

Impacto da Gestão de Dados na Aplicação de Inteligência Artificial na Pesquisa em Biodiversidade

The Impact of Data Management on the Application of Artificial Intelligence in Biodiversity

Guilherme Ataíde Dias, Universidade Federal da Paraíba.

Marckson Roberto Ferreira de Sousa, Universidade Federal da Paraíba.

Paulo Roberto Santos Costa, Instituto Federal da Paraíba

Janaína Nascimento de Araújo, Universidade Federal da Paraíba.

Alba Lúcia de Almeida Silva, Universidade Federal da Paraíba.

Resumo

O aumento do uso de ferramentas de inteligência artificial (IA) na pesquisa em biodiversidade ressalta a importância de dados de qualidade para a eficácia dessas tecnologias. Este estudo examinou como está estruturada a produção científica de pesquisadores na área de biodiversidade no que tange ao emprego de ferramentas de IA e investigou o posicionamento de pesquisadores brasileiros sobre práticas de gestão de dados de pesquisa. De natureza exploratória, a pesquisa utilizou métodos bibliográficos e questionários aplicados a pesquisadores vinculados a programas de pós-graduação com conceitos CAPES 6 e 7 no Brasil. Os resultados evidenciaram lacunas significativas, como a baixa adesão a Planos de Gestão de Dados e padrões de metadados, além de dificuldades relacionadas à coleta e ao reuso de dados. Apesar do crescente interesse no uso de IA, as práticas de gestão de dados permanecem insuficientes para atender às demandas dessa tecnologia. Conclui-se que a ausência de boas práticas de gestão de dados compromete a qualidade dos conjuntos de dados, podendo impactar negativamente a precisão e a eficácia das análises realizadas por ferramentas de IA. Esses achados reforçam a urgência de iniciativas voltadas à capacitação e à disseminação de boas práticas de gestão de dados no contexto da biodiversidade.

Palavras-chave: gestão de dados de pesquisa; biodiversidade; inteligência artificial; formulação de hipóteses.

Abstract

The increasing use of artificial intelligence (AI) tools in biodiversity research highlights the importance of quality data for the effectiveness of these technologies. This study examined how scientific production in the field of biodiversity is structured regarding the use of AI tools and investigated the perspectives of Brazilian researchers on research data management practices. Exploratory in nature, the research employed bibliographic methods and questionnaires administered to researchers affiliated with postgraduate programs rated CAPES 6 and 7 in Brazil. The results revealed significant gaps, such as low adherence to Data Management Plans and metadata standards, as well as challenges related to data collection and reuse. Despite the growing interest in the use of AI, data management practices remain insufficient to meet the demands of this technology. It is concluded that the absence of good data management practices compromises the quality of datasets, potentially negatively

impacting the accuracy and effectiveness of analyses performed by AI tools. These findings underscore the urgency of initiatives focused on training and promoting good data management practices within the biodiversity context.

Keywords: research data management; biodiversity; artificial intelligence; hypothesis formulation.

1. Introdução

O crescente uso de ferramentas de Inteligência Artificial (IA) por pesquisadores tem se consolidado como um fenômeno cada vez mais evidente no contexto científico contemporâneo. Historicamente, a incorporação de novas tecnologias tem promovido transformações significativas nas metodologias e práticas científicas (Bianchini et al., 2024). O avanço do conhecimento científico está intrinsecamente ligado à integração de tecnologias emergentes. Em especial, as ferramentas de IA apresentam um potencial único para revolucionar as dinâmicas de produção de conhecimento, alcançando um impacto superior ao proporcionado por tecnologias anteriores, como os métodos estatísticos clássicos e os sistemas de recuperação da informação tradicionais (Bianchini et al., 2024).

A emergência das ferramentas de IA está fundamentada em um outro fenômeno que se consolida continuamente: a produção e o tratamento de grandes volumes de dados, impulsionados pela disseminação de dispositivos de tecnologia da informação em todas as esferas da atuação humana. Nesse contexto, Aroyo et al. (2022, p.66) afirmam que: “A eficácia dos modelos de aprendizado de máquina (ML) depende tanto dos algoritmos quanto dos dados.”¹.

O crescimento exponencial na geração e no uso de dados por governos, indústrias e na ciência tem agregado valor significativo aos conjuntos de dados, um fato amplamente reconhecido no artigo clássico, do periódico *The Economist*, que destaca que os dados são o novo petróleo (The Economist, 2017). Para que a IA seja operacionalizada de forma eficiente, é indispensável que os conjuntos de dados sejam submetidos a processos de curadoria específicos, relacionados a modelos de Ciclo de Vida de Dados (CVD). Esses modelos agregam valor aos dados em todas as etapas de seu ciclo de vida (Rodrigues, 2021).

Dados que passam por processos adequados de curadoria estão mais bem preparados para serem processados por ferramentas de IA. Nesse sentido, vale reiterar a contribuição de Aroyo et al. (2022, p.66) que destacam: “Se ‘os dados são o novo petróleo’, ainda nos falta

¹ Texto na língua original: “The efficacy of machine learning (ML) models depends on both algorithms and data.”

trabalho nas refinarias pelas quais os próprios dados poderiam ser otimizados para um uso mais eficaz.”².

Essa abordagem destaca a necessidade de investir na qualidade dos dados como um fator crítico para maximizar o potencial das ferramentas de IA, evidenciando que os processos de curadoria desempenham um papel essencial no progresso científico e tecnológico.

A importância de conjuntos de dados de qualidade remete a um conceito clássico da computação, conhecido como *Garbage In, Garbage Out* (GIGO). De forma literal, a expressão idiomática em língua inglesa “*Garbage In, Garbage Out*” significa algo como “Lixo entra, Lixo sai” em língua portuguesa. Esse conceito é especialmente relevante no contexto atual, marcado pela crescente dependência de dados em todas as esferas das atividades humanas (Dias et al., 2024).

Embora os avanços no desenvolvimento de *software* tenham promovido maior consistência nos processos de tratamento de dados, os desafios relacionados à entrada de dados inconsistentes permanecem significativos, afetando desde sistemas tradicionais, como os de gestão empresarial, até os mais sofisticados sistemas IA, baseados em modelos de aprendizado profundo que dependem integralmente de vastos volumes de dados (Dias et al., 2024).

A biodiversidade constitui uma das áreas do conhecimento humano que faz uso intensivo de dados e, ao longo dos últimos anos, os pesquisadores desta área têm ampliado significativamente a aplicação de ferramentas de IA, conforme será discutido mais adiante. Pesquisas recentes que trazem a intersecção entre IA e biodiversidade incluem, de forma não exaustiva, as contribuições de Edosa et al. (2024), Grillo et al. (2024), Nel et al. (2024), Orejuela-Escobar et al. (2024) e Zhou et al. (2024).

A partir dessas considerações iniciais, indica-se que esta investigação tem como objetivos apresentar o posicionamento de pesquisadores da área da biodiversidade com relação a aspectos selecionados da gestão de dados de pesquisa – conhecida na literatura internacional com *Research Data Management* (RDM) – e apontar como está delineada a produção científica de pesquisadores na área da biodiversidade no que diz respeito ao emprego de ferramentas de inteligência artificial na sua área específica. A pesquisa proposta é exploratória com fito a dar suporte à proposição de questionamentos e hipóteses.

2. Qualidade dos Dados e o Ciclo de Vida de Dados: Fundamentos para Aplicações em

² Texto na língua original: “If “data is the new oil,” we are still missing work on the refineries by which the data itself could be optimized for more effective use.”

Biodiversidade

Aroyo et al. (2022) destacam que um dos aspectos menos glamourizados no ecossistema de IA atualmente são as questões relacionadas ao tratamento inicial dos dados, frequentemente consideradas atividades tediosas e que demandam tempo. No entanto, é fundamental reconhecer que as ferramentas de IA dependem diretamente da qualidade dos dados para realizar previsões de forma adequada. Esse fato está intrinsecamente conectado aos processos de curadoria, os quais se alinham aos modelos do Ciclo de Vida de Dados (CVD). Ademais, com base nos trabalhos de Aroyo et al. (2022) e Dias et al. (2024), destacamos alguns aspectos relacionados à qualidade dos dados que podem influenciar diretamente a precisão dos resultados obtidos pelo uso de ferramentas de inteligência artificial. Para fins didáticos, adotamos, neste caso específico, o modelo proposto pelo DataONE (DataONE, n.d.), embora, existam outras possibilidades de modelos, tais como o Ciclo de Vida de Dados para a Ciência da Informação (Sant’Ana, 2016). e o *Curation Lifecycle Model* (Digital Curation Centre, n.d.):

- **Planejar** (*Plan*): Nessa etapa, são mapeados os processos e recursos necessários para todo o CVD. É nesse momento que se deve elaborar o Plano de Gestão de Dados (PGD), garantindo que as diretrizes e estratégias para o manejo dos dados estejam nitidamente definidas;
- **Coletar** (*Collect*): Na etapa de coleta, os dados são obtidos de forma manual ou com o auxílio de dispositivos tecnológicos da informação. Sempre que necessário, esses dados são convertidos para o formato digital, assegurando sua adequação aos processos subsequentes das etapas do CVD;
- **Assegurar** (*Assure*): Essa etapa é dedicada à aplicação de procedimentos voltados à garantia e verificação da qualidade, abrangendo ações como a calibração dos instrumentos de coleta e a análise da consistência semântica e sintática dos dados. Trata-se de uma instância essencial para mitigar o fenômeno GIGO. Além disso, diversas ferramentas de software estão disponíveis para apoiar as atividades realizadas nessa etapa do CVD;
- **Descrever** (*Describe*): Os dados devem ser descritos de maneira explícita e detalhada, incorporando informações que favoreçam seu uso, reúso e compartilhamento. É fundamental considerar os padrões de metadados estabelecidos pelas comunidades específicas do domínio. No caso da área de biodiversidade, por exemplo, recomenda-

se a adoção do padrão *Darwin Core*;

- **Preservar** (*Preserve*): A etapa de preservação é essencial para garantir que os dados de pesquisa permaneçam disponíveis a longo prazo, assegurando a possibilidade de serem reutilizados sempre que necessário. Preservar não se resume a realizar *backups* em dispositivos de armazenamento de massa, mas constitui um processo que exige planejamento cuidadoso, incluindo a guarda de elementos descritivos que previnam a obsolescência de *hardware* e *software*. Além disso, recomenda-se a adoção de formatos abertos, que favorecem a interoperabilidade e a longevidade dos dados. Para a etapa de arquivamento permanente, sugere-se a aderência ao modelo *Open Archival Information System* (OAIS);
- **Descobrir** (*Discover*): etapa de descoberta está associada à identificação e atribuição de dados complementares que possam agregar valor aos dados vinculados ao projeto de pesquisa. Recomenda-se a atribuição de identificadores persistentes aos dados, bem como o seu depósito em repositórios de dados abertos. No contexto específico da biodiversidade, exemplos de repositórios incluem o *Global Biodiversity Information Facility*³ (GBIF), o *DATA PB*, entre outros;
- **Integrar** (*Integrate*): A etapa de integração consiste em combinar, de forma direcionada, outras fontes de dados necessárias para ampliar as análises e investigações relacionadas aos dados do projeto de pesquisa. Essa etapa depende do cumprimento rigoroso das boas práticas de gestão de dados ao longo do ciclo de vida adotado no projeto, bem como da qualidade e organização dos dados de terceiros, gerenciados por seus respectivos curadores. No contexto da biodiversidade, é frequentemente necessário integrar dados biológicos com dados que representam indicadores socioeconômicos;
- **Analisar** (*Analyse*): A etapa de análise está relacionada à execução de análises e à criação de visualizações com o objetivo de identificar padrões, testar hipóteses e ilustrar descobertas. Recomenda-se que, durante essa etapa, sejam registrados os métodos utilizados, bem como cada uma das etapas do processamento de dados. É igualmente importante documentar os produtos de *software* empregados nesse processo. Um caderno de laboratório pode ser utilizado para registrar todas as etapas da análise, servindo posteriormente como uma fonte probatória confiável das operações realizadas com os dados.

³ Disponível a partir da seguinte URL: <https://www.gbif.org/>.

É necessário destacar que, além da adoção de um modelo de CVD para uma gestão adequada de dados de pesquisa, a aderência a determinados padrões também desempenha um papel importante. Nesse sentido, destaca-se o uso dos princípios FAIR, que visam ampliar a encontrabilidade, acesso, interoperabilidade e reuso dos dados (Wilkinson et al., 2016).

3. Percurso metodológico

A pesquisa caracteriza-se como exploratória, permitindo o levantamento de questionamentos e a formulação de hipóteses que contribuam para um aprofundamento analítico acerca do impacto da gestão de dados de pesquisa nos resultados produzidos por ferramentas de IA.

Para analisar como se configura a produção científica na área da biodiversidade, no que se refere ao uso de IA, foi realizada uma pesquisa bibliográfica no Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) entre os dias 27/01/2025 e 28/01/2025. A busca não estabeleceu restrições temporais quanto à data de publicação dos artigos recuperados. O escopo da pesquisa incluiu periódicos de todas as bases de dados acessíveis por meio do portal, tendo sido selecionado o campo de busca referente ao título dos artigos.

A expressão de busca, elaborada por meio da ferramenta de busca avançada do portal, permitiu recuperar artigos cujos títulos contivessem os termos “Inteligência Artificial” e “Biodiversidade” ou “Machine Learning” e “Biodiversidade” ((**“Inteligência artificial” AND “Biodiversidade”**) OR (**“Machine Learning” AND “Biodiversidade”**)). Não foram aplicadas restrições quanto ao uso de maiúsculas ou minúsculas, nem quanto ao país ou idioma das publicações.

No total, foram recuperados 42 artigos. Destes, dois foram eliminados por não se enquadrarem tecnicamente como artigos científicos, resultando em um total de 40 artigos analisados (vide Figura 1). O critério de inclusão dos artigos foi que eles possuísem conteúdo substantivo relacionado à aplicação de inteligência artificial e aprendizado de máquina no domínio da biodiversidade.

Os dados utilizados nesta investigação, com o objetivo de avaliar o posicionamento de pesquisadores(as) da área de biodiversidade em relação a aspectos selecionados da gestão de dados de pesquisa, foram coletados por meio de um questionário aplicado a pesquisadores(as) vinculados(as) a programas de pós-graduação brasileiros classificados na área de biodiversidade e avaliados com conceitos 6 e 7 pela CAPES. Ressalta-se que o foco recai sobre pesquisadores(as) vinculados(as) a tais programas, cuja área de avaliação é

biodiversidade, ainda que sua formação acadêmica possa abranger distintas áreas do conhecimento. Tal característica é usual nesses programas, os quais, embora possuam um foco temático definido, frequentemente integram discentes oriundos(as) de áreas correlatas, o que enriquece a abordagem interdisciplinar característica da pesquisa em biodiversidade. Esses dados fornecem subsídios para compreender as práticas adotadas por esses pesquisadores no âmbito da gestão de dados de pesquisa. O questionário foi disponibilizado por meio da ferramenta *Google Forms* no período de 08/10/2024 a 20/01/2025. Dos 653 questionários enviados, objetou-se o retorno efetivo de 148 respondentes, o que corresponde a uma taxa de retorno de 22,7%. No que se refere à taxa de retorno de questionários disponibilizados de forma eletrônica, Nulty (2008) evidencia que estas são consideravelmente inferiores às obtidas por meio de questionários aplicados presencialmente. Ademais, quando a pesquisa tem por objetivo gerar *insights* iniciais ou explorar questões, taxas mais baixas de retorno podem ser consideradas aceitáveis.

Os dados coletados por meio dos questionários foram disponibilizados em acesso aberto no repositório de dados da Universidade Federal da Paraíba, denominado DATAPB, acessível através da URL <https://dataverse.ideal.ufpb.br/> (Costa et al., 2024). O instrumento de coleta dos dados, composto basicamente por questões objetivas, foi devidamente respaldado pelo Comitê de Ética do Centro de Ciências da Saúde da Universidade Federal da Paraíba, conforme o parecer nº 7.286.870. Os dados estavam anonimizados, em conformidade com a Lei Geral de Proteção de Dados (LGPD). O arquivo disponibilizado no repositório de dados estava no formato texto, estruturado no padrão *Comma-Separated Values* (CSV). Essa configuração permite a organização dos dados em formato tabular, com colunas delimitadas por vírgulas, o que facilita sua interpretação e integração com diversas ferramentas de análise e gestão de dados. Os dados obtidos foram inicialmente processados utilizando a ferramenta *Microsoft Excel*, enquanto os gráficos e tabelas foram elaborados com o auxílio do modelo de IA *ChatGPT* (Modelo *GPT-4o*). Considerando o risco de alucinação inerente ao uso de modelos de linguagem, as informações apresentadas nos gráficos foram conferidas pelos autores, assegurando a acurácia e a fidedignidade dos resultados reportados.

As variáveis selecionadas para análise no conjunto de dados abrangem a familiaridade e o uso do PGD nas pesquisas realizadas pelos respondentes, bem como aspectos relacionados às dificuldades enfrentadas na coleta de dados, à adoção de padrões de metadados e ao reuso de informações provenientes de repositórios de dados da biodiversidade. Tais variáveis apresentam uma relação mais significativa com as etapas de planejamento, coleta, descrição e descoberta, conforme estabelecido no modelo de CVD do DataONE (DataONE, n.d.).

A seleção de um conjunto de dados provenientes de pesquisadores na área da biodiversidade justifica-se pelo fato de estarmos, no momento, conduzindo pesquisas voltadas a esse domínio e, adicionalmente, por ser uma área em que há um uso significativo de dados por ferramentas de inteligência artificial. Essa escolha reflete tanto a relevância do tema nas investigações em andamento quanto o potencial da biodiversidade como campo de aplicação para tecnologias avançadas.

4. Apresentação e análise dos dados

Nesta seção, apresenta-se gráficos e tabelas que visam proporcionar uma compreensão mais detalhada dos dados coletados.

A análise do gráfico apresentado na Figura 1 indica um crescimento na quantidade de artigos publicados na área da biodiversidade que fazem uso de ferramentas de IA ao longo dos anos. A partir de 2019, observa-se um aumento significativo na produção científica, com uma tendência de crescimento contínuo até 2024.

Embora haja flutuações anuais, como a queda observada em 2020, verifica-se uma retomada no número de publicações a partir de 2021, culminando no pico em 2024. Além disso, a proporção relativa das publicações acompanha essa trajetória ascendente, sugerindo um crescente interesse na interseção entre biodiversidade e IA.

Contudo, para afirmar categoricamente a existência de uma tendência consolidada de crescimento, é necessário considerar fatores externos que podem influenciar a produção científica, tais como financiamento para pesquisas, avanços tecnológicos e eventos acadêmicos. Ainda assim, os dados analisados indicam um aumento do interesse e da aplicação de IA na biodiversidade, o que sugere uma tendência de crescimento contínua nos próximos anos.

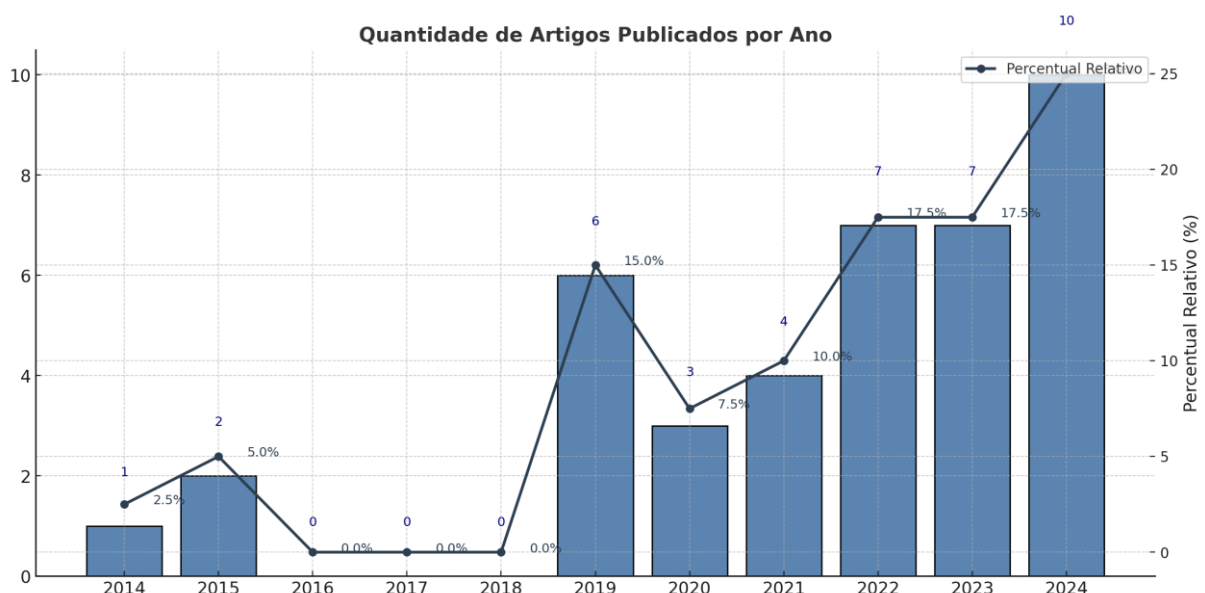


Figura 1: Artigos publicados na área da biodiversidade que fazem uso de ferramentas de IA.

Fonte: Elaborado pelos autores a partir de dados obtidos em Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, n.d.)

A seguir, apresentamos e analisamos algumas das variáveis associadas com o conjunto de dados provenientes do questionário. A primeira variável a ser analisada está relacionada com a familiaridade na elaboração do PGD por parte dos pesquisadores respondentes. A Tabela 1 apresenta as frequências absolutas e relativas relacionadas a familiaridade dos pesquisadores respondentes com a elaboração do PGD.

Tabela 1. Frequências absolutas e relativas relacionadas a familiaridade com a elaboração do PGD.

Categoria	Frequência Absoluta	Frequência Relativa (%)
Desconheço totalmente o tema	30	21,6 %
Tenho pouca familiaridade com o tema	29	20,9 %
Tenho familiaridade moderada com o tema	29	20,9 %
Estou razoavelmente familiarizado com o tema	28	20,1 %
Estou plenamente familiarizado com o tema	23	16,5 %
Total	139	100,0 %

Fonte: Elaborado pelo autor com base em Costa et al (2024).

A Figura 2 ilustra, por meio de um gráfico de colunas a familiaridade dos pesquisadores respondentes com a elaboração do PGD.

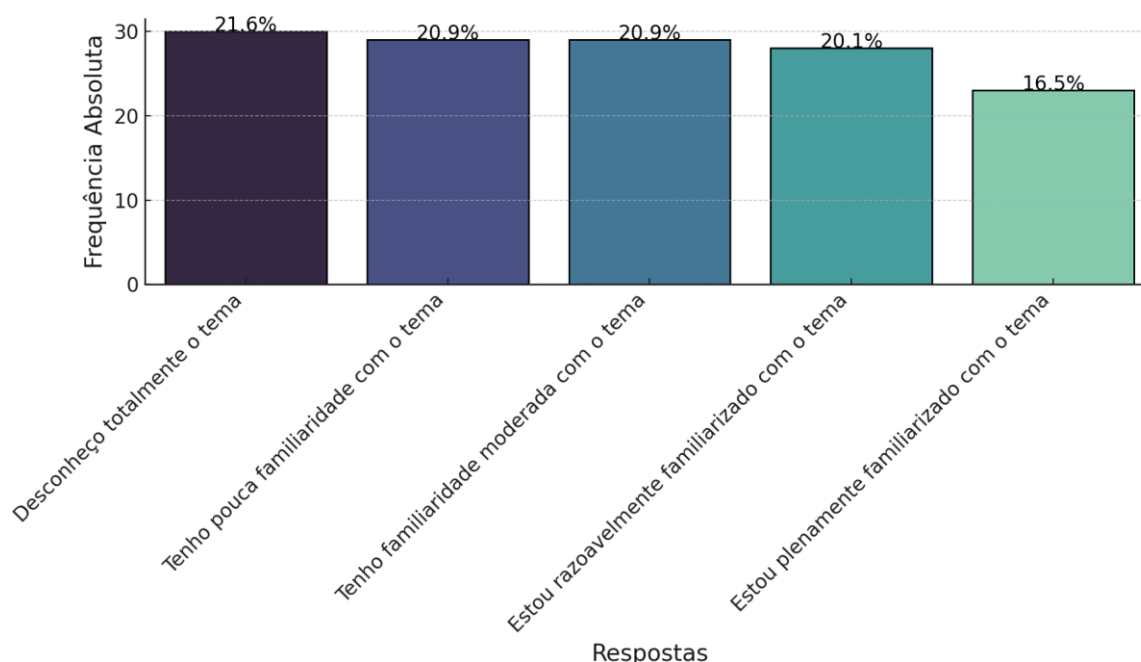


Figura 2: Familiaridade com a elaboração do plano de gestão de dados PGD.

Fonte: Elaborado pelos autores com base em Costa et al (2024).

Considerando a Tabela 1, observa-se que a categoria “Desconheço totalmente o tema” corresponde a 21,6 % das respostas (30 respostas), enquanto as categorias “Tenho pouca familiaridade com o tema” e “Tenho familiaridade moderada com o tema” representam, cada uma, 20,9 % das respostas (29 respostas cada). Com base na concentração de 63,3 % (88 respostas) nessas categorias, pode-se inferir que a maioria dos pesquisadores participantes da pesquisa apresenta limitações de conhecimento sobre os fundamentos e a importância de um PGD.

A Tabela 2 apresenta as frequências absolutas e relativas relacionadas com a variável que expressa a inclusão de um PGD nas pesquisas dos pesquisadores respondentes do instrumento de pesquisa.

Tabela 2. Frequências absolutas e relativas relacionadas com a inclusão de plano de PGD.

Categoria	Frequência Absoluta	Frequência Relativa (%)
Ocasionalmente	43	30,9 %
Nunca	28	20,1 %
Sempre	27	19,4 %
Raramente	26	18,7 %
Não sei informar	15	10,8 %
Total	139	100,0 %

Fonte: Elaborado pelo autor com base em Costa et al (2024).

A Figura 3 ilustra, por meio de um gráfico de colunas, a frequência com que os pesquisadores respondentes do instrumento de pesquisa incluem um PGD nas suas respectivas pesquisas.

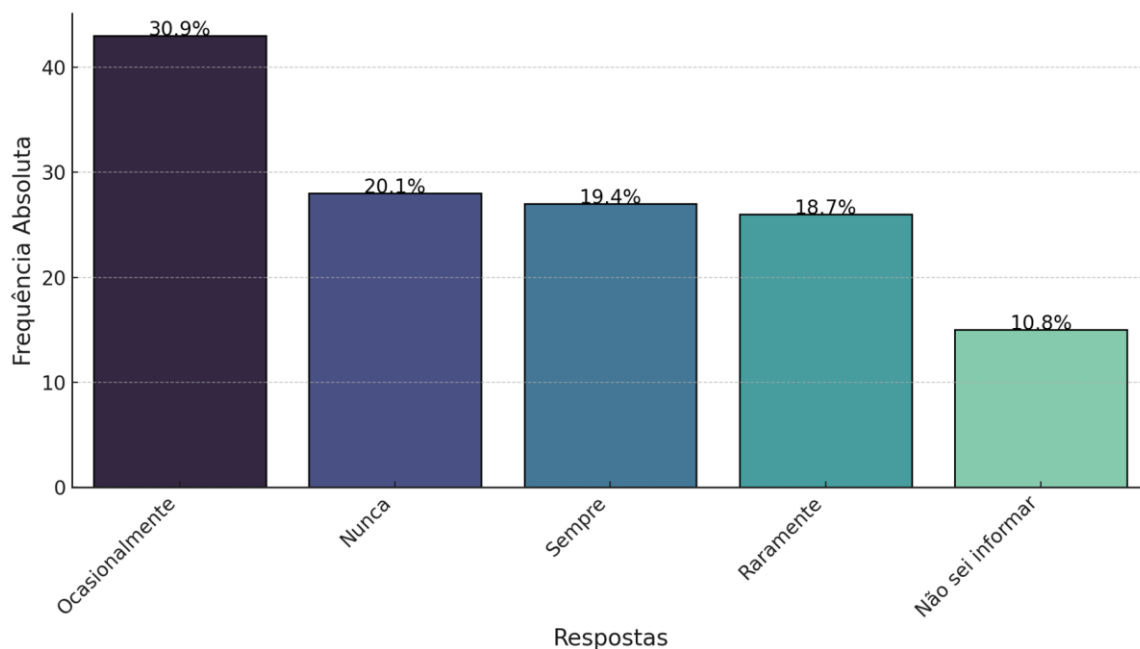


Figura 3: Inclusão de plano de gestão de dados PGD.

Fonte: Elaborado pelos autores com base em Costa et al (2024).

Observando a Tabela 2, verifica-se que a categoria “Ocasionalmente” corresponde a 30,9 % das respostas (43 respostas), enquanto as categorias “Nunca” e “Raramente” representam, respectivamente, 20,1 % (28 respostas) e 18,7 % (26 respostas). Com base na concentração de 69,8 % (97 respostas) nessas categorias, pode-se inferir que a maioria dos pesquisadores participantes da pesquisa apresenta limitações no que diz respeito à inclusão efetiva de um PGD, seja pela ausência total de sua utilização, seja pelo uso esporádico ou insuficiente.

A Tabela 3 apresenta as frequências absolutas e relativas relacionadas com a variável que expressa dificuldades na coleta dos dados de pesquisa por parte dos pesquisadores que responderam ao questionário.

Tabela 3. Frequências absolutas e relativas relacionadas com dificuldades na coleta dos dados de pesquisa.

Categoria	Frequência Absoluta	Frequência Relativa (%)
Ocasionalmente	92	66,2 %
Sempre	22	15,8 %
Raramente	19	13,7 %
Nenhuma	5	3,60 %
Não sei informar	1	0,7 %
Total	139	100,00 %

Fonte: Elaborado pelo autor com base em Costa et al (2024).

A Figura 4 ilustra, por meio de um gráfico de colunas, as dificuldades enfrentadas pelos pesquisadores na coleta de dados de pesquisa, conforme indicado nas respostas ao

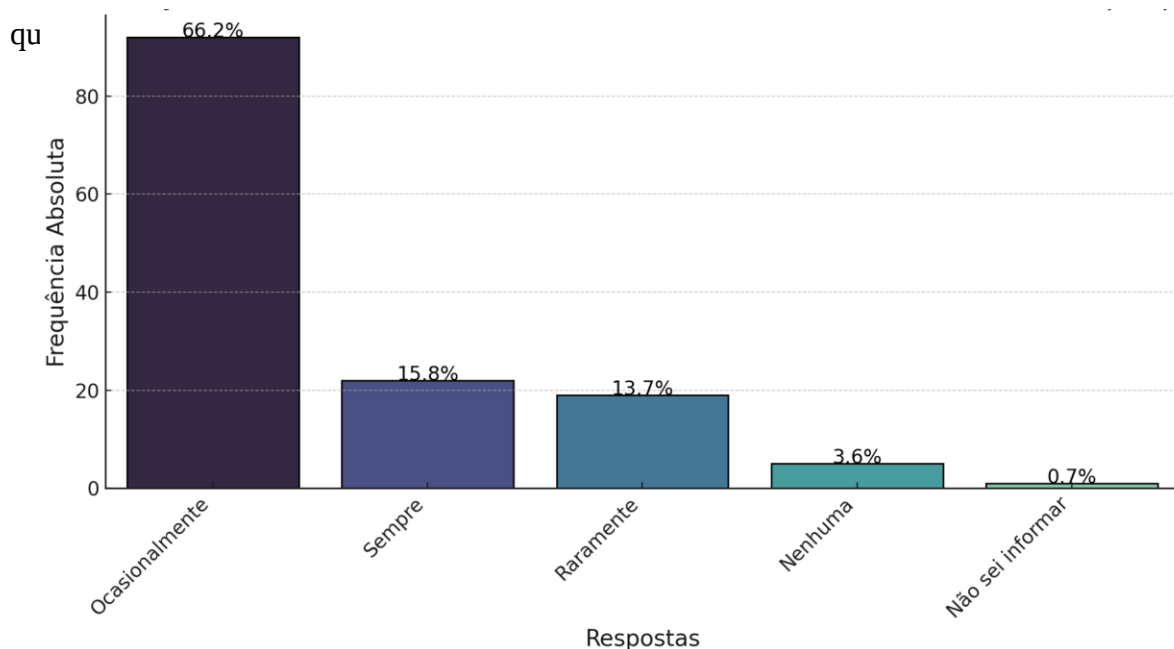


Figura 4: Dificuldades na coleta dos dados de pesquisa.

Fonte: Elaborado pelos autores com base em Costa et al (2024).

Observando a Tabela 3, verifica-se que a categoria “Ocasionalmente” corresponde a 66,2 % das respostas (92 respostas), enquanto a categoria “Sempre” representa 15,8 % das respostas (22 respostas). Com base na concentração de 82,0 % (114 respostas) nessas categorias, pode-se afirmar que a maioria dos pesquisadores participantes da pesquisa enfrenta algum grau de dificuldade relacionado à coleta de dados de pesquisa. A alta incidência na categoria “Ocasionalmente” sugere que, embora essas dificuldades não sejam constantes para muitos pesquisadores, elas ocorrem em situações específicas ou em determinados contextos de coleta, o que pode impactar a continuidade ou a eficiência do processo de pesquisa. Por outro lado, a representatividade da categoria “Sempre” indica que uma parcela significativa dos respondentes enfrenta desafios contínuos e generalizados na coleta de dados, o que pode limitar severamente o progresso de suas investigações.

A Tabela 4 apresenta as frequências absolutas e relativas relacionadas com a variável indicativa da adoção de padrões de metadados para descrever dados de pesquisa por parte dos pesquisadores que responderam ao questionário.

Tabela 4. Frequências absolutas e relativas relacionadas com adoção de padrões de metadados para descrever dados de pesquisa.

Categoria	Frequência Absoluta	Frequência Relativa (%)
Adoto ocasionalmente	45	32,4 %
Adoto sempre	35	25,2 %
Não sei informar	22	15,8 %
Não adoto	21	15,1 %
Adoto raramente	16	11,5 %
Total	139	100,00 %

Fonte: Elaborado pelo autor com base em Costa et al (2024).

A Figura 5 ilustra, por meio de um gráfico de colunas, a adoção de padrões de metadados para a descrição de dados de pesquisa pelos pesquisadores que responderam ao questionário.

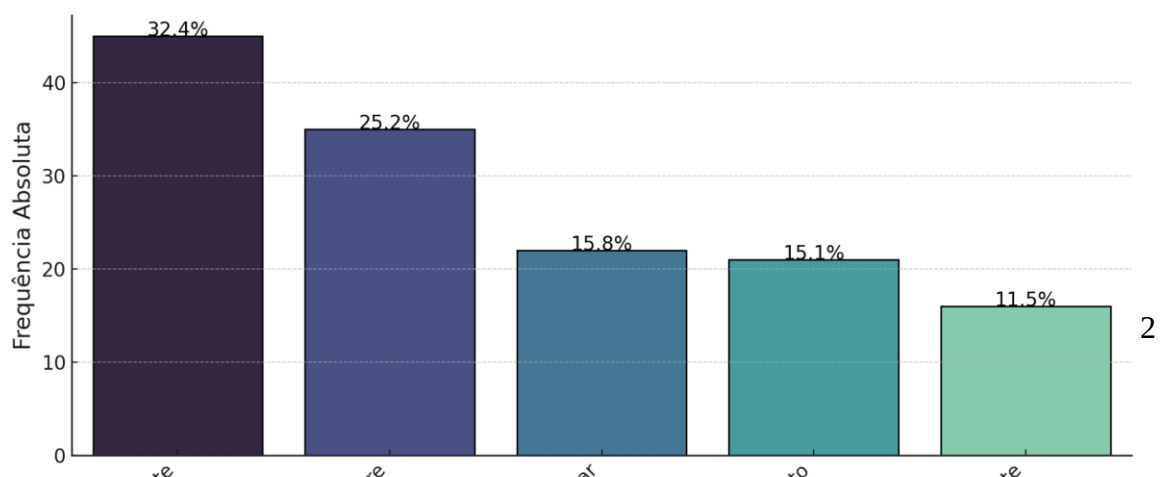


Figura 5: Adoção de padrões de metadados para descrever dados de pesquisa.

Fonte: Elaborado pelos autores com base em Costa et al (2024).

Observando a Tabela 4, verifica-se que a categoria “Adoto ocasionalmente” corresponde a 32,4 % das respostas (45 respostas), enquanto as categorias “Não sei informar”, “Não adoto” e “Adoto raramente” representam, respectivamente, 15,8 % (22 respostas), 15,1 % (21 respostas) e 11,5 % (16 respostas). Com base na concentração de 74,8 % (104 respostas) nessas categorias evidencia-se uma lacuna significativa no uso consistente de padrões de metadados para descrever dados de pesquisa. A predominância de respostas indicando adoção ocasional ou rara, bem como a presença de incerteza, reforça a necessidade de iniciativas voltadas para a capacitação dos pesquisadores nessa área. A disseminação de conhecimentos sobre os benefícios da adoção de padrões de metadados pode contribuir para maior uniformidade e qualidade na gestão, compartilhamento de dados de pesquisa e maior facilidade quando da análise dos dados.

A Tabela 5 apresenta as frequências absolutas e relativas referentes à variável que indica a frequência com que os pesquisadores, respondentes do questionário, reutilizam dados disponibilizados por repositórios de dados de biodiversidade.

Tabela 5. Frequências absolutas e relativas relacionadas com o reuso de dados de pesquisa disponíveis em repositórios de dados de biodiversidade.

Categoria	Frequência Absoluta	Frequência Relativa (%)
Ocasionalmente	69	49,6 %
Raramente	27	19,4 %
Sempre	26	18,7 %
Nunca	16	11,5 %
Não sei informar	1	0,7 %
Total	139	100,0 %

Fonte: Elaborado pelo autor com base em Costa et al (2024).

A Figura 6 ilustra, por meio de um gráfico de colunas, o reuso de dados de pesquisa disponibilizados em repositórios de biodiversidade por parte dos pesquisadores que participaram do questionário.

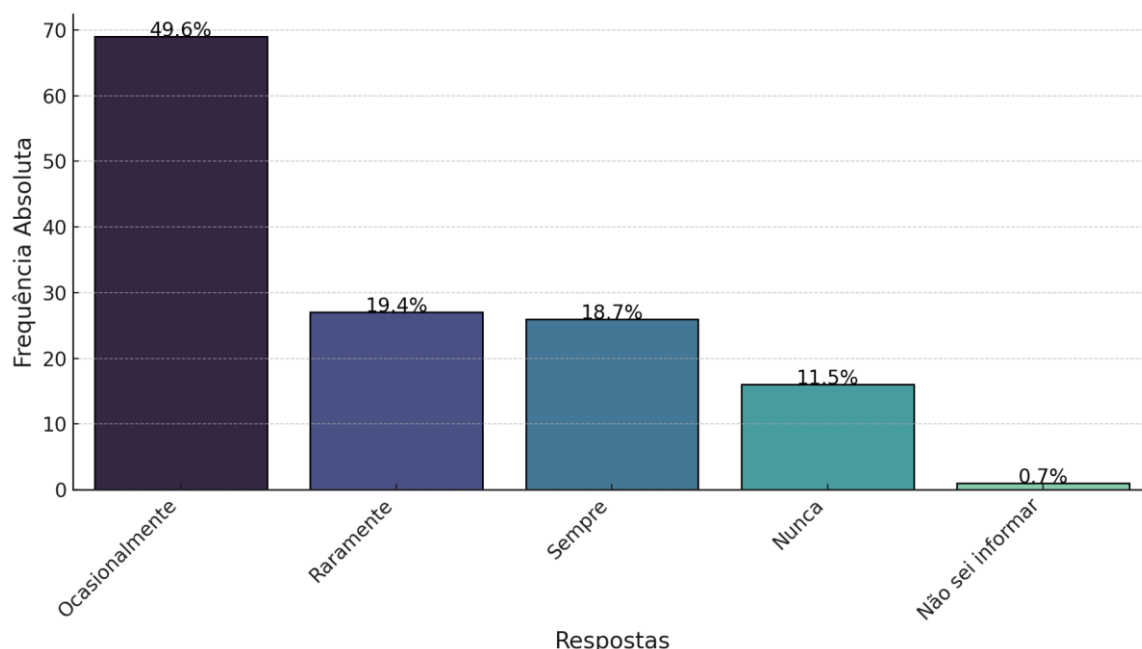


Figura 6: *Reuso de dados de pesquisa disponíveis em repositórios de dados de biodiversidade.*

Fonte: Elaborado pelos autores com base em Costa et al (2024).

A partir da observação dos dados disponibilizados na Tabela 5, verifica-se que a categoria “Ocasionalmente” corresponde a 49,6 % das respostas (69 respostas), enquanto as categorias “Raramente”, “Nunca” e “Não sei informar” representam, respectivamente, 19,4 % (27 respostas), 11,5 % (16 respostas) e 0,7 % (uma resposta). Com base na concentração de 81,3% (113 respostas) nessas categorias evidencia-se uma prática limitada do reuso de dados de repositórios de biodiversidade. Apesar de a categoria predominante indicar algum grau de utilização, a frequência ocasional sugere a necessidade de iniciativas que promovam a conscientização sobre os benefícios do reuso de dados, como a economia de recursos, maior reprodutibilidade, possibilidade de acesso a dados de qualidade e colaboração científica.

A Tabela 6 apresenta as frequências absolutas e relativas referentes à variável que indica como os pesquisadores da biodiversidade tratam os erros encontrados em seus dados de pesquisa.

Tabela 6. Frequências absolutas e relativas relacionadas a como os pesquisadores lidam com erros identificados nos seus dados de pesquisa.

Respostas	Frequência Absoluta	Frequência Relativa (%)
Corrijo manualmente	99	71,2 %

Outros membros da equipe tratam estes erros	21	15,1 %
Utilizo ferramentas automatizadas	16	11,5 %
Identifico a existência de erros, mas não os corrijo	2	1,4 %
Não sei informar como os erros são tratados	1	0,7 %
Total	139	100,0 %

Fonte: Elaborado pelo autor com base em Costa et al (2024).

A Figura 7 ilustra, por meio de um gráfico de colunas, como os pesquisadores respondentes do questionário lidam com erros identificados nos seus dados de pesquisa.

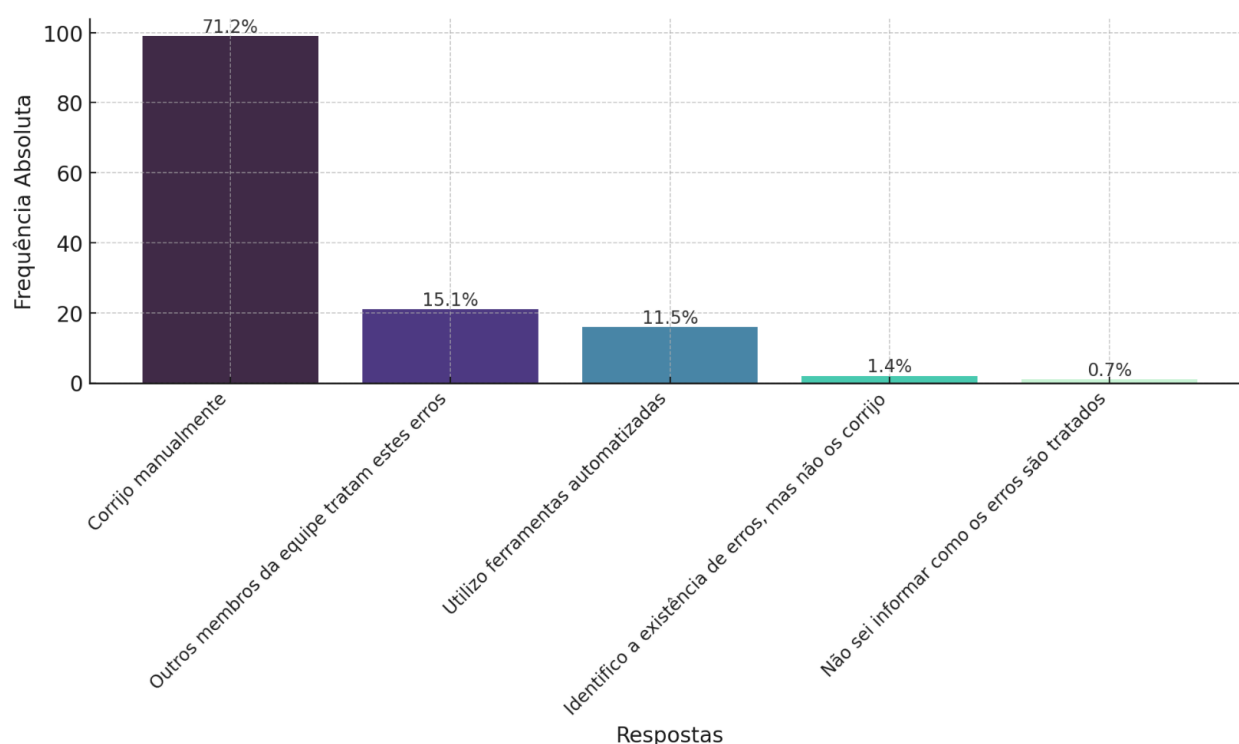


Figura 7: Como os pesquisadores lidam com erros encontrados nos seus dados de pesquisa.

Fonte: Elaborado pelos autores com base em Costa et al (2024).

A partir da observação dos dados disponibilizados no gráfico, verifica-se que a categoria “Corrijo manualmente” corresponde a 71,2% das respostas (99 respostas), enquanto as categorias “Identifico a existência de erros, mas não os corrijo”, “Não sei informar como os erros são tratados” e “Utilizo ferramentas automatizadas” representam, respectivamente, 1,4% (2 respostas), 0,7% (1 resposta) e 11,5% (16 respostas). Já a categoria “Outros membros da equipe tratam estes erros” totaliza 15,1% (21 respostas). Com base na concentração de 71,2% na categoria predominante, evidencia-se que a prática manual é o principal método utilizado para o tratamento de erros. Apesar da correção manual ser amplamente predominante, a baixa utilização de ferramentas automatizadas (11,5%) e a dependência de outros membros da equipe em alguns casos sugerem a necessidade de promover a adoção de práticas mais

eficientes e consistentes. O incentivo ao uso de ferramentas automatizadas pode potencializar a eficiência no tratamento de erros e minimizar a possibilidade de falhas humanas, contribuindo para maior qualidade e confiabilidade dos dados de pesquisa.

Conclusões

De acordo com a pesquisa bibliográfica realizada, verifica-se que existe um aumento no interesse e na aplicação de IA na biodiversidade. Os dados referentes às práticas de pesquisadores(as) vinculados(as) a programas de pós-graduação brasileiros, com conceitos CAPES 6 e 7, classificados na área de biodiversidade, indicam que as práticas de gestão de dados de pesquisa ainda apresentam um amplo potencial de evolução. Este achado é particularmente relevante, considerando que, conforme argumentado anteriormente e em alinhamento com a literatura, a área de biodiversidade já faz uso de ferramentas de IA. Entendemos que esse uso apresenta perspectivas promissoras de crescimento nos próximos anos, reforçando a importância de aprimorar a gestão de dados para maximizar os benefícios dessas tecnologias.

A despeito da articulação entre objetivos e resultados, é possível aprofundar a discussão sobre como falhas na gestão de dados afetam diretamente a acurácia dos modelos de IA. A baixa adoção de PGDs, por exemplo, pode resultar em conjuntos de dados incompletos ou mal documentados, o que compromete os processos de treinamento dos algoritmos. Do mesmo modo, a ausência de padrões de metadados dificulta a interoperabilidade entre bases de dados, limitando a capacidade dos modelos de IA de integrar informações relevantes. Assim, reforça-se a necessidade de estratégias mais concretas para promover uma governança eficaz dos dados, como a exigência de PGDs nas agências de fomento, a capacitação contínua de pesquisadores e o fortalecimento dos repositórios de dados abertos.

À guisa de conclusão, apresentamos um exemplo do impacto que a má qualidade dos dados pode gerar nos elementos de um conjunto de dados utilizado na área da biodiversidade, com base no relato de Ribeiro et al. (2022):

Dos ~31 milhões de registros incluídos nos bancos de dados originais, apenas ~2,7 milhões de registros foram considerados dados de alta qualidade após os processos de

limpeza e validação, representando uma redução de quase 91% do conjunto de dados inicial. (p. 1425).⁴

Algumas reflexões e indagações emergem a partir da constatação que dados associados a investigações na área da biodiversidade apresentam desafios relacionados à qualidade de desses insumos. Como garantir a precisão dos resultados das ferramentas de IA quando estas são alimentadas com conjuntos de dados problemáticos? Os resultados gerados a partir de microdados possuem a mesma precisão quando comparados aos resultados obtidos com conjunto de dados que passaram pelos processos adequados de gestão de dados de pesquisa?

Considerando o os pontos tratados nesse trabalhos, formulamos as seguintes hipóteses de pesquisa científica: 1 – *A não aderência às boas práticas de gestão de dados de pesquisa impacta negativamente a qualidade dos conjuntos de dados utilizados por ferramentas de inteligência artificial, comprometendo a precisão e a eficácia das análises na área da biodiversidade*; 2 - *A padronização dos processos de coleta e curadoria de dados na área da biodiversidade aumenta a confiabilidade das inferências realizadas por modelos de IA, ao reduzir inconsistências e garantir maior qualidade dos insumos utilizados nas análises.*

Conclui-se que os objetivos postos nessa investigação foram atingidos (1 - apresentar o posicionamento de pesquisadores da área da biodiversidade com relação a aspectos selecionados da gestão de dados de pesquisa e 2 - delinear a produção científica de pesquisadores na área da biodiversidade no que diz respeito ao emprego de ferramentas de inteligência artificial na sua área específica) e que foram gerados questionamentos e hipóteses para a proposição de outras pesquisas que possibilitem aprofundar questões que envolvam aa temáticas gestão de dados de pesquisa, biodiversidade e inteligência artificial.

Referências

- Aroyo, L., Lease, M., Paritosh, P., & Schaeckermann, M. (2022). Data excellence for AI: Why should you care? *Interactions*, 29(2), 66–69. <https://doi.org/10.1145/3517337>.
- Bianchini, S., Müller, M., & Pelletier, P. (2024). Drivers and barriers of AI adoption and use in scientific research. *arXiv*. <https://doi.org/10.48550/arXiv.2312.09843>.

⁴ Texto na língua original: “From ~31 million records included in the original databases, only ~2.7 million records were considered high-quality data after the data cleaning and validation processes, representing a reduction of nearly 91% of the initial dataset (p. 1425).”

- Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). (n.d.). Portal de Periódicos da CAPES. Ministério da Educação. Recuperado em 28 de janeiro de 2025, de <https://www.periodicos.capes.gov.br/>.
- Costa, P. R. S., Araújo, J. N., Dias, G. A., & Sousa, M. R. F. (2024). *Research data management: An investigation from the perspective of biodiversity, in the context of open science* (Versão 2) [Conjunto de dados]. DATAPB. <https://doi.org/10.48472/DATAPB/YDW3IM>.
- DataONE. (n.d.). *Data Management Skillbuilding Hub*. DataONE. Retrieved January 26, 2025, from <https://dataoneorg.github.io/Education/>.
- Dias, G. A., Araújo, W. J. de, Diniz, A. V. S., Córdula, F. R., & Costa, P. R. S. (2024). Qualidade dos dados no Programa Dinheiro Direto na Escola (PDDE): Superando o desafio do Garbage In, Garbage Out (GIGO). *Ciência da Informação*, 53, Artigo e6747. <https://doi.org/10.18225/ci.inf.v53i.6747>.
- Edosa, B. T., Erena, M. G., Wolteji, B. N., Werati, G. T., & Nagasa, M. D. (2024). Urban growth assessment using machine learning algorithms, GIS techniques, and its impact on biodiversity: The case of Sululta sub-city, Central Oromia, Ethiopia. *City and Environment Interactions*, 23, 100151. <https://doi.org/10.1016/j.cacint.2024.100151>.
- Digital Curation Centre. (n.d.). *Curation lifecycle model*. Digital Curation Centre. Retrieved January 26, 2025, from <https://www.dcc.ac.uk/guidance/curation-lifecycle-model>.
- Grillo, M., Schiaparelli, S., Durazzano, T., & et al. (2024). Machine learning applied to species occurrence and interactions: The missing link in biodiversity assessment and modelling of Antarctic plankton distribution. *Ecological Processes*, 13(56). <https://doi.org/10.1186/s13717-024-00532-6>.
- Nel, H., Mishra, A. K., & Schonken, F. (2024). EcoSonicML: Harnessing machine learning for biodiversity monitoring in South African wetlands. *SN Computer Science*, 5(513). <https://doi.org/10.1007/s42979-024-02859-5>.
- Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, 33(3), 301–314. <https://doi.org/10.1080/02602930701293231>.
- Orejuela-Escobar, L., Venegas-Vásconez, D., & Méndez, M. A. (2024). Opportunities of artificial intelligence in valorisation of biodiversity, biomass and bioresidues –

- Towards advanced bio-economy, circular engineering, and sustainability. *International Journal of Sustainable Energy and Environmental Research*, 13(2), 105–113. <https://doi.org/10.18488/13.v13i2.3868>.
- Ribeiro, B. R., Velazco, S. J. E., Guidoni-Martins, K. G., Tessarolo, G., Jardim, L., Bachman, S. P., & Loyola, R. (2022). bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods in Ecology and Evolution*, 13(7), 1421–1428. <https://doi.org/10.1111/2041-210x.13868>.
- Rodrigues, D. G. de A. (2021). *Elementos de ciclos de vida dos dados no percurso metodológico das teses brasileiras da área de Ciência da Informação: Um estudo diagnóstico* (Tese de doutorado, Universidade Federal da Paraíba). Universidade Federal da Paraíba. <https://repositorio.ufpb.br/jspui/handle/123456789/21993>.
- Sant’Ana, R. C. G. (2016). Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. *Informação & Informação*, 21(2), 116–142. <https://doi.org/10.5433/1981-8920.2016v21n2p116>.
- The Economist. (2017, 6 de maio). The world’s most valuable resource is no longer oil, but data. *The Economist*. Recuperado de <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>.
- Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 160018. <https://doi.org/10.1038/sdata.2016.18>
- Zhou, P., Bu, Y.-X., Fu, G.-Y., Wang, C.-S., Xu, X.-W., & Pan, X. (2024). Towards standardizing automated image analysis with artificial intelligence for biodiversity. *Frontiers in Marine Science*, 11. <https://doi.org/10.3389/fmars.2024.1349705>.