UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RICARDO HENRICKI DIAS BORGES

# Em Busca do Estado da Arte e da Prática sobre Schema Matching na Indústria Brasileira - Resultados Preliminares de uma Revisão de Literatura e uma Pesquisa de Opinião

Goiânia
2024

UFG

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

# TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

**1. Identificação do material bibliográfico**

[ x ] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

**Exemplos**: Estudo de caso ou Revisão sistemática ou outros formatos.

**2. Nome completo do autor**

Ricardo Henricki Dias Borges

**3. Título do trabalho**

Em Busca do Estado da Arte e da Prática sobre Schema Matching na Indústria Brasileira - Resultados Preliminares de uma Revisão de Literatura e uma Pesquisa de Opinião

**4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)**

Concorda com a liberação total do documento [ x ] SIM [ ] NÃO[1]

**[1]** Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
**a)** consulta ao(à) autor(a) e ao(à) orientador(a);
**b)** novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

**Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.**

Documento assinado eletronicamente por **Valdemar Vicente Graciano Neto**, **Professor do Magistério Superior**, em 18/09/2024, às 16:16, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ricardo Henricki Dias Borges**, **Discente**, em 19/09/2024, às 08:21, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4834695** e o código CRC **572621E3**.

**Referência:** Processo nº 23070.022893/2024-00 SEI nº 4834695

RICARDO HENRICKI DIAS BORGES

# Em Busca do Estado da Arte e da Prática sobre Schema Matching na Indústria Brasileira - Resultados Preliminares de uma Revisão de Literatura e uma Pesquisa de Opinião

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

**Área de concentração:** Ciência da Computação.

**Linha de pesquisa:** Metodologias e Técnicas de Computação.

**Orientador:** Prof. Dr. Valdemar V. Graciano Neto

**Co-Orientador:** Prof. Dr. Leonardo Andrade Ribeiro

Goiânia
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

BORGES, RICARDO HENRICKI DIAS
Em Busca do Estado da Arte e da Prática sobre Schema Matching na Indústria Brasileira - Resultados Preliminares de uma Revisão de Literatura e uma Pesquisa de Opinião [manuscrito] / RICARDO HENRICKI DIAS BORGES. - 2024.
107 f.

Orientador: Prof. Dr. Valdemar Vicente Graciano Neto ; co orientador Dr. Leonardo Andrade Ribeiro .
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2024.

1. Schema Matching. 2. Integração de Dados. 3. Inteligência Artificial. 4. Survey. I. , Valdemar Vicente Graciano Neto, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

# ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **23** da sessão de Defesa de Dissertação de **Ricardo Henricki Dias Borges**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos dois dias do mês de setembro de dois mil e vinte e quatro, a partir das catorze horas, na sala 150 do INF, realizou-se a sessão pública de Defesa de Dissertação intitulada **"O Estado da Prática Sobre Schema Matching em Processos de Integração de Dados na Indústria Brasileira"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Valdemar Vicente Graciano Neto (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Leonardo Andrade Ribeiro (INF/UFG), coorientador; Professor Doutor Rafael Zancan Frantz (UNIJUÍ), membro titular externo; e Professor Doutor Arlindo Rodrigues Galvão Filho (INF/UFG), membro titular interno. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Valdemar Vicente Graciano Neto, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos dois dias do mês de setembro de dois mil e vinte e quatro.

TÍTULO SUGERIDO PELA BANCA

Em Busca do Estado da Arte e da Prática sobre Schema Matching na Indústria Brasileira - Resultados Preliminares de uma Revisão de Literatura e uma Pesquisa de Opinião

---

Documento assinado eletronicamente por **Rafael Zancan Frantz**, **Usuário Externo**, em 27/09/2024, às 16:11, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ricardo Henricki Dias Borges**, **Discente**, em 27/09/2024, às 16:18, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Valdemar Vicente Graciano Neto**, **Professor do Magistério Superior**, em 27/09/2024, às 16:58, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Leonardo Andrade Ribeiro**, **Professora do Magistério Superior**, em 27/09/2024, às 17:16, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 30/09/2024, às 17:26, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

# Resumo

H. D. BORGES, RICARDO. **Em Busca do Estado da Arte e da Prática sobre Schema Matching na Indústria Brasileira - Resultados Preliminares de uma Revisão de Literatura e uma Pesquisa de Opinião**. Goiânia, 2024. 106p. Dissertação de Mestrado Relatório de Graduação. Programa de Pós Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

A integração de sistemas e a interoperabilidade entre diferentes bancos de dados são desafios críticos na tecnologia da informação, principalmente devido à diversidade de esquemas de dados. A técnica de *schema matching* é essencial para unificar esses esquemas, facilitando a pesquisa, análise e descoberta de conhecimento. Esta dissertação investiga a aplicação do *schema matching* na indústria de software brasileira, com foco em compreender as razões para sua baixa adoção. A pesquisa incluiu um mapeamento sistemático sobre o uso de algoritmos de Inteligência Artificial (IA) e técnicas de similaridade no *schema matching*, além de uma pesquisa de opinião com 35 profissionais do setor. Os resultados indicam que, embora o *schema matching* ofereça melhorias significativas na integração de dados, como a redução de tempo e o aumento da assertividade, a maioria dos profissionais desconhece o termo, mesmo entre aqueles que utilizam ferramentas semelhantes. A baixa adoção dessas técnicas pode ser atribuída à falta de ferramentas gratuitas ou *open source* e à ausência de planos de implementação nas empresas. A dissertação destaca a necessidade de iniciativas que superem essas barreiras, capacitando os profissionais e promovendo o uso mais amplo do *schema matching* na indústria brasileira.

**Palavras–chave**

Schema Matching, Integração de dados, Inteligência Artificial, Survey

# Abstract

H. D. BORGES, RICARDO. **The State of the Practice on Schema Matching in Data Integration Processes in Brazilian Industry**. Goiânia, 2024. 106p. MSc. Dissertation Relatório de Graduação. Programa de Pós Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

The integration of systems and interoperability between different databases are critical challenges in information technology, mainly due to the diversity of data schemas. The schema matching technique is essential for unifying these schemas, facilitating research, analysis, and knowledge discovery. This dissertation investigates the application of schema matching in the Brazilian software industry, focusing on understanding the reasons for its low adoption. The research included a systematic mapping of the use of Artificial Intelligence (AI) algorithms and similarity techniques in schema matching, as well as a survey with 35 professionals in the field. The results indicate that, although schema matching offers significant improvements in data integration processes, such as reducing time and increasing accuracy, most professionals are unfamiliar with the term, even among those who use similar tools. The low adoption of these techniques can be attributed to the lack of free or open source tools and the absence of implementation plans within companies. The dissertation highlights the need for initiatives that overcome these barriers, empower professionals, and promote broader use of schema matching in the Brazilian industry.

**Keywords**

Schema Matching, Data Integration, Artificial Intelligence, Survey

# Sumário

CAPÍTULO 1

# Introdução

Nos últimos anos, a indústria de software tem se deparado com desafios crescentes de integração de grandes volumes de dados. A integração de dados — processo de combinar dados de diferentes fontes para proporcionar uma visão unificada [Doan, Halevy e Ives 2012] — é uma necessidade crítica para muitas organizações que buscam consolidar informações de diversas fontes para apoiar suas decisões e atingir seus objetivos. Todavia, integrar dados pode ser um desafio, particularmente quando esses dados são provenientes de fontes com representações distintas [Calvanese et al. 2001]. Nesse contexto, as representações são denominadas de "esquemas", que são uma descrição lógica e estruturada da organização dos dados, podendo ser em um sistema de banco de dados relacional, sistema de banco de dados semi-estruturado e não estruturado. Por exemplo, em um banco de dados relacional, o esquema define a estrutura dos objetos de dados, como tabelas, colunas, relacionamentos e restrições [Date 2004].

Quando a integração de dados é realizada sem os devidos cuidados, podem surgir problemas como erros, duplicação de informações e inconsistências nos dados [Azeroual, Saake e Schallehn 2018], causando um comprometimento da qualidade dos dados, aumento dos custos operacionais, ineficiências analíticas, entre outros problemas. Além disso, o processo de integração pode se tornar demorado, complexo e altamente dependente, o que acaba impactando negativamente a eficiência das operações organizacionais.

Existem várias formas de integração de dados, cada uma com suas próprias vantagens e desvantagens, dependendo das necessidades da organização e das fontes de dados envolvidas. Uma das formas mais comuns de integração de dados entre sistemas de informação (SI) é a integração baseada em API (Interface de Programação de Aplicativos) [Yusof, Man e Ismail 2022]. As APIs oferecem uma forma padronizada de integração, podendo ser personalizada para atender às necessidades específicas de cada organização.

Outra forma popular de integração de dados entre SI é a integração de dados em tempo real [Tatbul 2010], também chamada de integração virtual em [Haas 2006]. Essa abordagem envolve a transmissão contínua de informações entre os sistemas, permitindo que as informações sejam atualizadas instantaneamente. É particularmente útil para or-

ganizações que precisam de informações atualizadas rapidamente para tomar decisões de negócios críticas. Essa abordagem pode ser implementada por meio de diferentes tecnologias, como o uso de mensagerias, *streaming* de dados e CDC (Change Data Capture). A escolha da tecnologia depende das necessidades específicas de cada organização, dos sistemas existentes e da complexidade dos processos de integração.

Além disso, a integração de dados por meio de ETL, do inglês *Extract, Transform, Load* (Extração, Transformação e Carga) é outra forma popular de integrar dados entre SI [Sreemathy et al. 2021]. Essa abordagem envolve a extração de informações de diferentes fontes, a transformação dessas informações em um formato padrão e a carga dessas informações em um único sistema. O ETL também é uma técnica comum para integração de dados em *data warehouses* e pode ser personalizado de acordo com as necessidades de cada organização.

Em todas abordagens de integração de dados, ainda há processos manuais que podem exigir grandes esforços da equipe de TI, entre elas uma das primeiras etapas, que é a descoberta e mapeamento dos dados envolvidos durante a integração, o que chamamos de *Schema Matching* [Rahm e Bernstein 2001]. Técnicas de *Schema Matching* surgem como soluções potenciais para facilitar a integração e promovendo a correspondência automática entre os esquemas de dados. Esta etapa é crucial, sendo envolvida em todo o processo de integração até mesmo na sua manutenção. Normalmente é executada de forma manual, se tornando um gargalo, e até um risco para projetos de integração [Do 2006].

O *Schema Matching* ajuda a garantir que os dados sejam integrados de forma consistente e precisa, envolvendo a comparação de elementos de esquema como tabelas, colunas e tipos de dados, para identificar correspondências entre eles. Uma vez identificadas as correspondências entre os esquemas, é possível mapear as diferenças semânticas e transformar os dados de forma a garantir a integridade das informações. Apesar de sua importância, suspeita-se que haja uma adoção limitada dessas técnicas no cenário brasileiro da indústria de software.

Ao elucidar os motivos pelos quais a indústria de software brasileira ainda não incorporou amplamente técnicas de *Schema Matching*, este trabalho contribui para o estudo de integração de dados no Brasil. Espera-se que os resultados desta pesquisa forneçam *insights* para acadêmicos, desenvolvedores de software e gestores de TI, auxiliando na superação dos desafios de integração de dados e promovendo uma inovação no setor.

## 1.1 Objetivos

Este trabalho tem como objetivo principal analisar os motivos pelos quais a indústria de software brasileira ainda não incorporou amplamente técnicas de *Schema*

*Matching* e diagnosticar os pontos críticos em projetos de integração de dados. Visa-se explorar como o *Schema Matching* pode oferecer soluções efetivas para esses desafios.

Com base nos achados de uma revisão de literatura *ad hoc* e de um mapeamento sistemático, conduziu-se análises empíricas com dados acadêmicos para avaliar a viabilidade do *Schema Matching*. O intuito dessas análises foi estabelecer um repertório de algoritmos, personalizáveis às particularidades e demandas específicas do setor, para otimizar a integração de dados. Os desfechos dessas análises foram cotejados com outras abordagens da literatura, visando ressaltar as inovações e benefícios da metodologia proposta, notadamente na mitigação de obstáculos na integração de dados e na geração de perspectivas para o campo.

Tendo em vista que os resultados evidenciaram a viabilidade do *Schema Matching* no processo de integração de dados, a fase conclusiva do estudo se dedicou a investigar o estado da prática da integração de dados na indústria de software brasileira. Nesse contexto, foi executado um *Survey* de opinião com profissionais do setor para discernir os principais fatores dessa resistência. Com essa pesquisa, almejou-se entender os obstáculos e desafios que o setor enfrenta e, por conseguinte, prover recomendações embasadas para promover a assimilação do *Schema Matching* no panorama nacional. Esse segmento do trabalho é crucial para validar a aplicabilidade das melhorias apontadas e para incentivar uma maior inserção dessas técnicas no cenário brasileiro, sob a premissa de que os resultados preliminares sugiram um impacto positivo na eficiência da integração de sistemas.

Os objetivos específicos deste trabalho são:

1. Investigar a literatura de Integração de dados e *Schema Matching* para entender o estado da arte.
2. Realizar um mapeamento sistemático para determinar o estado da arte em *Schema Matching*, destacando as técnicas e metodologias predominantes.
3. Investigar o estado da prática do *Schema Matching* na indústria de software brasileira.
4. Desenvolver um guia prático com recomendações para incentivar a adoção de *Schema Matching*.

## 1.2 Justificativa

A integração de dados desempenha um papel crucial na era da informação, permitindo que sistemas distintos compartilhem e processem informações de forma eficaz [Doan, Halevy e Ives 2012]. Tradicionalmente, este processo inclui etapas manuais, especialmente na correspondência de esquemas (*Schema Matching*), onde a identificação

manual de atributos envolvidos na integração demanda um esforço considerável da equipe de TI durante o ciclo de vida da aplicação integradora. Esta abordagem não apenas atrasa o desenvolvimento de soluções de integração de dados, mas também aumenta a probabilidade de erros, comprometendo a qualidade e a eficiência da integração.

O uso de métodos automatizados para o *Schema Matching* promete uma transformação significativa neste cenário [Rodrigues e Silva 2021], reduzindo as intervenções manuais e permitindo o monitoramento eficiente das variações e seus estágios de crescimento. A implementação de correspondência automática de esquemas tem o potencial de agilizar o desenvolvimento de soluções de integração, melhorando a eficiência operacional e a qualidade dos dados integrados.

No entanto, apesar dos avanços tecnológicos e dos benefícios evidentes das técnicas automatizadas de *Schema Matching*, sua adoção na indústria de software brasileira parece limitada. Este fato suscita questões importantes sobre as barreiras à implementação dessas tecnologias no contexto brasileiro. Portanto, este trabalho visa explorar por meio de um *survey* direcionado a profissionais da indústria, os fatores que contribuem para a baixa adesão a estas técnicas no Brasil. A compreensão dessas barreiras é fundamental para elaborar recomendações estratégicas que possam incentivar a adoção de práticas de *Schema Matching*, superando os desafios existentes e aproveitando plenamente o potencial dessas tecnologias para aprimorar a integração de dados na indústria de software brasileira.

O trabalho segue estruturado a partir do Capítulo 2, que versa sobre os desafios da integração de dados, destacando-se a importância crítica do *Schema Matching* como solução. Avança-se para o Capítulo 3, onde é realizada uma revisão abrangente sobre o uso atual da Inteligência Artificial (IA) no âmbito do *Schema Matching*, revelando as tendências e inovações mais recentes. Este estudo foi motivado pela nossa revisão bibliográfica inicial, na qual foi identificado um número significativo de pesquisas recentes que exploram o uso de técnicas de IA no *Schema Matching*. A sequência da investigação leva ao Capítulo 4, onde é apresentado o *survey* desenvolvido especificamente para este trabalho e pela análise detalhada dos resultados deste *survey*. Finalmente, no Capítulo 5, são reunidas as conclusões, refletindo sobre os *insights* obtidos e as implicações para a indústria de software e a comunidade acadêmica. Esta estrutura não apenas facilita uma compreensão clara da investigação realizada, mas também sublinha a contribuição do estudo para o campo da integração de dados e *Schema Matching*.

CAPÍTULO 2

# Desafios da Integração de Dados e a Contribuição do *Schema Matching*

Este capítulo tem como objetivo oferecer uma introdução aos conceitos fundamentais de integração de dados, *Schema Matching* e sua importância na integração de dados. Além disso, serão abordada algumas técnicas para aplicar o *Schema Matching*, incluindo a utilização de inteligência artificial e medidas de similaridade.

## 2.1 Integração de Dados

A integração de dados refere-se ao processo de combinar dados provenientes de diferentes fontes e sistemas, consolidando-os em um único local ou sistema. O objetivo principal da integração de dados é garantir a consistência, a precisão e a disponibilidade dos dados para uso em toda a organização [Doan, Halevy e Ives 2012], [Sreemathy et al. 2021]. Isso envolve a harmonização de formatos, a reconciliação de valores discrepantes, a eliminação de duplicações e a padronização dos dados.

### 2.1.1 Integração de Dados vs Integração de Sistemas

A integração de dados e a integração de sistemas, esta última conhecida na literatura como *Enterprise Application Integration* (EAI) [Hohpe e Woolf 2004], são conceitos relacionados, mas apresentam diferenças significativas em termos de escopo e foco. A integração de dados concentra-se na união e harmonização dos dados, enquanto a integração de sistemas concentra-se na conexão e comunicação entre os sistemas, podendo ser através do compartilhamento de dados ou de funcionalidades.

Na maioria dos casos, para integrar sistemas de forma eficaz, é necessário integrar os dados que esses sistemas utilizam e produzem. Quando diferentes sistemas estão envolvidos, eles geralmente têm suas próprias estruturas de dados, formatos de dados e regras de negócio específicas. A integração de sistemas requer a troca de

informações entre esses sistemas, garantindo que os dados sejam compartilhados e interpretados corretamente por todos os sistemas envolvidos.

Para que a troca de informações entre os sistemas seja bem-sucedida, é essencial que haja uma harmonização e mapeamento dos dados entre os diferentes formatos e estruturas usados pelos sistemas. Isso envolve a integração dos dados, onde as diferenças de esquemas e formatos são tratadas para permitir a comunicação adequada entre os sistemas.

Portanto, a integração de sistemas muitas vezes requer a integração de dados, garantindo que os sistemas possam compartilhar e utilizar as informações necessárias para seu funcionamento integrado. A integração de dados é um componente crítico para a integração bem-sucedida de sistemas.

### 2.1.2 Arquitetura de Referência para Integração de Dados

[Giordano 2010] define um processo de integração de dados desenvolvido ao longo do tempo por meio da observação de ambientes de aplicativos de integração de dados de alto desempenho e experiência no campo em projetar, construir e manter dados grandes e complexos ambientes de aplicativos de integração. De acordo com o autor, esta arquitetura de referência de integração de dados foi desenvolvida para garantir dois objetivos principais: simplicidade e escalabilidade. Na Figura 2.1 é detalhada a arquitetura, onde as etapas em preto são etapas do processo, e etapas em azul são áreas de armazém de dados da etapa anterior.

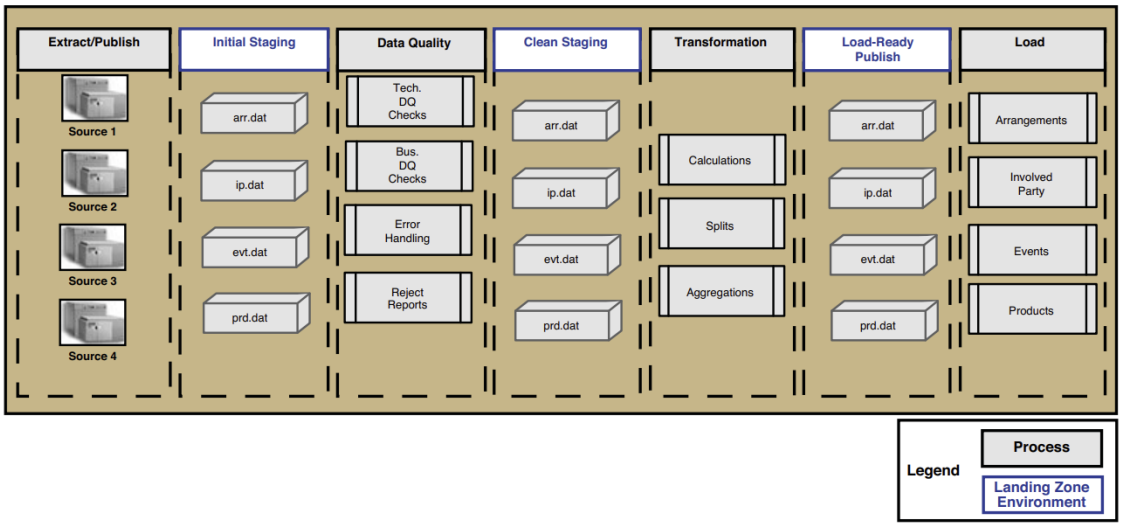


Figura 2.1: Arquitetura de referência de integração de dados [Giordano 2010].

### Processo de Integração de Dados

Abaixo são descritas as etapas do processo extraído da [Giordano 2010] conforme ilustrado na Figura 2.1.

**(i) *Extract/Publish*** (Extração de dados): Nesta etapa, os dados são coletados a partir das diferentes fontes, sejam elas bancos de dados, planilhas, arquivos de texto ou sistemas legados. Isso envolve a identificação das fontes de dados relevantes e a extração dos dados de forma estruturada.

**(ii) *Data Quality*** (Análise da qualidade dos dados): Os dados extraídos geralmente precisam ser limpos e qualificados para garantir a qualidade e consistência. Isso inclui a remoção de dados inválidos, a correção de erros, a padronização de formatos e a transformação de dados para uma representação comum, quando necessário.

**(iii) *Transformation*** (Transformação): É necessário realizar a transformação dos dados para que sejam compatíveis e possam ser integrados de forma consistente. Isso pode envolver a conversão de formatos, a agregação de dados, a normalização e a padronização de valores, entre outras operações. Nesta etapa ocorre o *Schema Matching*, envolvendo a comparação dos elementos de esquema, como tabelas, colunas e tipos de dados, a fim de identificar correspondências entre eles. Isso permite mapear as diferenças semânticas entre os esquemas e realizar a integração adequada dos dados.

**(iv) *Load*** (Carga): é um conjunto de etapas padronizadas em que se realiza a carga dos dados tratados.

## 2.2 Esquemas

De acordo com [Rahm e Bernstein 2001], os esquemas referem-se às estruturas de metadados que descrevem as características e a organização dos dados em diferentes fontes de informação, podendo ser representados em diferentes formatos, como diagramas ER (Entidade-Relacionamento), modelo Orientado a objetos (OO), XML (Extensible Markup Language), JSON (JavaScript Object Notation) ou DTD (Document Type Definition). Cada fonte de informação pode ter seu próprio esquema que descreve as estruturas dos dados.

Embora existam diversos tipos de esquemas amplamente utilizados na integração de dados, será explorado esquemas de banco de dados, particularmente no modelo relacional, onde os esquemas definem a estrutura, as restrições e as relações dos dados armazenados. Abaixo estão descritos alguns aspectos importantes dos esquemas em banco de dados, com foco particular no modelo relacional:

**Estrutura dos dados:** Os esquemas definem a estrutura dos dados em um banco de dados relacionais, especificando as entidades (tabelas), os atributos (colunas) e os

relacionamentos entre eles. Por exemplo, em um banco de dados de uma universidade, pode haver um esquema que define a entidade “Aluno” com atributos como “nome”, “matricula”, “endereço” e “email”[1].

**Tipos de dados:** Os esquemas também especificam os tipos de dados que podem ser armazenados em cada atributo. Por exemplo, um atributo pode ser do tipo texto, numérico, data, etc. Essas definições ajudam a garantir a consistência e a integridade dos dados armazenados.

**Restrições e validações:** Os esquemas permitem definir restrições e validações nos dados para garantir a qualidade e a consistência dos dados armazenados. Isso inclui restrições de chave primária, restrições de chave estrangeira, restrições de integridade referencial, restrições de domínio, entre outras.

**Relacionamentos:** Os esquemas definem os relacionamentos entre as entidades em um banco de dados. Isso inclui relacionamentos um-para-um, um-para-muitos e muitos-para-muitos. Os relacionamentos são estabelecidos por meio de chaves primárias e chaves estrangeiras, que são especificadas nos esquemas.

**Documentação e compreensão dos dados:** Os esquemas fornecem uma documentação clara e estruturada dos dados armazenados em um banco de dados. Eles ajudam os desenvolvedores, administradores de banco de dados e usuários a entender a estrutura e a semântica dos dados.

## 2.3 Operador Match

O operador Match é definido como uma função que recebe dois esquemas S1 e S2 como entrada e retorna um mapeamento entre esses dois esquemas como saída. Cada elemento de mapeamento do resultado do operador especifica que certos elementos do esquema S1 correspondem logicamente a certos elementos de S2. Os critérios para estabelecer essas correspondências baseiam-se em heurísticas. Essas heurísticas não são expressas como fórmulas matemáticas determinísticas, mas são aplicadas para guiar a implementação do processo de correspondência entre os esquemas [Rahm e Bernstein 2001].

## 2.4 Schema Matching

O *Schema Matching* pode ser descrito como uma técnica utilizada para identificar e relacionar elementos semânticos entre dois ou mais esquemas de dados a fim de integrá-los de forma eficiente. Ele é uma necessidade em projetos de integração de dados, onde múltiplas fontes de dados precisam ser combinadas em um único esquema [Do 2006].

[1] Nome dos campos em caixa baixa, sem acento e sem caracteres especiais

O objetivo do *Schema Matching* é encontrar correspondências semânticas entre elementos de dados de diferentes esquemas, identificando relações de equivalência e subordinação. Por exemplo, ele pode ajudar a identificar que um atributo “nome” em um esquema se relaciona semanticamente com um atributo “nome_completo” em outro esquema.

Durante o processo de *Schema Matching*, os esquemas das diferentes fontes de informação são comparados e analisados para identificar elementos correspondentes [Bilke e Naumann 2005], como ilustrado na Figura 2.2. Esses elementos podem incluir tabelas com nomes semelhantes, colunas com tipos de dados compatíveis, chaves primárias correspondentes, relacionamentos similares, entre outros critérios.

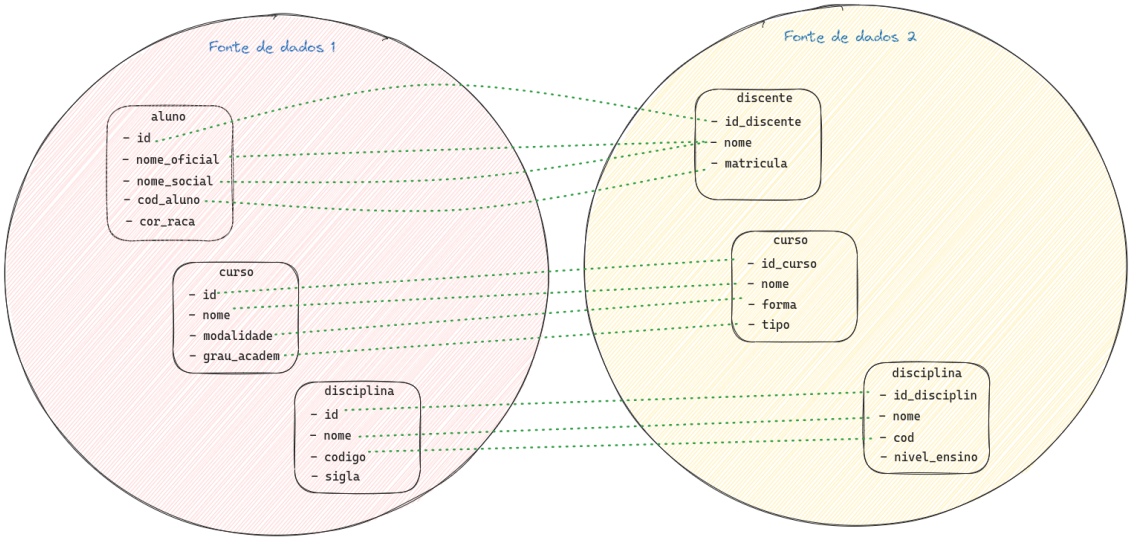


Figura 2.2: Exemplo de *Schema Matching* com dados da educação.

O *Schema Matching* é um processo crítico em projetos de integração de dados, pois pode afetar significativamente a qualidade dos dados integrados [Hai Do 2007]. Uma correspondência incorreta entre elementos semânticos pode levar a erros e inconsistências nos dados integrados. Portanto, é importante aplicar uma abordagem sistemática e rigorosa para o *Schema Matching*, a fim de garantir a integridade e a precisão dos dados integrados. Para realizar o *Schema Matching*, é possível utilizar várias técnicas, incluindo (i) comparação de valores semânticos, (ii) a análise sintática e (iii) a análise semântica. As técnicas de comparação de valores envolvem a comparação direta de valores de atributos, enquanto as técnicas de análise sintática e semântica envolvem a análise da estrutura dos esquemas e das relações semânticas entre eles.

A Figura 2.3 representa a técnica *Schema Matching* através da correspondência de quatro esquemas. São esquemas (simplificados) de bases de dados sobre informações acadêmicas de universidades, onde os elementos do esquema que representam o mesmo

conceito estão ligados por setas pontilhadas, em que os elementos podem ter correspondências em um ou mais esquemas. Por exemplo, o atributo “fname” do esquema **UNIB** é correspondido pelo atributo “NAME_FIRST” do esquema **UNIC**, “Prefix” do esquema **UNIA** e “FIRST_NAME” do esquema **UNID**.

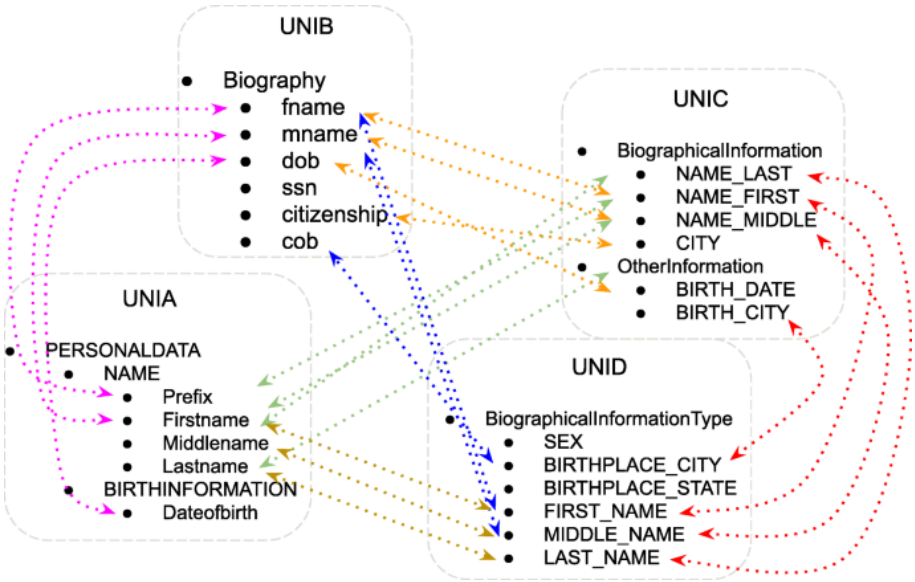


Figura 2.3: Um exemplo de *Schema Matching* retirado de [Rodrigues e Silva 2021].

Essa representação visual demonstra como o processo de correspondência de esquemas pode identificar correspondências entre elementos de diferentes esquemas, permitindo a integração adequada dos dados.

### 2.4.1 Abordagens de *Schema Matching*

Rahm e Bernstein (2001) apresentam uma classificação das principais abordagens para *Schema Matching*, quais sejam [Rahm e Bernstein 2001]:

**Instância vs esquema:** as abordagens de correspondência podem considerar dados de instância, ou seja, o conteúdo dos dados, ou apenas informações em nível de esquema.

**Correspondência de elemento vs estrutura:** a correspondência pode ser realizada para elementos de esquema individuais, como atributos, ou para combinações de elementos, como estruturas de esquema complexas.

**Linguagem vs restrição:** um *matcher* pode utilizar uma abordagem baseada em linguística, como nomes e descrições textuais de elementos do esquema, e uma abordagem baseada em restrições, como chaves e relacionamentos.

**Cardinalidade correspondente:** o resultado da correspondência pode relacionar um ou mais elementos de um esquema a um ou mais elementos do outro, resultando

em quatro casos: 1:1, 1:n, n:1, n:m. Além disso, cada elemento de mapeamento pode inter-relacionar um ou mais elementos dos dois esquemas. Também pode haver diferentes cardinalidades de correspondência no nível da instância.

**Informações auxiliares:** a maioria dos matchers depende não apenas dos esquemas de entrada S1 e S2, mas também de informações auxiliares, como dicionários, esquemas globais, decisões de correspondência anteriores e entrada do usuário.

#### Níveis das abordagens de *Schema Matching*

No contexto do *Schema Matching* em nível de esquema, são consideradas exclusivamente as informações relacionadas à estrutura e características do esquema, desconsiderando os dados de instância. Essas informações compreendem as propriedades fundamentais dos elementos do esquema, como nome, descrição, tipo de dados, tipos de relacionamento (por exemplo, “parte de” ou “é-um”), restrições e a própria estrutura do esquema. Em decorrência dessa abordagem, é comum que a correspondência de um atributo tenha potenciais candidatos que possam se correlacionar de maneira apropriada.

No que diz respeito ao *Schema Matching* em nível de instância, são levadas em consideração informações relativas aos dados em si, proporcionando *insights* sobre o conteúdo e o significado dos elementos do esquema. Esse enfoque é particularmente relevante quando as informações disponíveis no esquema são limitadas, como é frequentemente o caso em dados semiestruturados. Por exemplo, ao enfrentar ambiguidade entre correspondências igualmente plausíveis no nível do esquema, é possível utilizar as informações de instância para selecionar as correspondências cujas instâncias são mais semelhantes, a fim de reduzir a incerteza.

Além das abordagens mencionadas nesta seção e suas respectivas perspectivas, é possível combinar múltiplas estratégias visando obter resultados mais robustos. Essa combinação pode ser realizada por meio de um matcher híbrido que integra diferentes critérios de correspondência, bem como por meio de matchers compostos que combinam os resultados de matchers executados de forma independente. Essas abordagens combinadas podem oferecer maior precisão e acurácia no processo de correspondência automática de esquemas.

### 2.4.2 Métodos de aplicação do *Schema Matching*

Existem diversos métodos para aplicar o *Schema Matching*, que incluem abordagens manuais, automáticas e semi-automáticas [Rahm e Bernstein 2001]. A escolha do método adequado depende do contexto específico e dos requisitos do projeto. Os métodos comuns de aplicação do *Schema Matching* incluem:

**Método manual:** Neste método, os especialistas ou pesquisadores analisam e comparam manualmente os esquemas de dados para identificar correspondências. Isso envolve examinar a estrutura, os atributos e as semânticas dos esquemas, bem como utilizar conhecimento de domínio para identificar relacionamentos relevantes entre os elementos. Essa abordagem é mais adequada para projetos de menor escala ou quando o conhecimento especializado é fundamental.

**Método automático:** Este método utiliza algoritmos automatizados, como técnicas de Inteligência Artificial (IA), similaridade e outros métodos, para identificar correspondências entre os elementos dos esquemas de forma eficiente. Esses algoritmos são capazes de analisar as características estruturais, semânticas e de dados dos esquemas, comparando-os e encontrando padrões de correspondência. Essa abordagem permite a identificação rápida e precisa de correspondências, especialmente em projetos de grande escala, onde a análise manual seria inviável.

**Método semi-automático:** Combina o uso de técnicas automatizadas com a intervenção humana. Os especialistas podem fornecer orientações iniciais e definir regras, que são posteriormente refinadas e aprimoradas com o uso de técnicas de IA. Isso combina o conhecimento humano com a capacidade de aprendizagem automatizada.

## 2.5 Inteligência Artificial

A Inteligência Artificial (IA) é um campo da ciência da computação que se concentra no desenvolvimento de algoritmos e sistemas capazes de realizar tarefas que normalmente requerem raciocínio humano. Essas tarefas podem incluir reconhecimento de padrões, aprendizado, raciocínio, tomada de decisões e resolução de problemas complexos. Dentro da IA, o *Machine Learning* (ML) é uma abordagem fundamental que permite aos computadores aprenderem a partir de dados. O *Deep Learning*, uma subcategoria do ML, utiliza redes neurais profundas para modelar e resolver problemas complexos. Outras abordagens significativas incluem o Processamento de Linguagem Natural, entre outras. As descrições dessas técnicas são apresentadas a seguir:

1. **Aprendizado de Máquina (*Machine Learning*):** É uma abordagem central da IA que permite que os sistemas aprendam a partir de dados sem serem explicitamente programados. Os algoritmos de aprendizado de máquina são treinados com conjuntos de dados para reconhecer padrões e fazer previsões ou tomar decisões com base nesses padrões. O aprendizado de máquina pode ser supervisionado (com rótulos nos dados de treinamento), não supervisionado (sem rótulos) ou por reforço (baseado em feedback de recompensa).
2. **Redes Neurais Profundas (*Deep Learning*):** Inspiradas no funcionamento do cérebro humano, as redes neurais artificiais são estruturas compostas por neurônios

interconectados. Essas redes são capazes de aprender e generalizar informações a partir de dados de entrada, permitindo o reconhecimento de padrões complexos. São amplamente utilizadas em tarefas como visão computacional e reconhecimento de fala.

3. **Processamento de Linguagem Natural (*Natural Language Processing* - NLP):** É uma área da IA que se concentra na interação entre humanos e máquinas por meio da linguagem natural. As técnicas de NLP envolvem o processamento e a compreensão de textos e discursos, permitindo que as máquinas interpretem a linguagem humana, incluindo tarefas como tradução automática, resumo de textos, análise de sentimentos e chatbots.

Uma das vantagens mais claras de aplicar a IA no *Schema Matching* é aplicar a correspondência para um grande volume de dados, ou seja, os algoritmos de IA podem lidar com a análise e o processamento de dados em larga escala, podendo eliminar a correspondência de esquema totalmente manual.

Os algoritmos de IA podem analisar a estrutura e os padrões dos esquemas, identificando correspondências semânticas e sintáticas entre os elementos de dados. Outro benefício é a capacidade de aprendizado. Os algoritmos de ML podem ser treinados em conjuntos de dados de referência para aprender a identificar padrões de um domínio de sistema e relacionamentos entre os elementos de dados. Com o treino, esses algoritmos podem melhorar seu desempenho e precisão na correspondência de esquemas, adaptando-se a novos contextos e desafios.

Na literatura, uma variedade de abordagens de inteligência artificial têm sido aplicadas ao *Schema Matching*, conforme evidenciado em estudos como [Rodrigues e Silva 2021], [Zhang et al. 2023], [Bulygin e Stupnikov 2019], [Xue et al. 2021], entre outros. Essas abordagens incluem técnicas de aprendizado supervisionado, não supervisionado, redes neurais e processamento de linguagem natural (NLP). Cada abordagem possui características e vantagens distintas, e a escolha da técnica a ser utilizada depende do contexto e dos requisitos específicos do domínio. No próximo capítulo são demonstrados os resultados de um mapeamento sistemático realizado para identificar as técnicas de IA mais amplamente empregadas no *Schema Matching*, fornecendo uma visão abrangente do estado atual da pesquisa nesse campo. Isso permitirá uma compreensão do estado da arte sobre o uso da IA no *Schema Matching*.

## 2.6 Técnicas de *Schema Matching*, Ontologias e Similaridade

A similaridade é usada para medir a proximidade ou correspondência entre os elementos dos esquemas, como atributos, tabelas ou entidades. A similaridade é geralmente calculada com base em características específicas dos elementos, como nomes, tipos de dados, estruturas ou semânticas.

Além disso, assim como a IA, a similaridade também pode ser aplicada no contexto do *Schema Matching* conforme evidenciado em estudos como [Melnik, Garcia-Molina e Rahm 2002] e [Peukert, Massmann e Koenig 2010].

De acordo com [Le, Dieng-Kuntz e Gandon 2004], *Ontology Matching*, também conhecido como correspondência de ontologias, é uma área de pesquisa que se concentra em identificar e alinhar correspondências entre diferentes ontologias. Uma ontologia é uma representação formal e explícita de conceitos e suas inter-relações em um domínio específico.

O principal objetivo da correspondência de ontologias é mapear conceitos e relacionamentos entre duas ou mais ontologias, permitindo que sistemas que usam diferentes ontologias se comuniquem e compartilhem informações de maneira eficiente. Isso é especialmente útil quando se trata de integrar informações de diferentes fontes e domínios, como na integração de bases de dados, serviços da web ou aplicativos heterogêneos.

Embora a *Ontology Matching* e o *Schema Matching* lidem com a correspondência de elementos, o *Ontology Matching* se concentra na correspondência de conceitos e relacionamentos em ontologias, enquanto o *Schema Matching* se concentra na correspondência de elementos estruturais, como tabelas, colunas e tipos de dados, em esquemas de banco de dados ou outras fontes de dados [Uschold e Grüninger 2004]. Ao considerar os dois conceitos como sinônimos para esta pesquisa, a busca de soluções pode ser aplicável em uma variedade de contextos e cenários, abrindo oportunidades para explorar sinergias e aplicar técnicas de ambas as áreas para enriquecer os resultados.

*Schema Mapping* refere-se ao processo de relacionar ou traduzir a estrutura de dados de um esquema (*schema*) para outro. Esse conceito é fundamental em diversas áreas, incluindo bancos de dados, integração de sistemas e de dados.

Quando lidamos com diferentes fontes de dados que têm estruturas distintas, o *Schema Mapping* torna-se essencial para garantir que as informações possam ser entendidas e utilizadas de maneira consistente. Esse processo envolve a correspondência de elementos do esquema, como tabelas, atributos e relacionamentos, de uma fonte para outra. Pode ser necessário em situações como migração de dados entre sistemas, integração de bases de dados heterogêneas, ou mesmo ao realizar análises que envolvam dados de fontes diversas. Existem diferentes abordagens e técnicas para realizar o mapeamento de

esquemas, incluindo mapeamento manual, semi-automático e automático.

*Schema Matching* refere-se ao processo de identificar correspondências semânticas entre elementos de esquemas de dados provenientes de diferentes fontes. Por outro lado, o *Schema Mapping* refere-se ao processo de criar relações explícitas ou regras de correspondência entre elementos específicos de esquemas diferentes.

## 2.7 Considerações Finais do Capítulo

Neste capítulo, foram explorados diversos aspectos relacionados à integração de dados e, mais especificamente, ao processo de *Schema Matching* e compreendendo o conceito de integração de dados e sua estreita relação com a integração de sistemas. Foi discutida a arquitetura para a integração de dados e sua importância na definição de diretrizes e melhores práticas.

A compreensão sobre os esquemas de dados e o papel fundamental do operador *Match* no processo de integração foi demonstrada. O conceito de *Schema Matching* foi analisado detalhadamente, destacando-se sua relevância na reconciliação de diferenças semânticas entre os esquemas e na garantia da integridade das informações.

Explorou-se também as diferentes abordagens utilizadas no *schema matching* e como sua aplicação pode contribuir para a integração eficiente de dados. Além disso, discutiu-se o uso da inteligência artificial nesse contexto, reconhecendo sua capacidade de melhorar e automatizar o processo de correspondência de esquemas.

Por fim, destacou-se a importância do uso de medidas de similaridade no *schema matching*, como uma abordagem na identificação de correspondências entre elementos de esquema.

No próximo capítulo, será fornecido um mapeamento sistemático sobre o uso da inteligência artificial no *schema matching*, explorando os estudos e pesquisas relevantes nessa área, destacando as abordagens e técnicas mais utilizadas no emprego da inteligência artificial para aprimorar a precisão e a eficácia do processo de *schema matching*.

CAPÍTULO 3

# O Estado da Arte sobre o uso da Inteligência Artificial no *Schema Matching*

Este capítulo apresenta os resultados de um mapeamento sistemático sobre o uso da Inteligência Artificial em técnicas de *Schema Matching*[1]. A decisão de realizar este mapeamento foi motivada pelo significativo volume de estudos sobre a intersecção desses assuntos encontrados durante a revisão da literatura. Inicialmente, o foco da pesquisa era comparar algoritmos de IA para abordar o problema de *Schema Matching*. Seguindo um protocolo inspirado em [Petersen, Vakkalanka e Kuzniarz 2015], tal mapeamento busca fazer o levantamento do estado da arte e comunicar os principais tópicos e oportunidades de pesquisa.

O mapeamento sistemático consiste em uma abordagem baseada em evidência o qual tem o objetivo de buscar identificar, avaliar e sintetizar evidências relevantes de estudos existentes em uma área específica de conhecimento. Ele é utilizado para responder a questões de pesquisa, e pode ser útil para a identificação de lacunas no conhecimento para orientar o desenvolvimento de novas pesquisas e para apoiar a tomada de decisões baseadas em evidências. Para conduzir um Estudo Sistemático da Literatura (ESL), [Scannavino et al. 2017] definiu um protocolo genérico, apresentado na Figura 3.1.

[1]O conteúdo deste capítulo baseia-se na publicação: O Uso da Inteligência Artificial no Schema Matching: Um Mapeamento Sistemático [Borges, Ribeiro e Neto 2023]

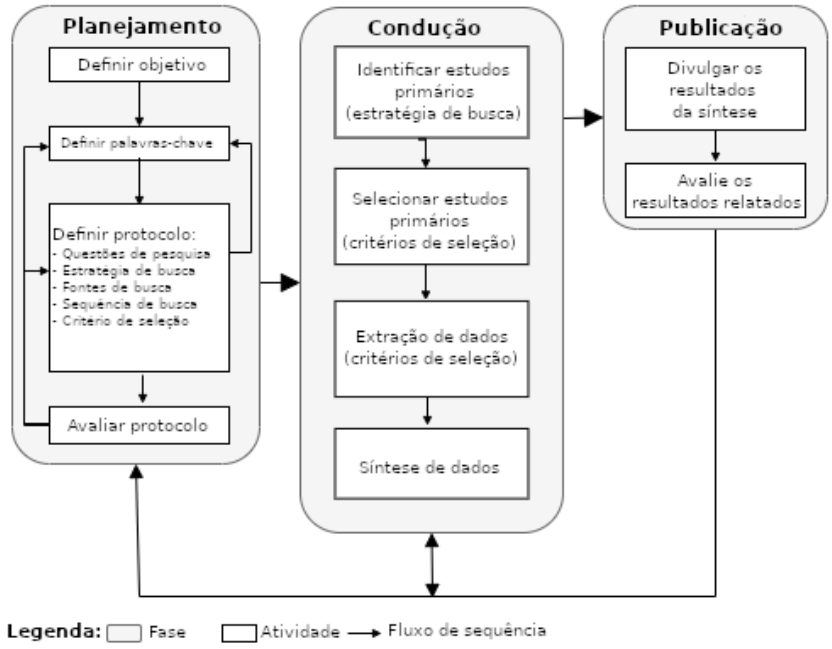


Figura 3.1: Fases e atividades do ESL, adaptado de [Scannavino et al. 2017].

O protocolo é separado em três grandes fases: Planejamento, Condução e Comunicação dos Resultados.

## 3.1 Planejamento

A fase de planejamento compreende três etapas fundamentais: Definição de Objetivo, Definição de Protocolo e Avaliação de Protocolo. Na etapa inicial, “Definição de Objetivo”, é necessário estabelecer claramente o propósito que justifica a realização da pesquisa secundária. Na segunda etapa, elaborar o Protocolo do Mapeamento Sistemático, o qual é dividido em etapas de definição que, em conjunto, constituem o protocolo de pesquisa. Na quarta etapa, denominada “Avaliação de Protocolo”, o objetivo é avaliar os passos previamente executados e efetuar ajustes no protocolo.

Ainda na primeira etapa, são incorporadas diversas subetapas fundamentais para o planejamento da pesquisa: Definição das Questões de Pesquisa, Elaboração da Estratégia de Busca, Determinação das Fontes de Pesquisa, Construção da String de Busca e Estabelecimento dos Critérios de Seleção. A seguir, é apresentada uma descrição detalhada de cada uma dessas subetapas.

### 3.1.1 Definição dos objetivos e palavras-chave

O objetivo deste mapeamento é identificar, avaliar e sintetizar as evidências relevantes de estudos existentes, facilitando uma visão abrangente do estado da arte em

técnicas ou algoritmos aplicados ao *schema matching*. Permitindo organizar e categorizar as informações encontradas, destacando lacunas de conhecimento e tendências emergentes, orientando futuras pesquisas.

As palavras-chave utilizadas para conduzir essa busca foram: *schema matching*, inteligência artificial, *machine learning* e *deep learning*.

### 3.1.2 Definição das Questões de Pesquisa

Durante a fase de planejamento, é essencial definir as questões que a pesquisa busca responder. No contexto da Inteligência Artificial, foram levantadas três questões de pesquisa, abordando cada uma das subáreas: *Deep Learning*, *Machine Learning* e NLP. Além disso, uma questão de pesquisa abrangente foi formulada para abordar a área como um todo, incluindo técnicas que não estão necessariamente ligadas a uma subárea específica. Essa abordagem garante uma cobertura ampla e abrangente das principais áreas da Inteligência Artificial, permitindo explorar de diferentes técnicas e abordagens em nossa pesquisa. Na Figura 3.2 mostra um diagrama de Venn da relação entre as subáreas da IA.

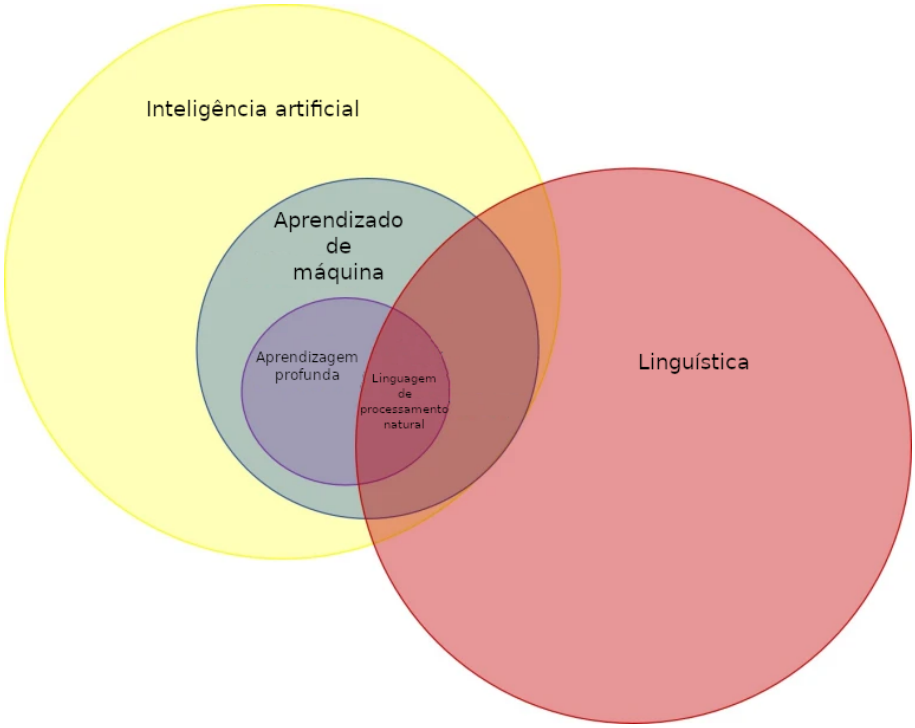


Figura 3.2: Diagrama de Veen de subáreas da Inteligência Artificial [MUYLAERT 2020].

No diagrama da Figura 3.2, é possível observar que *Deep Learning* está situado dentro do domínio do *Machine Learning*, uma vez que *Deep Learning* é uma abordagem específica que se baseia em redes neurais profundas para o aprendizado de máquina.

No entanto, o Processamento de Linguagem Natural (NLP) é representado como uma área separada, pois é um campo especializado que se concentra exclusivamente no processamento e compreensão da linguagem natural. Embora o NLP utilize técnicas de Inteligência Artificial para seu funcionamento, é importante reconhecer sua singularidade devido ao foco específico na linguagem e em desafios linguísticos complexos. A seguir as questões de pesquisa (QP) levantadas durante esta fase e sua justificativa.

**QP1:** ***Quais técnicas de Deep Learning são utilizadas na aplicação de schema matching?***
**Justificativa:** Pretende-se conhecer as técnicas de *Deep Learning* utilizadas no *schema matching* a fim de extrair conhecimento através dos dados, através de técnicas de estudo, comparações, etc.

**QP2:** ***Quais técnicas de Machine Learning têm sido utilizadas na aplicação de schema matching?***
**Justificativa:** Pretende-se conhecer as técnicas de *Machine Learning* (que não são *Deep Learning*) utilizadas no *schema matching*, afim de entender quais técnicas especificamente de *Machine Learning* foram aplicadas na técnica de *schema matching*.

**QP3:** ***Quais técnicas de NLP têm sido utilizadas na aplicação de schema matching?***
**Justificativa:** Pretende-se conhecer as técnicas de NLP utilizadas no *schema matching*, afim de entender quais técnicas especificamente de NLP foram aplicadas na técnica de *schema matching*.

**QP4:** ***Quais técnicas de IA têm sido utilizadas na aplicação de schema matching?***
**Justificativa:** Pretende-se conhecer as técnicas de Inteligência Artificial utilizadas no *schema matching*, afim de entender de forma abrangente quais outras técnicas que não estão no domínio da *Deep Learning* e *Machine Learning* que estão sendo aplicadas na técnica de *schema matching*.

### 3.1.3 Definição da Estratégia de Busca

Para este trabalho optou-se por adotar uma estratégia de busca automática, combinada com a técnica do *Snowballing* para os estudos selecionados. Essa abordagem nos permitirá explorar de forma mais detalhada os estudos relevantes e identificar referências adicionais por meio da análise de citações e referências cruzadas.

### 3.1.4 Definição das Fontes de Pesquisa

A definição das fontes envolve a seleção das bases de estudos indexados que serão utilizadas para realizar a pesquisa. Para este mapeamento sistemático em particular,

optou-se por utilizar quatro bases de dados renomadas: Scopus [2], IEEEXplore [3], ACM [4] e Engineering Village [5]. Essas bases foram escolhidas devido à sua abrangência, reputação e pertinência ao tema de estudo, fornecendo assim um conjunto abrangente de estudos para análise [Dyba, Kitchenham e Jorgensen 2005].

### 3.1.5 Definir String de Busca

Para a concepção da string de busca, optou-se por incluir os termos que são pertinentes às questões de pesquisa, ou seja, aqueles relacionados às palavras-chave: inteligência artificial, *Deep Learning*, *Machine Learning* e *schema matching*. Além disso, foi incluido de sinônimos para ampliar a abrangência dos resultados.

A string de busca definida para este trabalho foi:

(“deep learning”) OR (“Machine Learning”) OR (“Artificial Intelligence”)) AND ((“schema matching”) OR (“ontology matching”) OR (“ontology alignment”)

### 3.1.6 Definir Critérios de seleção

Os critérios de inclusão (**CI**) e critérios de exclusão (**CE**) definidos foram:

**CI1:** O estudo apresenta algoritmos ou técnicas de *Deep Learning* aplicado a *schema matching*/*Ontology Matching*/*Ontology Alignment*;

**CI2:** O estudo apresenta algoritmos ou técnicas de *Machine Learning* aplicado a *schema matching*/*Ontology Matching*/*Ontology Alignment*;

**CI3:** O estudo apresenta algoritmos ou técnicas de NLP aplicado a *schema matching*/*Ontology Matching*/*Ontology Alignment*;

**CI4:** O estudo apresenta conjunto de algoritmos ou técnicas IA aplicado a *schema matching*/*Ontology Matching*/*Ontology Alignment*.

**CE1:** O estudo não é um estudo primário;

**CE2:** O estudo não está disponível para acesso gratuito;

**CE3:** O estudo não está escrito em Inglês ou Português;

**CE4:** O estudo não foi publicado nos últimos 5 anos;

**CE5:** O estudo não apresenta algoritmos ou técnicas de Inteligencia Artificial aplicado a *schema matching*/*Ontology Matching*/*Ontology Alignment*;

**CE6:** O estudo não é um artigo.

[2] https://www.scopus.com
[3] https://ieeexplore.ieee.org
[4] https://www.acm.org
[5] https://www.engineeringvillage.com

## 3.2 Condução

Para apoio à condução, optou-se pelo uso da ferramenta Parsif.al. A escolha dessa ferramenta baseou-se em sua capacidade de facilitar a condução da pesquisa. Através do uso da ferramenta Parsif.al, foi possível otimizar o processo de condução da pesquisa, garantindo assim resultados confiáveis e de alta qualidade. A seguir, é apresentada uma descrição detalhada de cada subetapa da condução.

### 3.2.1 Identificação dos Estudos

Nessa etapa, aplicou-se a string de busca nas bases de estudos selecionadas, visando encontrar os estudos relevantes ao tema. Durante o desenvolvimento dessa etapa, calibrou-se a string de busca através de ajustes e refinamentos com base nos resultados obtidos e na análise dos estudos encontrados. Por meio desses ajustes, buscou-se ampliar a precisão e a abrangência da busca, garantindo a inclusão dos estudos mais pertinentes. Ao longo desse processo, foi possível observar a evolução da string de busca, à medida que foi-se refinando os termos-chave, considerando sinônimos, variações linguísticas e acrônimos relevantes para a nossa área de estudo. A seguir é apresentada a evolução da string de busca:

**Versão 1:** (“deep learning”) AND ((“schema matching”) OR (“ontology matching”) OR (“entity matching”));

**Versão 2:** (“deep learning”) AND ((“schema matching”) OR (“ontology matching”));

**Versão 3:** ((“deep learning”) OR (“deep neural network”)) AND ((“schema matching”) OR (“ontology matching”));

**Versão 4:** ((“deep learning”) OR (“machine learning”) OR (“artificial intelligence”)) AND ((“schema matching”) OR (“ontology matching”) OR (“ontology alignment”)).

Nesta etapa também podem ser necessários pequenos ajustes na string de busca, pois cada base possui comportamentos diferentes mediante ao formulário pesquisa.

### 3.2.2 Seleção dos Estudos

Nesta etapa, foi realizada a extração dos estudos obtidos por meio da aplicação da string de busca e, em seguida, aplicado os critérios de seleção, que estão divididos em critérios de inclusão (CI) e critérios de exclusão (CE). Estudos que atenderam pelo menos um critério de inclusão foram incluídos na seleção inicial e os estudos que atendem a pelo menos um critério de exclusão foram excluídos da seleção inicial. Inicialmente, foram aplicados os critérios de exclusão de limitar os estudos aos últimos 10 anos. No entanto,

percebe-se que o tema em estudo tem passado por uma evolução rápida, especialmente a partir dos anos de 2018 e 2019, com o surgimento de estudos que aplicam diversas técnicas inovadoras. Diante dessa percepção, foi reduzido o escopo do mapeamento para os últimos 5 anos. Essa redução tem como objetivo identificar os estudos mais recentes que possam fornecer uma contribuição significativa durante esse período de avanços. Dessa forma, buscou-se garantir que a análise inclua as pesquisas mais relevantes e atualizadas, refletindo as tendências e os desenvolvimentos recentes na área.

A partir da extração inicial de 644 estudos, foram retirados os duplicados (151 estudos), assim, realizando a leitura de todos os títulos, resumos e até conclusões de cada estudo, a fim de identificar os estudos que realmente atenderam o tema proposto. Após isso, foi realizada a seleção inicial dos estudos para realizar a leitura completa, nesta seleção foram extraídos inicialmente 82 estudos, porém com a leitura completa percebeu-se que havia estudos que fugiam do tema proposto. Desta forma, eles foram excluídos, restando ao final 59 estudos.

No dia 05 de janeiro de 2023, foi realizada uma busca que inseriu todos os estudos extraídos das bases indexadoras, marcados como aceitos ou não aceitos. Posteriormente, foram identificados os mesmos dados da planilha anterior, incluindo uma coluna de identificação para cada estudo extraído das bases indexadoras. Nesse processo, também foram localizados todos os estudos duplicados, bem como os estudos excluídos da seleção inicial. Uma seleção inicial foi feita de forma incremental, adicionando 10 estudos por dia à análise. Após essa etapa, houve a seleção de 59 estudos considerados finais. Por fim, foi realizada a extração de dados que respondem às questões de pesquisa, completando assim o ciclo de triagem e análise dos estudos coletados. No apêndice A estão listados os estudos incluídos com seu identificador atribuído durante a seleção de forma cronológica.

Nesta extração, respondeu-se a todas questões de pesquisa sendo realizado a segregação destes estudos através de suas respostas para cada questão de pesquisa.

Para sintetizar os dados, foi utilizado planilhas para criar gráficos com base nos dados obtidos, em resposta às questões de pesquisa definidas. Além disso, também foi identificado outros tipos de gráficos que agregam valor à análise, permitindo uma compreensão mais abrangente dos resultados obtidos. Essas representações visuais fornecem insights e facilitam a interpretação dos dados, contribuindo para uma melhor compreensão e apresentação dos resultados obtidos neste estudo.

Para identificar estudos relevantes que não foram encontrados por meio de buscas convencionais usando apenas uma busca bibliográfica padrão foi aplicado o método *Snowballing*. Onde os estudos identificados no mapeamento sistemático foram utilizados para encontrar outros estudos relevantes que não foram encontrados na busca inicial. Foi feito examinando as referências bibliográficas dos estudos identificados e também por

meio da busca de estudos que citam os estudos identificados.

Com a utilização do método *Snowballing*, foi possível identificar 26 novos estudos relevantes que não haviam sido incluídos no mapeamento sistemático inicial. Esses 26 estudos foram submetidos ao mesmo protocolo de análise do mapeamento sistemático e, ao final, apenas 9 foram considerados aptos para o trabalho em questão, totalizando **68 estudos**. A aplicação do método *Snowballing* permitiu a ampliação da base de estudos e a identificação de novas fontes relevantes para a pesquisa, evidenciando a importância de estratégias de busca adicionais para garantir uma revisão da literatura completa e abrangente.

## 3.3 Resultados e Discussão

Nesta seção, é apresentado os resultados em termos quantitativos e qualitativos do protocolo do mapeamento sistemático estabelecido e apresentado a discussão dos dados extraídos de acordo com as questões de pesquisa.

**Quais técnicas ou algoritmos de *Deep Learning* têm sido utilizadas na aplicação de *schema matching*?**

Entre as técnicas ou algoritmos da *Deep Learning* mais comuns encontradas, destacam-se as Redes Neurais Siameses, como mencionado nos estudos [Iyer, Agarwal e Kumar 2020], [Iyer, Agarwal e Kumar 2021], [Srinivas, Gale e Dolby 2018], [Sun, Takeuchi e Yamasaki 2020], [Chen et al. 2021], [Iyer, Agarwal e Kumar 2020] e [Xue et al. 2021]. Redes Neurais Convolucionais (CNN), como mencionado em estudos como [Bento, Zouaq e Gagnon 2020] e [Wang et al. 2022]. Também as Redes de Memória de Curto Prazo Longa (LSTM) como [Jiang e Xue 2020], [Maji, Rout e Choudhary 2021], [Sun e Shen 2022], [Mohamed et al. 2022], [Chakraborty et al. 2021], [Shraga e Gal 2022] e [Koutras et al. 2020]. Além disso, o Multi-layer Perceptron (MLP) foi identificado em estudos como [Xue et al. 2021], [Khan e Gubanov 2020], [Shraga, Gal e Roitman 2020] e [Bento, Zouaq e Gagnon 2020]. Rede Neural Recorrente (RNN) [Sun e Shen 2022], [Mohamed et al. 2022] e [Koutras et al. 2020]. Gráfico de Redes Convolucionais (GCN) os estudos [Wang et al. 2022], [Hao et al. 2021] e [Jurisch e Igler 2019]. Também foram encontrados diversos outros algoritmos de *Deep Learning*, incluindo Bidirectional LSTM o estudo [Zhang et al. 2021], Rede Neural Recursiva (RvNNs) o estudo [Chakraborty et al. 2021], Aprendizado Competitivo o estudo [Xue et al. 2021] e por ultimo o Multi-Input [Hulsebos et al. 2019]. Veja na Figura 3.3.

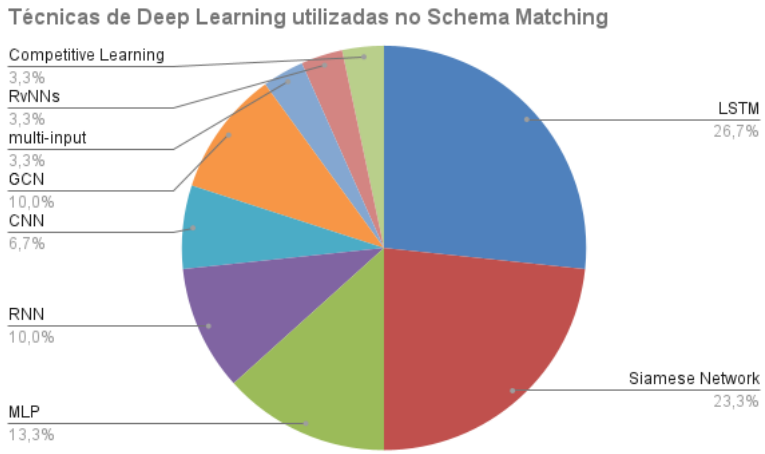


Figura 3.3: Técnicas de *Deep Learning* utilizadas no *schema matching*

**Quais técnicas ou algoritmos de *Machine Learning* têm sido utilizadas na aplicação de *schema matching*?**

A Figura 3.4 apresenta uma variedade de técnicas e algoritmos de *Machine Learning* identificados. Entre as técnicas mais comuns, destacam-se o classificador Naive Bayes, mencionado em estudos como [Lima et al. 2020], [Bulygin 2018], [Xue, Chen e Liu 2021], [Laadhar et al. 2019], [Schmidts et al. 2019], [Berlin e Motro 2002] e [Nikovski et al. 2012]. A árvore de decisão (Decision Tree), citada nos estudos [Lima et al. 2020], [Amrouch, Mostefai e Fahad 2016], [Nezhadi, Shadgar e Osareh 2011], [Nezhadi, Shadgar e Osareh 2011] e [Rodrigues et al. 2015]. A floresta aleatória (Random Forest), discutida em [Bulygin e Stupnikov 2019], [Lima et al. 2020], [Nkisi-Orji et al. 2019] e [Rodrigues e Silva 2021]. Além disso, outros algoritmos de *Machine Learning*, como k-means citado em [Jiménez-Ruiz et al. 2018], [Sahay, Mehta e Jadon 2020], [Belhadi et al. 2023] e [Li, Liu e Zhang 2005]. E outros como JRip, SVM e C4.5 (algoritmo para construção de árvores de decisão) mencionados em [Laadhar et al. 2019] e [Nezhadi, Shadgar e Osareh 2011]; Logistic Regression e Gradient Boosting, citados em [Bulygin e Stupnikov 2019], [Lima et al. 2020] e [Bulygin 2018]; Adaboost, citado em [Nezhadi, Shadgar e Osareh 2011] e [Lima et al. 2020]; K-Nearest citados em [Lima et al. 2020] e [Nezhadi, Shadgar e Osareh 2011]; Gaussian mixture, o [Przyborowski et al. 2021]; algoritmo evolutivo [Xue, Chen e Ren 2019]; e por fim o Kohonen Self-Organizing Map, o estudo [Sahay, Mehta e Jadon 2020]. Os resultados desses estudos evidenciam que há a predominancia do uso de algoritmos *Naive Bayes*, *Decision Tree* e *Random Forest* em aplicações de *schema matching*.

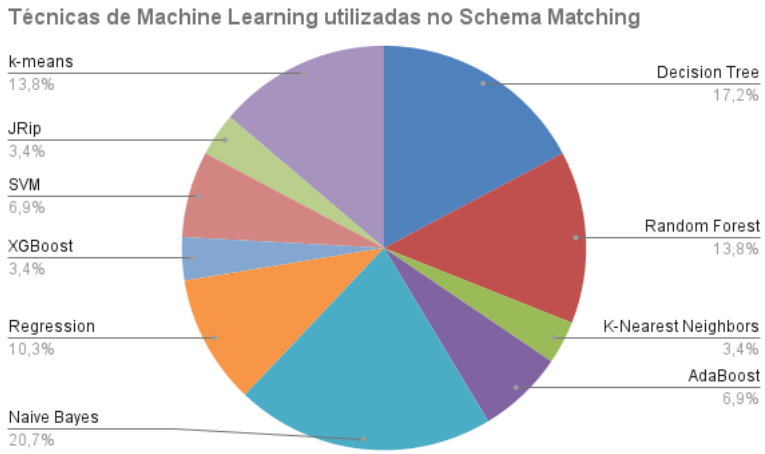


Figura 3.4: Técnicas de *Machine Learning* utilizadas no *schema matching*

**Quais técnicas ou algoritmos de NLP têm sido utilizadas na aplicação de *schema matching*?**

A Figura 3.5 relata diversas técnicas e algoritmos de Processamento de Linguagem Natural (NLP). Dentre as técnicas mais comuns encontradas, destacam-se o BERT (Bidirectional Encoder Representations from Transformers) um modelo de linguagem pré-treinado baseado na arquitetura Transformers. Estudos como [Maji, Rout e Choudhary 2021], [He et al. 2021], [Yorsh et al. 2022], [Neutel e Boer 2021], [Pan, Pan e Monti 2022] e [He et al. 2022] exploraram o uso do BERT para identificar correspondências entre esquemas de dados heterogêneos, aproveitando a semântica subjacente dos dados.

Além do BERT, outros algoritmos de NLP também foram identificados. Por exemplo, o Word2Vec foi utilizado em estudos como [Hertling, Portisch e Paulheim 2020], [Bulygin 2018], [Teslya e Savosin 2019], [Li 2020], [Li 2020], [Nkisi-Orji et al. 2019], [Nozaki, Hochin e Nomiya 2019], [Pan, Pan e Monti 2022], [Li 2020], [Chen et al. 2021], [Jurisch e Igler 2019], [Cappuzzo 2020], [Koutras et al. 2020], [Hättasch et al. 2022] e [Zhang et al. 2014], enquanto o FastText foi mencionado em estudos como [Yorsh et al. 2022], [Dhouib, Zucker e Tettamanzi 2019], [Pan, Pan e Monti 2022] e [Cappuzzo 2020]. GloVe os estudos [Ayala et al. 2022], [Pan, Pan e Monti 2022], [Cappuzzo 2020] e [Hättasch et al. 2022]. TransE os estudos [Jurisch e Igler 2019], [Li et al. 2019]. StarSpace [Jiménez-Ruiz et al. 2018] e [Jiménez-Ruiz et al. 2018]. Byte-Pair Encoding (BPE) e abordando também o uso de análise de sentimentos, o estudo [Zhang et al. 2021]. E técnicas de embedding sem especificar o algoritmo, os estudos [Iyer, Agarwal e Kumar 2020], [Iyer, Agarwal e Kumar 2021], [Iyer, Agarwal e Kumar 2020] e [Li et al. 2021].

Na grande maioria dos estudos, foi demonstrado que o uso de algoritmos de NLP pode ser uma possibilidade para melhorar a qualidade e a precisão de sistemas que lidam com dados heterogêneos.

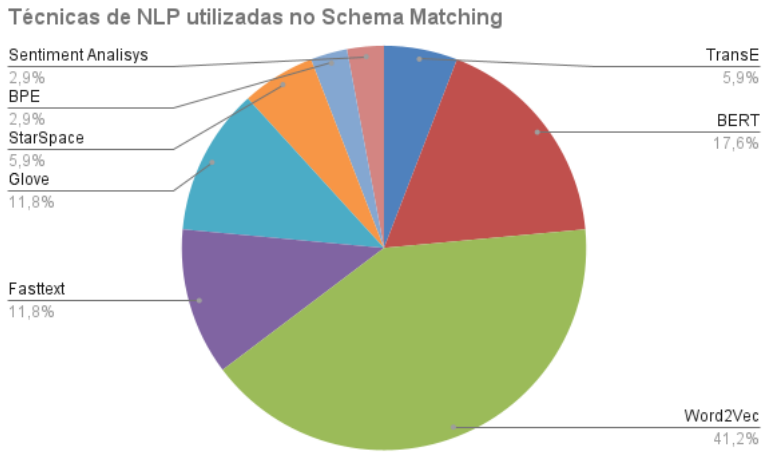


Figura 3.5: Técnicas de NLP utilizadas no *schema matching*

**Quais técnicas ou algoritmos de Inteligência artificial têm sido utilizadas na aplicação de *schema matching*?**

A Figura 3.6 mostra que foi encontrado também estudos isolados, onde utilizaram algoritmos mais generalistas da própria Inteligência Artificial. Em alguns desses estudos, não foi especificado qual algoritmo foi utilizado, enquanto outros mencionaram apenas a subárea (Machine Learning ou *Deep Learning*) sem detalhar o algoritmo específico, como os estudos [Laadhar et al. 2019], [AISSAOUI e OUGHDIR 2020], [Doan et al. 2020], [Mukherjee et al. 2021] e [Shraga 2022]. *Deep Learning*, [Cappuzzo 2020]. Grasshopper Algorithm, [Lv 2022]. Enfim, o Artificial Bee Colonies [Rangel et al. 2015]. Além disso, há casos em que foram utilizados algoritmos que não se enquadram nessas subáreas mencionadas.

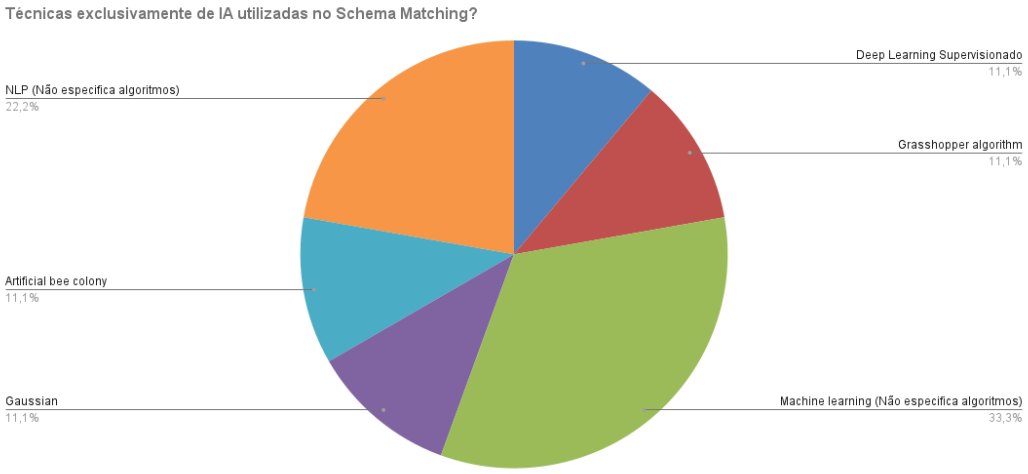


Figura 3.6: Técnicas de IA utilizadas no *schema matching*

### 3.3.1 Discussão

Durante a análise dos estudos, observou-se a importância e o amplo uso de IA aplicada no *schema matching* para a integração de dados. A complexidade e o custo de identificar correspondências em uma massa de dados grande pode expandir significativamente o impacto de uma falhas, tornando o custo e o tempo necessários para resolvê-lo um problema relevante, principalmente quando executada de forma manual. Embora haja a necessidade da automatização desta correspondência, é possível perceber pela Figura 3.7 que nos últimos 5 anos houve um pico de estudos até o ano 2020 e um decréscimo a partir de 2021, o que é inesperado.

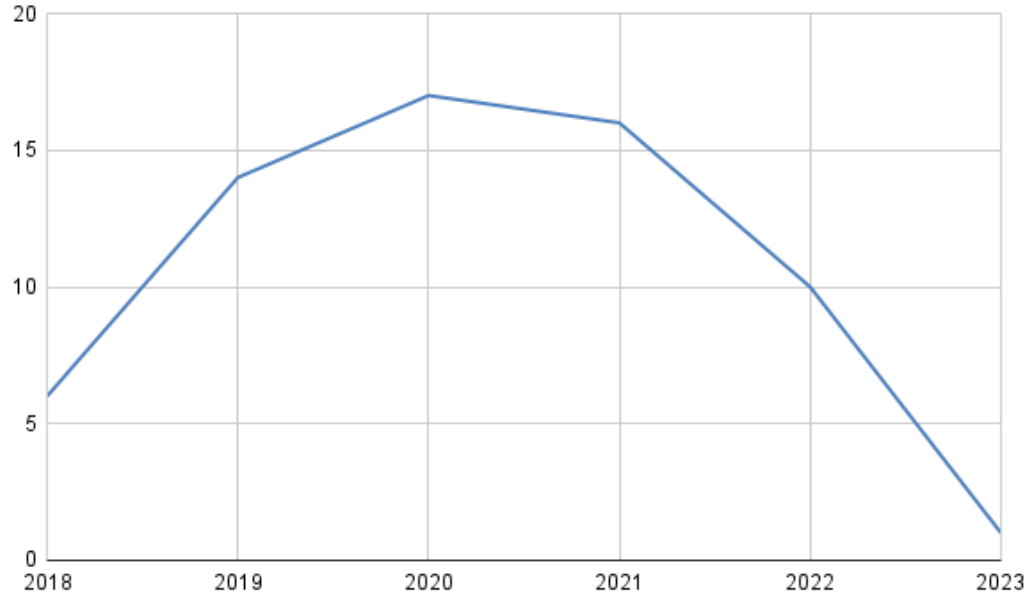


Figura 3.7: Distribuição de estudos por ano

Além disso, ao analisar os dados extraídos do mapeamento, constata-se a ampla variedade de técnicas de Inteligência Artificial (IA) aplicadas no *schema matching*, como algoritmos de aprendizado de máquina, *Deep Learning* e NLP. Sendo uma grande tendência o uso de técnicas de NLP.

Também observou-se a presença de abordagens como similaridade e análise de ontologias semânticas junto com técnicas de IA. Essa diversidade de técnicas demonstra que ainda não existe uma única solução ideal para o problema de *schema matching*. Pelo contrário, é possível aplicar uma combinação de várias técnicas, adaptando-as ao contexto específico e às necessidades do projeto em questão. A escolha da abordagem mais adequada dependerá das características dos dados, dos requisitos do problema e das metas de qualidade estabelecidas.

## 3.4 Ameaças à Validade

Ameaças à validade deste estudo foram identificadas, e foram categorizados de acordo com [Hyman 1982] e [Wohlin et al. 2012]:

**Validade da construção:** Pode haver uma possível exclusão de estudos relevantes. Para mitigar esse problema, foi realizada revisões de cada etapa de condução e extração de dados deste estudo mais de uma vez. Também houve a necessidade de realizar ajustes na string de busca de acordo com a base, pois há divergências no tratamento das combinações de operadores lógicos com parênteses de forma específica em cada base.

A falta da especificidade do tema também foi uma ameaça encontrada, assim, encontra-se estudos com foco em técnicas e algoritmos de IA aplicado ao *schema matching*, porém sem especificar o algoritmo utilizado. Sendo assim, é alcançado a transposição destes resultados identificando pontos aplicados ao *schema matching*.

**Validade Interna:** Possíveis ameaças internas podem ter surgido devido aos métodos de busca escolhidos. Por exemplo, a opção de não incluir algumas bibliotecas digitais pode levar à exclusão de estudos relevantes e ao número relativamente baixo de estudos incluídos. Para mitigar essa ameaça, o protocolo foi previamente avaliado pelos orientadores para identificar possíveis erros.

**Validade Externa:** Questões externas como a indisponibilidade de estudos foram resolvidas por meio de pesquisa utilizando a Portal de Periódicos da CAPES [6], também utilizou-se a literatura cinza como Google Acadêmico [7], Google [8], Researchgate [9].

**Validade de Conclusão:** Possíveis ameaças estão relacionadas ao viés durante a condução e extração de dados, o que pode causar imprecisão na extração de dados, ameaçando a conclusão dos resultados do estudo. Para mitigar essas ameaças, foi apresentado os resultados em uma disciplina de Metogolodia Científica e a um grupos de estudo e pesquisa.

## 3.5 Considerações Finais do Capítulo

Neste capítulo, foi apresentado um mapeamento sistemático sobre a aplicação da Inteligência Artificial (IA) no processo de *schema matching*. Foi explorado diversos estudos e pesquisas que utilizaram técnicas de IA em *Deep Learning*, *Machine Learning* e NLP, e identificado o uso de técnicas como Redes Neurais Convolucionais (CNN), Multilayer Perceptron (MLP), dentre outros para abordar o desafio de *schema matching*.

Durante a discussão dos resultados, foi possível observar o amplo uso de IA no contexto do *schema matching*. Ficando claro que a IA pode ter um papel crucial na otimização do *schema matching*, permitindo a identificação de correspondências entre elementos de esquemas heterogêneos. Essas descobertas enfatizam a necessidade de continuar explorando e aprimorando as técnicas de IA aplicadas ao *schema matching*, com o objetivo de compará-las com abordagens não baseadas em IA. Esse enfoque contínuo na evolução das técnicas de IA certamente contribuirá para um avanço significativo no campo do *schema matching* e suas aplicações práticas.

[6] https://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br
[7] https://scholar.google.com/
[8] https://www.google.com.br
[9] https://www.researchgate.net

No próximo capítulo, será apresentado o protocolo utilizado para a aplicação de uma pesquisa de opinião (Survey) sobre o estado da prática de *Schema Matching* na indústria brasileira. Foram detalhados o protocolo, as questões do formulário, o público-alvo e os resultados obtidos, acompanhados de uma análise para cada pergunta.

CAPÍTULO 4

# Estado da Prática sobre *Schema Matching* na Indústria Brasileira

Um *survey* é uma técnica de coleta de dados que visa obter informações de uma amostra representativa de uma população maior por meio de questionários, entrevistas ou outras formas de interação estruturada. É uma técnica essencial para contextualizar a pesquisa no ambiente prático e obter *insights* [Kasunic 2005], [Linåker et al. 2015], [Lebtag et al. 2022].

A diversidade de experiências na indústria é vasta, com diferentes setores apresentando nuances específicas nas suas necessidades de integração de dados. O *survey* permite capturar essa diversidade, garantindo uma análise abrangente que considera diferentes contextos e práticas, fornecendo uma visão realista dos desafios enfrentados e das soluções implementadas.

Ao coletar respostas de uma amostra representativa, é possível identificar tendências e padrões emergentes no uso de *schema matching* e *schema mapping*, e com isso realizar análises para revelar práticas comuns, áreas onde há inovação ou estão esfrentando desafios semelhantes.

Além disso, o *survey* valida a relevância da pesquisa em relação às demandas práticas da indústria. Entender como estas organizações abordam a integração de dados confirma se a pesquisa está alinhada com as necessidades reais do campo, aumentando a aplicabilidade e utilidade dos resultados. A coleta de dados sobre os desafios enfrentados pelas organizações na integração de dados, juntamente com as estratégias utilizadas para superar esses desafios, permite identificar áreas específicas que podem se beneficiar de avanços na pesquisa ou de melhores práticas compartilhadas.

Envolver profissionais da indústria no *survey* cria uma ponte direta entre a academia e a prática, fornecendo uma colaboração entre as partes envolvidas. Esta interação garante que as descobertas sejam relevantes e aplicáveis, fortalecendo a integração entre teoria e realidade.

O processo de condução de um *survey* geralmente envolve várias etapas. Um protocolo foi elaborado inspirado nas diretrizes de [Molléri, Petersen e Mendes 2016],

representado na figura 4.1 e é composto das seguintes etapas:

1. Definir objetivo da pesquisa;
2. Definir a justificativa da pesquisa;
3. Definir questões de pesquisa;
4. Definir público-alvo da pesquisa;
5. Elaborar questionário;
6. Teste piloto do questionário;
7. Distribuir e acompanhar o questionário;
8. Analisar resultados.

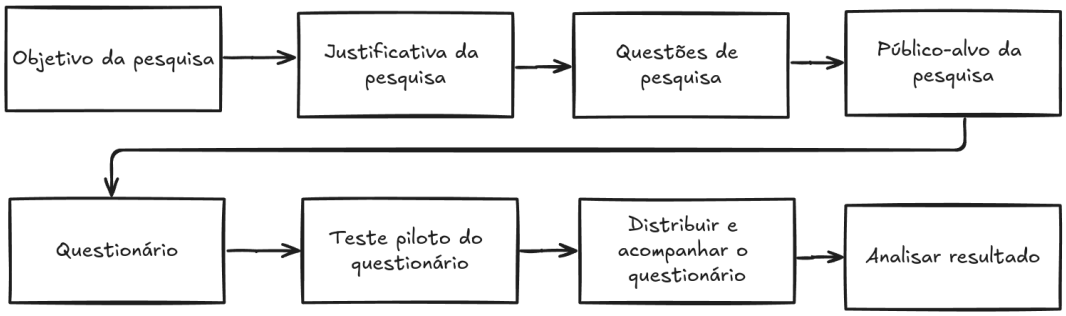


Figura 4.1: Protocolo do survey

Nas próximas seções serão detalhadas as etapas do protocolo do Survey.

## 4.1 Protocolo do *Survey*

**Objetivo de pesquisa**

O objetivo da pesquisa é investigar o estado da prática em integração de dados, e em particular, *schema matching* na indústria brasileira, ou seja, é descobrir se a indústria brasileira utiliza o *schema matching* em seus processos de integração de dados.

**Definição da justificativa da pesquisa**

A justificativa da pesquisa é compreender como a integração de software é abordada na indústria brasileira e como o *schema matching* é aplicado, com o objetivo de investigar a sua adesão no setor.

**Definição das questões de pesquisa**

As questões de pesquisa levantadas foram:

**QP1:** Quais são as abordagens predominantes que as empresas utilizam atualmente para *schema matching*?

**QP2:** Quais são os benefícios que as empresas percebem na abordagem que elas utilizam para o *schema matching*?

**QP3:** Quais são os principais desafios ou dificuldades enfrentados ao realizar *schema matching*?

**QP4:** Qual é o grau de esforço desempenhado pelas empresas na integração de dados?

**QP5:** Qual é o nível de complexidade das integrações de dados com que as empresas estão atualmente lidando?

**QP6:** Qual é o processo de *schema matching* que as empresas estão empregando?

**QP7:** Qual é o nível de satisfação das empresas ao utilizar o *schema matching*?

**Definição do público-alvo da pesquisa**

A definição do público-alvo da pesquisa foi um passo crucial, demandando a escolha estratégica de uma amostra de profissionais que aplicam a integração de dados no seu cotidiano. Essa abordagem cuidadosa na seleção da amostra fortalece a validade externa dos resultados, e também propicia uma análise mais abrangente e generalizável das conclusões obtidas. O público-alvo deste *survey* incluiu profissionais especializados em diversas áreas, incluindo desenvolvedores de software, analistas de dados, engenheiros de dados, administradores de banco de dados, cientistas de dados, pesquisadores, professores e arquitetos de soluções da indústria de software brasileira.

**Elaboração do questionário**

Durante a fase inicial de construção do questionário, empreendeu-se esforços para desenvolver perguntas que atendessem efetivamente aos objetivos delineados pelas questões de pesquisa. Este processo não se limitou apenas à formulação de questões, mas também incorporou a validação por especialistas no campo de integração de dados, garantindo que as perguntas fossem tecnicamente precisas e alinhadas com as práticas contemporâneas do setor.

Além disso, como parte integrante deste processo, dedicou-se atenção especial à elaboração do Termo de Consentimento Livre e Esclarecido (TCLE). Tal passo foi essencial pois, devido ao tempo exíguo para finalização do trabalho, não foi viável passar o trabalho por Comitê de Ética. De todo modo, recorre-se aqui também à dispensa de apreciação de Comitê de Ética, por enquadrar-se na categoria Pesquisa de opinião pública com participantes não identificáveis, conforme o Ofício Circular No 17/2022/CONEP/SECNS/MS, de julho de 2022 e OFÍCIO CIRCULAR No 12/2023/CONEP/SECNS/DGIP/SE/MS. Ademais este documento foi concebido com o propósito de assegurar a compreensão plena e informada por parte dos participantes, destacando seus direitos e delineando claramente os objetivos e procedimentos da pesquisa. A seguir, é apresentada uma versão resumida do TCLE para referência.

*"Este questionário é direcionado para profissionais que atuam na indústria brasileira de software. A pesquisa em questão tem como objetivo analisar o uso de técnicas e tecnologias para integração de dados e, em particular schema matching, pelos*

*profissionais da indústria. As respostas coletadas servirão como subsídio para confecção deste estudo. Os resultados obtidos serão utilizados apenas para fins acadêmicos.*

*Sua participação é voluntária, ou seja, você poderá desistir a qualquer momento, retirando seu consentimento, sem que isso lhe traga nenhum prejuízo ou penalidade. Se optar por aceitar o convite você será conduzido em seguida a preencher um questionário online. O tempo estimado para responder as questões é de 5 minutos. Não existem respostas certas ou erradas.*

*Todas as informações obtidas serão sigilosas e seu nome ou qualquer outro dado pessoal não será solicitado. Porém, se quiser receber informações dos resultados deste estudo, você pode informar o seu e-mail ao final do formulário. Os dados dessa pesquisa ficarão arquivados com os pesquisadores por um período de cinco anos. Após esse período, serão descartados conforme a legislação vigente.*

*O formulário dispensa apreciação de Comitê de Ética por enquadrar-se na categoria Pesquisa de opinião pública com participantes não identificáveis, conforme o Ofício Circular No 17/2022/CONEP/SECNS/MS, de julho de 2022 e OFÍCIO CIRCULAR No 12/2023/CONEP/SECNS/DGIP/SE/MS."*

A estrutura do questionário foi concebida conforme ilustrado na Tabela 4.1. Foram elaboradas (i) questões demográficas (QD), ou seja, questões de caracterização para obter o perfil dos participantes e (ii) questões de estudo (QE) para solicitar aos participantes informações específicas sobre sua prática com *schema matching*.

O questionário utilizou-se somente de questões fechadas (QF). As QF foram utilizados para padronizar as respostas, facilitar a leitura e sintetizar as informações necessárias.

Tabela 4.1: Leiaute do questionário

| **Etapas** | **Questões** |
|---|---|
| Etapa 1 | Explicações dos termos do questionários |
| | Termo de Consentimento Livre e Esclarecido (TCLE) |
| | Opção de concordar com os termos do questionário |
| Etapa 2 | Questionário demográfico fechado |
| Etapa 3 | Questionário de estudo fechado |

Na Tabela 4.2, é apresentada a correlação entre as questões de pesquisa e as perguntas do questionário. Cada pergunta do questionário está associada a uma questão de pesquisa específica, evidenciando a correspondência entre os objetivos da pesquisa e as indagações direcionadas aos participantes.

Cada célula da tabela representa uma ponte entre a coleta de dados e os objetivos mais amplos da pesquisa. Ao examinar a tabela, torna-se evidente como as respostas obtidas a partir do questionário alimentam diretamente os elementos fundamentais que compõem as questões de pesquisa.

Tabela 4.2: Relação das questões de pesquisa com as questões do formulário

| **ID** | **Questões de pesquisa** | **Questão do formulário** |
|---|---|---|
| QP1 | Quais são as abordagens predominantes que as empresas utilizam atualmente para *schema matching*? | QE10, QE11, QE13, QE17 |
| QP2 | Quais são os benefícios que as empresas percebem na abordagem que elas utilizam para o *schema matching*? | QE12 |
| QP3 | Quais são os principais desafios ou dificuldades enfrentados ao realizar *schema matching*? | QE3, QE12, QE14, QE16 |
| QP4 | Qual é o grau de esforço desempenhado pelas empresas na integrações de dados? | QE1, QE6, QE18, QE19 |
| QP5 | Qual é o nível de complexidade das integrações de dados com que as empresas estão atualmente lidando? | QE3, QE4, QE5, QE7 |
| QP6 | Qual é o processo de *schema matching* que as empresas estão empregando? | QE8, QE9, QE23 |
| QP7 | Qual é o nível de satisfação das empresas ao utilizar *schema matching*? | QE20, QE21, QE22 |

Nas Tabelas 4.3 e 4.4 são mostradas as perguntas de cada etapa do questionário, respectivamente das etapas 2 e 3. No apêndice B, são mostradas as imagens do questionário no Google Forms.

Tabela 4.3: Questões demográficas fechadas (QD)

| **ID** | **Questões de Caracterização** |
|---|---|
| QD1 | Qual sua faixa etária? |
| QD2 | Qual é o seu nível de escolaridade? |
| QD3 | Em qual unidade federativa você reside? |
| QD4 | Qual é o seu sexo? |
| QD5 | Quanto tempo de experiência profissional você tem no mercado de TI? |
| QD6 | Qual é a sua principal ocupação no mercado de TI? |
| QD7 | Quanto tempo de experiência profissional você tem na integração de dados? |
| QD8 | Qual é o seu nível de familiaridade com o conceito de *schema matching*? |

Tabela 4.4: Questões de estudo (QE)

| ID | Questões de estudo |
|---|---|
| QE1 | De quantos projetos de integração de dados você já participou? |
| QE2 | Em qual contexto sua empresa realiza integração de dados? |
| QE3 | Qual a complexidade (quantidade de atributos envolvidos) da fonte de dados com a qual você já trabalhou em uma atividade de integração de dados? |
| QE4 | Qual a quantidade de pessoas envolvidas em projetos de integração na sua empresa? |
| QE5 | Qual o esforço desempenhado em horas semanais dedicadas em projetos de integração na sua empresa? |
| QE6 | Quais tipos de dados ou esquemas sua empresa precisa integrar com mais frequência? |
| QE7 | Quais fatores você acredita que contribuem para o aumento do esforço em projetos de integração? |
| QE8 | Que estratégias sua empresa adota para gerenciar a complexidade na integração de dados? |
| QE9 | Qual é o impacto da complexidade da etapa de integrações de dados no tempo necessário para concluir um projeto? |
| QE10 | Sua empresa atualmente está utilizando *schema matching* automático ou semi automático? |
| QE11 | Qual o principal objetivo ou caso de uso para o *schema matching* na integração de sistemas em sua organização? |
| QE12 | Como você avaliaria o desempenho do *schema matching* em sua organização em uma escala de 1 a 5, sendo 1 muito ineficaz e 5 muito eficaz? |
| QE13 | Quais ferramentas ou tecnologias de *schema matching* sua organização utiliza atualmente? |
| QE14 | Quais são os principais desafios que sua organização enfrentou ao implementar o uso do *schema matching*? |
| QE15 | Em sua opinião, quais são os principais benefícios do uso de *schema matching* automático ou semi automático? |
| QE16 | Quais são as principais razões para a não utilização do *schema matching*? |
| QE17 | Quais ferramentas de *schema matching* você utilizou em projetos de integração de dados? |
| QE18 | Como sua empresa lida com conflitos de correspondência ao usar *schema matching*? |
| QE19 | Sua empresa tem estratégias ou planos para melhorar o uso de *schema matching* no futuro? Se sim, quais são esses planos? |
| QE20 | Que práticas sua empresa adota para documentar e monitorar as correspondências de esquemas ao longo do tempo? |
| QE21 | Sua empresa realiza alguma validação ou verificação posterior ao processo de *schema matching* para garantir a precisão e a integridade dos dados? |
| QE22 | Em uma escala de 1 a 5, onde 1 representa "totalmente insatisfeito"e 5 representa "totalmente satisfeito", quão satisfeito você está com o processo de *schema matching* em sua empresa? |
| QE23 | Quais ferramentas de *schema mapping* (Transformação) você utiliza na sua empresa? |

As questões do questionário constituem o cerne da coleta de dados, visando extrair informações pertinentes diretamente relacionadas aos objetivos da pesquisa. Por sua vez, as questões demográficas podem oferecer *insights* sobre as características sociodemográficas dos participantes.

**Teste piloto do questionário**

Na etapa de desenvolvimento do questionário, buscou-se garantir a qualidade dos dados coletados por meio de validações, contando com a colaboração de especialistas experientes de diversas empresas, visando garantir que as perguntas fossem relevantes e abrangessem os temas de integração de dados, *schema matching* e *schema mapping*.

Além disso, reconhecendo a sensibilidade de algumas questões pessoais contidas no questionário, foi realizada uma etapa de validação em colaboração com especialistas em anonimização de formulários da Pro-reitoria de Graduação da UFG e Procurador Educacional Institucional. Esse processo teve como objetivo garantir que as informações de cunho pessoal dos respondentes fossem devidamente protegidas e que o questionário respeitasse as diretrizes éticas necessárias para a pesquisa.

A participação ativa desses especialistas contribuiu para a refinamento do questionário, assegurando que as perguntas fossem claras, relevantes e que a coleta de dados fosse conduzida de maneira ética e respeitosa.

**Distribuição e acompanhamento do questionário**

A aplicação do questionário à amostra selecionada exigiu atenção para garantir que os participantes compreendessem plenamente as perguntas e pudessem fornecer respostas de maneira adequada. Este estágio não apenas representou a interação direta com os respondentes, mas também influenciou diretamente a confiabilidade dos dados obtidos.

Durante a aplicação do questionário, observou-se com cautela a criação de um ambiente propício para que os participantes se sentissem confortáveis e confiantes ao responder. Isso envolveu a inclusão de instruções claras, garantindo uma linguagem acessível ao público-alvo. Além disso, a disposição em fornecer orientações adicionais contribuiu para a compreensão completa das perguntas.

A verificação da compreensão das perguntas foi uma preocupação crucial para evitar equívocos nas respostas. Isso foi alcançado por meio de aplicação de testes anteriores à aplicação efetiva do questionário. Essas estratégias ajudaram a garantir que as perguntas fossem interpretadas conforme a intenção do pesquisador, aumentando a validade dos dados coletados.

Além disso, outra preocupação foi com fatores que poderiam influenciar a qualidade das respostas, como a fadiga do respondente ou distrações externas. Para isso evitou-se aplicar várias perguntas no *survey*, fazendo com que uma pergunta respondesse outras relacionadas.

A comunicação transparente sobre a finalidade da pesquisa, a garantia de anonimato e confidencialidade promoveram a sinceridade e a participação ativa dos respondentes.

O questionário da pesquisa foi divulgado por meio de um link público [1] no *Google Forms*, amplamente compartilhado em diferentes canais como listas de e-mail, *LinkedIn* [2], grupos de analista de dados no *Telegram* [3], no *WhatsApp* [4] e convites diretos. O período de coleta de dados estendeu-se de 21 de novembro de 2023 a 22 de março de 2024. Essa abordagem visou maximizar a participação e garantir a diversidade de respondentes. Houve um esforço deliberado para alcançar profissionais de diversas regiões do país, a fim de assegurar uma representação abrangente do mercado de tecnologia brasileiro. Durante a aplicação do questionário, foi realizado o monitoramento do número de participantes e da distribuição geográfica dos respondentes, com o objetivo de alcançar uma amostra diversificada. Ao todo foram convidadas cerca de 30 pessoas de forma direta e mais de 5000 de forma indireta, onde apenas 35 responderam ao *survey*. Nota-se, portanto, uma maior efetividade na estratégia de convite direto para a obtenção de respostas.

**Análise dos resultados**

A condução de análises estatísticas visou extrair padrões, identificar tendências e derivar conclusões substanciais a partir dos dados coletados. Este estágio não apenas conferiu uma dimensão quantitativa das informações adquiridas, mas também proporcionou visões que foram essenciais para a compreensão do tema.

A identificação dos padrões nos dados proporcionou uma visão sobre as relações entre variáveis, enquanto a análise de tendências revelou a direção geral das mudanças ao longo do tempo. Além disso, ao extrair inferências estatísticas, realizou-se validações das hipóteses e fornecido as respostas para as questões de pesquisa.

As análises estatísticas foram conduzidas com rigor e transparência, utilizando métodos estatísticos apropriados e considerando possíveis viéses. Além disso, a interpretação dos resultados foi feita com cautela, evitando generalizações precipitadas e destacando as limitações do estudo.

A visualização dos resultados foi viabilizada por meio de gráficos e tabelas, facilitando a compreensão e comunicação efetiva. Houve também um tratamento de alguns dados quando necessário, visto que encontrou-se a necessidade de padronizar respostas em questões com campo aberto.

Entretanto, buscou-se a apresentação clara dos resultados a fim de contribuir não apenas para a validade interna do estudo, mas também para sua utilidade prática e aplicabilidade em contextos acadêmicos e profissionais.

[1] https://forms.gle/dzGJdbd9Mdri5dvu8
[2] https://www.linkedin.com
[3] https://web.telegram.org
[4] https://web.whatsapp.com

## 4.2 Resultados

Nas próximas seções serão apresentados em detalhes os resultados obtidos por meio do *survey* realizado. Inicialmente, será destacado os resultados específicos para cada pergunta, proporcionando uma visão das respostas obtidas. Em seguida, serão exploradas as descobertas relacionadas, buscando correlações que possam surgir ao analisar as inter-relações das perguntas. Esta seção oferece uma visão das percepções dos profissionais envolvidos, contribuindo para a compreensão do cenário atual de *schema matching* e suas implicações nos processos de integração de dados dentro do contexto industrial brasileiro.

### 4.2.1 Questões do *Survey*

#### Questões demográficas

No âmbito das questões demográficas, foram abordados diversos aspectos relacionados ao perfil dos respondentes, proporcionando uma compreensão abrangente do contexto pessoal dos participantes

Nas próximas seções, serão demonstrados os resultados relativos às respostas a cada questão.

#### Qual sua faixa etária?

A primeira questão, destinada a mapear a distribuição etária da amostra, revela-se crucial para a compreensão do perfil demográfico dos participantes do *survey*. A análise desta questão proporciona *insights* sobre a representatividade de diferentes faixas etárias, a Figura 4.2 apresenta de maneira visual. É possível identificar a prevalência de profissionais com idade média de 31 anos à 40 anos, cerca de 40%, e também uma diversidade de faixas etárias na amostra com profissionais a partir dos 18 anos de idade.

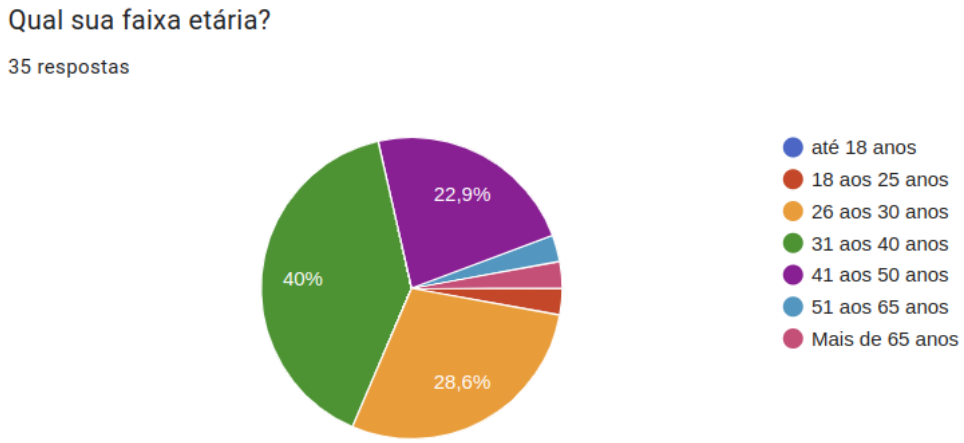


Figura 4.2: Qual sua faixa etária?

**Qual seu nível de escolaridade?**

A segunda questão foi estrategicamente formulada para avaliar o *background* educacional dos profissionais participantes. Compreender o nível de escolaridade dos respondentes é fundamental para contextualizar as perspectivas e abordagens adotadas em relação ao conhecimento sobre o termo *schema matching*, onde muitas vezes é desconhecido ou conhecido com uma outra nomenclatura.

Os resultados dessa investigação estão detalhadamente apresentados na Figura 4.3, que fornece uma representação gráfica da distribuição dos respondentes com base em seus níveis de escolaridade. Esta visualização permite identificar que a ampla maioria dos participantes possuem “Pós graduação completa”.

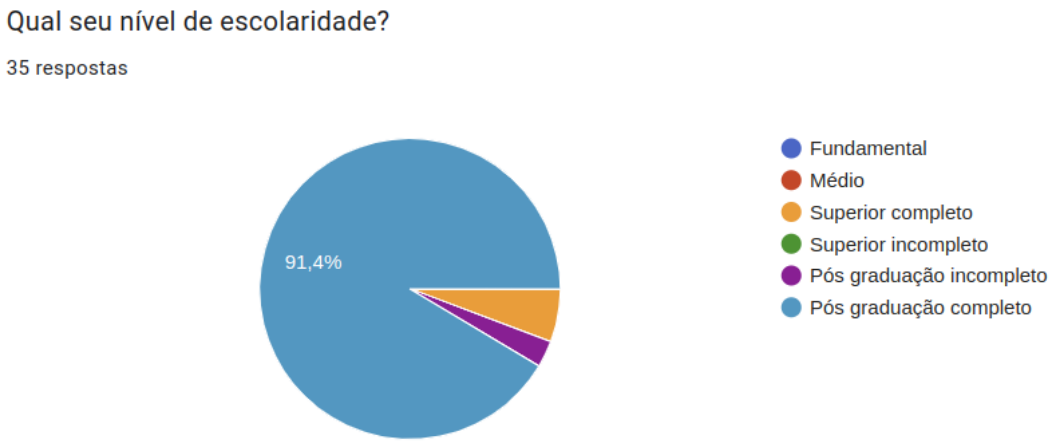


Figura 4.3: Qual seu nível de escolaridade?

**Em qual unidade federativa você reside?**

A partir do gráfico da Figura 4.4 de setores fornecido, é possível realizar uma análise quantitativa por unidade federativa (UF). A unidade federativa com a maior representatividade é Goiás (GO), concentrando 15 respondentes (42,9%), indicando um possível viés na aplicação do *survey*, a possível centralização da adoção ou do interesse em práticas de *schema matching* nesta região. O Rio de Janeiro (RJ) segue com 4 (11,4%). São Paulo (SP) frequentemente reconhecido como um centro tecnológico no Brasil, representa 8,6% da amostra.

As demais unidades federativas apresentam uma participação mais homogênea, variando de 2,9% a 5,7%. Tais percentuais são indicativos de uma dispersão moderada de respondentes pelo território nacional. Os estados do Maranhão (MA), Pernambuco (PE), Paraíba (PB), Bahia (BA), Rio Grande do Norte (RN), Distrito Federal (DF), Santa Catarina (SC) e Minas Gerais (MG) contribuem com 5,7% cada um. Enquanto isso, os estados do Rio Grande do Sul (RS) e Paraná (PR) apresentam uma contribuição de 2,9%.

Esta distribuição pode sugerir uma presença significativa de práticas de *schema matching* em todo cenário brasileiro, o que pode refletir as tendências atuais da indústria de software brasileira em termos de diversificação geográfica das tecnologias de informação.

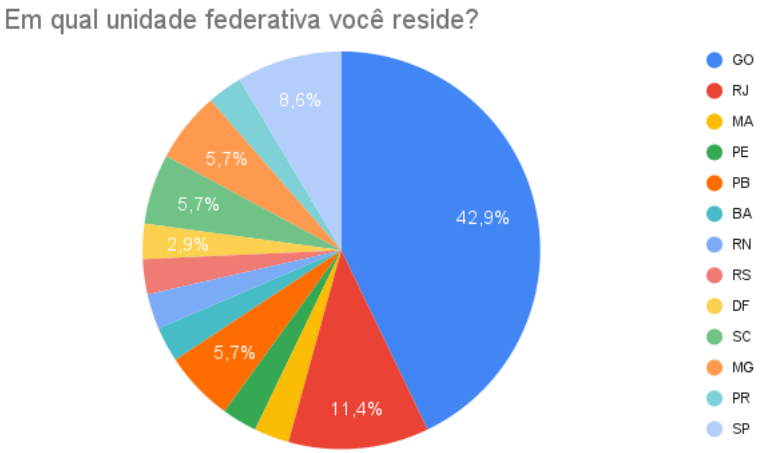


Figura 4.4: Em qual unidade federativa você reside?

**Qual é o seu gênero?**

Esta questão foi elaborada com o objetivo de obter uma perspectiva mais abrangente da diversidade de gênero entre os participantes. Os resultados desta análise de gênero são apresentados na Figura 4.5, oferecendo uma visualização gráfica que destaca a

predominância de identidades do gênero masculino entre os respondentes, o que é típico no ambiente da computação.

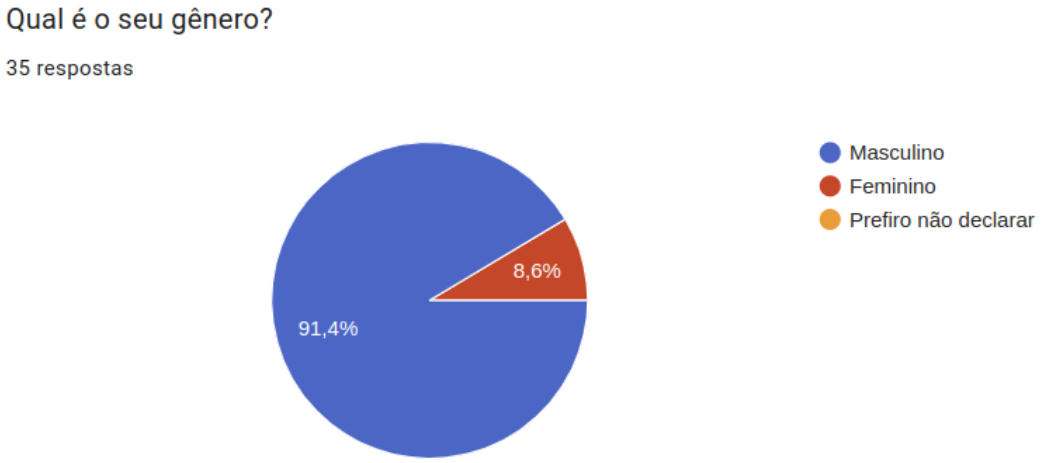


Figura 4.5: Qual é o seu gênero?

**Quanto tempo de experiência profissional você tem no mercado de TI?**

Voltada para a mensuração do tempo de experiência profissional no mercado de Tecnologia da Informação (TI), esta pergunta destinou-se à compreensão do tempo dedicado ao setor de TI pelos respondentes. Os resultados estão apresentados de maneira detalhada na Figura 4.6. Notavelmente, observa-se que impressionantes 88% dos participantes acumulam mais de 5 anos de experiência no setor. Esta constatação sugere uma amostra predominantemente composta por profissionais experientes, o que contribui para uma visão mais consolidada das práticas e desafios enfrentados no contexto de integração de dados.


Figura 4.6: Quanto tempo de experiência profissional você tem no mercado de TI?

**Qual é a sua principal ocupação no mercado de TI?**

Direcionada à identificação da ocupação principal no mercado de Tecnologia da Informação (TI) por parte dos participantes, visa contextualizar as diversas funções desempenhadas pelos profissionais envolvidos no *survey*. Aqui é importante destacar que o *survey* possibilitou o preenchimento de mais de uma opção para a principal ocupação.

O gráfico da Figura 4.7 é um gráfico que representa as principais ocupações dos respondentes no mercado de TI.

Desenvolvedor de Software é a ocupação mais comum. Com 14 respondentes, correspondendo a 40% do total, esta é claramente a ocupação mais prevalente entre os participantes da pesquisa. Isso sugere que o desenvolvimento de software é um papel significativo dentro do setor de TI entre esse grupo.

Com 7 Analistas de Dados e 8 Engenheiros de Dados, o que representa em torno de 20% do total para cada categoria. Isso indica que as funções focadas em dados são bastante importantes e bem representadas no mercado de TI.

Além dos Analistas e Engenheiros de Dados, há também Arquitetos de Soluções e Cientistas de Dados, cada um com 7 e 6 representantes respectivamente. Isso mostra uma diversidade nas ocupações relacionadas ao campo de dados dentro do mercado de TI.

Gestores, como Gerente de Operações e Gestor de TI, têm apenas 1 representante cada, o que é 4% do total para cada um. Isso pode sugerir que a amostra é mais técnica e menos focada em funções de gestão.

Ainda uma representação significativa de Professores e Pesquisadores (8 e 9 respondentes, respectivamente), o que sugere que há também um enfoque em ensino e pesquisa no mercado de TI entre os entrevistados.

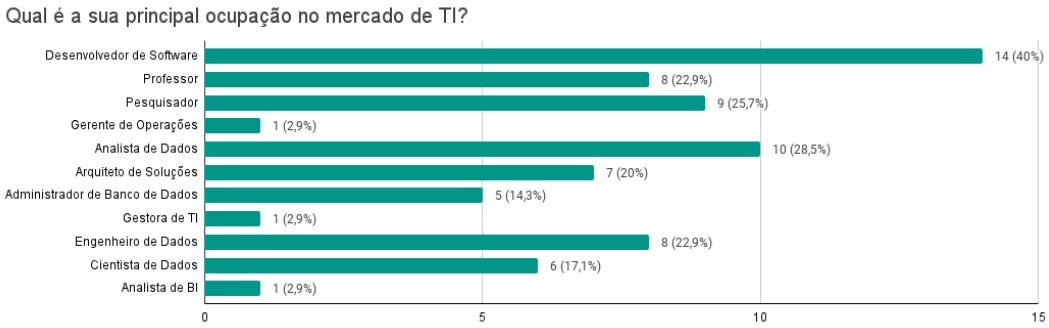


Figura 4.7: Qual é a sua principal ocupação no mercado de TI?

**Quanto tempo de experiência profissional você tem na integração de dados?**

No gráfico da Figura 4.8 é mostrado a distribuição do tempo de experiência profissional que os entrevistados têm na área de integração de dados. As porcentagens indicam as seguintes distribuições: 57,1% (20) dos entrevistados têm mais de 5 anos de experiência; 25,7% (9) têm entre 2 a 5 anos de experiência e 16% (6) têm até 2 anos de experiência.

Há uma predominância de profissionais com experiência relativamente alta na área (mais de 5 anos), o que sugere que muitos profissionais podem ser relativamente novos no campo da integração de dados ou que a área tem atraído novos profissionais recentemente. Esta alta proporção de profissionais com mais de 5 anos de experiência pode indicar que a indústria de integração de dados está crescendo.

Uma menor proporção de profissionais com até 2 anos de experiência pode sugerir uma lacuna de habilidades ou uma oportunidade de mercado para programas de treinamento e educação voltados para profissionais de integração de dados com experiência mais extensa.

Com mais profissionais experientes, pode haver impactos na inovação e na liderança dentro da área, pois a experiência prolongada muitas vezes contribui para a capacidade de liderar projetos complexos e inovar.


Figura 4.8: Quanto tempo de experiência profissional você tem na integração de dados?

**Qual é o seu nível de familiaridade com o conceito de *schema matching*?**

O gráfico da Figura 4.9 mostra a familiaridade dos entrevistados com o conceito. Onde a maior parte dos entrevistados, sendo 14 (40%) relatam estarem Familiarizados com o conceito de *schema matching*. Isso sugere que a maioria dos profissionais tem uma compreensão básica ou melhor do conceito. No entanto, apenas uma pequena fração de 25,7% (9) se considera "Muito familiar", o que pode indicar que, embora muitos tenham algum conhecimento, poucos são altamente proficientes ou especializados neste conceito.

Com 11,4% (4) dos entrevistados sentindo-se "Pouco familiar", "Neutro" ou "Não familiar" com o conceito, existe uma oportunidade para o desenvolvimento de treinamento ou recursos educacionais para aumentar o conhecimento nessa área. Comparando com o gráfico anterior sobre a experiência em integração de dados, pode-se explorar se existe uma correlação entre o nível de experiência em integração de dados e a familiaridade com *schema matching*. Por exemplo, pode-se investigar se aqueles com mais experiência em integração de dados têm maior familiaridade com o conceito.

O fato de uma proporção significativa dos entrevistados estar familiarizada com o conceito sugere que *schema matching* é relevante para a indústria atual e é provável que seja um tópico importante em muitos ambientes de trabalho relacionados a dados.

A distribuição da familiaridade com *schema matching* pode refletir a diversidade de papéis dentro do campo de integração de dados, onde alguns papéis podem exigir conhecimento profundo de *schema matching* enquanto outros não.

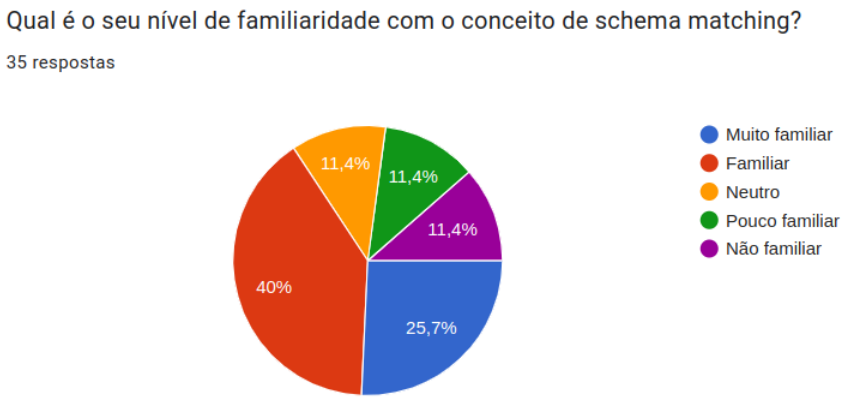


Figura 4.9: Qual é o seu nível de familiaridade com o conceito de *schema matching*?

**Questões de estudo**

As questões apresentadas têm como objetivo investigar a experiência dos participantes em projetos de integração de dados. Elas buscam entender o contexto, a complexidade das fontes de dados, o esforço dedicado, e as estratégias adotadas. Além disso, focam no uso de *schema matching* e *schema mapping*, explorando ferramentas, satisfação com o processo, desafios enfrentados e planos futuros. Essas informações são essenciais para mapear práticas atuais, identificar tendências, e orientar melhorias contínuas na integração de dados, visando eficiência e eficácia aprimoradas. As respostas dos participantes contribuirão para uma compreensão mais clara do panorama da integração de dados e fornecerão insights valiosos para o desenvolvimento futuro dessas práticas.

Nas próximas seções, serão demonstrados os resultados de cada questão.

**De quantos projetos de integração de dados você já participou?**

O gráfico da Figura 4.10 exibe a quantidade de projetos de integração de dados dos quais os entrevistados participaram. A maior parte dos entrevistados, sendo 34,3% (12) esteve envolvida em 3 a 5 projetos de integração de dados, o que sugere um nível de experiência moderada a alta entre os participantes da pesquisa. Apenas 25,7% (9) dos entrevistados participaram de mais de 10 projetos. Isso indica que profissionais com um grande número de projetos de integração de dados são menos comuns, o que pode refletir a complexidade e o escopo desses projetos ou a demanda do mercado. Com 28,6% (10) dos entrevistados tendo participado de apenas 1 a 2 projetos, isso pode indicar que um número significativo de profissionais está no início de sua carreira na integração de dados ou que alguns profissionais não se dedicam exclusivamente a essa área.

A distribuição sugere que há espaço para o crescimento profissional, com muitos participantes tendo a oportunidade de aumentar seu número de projetos ao longo do

tempo. Comparando com os gráficos anteriores sobre experiência em integração de dados e familiaridade com *schema matching*, pode-se analisar a correlação entre a quantidade de projetos dos quais os participantes fizeram parte e seu nível de experiência e familiaridade.

A participação em um número maior de projetos pode estar associada a um aumento na qualificação e no conhecimento prático, o que pode influenciar a eficácia e a qualidade do trabalho em integração de dados.

As organizações e instituições educacionais podem usar essas informações para orientar o desenvolvimento de carreira, programas de treinamento, e mentorias para ajudar profissionais a adquirir experiência prática em mais projetos.

A distribuição também pode refletir a maturidade do campo de integração de dados, onde muitos profissionais têm experiência moderada, mas ainda há uma necessidade de acumular mais experiências e conhecimentos ao longo do tempo.

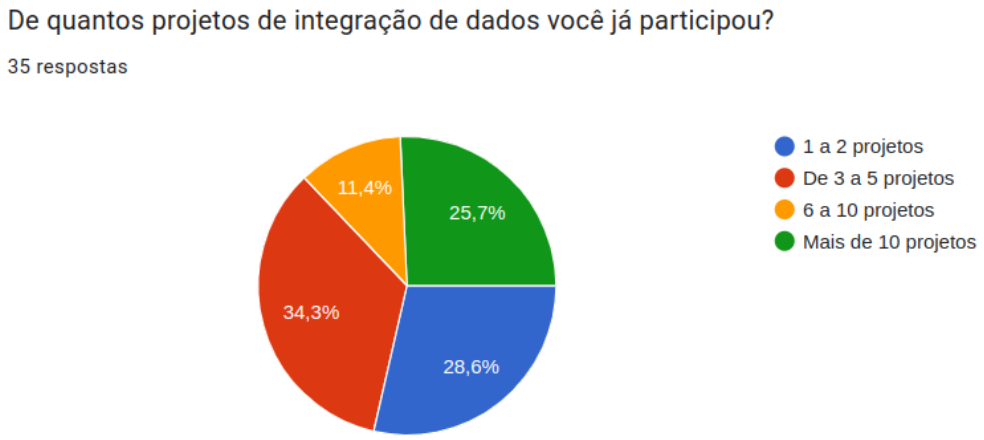


Figura 4.10: De quantos projetos de integração de dados você já participou?

**Em qual contexto sua empresa realiza integração de dados?**

O gráfico da Figura 4.11 mostra o contexto em que as empresas dos entrevistados realizam integração de dados. Uma grande maioria (61,8%) das empresas realiza integração de dados em contextos de *Data Warehouses*. Metade das empresas também utiliza *Data Lakes*, cerca de 50%, o que indica que essa tecnologia tem uma adoção significativa, possivelmente devido à sua flexibilidade e capacidade de armazenar grandes volumes de dados não estruturados. Isso sugere que essas duas abordagens são as mais comuns ou preferidas pelas empresas representadas na amostra.

*Lake House*, uma arquitetura emergente que combina elementos de *Data Lakes* e *Data Warehouses*, parece ser menos comum entre os entrevistados, o que pode ser devido ao fato de ser um conceito mais novo e menos estabelecido no mercado.

A migração de sistemas é raramente mencionada, cerca de 2,9%, o que pode indicar que as empresas estão mais focadas em manter e integrar dados entre sistemas existentes do que em migrar para novos sistemas.

É importante notar que as porcentagens somam mais de 100%, o que mostra que as empresas usam múltiplas abordagens de integração de dados simultaneamente. Empresas com integração virtual e *Data Warehouses* podem ser mais maduras em suas operações de dados ou podem ter requisitos específicos que justificam o uso dessas abordagens.

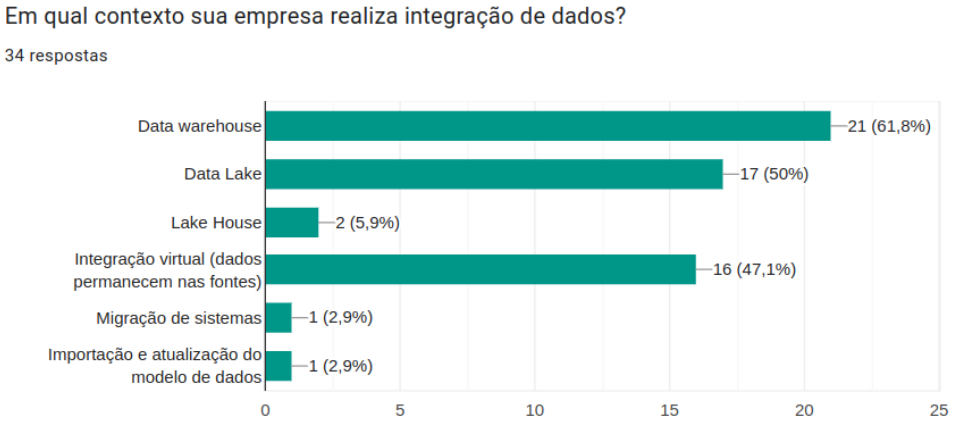


Figura 4.11: Em qual contexto sua empresa realiza integração de dados?

**Qual a complexidade (quantidade de atributos envolvidos) da fonte de dados com a qual você já trabalhou em uma atividade de integração de dados?**

O gráfico da Figura 4.12 mostra a complexidade das fontes de dados, em termos da quantidade de atributos envolvidos, com as quais os entrevistados já trabalharam em atividades de integração de dados. A maioria dos entrevistados, cerca de 51,4% trabalharam com fontes de dados de alta complexidade, o que pode sugerir que há uma grande quantidade de projetos de integração de dados que envolvem fontes de dados mais complexos ou que a maioria das empresas lidam com grandes volumes de atributos em seus dados.

Entretanto, menos da metade dos entrevistados (17,1%) enfrentaram fontes de baixa complexidade, indicando que os profissionais lidam com integrações menos desafiadoras, mas que podem requerir uma gestão de dados mais robusta e habilidades avançadas.

Além disso, com uma proporção significativa de 31,4% trabalhando com baixa complexidade, há um potencial mercado para especialistas que possam lidar com baixa complexidade em integração de dados. A distribuição também sugere que as empresas

podem precisar de estratégias de integração de dados que sejam escaláveis para lidar com diferentes níveis de complexidade à medida que seus dados crescem em volume e variedade.

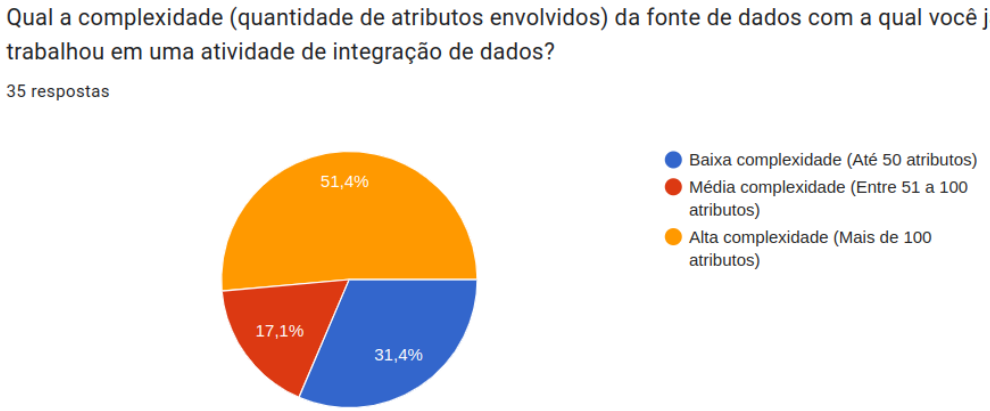


Figura 4.12: Qual a complexidade (quantidade de atributos envolvidos) da fonte de dados com a qual você já trabalhou em uma atividade de integração de dados?

**Qual a quantidade de pessoas envolvidas em projetos de integração na sua empresa?**

O gráfico da Figura 4.13 apresenta a quantidade de pessoas envolvidas em projetos de integração de dados nas empresas dos entrevistados. Uma maioria significativa (31,4%) das respostas indica que os projetos de integração de dados geralmente envolvem equipes de até 3 pessoas, o que pode sugerir que muitos projetos de integração de dados são de escala menor ou que as empresas preferem equipes menores e mais ágeis.

Há uma distribuição variada no tamanho das equipes envolvidas, o que pode refletir uma diversidade na escala dos projetos de integração de dados. Além disso, proporcionalmente projetos que envolvem uma grande quantidade de pessoas são mais raros (11,4%), indicando que projetos de integração de grande escala são menos comuns ou que apenas empresas maiores tendem a ter projetos desse tamanho.

Pode existir uma correlação entre o tamanho da equipe e a complexidade do projeto mencionada anteriormente; projetos com maior complexidade podem exigir mais pessoas. O tamanho das equipes também pode refletir o tamanho geral da empresa e sua cultura organizacional em relação à gestão de projetos de integração de dados.

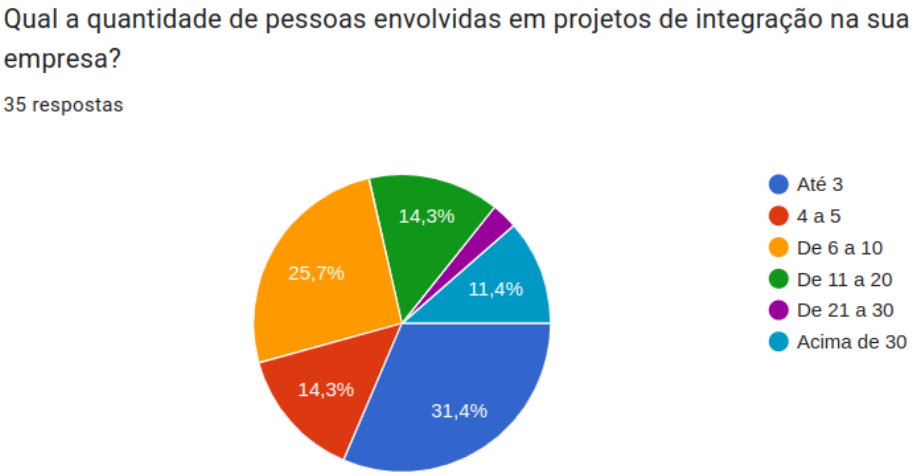


Figura 4.13: Qual a quantidade de pessoas envolvidas em projetos de integração na sua empresa?

**Qual o esforço desempenhado em horas semanais dedicadas em projetos de integração na sua empresa?**

O gráfico da Figura 4.14 exibe a distribuição do tempo semanal dedicado pelos entrevistados a projetos de integração de dados nas suas empresas. A maioria dos entrevistados, cerca de 11 respondentes (31,4%) dedica de 6 a 10 horas por semana em projetos de integração de dados, o que sugere que, para a maioria dos entrevistados, a integração de dados não é uma atividade de tempo integral. Assim, existe uma diversidade considerável no tempo dedicado aos projetos de integração de dados, o que pode refletir diferenças no escopo dos projetos, no tamanho das empresas ou no papel dos indivíduos dentro de suas organizações.

O fato de que uma grande percentagem de profissionais dedica uma parcela menor do seu tempo à integração de dados pode indicar que essa atividade é frequentemente realizada juntamente com outras tarefas de um projeto. Há uma porcentagem de 37,2% de profissionais (13) que dedicam mais de 20 horas semanais, o que pode indicar a existência de projetos de integração de maior envergadura ou de papéis específicos focados integralmente nessa atividade.

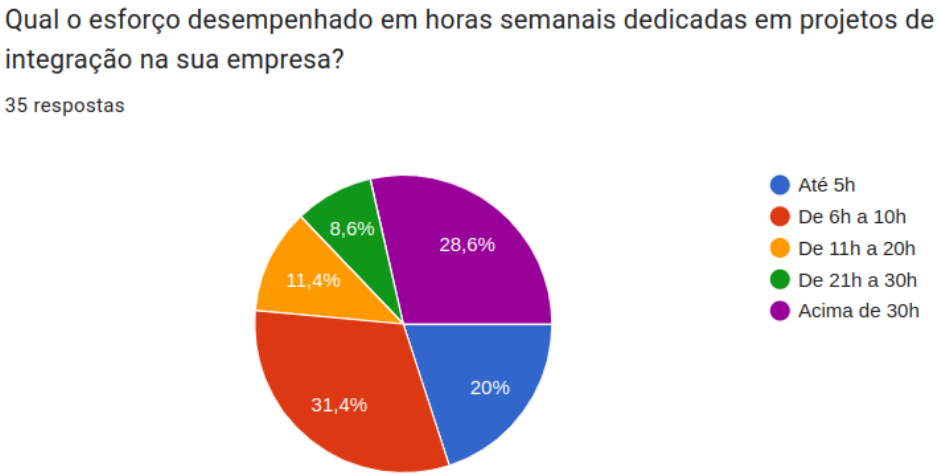


Figura 4.14: Qual a quantidade de pessoas envolvidas em projetos de integração na sua empresa?

**Quais tipos de dados ou esquemas sua empresa precisa integrar com mais frequência?**

O gráfico da Figura 4.15 mostra a frequência com que diferentes tipos de dados ou esquemas são integrados nas empresas dos entrevistados. A integração de dados estruturados é, de longe, a mais comum nas empresas representadas na amostra, representando cerca de 74,3%, o que sugere que bancos de dados relacionais continuam sendo uma parte central das operações de dados empresariais. Dados semiestruturados, como JSON e XML, também são frequentemente integrados, cerca de 11,4%, o que pode indicar o uso de APIs e trocas de dados entre sistemas modernos e web services. Já a presença de dados não estruturados (5,7%), mesmo que pequena, mostra que as empresas estão trabalhando com uma variedade de fontes de dados, o que pode exigir competências diversificadas e ferramentas de integração especializadas.

Há também a necessidade de selecionar múltiplos tipos de dados. Isso indica que as empresas podem muitas vezes precisar integrar vários tipos de dados, exigindo flexibilidade nas ferramentas de integração e nas abordagens.

A predominância de dados estruturados não diminui a complexidade dos desafios de integração, pois mesmo dentro dos dados estruturados, pode haver uma grande variedade de esquemas e formatos a serem considerados. A distribuição também pode refletir tendências atuais no setor de integração de dados, com um foco contínuo em dados estruturados, mas também uma necessidade reconhecida de lidar com dados semiestruturados e não estruturados.

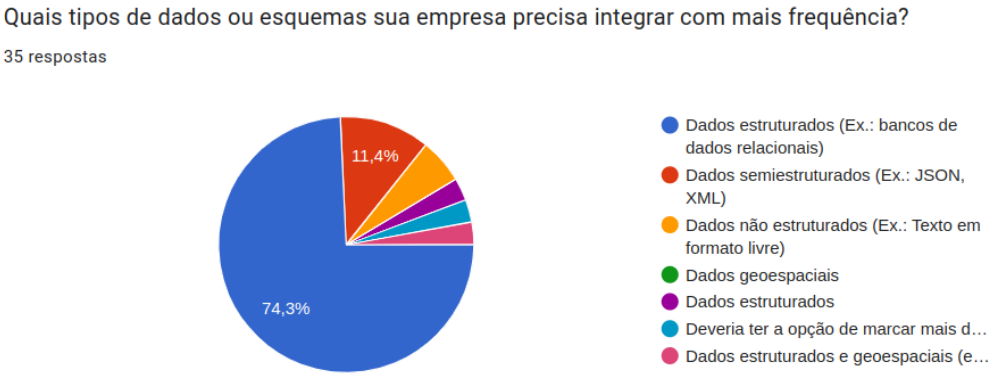


Figura 4.15: Quais tipos de dados ou esquemas sua empresa precisa integrar com mais frequência?

**Quais fatores você acredita que contribuem para o aumento do esforço em projetos de integração?**

O gráfico da Figura 4.16 mostra a percepção dos entrevistados sobre os fatores que contribuem para o aumento do esforço em projetos de integração de dados.

A falta de documentação adequada é vista como o principal fator que aumenta o esforço nos projetos de integração, representando cerca de 77,1%, sugerindo que o acesso a informações claras e precisas sobre os dados e os processos é fundamental para o sucesso da integração. A alta porcentagem de respostas citando a falta de padronização nos formatos de dados como um fator de aumento de esforço, cerca de 68,6%, destaca a importância de padronização para facilitar a integração de dados eficiente.

Mudanças frequentes nos requisitos são identificadas como um fator significativo, cerca de 60%, o que pode indicar um ambiente de projeto dinâmico onde os objetivos e necessidades estão constantemente evoluindo. A complexidade dos sistemas ou plataformas (45,7%) é também uma consideração importante, sugerindo que a integração de dados muitas vezes envolve sistemas complexos que podem ser desafiadores para trabalhar. Já o volume elevado de dados (34,3%) é citado por um terço dos entrevistados, indicando que o esforço para lidar com grandes volumes de dados é um desafio significativo para muitas empresas.

Uma proporção menor de entrevistados aponta a falta de experiência ou treinamento (20%) como um fator, o que pode refletir a necessidade de educação e desenvolvimento profissional contínuos na área de integração de dados. A integração de dados de terceiros e a dependência de ferramentas ou plataformas externas (34,3%) também são consideradas fatores relevantes, destacando a complexidade adicional trazida pela necessidade de integrar sistemas que estão fora do controle direto da empresa.

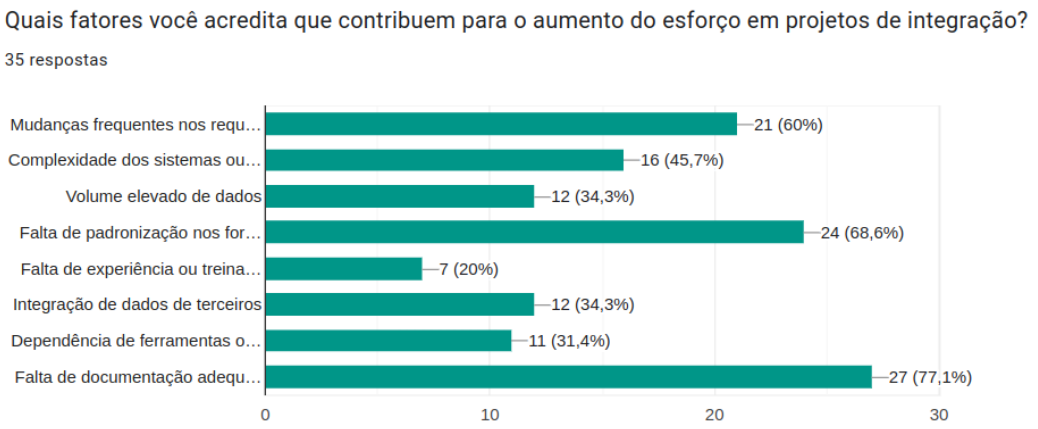


Figura 4.16: Quais fatores você acredita que contribuem para o aumento do esforço em projetos de integração?

**Que estratégias sua empresa adota para gerenciar a complexidade na integração de dados?**

O gráfico da Figura 4.17 representa as estratégias adotadas por empresas para gerenciar a complexidade na integração de dados. A padronização de processos é a estratégia mais adotada, representando cerca de 65,7%, sugerindo que as empresas veem a uniformidade nos procedimentos como fundamental para gerenciar a complexidade e melhorar a eficiência.

Entretanto, o uso de ferramentas específicas para integração de dados é também comum, cerca de 48,6%, o que implica que a tecnologia desempenha um papel importante em enfrentar os desafios de integração. A contratação de especialistas em integração de dados (37,1%) também é uma estratégia empregada por quase metade das empresas, o que indica que o conhecimento especializado é crucial para gerenciar a complexidade dos dados.

Outro fato imporante é que menos empresas adotam a estratégia de atualização regular dos requisitos de integração, cerca de 14,3%, o que pode sugerir que essa é uma área com potencial para melhoria. Tabém há empresas que não gerenciam ou não adotam estratégias específicas para lidar com a complexidade, o que pode indicar uma oportunidade de mercado para consultorias e provedores de soluções de integração de dados.

As informações do gráfico também sugerem que há uma demanda contínua por educação e formação em padronização de processos e no uso de ferramentas de integração de dados.

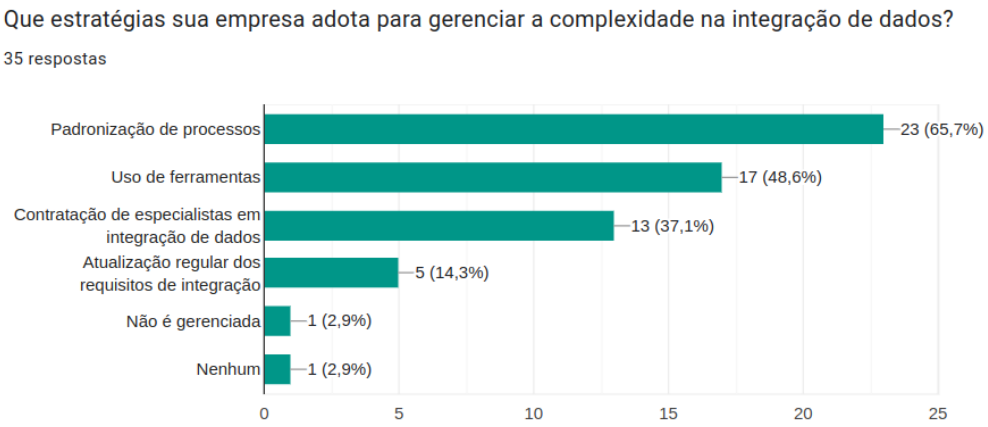


Figura 4.17: Que estratégias sua empresa adota para gerenciar a complexidade na integração de dados?

**Qual é o impacto da complexidade da etapa de integrações de dados no tempo necessário para concluir um projeto?**

O gráfico da Figura 4.18 fornece dados sobre a percepção do impacto da complexidade da etapa de integrações de dados no tempo necessário para concluir um projeto. Uma grande maioria, cerca de 29 (82,9%) dos entrevistados, acreditam que a complexidade na etapa de integração de dados aumenta o tempo necessário para concluir um projeto, indicando que a complexidade é um dos principais obstáculos para a eficiência do projeto. Uma possível explicação para isso é que para gestores de projeto, é essencial considerar a complexidade das integrações de dados no planejamento e na estimativa de prazos, alocando tempo adicional para lidar com possíveis complicações.

As empresas podem precisar desenvolver ou aprimorar estratégias para mitigar o impacto da complexidade nos cronogramas de seus projetos, talvez adotando melhores práticas ou ferramentas especializadas. O dado reforça a necessidade de treinamento adequado e desenvolvimento de competências nas equipes de integração de dados para lidar efetivamente com a complexidade e minimizar seu impacto nos cronogramas de projeto.

O fato de que uma parcela tão grande de entrevistados percebe um aumento no tempo necessário pode sugerir que há uma tendência de subestimar a complexidade nas fases iniciais do projeto. Os dados sugerem que uma avaliação mais precisa da complexidade da integração de dados poderia ajudar a melhorar a precisão das estimativas de duração de projetos futuros.

Embora a redução do tempo necessário seja raramente mencionada, poderia ser um objetivo para empresas inovadoras que buscam ganhar vantagem competitiva por meio de eficiência operacional.

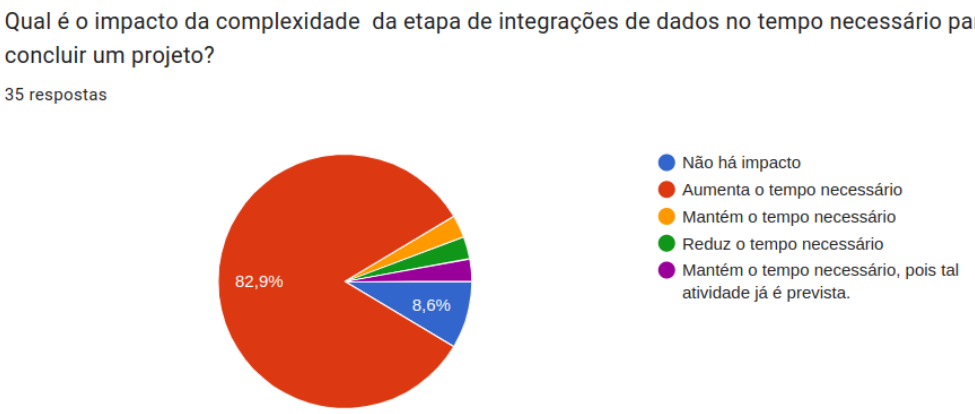


Figura 4.18: Qual é o impacto da complexidade da etapa de integrações de dados no tempo necessário para concluir um projeto?

**Sua empresa atualmente está utilizando *schema matching* automático ou semi automático?**

Indica que 7 respondentes (20%) das empresas estão utilizando *schema matching* automático ou semi automático, enquanto 28 (80%) não estão. Isso mostra que a maioria das empresas na amostra ainda não adotou nenhuma ferramenta.

A maioria das empresas pode estar hesitante em adotar *schema matching* automático ou enfrenta barreiras para implementá-lo, como custos, complexidade ou falta de expertise técnica. Há uma oportunidade significativa para o desenvolvimento e a adoção de ferramentas de *schema matching* automático ou semi automático nas empresas que ainda não o utilizam.

Existe uma necessidade clara de conscientização e educação sobre os benefícios e o uso de *schema matching* automático, bem como treinamento sobre como implementar e utilizar essas ferramentas efetivamente. Para os fornecedores de soluções de integração de dados, o mercado parece estar amplamente aberto para a introdução de ferramentas automatizadas de *schema matching* que possam atender às necessidades dessas empresas.

A falta de uso de *schema matching* automático pode também sugerir que questões de qualidade de dados e compatibilidade entre diferentes sistemas são gerenciadas manualmente ou com menos eficiência. Também a adoção de *schema matching* automático pode se tornar um diferencial competitivo importante para as empresas, o que poderia estimular outras a seguir o exemplo.

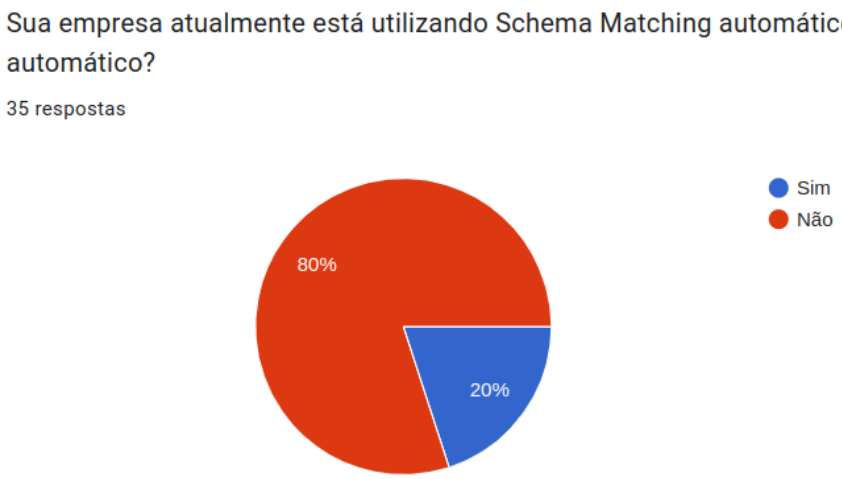


Figura 4.19: Sua empresa atualmente está utilizando *schema matching* automático ou semi automático?

**Qual o principal objetivo ou caso de uso para o *schema matching* na integração de sistemas em sua organização?**

O gráfico da Figura 4.20 fornece informações sobre os principais objetivos ou casos de uso para *schema matching* na integração de sistemas em diferentes organizações. Onde a maioria das organizações utilizam *schema matching* principalmente para construir Data Warehouses, representando cerca de 48,6%, o que sugere que esse é um caso de uso crítico onde a correspondência de esquemas é essencial para consolidar dados de várias fontes.

Outro fator é que a migração de banco de dados (42,9%) é também um caso de uso frequente, indicando que muitas empresas estão em processo de transição entre sistemas de dados ou atualizando suas infraestruturas de dados. Entretanto, uma proporção significativa das empresas está focada em melhorar a qualidade dos dados (31,4%), onde o *schema matching* pode ajudar a garantir a consistência e a precisão dos dados integrados.

A integração de banco de dados também é outro caso de uso importante (45,7%), destacando o papel do *schema matching* na combinação de dados de fontes distintas para análise e relatórios.

Observou-se que poucas organizações relatam o uso de *schema matching* para modificações de esquema em bancos de dados existentes, cerca de 11,4%, o que pode indicar que essas atividades são menos comuns ou realizadas sem a necessidade de correspondência de esquemas.

A exploração e descoberta de dados, bem como a construção de bases de conhecimento (31,4%), são casos de uso representativos, mas com menor frequência em

comparação com outros, o que pode refletir uma aplicação mais especializada de *schema matching*.

Áreas com menor número de respostas, como modificações de esquema (11,4%) e construção de base de conhecimento (20%), podem representar oportunidades de crescimento para o uso de *schema matching* conforme as organizações expandem suas capacidades analíticas.

Visto isto, as respostas indicam que o *schema matching* é uma ferramenta estratégica utilizada em várias facetas da gestão de dados, com ênfase em grandes iniciativas como construção de Data Warehouses e migração de bancos de dados.

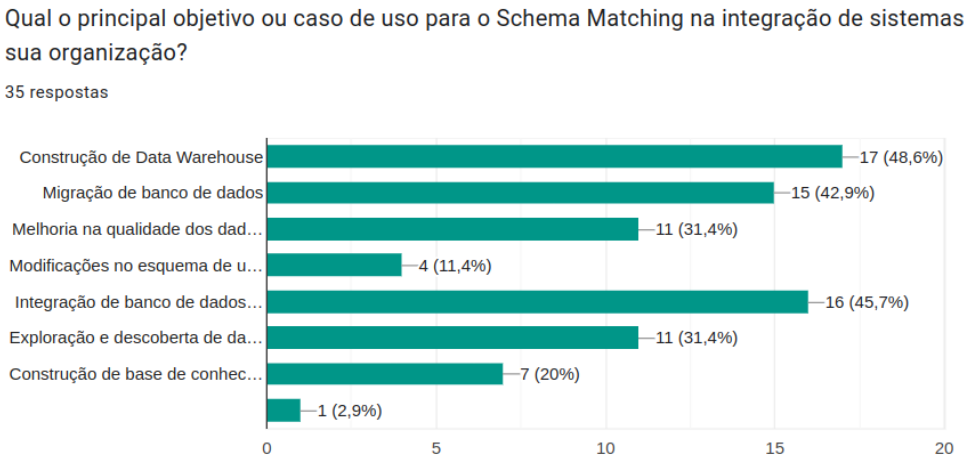


Figura 4.20: Qual o principal objetivo ou caso de uso para o *schema matching* na integração de sistemas em sua organização?

**Como você avaliaria o desempenho do *schema matching* em sua organização em uma escala de 1 a 5, sendo 1 muito ineficaz e 5 muito eficaz?**

O gráfico da Figura 4.21 mostra a avaliação do desempenho do *schema matching* em organizações, com base em uma escala de 1 a 5, onde 1 é “muito ineficaz” e 5 é “muito eficaz”.

Cerca de 34,3% (11) classificou o desempenho do *schema matching* como abaixo da média (notas 1 e 2), indicando que muitas organizações enfrentam desafios na eficiência dessa ferramenta ou processo.

Observa-se uma distribuição relativamente equilibrada nas categorias de avaliação, com um peso ligeiramente maior para as avaliações positivas. Isso pode sugerir que há uma variação significativa nas experiências de desempenho do *schema matching* entre diferentes organizações.

Com 34,3% (12) das organizações atribuindo notas 4 e 5, há claramente espaço para melhorias no *schema matching* em muitas organizações. As avaliações podem refletir uma necessidade de desenvolvimento e treinamento mais aprofundados nas ferramentas e técnicas de *schema matching*.

A complexidade dos processos de integração de dados pode estar impactando negativamente a eficácia do *schema matching*, conforme indicado pelas avaliações mais baixas.

Um número igual de organizações 8,6% (3) sentem que o *schema matching* é muito ineficaz, e 11,4% (4) muito eficaz, o que pode apontar para uma divisão em como diferentes organizações são capazes de implementar e tirar proveito dessa tecnologia.

Pode haver uma correlação entre a eficácia percebida do *schema matching* e as estratégias específicas de integração de dados adotadas pelas organizações, sugerindo que aquelas com processos mais maduros e padronizados podem ter melhores resultados.

O tipo e a qualidade dos dados com os quais as organizações estão trabalhando podem influenciar a avaliação do desempenho do *schema matching*, com dados mais estruturados e padronizados potencialmente levando a uma avaliação mais positiva.

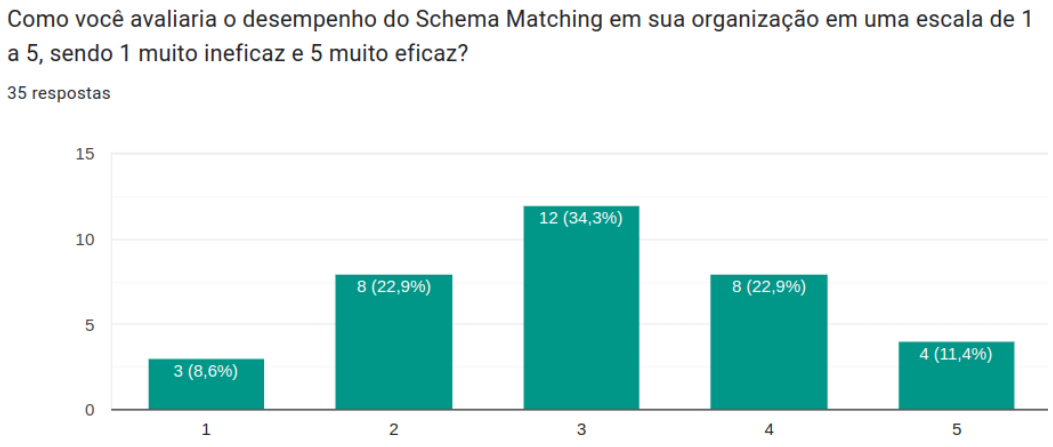


Figura 4.21: Como você avaliaria o desempenho do *schema matching* em sua organização em uma escala de 1 a 5, sendo 1 muito ineficaz e 5 muito eficaz?

**Quais ferramentas ou tecnologias de *schema matching* sua organização utiliza atualmente?**

O gráfico da Figura 4.22 mostra as ferramentas ou tecnologias de *schema matching* utilizadas pelas organizações dos entrevistados. Observa-se que há uma preferência por soluções internas. A maioria das organizações prefere desenvolver suas próprias ferramentas de *schema matching* personalizadas, o que sugere que elas valorizam soluções

que são adaptadas às suas necessidades específicas ou que o mercado ainda não oferece soluções que atendam completamente aos seus requisitos.

Há também o uso limitado de ferramentas comerciais e de código aberto, com uma adoção relativamente baixa, o que pode indicar uma oportunidade para fornecedores de soluções de *schema matching* melhorarem e promoverem suas ofertas para atender às necessidades dessas organizações.

Um número substancial de organizações, cerca de 34,3% (12) não utiliza ferramentas de *schema matching*, o que pode apontar para uma falta de consciência sobre os benefícios dessas ferramentas ou para a satisfação com métodos manuais ou outras abordagens para a integração de dados.

Há um claro potencial para inovação e desenvolvimento de novas ferramentas de *schema matching* que possam ser mais atraentes para as organizações, seja através da facilidade de uso, custo-benefício ou funcionalidades avançadas.

As organizações podem enfrentar barreiras para adotar ferramentas de terceiros, como preocupações com a segurança, integração com sistemas existentes, ou a complexidade de implementação. A alta taxa de desenvolvimento interno de ferramentas personalizadas pode refletir uma necessidade de customização devido à complexidade específica dos dados ou dos processos de negócios das organizações.

Existe uma oportunidade para fornecedores de ferramentas de *schema matching* realizarem parcerias e oferecerem treinamento para organizações que atualmente não utilizam essas tecnologias.

A preferência por ferramentas internas também pode indicar que as organizações veem o desenvolvimento de competências internas em *schema matching* como uma vantagem estratégica.

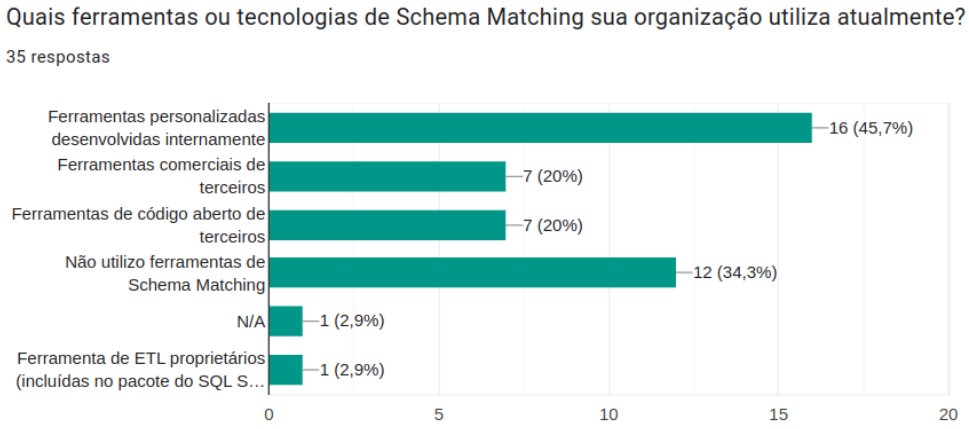


Figura 4.22: Quais ferramentas ou tecnologias de *schema matching* sua organização utiliza atualmente?

**Quais são os principais desafios que sua organização enfrentou ao implementar o uso do *schema matching*?**

O gráfico representa os principais desafios enfrentados por organizações ao implementar o uso do *schema matching*. A maioria das respostas considera a complexidade como o desafio mais significativo, representando cerca de 51,4%, o que destaca a necessidade de abordagens mais eficazes e talvez a automação avançada no processo de *schema matching*. Quase metade das respostas indicam que enfrentam problemas ao integrar *schema matching* com sistemas existentes, cerca de 42,9%, o que pode indicar uma necessidade de soluções mais integradas.

A falta de recursos técnicos ou humanos (37,1%) sugere que muitas organizações podem não ter a capacidade interna suficiente para implementar eficientemente o *schema matching*, o que pode ser uma barreira para a adoção. Outro fator preocupante é a segurança (14,3%), o que pode refletir a importância da proteção de dados e a necessidade de soluções de *schema matching* que sejam seguras por design.

Pode-se inferir também que a falta de iniciativas (2,9%) em tentar padronizar esquemas de dados pode indicar problemas organizacionais ou culturais que impedem a adoção efetiva de *schema matching*. Existe uma clara oportunidade para melhorar as soluções de *schema matching* e o suporte fornecido às organizações, abordando diretamente os desafios mais comuns.

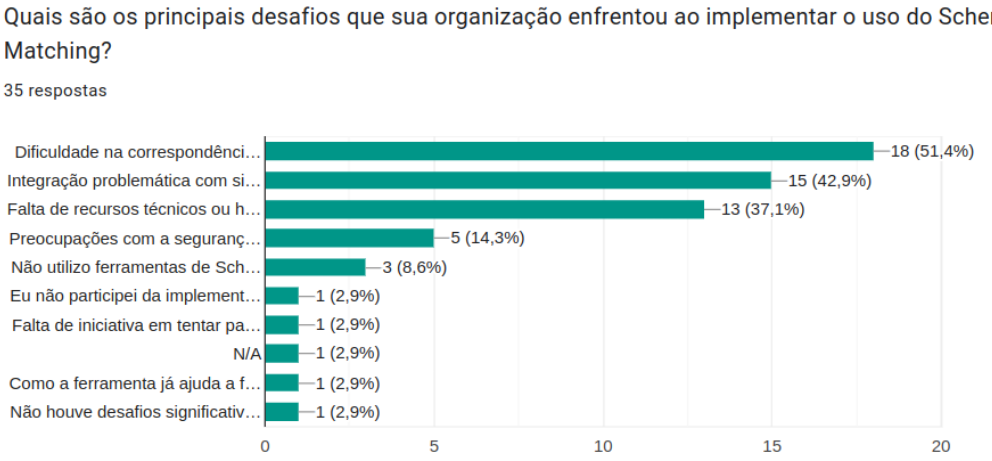


Figura 4.23: Quais são os principais desafios que sua organização enfrentou ao implementar o uso do *schema matching*?

**Em sua opinião, quais são os principais benefícios do uso de *schema matching* automático ou semi automático?**

O gráfico da Figura 4.24 apresenta a opinião sobre os principais benefícios do uso de *schema matching* automático ou semi automático. A maior eficiência no processo de integração de dados e a redução no tempo gasto são os benefícios mais citados, representando cerca de 68,6% e 80% respectivamente, destacando que o *schema matching* automático ou semi automático é valorizado principalmente pela sua capacidade de acelerar e otimizar o processo de integração. Mais da metade dos respondentes (51,4%) reconhece que o *schema matching* automático contribui para uma menor taxa de erro, o que pode levar a uma integração de dados mais confiável.

A facilitação da análise de dados (31,4%) é um benefício importante para aproximadamente um terço dos respondentes, indicando que a automatização do *schema matching* pode contribuir para tornar os dados mais acessíveis e prontos para análise. A melhoria na qualidade dos dados (28,6%), embora não seja o benefício mais citado, ainda é reconhecida como uma vantagem substancial, sugerindo que a automação pode ajudar a garantir que os dados sejam precisos e úteis para as necessidades da empresa.

A maioria dos respondentes concorda que o *schema matching* automático ou semi automático oferece benefícios significativos, com apenas 1 resposta com abstenção (2,9%), o que pode incentivar mais organizações a adotarem essas tecnologias. Os dados também indicam que os fornecedores de ferramentas de *schema matching* devem se concentrar em comunicar como suas soluções podem melhorar a eficiência e reduzir o tempo e os erros na integração de dados.

Mesmo com os benefícios reconhecidos, o treinamento e desenvolvimento contínuos podem ser necessários para maximizar a eficiência e a qualidade que o *schema matching* automático ou semi automático pode trazer.

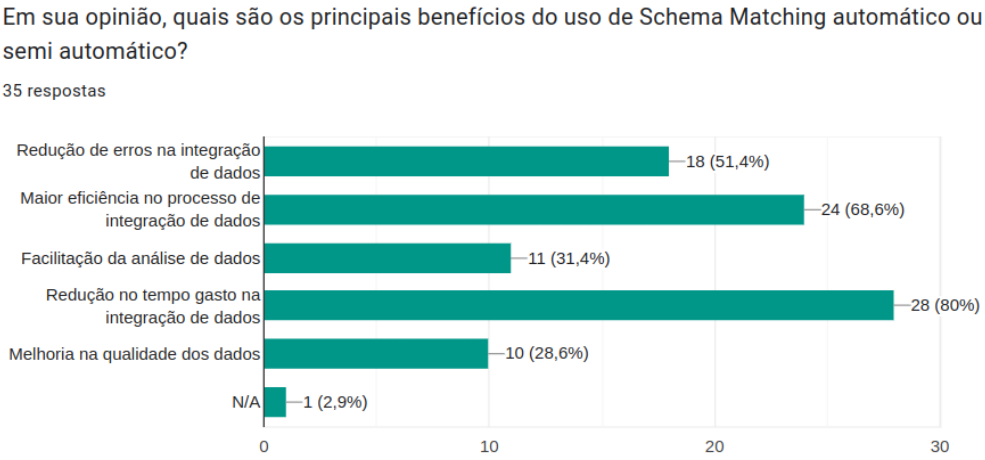


Figura 4.24: Em sua opinião, quais são os principais benefícios do uso de *schema matching* automático ou semi automático?

**Quais são as principais razões para a não utilização do *schema matching*?**

Detalha as principais razões pelas quais as organizações não utilizam *schema matching*. A maior barreira para a utilização do *schema matching* é a falta de conhecimento, representando cerca de 51,4%, o que sugere uma grande necessidade de educação e treinamento na área. A desconfiança na precisão das ferramentas (14,3%) também indica que a confiabilidade é uma preocupação significativa para as organizações.

Outro fator é o custo e a falta de recursos (28,6%), que são fatores limitantes para algumas organizações, sugerindo que pode ser percebido como um investimento significativo.

Entretanto, a alta complexidade é uma das razões menos citada, juntamente com o custo elevado, representando certa de 8,6%, o que pode indicar que embora seja um problema para algumas organizações, não é a principal preocupação para a maioria delas.

Há uma oportunidade para fornecedores de ferramentas de *schema matching* promoverem o valor e a eficiência de suas soluções para superar a falta de conhecimento e confiança. E também a diversidade das razões aponta para a necessidade de abordagens personalizadas para incentivar a adoção do *schema matching*, considerando as necessidades específicas de cada organização.

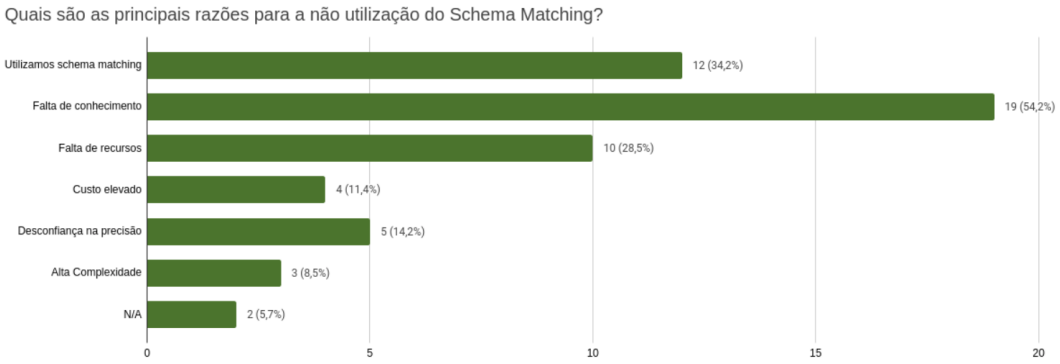


Figura 4.25: Quais são as principais razões para a não utilização do *schema matching*?

**Quais ferramentas de *schema matching* você utilizou em projetos de integração de dados?**

O gráfico da Figura 4.26 apresenta as ferramentas de *schema matching* utilizadas em projetos de integração de dados por diferentes organizações. Ferramentas de grandes fornecedores como Microsoft SQL Server Data Tools (SSDT) (25,7%), Oracle Data Integrator (ODI) (17,1%) e Talend (17,1%), são as mais usadas, o que sugere uma preferência por soluções estabelecidas, com uma grande comunidade contribuidora e consequentemente mais confiáveis no mercado.

A variedade de ferramentas mencionadas indica uma diversidade de soluções utilizadas e que não há preferências de soluções entre as organizações quando se trata de *schema matching*. Entretanto uma parcela de respostas indicam o uso de soluções desenvolvidas internamente, o que pode refletir uma necessidade de personalização ou de integração com sistemas internos específicos.

A existência de respostas onde não utilizam ferramentas de *schema matching* sugere que pode haver subutilização de tais ferramentas ou falta de conhecimento sobre as mesmas. Há também a presença de ferramentas de código aberto, mostra que há uma adesão significativa a soluções deste segmento, embora não seja tão predominante quanto as soluções de grandes fornecedores.

A variedade de respostas sugere que há espaço para expansão no mercado de ferramentas de *schema matching*, tanto para soluções comerciais quanto de código aberto.

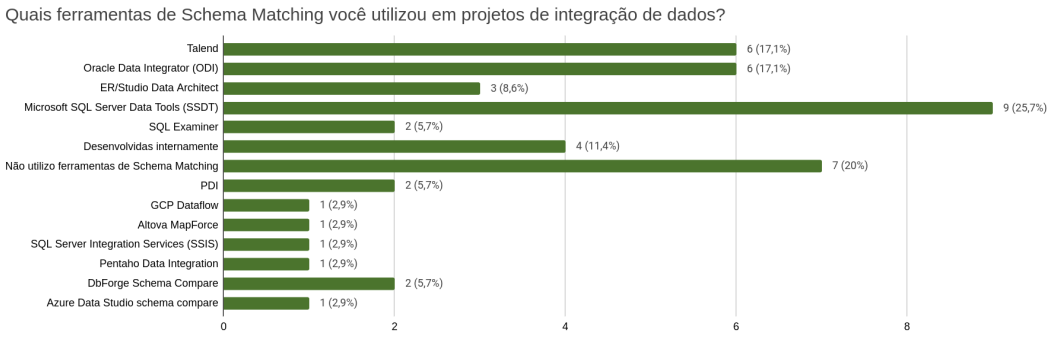


Figura 4.26: Quais ferramentas de *schema matching* você utilizou em projetos de integração de dados?

**Como sua empresa lida com conflitos de correspondência ao usar *schema matching*?**

O gráfico da Figura 4.27 apresenta como as empresas lidam com conflitos de correspondência ao usar *schema matching*. Em que, a grande maioria das empresas ainda lida com conflitos de correspondência de forma manual, o que sugere que, apesar da disponibilidade de ferramentas automatizadas, a intervenção humana é frequentemente preferida ou visto como necessário para a resolução de conflitos complexos.

É possível observar também que ferramentas de resolução de conflitos automáticas ou semi-automáticas são subutilizadas. Isso pode ser devido à falta de conhecimento dessas ferramentas, ou mesmo à preferência por controle manual sobre processos automatizados.

Há uma pequena porcentagem (8,6%) de empresas que afirmam não enfrentar conflitos durante o *schema matching*, o que pode indicar processos bem estabelecidos ou um domínio de dados mais simples. Pode-se observar também que há uma oportunidade

para melhorar e promover ferramentas de resolução de conflitos que possam integrar automação com a flexibilidade e o discernimento dos métodos manuais. A predominância do método manual pode também refletir uma necessidade de educação e treinamento sobre as ferramentas automatizadas disponíveis no mercado.

Existe também um potencial para inovação no desenvolvimento de ferramentas de resolução de conflitos que sejam mais intuitivas, confiáveis e fáceis de integrar nos fluxos de trabalho existentes. Uma vez que o custo e os recursos necessários para implementar e manter ferramentas automatizadas podem ser um fator limitante para algumas empresas.

A dependência de processos manuais pode indicar uma preocupação com a qualidade e a confiabilidade da resolução de conflitos realizada por ferramentas automatizadas.

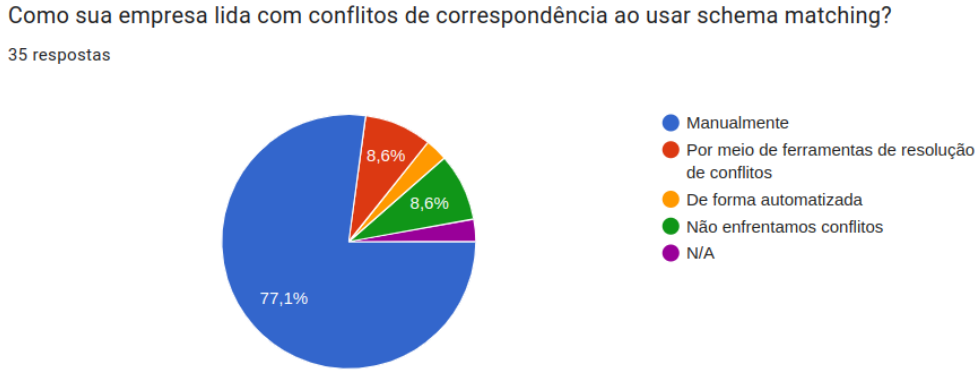


Figura 4.27: Como sua empresa lida com conflitos de correspondência ao usar *schema matching*?

**Sua empresa tem estratégias ou planos para melhorar o uso de *schema matching* no futuro? Se sim, quais são esses planos?**

O gráfico mostra a distribuição das respostas sobre seus planos para melhorar o uso de *schema matching* no futuro, onde uma parcela das empresas estão buscando adotar novas ferramentas para melhorar o uso de *schema matching*, cerca de 28,6%, o que indica uma tendência para a modernização e busca por soluções mais eficientes. Há também um foco considerável no aprimoramento das habilidades da equipe (20%), sugerindo que as empresas reconhecem a importância do desenvolvimento profissional contínuo e do treinamento especializado.

Cerca de 45,7% das empresas não tem planos de mudar suas práticas atuais, o que pode refletir satisfação com os sistemas atuais ou uma resistência a alterações no processo de trabalho existente. A falta de conhecimento sobre os planos futuros ou a ausência de respostas sugere que pode haver incerteza dentro de algumas organizações sobre como

ou se devem evoluir suas práticas de *schema matching*. O interesse em novas ferramentas destaca oportunidades para fornecedores de soluções de *schema matching* apresentarem suas inovações ao mercado. As empresas também podem precisar de estratégias mais claras e definidas para o uso futuro de *schema matching*, abordando tanto a tecnologia quanto o desenvolvimento de competências.

A liderança dentro das empresas pode precisar de mais informações para tomar decisões informadas sobre o investimento em *schema matching*, como evidenciado pela parcela de respondentes que não têm certeza sobre os planos futuros.

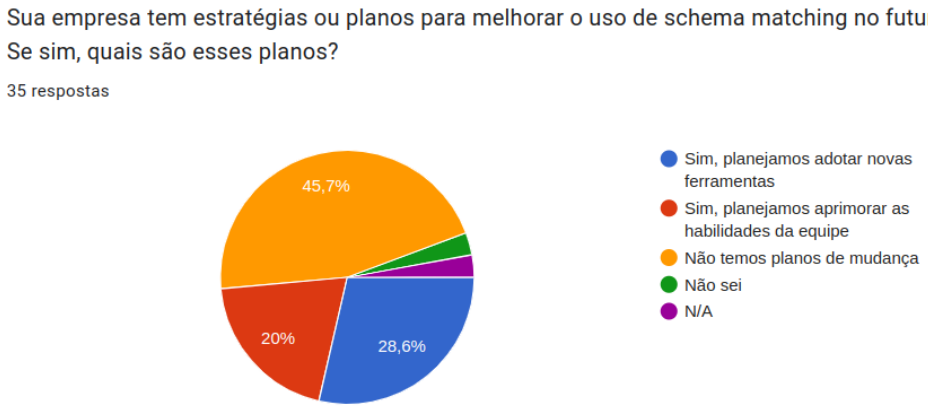


Figura 4.28: Sua empresa tem estratégias ou planos para melhorar o uso de *schema matching* no futuro? Se sim, quais são esses planos?

**Que práticas sua empresa adota para documentar e monitorar as correspondências de esquemas ao longo do tempo?**

O gráfico da Figura 4.29 apresenta as práticas adotadas por empresas para documentar e monitorar as correspondências de esquemas ao longo do tempo, onde quase metade das empresas enfatizam a documentação de esquemas como uma prática fundamental, indicando que é uma prioridade manter um registro detalhado das estruturas de dados. Já 5,7% (2) das empresas realizam monitoramento contínuo das correspondências de esquemas, o que pode refletir um esforço para manter a integridade dos dados ao longo do tempo e responder prontamente a mudanças. Uma pequena fração de 11,4% (4) das empresas mantém registros históricos, o que pode sugerir que a prática não é tão difundida ou valorizada quanto a documentação atualizada e o monitoramento contínuo.

A existência de empresas que não documentam as correspondências de esquemas destaca uma área de risco potencial e uma oportunidade de melhoria na gestão de dados. A falta de documentação pode também indicar desafios de conformidade, especialmente em indústrias regulamentadas onde a documentação completa é geralmente necessária. Há

espaço para melhorar as práticas de documentação e monitoramento, dada a importância dessas atividades para a gestão de dados de longo prazo.

O gráfico também sugere uma oportunidade de mercado para ferramentas que possam facilitar e automatizar a documentação e o monitoramento de correspondências de esquemas. As estratégias para documentar e monitorar as correspondências de esquemas variam, o que pode refletir diferentes níveis de maturidade das práticas de governança de dados entre as empresas.

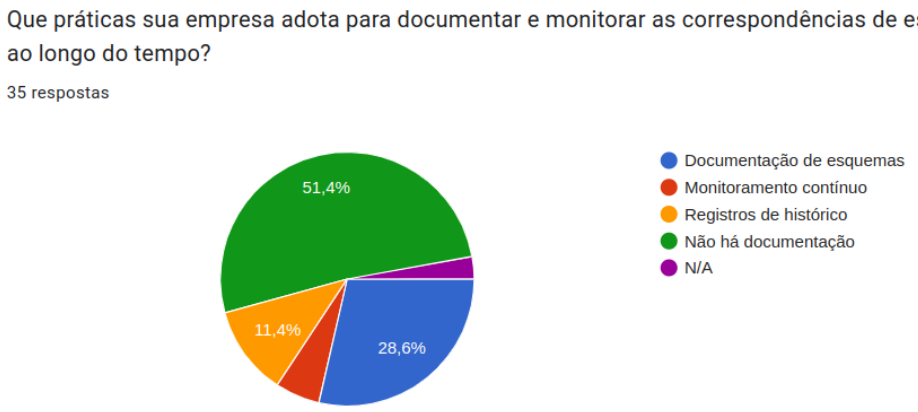


Figura 4.29: Que práticas sua empresa adota para documentar e monitorar as correspondências de esquemas ao longo do tempo?

**Sua empresa realiza alguma validação ou verificação posterior ao processo de *schema matching* para garantir a precisão e a integridade dos dados?**

O gráfico da Figura 4.30 mostra como as empresas realizam validação ou verificação após o processo de *schema matching* para garantir a precisão e a integridade dos dados.

A maior parte das empresas, cerca de 57,1% (20) não realiza validações regulares após o *schema matching*, indicando que a validação não é uma etapa crítica no processo de integração de dados para muitas organizações. Além de 37,1% (13) das empresas realizam validação após o *schema matching*, o que pode sugerir uma oportunidade pelas empresas em melhorar as práticas de garantia de qualidade. Há também a existência de respostas que indicam desconhecimento sobre a realização de validação pode refletir uma falta de comunicação interna ou uma compreensão limitada dos processos de gestão de dados.

As respostas indicam também uma oportunidade para consultores e educadores enfatizarem a importância da validação de dados após o *schema matching*. A adoção de validações regulares sugere uma demanda por ferramentas que possam automatizar e simplificar este processo. A preocupação com a precisão e a integridade dos dados é

evidente, destacando a importância da validação como uma etapa no ciclo de vida da integração de dados.

Contudo, a validação regular pode fazer parte de uma estratégia de melhoria contínua, na qual as empresas buscam otimizar e aperfeiçoar constantemente seus processos de integração de dados.

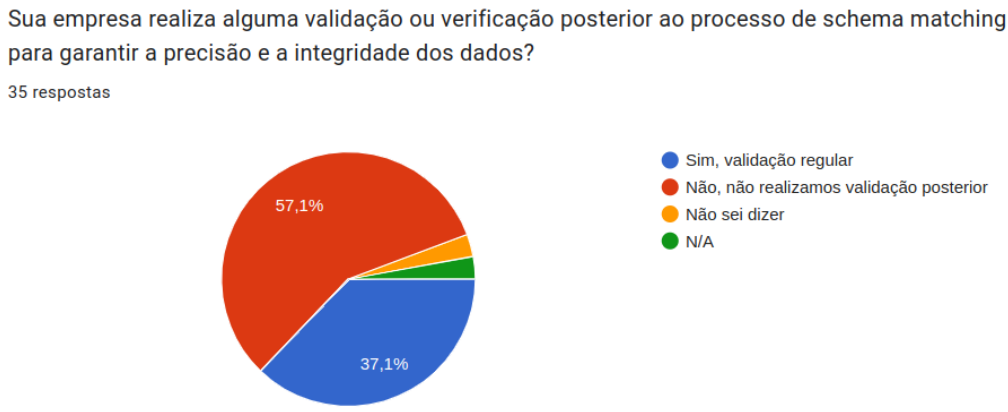


Figura 4.30: Sua empresa realiza alguma validação ou verificação posterior ao processo de *schema matching* para garantir a precisão e a integridade dos dados?

**Em uma escala de 1 a 5, onde 1 representa "totalmente insatisfeito"e 5 representa "totalmente satisfeito", quão satisfeito você está com o processo de *schema matching* em sua empresa?**

Há um nível notável de insatisfação (40% das respostas nas notas 1 e 2), o que sugere que muitas empresas encontram dificuldades ou estão descontentes com seus processos atuais de *schema matching*. Já a maior porcentagem de respostas está no meio da escala (nota 3), indicando sentimentos mistos ou uma avaliação mediana da satisfação com o *schema matching*. Isso pode refletir uma eficácia variável ou inconsistente dos processos de *schema matching*. E apenas uma pequena porção das empresas está totalmente satisfeita (nota 4 e 5), o que pode indicar que há espaço significativo para melhorias no processo de *schema matching*.

A distribuição das respostas sugere que há potencial para otimização nos processos de *schema matching*, com um foco em aumentar a satisfação geral. A insatisfação e as avaliações médias podem refletir a necessidade de melhores ferramentas, práticas mais eficientes ou treinamento adicional para os envolvidos no processo de *schema matching*. A presença de insatisfação pode ser vista como uma oportunidade para inovação no campo de *schema matching*, seja através do desenvolvimento de novas ferramentas, aprimoramento das existentes, ou oferecendo serviços especializados de consultoria.

Empresas podem beneficiar-se de avaliações contínuas e feedback sobre seus processos de *schema matching* para identificar áreas de melhoria e alinhar suas estratégias com as expectativas dos usuários. As respostas indicam uma necessidade de estratégias de melhoria focadas em aumentar a satisfação com o *schema matching*, o que pode incluir a revisão de processos, adoção de novas tecnologias ou aumento da colaboração entre equipes.

O gráfico da Figura 4.31 fornece uma visão geral da satisfação com o processo de *schema matching* em diferentes empresas, em uma escala de 1 a 5, onde 1 é "totalmente insatisfeito"e 5 é "totalmente satisfeito".

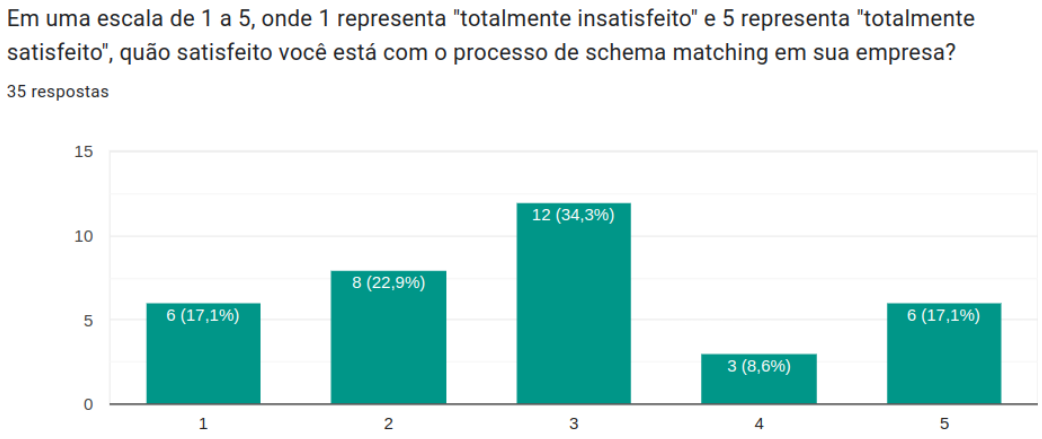


Figura 4.31: Quão satisfeito você está com o processo de *schema matching* em sua empresa?

**Quais ferramentas de *schema mapping* (Transformação) você utiliza na sua empresa?**

O gráfico da Figura 4.32 representa as ferramentas de *schema mapping* (Transformação) utilizadas pelas empresas dos entrevistados. Pentaho PDI domina claramente como a ferramenta de escolha, possivelmente devido à sua funcionalidade robusta, comunidade ativa ou custo-benefício. Existe uma diversidade de ferramentas utilizadas, mas muitas têm uma adoção limitada (4% ou 8%), o que pode indicar preferências específicas de negócio ou requisitos técnicos. Além do Pentaho PDI, ferramentas de código aberto como Apache NiFi e Apache Camel estão sendo utilizadas, refletindo uma tendência de adotar soluções de código aberto. Microsoft SQL Server Integration Services (SSIS) também é relativamente bem utilizado, mostrando que ferramentas especializadas ainda têm um lugar significativo em algumas organizações.

Há espaço no mercado para novas ferramentas ou para a expansão de ferramentas existentes, dada a variedade de ferramentas com baixa adoção. Com a utilização de várias ferramentas diferentes, há uma necessidade contínua de capacitação e desenvolvimento profissional para que as equipes possam utilizar efetivamente essas tecnologias. A seleção de ferramentas pode ser influenciada pelos recursos de TI existentes, experiência da equipe e integração com sistemas atuais.

Algumas empresas optam por desenvolver soluções internas ou não adotar ferramentas específicas de *schema mapping*, o que pode indicar necessidades únicas que as soluções de mercado não atendem.

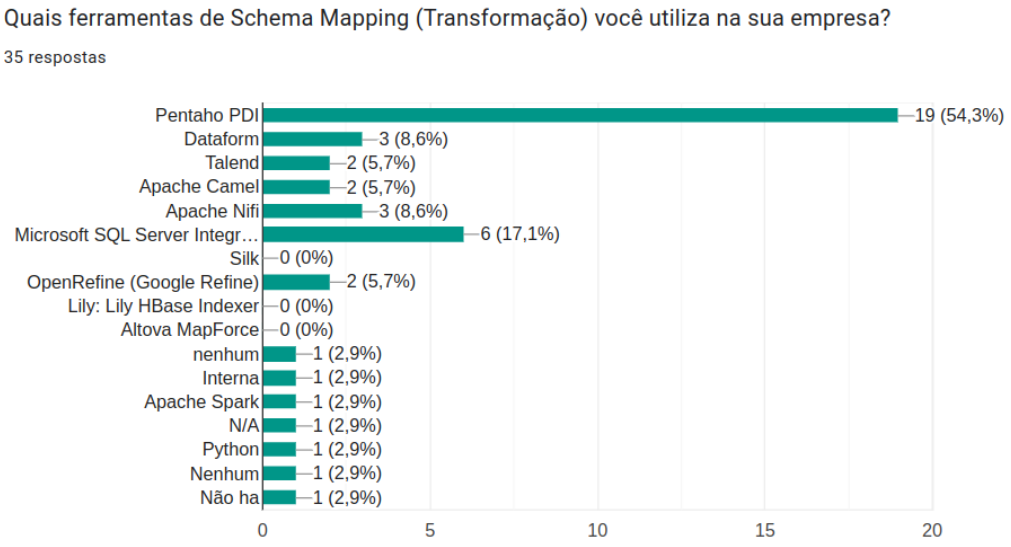


Figura 4.32: Quais ferramentas de *schema mapping* (Transformação) você utiliza na sua empresa?

Nas próximas seções, serão abordadas as ameaças e as considerações finais, com base em [Lebtag et al. 2022], apresentando uma síntese dos resultados do survey.

### 4.2.2 Ameaças à Validade

Nesta seção, serão discutidas as principais ameaças à validade das etapas envolvidas neste estudo e como mitigá-las.

**Limitações relacionadas à pesquisa com profissionais da indústria de software brasileira:**

**População:** Uma população de 35 profissionais da indústria de software brasileira foi consultada neste estudo. Para maximizar o número de participantes, foi feito o

convite ao maior número possível de profissionais nacionais dentro de um período fixo. Buscou-se obter uma representação diversificada das diferentes realidades regionais e nacionais. No entanto, o estudo foi limitado pelo número de respostas, uma dificuldade também observada em outros estudos de engenharia de software[Smith et al. 2013]. Além disso, o número e o tipo de problemas apresentados aos participantes podem ser considerados limitações do estudo, o que limita o potencial de generalização dos resultados.

**Fadiga:** A generalização dos resultados para toda a indústria de software brasileira pode ser limitada devido a pequena parcela de participantes, as empresas envolvidas e até aos níveis de experiência dos profissionais.

**Limitações relacionadas as inferências:**

**Análises e Inferências:** As inferências obtidas também estão inerentemente sujeitas a interpretações e opiniões humanas. Para mitigar esta ameaça, três engenheiros de software e especialistas em integração de dados apoiaram a elaboração e avaliação das inferências.

### 4.2.3 Considerações Finais do Capítulo

As principais contribuições deste *survey* foram comunicar os resultados da pesquisa realizada para responder as questões de pesquisa. As respostas à pesquisa foram analisadas quantitativa e qualitativamente. De um conjunto de 35 participantes, mais da metade (precisamente 23 participantes) tem familiaridade com o conceito de *schema matching*. Além disso, vários participantes (mesmo os menos experientes em *schema matching*) identificaram os benefícios na redução no tempo gasto com a integração de dados, redução de erros e uma maior eficiência no uso da técnica de forma automática. A análise realizada permitiu identificar que (i) os profissionais vislumbram diversas vantagens e oportunidades para aplicar o *schema matching* automático na integração de dados, (ii) o tamanho e a diversidade da base de dados deve fornecer um custo-benefício eficaz para a aplicação da técnica antes de sua adoção, e (iii) qualquer estratégia de adoção bem-sucedida deve ser avaliada de acordo com a complexidade da integração de dados.

Com base nesses resultados, foi possível compilar uma lista agrupada para responder às questões de pesquisa, originada das respostas fornecidas pelos participantes da pesquisa.

As respostas foram categorizadas em quatro grupos principais: (i) complexidade das bases de dados, (ii) benefícios percebidos, (iii) desafios e dificuldades encontrados, e (iv) nível de satisfação com o uso do *schema matching*.

A primeira categoria, complexidade das bases de dados, analisa o grau de complexidade envolvido na integração das bases de dados. Os resultados indicam que

respondentes envolvidos em projetos de maior complexidade tendem a perceber de forma mais clara os benefícios do uso de técnicas de *schema matching*. Isso sugere que, em contextos onde a integração de dados é mais desafiadora, a aplicação dessas técnicas pode resultar em melhorias significativas na eficiência dos processos.

Os benefícios identificados pelos participantes estão predominantemente associados à otimização de tempo. A aplicação de *schema matching* automático reduz o tempo gasto na integração de dados, facilitando a conciliação de esquemas divergentes e acelerando a disponibilidade de informações para tomada de decisão. Essa eficiência é especialmente valorizada em ambientes de negócios onde a rapidez na manipulação de grandes volumes de dados é crucial.

Quanto aos desafios e dificuldades, estes foram frequentemente relacionados à complexidade dos esquemas de dados envolvidos, à presença de dados inconsistentes que dificultam a integração automática, e à falta de treinamento específico para as equipes técnicas. A complexidade dos esquemas muitas vezes exige abordagens mais sofisticadas, como a intervenção humana em conjunto com o *schema matching* automático, enquanto a inconsistência dos dados pode demandar etapas adicionais de limpeza e validação, aumentando o esforço necessário para a integração eficaz.

Por fim, o nível de satisfação dos respondentes com o uso do *schema matching* automático revelou-se majoritariamente positivo. De um total de 35 respostas, 21 profissionais expressaram satisfação, destacando o impacto positivo da técnica na simplificação de processos e na melhoria da qualidade dos dados integrados. Esta percepção positiva reflete o potencial do *schema matching* como uma ferramenta valiosa para enfrentar os desafios da integração de dados em ambientes corporativos diversificados.

Essas descobertas fornecem uma visão importante sobre como as técnicas de *schema matching* são percebidas e aplicadas na indústria, ressaltando tanto as oportunidades quanto os desafios que moldam sua adoção.

CAPÍTULO 5

# Conclusão

Este trabalho explorou o estado da prática de *schema matching* em processos de integração de dados na indústria brasileira de software, evidenciando tanto os desafios como as potenciais estratégias de melhoria. A integração eficaz de dados é um componente crítico da infraestrutura tecnológica moderna, permitindo que entidades diversas — desde corporações até instituições acadêmicas — sintetizem e explorem vastos repositórios de conhecimento. A adoção de técnicas de *schema matching*, reforçadas por métodos avançados de Inteligência Artificial, emergiu como uma solução viável para superar as discrepâncias semânticas inerentes aos esquemas de dados heterogêneos.

Através de um *survey*, esta pesquisa identificou que a maioria dos respondentes situava-se em Goiás, possivelmente pela proximidade dos pesquisadores. Embora também tenha sido possível alcançar participantes de outras regiões brasileiras, como Rio de Janeiro e São Paulo, o número de respostas foi limitado. Portanto, os dados obtidos não permitem generalizações robustas sobre a distribuição geográfica das estratégias de integração de dados no país. Ainda assim, os resultados sugerem a necessidade de futuras investigações com amostras mais amplas para compreender melhor o panorama regional e o estado de inovação e recursos de TI em diferentes localidades.

Além disso, a pesquisa destacou a importância de aplicar abordagens automatizadas e semi-automatizadas de *schema matching*. O questionário revelou um alto nível de satisfação entre as organizações que utilizam a técnica, demonstrando que ela representa um avanço significativo em relação aos métodos manuais tradicionais, proporcionando maior eficiência e precisão na integração de dados.

No entanto, a implementação dessas técnicas enfrenta barreiras significativas, como a escassez de profissionais qualificados, a falta de interesse da maioria das organizações em investir em ferramentas e fornecer treinamento, resistência às mudanças nos fluxos de trabalho estabelecidos, além das limitações das ferramentas atualmente disponíveis. Compreender essas barreiras, a partir do *survey* e das análises realizadas, oferece um caminho para a elaboração de estratégias que promovam uma adoção mais ampla das práticas de *schema matching* no Brasil.

Reconhecendo as limitações deste estudo, principalmente no que tange à amostra

do *survey* e à possibilidade de vieses regionais, recomenda-se a condução de pesquisas adicionais com amostras mais extensas e diversificadas. A exploração de casos de estudo em diferentes contextos e indústrias poderá fornecer uma visão mais holística sobre a adoção de *schema matching* e sua eficácia.

A análise dos dados coletados aponta para um cenário onde a eficiência operacional e a precisão da integração de dados são potencialmente aumentadas pela implementação de *schema matching*. A adoção de técnicas automáticas e semi-automáticas de *schema matching* também se alinha com a tendência global de transformação digital, onde a otimização de processos mediante a automação e a inteligência de dados constitui uma vantagem competitiva estratégica. Os desafios identificados, inclusive a falta de expertise técnica e as barreiras culturais, refletem uma oportunidade para o desenvolvimento de programas educacionais e de conscientização, visando fomentar a capacitação e a inovação tecnológica em todo o Brasil.

Através dos resultados do *survey* e do mapeamento sistemático da literatura, este estudo contribuiu para uma melhor compreensão das práticas atuais e das perspectivas de *schema matching* na indústria brasileira de software. A aplicação de abordagens baseadas em Inteligência Artificial emerge como uma ferramenta poderosa, oferecendo caminhos para a otimização de processos de integração de dados e para o manejo efetivo da complexidade inerente aos esquemas heterogêneos. No entanto, a implementação efetiva dessas técnicas requer um ecossistema de suporte robusto, que inclua treinamento especializado, desenvolvimento de ferramentas adequadas e um ambiente organizacional adaptativo.

Embora a presente pesquisa tenha delineado o panorama do *schema matching* no Brasil, reconhece-se a necessidade de investigações futuras que possam abordar as práticas de *schema matching* em contextos variados, além de estender a compreensão sobre a aplicabilidade e os benefícios das técnicas de Inteligência Artificial em diversas indústrias. Seria particularmente benéfico analisar como as práticas de *schema matching* são influenciadas pelas particularidades de cada setor e como as soluções automatizadas podem ser personalizadas para atender a esses contextos específicos.

Adicionalmente, sugere-se o desenvolvimento de estudos longitudinais que possam acompanhar a evolução da adoção de *schema matching* ao longo do tempo, assim como pesquisas que possam avaliar o impacto econômico e estratégico dessa adoção nas organizações. Tal conhecimento é vital para a tomada de decisão estratégica, para a priorização de investimentos em TI e para a formulação de políticas públicas que visem à promoção de uma indústria de software mais inovadora e competitiva no Brasil.

Em vista dos resultados obtidos, conclui-se que a aplicação do *schema matching* proporciona melhorias significativas nos processos de integração de dados, como a redução do tempo gasto e o aumento da assertividade. No entanto, é crucial que empresas

e a comunidade acadêmica se mobilizem para reduzir as barreiras à adoção dessas técnicas e capacitar os profissionais. A implementação de uma ferramenta para apoiar o processo de *schema matching* é de fundamental importância.

Durante a aplicação do questionário, observou-se que a maioria dos respondentes desconhecia o termo em estudo. Alguns até utilizavam ferramentas de *schema matching*, mas não estavam familiarizados com a nomenclatura. Outra conclusão importante é que a maioria das ferramentas de *schema matching* não são gratuitas ou open source, o que pode explicar a baixa adoção, especialmente considerando que a maioria das empresas não tem planos para implementar *schema matching* no futuro.

## 5.1 Produção Científica

Um artigo intitulado "O Uso da Inteligência Artificial no *schema matching*: Um Mapeamento Sistemático" foi publicado e apresentado na XI Escola Regional de Informática de Goiás (ERI-GO 2023)[1]. Além disso, este mesmo artigo foi selecionado para submissão em uma versão estendida na revista TECNIA-IFG[2], classificada como B1 no Qualis Capes. Planeja-se também submeter um artigo sobre os resultados do survey ao XXI Simpósio Brasileiro de Sistemas de Informação (SBSI) 2025[3].

## 5.2 Trabalhos Futuros

Para trabalhos futuros, podem ser exploradas diversas direções. Uma possibilidade é realizar provas de conceito utilizando ferramentas de *schema matching* open source, demonstrando sua eficácia em cenários reais e facilitando sua adoção. Além disso, é essencial desenvolver um estudo para adotar formas da promoção do conhecimento sobre *schema matching* dentro das organizações, podendo ser através de workshops, seminários e materiais educacionais. Um estudo comparativo detalhado entre ferramentas de *schema matching* automáticas e semi-automáticas, avaliando aspectos como precisão, eficiência e custos. Análise de casos de uso reais em diferentes indústrias, fornecendo *insights* sobre desafios e melhores práticas. O desenvolvimento de frameworks para avaliar a eficácia das ferramentas de *schema matching*. Além disso, a implementação de projetos piloto em organizações e o estudo das barreiras organizacionais e culturais que afetam a adoção de *schema matching* podem oferecer novas perspectivas. A criação de ferramentas de suporte e capacitação para profissionais, bem como a avaliação do impacto de tecno-

[1] http://erigo.sbc.org.br/
[2] https://periodicos.ifg.edu.br/tecnia/index
[3] https://sbbd.org.br/2024/

logias emergentes, sobre as práticas de *schema matching*, também são áreas de interesse para futuras pesquisas.

# Referências Bibliográficas


[AISSAOUI e OUGHDIR 2020]AISSAOUI, O. E.; OUGHDIR, L. A learning style-based ontology matching to enhance learning resources recommendation. In: IEEE. *2020 1st international conference on innovative research in applied science, engineering and technology (IRASET)*. [S.l.], 2020. p. 1–7.

[Amrouch, Mostefai e Fahad 2016]AMROUCH, S.; MOSTEFAI, S.; FAHAD, M. Decision trees in automatic ontology matching. *International Journal of Metadata, Semantics and Ontologies*, Inderscience Publishers (IEL), v. 11, n. 3, p. 180–190, 2016.

[Ayala et al. 2022]AYALA, D. et al. Leapme: Learning-based property matching with embeddings. *Data & Knowledge Engineering*, Elsevier, v. 137, p. 101943, 2022.

[Azeroual, Saake e Schallehn 2018]AZEROUAL, O.; SAAKE, G.; SCHALLEHN, E. Analyzing data quality issues in research information systems via data profiling. *International Journal of Information Management*, v. 41, p. 50–56, 2018. ISSN 0268-4012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0268401218300975>.

[Belhadi et al. 2023]BELHADI, A. et al. Fast and accurate framework for ontology matching in web of things. *ACM Transactions on Asian and Low-Resource Language Information Processing*, ACM New York, NY, v. 22, n. 5, p. 1–19, 2023.

[Bento, Zouaq e Gagnon 2020]BENTO, A.; ZOUAQ, A.; GAGNON, M. Ontology matching using convolutional neural networks. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. [S.l.: s.n.], 2020. p. 5648–5653.

[Berlin e Motro 2002]BERLIN, J.; MOTRO, A. Database schema matching using machine learning with feature selection. In: SPRINGER. *Advanced Information Systems Engineering: 14th International Conference, CAiSE 2002 Toronto, Canada, May 27–31, 2002 Proceedings 14*. [S.l.], 2002. p. 452–466.

[Bilke e Naumann 2005]BILKE, A.; NAUMANN, F. Schema matching using duplicates. p. 69–80, 2005.

[Borges, Ribeiro e Neto 2023]BORGES, R. H.; RIBEIRO, L. A.; NETO, V. V. G. O uso da inteligência artificial no schema matching: Um mapeamento sistemático. In: SBC. *Anais da XI Escola Regional de Informática de Goiás*. [S.l.], 2023.

[Bulygin 2018]BULYGIN, L. Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem. In: *Proc. Int. Conf. Data Anal. Manage. Data Intensive Domains (DAMDID/RCDL)*. [S.l.: s.n.], 2018. p. 245–249.

[Bulygin e Stupnikov 2019]BULYGIN, L.; STUPNIKOV, S. A. Applying of machine learning techniques to combine string-based, language-based and structure-based similarity measures for ontology matching. In: *DAMDID/RCDL*. [S.l.: s.n.], 2019. p. 129–147.

[Calvanese et al. 2001]CALVANESE, D. et al. Data integration in data warehousing. *International Journal of Cooperative Information Systems*, World Scientific, v. 10, n. 03, p. 237–271, 2001.

[Cappuzzo 2020]CAPPUZZO, R. Creating embeddings of heterogeneous relational datasets for data integration tasks. In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. [S.l.: s.n.], 2020. p. 1335–1349.

[Chakraborty et al. 2021]CHAKRABORTY, J. et al. Ontoconnect: Unsupervised ontology alignment with recursive neural network. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. [S.l.: s.n.], 2021. p. 1874–1882.

[Chen et al. 2021]CHEN, J. et al. Augmenting ontology alignment by semantic embedding and distant supervision. In: SPRINGER. *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*. [S.l.], 2021. p. 392–408.

[Cherradi 2024]CHERRADI, M. Data lakehouse: Next generation information system. In: *Seminars in Medical Writing and Education*. [S.l.: s.n.], 2024. v. 3, p. 67–67.

[Date 2004]DATE, C. J. *Introdução a sistemas de bancos de dados*. [S.l.]: Elsevier Brasil, 2004.

[Dhouib, Zucker e Tettamanzi 2019]DHOUIB, M. T.; ZUCKER, C. F.; TETTAMANZI, A. G. An ontology alignment approach combining word embedding and the radius measure. In: SPRINGER INTERNATIONAL PUBLISHING. *Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings 15*. [S.l.], 2019. p. 191–197.

[Do 2006]DO, H.-H. Schema matching and mapping-based data integration. 2006.

[Doan, Halevy e Ives 2012]DOAN, A.; HALEVY, A.; IVES, Z. *Principles of data integration*. [S.l.]: Elsevier, 2012.

[Doan et al. 2020]DOAN, A. et al. Magellan: toward building ecosystems of entity matching solutions. *Communications of the ACM*, ACM New York, NY, USA, v. 63, n. 8, p. 83–91, 2020.

[Domingos 2012]DOMINGOS, P. A few useful things to know about machine learning. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 55, n. 10, p. 78–87, oct 2012. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/2347736.2347755>.

[Dyba, Kitchenham e Jorgensen 2005]DYBA, T.; KITCHENHAM, B. A.; JORGENSEN, M. Evidence-based software engineering for practitioners. *IEEE software*, IEEE, v. 22, n. 1, p. 58–65, 2005.

[Fabbri et al. 2013]FABBRI, S. C. P. F. et al. Externalising tacit knowledge of the systematic review process. *IET Softw.*, v. 7, n. 6, p. 298–307, 2013. Disponível em: <https://doi.org/10.1049/iet-sen.2013.0029>.

[Galvão, Pansani e Harrad 2015]GALVãO, T. F.; PANSANI, T. de S. A.; HARRAD, D. Principais itens para relatar revisões sistemáticas e meta-análises: A recomendação PRISMA. *Epidemiol. Serv. Saúde*, scielo, v. 24, p. 335 – 342, 06 2015. ISSN 2237-9622. Disponível em: <http://www.scielo.br/scielo.php?script=sci$_arttextpid = S2237 - 96222015000200335nrm = iso$>.

[Giordano 2010]GIORDANO, A. D. *Data integration blueprint and modeling: techniques for a scalable and sustainable architecture*. [S.l.]: Pearson Education, 2010.

[Haas 2006]HAAS, L. Beauty and the beast: The theory and practice of information integration. In: SPRINGER. *Database Theory–ICDT 2007: 11th International Conference, Barcelona, Spain, January 10-12, 2007. Proceedings 11*. [S.l.], 2006. p. 28–43.

[Hai Do 2007]HAI DO, H. *Schema matching and mapping-based data integration: Architecture, approaches and evaluation*. [S.l.]: VDM Verlag, 2007.

[Hao et al. 2021]HAO, J. et al. Medto: Medical data to ontology matching using hybrid graph neural networks. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. [S.l.: s.n.], 2021. p. 2946–2954.

[Hättasch et al. 2022]HÄTTASCH, B. et al. It's ai match: A two-step approach for schema matching using embeddings. *arXiv preprint arXiv:2203.04366*, 2022.

[He et al. 2021]HE, Y. et al. Biomedical ontology alignment with bert. 2021.

[He et al. 2022]HE, Y. et al. Bertmap: a bert-based ontology alignment system. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2022. v. 36, n. 5, p. 5684–5691.

[Hertling, Portisch e Paulheim 2020]HERTLING, S.; PORTISCH, J.; PAULHEIM, H. Supervised ontology and instance matching with melt. *arXiv preprint arXiv:2009.11102*, 2020.

[Hohpe e Woolf 2004]HOHPE, G.; WOOLF, B. *Enterprise integration patterns: Designing, building, and deploying messaging solutions*. [S.l.]: Addison-Wesley Professional, 2004.

[Hulsebos et al. 2019]HULSEBOS, M. et al. Sherlock: A deep learning approach to semantic data type detection. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. [S.l.: s.n.], 2019. p. 1500–1508.

[Hyman 1982]HYMAN, R. Quasi-experimentation: Design and analysis issues for field settings (book). *Journal of Personality Assessment*, Taylor & Francis, v. 46, n. 1, p. 96–97, 1982.

[IBGE 2019]IBGE. Nov 2019. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/25885-11-8-dos-jovens-com-menores-rendimentos-abandonaram-a-escola-sem-concluir-a-educacao-basica-em-2018>.

[Iyer, Agarwal e Kumar 2020]IYER, V.; AGARWAL, A.; KUMAR, H. Multifaceted context representation using dual attention for ontology alignment. *arXiv preprint arXiv:2010.11721*, 2020.

[Iyer, Agarwal e Kumar 2020]IYER, V.; AGARWAL, A.; KUMAR, H. Veealign: a supervised deep learning approach to ontology alignment. In: *OM@ ISWC*. [S.l.: s.n.], 2020. p. 216–224.

[Iyer, Agarwal e Kumar 2021]IYER, V.; AGARWAL, A.; KUMAR, H. Veealign: multifaceted context representation using dual attention for ontology alignment. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2021. p. 10780–10792.

[Jiang e Xue 2020]JIANG, C.; XUE, X. Matching biomedical ontologies with long short-term memory networks. In: IEEE. *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*. [S.l.], 2020. p. 2484–2489.

[Jiménez-Ruiz et al. 2018]JIMÉNEZ-RUIZ, E. et al. Breaking-down the ontology alignment task with a lexical index and neural embeddings. *arXiv preprint arXiv:1805.12402*, 2018.

[Jiménez-Ruiz et al. 2018]JIMÉNEZ-RUIZ, E. et al. We divide, you conquer: from large-scale ontology alignment to manageable subtasks with a lexical index and neural embeddings. In: *CEUR Workshop Proceedings*. [S.l.: s.n.], 2018. v. 2288, p. 13–24.

[Jurafsky e Martin 2009]JURAFSKY, D.; MARTIN, J. H. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall, 2009. ISBN 9780131873216 0131873210. Disponível em: <http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd$_b$x$g$y$_{bi}$m$g_y$>.

[Jurisch e Igler 2019]JURISCH, M.; IGLER, B. Graph-convolution-based classification for ontology alignment change prediction. In: *DL4KG@ ESWC*. [S.l.: s.n.], 2019. p. 11–20.

[Kasunic 2005]KASUNIC, M. *Designing an effective survey*. [S.l.]: Citeseer, 2005.

[Khan e Gubanov 2020]KHAN, R.; GUBANOV, M. Weblens: Towards web-scale data integration, training the models. In: IEEE. *2020 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2020. p. 5727–5729.

[Koutras et al. 2020]KOUTRAS, C. et al. Rema: Graph embeddings-based relational schema matching. In: *EDBT/ICDT Workshops*. [S.l.: s.n.], 2020.

[Koutras et al. 2021]KOUTRAS, C. et al. Valentine: Evaluating matching techniques for dataset discovery. In: *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. [S.l.: s.n.], 2021. p. 468–479.

[Kudo et al. 2019]KUDO, T. et al. A revisited systematic literature mapping on the support of requirement patterns for the software development life cycle. *Journal of Software Engineering Research and Development*, v. 7, p. 9:1–9:11, 2019. ISSN 2195-1721. Disponível em: <https://sol.sbc.org.br/journals/index.php/jserd/article/view/458>.

[Kudo, Neto e Vincenzi 2020]KUDO, T. N.; NETO, R. F. Bulcão; VINCENZI, A. M. R. Requirement patterns: A tertiary study and a research agenda. *IET Software*, Institution of Engineering and Technology, v. 14, n. 1, p. 18–26, September 2020. ISSN 1751-8806. https://doi.org/10.1049/iet-sen.2019.0016.

[Laadhar et al. 2019]LAADHAR, A. et al. Partitioning and local matching learning of large biomedical ontologies. In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. [S.l.: s.n.], 2019. p. 2285–2292.

[Laadhar et al. 2019]LAADHAR, A. et al. Pomap++ results for oaei 2019: fully automated machine learning approach for ontology matching. In: *14th International Workshop on*

*Ontology Matching co-located with the International Semantic Web Conference (OM@ ISWC 2019)*. [S.l.: s.n.], 2019. p. 169–174.

[Le, Dieng-Kuntz e Gandon 2004]LE, B. T.; DIENG-KUNTZ, R.; GANDON, F. On ontology matching problems. *ICEIS (4)*, p. 236–243, 2004.

[Lebtag et al. 2022]LEBTAG, B. G. A. et al. Strategies to evolve exm notations extracted from a survey with software engineering professionals perspective. *Journal of Software Engineering Research and Development*, v. 10, p. 2–1, 2022.

[Lebtag et al. 2022]LEBTAG, B. G. A. et al. Strategies to evolve exm notations extracted from a survey with software engineering professionals perspective. *Journal of Software Engineering Research and Development*, v. 10, p. 2–1, 2022.

[Li 2020]LI, G. Deepfca: Matching biomedical ontologies using formal concept analysis embedding techniques. In: *proceedings of the 4th International Conference on Medical and Health Informatics*. [S.l.: s.n.], 2020. p. 259–265.

[Li 2020]LI, G. Improving biomedical ontology matching using domain-specific word embeddings. In: *Proceedings of the 4th International Conference on Computer Science and Application Engineering*. [S.l.: s.n.], 2020. p. 1–5.

[Li et al. 2019]LI, W. et al. Multi-view embedding for biomedical ontology matching. *OM@ ISWC*, v. 2536, p. 13–24, 2019.

[Li et al. 2021]LI, X. et al. Heterogeneous embeddings for relational data integration tasks. In: SPRINGER. *Web Information Systems and Applications: 18th International Conference, WISA 2021, Kaifeng, China, September 24–26, 2021, Proceedings 18*. [S.l.], 2021. p. 680–692.

[Li, Liu e Zhang 2005]LI, Y.; LIU, D.-B.; ZHANG, W.-M. Schema matching using neural network. In: IEEE. *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. [S.l.], 2005. p. 743–746.

[Liberati et al. 2009]LIBERATI, A. et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med.*, v. 6, n. 7, p. e1000–100, 2009.

[Lima et al. 2020]LIMA, B. et al. Learning reference alignments for ontology matching within and across domains. In: *OM@ ISWC*. [S.l.: s.n.], 2020. p. 72–76.

[Linåker et al. 2015]LINÅKER, J. et al. Guidelines for conducting surveys in software engineering. Department of Computer Science, Lund University, 2015.

[LINS e Arbix 2011]LINS, L. M.; ARBIX, G. Educação, qualificação, produtividade e crescimento econômico: a harmonia colocada em questão. *IPEA: Anais do I Circulo de Debates Acadêmicos*, 2011.

[Lv 2022]LV, Z. An effective approach for large ontology matching using multi-objective grasshopper algorithm. In: *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*. [S.l.: s.n.], 2022. p. 110–116.

[Maji, Rout e Choudhary 2021]MAJI, S.; ROUT, S. S.; CHOUDHARY, S. Dcom: A deep column mapper for semantic data type detection. *arXiv preprint arXiv:2106.12871*, 2021.

[Melnik, Garcia-Molina e Rahm 2002]MELNIK, S.; GARCIA-MOLINA, H.; RAHM, E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In: IEEE. *Proceedings 18th international conference on data engineering*. [S.l.], 2002. p. 117–128.

[Mohamed et al. 2022]MOHAMED, A. et al. Schema matching based on deep learning using lstm model. In: IEEE. *2022 IEEE 3rd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*. [S.l.], 2022. p. 1–5.

[Molléri, Petersen e Mendes 2016]MOLLÉRI, J. S.; PETERSEN, K.; MENDES, E. Survey guidelines in software engineering: An annotated review. In: *Proceedings of the 10th ACM/IEEE international symposium on empirical software engineering and measurement*. [S.l.: s.n.], 2016. p. 1–6.

[Mukherjee et al. 2021]MUKHERJEE, D. et al. Learning knowledge graph for target-driven schema matching. In: *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*. [S.l.: s.n.], 2021. p. 65–73.

[MUYLAERT 2020]MUYLAERT, R. *Pandemia do novo coronavírus, Parte 6: inteligência artificial (NLP) — marcoarmello.wordpress.com*. 2020. https://marcoarmello.wordpress.com/2020/08/19/coronavirus6/. [Accessed 09-Jul-2023].

[Neutel e Boer 2021]NEUTEL, S.; BOER, M. H. de. Towards automatic ontology alignment using bert. In: *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. [S.l.: s.n.], 2021.

[Nezhadi, Shadgar e Osareh 2011]NEZHADI, A. H.; SHADGAR, B.; OSAREH, A. Ontology alignment using machine learning techniques. *International Journal of Computer Science & Information Technology*, Academy & Industry Research Collaboration Center(AIRCC), v. 3, n. 2, p. 139, 2011.

[Nikovski et al. 2012]NIKOVSKI, D. et al. Bayesian networks for matcher composition in automatic schema matching. In: SCITEPRESS. *International Conference on Enterprise Information Systems*. [S.l.], 2012. v. 2, p. 48–55.

[Nkisi-Orji et al. 2019]NKISI-ORJI, I. et al. Ontology alignment based on word embedding and random forest classification. In: SPRINGER. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. [S.l.], 2019. p. 557–572.

[Nozaki, Hochin e Nomiya 2019]NOZAKI, K.; HOCHIN, T.; NOMIYA, H. Semantic schema matching for string attribute with word vectors. In: IEEE. *2019 6th International Conference on Computational Science/Intelligence and Applied Informatics (CSII)*. [S.l.], 2019. p. 25–30.

[OECD 2021]OECD. *Education at a Glance 2021*. [s.n.], 2021. 474 p. Disponível em: <https://www.oecd-ilibrary.org/content/publication/b35a14e5-en>.

[Pan, Pan e Monti 2022]PAN, Z.; PAN, G.; MONTI, A. Semantic-similarity-based schema matching for management of building energy data. *Energies*, MDPI, v. 15, n. 23, p. 8894, 2022.

[Petersen, Vakkalanka e Kuzniarz 2015]PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, Elsevier, v. 64, p. 1–18, 2015.

[Peukert, Massmann e Koenig 2010]PEUKERT, E.; MASSMANN, S.; KOENIG, K. Comparing similarity combination methods for schema matching. *INFORMATIK 2010. Service Science–Neue Perspektiven für die Informatik. Band 1*, Gesellschaft für Informatik eV, 2010.

[Przyborowski et al. 2021]PRZYBOROWSKI, M. et al. Schema matching using gaussian mixture models with wasserstein distance. *arXiv preprint arXiv:2111.14244*, 2021.

[Rahm e Bernstein 2001]RAHM, E.; BERNSTEIN, P. *On Matching Schemas Automatically*. [S.l.], February 2001. 22 p. Disponível em: <https://www.microsoft.com/en-us/research/publication/on-matching-schemas-automatically/>.

[Rahm e Bernstein 2001]RAHM, E.; BERNSTEIN, P. A survey of approaches to automatic schema matching. *VLDB J.*, v. 10, p. 334–350, 12 2001.

[Rangel et al. 2015]RANGEL, C. et al. An approach for the emerging ontology alignment based on the bees colonies. In: THE STEERING COMMITTEE OF THE WORLD CONGRESS IN COMPUTER SCIENCE, COMPUTER .... *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. [S.l.], 2015. p. 536.

[Refoufi e Benarab 2018]REFOUFI, A.; BENARAB, A. A robust approach to the ontology matching problem. In: *Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence*. [S.l.: s.n.], 2018. p. 76–83.

[Rodrigues e Silva 2021]RODRIGUES, D.; SILVA, A. A study on machine learning techniques for the schema matching network problem. *Journal of the Brazilian Computer Society*, v. 27, 12 2021.

[Rodrigues et al. 2015]RODRIGUES, D. et al. Using active learning techniques for improving database schema matching methods. In: IEEE. *2015 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2015. p. 1–8.

[Sahay, Mehta e Jadon 2020]SAHAY, T.; MEHTA, A.; JADON, S. Schema matching using machine learning. In: IEEE. *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. [S.l.], 2020. p. 359–366.

[Scannavino et al. 2017]SCANNAVINO, K. R. F. et al. *Revisão sistemática da literatura em Engenharia de Software: teoria e prática*. first. [S.l.]: Elsevier, 2017.

[Schmidts et al. 2019]SCHMIDTS, O. et al. Schema matching with frequent changes on semi-structured input files: A machine learning approach on biological product data. In: *ICEIS (1)*. [S.l.: s.n.], 2019. p. 208–215.

[Shraga 2022]SHRAGA, R. Humanal: Calibrating human matching beyond a single task. In: *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. [S.l.: s.n.], 2022. p. 1–8.

[Shraga e Gal 2022]SHRAGA, R.; GAL, A. Powarematch: A quality-aware deep learning approach to improve human schema matching. *ACM Journal of Data and Information Quality (JDIQ)*, ACM New York, NY, v. 14, n. 3, p. 1–27, 2022.

[Shraga, Gal e Roitman 2020]SHRAGA, R.; GAL, A.; ROITMAN, H. Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 13, n. 9, p. 1401–1415, 2020.

[Smith et al. 2013]SMITH, E. et al. Improving developer participation rates in surveys. In: IEEE. *2013 6th International workshop on cooperative and human aspects of software engineering (CHASE)*. [S.l.], 2013. p. 89–92.

[Sreemathy et al. 2021]SREEMATHY, J. et al. Data integration and etl: a theoretical perspective. In: IEEE. *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. [S.l.], 2021. v. 1, p. 1655–1660.

[Srinivas, Gale e Dolby 2018]SRINIVAS, K.; GALE, A.; DOLBY, J. Merging datasets through deep learning. *arXiv preprint arXiv:1809.01604*, 2018.

[Sun e Shen 2022]SUN, C.; SHEN, D. Towards deep entity resolution via soft schema matching. *Neurocomputing*, Elsevier, v. 471, p. 107–117, 2022.

[Sun, Takeuchi e Yamasaki 2020]SUN, J.; TAKEUCHI, S.; YAMASAKI, I. Few-shot ontology alignment model with attribute attentions. In: IEEE. *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*. [S.l.], 2020. p. 1218–1222.

[Tatbul 2010]TATBUL, N. Streaming data integration: Challenges and opportunities. In: IEEE. *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*. [S.l.], 2010. p. 155–158.

[Teslya e Savosin 2019]TESLYA, N.; SAVOSIN, S. Matching ontologies with word2vec-based neural network. In: SPRINGER. *Computational Science and Its Applications–ICCSA 2019: 19th International Conference, Saint Petersburg, Russia, July 1–4, 2019, Proceedings, Part I 19*. [S.l.], 2019. p. 745–756.

[Uschold e Grüninger 2004]USCHOLD, M.; GRÜNINGER, M. Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, v. 33, n. 4, p. 58–64, 2004. Disponível em: <https://doi.org/10.1145/1041410.1041420>.

[Wang et al. 2022]WANG, P. et al. Matching biomedical ontologies with gcn-based feature propagation. *Math. Biosci. Eng*, v. 19, p. 8479–8504, 2022.

[Wohlin et al. 2012]WOHLIN, C. et al. *Experimentation in software engineering*. [S.l.]: Springer Science & Business Media, 2012.

[Xue, Chen e Liu 2021]XUE, X.; CHEN, D.; LIU, W. Naive bayesian classifier based semi-supervised learning for matching ontologies. In: IEEE. *2021 17th International Conference on Computational Intelligence and Security (CIS)*. [S.l.], 2021. p. 162–167.

[Xue, Chen e Ren 2019]XUE, X.; CHEN, J.; REN, A. Interactive ontology matching based on evolutionary algorithm. In: IEEE. *2019 15th International Conference on Computational Intelligence and Security (CIS)*. [S.l.], 2019. p. 1–5.

[Xue et al. 2021]XUE, X. et al. Artificial neural network based sensor ontology matching technique. In: *Companion Proceedings of the Web Conference 2021*. [S.l.: s.n.], 2021. p. 44–51.

[Xue et al. 2021]XUE, X. et al. Matching sensor ontologies through siamese neural networks without using reference alignment. *PeerJ Computer Science*, PeerJ Inc., v. 7, p. e602, 2021.

[Yorsh et al. 2022]YORSH, U. et al. Text-to-ontology mapping via natural language processing models. 2022.

[Yusof, Man e Ismail 2022]YUSOF, M. K.; MAN, M.; ISMAIL, A. Design and implement of rest api for data integration. In: *2022 International Conference on Engineering and Emerging Technologies (ICEET)*. [S.l.: s.n.], 2022. p. 1–4. ISSN 2831-3682.

[Zhang et al. 2021]ZHANG, J. et al. Smat: An attention-based deep learning solution to the automation of schema matching. In: SPRINGER. *Advances in Databases and Information Systems: 25th European Conference, ADBIS 2021, Tartu, Estonia, August 24–26, 2021, Proceedings 25*. [S.l.], 2021. p. 260–274.

[Zhang et al. 2023]ZHANG, Y. et al. Schema matching using pre-trained language models. In: IEEE. *ICDE*. 2023. Disponível em: <https://www.microsoft.com/en-us/research/publication/schema-matching-using-pre-trained-language-models/>.

[Zhang et al. 2014]ZHANG, Y. et al. Ontology matching with word embeddings. In: SPRINGER. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 13th China National Conference, CCL 2014, and Second International Symposium, NLP-NABD 2014, Wuhan, China, October 18-19, 2014. Proceedings*. [S.l.], 2014. p. 34–45.

APÊNDICE A

# Tabela de estudos selecionados no mapeamento sistemático

Tabela A.1: Artigos selecionados

| ID | Título |
|---|---|
| A1 | Matching Biomedical Ontologies with Long Short-Term Memory Networks |
| A2 | LEAPME: Learning-based Property Matching with Embeddings |
| A3 | VeeAlign: A supervised deep learning approach to ontology alignment |
| A4 | Dcom: A deep column mapper for semantic data type detection |
| A5 | Biomedical Ontology Alignment with BERT |
| A6 | Towards deep entity resolution via soft schema matching |
| A7 | Applying of machine learning techniques to combine string-based, language-based and structure-based similarity measures for ontology matching |
| A8 | POMAP++ results for OAEI 2019: Fully automated machine learning approach for ontology matching |
| A9 | Supervised ontology and instance matching with MELT |
| A260 | Learning reference alignments for ontology matching within and across domains |
| A261 | Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem |
| A267 | Matching Ontologies with Word2Vec-Based Neural Network |
| A282 | Text-to-Ontology Mapping via Natural Language Processing Models |
| A313 | Ontology matching using convolutional neural networks |
| A314 | VeeAlign: Multifaceted context representation using dual attention for ontology alignment |
| A319 | DeepFCA: Matching Biomedical Ontologies Using Formal Concept Analysis Embedding Techniques |
| A323 | An Effective Approach for Large Ontology Matching Using Multi-objective Grasshopper Algorithm |
| A324 | SCHEMA MATCHING USING GAUSSIAN MIXTURE MODELS WITH WASSERSTEIN DISTANCE |
| A329 | Fast and Accurate Framework for Ontology Matching in Web of Things |

| | |
|---|---|
| A330 | Artificial Neural Network Based Sensor Ontology Matching Technique |
| A348 | Graph-convolution-based classification for ontology alignment change prediction |
| A355 | Towards automatic ontology alignment using BERT |
| A362 | Multi-view embedding for biomedical ontology matching |
| A363 | An approach for the emerging ontology alignment based on the bees colonies |
| A364 | Ontology alignment based on word embedding and random forest classification |
| A366 | An Ontology Alignment Approach Combining Word Embedding and the Radius Measure |
| A369 | Partitioning and local matching learning of large biomedical ontologies |
| A370 | Merging datasets through deep learning |
| A388 | Schema Matching Based On Deep Learning Using LSTM Model |
| A390 | Naive Bayesian Classifier Based Semi-supervised Learning for Matching Ontologies |
| A392 | Semantic Schema Matching for String Attribute with Word Vectors |
| A393 | Interactive Ontology Matching Based on Evolutionary Algorithm |
| A394 | Sherlock: A deep learning approach to semantic data type detection |
| A398 | Semantic-Similarity-Based Schema Matching for Management of Building Energy Data |
| A399 | Schema matching with frequent changes on semi-structured input files: A machine learning approach on biological product data |
| A400 | Schema matching using machine learning |
| A423 | Matching biomedical ontologies with GCN-based feature propagation |
| A428 | We divide, you conquer: From large-scale ontology alignment to manageable subtasks with a lexical index and neural embeddings |
| A429 | Matching sensor ontologies with unsupervised neural network with competitive learning |
| A433 | Breaking-down the ontology alignment task with a lexical index and neural embeddings |
| A435 | A learning style-based Ontology Matching to enhance learning resources recommendation |
| A438 | Few-Shot Ontology Alignment Model with Attribute Attentions |
| A331 | Improving Biomedical Ontology Matching Using Domain-Specific Word Embeddings |
| A442 | Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision |
| A443 | Multifaceted context representation using dual attention for ontology alignment |
| A445 | BERTMap: A BERT-Based Ontology Alignment System |
| A447 | A study on machine learning techniques for the schema matching network problem |
| A477 | Matching sensor ontologies through siamese neural networks without using reference alignment |
| A505 | WebLens: Towards Web-scale Data Integration, Training the Models |
| A576 | Magellan: Toward Building Ecosystems of Entity Matching Solutions |
| A395 | Learning Knowledge Graph for Target-driven Schema Matching |
| A584 | HumanAL: Calibrating Human Matching beyond a Single Task |

| | |
|---|---|
| A586 | Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks |
| A593 | ADnEV: Cross-Domain Schema Matching Using Deep Similarity Matrix Adjustment and Evaluation |
| A597 | OntoConnect: Unsupervised Ontology Alignment with Recursive Neural Network |
| A602 | MEDTO: Medical Data to Ontology Matching Using Hybrid Graph Neural Networks |
| A606 | SMAT: An Attention-Based Deep Learning Solution to the Automation of Schema Matching |
| A608 | PoWareMatch: A Quality-aware Deep Learning Approach to Improve Human Schema Matching |
| A630 | Heterogeneous Embeddings for Relational Data Integration Tasks |

Tabela A.2: Artigos selecionados pelo SnowBalling

| **ID** | **Título** |
|---|---|
| SN1 | REMA: Graph Embeddings-based Relational Schema Matching. |
| SN2 | Decision trees in automatic ontology matching |
| SN3 | It's AI Match: A Two-Step Approach for Schema Matching Using Embeddings |
| SN4 | Ontology matching with word embeddings |
| SN5 | Bayesian Networks for Matcher Composition in Automatic Schema Matching |
| SN6 | Schema matching using neural network |
| SN7 | Using active learning techniques for improving database schema matching methods |
| SN8 | Database Schema Matching Using Machine Learning with Feature Selection |
| SN9 | Ontology Alignment Using Machine Learning Techniques |

APÊNDICE B

# Imagens do questionário aplicado

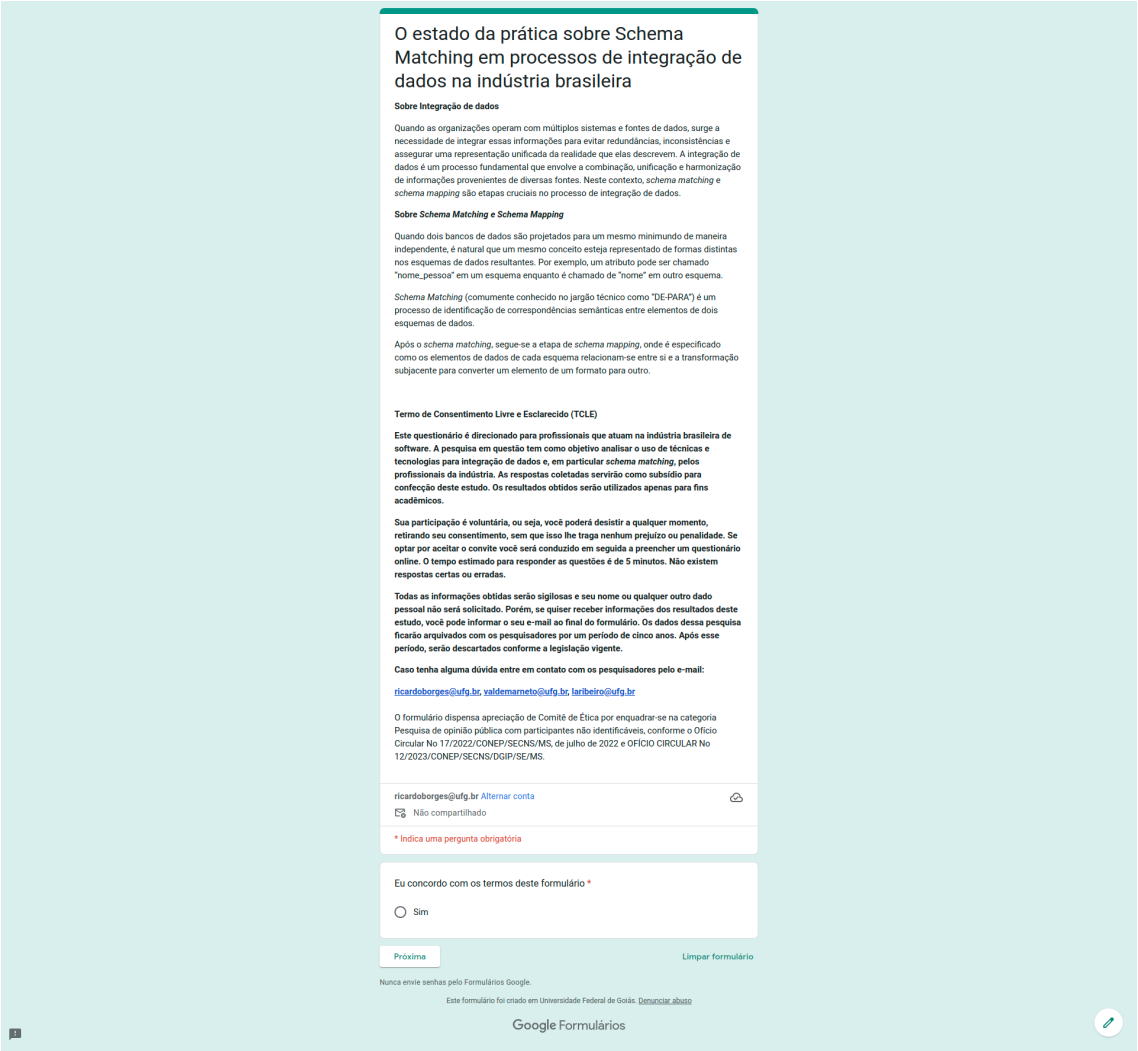


Figura B.1: Etapa 1 do questionário no Google Forms

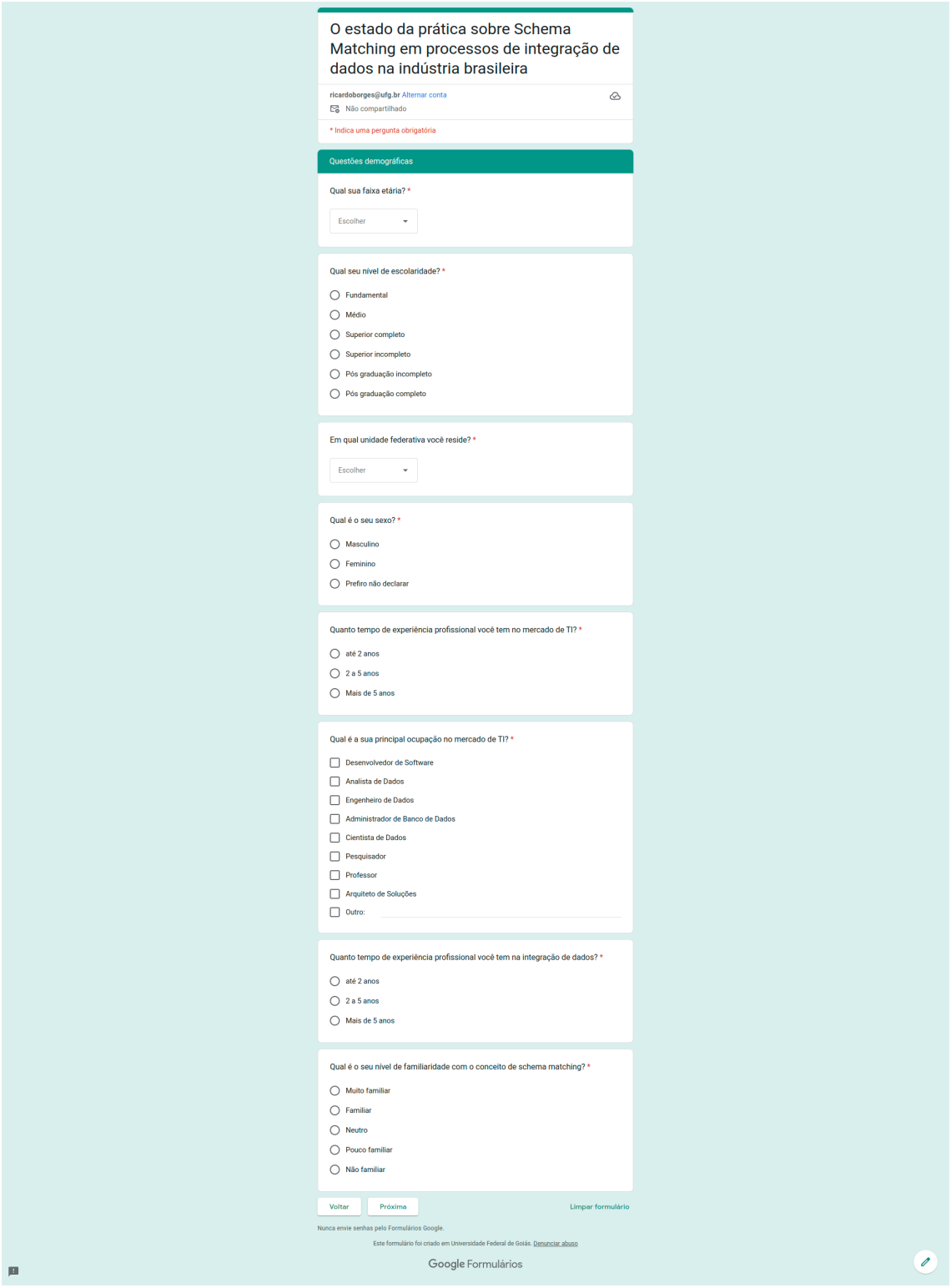


Figura B.2: Etapa 2 do questionário no Google Forms

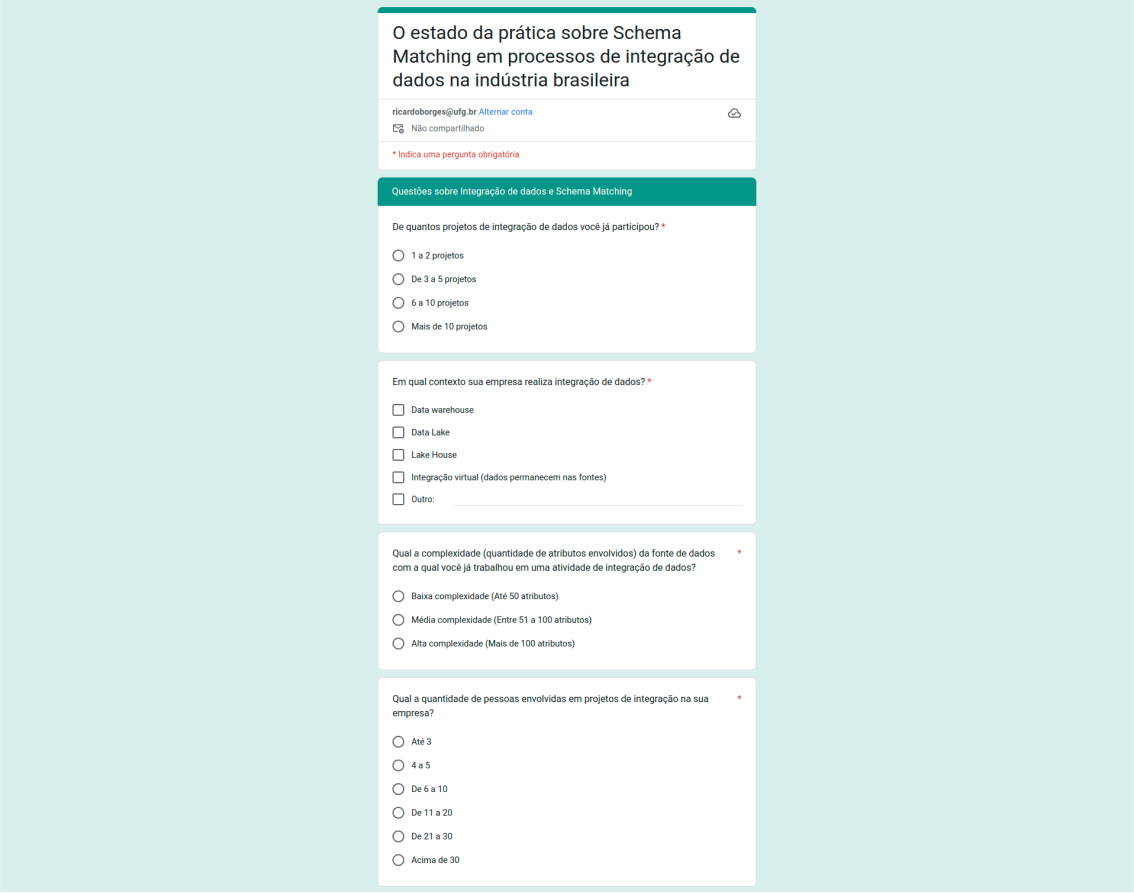


Figura B.3: Etapa 3, parte 1 do questionário no Google Forms

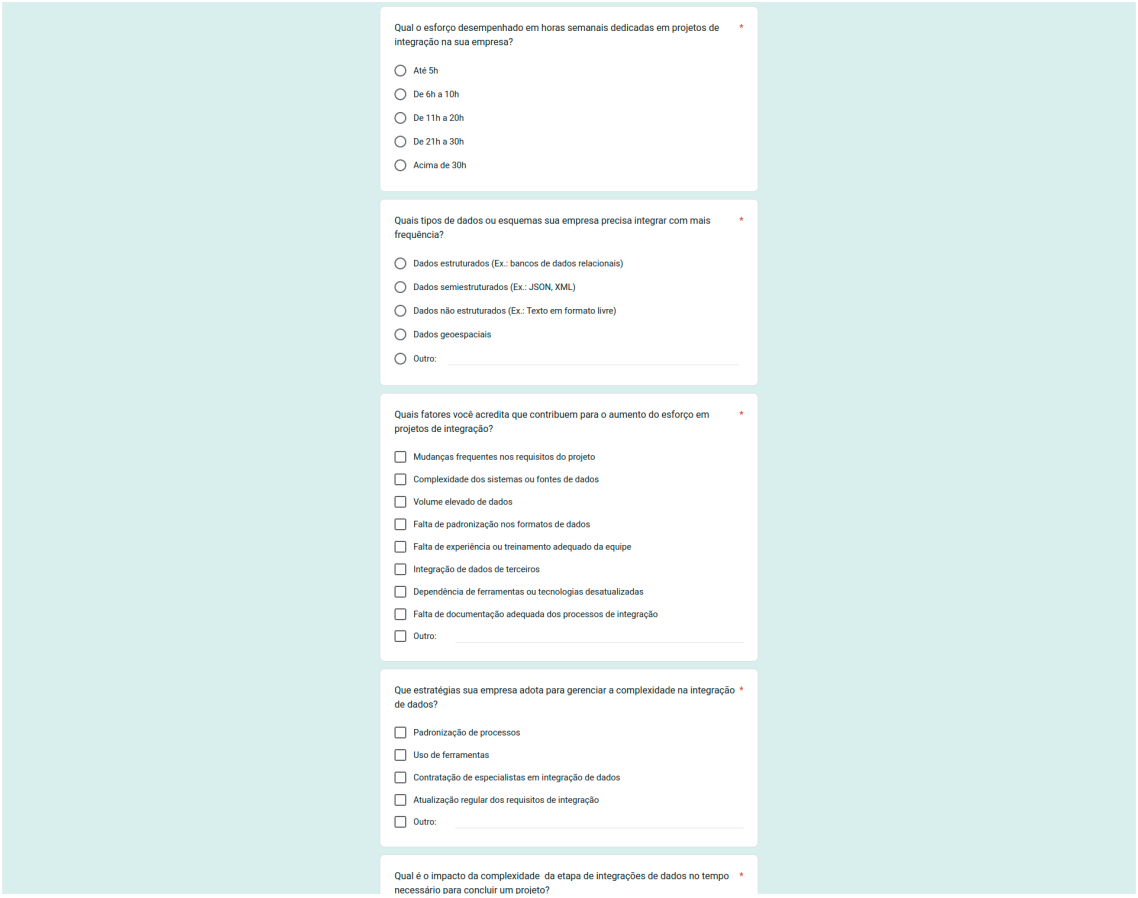


Figura B.4: Etapa 3, parte 2 do questionário no Google Forms

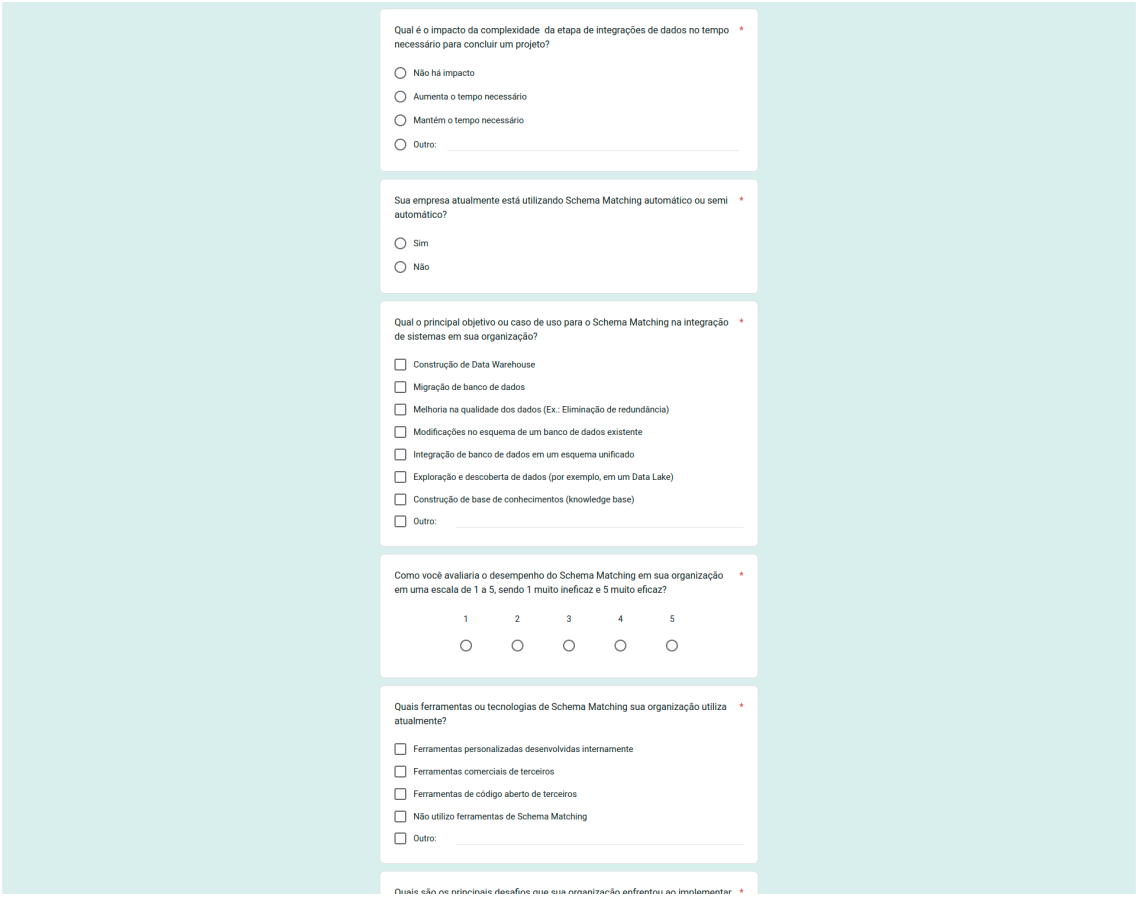


Figura B.5: Etapa 3, parte 3 do questionário no Google Forms

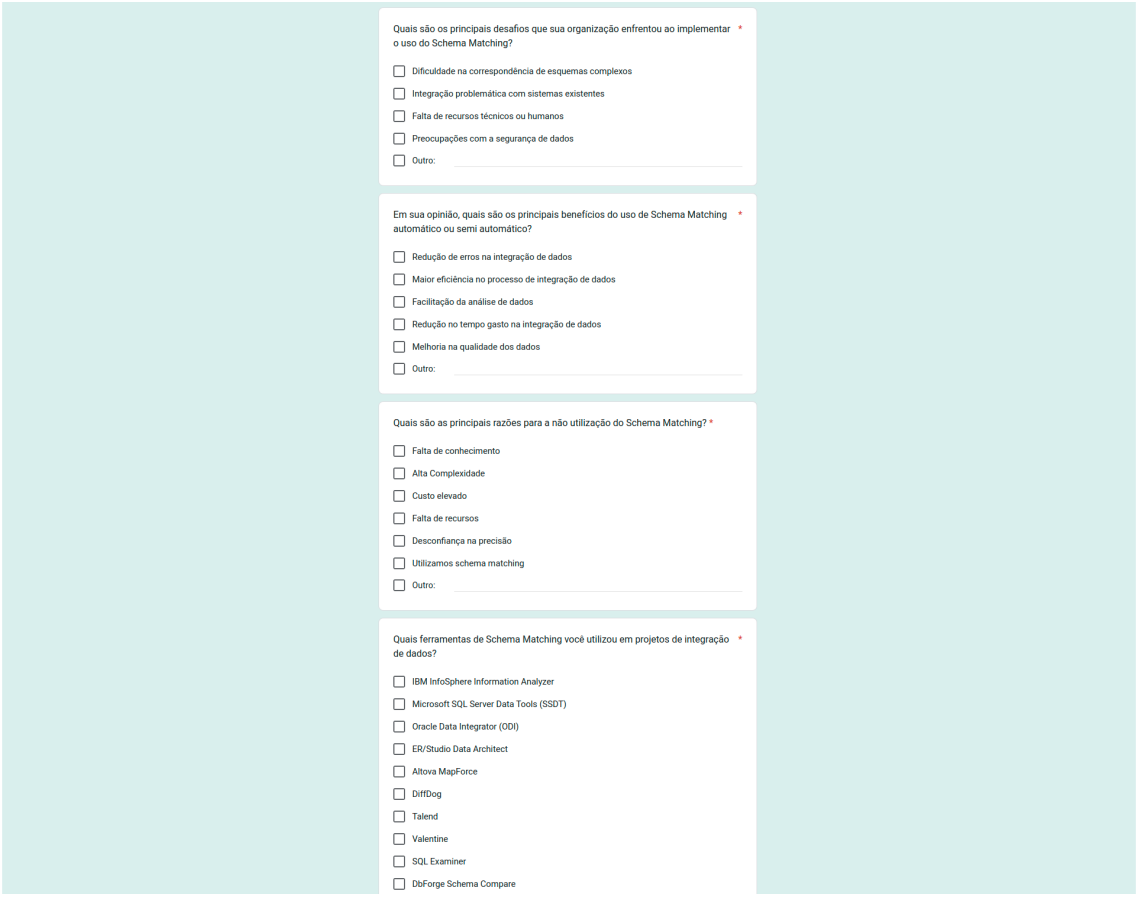


Figura B.6: Etapa 3, parte 4 do questionário no Google Forms

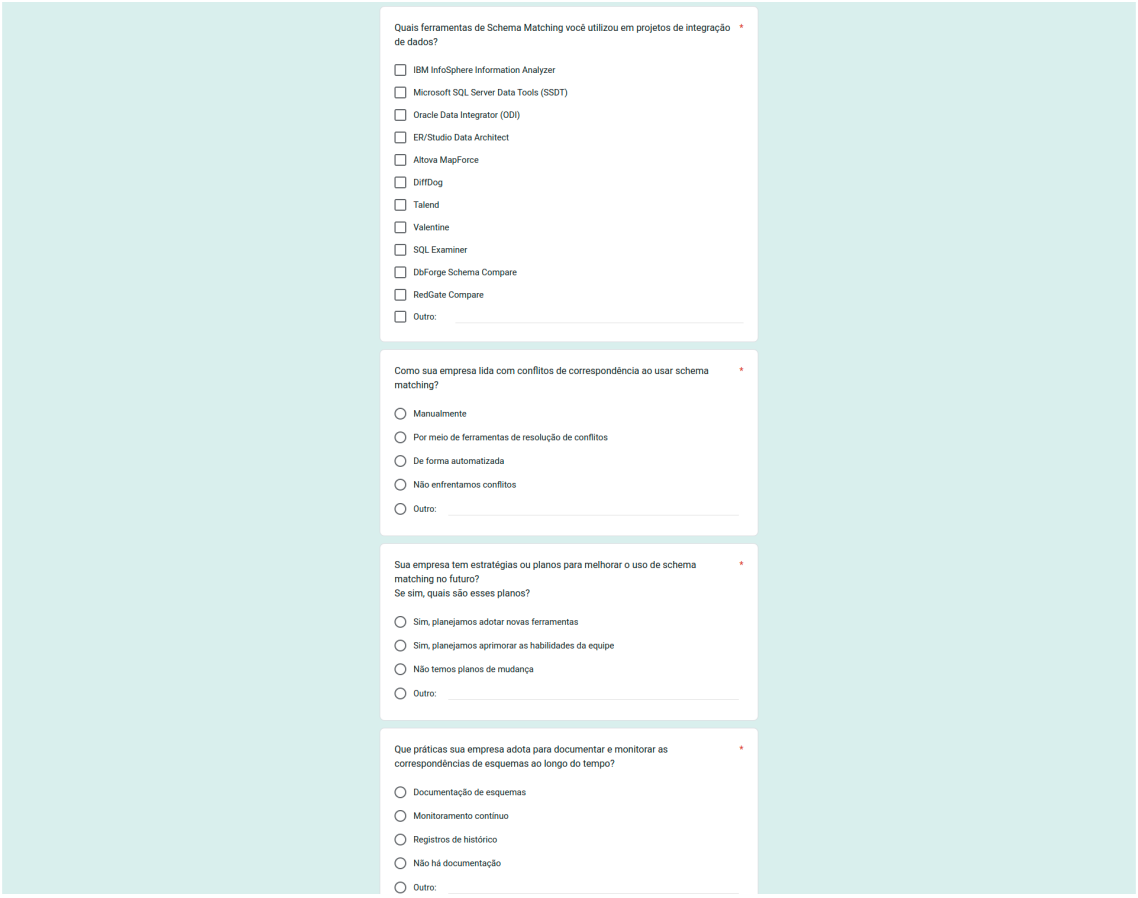


Figura B.7: Etapa 3, parte 5 do questionário no Google Forms

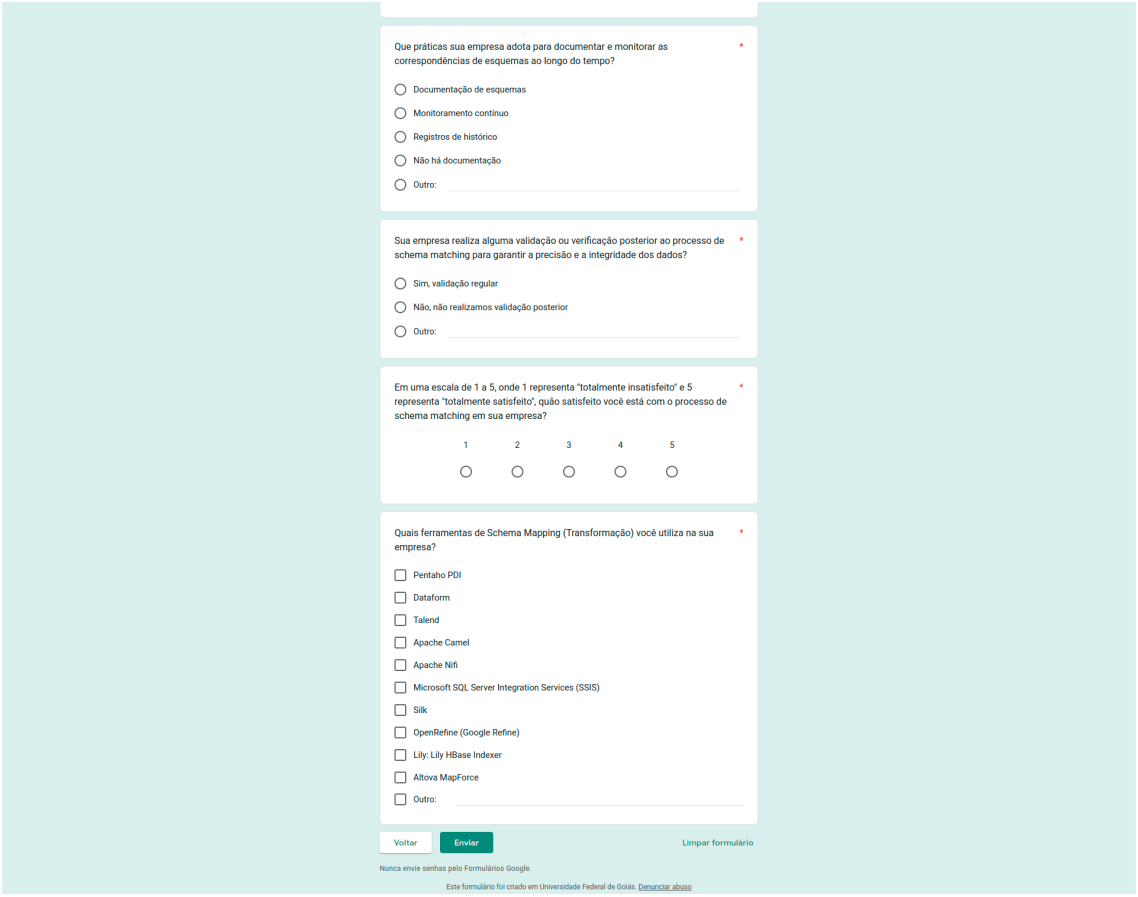


Figura B.8: Etapa 3, parte 6 do questionário no Google Forms