26º ENANCIB
Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
A Amazônia na sociedade global em rede: assimetrias e potencialidades informacionais, tecnológicas e ambientais
Belém, Pará, 09 a 13 de novembro de 2026
Universidade Federal do Pará

ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
ISSN 2177-3688

GT 8 – Dados, Informação e Tecnologia

PRÁTICA DE CURADORIA E GESTÃO DE DADOS CIENTÍFICOS EM REPOSITÓRIOS: UM ESTUDO DE CASO DE DADOS CLIMATOLÓGICOS NA UNIVERSIDADE FEDERAL DE GOIÁS

CURATION AND MANAGEMENT OF SCIENTIFIC DATA IN REPOSITORIES: *A CASE STUDY OF CLIMATOLOGICAL DATA AT THE UNIVERSIDADE FEDERAL DE GOIÁS*

Anne Oliveira – Universidade Federal de Goiás (UFG)
Geisa Müller de Campos Ribeiro – Universidade Federal de Goiás (UFG)
Laura Vilela Rodrigues Rezende – Universidade Federal de Goiás (UFG)
Kátia Kelvis Cassiano – Universidade Federal de Goiás (UFG)
Angel Domínguez Chovert – Universidade Federal de Goiás (UFG)

Modalidade: Trabalho Completo

Resumo: Compreendendo que a Ciência Aberta impõem às Instituições de Ciência e Tecnologia (ICTs) novos desafios relacionados à gestão de dados científicos, este trabalho tem como objetivo avaliar e validar estratégias para a integração, gestão e disponibilização de grandes volumes de dados científicos em repositórios institucionais, por meio de um estudo de caso do Centro de Excelência em Estudos, Monitoramento e Previsões Ambientais do Cerrado da Universidade Federal de Goiás (CEMPA-Cerrado/UFG), considerando os princípios FAIR e o uso da Inteligência Artificial na curadoria de dados. Adota-se a abordagem qualitativa, de natureza aplicada, configurando-se como estudo de caso. A investigação foi conduzida por uma equipe multidisciplinar de pesquisadores nas áreas de Tecnologia e Ciência da Informação, que desenvolveu uma prova de conceito para validar a integração do *workflow* científico do CEMPA-Cerrado/UFG ao Repositório de Dados de Pesquisa da UFG (RDP/UFG). Os resultados evidenciaram desafios de infraestrutura, interoperabilidade e curadoria semântica, bem como oportunidades relacionadas à visibilidade dos dados, customização de metadados, uso de IA e replicabilidade da prova de conceito. A atuação de um gestor de dados emergiu como fator relevante para a implementação efetiva dos princípios FAIR. O estudo contribui com uma modelagem conceitual replicável para outras ICTs que lidam com demandas semelhantes.
Palavras-chave: dados científicos; repositórios digitais; acesso aberto; curadoria de dados; inteligência artificial.

Abstract: Recognizing that Open Science presents Science and Technology Institutions (STIs) with new challenges related to the management of scientific data, this study aims to evaluate and validate strategies for the integration, management, and dissemination of large volumes of scientific data in institutional repositories, through a case study of the Center of Excellence in Studies, Monitoring, and Environmental Forecasting of the Cerrado at the Universidade Federal de Goiás (CEMPA-Cerrado/UFG), considering the FAIR principles and the use of Artificial Intelligence in data curation. A qualitative, applied approach is adopted, structured as a case study. The research was conducted by a multidisciplinary team of researchers in the fields of Information Technology and Information Science, who developed a proof of concept to validate the integration of CEMPA-Cerrado/UFG's scientific workflow with the UFG Research Data Repository (RDP/UFG). The results highlighted challenges related to infrastructure, interoperability, and semantic curation, as well as opportunities regarding data visibility, metadata customization, the use of AI, and the replicability of the proof of concept. The role of a data manager emerged as a key factor for the effective implementation of FAIR principles. The study contributes a replicable conceptual model for other ICTs dealing with similar demands.
Keywords: scientific data; digital repositories; open access; data curation; artificial intelligence.

1 INTRODUÇÃO

A intensificação do uso de tecnologias digitais na produção científica tem dado origem a um novo paradigma, frequentemente denominado e-Ciência, que exige uma transformação radical da prática científica tradicional por meio do uso de redes eletrônicas e infraestrutura computacional de alta velocidade (Jankowski, 2007) e distribuída. Nesse contexto, a geração, o processamento e o compartilhamento de dados científicos em larga escala se tornaram centrais no ciclo de vida da pesquisa, demandando infraestruturas tecnológicas robustas e práticas de gestão da informação alinhadas às melhores diretrizes internacionais.

O surgimento destas novas demandas impõe às Instituições de Ciência e Tecnologia (ICTs) desafios de natureza técnica, organizacional e informacional. Questões relacionadas à preservação e disponibilização dos conjuntos de dados de grande volume, bem como a adequação dos sistemas de descrição e de metadados, garantindo o atendimento às especificidades de diferentes domínios científicos e a promoção do acesso aberto aos dados produzidos com financiamento público, em consonância com os princípios da Ciência Aberta, constituem-se como desafios centrais nesse contexto. Tais desafios tornam-se ainda mais críticos quando os dados derivam de modelagem numérica, sensoriamento remoto ou monitoramento ambiental contínuo, gerando bases que frequentemente superam os limites operacionais das infraestruturas convencionais de repositórios.

É nesse quadro que se insere o presente trabalho[1], cujo objeto de investigação é a gestão de dados científicos produzidos pelo Centro de Excelência em Estudos, Monitoramento e Previsões Ambientais do Cerrado[2] da Universidade Federal de Goiás (CEMPA-Cerrado/UFG), com foco em sua integração ao repositório institucional de dados de pesquisa. O conjunto de dados utilizado é composto por variáveis meteorológicas horárias com resolução espacial de 3 km, cobrindo 30 anos de série histórica e totalizando aproximadamente 1,011 GB de volume. Sua dimensão, complexidade técnica e potencial de reutilização por diferentes áreas do conhecimento o tornam um caso exemplar para a análise dos desafios envolvidos na publicação de dados científicos de grande volume em repositórios institucionais.

Nesse cenário, observa-se uma lacuna entre a infraestrutura tecnológica e as práticas de gestão informacional disponíveis nos repositórios institucionais convencionais e as demandas específicas impostas por dados científicos de grande volume, alta complexidade técnica e especificidade disciplinar, como os produzidos por centros de pesquisa voltados à modelagem numérica e ao monitoramento ambiental. Tais conjuntos de dados, frequentemente caracterizados por múltiplas variáveis, elevada resolução espaço-temporal e formatos pouco convencionais, colocam em evidência os limites das soluções tradicionais de curadoria, armazenamento e descrição, dificultando sua adequação plena aos princípios FAIR (Findable, Accessible, Interoperable e Reusable).

Diante desse contexto, questiona-se: quais estratégias de curadoria, integração, gestão e disponibilização são necessárias para viabilizar a publicação de dados científicos de grande volume em repositórios, em conformidade com os princípios FAIR e considerando o potencial de recursos de Inteligência Artificial como suporte a esse processo? Para investigar essa questão, toma-se como caso exemplar a experiência do Centro de Excelência em Estudos, Monitoramento e Previsões Ambientais do Cerrado da Universidade Federal de Goiás (CEMPA-Cerrado/UFG) e sua integração ao Repositório de Dados de Pesquisa da UFG (RDP/UFG).

Portanto, tem-se como objetivo central identificar as estratégias de curadoria, integração, gestão e disponibilização de dados científicos de grande volume em repositórios institucionais, considerando as especificidades do caso analisado.

---

[1] O trabalho é resultado de um projeto de Pesquisa e Desenvolvimento, realizado em parceria com a Rede Nacional de Ensino e Pesquisa (RNP), no âmbito da Rede de e-Ciência.

[2] https://cempa.ufg.br/.

O estudo adota uma abordagem qualitativa, de natureza aplicada, configurando-se como um estudo de caso. Trata-se de uma pesquisa que pode contribuir com subsídios práticos e teóricos no âmbito da Ciência da Informação e para Instituições de Ciência e Tecnologia (ICTs) que enfrentam demandas similares no contexto da e-Ciência e da Ciência Aberta. Ao abordar processos de representação, armazenamento, recuperação, gestão, governança, processamento e preservação da informação em ambientes digitais, a pesquisa amplia as discussões sobre curadoria e gestão de dados científicos.

2 E-CIÊNCIA: UM NOVO PARADIGMA PARA A PESQUISA

O conceito de e-Ciência emergiu no início dos anos 2000 para designar uma forma de produção científica caracterizada pelo uso intensivo de infraestruturas computacionais distribuídas, redes de alta velocidade e ambientes colaborativos para o processamento e o compartilhamento de grandes volumes de dados. Na obra "*The Fourth Paradigm*" (2009) os autores indicam que o termo foi concebido por John Taylor, então diretor geral dos Conselhos de Pesquisa do Reino Unido, ao evidenciar o papel crescente da tecnologia da informação na pesquisa científica do século XXI, que se tornava mais colaborativa, multidisciplinar e intensiva em dados. Esse movimento resultou no que Jim Gray definiu como o quarto paradigma, ao referir-se à ciência intensiva em dados (*data-intensive science*).

Após a ciência empírica, a teórica e a computacional, emerge então uma ciência orientada pelos dados, na qual a exploração de grandes bases de dados digitais constitui o principal método investigativo (Hey; Tansley; Tolle, 2009). Esse paradigma requer infraestruturas de rede e computação de alto desempenho, colaboração interinstitucional e interdisciplinar em "colaboratórios" distribuídos (Karmous-Edwards, 2005), e práticas sistemáticas de curadoria e compartilhamento de dados ao longo de todo o ciclo de vida da pesquisa. Centros de pesquisa como o CEMPA-Cerrado/UFG são exemplos representativos desse contexto no Brasil, ao produzirem conjuntos de dados que exigem soluções específicas de gestão e disseminação que vão além das práticas convencionais de publicação científica. É importante considerar que a gestão de dados na e-Ciência não se reduz a uma questão meramente técnica, mas engloba também os âmbitos informacional e organizacional. Demanda, nesse sentido, a articulação entre infraestrutura computacional, políticas institucionais, instrumentos adequados de descrição, preservação e acesso e principalmente, a presença de um gestor de dados (*data steward)* com competências especializadas para atuar

tanto na operação técnica quanto na governança estratégica dos ativos de informação (Oladipo *et al.*, 2022).

Sob a perspectiva da Ciência Aberta, esses desafios se intensificam diante da necessidade de conciliar a abertura dos dados com os princípios de transparência, acessibilidade e reprodutibilidade da pesquisa, considerando as especificidades dos conjuntos de dados, as restrições éticas e legais e os interesses dos diferentes atores envolvidos em sua produção e utilização. A gestão de dados científicos se destaca nesse contexto, sendo fundamental para estabelecer as melhores práticas e infraestruturas que viabilizem a abertura responsável dos dados produzidos no contexto da e-Ciência.

2.1 CIÊNCIA ABERTA E GESTÃO DE DADOS CIENTÍFICOS

A Ciência Aberta representa uma ruptura social e cultural que propõe a abertura de todas as etapas da pesquisa científica (Anglada; Abadal, 2018), buscando tornar os processos e produtos da pesquisa mais transparentes, acessíveis e reutilizáveis. Tenciona, portanto, as estruturas acadêmicas do modelo tradicional de pesquisa, visando ao estabelecimento de uma cultura de abertura, conforme recomendações da Unesco (2022), OCDE e Comissão Europeia.

Dentre suas diversas vertentes, há uma intensa mobilização em prol da abertura de dados de pesquisa científicos, considerados como recursos intelectuais que constituem parte importante da produção científica acumulada pela comunidade de pesquisa. Para a Organisation for Economic Co-operation and Development (OECD), (2007, tradução nossa)[3], dados de pesquisa são definidos como registros factuais usados como fontes primárias para pesquisa científica e aceitos como necessários para validar os resultados.

Para Sayão e Sales (2014), o valor dos dados está relacionado à sua possibilidade de uso e reinterpretação em áreas e contextos distintos daqueles em que foram originalmente gerados. Assim, quando tratados, organizados e preservados adequadamente, tornam-se fontes valiosas de informação. Os repositórios de dados emergem como infraestruturas centrais para operacionalizar esse processo, em consonância com os princípios FAIR (*Findable, Accessible, Interoperable e Reusable*).

[3] Tradução original: "[..] re defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings".

Cabe destacar que, a implementação dos princípios FAIR em contextos de e-Ciência enfrenta desafios técnicos e semânticos que se intensificam diante do crescimento exponencial dos dados científicos. Hey, Tansley e Tolle (2009) demonstram que a ciência contemporânea é marcada por um "dilúvio de dados", no qual instrumentos, sensores e simulações produzem continuamente um volume massivo de dados, frequentemente em escalas de terabytes e petabytes, impondo limitações de armazenamento, processamento e transferência de dados. Nesse contexto, As práticas tradicionais de gerenciamento tornam-se insuficientes, uma vez que conjuntos de dados dessa magnitude exigem arquiteturas distribuídas, processamento paralelo e infraestruturas capazes de aproximar a computação dos próprios dados. Além do volume, a heterogeneidade dos formatos e a ausência de padronização semântica dificultam a interoperabilidade e a reutilização dos dados, uma vez que muitos conjuntos permanecem organizados sem esquemas ou metadados suficientemente descritivos. Embora formatos como HDF e NetCDF incorporem mecanismos de descrição estrutural, ainda faltam ferramentas e padrões capazes de representar adequadamente a complexidade semântica dos dados científicos e possibilitar sua interpretação interoperável entre diferentes áreas do conhecimento.

No plano semântico, esquemas de metadados generalistas frequentemente não capturam a riqueza descritiva de dados com múltiplas variáveis, diferentes escalas espaço-temporais e metodologias específicas. Essa lacuna é agravada pela ausência de profissionais especializados para mediar o processo junto às equipes de pesquisa. Os gestores de dados, tradução adotada neste estudo, internacionalmente denominados como *data stewards*, atuam essencialmente como facilitadores e garantem que o ciclo de vida dos dados seja gerido de forma eficiente por meio da comunicação e da competência técnica (Gruber *et al*., 2021). Essa atuação entre os domínios científico e informacional, contribui para a tradução de práticas e vocabulários disciplinares para os sistemas de metadados. No plano organizacional, a ausência de *workflows* consolidados resulta em compartilhamento reativo de dados, sem descrição sistemática e rastreabilidade, uma vez que esses elementos são considerados essenciais para gerenciar a complexidade crescente dos dados em aplicações de larga escala e contribuem para a definição dos metadados, essenciais para reutilização futura (Simmhan; Plale; Gannon, 2005). Diante dessas problemáticas, a IA emerge como recurso complementar relevante para auxiliar na curadoria e na consulta em linguagem natural, embora sua

integração com repositórios existentes ainda represente desafios operacionais que precisam ser investigados em contextos institucionais concretos.

3 PROCEDIMENTOS METODOLÓGICOS

A presente investigação adota uma abordagem qualitativa, de natureza aplicada, configurando-se como um estudo de caso com implementação de prova de conceito que visa identificar estratégias de curadoria, integração, gestão e disponibilização de dados científicos de grande volume em repositórios institucionais, por meio do estudo de caso do CEMPA-Cerrado/UFG.

O projeto foi conduzido por uma equipe multidisciplinar composta por bibliotecárias, cientistas da computação e um pesquisador da área de meteorologia, todos vinculados à UFG. Trata-se de um estudo originado como demanda da Rede Nacional de Ensino e Pesquisa (RNP), visando entender demandas de Ciência Aberta no escopo da Rede e-Ciência[4]. A equipe foi organizada em dois grupos de trabalho (GT1 e GT2), que atuaram de forma articulada ao longo das etapas do projeto (Figura 1).

Figura 1 - Atividades desenvolvidas pelos GTs



Fonte: elaboração própria (2026).

Ao longo do estudo, os principais procedimentos investigativos foram pautados na observação participante das atividades de implementação, especialmente no mapeamento do *workflow* científico específico, com registro sistemático das decisões técnicas e dos

[4] https://www.rnp.br/projetos/rede-de-e-ciencia/.

desafios enfrentados; e na análise documental dos registros produzidos ao longo do projeto, incluindo relatórios de atividades, logs de instalação e configuração e *templates* de metadados desenvolvidos.

As análises foram orientadas pelos princípios FAIR como referencial avaliativo, permitindo examinar em que medida as soluções implementadas poderiam contribuir para tornar os dados localizáveis, acessíveis, interoperáveis e reutilizáveis. Os desafios e oportunidades identificados ao longo da implementação foram sistematizados, buscando articular as evidências empíricas do caso com as problemáticas mais amplas da gestão de dados científicos em contextos de e-Ciência.

4 ESTUDO DE CASO

Esta seção apresenta o estudo de caso desenvolvido, descrevendo o contexto institucional, o conjunto de dados selecionado, o *workflow* científico do CEMPA-Cerrado/UFG, a modelagem conceitual de integração ao repositório, o processamento e enriquecimento dos dados, o uso de recursos de IA na gestão dos dados, a implementação da coleção no RDP/UFG e a experiência de compartilhamento com pesquisadores e estudantes de instituições parceiras.

4.1 CONTEXTO INSTITUCIONAL: O CEMPA-CERRADO/UFG E O RDP/UFG

O CEMPA-Cerrado/UFG é um centro de pesquisa voltado à geração de previsões meteorológicas de alta precisão para a região Centro-Oeste, além de cenários climáticos e modelos agroclimáticos. Para tanto, o Centro emprega modelagem numérica avançada e dados de satélites, produzindo conjuntos de dados atmosféricos de elevada resolução espacial e temporal. Suas atividades estão associadas ao Laboratório Multiusuário de Computação de Alto Desempenho da UFG (LaMCAD/UFG), que provê a infraestrutura computacional necessária para o processamento das simulações e o armazenamento dos dados gerados.

A disponibilização dos dados produzidos pelo CEMPA-Cerrado/UFG, até o início deste projeto, ocorria de forma reativa. Isso significa que os dados eram compartilhados quando havia manifestação explícita de necessidade por parte de pesquisadores ou programas institucionais, sem um fluxo de trabalho sistematizado que assegurasse a geração dos dados com descrição qualificada, identificação persistente e controle de acesso. Esse modelo de compartilhamento informal limitava a visibilidade dos dados, dificultando sua citação formal

em publicações científicas e ~~não garante~~ inviabilizando a rastreabilidade sobre seu uso. A ausência de uma equipe dedicada ao atendimento de solicitações externas e ao processamento personalizado dos dados reforçava essa condição, cabendo geralmente ao solicitante realizar o pós-processamento dos arquivos, que eram gerados e disponibilizados em formatos específicos da área, para extrair as informações de interesse.

O Repositório de Dados de Pesquisa da UFG (RDP/UFG)[5], foi implantado em 2024, constituindo a infraestrutura institucional disponível para o armazenamento e compartilhamento de dados científicos produzidos pela universidade. O RDP utiliza a plataforma Dataverse[6], mundialmente reconhecida por sua infraestrutura adequada à gestão de conjuntos de dados e ~~que~~ pelo suporte~~a~~ a padrões de metadados específicos de diferentes domínios ~~específicos~~. À época do projeto, a instalação do RDP/UFG sediada no LaMCAD dispunha de 500 GB de armazenamento e 16GB de memória RAM, configuração que, como se verá adiante, revelou-se insuficiente para a ingestão do conjunto de dados de maior volume pelo Dataverse, produzido pelo CEMPA-Cerrado/UFG.

4.2 O CONJUNTO DE DADOS, O *WORKFLOW* CIENTÍFICO E A INTEGRAÇÃO AO REPOSITÓRIO

O Conjunto de Dados Atmosféricos ERA5-MONAN-IA (EMAI-Goiás v1.0) é uma base meteorológica de alta resolução composta por informações horárias de duas variáveis primárias (temperatura a 2 m e vento a 10 m) e sete variáveis derivadas (temperaturas máxima, mínima e média, velocidade total e componentes do vento, e direção do vento), com resolução espacial de 3 km, cobrindo o estado de Goiás e áreas adjacentes no período 1994–2024 (30 anos, conforme diretrizes da OMM para normais climatológicas). O conjunto apresenta um volume total de aproximadamente 1.011 GB. A partir desse conjunto primário, foi gerado o EMAI-Goiás Climatologia (v1.0), subconjunto derivado com valores médios em escalas mensal, trimestral e anual por ponto de grade de 3 km × 3 km, desenvolvido em resposta a demandas recorrentes recebidas pelo Centro.

O EMAI-Goiás (v1.0) resultou da integração de dados do MONAN (Model for Ocean–LaNd–Atmosphere Prediction), modelo numérico brasileiro de simulação acoplada, e do ERA5, reanálise global do ECMWF com séries horárias de 1940 ao presente (Hersbach et al., 2020).

[5] https://dadosdepesquisa.ufg.br/.

[6] Desenvolvido pelo Institute for Quantitative Social Science (IQSS) de Harvard, o Dataverse é uma aplicação web de código aberto para compartilhar, preservar, citar, explorar e analisar dados de pesquisa.

Os dados do ERA5 foram usados como condições de contorno para o MONAN, executado com horizonte de previsão de 120 horas e saídas horárias em grade de 3 km × 3 km. Os pares de dados nas resoluções de 27 km (ERA5) e 3 km (MONAN) serviram de insumo para treinar um modelo de IA baseado em *Conditional Generative Adversarial Network* (cGAN), que aprendeu relações não lineares entre campos meteorológicos de baixa e alta resolução e, após o treinamento, gerou automaticamente saídas em 3 km a partir do ERA5, eliminando o alto custo computacional de novas integrações. Aplicado ao período 1994–2024, produziu a série histórica do EMAI-Goiás (v1.0).

A partir do mapeamento desse *workflow* científico, a equipe do projeto construiu uma modelagem conceitual de integração do *workflow* do Centro ao RDP/UFG, com o objetivo de consolidar a entrega de conjuntos de dados em conformidade com os princípios FAIR. A modelagem revelou a importância central do processo de curadoria de dados, uma vez que os dados produzidos pelo CEMPA-Cerrado/UFG apresentam especificidades técnicas, metodológicas e disciplinares que demandam tratamento diferenciado e apoio contínuo aos pesquisadores nas etapas de organização, documentação e disponibilização.

O EMAI-Goiás Climatologia (v1.0), originalmente em formato NetCDF (250 variáveis, 435 latitudes × 405 longitudes, aproximadamente 44 milhões de registros), foi transformado em formatos tabulares (CSV e Parquet) por *scripts* em *Python*, ampliando a acessibilidade para usuários sem familiaridade com arquivos matriciais multidimensionais. Os dados foram enriquecidos com bases geoespaciais do IBGE, permitindo a geração de subconjuntos temáticos como as normais climatológicas para os 246 municípios goianos.

O estudo também contemplou o início de testes com ferramentas de IA que foram integradas à instância de desenvolvimento do Dataverse: o *Data Explorer*[7] (visualização interativa dos dados depositados e estatísticas descritivas), o *Ask the Data*[8] (consultas aos dados tabulares em linguagem natural) e o *TurboCurator*[9] (curadoria assistida e automatização de metadados). Em relação à aplicação de recursos integrados ao Dataverse, utilizou-se também a aplicação da ferramenta de vocabulário externo, tendo sido

[7] https://github.com/scholarsportal/Dataverse-Data-Explorer.
[8] https://github.com/IQSS/askdataverse/tree/main/askthedata.
[9] https://turbocurator.icpsr.umich.edu/tc/adminabout/.

implementado o GEMET (General Multilingual Environmental Thesaurus)[10] para padronização semântica dos metadados, alinhada ao princípio FAIR de interoperabilidade.

4.2.1 A coleção do CEMPA-CERRADO/UFG no RDP/UFG

Foram criados dois *datasets* na coleção. O primeiro, denominado “Normais climatológicas para Goiás, EMAI-Goiás (v1.0)”, recebeu três arquivos: o conjunto de dados EMAI-Goiás Climatologia (v1.0) em formato NetCDF; o arquivo "EmainormaisTmax.tab", gerado para o atendimento de demanda específica com valores médios mensais de temperatura máxima para os 246 municípios goianos; e o arquivo “variaveis_climaticas_meta_urbanas_periodo_anual.tab”, produzido pelo GT2 a partir da transformação e enriquecimento do arquivo original com dados geoespaciais do IBGE. O segundo *dataset*, "Conjunto de Dados Atmosféricos ERA5-MONAN-IA (EMAI-Goiás v1.0)", foi concebido como catálogo de dados, dada a inviabilidade de ingestão do conjunto primário (aproximadamente 1 TB) em repositório com limite de 500 GB. O *dataset* redireciona o usuário para a URL de disponibilização original, assegurando a atribuição de DOI, a citação formal e a visibilidade dos dados.

Para aprimorar o processo de submissão e curadoria, foi desenvolvido no Dataverse, um *template* de metadados específico para dados meteorológicos, customizável e extensível a outras áreas temáticas. Foi também habilitada a função “*Submit for review*”, que introduz uma etapa intermediária no fluxo de publicação, entre a criação do *dataset* pelo pesquisador e sua publicação final no repositório, permitindo verificação institucional e curadoria antes da disponibilização pública.

4.3 A EXPERIÊNCIA DE COMPARTILHAMENTO ENTRE PARES

Após a criação dos *datasets*, a coleção foi apresentada a pesquisadores e estudantes da área, vinculados a instituições parceiras, especialmente à Universidade Federal do Pará (UFPA), ao Instituto Nacional de Pesquisas Espaciais (INPE) e à própria UFG. As interações foram organizadas em três grupos de análise, permitindo uma avaliação comparada das percepções sobre o repositório e os dados disponibilizados.

Pesquisadores de meteorologia consideraram os metadados adequados e completos, destacando a facilidade de disponibilização pública, a organização estruturada dos conjuntos

[10] https://www.eionet.europa.eu/gemet/en/about/

e a possibilidade de citação formal. Um pesquisador ressaltou que informações de alta relevância científica e social do CEMPA-Cerrado/UFG frequentemente permanecem restritas a públicos especializados e que o repositório pode ampliar sua disseminação; outro sugeriu incluir informações sobre estudos que utilizaram as bases, evidenciando o impacto científico dos conjuntos. Pesquisadores de áreas afins valorizaram o controle de acesso a downloads e o versionamento, reconhecidos como mecanismos de transparência e rastreabilidade. Estudantes de pós-graduação em Meteorologia da UFPA avaliaram que a disponibilização organizada e documentada dos dados amplia as possibilidades de análise em iniciação científica e trabalhos de pós-graduação. De modo geral, a recepção foi positiva, e as interações evidenciaram demandas de aprimoramento nos padrões de metadados da área de meteorologia, que podem ser contemplados por meio dos recursos de customização do Dataverse.

5 DISCUSSÃO

A implementação da prova de conceito evidenciou desafios articulados e oportunidades estratégicas para a consolidação da gestão de dados científicos em ICTs.

5.1 DESAFIOS

O caso do CEMPA-Cerrado/UFG evidencia que a limitação de infraestrutura permanece como um dos principais obstáculos para a institucionalização da gestão de dados científicos de grande volume em ICTs brasileiras. O EMAI-Goiás (v1.0), com aproximadamente 1.011 GB e exigência mínima estimada de 1,5 TB de armazenamento e mínimo de 32 GB de RAM para ingestão adequada, excede a capacidade disponível no RDP/UFG, restrita a 500 GB e 16 GB de RAM. A inviabilidade do depósito integral e a adoção da solução de catálogo revelam que os desafios associados aos dados científicos massivos ultrapassam a simples disponibilização em repositórios, exigindo arranjos alternativos. Os resultados também indicam que a interoperabilidade entre sistemas permanece um desafio relevante. A integração do *TurboCurator* ao Dataverse demandou migração para HTTPS e, mesmo após essa adequação, persistiram incompatibilidades relacionadas à interpretação das rotas pelo serviço externo. Esse resultado sugere que a interoperabilidade, frequentemente tratada como requisito técnico abstrato, depende de compatibilização contínua entre plataformas, versões e serviços

complementares, revelando a fragilidade de soluções apoiadas em múltiplas dependências tecnológicas.

No plano informacional, o estudo demonstra que a representação e a descrição de dados científicos complexos não se resolvem pela adoção automática de esquemas padronizados de metadados. Dados meteorológicos produzidos por modelagem numérica envolvem múltiplas variáveis, diferentes escalas espaço-temporais e procedimentos metodológicos cuja inteligibilidade depende de contextualização especializada. O formato NetCDF ilustra essa condição, exigindo documentação complementar e estratégias de enriquecimento informacional capazes de ampliar sua interoperabilidade semântica e seu potencial de reutilização interdisciplinar.

Por fim, o caso evidencia que os limites da gestão de dados científicos não são apenas tecnológicos ou documentários, mas também organizacionais. A ausência de gestores de dados atuando ao longo do ciclo de vida dos dados emergiu como fator crítico para a qualidade descritiva dos metadados, para a continuidade das práticas de curadoria e a sustentabilidade institucional do depósito em repositórios. Mais do que uma dificuldade localizada, esse achado sugere uma lacuna estrutural nas ICTs brasileiras, nas quais a institucionalização da gestão de dados ainda depende de iniciativas pontuais e equipes de projeto, em vez de competências e funções formalmente incorporadas às políticas e infraestruturas de pesquisa.

5.2 OPORTUNIDADES

A disponibilização estruturada dos conjuntos de dados com DOI desloca dados antes compartilhados informalmente para a condição de produtos científicos formalmente identificáveis, citáveis e rastreáveis, ampliando o reconhecimento da autoria, a mensuração do impacto e a visibilidade institucional. A flexibilidade do Dataverse para criar *templates* de metadados específicos ao domínio meteorológico e integrar vocabulários controlados, como o GEMET, evidencia que a interoperabilidade FAIR não depende de padronização rígida, mas da capacidade de adaptação das infraestruturas às especificidades disciplinares. Nesse sentido, o *template* desenvolvido apresenta potencial de replicação em outras ICTs com perfis temáticos semelhantes.

As ferramentas de IA testadas indicaram potencial para reduzir barreiras de acesso e ampliar a mediação informacional, seja por visualização interativa e estatísticas automáticas,

seja por consultas em linguagem natural. Embora as limitações operacionais identificadas demandem investigação adicional, os resultados sugerem que a integração entre *workflows* científicos e repositórios constitui uma estratégia promissora para ampliar a acessibilidade e a usabilidade dos dados. O controle de versões nativo do Dataverse reforça a transparência sobre atualizações e a confiabilidade dos conjuntos disponibilizados, aspecto particularmente relevante para bases dinâmicas e em contínua atualização.

Figura 2 - Atuação do gestor de dados



Fonte: elaboração própria (2026).

A atuação de uma equipe multidisciplinar revelou a importância da atuação do gestor de dados (*data steward)* em todo o processo de gestão e compartilhamento dos dados. A gestão de dados é uma estratégia crucial e um investimento sustentável que permite que a ciência aberta evolua de um conceito teórico para uma prática operacional robusta e confiável (Rosenbaum, 2010). Essa atuação pode ser compreendida a partir de quatro dimensões interdependentes, executadas no âmbito desse projeto, cuja articulação é fundamental para a implementação efetiva dos princípios FAIR (Figura 2).

6 CONSIDERAÇÕES FINAIS

Este trabalho identificou as estratégias de integração, gestão e disponibilização de dados científicos de grande volume em repositórios institucionais, considerando os princípios FAIR e o uso de recursos de IA na curadoria e enriquecimento dos dados. Os resultados confirmaram que repositórios convencionais enfrentam limitações significativas diante das demandas da e-Ciência. Essa demanda é acentuada pela crescente mobilização, promovida

pelo movimento de Ciência Aberta, acerca das exigências para abertura e compartilhamento dos dados de pesquisa.

A solução de catálogo adotada, com a disponibilização de um *dataset* que redireciona para a URL de armazenamento original, demonstrou ser uma alternativa viável para assegurar a identificação e a visibilidade dos dados sem ingestão integral na plataforma, e pode servir como modelo para situações análogas.

Do ponto de vista semântico e informacional, a conformidade com os princípios FAIR não se obtém apenas pela adoção de plataforma tecnicamente adequada, mas requer competências especializadas de curadoria. A figura do gestor de dados emerge como elemento estruturante para a qualidade, reutilização e sustentabilidade dos dados das ICTs. Sua ausência foi o principal fator limitante identificado, e sua institucionalização representa agenda prioritária para a consolidação da ciência aberta no contexto das ICTs brasileiras. Do ponto de vista organizacional, a integração do *workflow* científico ao repositório de dados, com etapas sistemáticas de preparação e depósito, representa uma mudança de paradigma que demanda infraestrutura, ferramentas e políticas institucionais que reconheçam a produção de dados como atividade científica legítima.

O uso de IA na curadoria revelou-se promissor, mas ainda permeado por desafios de integração, escalabilidade e sustentabilidade que constituem uma agenda de pesquisa relevante para a Ciência da Informação e a Computação. A experiência de compartilhamento com os pares inaugurou um ciclo de retroalimentação que pode contribuir para a melhoria contínua da qualidade descritiva e da usabilidade das bases disponibilizadas. Como limitação, embora os resultados se refiram a um único caso institucional, estudos futuros que ampliem a análise para outras ICTs poderão validar e refinar o modelo desenvolvido. O caso do CEMPA-Cerrado/UFG oferece um exemplo concreto e documentado de como ICTs podem avançar na gestão de dados científicos orientada pelos princípios da Ciência Aberta, com contribuições que apontam para um modelo replicável.

REFERÊNCIAS

ANGLADA, Lluís; ABADAL, Ernest. ¿Qué es la ciencia abierta? **Anuario ThinkEPI**, Barcelona, v. 12, p. 292-298, 2018. Disponível em: https://doi.org/10.3145/thinkepi.2018.43. Acesso em: 24 maio 2026.

GRUBER, Alexander; SCHRANZHOFER, Hermann; KNOPPER, Sabrina; STRYECK, Sarah *et al*. Kompetenzen von Data Stewards an österreichischen Universitäten. **Mitteilungen der**

Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare, Wien, v. 74, n. 1, 2021. Disponível em: https://doi.org/10.31263/voebm.v74i1.6255. Acesso em: 27 maio 2026.

HERSBACH, Hans; BELL, Bill; BERRISFORD, Paul; HIRAHARA, Shoji *et al*. The ERA5 global reanalysis. **Quarterly Journal of the Royal Meteorological Society**, Hoboken, v. 146, n. 730, p. 1999-2049, 2020. Disponível em: https://doi.org/10.1002/qj.3803. Acesso em: 10 abr. 2026.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (ed.). **The fourth paradigm**: data-intensive scientific discovery. Redmond: Microsoft Research, 2009.

JANKOWSKI, Nicholas W. Exploring e-science: an introduction. **Journal of Computer-Mediated Communication**, Oxford, v. 12, n. 2, p. 549-562, 2007. Disponível em: https://doi.org/10.1111/j.1083-6101.2007.00337.x. Acesso em: 24 maio 2026.

KARMOUS-EDWARDS, Gigi. Global e-science collaboration. **Computing in Science & Engineering**, [*s. l.*], v. 7, n. 2, p. 67-74, 2005. Disponível em: https://doi.org/10.1109/MCSE.2005.32. Acesso em: 23 maio 2026.

OLADIPO, Francisca; FOLORUNSO, Sakinat; OGUNDEPO, Ezekiel; OSIGWE, Obinna *et al*. Curriculum development for FAIR data stewardship. **Data Intelligence**, Cambridge, v. 4, n. 4, p. 991-1012, 2022. Disponível em: https://www.sciengine.com/DI/doi/10.1162/dint_a_00183. Acesso em: 29 maio 2026.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **OECD principles and guidelines for access to research data from public funding**. Paris: OECD, 2007. Disponível em: https://doi.org/10.1787/9789264034020-en-fr. Acesso em: 24 maio 2026.

ROSENBAUM, Sara. Data governance and stewardship: designing data stewardship entities and advancing data access. **Health Services Research**, Hoboken, v. 45, n. 5, p. 1442-1455, 2010. Disponível em: https://doi.org/10.1111/j.1475-6773.2010.01140.x. Acesso em: 29 maio 2026.

SAYÃO, Luis Fernando; SALES, Luana Farias. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **RECIIS**: Revista Eletrônica de Comunicação, Informação e Inovação em Saúde, Rio de Janeiro, v. 8, n. 2, 2014. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/611. Acesso em: 29 maio 2026.

SIMMHAN, Yogesh L.; PLALE, Beth; GANNON, Dennis. A survey of data provenance in e-science. **ACM SIGMOD Record**, New York, v. 34, n. 3, p. 31-36, 2005. Disponível em: https://doi.org/10.1145/1084805.1084812. Acesso em: 21 maio 2026.

AGRADECIMENTOS

Agradecemos à Rede Nacional de Ensino e Pesquisa (RNP) pelo financiamento desta pesquisa.