BRAZIL DATA COMMONS: Plataforma Unificada para Análise de Dados Públicos de Diferentes Repositórios

Isadora C. Rodrigues[1], **Julio C. S. Reis**[2], **Bernardo L. Queiroz**[3], **Fabrício Benevenuto**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

[2]Departamento de Informática – Universidade Federal de Viçosa (UFV)

[3]Departamento de Demografia – Universidade Federal de Minas Gerais (UFMG)

{isadorarodrigues,fabricio}@dcc.ufmg.br, jreis@ufv.br, blanza@ufmg.br

Abstract. *Addressing urgent social and environmental challenges requires solid research grounded in reliable data. In Brazil, despite the abundance of public data available online, fragmentation, inconsistent standards, and limited interoperability hinder effective research and policymaking. To overcome these challenges, we present* BRAZIL DATA COMMONS *— a platform that unifies Brazilian datasets under a common semantic framework. By leveraging global ontologies and interoperable standards,* BRAZIL DATA COMMONS *enables the discovery, integration, and visualization of data across domains, situating Brazilian data in a global context. With intuitive interfaces and flexible data access, the platform democratizes data use and supports informed decision-making.*

Resumo. *Enfrentar desafios sociais e ambientais urgentes exige pesquisas sólidas baseadas em dados confiáveis. No Brasil, apesar da abundância de dados públicos disponíveis online, sua fragmentação, padrões inconsistentes e baixa interoperabilidade dificultam a pesquisa e a formulação de políticas eficazes. Para superar esses desafios, apresentamos o* BRAZIL DATA COMMONS *— uma plataforma que unifica conjuntos de dados brasileiros sob um arcabouço semântico comum. Utilizando ontologias globais e padrões interoperáveis, o* BRAZIL DATA COMMONS *facilita a descoberta, integração e visualização de dados de diferentes domínios, colocando o Brasil em um contexto internacional. Com interfaces intuitivas e acesso flexível aos dados, a plataforma democratiza o uso de dados e promove decisões informadas.*

1. Introdução

Abordar desafios sociais, econômicos e ambientais urgentes no mundo moderno requer pesquisas sólidas que, por sua vez, dependem crucialmente do acesso a dados confiáveis e de alta qualidade [Herrera and Kapur 2007]. Embora uma abundância de dados públicos esteja disponível online, particularmente no Brasil, o grande volume e a fragmentação dessas informações frequentemente dificultam sua descoberta, integração e uso eficaz [Shikida et al. 2021, Passos 2022].

Embora diversas iniciativas tenham buscado mitigar os desafios da unificação e organização de dados públicos [Dang et al. 2023], elas diferem fundamentalmente da proposta apresentada neste estudo. Plataformas governamentais frequentemente operam como catálogos de dados [Governo Federal do Brasil 2025], sem reduzir significativamente as tarefas demoradas de preparação e padronização [Press 2016], o que impõe barreiras para usuários sem expertise técnica antes de viabilizar pesquisas sofisticadas [da Cruz Martins et al. 2013]. Além disso, os portais e repositórios de dados públicos atualmente disponíveis tendem a ser estruturados como bancos de dados monolíticos [Freitas et al. 2023, Base dos Dados 2025], carecendo de padrões consistentes, oferecendo metadados limitados e concentrando-se predominantemente em um único domínio, o que dificulta a integração eficiente de dados entre diferentes áreas. Iniciativas internacionais [Eurostat 2025, World Bank 2025], por sua vez, costumam disponibilizar dados apenas em nível nacional, com alto grau de agregação, o que restringe análises detalhadas em contextos regionais e setoriais. Vale ressaltar, por fim, que fontes de dados abertos não costumam promover, de forma aberta e consistente, um formato unificado de dados e, ao dependerem de identificadores específicos, tornam ainda mais complexos os esforços para consolidar fontes de dados dispersas.

Para lidar com essas limitações, apresentamos o BRAZIL DATA COMMONS, disponível em https://brazildatacommons.com.br, uma plataforma que organiza uma ampla gama de conjuntos de dados públicos brasileiros—abrangendo desde indicadores socioeconômicos até estatísticas ambientais—e os integra em um sistema unificado, agregando um valioso nível de interoperabilidade aos dados públicos. Ao coletar, processar e disponibilizar dados por meio de esquemas padronizados e interfaces de programação de aplicações (APIs), qualquer pessoa com conexão à Internet pode interagir facilmente com informações de alta qualidade sobre o Brasil por meio de interfaces gráficas intuitivas ou até mesmo ter acesso a conjuntos de dados inteiros com habilidades mínimas de programação.

Embora o BRAZIL DATA COMMONS esteja firmemente ancorado no contexto brasileiro, ele também se baseia nos princípios fundamentais do ecossistema global do DATA COMMONS [Guha et al. 2023]. Ao adotar uma estrutura semântica consistente e utilizar ontologias reconhecidas globalmente—endossadas por importantes ferramentas de indexação, como Google[1], Microsoft[2], Pinterest[3], Yandex[4], e adotadas por mais de 45 milhões de *websites* [Schema.org 2025]—o BRAZIL DATA COMMONS transforma dados públicos previamente fragmentados em um recurso coeso e interoperável, o que favorece sua gestão e análise. Além disso, ao integrar-se ao ecossistema global de DATA COMMONS, o BRAZIL DATA COMMONS se beneficia da rede colaborativa de APIs como um todo e permite que os usuários contextualizem os dados brasileiros em um panorama internacional, ajudando a reduzir a lacuna de pesquisa que existe entre os países do Sul Global devido ao acesso limitado à informação [Blicharska et al. 2017].

Por meio de interfaces amigáveis, mecanismos de consulta simplificados e visualizações prontas para uso, a plataforma democratiza ainda mais o acesso e a análise de dados, capacitando pesquisadores, formuladores de políticas e entusiastas a extrair *insights* significativos e tomar decisões embasadas em dados. Neste artigo, discutimos como o BRAZIL DATA COMMONS tem potencial para transformar informações dispersas em um recurso coeso e acessível, essencial para compreender desafios atuais enfrentados pelo Brasil.

[1]https://www.google.com/
[2]https://www.microsoft.com/pt-br/
[3]https://br.pinterest.com
[4]https://yandex.com

O restante deste artigo está organizado da seguinte forma. Na próxima seção, apresentamos a arquitetura do BRAZIL DATA COMMONS. Em seguida, descrevemos exemplos de análises que podem ser realizadas com a plataforma proposta, incluindo considerações éticas e limitações dos dados. Por fim, apresentamos uma demonstração do BRAZIL DATA COMMONS, concluímos o estudo e discutimos direções para trabalhos futuros.

2. Brazil Data Commons

Em vez de funcionar como um grande repositório centralizado, o DATA COMMONS se manifesta como uma rede distribuída e interoperável de APIs [Guha et al. 2023]. No centro de sua arquitetura está um conjunto de esquemas bem definidos – um superconjunto do Schema.org [Guha et al. 2016] – que descrevem entidades, propriedades e relações entre os dados. Esses esquemas atuam como um "vocabulário compartilhado", garantindo que dados de diferentes fontes possam ser combinados e referenciados de forma coerente por qualquer API da rede, mesmo que tenham sido construídos de maneiras completamente distintas. Essa integração resulta em um grafo de conhecimento: um conjunto de nós e arestas que representam entidades (como países, estados, cidades) e suas propriedades (população, área, índice de desenvolvimento humano - IDH).

Vale destacar que o DATA COMMONS se baseia no princípio de *Referência por Descrição* [Guha and Gupta 2015], o que significa que, em vez de identificadores arbitrários, são utilizadas descrições semânticas claras e padronizadas. Por exemplo, em vez de um simples "ID 1234" para um local, a referência é feita por meio de uma descrição que esclarece qual entidade está sendo referida (por exemplo, "A cidade de Manaus no Brasil"). Essa abordagem semântica facilita a compreensão humana e elimina a necessidade de que diferentes fontes de dados compartilhem os mesmos identificadores.

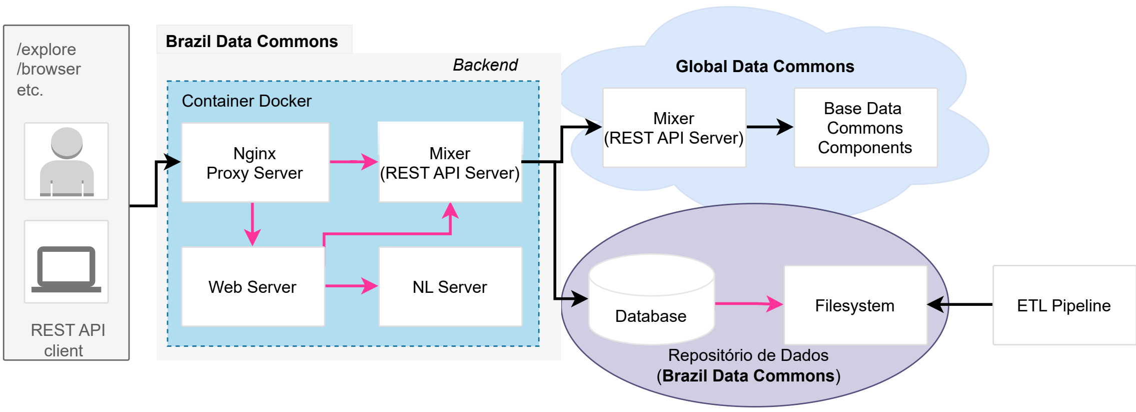


Figura 1. Arquitetura do BRAZIL DATA COMMONS.

O BRAZIL DATA COMMONS possui uma arquitetura modular que integra componentes personalizados à infraestrutura básica da rede. O sistema é composto por elementos-chave, conforme ilustrado na Figura 1: o NGINX[5], que roteia requisições externas para os servidores internos; o Web Server, responsável pela interface do usuário; o

[5]https://nginx.org/

NL Server, que possibilita consultas em linguagem natural; e o componente de integração ("Mixer"), que vai além da simples integração de dados, operando como um servidor REST API que conecta o repositório de dados aos usuários finais. Por meio dessa API, a interface Web do BRAZIL DATA COMMONS acessa e exibe dados integrados, ao mesmo tempo em que fornece aos clientes externos uma interface padronizada para acesso programático direto ao sistema.

Diferentemente da implementação padrão do DATA COMMONS, que depende fortemente de serviços em nuvem, o repositório de dados, que contém um sistema de arquivos e um banco de dados, foi especificamente adaptado para operar em ambientes computacionais locais. Essa adaptação permite que instituições implantem e mantenham suas próprias instâncias do sistema sem dependências de infraestrutura em nuvem. Alimentado por um pipeline de integração de dados - ETL, a principal função do repositório de dados é armazenar arquivos-fonte obtidos de portais e bancos de dados governamentais. Esses arquivos são processados para aderirem estritamente à padronização proposta pelo DATA COMMONS, garantindo a compatibilidade entre os dados e as ferramentas disponíveis na rede — como visualizações gráficas geradas dinamicamente. O banco de dados, construído a partir das informações do repositório, é otimizado para atualizações incrementais, reduzindo custos computacionais e operacionais a cada ciclo de atualização. Isso melhora a usabilidade do pipeline e incentiva contribuições de múltiplas fontes.

Essa arquitetura permite que o BRAZIL DATA COMMONS mantenha sua autonomia na gestão dos dados nacionais, ao mesmo tempo em que se beneficia da infraestrutura da rede como um todo, incentivando o desenvolvimento de estudos nacionais e tornando os dados brasileiros acessíveis para pesquisas em nível global[6].

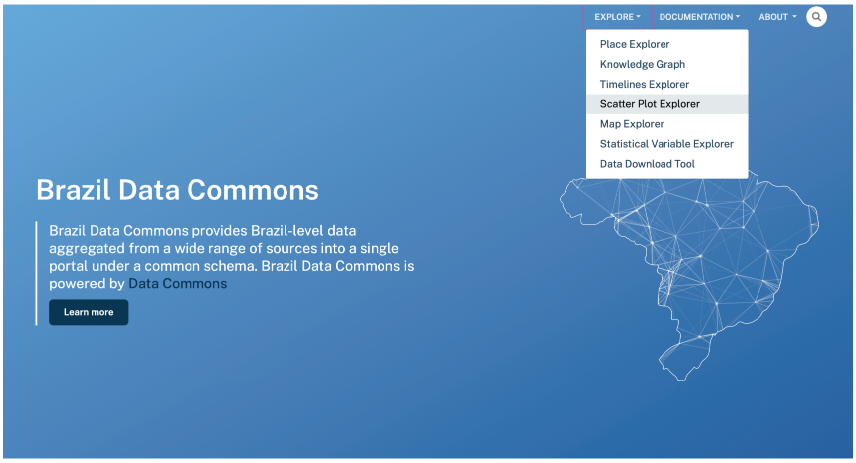
3. Demonstração da Plataforma

Nas Figuras 2(a), 2(b) e 2(c), apresentamos capturas de tela da interface Web do BRAZIL DATA COMMONS. A Figura 2(a) mostra a página inicial da plataforma, onde os usuários podem, entre outras opções, acessar as principais ferramentas clicando no menu suspenso "Explore", como o explorador de locais e as ferramentas de visualização gráfica.

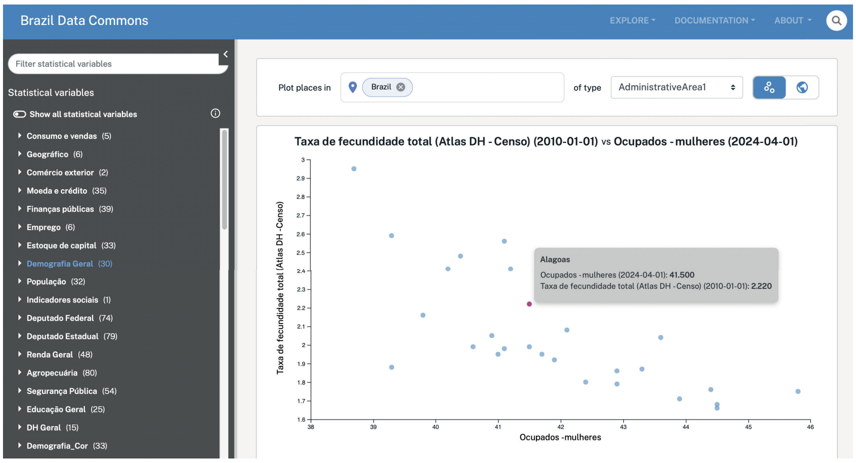
A Figura 2(b) ilustra um cenário em que o usuário seleciona a opção "Scatter Plot Explorer". Após ser direcionado para um novo painel, o usuário insere o nome de um local —um país, um município ou qualquer outra área administrativa definida por cada região— para visualizar os conjuntos de dados disponíveis. Em seguida, o usuário seleciona o nível de agregação no menu suspenso à esquerda para definir como a distribuição dos dados será exibida. Por fim, escolhe duas variáveis estatísticas, uma para o eixo x e outra para o eixo y, no lado direito da interface. Neste exemplo, vemos "Taxa de Fecundidade Total" vs. "Mulheres Empregadas" entre todos os estados brasileiros, com o ponto representando o estado de Alagoas destacado.

A Figura 2(c) mostra um mapa, outra ferramenta de visualização gráfica. O uso do "Map Explorer" é semelhante ao do "Scatter Plot Explorer", com a diferença crucial de que apenas uma variável estatística pode ser selecionada por vez. Neste exemplo, a variável apresentada é a "Expectativa de Vida ao Nascer" para cidades no Brasil. Por fim, ambas as ferramentas de visualização oferecem opções adicionais de personalização,

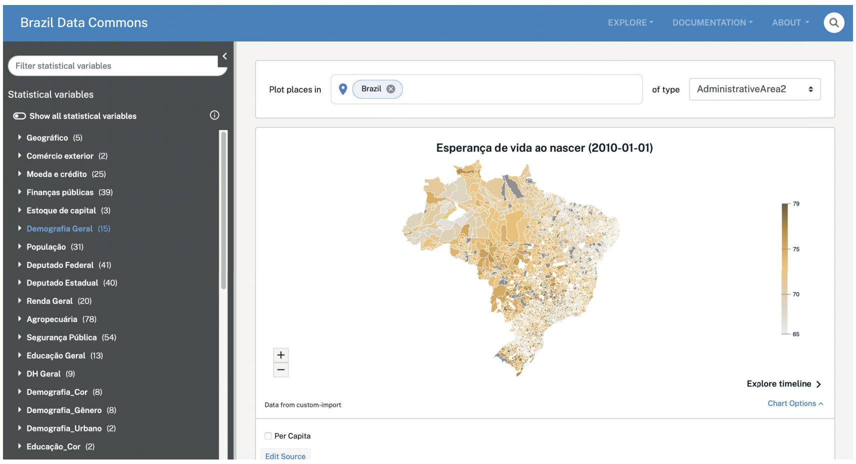
[6]O repositório da plataforma está disponível em: https://github.com/iscris/data-commons-brasil.



(a) Página inicial



(b) Visualização – Scatter plot



(c) Visualização – Map

Figura 2. Capturas de tela do BRAZIL DATA COMMONS: (a) Página inicial; (b) Exemplo de visualização gráfica – Scatter plot: "Taxa de Fecundidade Total" vs. "Mulheres Empregadas" entre todos os estados brasileiros, e (c) Exemplo de visualização gráfica – Map: "Expectativa de Vida ao Nascer"para cidades no Brasil.

como trocar variáveis entre os eixos, visualizar dados com dimensões territoriais e adicionar rótulos aos pontos, entre outras. Uma demonstração da plataforma BRAZIL DATA COMMONS pode ser acessada em: https://youtu.be/cbHIzOPqiIQ.

4. Exemplos de Uso dos Dados e Limitações

Esta seção apresenta alguns estudos de caso potenciais que podem ser realizados com os dados disponíveis no BRAZIL DATA COMMONS, os quais podem ser relacionados aos dados disponíveis no DATA COMMONS global, assim como considerações éticas e as possíveis limitações envolvidas no uso desses dados.

4.1. Análises Descritivas Básicas

Uma possível área de estudo é investigar a evolução recente da renda mensal do trabalho no Brasil. A análise básica permite observar como os salários estão evoluindo no país e identificar períodos de crescimento acelerado ou de estagnação. A linha do tempo que ilustra a página principal do site mostra o impacto da pandemia de Covid-19 no aumento da renda dos trabalhadores brasileiros.

4.2. Análise Espacial em Nível Local

Os dados coletados, ao contrário de outras iniciativas, possuem informações em nível municipal no Brasil. O projeto contém informações sobre a expectativa de vida ao nascer para todas as 5.500 localidades brasileiras. A expectativa de vida é uma medida interessante de desenvolvimento social e econômico, permitindo identificar diferenças regionais de forma mais ampla.

4.3. Visualização e Comparação de Dados

O projeto fornece dados para o Brasil, mas também é possível acessar informações semelhantes para diferentes países. Uma área de interesse é comparar a evolução do nível educacional no Brasil com a de outros países desenvolvidos e em desenvolvimento ao longo do tempo.

4.4. Comparação Internacional

Os dados brasileiros disponibilizados no BRAZIL DATA COMMONS permitem comparações internacionais com informações de outros países que já fazem parte do DATA COMMONS. Conduzimos uma comparação entre diferentes países em relação à média de anos de escolaridade e observamos que o Brasil está atrás tanto de economias desenvolvidas quanto de economias em desenvolvimento.

4.5. Uso Ético dos Dados e Limitações

Por fim, o BRAZIL DATA COMMONS disponibiliza os dados em sua forma original, sem quaisquer correções ou ajustes. Há esforços na literatura que apontam as limitações de certos conjuntos de dados brasileiros [Shikida et al. 2021], mas que também indicam melhorias nesses dados ao longo do tempo. Por exemplo, os dados de mortalidade — tanto gerais quanto por causa — são conhecidos por suas limitações, como erros na declaração de idade, sub-registro de óbitos, entre outros. Os dados de natalidade também podem apresentar problemas de registro e erros semelhantes. Os usuários do nosso sistema devem estar cientes dessas limitações dos dados e empregar métodos adequados de avaliação e correção, quando necessário.

5. Conclusão e Trabalhos Futuros

Ao fornecer uma plataforma unificada e interoperável para acessar e comparar conjuntos de dados públicos brasileiros, o BRAZIL DATA COMMONS dá um passo significativo rumo à democratização da pesquisa baseada em dados e da tomada de decisões no Brasil, o que favorece pesquisas no contexto de sistemas colaborativos. A integração de diversas fontes de dados em um arcabouço semântico consistente está alinhada com padrões e melhores práticas globais, enriquecendo o ecossistema internacional de dados. Por meio de interfaces intuitivas, APIs acessíveis e adesão aos princípios de dados abertos, a plataforma incentiva uma participação mais ampla da comunidade, promovendo uma cultura de alfabetização em dados e formulação de políticas baseadas em evidências.

Olhando para o futuro, vislumbramos duas áreas principais de desenvolvimento. Primeiro, a arquitetura da plataforma será expandida para suportar contribuições de colaboradores externos, convidando pesquisadores, desenvolvedores e especialistas de diversas áreas a enriquecer e refinar tanto os dados quanto os esquemas subjacentes. Ao adotar um modelo participativo, o BRAZIL DATA COMMONS busca ampliar a diversidade e profundidade dos conjuntos de dados disponíveis, melhorar sua qualidade e garantir a relevância da plataforma ao longo do tempo.

Segundo, à medida que a plataforma amadurece e acumula uma vasta quantidade de informações, ela possibilita análises mais complexas. Pesquisadores poderão aprofundar investigações sobre questões críticas na interseção entre transparência e privacidade, explorando o equilíbrio ideal entre iniciativas de dados abertos e a proteção de informações sensíveis. Da mesma forma, o crescente volume de conjuntos de dados integrados fomenta pesquisas sobre temas como equidade, detecção de vieses e medidas teóricas da informação, que podem orientar uma melhor governança de dados e enriquecer metodologias de ciências sociais quantitativas.

Em suma, o BRAZIL DATA COMMONS representa uma base essencial para trabalhos futuros, com potencial para evoluir em um ecossistema colaborativo dinâmico que apoie usos mais aprofundados, responsáveis e impactantes dos dados públicos brasileiros. No entanto, os usuários devem estar cientes de que a qualidade dos dados varia significativamente entre as fontes, tornando essencial a avaliação criteriosa antes do uso e da publicação de estudos baseados nesses dados. Novas fontes de dados abrem uma infinidade de oportunidades para compreender mudanças sociais e econômicas, mas métodos analíticos tradicionais ainda são necessários [Breen and Feehan 2024].

De maneira geral, o exemplo do Brasil pode ser útil para outros países do Sul Global. Investimentos em fontes públicas de dados acessíveis são fundamentais para promover o desenvolvimento em todas as áreas. Além disso, o aumento do número de usuários de dados por meio do DATA COMMONS pode também contribuir para a melhoria da qualidade dos dados administrativos.

Agradecimentos

Os autores agradecem à CAPES, ao CNPq, à FAPEMIG e, especialmente, à Google pelo financiamento de diferentes partes deste trabalho.

Referências

Base dos Dados (2025). Base dos Dados: Dados Abertos e Tratados para o Brasil. Disponível em: https://basedosdados.org. Acessado em 20/03/2025.

Blicharska, M., Smithers, R. J., Kuchler, M., Agrawal, G. K., Gutiérrez, J. M., Hassanali, A., Huq, S., Koller, S. H., Marjit, S., Mshinda, H. M., et al. (2017). Steps to overcome the north–south divide in research relevant to climate change policy and practice. *Nature Climate Change*, 7(1):21–27.

Breen, C. F. and Feehan, D. M. (2024). New data sources for demographic research. *Population and Development Review*.

da Cruz Martins, S., Mauritti, R., and da Costa, A. F. (2013). Acesso a bases de microdados: aplicações e impactos nas pesquisas em ciências sociais. *Mediações-Revista de Ciências Sociais*, 18(1):66–82.

Dang, H.-A. H., Pullinger, J., Serajuddin, U., and Stacy, B. (2023). Statistical performance indicators and index—a new tool to measure country statistical capacity. *Scientific Data*, 10(1):146.

Eurostat (2025). Eurostat - Statistical Office of the European Union. Disponível em: https://ec.europa.eu/eurostat. Acessado em 20/03/2025.

Freitas, E. E., Romero, J. P., Britto, G., de Queiroz Stein, A., and Torres, R. (2023). Dataviva: espaço de atividades e indicadores regionais de complexidade econômica. Textos para discussão cedeplar-ufmg, Cedeplar, Universidade Federal de Minas Gerais.

Governo Federal do Brasil (2025). Portal Brasileiro de Dados Abertos. Disponível em: https://dados.gov.br/home. Acessado em 20/03/2025.

Guha, R. V., Brickley, D., and Macbeth, S. (2016). Schema. org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51.

Guha, R. V. and Gupta, V. (2015). Communicating semantics: Reference by description. *arXiv preprint arXiv:1511.06341*.

Guha, R. V., Radhakrishnan, P., Xu, B., Sun, W., Au, C., Tirumali, A., Amjad, M. J., Piekos, S., Diaz, N., Chen, J., Wu, J., Ramaswami, P., and Manyika, J. (2023). Data commons. *arXiv preprint arXiv:2309.13054*.

Herrera, Y. M. and Kapur, D. (2007). Improving data quality: Actors, incentives, and capabilities. *Political Analysis*, 15(4):365–386.

Passos, J. (2022). Falta de integração e distribuição das bases de dados fragiliza sistemas de informação em saúde no país. Disponível em: https://www.epsjv.fiocruz.br/noticias/reportagem/falta-de-integracao-e-distribuicao-das-bases-de-dados-fragiliza-sistemas-de. Acessado em 06/12/2024.

Press, G. (2016). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. Disponível em: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=41cfc7606f63. Acessado em 28/07/2024.

Schema.org (2025). Schema.org - Structured Data on the Web. Disponível em: https://schema.org. Acessado em 20/03/2025.

Shikida, C. D., Monasterio, L., and Nery, P. F. (2021). Guia brasileiro de análise de dados: armadilhas & soluções.

World Bank (2025). World Bank Open Data. Disponível em: https://data.worldbank.org. Acessado em 20/03/2025.