CRIAÇÃO DE UM BANCO DE DADOS PARA CIDADES

Daniel Luis Notari, Eduardo Eberhardt Pereira, Stefani Caprioli Gonçalves, Ana Paula Fermiano, Ana Cristina Fachinelli

RESUMO

Este artigo aborda a criação e atualização de um banco de dados de indicadores para cidades, utilizando dados abertos de todas as cidades brasileiras. O objetivo é fornecer uma base de dados que permita a análise e comparação de indicadores usando diferentes modelos. A pesquisa destaca a importância do Desenvolvimento Baseado em Conhecimento (DBC) e dos Sistemas de Capitais (SC) para promover cidades do conhecimento, onde o uso de tecnologia e dados promove a sustentabilidade e qualidade de vida visando tornar as cidades inteligentes. O processo de atualização do banco de dados envolveu a coleta, extração, transformação e carga (ETL) de dados de fontes diversas, enfrentando desafios como a diversidade de formatações e a necessidade de normalização dos dados. Foram utilizados dados de diversas fontes de dados abertos. O banco de dados será utilizado para atualizar o Observatório de Cidades e a Plataforma Web de Indicadores, desenvolvidas pela equipe do City Living Lab da Universidade de Caxias do Sul. Os resultados demonstram a eficácia do uso de técnicas de ETL para consolidar grandes volumes de dados e criar um sistema robusto para suporte à decisão de gestores públicos e pesquisadores.

Palavras-chave: Dados Abertos, ETL, Cidades Inteligentes, Desenvolvimento Sustentável, Sistema de Capitais.

1 INTRODUÇÃO

Em 2021 cerca de 56% da população mundial vivia em áreas urbanas. E segundo projeções da Organização das Nações Unidas (ONU) até 2050 este percentual será de 68%. Esta projeção ressalta ainda que uma em cada três pessoas viverá em cidades com pelo menos 500 mil habitantes (HABITAT, 2022). Entender essa dinâmica é crucial para um desenvolvimento sustentável. A superlotação das cidades pode extrapolar limites econômicos, psicológicos e socioculturais, ocasionando aumento em índices de violência, pobreza e até mesmo maior vulnerabilidade a pandemias (ERGAZAKIS; METAXIOTIS; PSARRAS, 2004).

A aglomeração urbana representa um desafio significativo para as autoridades públicas, exigindo eficácia e rapidez no atendimento das necessidades sociais. Considerando ainda as restrições econômicas e legais para a destinação de recursos públicos e a intensa disputa entre as cidades para atração de investimentos, o cenário tende a ser ainda mais desafiador (WEISS; BERNARDES; CONSONI, 2015).

Uma cidade inteligente pode auxiliar no desenvolvimento das cidades tornando-as mais limpas, seguras e funcionais (BRIA; MOROZOV, 2020). Uma cidade inteligente utiliza de

tecnologia com o objetivo de otimizar o uso de recursos, produzir novos recursos, modificar o comportamento dos usuários e promover a sustentabilidade (BRIA; MOROZOV, 2020).

É comum pensar em uma cidade inteligente associando sensores, dispositivos responsivos ou microcomputadores, porém um dos recursos que estas dispõem é o Desenvolvimento Baseado em Conhecimento (DBC), como uma nova forma de fazer a gestão pública para enfrentar estes desafios (CARRILLO, 2002). O DBC possibilita a criação de uma cultura econômica que inclui as dimensões social, econômica e ambiental. Outra forma de auxiliar nisso é a abordagem de Sistemas de Capitais (SC), proposto por Carrillo (2002), que tem como objetivo transacionar o sistema de valor, baseado em produção material, para um sistema com foco na produção de conhecimento.

No artigo de Fachinelli, Carrillo e D'Arisbo (2014) é possível verificar o uso do SC como meio para identificar uma possível Cidade do Conhecimento (CC). Neste estudo foi definido categorias de capitais, isto é, os indicadores a serem utilizados. Na etapa posterior os dados foram obtidos de diferentes fontes de informação e adicionados a uma planilha eletrônica para serem analisados, tudo isso de forma manual. Aplicando os conceitos de DBC pode-se identificar uma CC, que é definida como "áreas urbanas que têm seu desenvolvimento centrado no conhecimento, onde a análise socioeconômica dos elementos e estratégias de gestão do conhecimento se concentra na avaliação de um sistema de valores fundamentado na criação, compartilhamento e aplicação de valor" (CARRILLO, 2006).

As CC são estabelecidas com objetivo de alcançar a sustentabilidade e a melhoria da qualidade de vida, fornece os serviços necessários, enriquecer a cultura e o conhecimento, e aumentar as competências da população (YIGITCANLAR; O'CONNOR; WESTERMAN, 2008).

Outra forma de olhar para as cidades é através dos Objetivos de Desenvolvimento Sustentável (ODS), adotados pelos países-membros das Nações Unidas, que traçam um caminho em direção à sustentabilidade, estabelecendo a meta de melhorar as condições de vida de toda a população do planeta até 2030 (ONU, 2016).

Além de aplicar o DBC e ter um retorno de dados, é igualmente importante fazer uma análise destas informações. Por isso, é fundamental o uso de indicadores confiáveis, que possam ser comparados e servir para embasar tomada de decisões, seja de empresários, gestores ou administradores públicos.

Exemplos do uso de dados abertos para perceber como as cidades podem ser verificadas quanto aos modelos citados são o Observatório de Cidades do Corede Serra (Fachinelli et. al, 2023) e a Plataforma Web de Indicadores (Notari et. al, 2020) desenvolvidos pelos pesquisadores do City Living Lab[1] da Universidade de Caxias do Sul. Os projetos citados fizeram uso de um banco de dados criado a partir da coleta de dados abertos entre os anos de 2017 e 2022. Dessa forma, o problema de pesquisa aborda a necessidade de atualizar esse banco de dados a partir de uma nova coleta de dados.

[1] www.citylivinglab.com

O objetivo deste artigo é descrever o processo de atualização do banco de dados de indicadores do City Living Lab a partir de dados abertos de todas as cidades brasileiras visando uma futura atualização dos projetos desenvolvidos.

O presente artigo está organizado com o referencial teórico e trabalhos relacionados sendo apresentado na seção 2; os procedimentos metodológicos na seção 3; os resultados e discussões na seção 4; e as considerações finais na seção 5.

2 REFERENCIAL TEÓRICO

Esta seção apresenta os conceitos sobre dados abertos, sobre o processo de extração, transformação e carga de dados em banco de dados, bem como descreve como trabalhos relacionados os projetos citados na Introdução do City Living Lab.

2.1 DADOS ABERTOS

O conceito de dado aberto, conforme definido por Open Knowledge Foundation (2024), especifica que os dados disponibilizados devem ser acessíveis, utilizáveis, modificáveis e compartilháveis livremente pelos cidadãos. No entanto, esses dados devem seguir requisitos que preservem sua procedência e garantam sua abertura. Para a distribuição dessas informações, são necessários alguns critérios:

a) Devem ser de domínio público ou possuir uma licença aberta;
b) Devem ser acessíveis pela internet sem custo;
c) Devem ser disponibilizados em formatos processáveis por computadores, permitindo que elementos individuais dos dados sejam acessados e modificados facilmente;
   Alguns exemplos de formatos incluem CSV (Comma-separated values), XML (Extensible Markup Language), JSON (JavaScript Object Notation) e RDF (Resource Description Framework); e,
d) Devem estar em formatos abertos, sem restrições financeiras ou de uso, e deve ser possível processá-los por ferramentas de software livre.

Quanto aos dados conectados, estes descrevem um conjunto de boas práticas voltadas para a publicação e interconexão de dados na web. Essas práticas utilizam tecnologias para representação de dados estruturados, visando facilitar o acesso e a utilização dos dados tanto por seres humanos quanto por máquinas (BANDEIRA et al., 2015). A adoção dessas boas práticas contribui para a criação de um ecossistema de dados mais acessível e integrado. As principais práticas incluem:

a) Empregar URI (Uniform Resource Identifier) como identificadores para entidades e conceitos;
b) Utilizar HTTP (Hypertext Transfer Protocol) URI para que as entidades e/ou conceitos possam ser encontrados com maior facilidade.
c) Adotar padrões estabelecidos pela W3C (World Wide Web Consortium).
d) Incorporar links para outros URI relacionados, expandindo o conhecimento disponível sobre os temas abordados.

Adicionalmente, foi desenvolvida a classificação chamada "Cinco estrelas dos Dados Abertos" (5 Stars Open Data, 2012), ilustrada no Quadro 1. Esta classificação ajuda a compreender os níveis de legibilidade, estrutura, disponibilidade e a facilidade com que as informações podem ser interpretadas por máquinas.

Quadro 1 – Classificação dos dados abertos

Classificação	Descrição
1	O dado está disponível na internet em vários formatos
2	O dado é legível por máquinas e está estruturado
3	O dado é oferecido em formato aberto (CSV)
4	O dado utiliza utiliza padrões abertos da W3C como RDF para a criação de identificadores únicos
5	Possui todas as características anteriores, incluindo conexões com outras fontes de dados

Fonte: O Autor (2024).

Com base nas definições anteriores, os dados abertos conectados representam a integração dos conceitos de dados abertos e dados conectados. Assim, os dados abertos conectados seguem os mesmos princípios discutidos nesta seção. A utilização dessas boas práticas, tanto na publicação quanto no consumo dessas informações, facilita significativamente a exploração de dados governamentais abertos. Além de aumentar a transparência nas transações governamentais, esta abordagem garante que os dados sejam acessíveis e legíveis por máquinas, o que facilita sua integração e processamento automatizado (BANDEIRA et al., 2015).

2.2 EXTRAÇÃO, TRANSFORMAÇÃO E CARGA DE DADOS (ETL)

O processo de carga de dados em um sistema como o de cidades do conhecimento é uma etapa crucial, pois a obtenção e extração dos dados de uma ou mais fontes podem ser trabalhosas e demoradas, como observado por Silva (2022). O processo ETL (Extract, Transform, Load), definido por Bansal e Kagemann (2015), desempenha um papel fundamental nessa atividade, envolvendo a extração dos dados das fontes originais, a transformação para atender às necessidades operacionais e, por fim, o carregamento dos dados no banco de dados de destino.

O modelo ETL possui características como simplicidade, completude e alta customização, conforme afirmado por El-Sappagh, Hendawi e El Bastawissy (2011). Suas etapas são funcionais e consecutivas, seguindo a ordem de extração, transformação e carga dos dados.

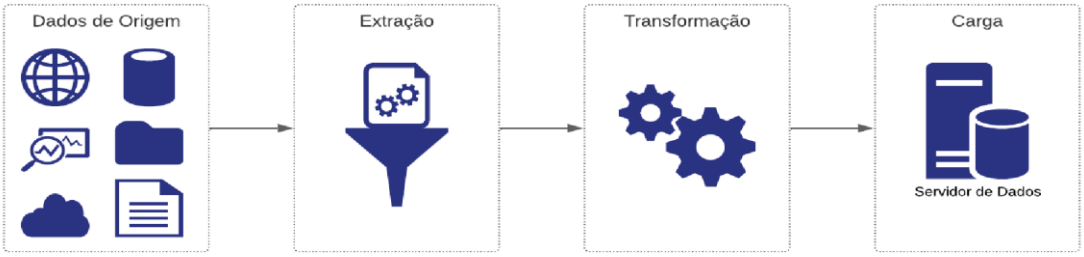
Na etapa de extração os dados são obtidos das fontes de dados originais, que podem incluir bancos de dados operacionais, sistemas legados, arquivos de texto, planilhas, entre outros. Durante a extração, é importante considerar as características específicas de cada fonte de dados e garantir a integração eficaz entre diferentes plataformas e sistemas. Existem duas fases principais na extração de dados: a extração inicial, que é realizada uma única vez para preencher o *data warehouse* com uma grande quantidade de dados, e a extração de dados modificados, que ocorre periodicamente para capturar apenas os dados que foram modificados ou adicionados desde a última extração.

A etapa de transformação é responsável por limpar, conformar e manipular os dados extraídos, preparando-os para serem carregados no sistema. Nessa etapa, são realizadas

atividades como limpeza de dados, conformação de formatos, conversão de unidades, aplicação de regras de negócios e integração de dados. Também podem ocorrer processos de enriquecimento de dados, adicionando informações extras, e criação de estruturas dimensionais, especificando a granularidade e hierarquia dos dados.

A etapa de carga envolve a inserção dos dados transformados nas estruturas dimensionais que serão acessadas pelos usuários finais e sistemas de aplicação. Existem dois tipos principais de carga: carga de tabelas de dimensões e carga de tabelas de fatos. Na carga de tabelas de dimensões, os dados das dimensões são carregados ou atualizados nas tabelas de dimensões do banco de destino, envolvendo a verificação de registros existentes e a inserção de novos registros. Na carga de tabelas de fatos, os dados dos fatos são carregados nas tabelas de fatos, podendo incluir agregação dos dados em diferentes níveis de granularidade, dependendo das necessidades de relatórios e análises. Durante a carga, é essencial garantir a integridade referencial, a consistência dos dados e o desempenho adequado do banco.

Figura 1 – Processo ETL



Fonte: O Autor (2024).

A Figura 1 representa graficamente as três etapas fundamentais do processo ETL. No contexto da extração, é possível observar a origem dos dados, incluindo bancos de dados operacionais, sistemas legados, arquivos de texto e planilhas. A transformação é apresentada como o estágio intermediário, onde os dados são processados, limpos e adaptados conforme necessário. Por fim, a carga demonstra a inserção dos dados processados no banco de dados de destino, para armazenamento e posterior análise. Essas etapas formam o núcleo do processo ETL e são essenciais para a construção e manutenção de um *data warehouse* eficiente e confiável, possibilitando a obtenção e utilização dos dados de forma organizada e útil para a tomada de decisões.

2.3 TRABALHOS RELACIONADOS

2.3.1 Aplicação Web Indicadores

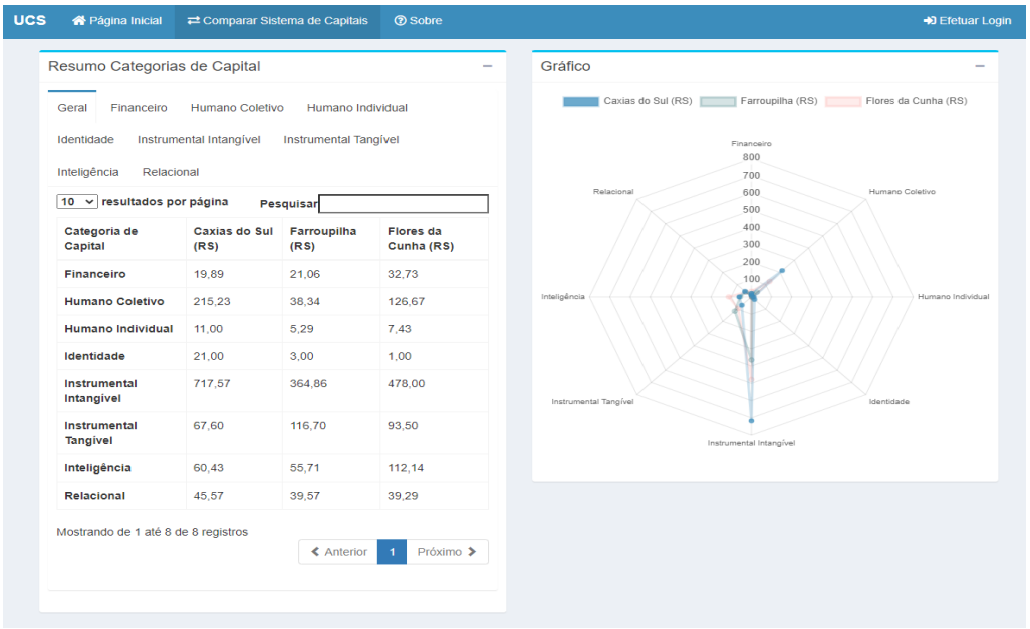
O projeto de pesquisa intitulado "Sistema de Capitais para Cidades do Conhecimento no Brasil: Um Modelo para o Desenvolvimento Baseado em Conhecimento" estudou o desenvolvimento de cidades, utilizando dados como PIB e IDH. Essa pesquisa representou a primeira aplicação completa desse método no Brasil, com seu campo de estudo situado na

cidade de Bento Gonçalves, no Rio Grande do Sul (FACHINELLI; CARRILLO; D'ARISBO, 2014).

A primeira versão do projeto utilizou uma planilha eletrônica para a criação do banco de dados, processamento e geração dos resultados. Devido a relevância da utilização de sistemas de desenvolvimento baseado em conhecimento (CARRILLO, 2002; CARRILLO; 2006) baseados em valor para cidades (CARRILLO, 2014; CARRILLO, 2015), e com o objetivo de automatizar o armazenamento, manipulação e apresentação de dados obtidos por meio de sistemas de valores, Battistelo (2018) e Notari et al (2020) propuseram a criação de uma plataforma *web*. A plataforma já conta com o cadastro de todas as cidades, e os dados da planilha eletrônica foram devidamente importados. Com isso, é possível comparar as cidades utilizando as dimensões e indicadores do sistema de capitais.

A Figura 2 (Notari et. al., 2020) apresenta o resultado usando o gráfico de radar das cidades de Caxias do Sul, Farroupilha e Flores da Cunha, onde cada cidade pode verificar o seu posicionamento nas dimensões do Sistema de Capitais usando os dados organizados em 2014. As dimensões envolvidas contêm Financeiro, Humano Coletivo, Humano Individual, Identidade, Instrumental Intangível, Instrumental Tangível, Inteligência e Relacional. Cada dimensão, acompanhada de seus indicadores e valores correspondentes, é apresentada.

Figura 2 – Plataforma Web Sistema de Capitais - Resultado



Fonte: (Notari et. al, 2020)

2.3.2 Observatório Cidade do Conhecimento

O desenvolvimento de cidades sustentáveis tem na coleta de dados, e sua devida análise, uma etapa imprescindível para o alcance dos seus objetivos. A confiabilidade dos dados é fundamental para a tomada de decisão por parte dos gestores públicos e demais atores do desenvolvimento regional.

O objetivo deste projeto foi o de promover o desenvolvimento da região do COREDE Serra através da implementação de uma plataforma de dados dinâmicos com informações coletadas na região, o que chamamos de Observatório de Cidades (Fachinelli et. Al, 2023).

Na plataforma do Observatório de Cidades do Corede Serra, as cidades podem ser analisadas por meio de seus indicadores, os quais são agrupados de acordo com 3 modelos: o Sistema de Capitais que define uma Cidade do Conhecimento; os Objetivos de Desenvolvimento Sustentável (ONU) e as dimensões de Cidades Inteligentes. Além disso, o Observatório de Cidades apresenta dados da percepção da população da região sobre as dimensões que caracterizam cada um dos modelos.

Os dados disponíveis na plataforma podem servir tanto como instrumento para os gestores das cidades, como para os cidadãos conhecerem dados sobre suas cidades de maneira visual, amigável, responsiva e cujas informações são abordadas em diferentes dimensões. O presente relatório apresenta uma síntese desses dados com o intuito de proporcionar subsídios para o desenvolvimento de políticas públicas para inovação, planejamento estratégico, elaboração de planos diretores, visão de futuro e especializações estratégicas (Fachinelli et. al, 2023).

Figura 3 – Dashboard Observatório de Cidades

Fonte: (Fachinelli et. Al, 2023).

3 PROCEDIMENTOS METODOLÓGICOS

O estudo adota um método centrado na coleta, análise e interpretação de dados numéricos para retratar a situação atual das cidades. O objetivo principal é construir um banco de dados com um conjunto de indicadores de todas as cidades brasileiras. Nas próximas seções é explicado como foi aplicado o processo baseado no modelo ETL, seguindo suas etapas para desenvolver o modelo de dados que atualiza o banco de dados existente. A Figura 4 apresenta o desenho metodológico de construção do banco de dados.

Figura 4 – Processo de ETL para construção do banco de dados

Fonte: O Autor.

3.1 EXTRAÇÃO

A primeira etapa realizada foi a extração dos dados. Inicialmente, foram verificadas as fontes utilizadas na versão anterior, validando se os *links* permaneciam ativos e se os dados ainda estavam disponíveis. Nos casos em que a fonte original não atendia mais às necessidades, novas fontes foram identificadas para suprir os respectivos indicadores. Para manter o controle dessas informações, foi criada uma planilha eletrônica contendo todos os indicadores, suas respectivas fontes e o status no desenvolvimento. O Quadro 1 apresenta uma parte dessa planilha. Por exemplo, as quatro primeiras linhas referem-se aos identificadores e informações do município. Depois segue com as informações dos indicadores como densidade demográfica, despesas, entre outros.

A proposta inicial era de automatizar o *download* dos dados reunidos na planilha utilizando uma plataforma de *web scraping* (Broucke & Baesens, 2018). No entanto, o modo de operação diverso e irregular entre as diferentes páginas fez da ferramenta ineficaz, tornando o *download* manual uma opção mais viável para a extração dos dados. A irregularidade mencionada decorre do fato de que cada fonte possui uma forma distinta de disponibilizar os dados.

O processo manual, no entanto, trouxe complicações devido ao volume de trabalho. No projeto, havia a intenção de realizar uma série histórica, isto é, reunir dados de diversos anos para cada indicador, conforme disponibilidade. Devido a isto, cada indicador criou a demanda de um grande volume de dados que precisou ser baixado de forma manual.

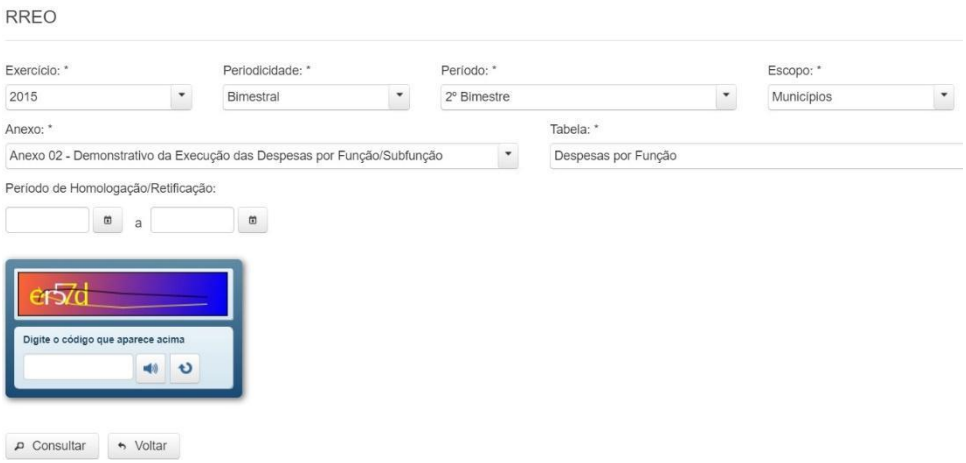
Além das dificuldades relacionadas à obtenção dos dados, o formato em que as fontes disponibilizavam os arquivos também exigiu um esforço adicional. Havia divergências significativas nas formatações, tanto entre diferentes fontes quanto entre arquivos de mesma fonte em anos distintos. Essas variações incluíam o tamanho dos cabeçalhos e rodapés, a extensão dos arquivos (CSV, XLS, entre outros) e a codificação fornecida pelos dados (UTF -8, UTF -16, ANSI, ISO-8859).

Quadro 1 – Planilha de controle dos indicadores

Indicadores	Cálculo feito sobre os dados originais	Link site	Link diteto
Código de 6 dígitos	Nenhum		ND
Código de 7 dígitos	Nenhum		ND
Município	Nenhum		ND
Sigla	Nenhum		ND
Densidade demográfica	Habitantes / Área		ND
IFDM	Nenhum		Na tabela ao lado
Despesas municipais com cultura	Nenhum (Já vem per capita)		ND: Ver filtro na imagem ao lado
Vínculos ativos	Nenhum		Id: Básico Senha: 12345679
Proporção de trabalhadores formais para cada cem habitantes	Número de vínculos ativos / População		Id: Básico Senha: 12345679
Importação	Valor / População		ND
Exportação	Valor / População		ND
Saldo de empregos por cem vínculos ativos	Saldo * 100 / Vínculos ativos		Usuário: basico Senha: 12345678
Despesas municipais com planejamento e orçamento	Nenhum (Já vem per capita)		ND: Ver filtro na imagem ao lado
Densidade de Telefonia Fixa	Nenhum		
Densidade de TV por Assinatura	Nenhum		

Fonte: O Autor (2024).

Figura 5 – Extração dos dados



Fonte: O Autor (2024).

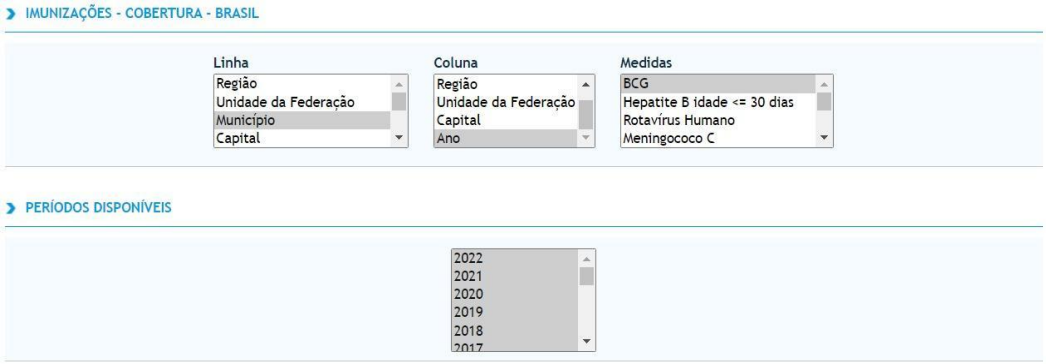
A Figura 5 apresenta uma consulta que deve ser feita para baixar os indicadores do SISCONFI[2]. Esse portal de dados aberto disponibiliza dados contábeis aplicadas ao setor público brasileiro. Nessa consulta deve-se selecionar um ano, o período, o escopo, o anexo e a Tabela, além de preencher o *captcha*. É visível que não tem mecanismos diretos de baixar um arquivo de dados sem configurar as informações explanadas.

A Figura 6 apresenta uma consulta que deve ser feita para baixar os indicadores Do Ministério da Saúde através do banco de dados DATASUS[3]. Nessa consulta deve-se selecionar

[2] https://siconfi.tesouro.gov.br/
[3] https://datasus.saude.gov.br/

informações diversas para ser montada uma planilha eletrônica, como por exemplo selecionar o município, o ano e o dado a ser consultado. Semelhante ao exemplo anterior não há mecanismos diretos de baixar um arquivo de dados sem configurar as informações explanadas.

Figura 6 – Extração dos dados



Fonte: O Autor (2024).

3.2 TRANSFORMAÇÃO

Para a transformação dos dados, foi utilizada a linguagem de Python[4]. A biblioteca Pandas[5] foi a ferramenta utilizada para a limpeza dos dados, permitindo o processamento em massa de diferentes tipos de dados e planilhas. Assim, foi possível carregar os arquivos extraídos para o ambiente de desenvolvimento e manipular o conteúdo bruto fornecido pelas entidades, qualificando e estruturando os dados.

O processo de tratamento dos dados apresentou algumas dificuldades, uma vez que as fontes eram variadas e, consequentemente, o tratamento necessário também. Cada conjunto de arquivos precisou ser analisado individualmente, e o procedimento para torná-los operacionais foi adaptado a cada caso. Além disso, a criação de uma série histórica trouxe uma complexidade adicional, visto que as fontes de dados frequentemente alteram o seu padrão de distribuição dos dados de um ano para o outro. Por exemplo, um estudo que abrange o período de 1994 a 2023 pode apresentar até quatro formatações diferentes para os arquivos anuais.

Além do Pandas, a biblioteca Numpy[6] foi usada para realizar cálculos matemáticos. A biblioteca SYS[7] permitiu o acesso a informações do sistema operacional, uma vez que os procedimentos para manipulação de uma grande quantidade de arquivos tiveram de ser feitos através de linha de comando do sistema operacional Windows.

Além de adequar os dados a uma estrutura mais simples e condizente com o indicador proposto, foi necessário colocá-los em uma mesma escala. Por exemplo, o número de ônibus em uma cidade, isoladamente, pode não oferecer muito significado, mas em uma escala comparativa, onde se mostra o desenvolvimento em relação a outras cidades, esse valor ganha relevância. Para isso, os dados passaram por um processo de interpolação ou normalização, onde o maior valor foi estabelecido como 1 e o menor como 0. Os demais valores foram convertidos em números entre 0 e 1, representando sua posição relativa no comparativo.

[4] https://www.python.org/
[5] https://pandas.pydata.org/
[6] https://numpy.org/
[7] https://docs.python.org/3/library/sys.html

3.3 CARGA

Após o tratamento dos dados para adequá-los ao projeto, eles foram carregados em um Sistema Gerenciador de Banco de Dados Relacional (SGBDR) PostgreSQL[8]. A disponibilidade dos dados em um SGBDR torna as consultas muito mais eficazes, pois o gerenciamento das informações é centralizado (Elmasri & Navathe, 2018). Durante o processo de transformação, os dados tratados foram armazenados em formato CSV visando simplificar a importação das planilhas para o SGBDR e a conversão para o formato de tabelas de banco de dados (Elmasri & Navathe, 2018).

Os dados foram inseridos no banco de dados através de um código criado na linguagem SQL (Structured Query Language) utiliza para manipular os dados em SGBDR (Elmasri & Navathe, 2018). Através de SQL, as tabelas foram inicialmente geradas de forma desestruturada, e em seguida foram criadas estruturas mais sólidas de relacionamento entre as tabelas.

Além disso, novas tabelas foram adicionadas para representar o escopo de cada dado dentro dos indicadores definidos. Cada dado estava relacionado a um indicador, cada indicador a algumas dimensões, e cada dimensão a um modelo (ou norma) conforme pode ser visto na Figura 8. Essa representação permite a adição de novos modelos de indicadores ao sistema, como novas inserções (linhas) no banco de dados existente, aproveitando toda a estrutura já estabelecida.

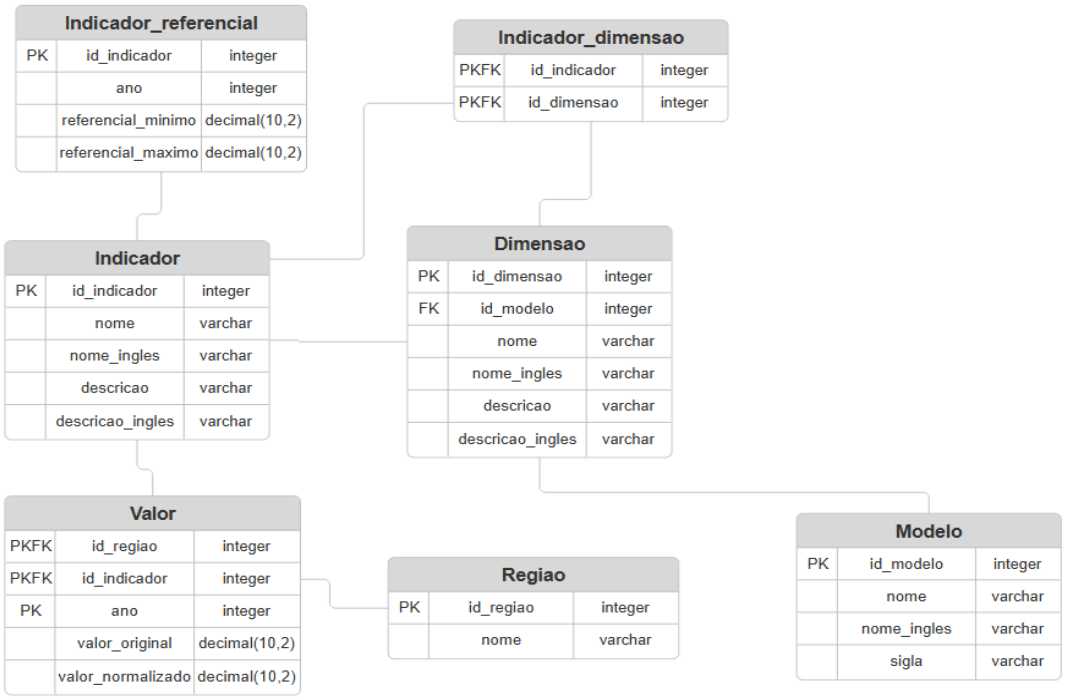
A partir da base de dados criada, o modelo relacional (Heuser, 2009) foi gerado com o intuito de permitir a visualização da estrutura das tabelas no SGBDR. Devido ao modelo utilizado, com diversas tabelas para representar os indicadores, uma generalização foi criada para compor as tabelas com dados referentes aos indicadores. O diagrama não demonstra as especializações dessa generalização, isto é, não demonstra quais entidades fazem parte daquela generalização. Como há um total de 34 tabelas representadas nessa entidade, colocar o modelo completo fica inviável para visualização.

Na Figura 7 pode-se verificar que os municípios estão representados na tabela Região, que se liga com as tabelas de Valores. Essa tabela possui qual indicador está referenciando em cada instância.

A Figura 8 apresenta uma tela administrativo do SGBDR PostgreSQL. No lado esquerdo é apresentado uma lista de tabelas criadas após a carga dos dados. A direita é apresentada o comando de criação da tabela para o indicador de *vínculos ativos e seu saldo*, logo abaixo é listado alguns dos dados inseridos.

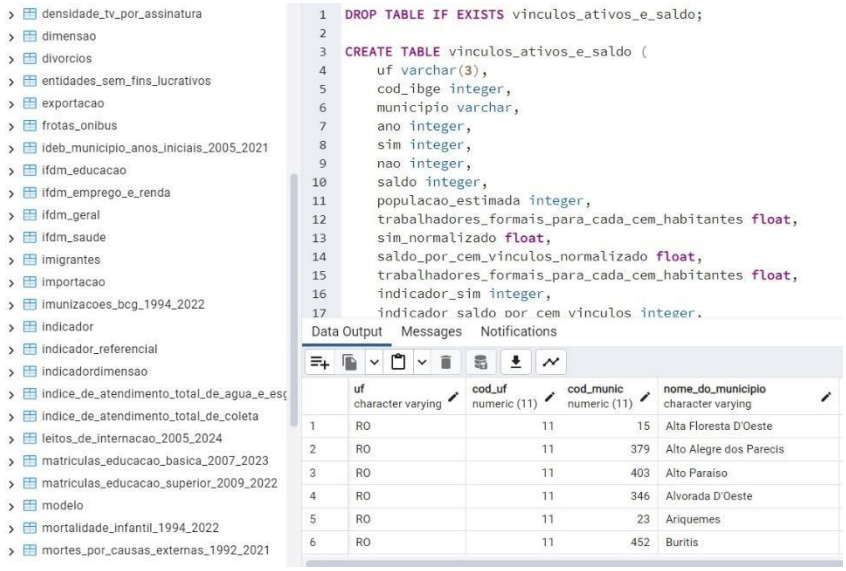
[8] https://www.postgresql.org/

Figura 7 – Modelo Relacional



Fonte: O Autor (2024).

Figura 8 – Extração dos dados



Fonte: O Autor (2024).

4 RESULTADOS E DISCUSSÕES

A criação do banco de dados usando os dados abertos das cidades brasileiras possibilitou a compilação de diversas informações provenientes de diferentes fontes em um único conjunto de dados, capaz de concentrar um conjunto de dados de nossas. O processo de coleta quanto e carregamento de dados foi desafiador, envolvendo questões como a quantidade massiva de dados, a diversidade de fontes, as diferentes formatações dos arquivos, e a criação de séries

históricas. Ainda assim, resultou em um banco de dados que oferece maior objetividade na visualização dessas informações.

A Tabela 1 resume os resultados do Processo de ETL realizado. Foram definidos um total de 66 indicadores provenientes do processamento de 546 arquivos diferentes. Esses arquivos totalizavam 14.139 colunas extraídas das planilhas eletrônicas em formato CSV. Após o processo de transformação, os dados foram organizados em 34 tabelas com um total 273 colunas, onde cada coluna representa um dado.

Tabela 1 – Resultados do Processo ETL

Número de indicadores	66
Número de Estudos (Fontes)	24
Número de Arquivos de Entrada	546
Número de Colunas de Entrada	14139
Número de tabelas geradas	34
Número de colunas geradas	273

Fonte: O Autor (2024).

O conjunto total extraído foi de 27 GB em arquivos, que, após tratamento, foram reduzidos para 180 MB em arquivos CSV e 123 MB em um arquivo .SQL no PostgreSQL. A redução final para 0,6% do tamanho original melhorou a eficiência de armazenamento e consultas.

Além de uma visão geral, é possível observar a importância do tratamento de dados em casos específicos. Os dados da plataforma do Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro (SICONFI), destacam-se por sua grande disparidade entre o tamanho necessário e o tamanho efetivo após o tratamento. Muitos indicadores de cidades estão relacionados aos gastos municipais em diferentes áreas, e os dados de despesa por função do tesouro atenderam a todos esses indicadores.

OS dados do SICONFI representam 23 indicadores (aproximadamente ⅓ do total), e totalizam 55 arquivos com dados bimestrais, do ano de 2015 até o primeiro bimestre de 2024. Cada arquivo possui oito colunas, totalizando 440 colunas utilizadas para representar os dados, que foram reduzidas para uma única tabela com nove colunas, incluindo colunas extras que representam os valores normalizados e a ligação com a tabela de indicadores. A Tabela 2 permite uma visualização mais detalhada dessas estatísticas.

O armazenamento é a questão mais significativa para os dados do SICONFI. Sozinhos, esses dados, em sua forma original, ocupavam 20GB. Após o processamento com as técnicas de ETL, as mesmas informações foram agrupadas por cidade e ano, e representadas em apenas 80MB.

Mesmo quando um estudo é bem estruturado e seus dados são disponibilizados de forma eficiente, o tratamento ainda pode melhorar a estrutura para casos específicos. No caso da Relação Anual de Informações Sociais (RAIS), os dados foram redistribuídos, permitindo que

várias colunas fossem consolidadas em uma só, aumentando o número de registros e facilitando o uso dessas informações em projetos futuros.

Tabela 2 – Resultados do Processo ETL nos dados do SICONFI

Número de indicadores	23
Número de Estudos (Fontes)	1
Número de Arquivos de Entrada	55
Número de Colunas de Entrada	440
Número de tabelas geradas	1
Número de colunas geradas	9

Fonte: O Autor (2024).

A RAIS é um relatório de informações socioeconômicas solicitado pelo Ministério do Trabalho e Emprego brasileiro às pessoas jurídicas e outros empregadores anualmente, e permite informações como o total de vínculos empregatícios ativos, distribuição de gênero, entre outras informações que mostram lados essenciais de qualquer cidade. A Tabela 3 demonstra como, mesmo quando a melhoria não é imediatamente evidente, ela pode ser significativa.

Tabela 3 – Resultados do Processo ETL nos dados do RAIS

Número de indicadores	5
Número de estudos (Fontes)	1
Número de arquivos de entrada	4
Número de colunas de entrada	205
Número de tabelas geradas	3
Número de colunas geradas	34

Fonte: O Autor (2024).

Os dados RAIS representam cinco indicadores, e foram necessários quatro arquivos para reunir todas as informações. Esses quatro arquivos continham 205 colunas no total, que, utilizando técnicas de colunas pivô, foram transformadas em três tabelas com 34 colunas no total. Curiosamente, o caso dos dados do RAIS é um dos poucos em que a versão final ficou maior que a original. Os quatro arquivos somavam 4 MB, enquanto os três arquivos finais que geraram as tabelas somaram 11 MB. Fatores como a adição de colunas para manter a estrutura do modelo contribuíram para esse aumento, mas o principal motivo foi a reestruturação dos dados, que originalmente tinham uma coluna para cada ano, com a quantidade de instâncias (linhas) igual ao total de municípios no país. Na versão processada, uma única coluna

denominada ‘ano’ foi criada, e cada cidade passou a ter uma entrada na tabela para cada ano, tornando os dados mais práticos para programação e criação de visualizações, e mais coerentes com o projeto do que a abordagem do relatório inicial.

5 CONSIDERAÇÕES FINAIS

A construção e atualização de um banco de dados com base em dados abertos para aplicação em diferentes modelos, como Cidades Inteligentes, Sistemas de Capitais e Objetivos de Desenvolvimento Sustentável (ODS), é um empreendimento complexo e exigente, mas essencial para o desenvolvimento urbano sustentável. O processo de criação inicial da base de dados, que levou seis meses, e sua subsequente atualização, que demorou quatro meses, destacam os desafios envolvidos, como a heterogeneidade das fontes de dados, a diversidade de formatações e a necessidade de um tratamento minucioso para garantir a integridade e usabilidade dos dados.

Apesar dos desafios, o uso de técnicas de ETL (Extração, Transformação e Carga) provou ser eficaz na consolidação de grandes volumes de dados em uma estrutura coesa e acessível. O banco de dados atualizado não só facilita a visualização e comparação de indicadores urbanos, mas também serve como uma ferramenta estratégica para gestores públicos e pesquisadores no desenvolvimento de políticas públicas inovadoras, planejamento estratégico e especializações para o futuro das cidades brasileiras.

Futuramente, espera-se que a base de dados sirva como um alicerce para novas versões do Observatório de Cidades e a Plataforma Web de Indicadores, oferecendo um potencial significativo para o avanço da pesquisa e inovação em governança urbana. Além disso, a integração de técnicas avançadas de inteligência artificial para a automação da coleta e processamento de dados pode otimizar ainda mais este processo, reduzindo custos e melhorando a precisão e rapidez na atualização das informações. Este trabalho contribui significativamente para o campo do desenvolvimento urbano sustentável, promovendo o uso de dados abertos e ferramentas digitais como pilares da gestão moderna das cidades. Por fim, o banco de dados criado será utilizado para uma nova versão para o Observatório de Cidades (Fachinelli et. al, 2023) e a Plataforma Web de Indicadores (Notari et. al, 2020).

REFERÊNCIAS

5 Stars Open Data. (2012). 5 Stars Open Data. Disponível em http://5stardata.info/ (Acessado em abril de 2024).

Bandeira, J. M., et al. (2015). Dados abertos conectados. III Simpósio Brasileiro de Tecnologia da Informação. Disponível em https://www.researchgate.net/profile/Thiago-Avila-3/publication/283569633_Dados_Abertos_Conectados/links/563fa41008aec6f17ddb819b/Dados-Abertos-Conectados.pdf.

Bansal, S. K., & Kagemann, S. (2015). Integrating big data: A semantic extract-transform-load framework. Computer, 48(3), 42–50.

Battistelo, R. (2018). Aplicação web para indicadores de cidades do conhecimento (Monografia, Universidade de Caxias do Sul). Caxias do Sul, RS, Brasil.

Bria, F., & Morozov, E. (2020). A cidade inteligente: Tecnologias urbanas e democracia. [S.l.]: Ubu Editora.

Carrillo, F. (2006). Knowledge cities: Approaches, perspectives and experiences. Oxford, Britain: Butterworth-Heinemann. https://doi.org/10.9780080460628

Carrillo, F. J. (2002). Capital systems: Implications for a global knowledge agenda. Journal of Knowledge Management, 6(4), 379–399.

Carrillo, F. J. (2014). What 'knowledge-based' stands for? A position paper. International Journal of Knowledge-Based Development, 5(4), 402–421.

Carrillo, F. J. (2015). Knowledge-based development as a new economic culture. Journal of Open Innovation: Technology, Market, and Complexity, 1(2), 1–17.

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. A. (2011). A proposed model for data warehouse ETL processes. Journal of King Saud University - Computer and Information Sciences, 23(2), 91–104. Disponível em https://www.sciencedirect.com/science/article/pii/S131915781100019X.

Elmasri, R., & Navathe, S. B. (2018). Sistemas de banco de dados (7ª ed.). São Paulo: Pearson.

Ergazakis, K., Metaxiotis, K. S., & Psarras, J. E. (2004). Towards knowledge cities: Conceptual analysis and success stories. Journal of Knowledge Management, 8, 5–15. Disponível em https://api.semanticscholar.org/CorpusID:38317272.

Fachinelli, A. C., Carrillo, F. J., & D'Arisbo, A. (2014). Capital system, creative economy and knowledge city transformation: Insights from Bento Goncalves, Brazil. Expert Systems with Applications, 41(12), 5614–5624.

Heuser, C. A. (2009). Projeto de banco de dados. Volume 4 da Série Livros didáticos informática UFRGS (Vol. 4). Bookman Editora.

Notari, D. L., Battistelo, R., Molin, L. W., Silva, S. de Á. e., & Fachinelli, A. C. (2019). Aplicação web para indicadores de cidades do conhecimento | Web application for knowledge cities indicators. Brazilian Journal of Management and Innovation (Revista Brasileira de Gestão e Inovação), 7(2), 95–118. Disponível em https://sou.ucs.br/etc/revistas/index.php/RBGI/article/view/7192.

Organização das Nações Unidas (ONU). (2010). Objetivos de Desenvolvimento do Milênio. Disponível em https://brasil.un.org/pt-br/66851-os-objetivos-de-desenvolvimento-do-mil%C3%AAnio (Acessado em 30 de agosto de 2023).

Organização das Nações Unidas (ONU). (2016). Os objetivos de desenvolvimento sustentável: Dos ODM aos ODS: Programa das Nações Unidas para o Desenvolvimento (PNUD). Disponível em http://www.pnud.org.br/ODS.aspx (Acessado em 29 de agosto de 2023).

Open Knowledge Foundation. (2024). The open definition. Disponível em https://opendefinition.org/ (Acessado em 18 de abril de 2024).

Silva, A. G. da. (2022). Desenvolvimento da automatização da coleta de dados na plataforma cidades do conhecimento (Monografia, Universidade de Caxias do Sul). Caxias do Sul, RS, Brasil.

Vanden Broucke, S., & Baesens, B. (2018). Practical Web scraping for data science. New York, NY: Apress.

Weiss, M. C., Bernardes, R. C., & Consoni, F. L. (2015). Cidades inteligentes como nova prática para o gerenciamento dos serviços e infraestruturas urbanos: A experiência da cidade de Porto Alegre. Urbe. Revista Brasileira de Gestão Urbana, 7, 310–324.

Yigitcanlar, T., O'Connor, K., & Westerman, C. (2008). The making of knowledge cities: Melbourne's knowledge-based urban development experience. Cities, 25(2), 63–72.