www.sas.com/pt_br/insights/analytics/etl-extract-transform-load.html. Acesso em: 21 fev. 2023.

SEMIDAO, R. A. M. **A gestão do conhecimento no contexto de Business Intelligence**: estudo de caso em uma empresa de telecomunicações. 2014. Dissertação (Mestrado) - Universidade Estadual Paulista, 2014.

SERRA, L. **A essência do Business Intelligence**. São Paulo: Editora Berkely Brasil, 2002.

SILVA, Caleb Siloé Ben; WILDAUER, Egon Walter. **Modelo de governança de dados em uma plataforma de pagamentos digitais**. 2022. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BJB/article/view/55448. Acesso em: 1 out. 2023.

SIVARAJAH, U. *et al*. The role of e-participation and open data in evidence-based policy decision making in local government. **Journal of Organizational Computing and Electronic Commerce**, v. 16, n.1-2, p. 64-79, 2016. DOI: https://doi.org/10.1080/10919392.2015.1125171. 2016. Disponível em: https://www.tandfonline.com/doi/epdf/10.1080/10919392.2015.1125171?needAccess=true&role=button. Acesso em: 07 jun. 2023.

SOUZA, E. D.; DIAS, E. J. W. A integração disciplinar na ciência da informação: os não-ditos sobre essa familiar desconhecida. **Ciência da Informação**, v. 40, n. 1,

2019. Disponível em: https://brapci.inf.br/index.php/res/v/18835. Acesso em: 07 jun. 2022

SVENONIUS, E. **The intellectual foundation of information organization**. Cambridge, MA: MIT Press, 2000.

WERSIG, G. Information science: the study of postmodern knowledge usage. **Information Processing and Management**: an International Journal, Tarrytown-Nova Iorque, v. 29, n. 2, p. 229-239, Mar./Apr. 1993.

TENOPIR, C. *et al*. **Data sharing, management, use, and reuse**: practices and perceptions of scientists worldwide. 11 mar. 2020. Disponível em: https://doi.org/10.1371/journal.pone.0229003.

TOTVS. **BPMN**: entenda o que é a modelagem de processos de negócios, como fazer e sua importância! Brasília: TCU, 2015. Disponível em: https://www.totvs.com/blog/gestao-industrial/bpmn/. Acesso em: 26 set. 2022.

TRIBUNAL DE CONTAS DA UNIÃO. Secretaria de Fiscalização de Tecnologia da Informação. **5 motivos para a abertura de dados na Administração Pública**. Brasília: TCU, 2015. Disponível em: https://portal.tcu.gov.br/lumis/portal/file/fileDownload.jsp?fileId=8A8182A24F0A728E014F0B36E7016F34. Acesso em: 21 set. 2022.

TUKEY, J. W. **Exploratory data analysis**. Reading, MA: Addison-Wesley, 1977.

VICTORINO, M. C. **Organização da informação para dar suporte à arquitetura orientada a serviços**: reuso da informação nas organizações. 2011. 276 p. Dissertação (Doutorado). Departamento de Ciência da Informação e Documentação, Universidade de Brasília. Brasília. 2011.

VILA, R. A.; ESTEVEZ, E.; FILLOTTRANI, P. R. The design and use of dashboards for driving decision-making in the public sector. **Government Information Quarterly**, v. 35, n. 1, p. 85-96, 2018.

VILLENEUVE, C. Cybersecurity and cyberwar: what everyone needs to know. Oxford: **Oxford University Press**, 2014.

WEERAKKODY, Vishanth *et al*. Open data and its usability: an empirical view from the Citizen's perspective. **Information Systems Frontiers**, v. 19, p. 285–300, 2017. DOI: https://doi.org/10.1007/s10796-016-9679-1. Disponível em: https://link.springer.com/article/10.1007/s10796-016-9679-1. Acesso em: 07 jun. 2023.

WORDCLOUDS. **Free online Wordcloud generator**. [2023]. Disponível em: https://www.wordclouds.com/. Acesso em: 07 jun. 2023.

WORLD WIDE WEB CONSORTIUM. **W3C:** The World Wide Web Consortium. 2023. Disponível em: https://www.w3.org/. Acesso em: 2 mar. 2022.

WURMAN, Richard Saul. **Information Architects**. Zurich: Switzerland: Graphis Press; 1996. Disponível em: https://www.amazon.com/Information-Architects-Richard-Saul-Wurman/dp/1888001380.

ZANON, Sandra Buth. **Management and electronic information security**: requirements for effective document management in Brazil. 2015. Disponível em: http://biblios.pitt.edu/ojs/biblios/article/view/185. Acesso em: 1 set. 2023.

ZIMAN, J. M. **Public Knowledge**: An Essay Concerning the Social Dimension of Science. London: Cambridge University Press, 1968.