8. COLETA, PROCESSAMENTO E ARMAZENAMENTO: O TRATAMENTO DE DADOS EM OBSERVATÓRIOS DIGITAIS

Caio Saraiva Coneglian
Fernanda Maciel Rufino
Larissa Moreno Silva
Macela Virginia Cavalcanti de Albuquerque
Felipe da Rocha Ferreira
Diego José Macêdo

8.1 INTRODUÇÃO

Os Observatórios Digitais destacam-se por sua capacidade de coletar, processar e apresentar uma ampla gama de dados, os quais se constituem como um dos elementos centrais desses sistemas. Entretanto, a diversidade de fontes e formatos de dados representa um dos principais desafios nesse ambiente, exigindo soluções tecnológicas diversas para o tratamento adequado. A heterogeneidade dos dados coletados requer uma arquitetura capaz de suportar desde dados estruturados até dados não estruturados, como os provenientes de redes sociais, o que demanda um processamento eficaz e um armazenamento escalável.

Assim, os Observatórios Digitais, além de armazenarem dados, também atuam como sistemas que contemplam indicadores e dados relevantes, oferecendo informações e conhecimento que apoiam a tomada de decisão estratégica. Dessa forma, os observatórios passam a ter o papel que extrapola a exibição de dados brutos, pois envolve a análise e a fusão de múltiplas fontes de dados, permitindo que as organizações identifiquem padrões, tendências e áreas de intervenção.

No contexto do desenvolvimento de um Observatório Digital, é necessário refletir sobre os processos que envolvem a coleta, o armazenamento e o processamento dos dados, assegurando que o sistema seja capaz de lidar com grandes volumes de informações de forma eficiente e segura. A qualidade desses processos impacta diretamente a capacidade do observatório de gerar valor e promover elementos para seus usuários, garantindo sua relevância em um ambiente cada vez mais orientado por dados.

8.2 COLETA E ARMAZENAMENTO DOS DADOS

No âmbito de Observatórios, é importante definir os processos e técnicas necessárias para a realização da coleta e da seleção dos dados que serão obtidos, utilizados, analisados e expostos. Ademais, identifica-se que existe um grande desafio em tal processo, devido às diversas tipologias e objetivos dos documentos que os observatórios utilizam.

Dessa forma, a discussão e a apresentação dos métodos e ferramentas para coleta e seleção de dados devem considerar, primeiramente, a tipologia documental, e em um segundo momento, o destino que será dado aos dados e às informações obtidas.

Partindo de tais premissas, apresenta-se a seguir a relação de tipologia documental identificada dos observatórios, para que a partir disso, verifique-se os métodos para coleta.

Quadro 1 – Tipologia Documental encontrada nos Observatórios Brasileiros analisados

Tipos	Características
Boletins	Documento em formato de texto (PDF)
Panorama/Estatística/Indicadores/Painéis/Gráficos/ Dashboards (números)	Planilhas e Dados Estruturados (*Comma-Separated Values* (CSV), XLS)
Relatórios/Resumos/Anuário	Documento em formato de texto (PDF)

Tipos	Características
Pesquisas/Projetos	Documento em formato de texto (PDF)
Publicações	Documento em formato de texto (PDF)
Dossiê/Legislação	Documento em formato de texto (PDF)
Artigos	Documento em formato de texto (PDF)
Notícias	Documento em formato de texto (PDF)
Mapa/Atlas	Documento em formato de texto (PDF) ou Imagens (PNG, JPG)

Fonte: Dados da pesquisa (2023).

O Quadro 1 apresenta, além dos tipos documentais, uma descrição da característica do documento, destacando os seus possíveis formatos e características. A seguir, apresenta-se as técnicas e as ferramentas necessárias para a coleta de dados de diferentes formatos e estruturas.

8.2.1 COLETA DOS DADOS

8.2.1.1 COLETA DE DADOS ESTRUTURADOS

Os dados são classificados em três tipos principais: estruturados, semiestruturados e não estruturados (Eberendu, 2016). Nesse contexto, apresenta-se o processo de coleta de dados estruturados. Eberendu (2016, p. 48, tradução nossa) aponta que "os dados estruturados se referem aos dados que possuem formato definido e comprimento, fácil de armazenar e analisar com alto grau de organização".

No que tange a esse tipo de dado, há técnicas específicas que possibilitam a coleta de dados estruturados, os quais desempenham um papel essencial nos observatórios, fornecendo informações relevantes e atualizadas

para análises e tomada de decisões. Esses dados concentram, especialmente, planilhas, dados estruturados no formato CSV e TXT, entre outros. O Quadro 2 apresenta, de forma sumarizada, o processo de coleta de dados estruturados.

Quadro 2 – Forma de Coleta de Dados Estruturados

TIPO	FORMATO	FORMA DE COLETA
Planilhas	CSV, XLS, XLSX, TXT	Obtenção de forma manual ou por meio de técnicas de RPA (*Robotic Process Automation*) para coleta automatizada.

Fonte: Dados da pesquisa (2023).

Destaca-se, no Quadro 2, que os dados podem ser coletados manualmente ou por meio de técnicas automatizadas, com o apoio de RPA, que permite a atualização dos dados de forma controlada e sistematizada. O detalhamento de RPA está contido na subseção 8.2.1.4, que explica as técnicas de coleta.

8.2.1.2 COLETA DE DADOS EM AMBIENTES WEB

Complementarmente, no contexto de Observatórios, é essencial a coleta de dados em ambientes web. Essa abordagem é amplamente utilizada para obter informações de fontes on-line, envolvendo a extração sistemática de dados de páginas web, como textos, imagens, tabelas e links. A seguir, no Quadro 3, destaca-se o processo de coleta em ambientes web.

Quadro 3 – Forma de Coleta de Dados em Ambientes Web

TIPO	FORMATO	FORMA DE COLETA
Dados estruturados em páginas Web.	HTML	Uso de ferramentas de *Web Scraping*, como *Beautiful Soup*[17], *Selenium* [18] e *Scrapy* [19].
Dados estruturados em ambientes Web.	*JavaScript Object Notation* (JSON) ou *eXtensible Markup Language* (XML)	Uso de *Application Programming Interface* (APIs) para a coleta de dados.

Fonte: Dados da pesquisa (2023).

Como apresentado no Quadro 3, existem diversas ferramentas e técnicas para realizar a coleta de dados em ambientes web, como o *Web Scraping*, que envolve o uso de bibliotecas e *frameworks* como *Beautiful Soup*, *Selenium* e *Scrapy* para extrair informações de páginas web de forma automatizada. Além disso, há a opção de coleta dos dados por meio de API, em que a coleta é feita diretamente a partir da solicitação a algum serviço. Na seção 8.2.1.4, detalham-se as abordagens utilizadas em *Web Scraping* e APIs.

8.2.1.3 COLETA DE DADOS EM BASES DE DADOS

A coleta de dados em bases de dados envolve a extração de informações de sistemas de armazenamento estruturados, como bancos de dados relacionais ou bancos de dados NoSQL. Nessa abordagem, as consultas e os comandos são utilizados para recuperar os dados desejados. As linguagens de consulta mais comuns para coleta de dados em bases de dados são *Structured Query Language* (SQL) para bancos de dados relacionais e consultas específicas para bancos de dados NoSQL, como o MongoDB. Essa abordagem é ideal quando os dados estão armazenados em bases

17 Disponível em: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Acesso em: 4 dez. 2024.

18 Disponível em: https://www.selenium.dev/. Acesso em: 4 dez. 2024.

19 Disponível em: https://scrapy.org/. Acesso em: 4 dez. 2024.

de dados e é necessário realizar consultas complexas para obter as informações desejadas.

Quadro 4 – Forma de Coleta de Dados em Bases de Dados

TIPO	FORMATO	FORMA DE COLETA
Dados contidos em ambientes de banco de dados relacionais.	SQL	Uso de ferramentas de apoio aos Sistemas de Gerenciamento de Banco de Dados (SGBDs), como MySQL *Workbench*, que permite a consulta utilizando SQL.
Dados contidos em ambientes de banco de dados não relacionais.	JSON ou outros	Uso de ferramentas gerenciadoras que permite a consulta aos dados.

Fonte: Dados da pesquisa (2023).

No Quadro 4, destaca-se que há duas formas principais para coleta de acordo com o tipo de banco de dados. Tais informações são detalhadas na seção 8.2.1.4.

8.2.1.4 FERRAMENTAS PARA COLETA DE DADOS

A partir dos aspectos apresentados, destacam-se quatro técnicas principais para a realização de coleta de dados: *Web Scraping*, APIs, RPA e Coleta em Banco de Dados.

8.2.1.4.1 TÉCNICA 1: *WEB SCRAPING*

A técnica de *Web Scraping* é amplamente utilizada para a extração de conteúdos disponíveis em páginas HTML, consistindo na extração automatizada de dados de páginas web. Essa técnica permite extrair informações estruturadas, como tabelas, listas, textos e imagens, diretamente dos elementos HTML das páginas.

Essa técnica é definida como "[...] uma ótima técnica de extração de dados não estruturados dos sites e transformação desses dados em dados

estruturados que podem ser armazenados e analisados em um banco de dados" (Sirisuriya, 2015, p. 135, tradução nossa).

Existem várias bibliotecas e *frameworks* disponíveis que facilitam a implementação do *Web Scraping*, oferecendo recursos avançados para navegar em páginas web, localizar elementos específicos e extrair os dados desejados. Algumas das bibliotecas populares para *Web Scraping* incluem:

- ***Beautiful Soup:*** Biblioteca *Python* que permite analisar e extrair dados de documentos HTML e XML. Ela facilita a navegação na estrutura da página, busca de elementos por meio de seletores e extração dos dados desejados;

- ***Selenium:*** Ferramenta amplamente utilizada para automação de testes em navegadores web. Ademais, o *Selenium* também pode ser empregado para o *Web Scraping*, pois permite interagir com páginas web de forma programática, preenchendo formulários, clicando em botões e coletando os dados resultantes;

- ***Scrapy:*** *Framework Python* dedicado ao *Web Scraping*, que fornece uma estrutura completa para a construção de *spiders* (robôs de coleta de dados) altamente personalizáveis. O *Scrapy* permite definir as regras de navegação, extração e persistência dos dados de maneira flexível e escalável.

8.2.1.4.2 TÉCNICA 2: INTERFACES DE PROGRAMAÇÃO DE APLICAÇÕES (APIS)

Outro modo de coleta de dados em ambientes web é por meio de APIs. As APIs são conjuntos de regras e protocolos que permitem a comunicação entre diferentes sistemas de software. Assim, diversas organizações e empresas disponibilizam APIs para permitir a coleta de dados de forma estruturada e programática.

A definição a seguir apresenta o conceito de API com mais detalhes:

> API significa Application Programming Interface (Interface de Programação de Aplicação). No contexto de APIs, a palavra Aplicação refere-se a qualquer software com uma função distinta. A interface pode

> ser pensada como um contrato de serviço entre duas aplicações. Esse contrato define como as duas se comunicam usando solicitações e respostas. A documentação de suas respectivas APIs contém informações sobre como os desenvolvedores devem estruturar essas solicitações e respostas (AWS, c2024).

As APIs são uma maneira eficiente e confiável de obter dados de fontes externas, pois fornecem *endpoints* (pontos de acesso) que podem ser acessados com solicitações HTTP. Esses *endpoints* retornam os dados solicitados em formato padronizado, como JSON, XML, entre outros formatos estabelecidos.

8.2.1.4.3 TÉCNICA 3: COLETA DE DADOS COM RPA (AUTOMAÇÃO ROBÓTICA DE PROCESSOS)

Considerando que os Observatórios podem utilizar diversos dados disponíveis publicamente em formatos de planilha, uma forma de automatizar o processo de coleta de dados é com o uso da técnica de RPA. O RPA é uma abordagem que utiliza softwares robóticos para automatizar tarefas repetitivas e baseadas em regras, incluindo a extração de dados de planilhas e sistemas governamentais.

Uma definição de RPA aponta que:

> RPA é um termo abrangente para ferramentas que operam na interface do usuário de outros sistemas de computador da maneira que um ser humano faria. O RPA visa substituir pessoas por automação feita de maneira 'de fora para dentro' (Aalst; Bichler; Heinzl, 2018, p. 269, tradução nossa).

Uma das vantagens do RPA é a capacidade de interagir com interfaces de usuário, como planilhas eletrônicas, navegadores web e aplicativos *desktop*, de forma semelhante a um usuário humano. Isso permite que os robôs executem ações, como abrir planilhas, localizar informações específicas, copiar e colar dados, e até mesmo preencher formulários automaticamente.

No contexto da coleta de dados em ambientes governamentais ou de transparência, o RPA pode ser aplicado para automatizar o processo de busca

e extração de dados de planilhas disponibilizadas pelos órgãos públicos. Os robôs de RPA podem ser configurados para acessar regularmente as fontes de dados, identificar atualizações ou novas informações e realizar a extração dos dados relevantes.

8.2.1.4.4 TÉCNICA 4: COLETA DE DADOS EM BANCO DE DADOS RELACIONAIS E NOSQL

A coleta de dados em bancos de dados pode apoiar a obtenção de dados para os observatórios. No entanto, há dois tipos principais de banco de dados: bancos de dados relacionais e bancos de dados NoSQL.

Os bancos de dados relacionais são amplamente utilizados para armazenar e gerenciar dados estruturados em tabelas, seguindo um modelo de dados predefinido. Para realizar a coleta de dados nesse tipo de banco de dados, é necessário utilizar consultas SQL para recuperar as informações desejadas.

A principal forma de consultar banco de dados é por meio de consultas SQL personalizadas. É possível escrever consultas SQL específicas para extrair dados de tabelas e relacionamentos no banco de dados. Essas consultas podem ser executadas diretamente no banco de dados ou por meio de ferramentas de acesso, como interfaces de linha de comando ou interfaces gráficas.

Já no âmbito dos bancos de dados NoSQL, estes são projetados para lidar com volumes massivos de dados, estruturas flexíveis e requisitos de escalabilidade. Eles oferecem uma variedade de modelos de dados, como documentos, chave-valor, colunas amplas e grafos. A coleta de dados em bancos de dados NoSQL requer abordagens específicas para cada modelo de dados.

Para realizar buscas e coletar dados em bancos NoSQL, verifica-se que cada modelo de dados NoSQL tem suas próprias consultas e métodos de acesso aos dados. Por exemplo, em bancos de dados de documentos, podem ser usadas consultas baseadas em JSON para recuperar documentos específicos ou executar operações de agregação. Em bancos de dados de chave-valor, é possível obter dados diretamente por meio das chaves.

8.2.2 INDICAÇÃO DE TÉCNICAS PARA TRATAMENTO E ARMAZENAMENTO

As técnicas e ferramentas recomendadas para o tratamento e armazenamento de informações no contexto da modelagem de um sistema para Observatório são essenciais para a compreensão de Observatórios no contexto atual, que se vincula fortemente à utilização de dados para a tomada de decisões nos mais diversos cenários. Complementarmente, destaca-se que o processo de tratamento e armazenamento se vincula diretamente à qualidade da informação, pois ela é de extrema importância para garantir a confiabilidade e a eficiência das operações do Observatório. Portanto, é essencial adotar abordagens adequadas que permitam o correto processamento, organização e armazenamento dos dados coletados.

No que tange às técnicas de tratamento de informações, identifica-se que há diversas etapas e elementos que devem ser considerados, como é apresentado a seguir:

8.2.2.1 ARMAZENAMENTO DOS DADOS

O primeiro aspecto após a coleta de dados é a preocupação com o processo de armazenamento dos dados. Nesse contexto, é crucial adotar técnicas adequadas para garantir a eficiência e a segurança das informações, considerando o contexto de um sistema para Observatório. Uma abordagem recomendada é a implementação de técnicas de particionamento de dados, que permitem dividir grandes conjuntos de dados em partes menores, facilitando o acesso e a recuperação de informações específicas. Além disso, a replicação de dados pode ser empregada para aumentar a disponibilidade e a tolerância a falhas do sistema, garantindo que os dados permaneçam acessíveis mesmo em caso de falhas em componentes de *hardware* ou conexões de rede.

O uso de técnicas de compressão de dados também é relevante, pois reduz o espaço de armazenamento necessário, minimizando os custos associados e otimizando o desempenho na recuperação de informações. A escolha das técnicas de armazenamento deve ser feita de acordo com os requisitos e características específicas do Observatório, levando em consideração a escala dos dados, as necessidades de desempenho e a segurança das informações.

8.2.2.1.1 FERRAMENTAS PARA ARMAZENAMENTO DOS DADOS

Há diferentes abordagens e formas para realizar o processo de armazenamento dos dados. A seguir, serão apresentas três propostas.

8.2.2.1.2 BANCOS DE DADOS RELACIONAIS

Os bancos de dados relacionais são amplamente utilizados para o armazenamento de informações estruturadas, oferecendo recursos de consulta e integridade referencial. Recomenda-se a adoção de SGBDs populares, como MySQL, PostgreSQL ou Oracle, que possuem robustez e escalabilidade para lidar com grandes volumes de dados.

No contexto dos Observatórios, alguns tipos de dados são adequados para serem armazenados em Bancos Relacionais, em especial aqueles que são obtidos a partir de sistemas de informação transacionais. No entanto, uma proposta para conjuntos de dados menos estruturados, que muitas vezes são utilizados em Observatórios, é a utilização de Bancos de Dados NoSQL, como apresentado a seguir.

8.2.2.1.2 BANCOS DE DADOS NOSQL

Para o armazenamento de informações não estruturadas ou semiestruturadas, os bancos de dados NoSQL são uma opção viável. Esses sistemas permitem a flexibilidade no armazenamento de diferentes tipos de dados, como documentos, gráficos e dados em formato de chave-valor. Exemplos de bancos de dados NoSQL incluem MongoDB[20], Apache Cassandra[21] e Redis[22].

20 Disponível em: https://www.mongodb.com/. Acesso em: 4 dez.

21 Disponível em: https://cassandra.apache.org/_/index.html. Acesso em: 4 dez.

22 Disponível em: https://redis.io/. Acesso em: 4 dez.

8.2.2.1.3 DATA WAREHOUSES

Complementando as soluções de Bancos de Dados Relacionais e NoSQL, é essencial que exista um processo de consolidação e junção desses dados para que possam ser realizadas análises e outros processos. Dessa forma, os *Data Warehouses* são soluções que permitem a consolidação e o armazenamento de grandes volumes de dados de diversas fontes em um único local. Essa abordagem facilita a análise e a geração de relatórios, fornecendo um ambiente centralizado para consulta e processamento dos dados.

Data Warehouse são definidos como: "um sistema [que] armazena dados históricos integrados e preparados para serem analisados por OLAP [OnLine Analytical Processing] e outras ferramentas" (Perez Martinez *et al.*, 2008, p. 940, tradução nossa).

8.2.2.2 NORMALIZAÇÃO DE DADOS

A primeira técnica apresentada é a normalização de dados, que trata de um método essencial para garantir a consistência e a uniformidade dos dados armazenados nos mais diversos contextos. Por meio desse processo, é possível evitar redundâncias e inconsistências, facilitando a integração e análise dos dados. Recomenda-se a utilização de padrões de normalização estabelecidos, como o modelo de normalização de banco de dados relacional.

Complementarmente, aponta-se que normalização de dados é definido como o processo "[...] onde os dados são dimensionados para uniformidade. A normalização de dados é necessária para estudar as melhores características dos dados" (Sree; Bindu, 2018, p. 209, tradução nossa).

8.2.2.2.1 FERRAMENTAS PARA NORMALIZAÇÃO DE DADOS

Há diversas ferramentas com objetivos distintos para realizar o processo de normalização dos dados. Além da própria normalização, utilizando softwares de gerenciamento de banco de dados relacional que permitem a definição de esquemas de banco de dados — incluindo a criação de tabelas, relacionamentos e restrições de integridade referencial — há outras ferramentas para a normalização de dados.

Em especial, além das ferramentas específicas de gerenciamento e modelagem de dados, existem também bibliotecas e *frameworks* de programação que oferecem funcionalidades para a normalização de dados. Essas ferramentas permitem a implementação de algoritmos e rotinas personalizadas para realizar a normalização dos dados, considerando as regras de negócio específicas do Observatório.

A adoção das ferramentas adequadas para a normalização de dados contribuirá para a garantia da consistência e da integridade dos dados no sistema de Observatório. Ao eliminar redundâncias e inconsistências, as ferramentas de normalização proporcionam uma base sólida para a análise e a tomada de decisões com base nos dados coletados, resultando em informações confiáveis e relevantes para as atividades do Observatório.

A seguir, serão apresentadas as ferramentas que podem ser utilizadas para a normalização de dados:

- ***OpenRefine*** [23]: é uma ferramenta de código aberto que permite limpar e transformar dados de forma interativa. Ela oferece recursos avançados para detecção e correção de erros, padronização de formatos, remoção de duplicatas e enriquecimento dos dados;

- ***Alteryx Designer Cloud*** [24] : essa ferramenta é voltada para a preparação de dados e possui uma interface intuitiva para limpar, transformar e estruturar dados de forma visual. Ela oferece recursos de sugestão automática, detecção de padrões e visualizações interativas para facilitar o processo de normalização;

- ***Microsoft SQL Server Integration Services (SSIS)*** [25] : é uma ferramenta de integração de dados da Microsoft que permite criar fluxos de trabalho para Extração, Transformação e Carga (ETL) de dados. Ela oferece recursos para mapeamento, limpeza e normalização de dados durante o processo de integração.

23 Disponível em: https://openrefine.org/. Acesso em: 4 dez.

24 Disponível em: https://www.trifacta.com/. Acesso em: 4 dez.

25 Disponível em: https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver16. Acesso em: 4 dez.

8.2.2.3 LIMPEZA DE DADOS

Junto ao processo de normalização, é essencial a realização da limpeza de dados, sendo um processo necessário para identificar e corrigir erros, omissões e inconsistências nos dados coletados. Isso inclui a remoção de valores nulos, a padronização de formatos e a detecção de *outliers* [26].

A limpeza de dados é definida como

> [...] uma abordagem inicial em que os conjuntos de dados são limpos para identificar quaisquer dados ausentes, remover os dados ruidosos e preparar os dados para análise. A limpeza de dados é necessária para resolver o problema de qualidade de dados. O problema de qualidade de dados é onde a análise pode dar errado em dados confusos (Sree; Bindu, 2018, p. 209, tradução nossa).

Para o processo de limpeza, existem diversas ferramentas e algoritmos disponíveis para auxiliar nesse processo, como técnicas de análise estatística e algoritmos de detecção de anomalias, como apresentadas a seguir.

8.2.2.3.1 FERRAMENTAS PARA LIMPEZA DE DADOS

O processo de limpeza de dados exige ferramentas capazes de apoiar a exclusão de adequação da base de dados. Os softwares de limpeza e transformação de dados oferecem recursos avançados para identificar e corrigir erros, inconsistências, valores ausentes e duplicatas nos conjuntos de dados. Eles permitem a aplicação de regras e algoritmos para padronização, normalização e validação dos dados, garantindo sua conformidade com os requisitos estabelecidos.

Um software que permite a realização de uma série de atividades para realizar a limpeza dos dados é o *OpenRefine*. O *OpenRefine* permite que sejam realizados filtros, exclusão dos dados, preenchimento de células em branco, além de permitir ao final o processo de exportação.

26 Outlier é um dado ou um conjunto de dados que se distancia radicalmente dos demais que compõem o conjunto total analisado.

Outra ferramenta útil é um software de deduplicação de dados. Essas ferramentas permitem identificar registros duplicados em um conjunto de dados, seja por meio de comparação de campos-chave ou por algoritmos de similaridade. A deduplicação é crucial para garantir a precisão e a integridade dos dados, evitando a duplicação de informações e a distorção dos resultados das análises.

Destaca-se que o *OpenRefine* se mostra como uma ferramenta capaz de realizar o processo de deduplicação dos dados de forma efetiva.

Cabe salientar que, no âmbito dos Observatórios, a realização do processo de limpeza deverá ser executado de forma constante, com um ciclo contínuo de monitoramento e melhoria da qualidade dos dados. As ferramentas selecionadas devem permitir a automatização e a repetibilidade dessas tarefas, a fim de reduzir o tempo e os esforços necessários para manter a integridade dos dados.

No geral, o uso do software como o *OpenRefine* permite a garantia de dados confiáveis e consistentes, a redução de erros e a melhoria da precisão das análises e dos processos de tomada de decisão. Investir em ferramentas de limpeza de dados é essencial para obter insights valiosos e embasar as atividades do Observatório em informações confiáveis e de alta qualidade.

Destaca-se que há outros softwares que permitem a realização dos processos de limpeza, como o *Microsoft PowerBI*[27] e o Tableau, que possibilitam o tratamento e a limpeza dos dados no momento de sua importação. No entanto, o *OpenRefine* sobressai como uma ferramenta mais completa, devido ao seu caráter específico e ao foco exclusivo nessa etapa anterior à análise.

8.2.2.4 TRANSFORMAÇÃO E ENRIQUECIMENTO DE DADOS

Após o processo de normalização e limpeza dos dados, a próxima etapa envolve a transformação e o enriquecimento de dados, que busca aplicar

27 Disponível em: https://powerbi.microsoft.com/pt-br/. Acesso em: 4 dez.

técnicas para converter dados brutos em formatos mais adequados e incorporar informações adicionais que enriquecem o conjunto de dados. Isso pode incluir a agregação de dados, a aplicação de regras de negócio e a incorporação de dados provenientes de outras fontes confiáveis.

8.2.2.4.1 FERRAMENTAS PARA TRANSFORMAÇÃO E ENRIQUECIMENTO DE DADOS

Além das ferramentas de normalização e limpeza, aponta-se a necessidade de possuir ferramentas adequadas para a transformação e o enriquecimento de dados. Essas ferramentas desempenham um papel fundamental na preparação e no aprimoramento dos dados coletados, permitindo que sejam convertidos em formatos mais adequados e enriquecidos com informações adicionais.

Uma das ferramentas amplamente utilizadas nesse contexto é um software de ETL, que possibilita a extração dos dados de diversas fontes, como bancos de dados, arquivos CSV ou APIs, e a aplicação de transformações para limpeza, normalização e padronização dos dados. Além disso, o ETL permite a carga dos dados transformados em um novo destino, como um banco de dados ou um *Data Warehouse*.

Outro importante tipo de software é voltado à integração de dados. Essas ferramentas facilitam a integração de dados provenientes de diferentes fontes, permitindo a harmonização e a unificação de formatos, a reconciliação de registros duplicados e a resolução de conflitos. Com essa integração, torna-se possível obter uma visão holística e unificada dos dados, promovendo uma compreensão mais completa e precisa.

Adicionalmente, o enriquecimento de dados pode ser realizado por meio de ferramentas externas de enriquecimento de dados. Essas ferramentas permitem a incorporação de informações adicionais aos dados existentes, provenientes de fontes confiáveis, como dados geográficos, dados demográficos ou informações atualizadas de terceiros. Esse enriquecimento de dados pode aumentar a qualidade, a relevância e o valor dos dados, fornecendo uma visão mais completa e enriquecida do contexto em que estão inseridos.

Alguns dos softwares de transformação mencionados no contexto da normalização e limpeza são capazes também de apoiar o processo de enriquecimento, conforme a relação a seguir:

- ***OpenRefine:*** é uma ferramenta de código aberto para limpeza e transformação de dados. Além de oferecer esses recursos, o *OpenRefine* também permite a transformação de dados por meio da execução de operações em massa, como separação de colunas, remoção de espaços em branco, conversão de formatos e muito mais;

- ***Talend Data Integration***[28]: plataforma de integração de dados que inclui recursos avançados de transformação de dados. Ela permite a criação de fluxos de trabalho para transformar dados de várias fontes, combinando, filtrando, agregando e aplicando regras de negócio. A plataforma suporta diferentes formatos de dados e oferece uma interface intuitiva para a criação e gerenciamento das transformações;

- ***RapidMiner***[29]: é uma plataforma de análise de dados que inclui recursos de transformação e enriquecimento de dados. Ela permite a criação de fluxos de trabalho para a preparação de dados, incluindo operações de transformação, filtragem, agregação e enriquecimento com dados externos. O *RapidMiner* suporta diferentes fontes de dados e oferece uma ampla gama de algoritmos e técnicas de análise;

- **Bibliotecas Pandas**[30] **e NumPy**[31] **(Linguagem Python):** tais bibliotecas oferecem recursos para transformação e manipulação de dados. Essas bibliotecas fornecem funções e métodos para realizar operações de transformação, limpeza, agregação e enriquecimento de dados de maneira eficiente.

28 Disponível em: https://www.talend.com/products/integrate-data/. Acesso em: 4 dez.

29 Disponível em: https://rapidminer.com/. Acesso em: 4 dez.

30 Disponível em: https://pandas.pydata.org/. Acesso em: 21 nov. 2024.

31 Disponível em: https://numpy.org/. Acesso em: 21 nov. 2024.

8.2.3 FERRAMENTAS PARA DADOS GEORREFERENCIAIS

Já no contexto de dados georreferenciais, que são informações relacionadas à localização geográfica de eventos, objetos ou fenômenos, destaca-se que a utilização de ferramentas adequadas para análise e visualização desses dados é essencial em um observatório. Apresentam-se algumas ferramentas que podem ser utilizadas para dados georreferencias:

- **Sistemas de Informação Geográfica (SIG):** são plataformas que permitem a captura, armazenamento, análise e visualização de dados georreferenciais. Alguns exemplos de SIG: ArcGIS, QGIS e MapInfo;

- **Visão**[32]: desenvolvido pelo Ibict, o Visão é um sistema para análise e visualização de dados georreferenciais, que oferece recursos para manipulação de dados espaciais e criação de mapas interativos.

8.3 CONSIDERAÇÕES FINAIS

As ferramentas e técnicas para coleta e armazenamento de dados são elementos essenciais para o desenvolvimento e funcionamento de observatórios digitais. Como apresentado, foram descritos elementos e principais abordagens para que o processo de coleta de dados seja realizado de maneira sistemática e organizada, buscando trazer para os observatórios a qualidade e a integridade das informações.

A coleta de dados é um meio de adquirir dados de diferentes fontes, utilizando tecnologias e processos de APIs ou extração de dados dos *websites*, com o objetivo de estruturar e organizar essas informações para depois apresentar em tecnologias de visualização. Além disso, a manipulação desses dados durante o processo de extração, utilizando técnicas de transformação e enriquecimento, agrega valor e contexto, o que é essencial para garantir que os dados capturados estejam adequados aos objetivos do observatório.

32 Disponível em: https://visao.ibict.br. Acesso em: 21 nov. 2024.

No que diz respeito ao armazenamento, há a possibilidade de utilização de bancos de dados relacionais ou bancos NoSQL, que podem ser combinados ou escolhidos conforme as especificidades do observatório, tais como a natureza dos dados e as necessidades de escalabilidade e performance. Ferramentas como MySQL e MongoDB oferecem soluções para diferentes tipos de dados e possibilitam o armazenamento dos dados coletados.

Com base nos elementos expostos, destaca-se que a combinação de ferramentas adequadas para coleta e armazenamento de dados permite que um Observatório Digital atue apoiando as comunidades, possibilitando o monitoramento e a análise de informações. Ademais, ressalta-se a necessidade de integrar as soluções para garantir a coleta e o armazenamento dos dados, além de organizá-los de maneira que facilitem sua utilização em processos de análise e tomada de decisão.

REFERÊNCIAS

AALST, W. M. P.; BICHLER, M.; HEINZL, A. Robotic process automation. **Business & information systems engineering,** [*s. l.*], v. 60, n. 4, p. 269-272, May 2018. DOI: https://doi.org/10.1007/s12599-018-0542-4. Disponível em: https://link.springer.com/article/10.1007/s12599-018-0542-4. Acesso em: 21 nov. 2024.

AWS. **O que é uma API?.** [*s. l.*]: Amazon Web Services, c2024. https://aws.amazon.com/pt/what-is/api/. Acesso em: 25 maio 2024.

EBERENDU, A. C. Unstructured Data: an overview of the data of Big Data. **International Journal of Computer Trends and Technology**, [*s. l.*], v. 38, n. 1, p. 46-50, Aug. 2016. DOI: 10.14445/22312803/IJCTT-V38P109. Disponível em: https://www.ijcttjournal.org/archives/ijctt-v38p109. Acesso em: 14 mar. 2024.

PEREZ MARTINEZ, J. M.; BERLANGA, R.; JOSE ARAMBURU, M.; PEDERSEN, T. B. Integrating data warehouses with web data: a survey. **IEEE Transactions on Knowledge and Data Engineering**, [*s. l.*], v. 20, n. 7, p. 940-955, July 2008. DOI: 10.1109/TKDE.2007.190746. Disponível em: https://ieeexplore.ieee.org/document/4490177. Acesso em: 4 fev. 2024.

SIRISURIYA, S. C. M. S. A Comparative Study on *Web Scraping*. *In*: INTERNATIONAL RESEARCH CONFERENCE, 8., 2015, Sri Lanka. **Proceedings [...]** . Ratmalana: KDU, 2015. p. 135 - 140. Disponível em: http://ir.kdu.ac.lk/handle/345/1051. Acesso em: 4 fev. 2024.

SREE, K. D.; BINDU, C. S. Data analytics: Why data normalization. **International Journal of Engineering and Technology (UAE)**, [*s. l.*], v. 7, n. 4.6, p. 209-213, 2018. Disponível em: https://www.sciencepubco.com/index.php/ijet/article/view/20464. Acesso em: 30 abr. 2024.

Como citar o capítulo: CONEGLIAN, Caio Saraiva; RUFINO, Fernanda Maciel; SILVA, Larissa Moreno; ALBUQUERQUE, Marcela Virginia Cavalcanti de; FERREIRA, Felipe da Rocha; MACÊDO, Diego José. Coleta, processamento e armazenamento: o tratamento de dados em observatórios digitais. *In*: MACÊDO, Diego José; CONEGLIAN, Caio Saraiva (org.). **Estudos em observatórios:** conceitos, modelo e aplicações. Brasília, DF: Editora Ibict, 2025. Cap. 8, p. 141-160. DOI: 10.22477/9788570131973.cap8.