25°enancib
Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação
Informação, Decolonialidade e Direitos Difusos para o Desenvolvimento Sustentável
3 a 7 de novembro de 2025 - Rio de Janeiro

XXV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - XXV ENANCIB

GT-8 – Dados, Informação e Tecnologia

METODOLOGIA DE PRESERVAÇÃO DIGITAL APLICADA À LITOTECA DO CECO/UFRGS

DIGITAL PRESERVATION METHODOLOGY APPLIED TO THE CECO/UFRGS LITOTECA

Bryan Nicollas Soares Costa - Universidade Federal do Rio Grande do Sul (UFRGS)
Fabiano Couto Corrêa da Silva - Universidade Federal do Rio Grande do Sul (UFRGS)
Ana Paula Sehn - Universidade Federal do Rio Grande do Sul (UFRGS)
Luciana Monteiro-Krebs - Universidade Federal do Rio Grande do Sul (UFRGS)

Modalidade: Trabalho Completo

Resumo: apresenta metodologia de preservação digital do acervo composto por testemunhos geológicos sobre a costa e o fundo marinho do sul do Brasil, coletado e salvaguardado na Litoteca do Centro de Estudos de Geologia Costeira e Oceânica, da Universidade Federal do Rio Grande do Sul. Objetiva desenvolver uma experiência-piloto de curadoria digital e abertura de dados científicos, alinhada aos princípios da Ciência Aberta, promovendo a reprodutibilidade e ampliando o acesso ao conhecimento gerado na Litoteca. Os procedimentos metodológicos envolvem cinco etapas principais: diagnóstico técnico, digitalização e extração de dados, consolidação e organização dos dados com o uso de inteligência artificial, produção científica, preservação e gestão digital. Resulta uma contribuição importante para o campo da curadoria digital de dados, documentos e informação aplicada às Geociências e às Ciências Naturais. O trabalho interdisciplinar entre especialistas em Geociências e Ciência da Informação permitiu a construção de um fluxo de trabalho potente, escalável e replicável em outras instituições com acervo similar. Ao transformar documentos físicos em dados estruturados, abertos e reutilizáveis, o projeto, além de salvaguardar o patrimônio científico da universidade, reforça os compromissos institucionais do CECO/IGEO e do Datalab/PPGCIN com a Ciência Aberta, a Reprodutibilidade e a sustentabilidade do conhecimento científico. Ademais, os dados e informações do acervo composto por testemunhos geológicos sobre a costa e o fundo marinho do sul do Brasil do CECO, podem subsidiar e instigar novas pesquisas no âmbito das geociências, oceanografia e áreas interdisciplinares.

Palavras-chave: preservação digital; lithoteca; testemunhos geológicos, inteligência artificial.

Abstract: it presents a methodology for the digital preservation of the collection of geological testimonies from the coast and seabed of southern Brazil, collected and safeguarded in the Lithoteca of the Center for Coastal and Oceanic Geology Studies at the Federal University of Rio Grande do Sul. The aim is to develop a pilot experiment in digital curation and the opening up of scientific data, in line with the principles of Open Science, promoting reproducibility and expanding access to the knowledge generated in the litho-library. The methodological procedures cover five main stages: technical diagnosis, digitization and data extraction, data consolidation and organization using artificial intelligence, scientific production, preservation and digital management. The result is an important contribution to the field of digital curation of data, documents and information applied to the Geosciences and Natural Sciences. The interdisciplinary work between specialists in Geosciences and

Information Science has made it possible to build a powerful workflow that is scalable and replicable in other institutions with similar collections. By transforming physical documents into structured, open and reusable data, the project, in addition to safeguarding the university's scientific heritage, reinforces the institutional commitments of CECO/IGEO and Datalab/PPGCIN to Open Science, Reproducibility and the sustainability of scientific knowledge. In addition, the data and information in CECO's collection of geological testimonies from the coast and seabed of southern Brazil can support and instigate new research in the geosciences, oceanography and interdisciplinary areas.

Keywords: digital preservation; litho library; geological evidence, artificial intelligence.

1 INTRODUÇÃO

Nas geociências, pesquisadores e instituições geram, gerenciam e preservam vastas quantidades de dados e materiais diversos, desde amostras de rochas e fósseis até imagens de satélite e modelos climáticos. A interseção entre preservação digital, coleções científicas, gestão do patrimônio e práticas de ciência aberta constitui uma estrutura conceitual importante para garantir que esses recursos permaneçam acessíveis e utilizáveis para as futuras gerações de cientistas. A preservação digital abrange os métodos e tecnologias utilizados para garantir o acesso a longo prazo à informação digital, o que se tornou cada vez mais importante à medida que a coleta e a análise de dados geocientíficos migraram para o mundo digital.

As coleções científicas e o patrimônio em geociências incluem espécimes físicos, documentos históricos, mapas e conjuntos de dados que documentam a história e os processos da Terra. Essas coleções representam registros insubstituíveis que embasam pesquisas, educação e decisões políticas sobre o planeta. As práticas de ciência aberta promovem, nesta direção, transparência, reprodutibilidade e acessibilidade dos resultados da pesquisa, permitindo maior colaboração e inovação dentro da comunidade de geociências. Juntos, esses domínios interconectados criam uma abordagem abrangente para a gestão do conhecimento e do patrimônio geocientífico que abrange tanto objetos físicos quanto informações digitais, conectando observações passadas com pesquisas presentes e descobertas futuras. Essas técnicas visam tornar os dados públicos ou, ao menos, mais acessíveis e fáceis de analisar, uma vez que podem fornecer metadados valiosos sobre informações que anteriormente estavam armazenadas em arquivos internos.

A Litoteca do Centro de Estudos de Geologia Costeira e Oceânica (CECO), da Universidade Federal do Rio Grande do Sul (UFRGS) é um acervo singular que reúne testemunhos geológicos coletados ao longo de décadas de pesquisa sobre a costa e o fundo

marinho do sul do Brasil. Seu conteúdo subsidia estudos em áreas como sedimentologia, paleoclima e dinâmica costeira, sendo um patrimônio científico de valor histórico e estratégico para a universidade (UFRGS, 2025). Apesar de lidar com materiais de valor inestimável cientificamente, atualmente carece de uma infraestrutura e fluxo sistemáticos de preservação digital. Em função disso, os dados coletados acabam ficando restritos a localização geográfica e equipe, além de serem vulneráveis às ações do tempo. Neste contexto, apresenta-se uma abordagem metodológica integrada para a preservação digital da Litoteca do CECO/UFRGS, articulando saberes da Ciência da Informação (CI) e das Geociências. Desenvolver e validar uma metodologia reprodutível de preservação digital para litotecas universitárias, integrando OCR+PLN, curadoria de metadados e depósito no Tainacan, com monitoramento de qualidade (OCR e NER), plano de preservação (fixidez, redundância, formatos) e publicação de data papers como mecanismo de disseminação e citação dos dados.. Essa proposta está inserida no esforço institucional de valorização do patrimônio científico e da adoção de boas práticas de gestão de dados de pesquisa. Este trabalho diferencia-se por aplicar práticas de preservação digital, curadoria de dados e inteligência artificial a um acervo geocientífico físico relevante. Trata-se de uma proposta inovadora no contexto da UFRGS, com potencial de replicação em outras litotecas e centros de pesquisa. Ao ampliar o acesso ao acervo da Litoteca, o projeto contribui para a valorização da memória científica da universidade e para o fortalecimento das políticas de ciência aberta no Brasil. A contribuição reside no encadeamento operacional entre digitalização, extração assistida por PLN, curadoria com métricas e depósito preservável no Tainacan, configurando um caminho replicável para litotecas acadêmicas.

2 PRESERVAÇÃO DIGITAL EM GEOCIÊNCIAS

A preservação digital nas geociências envolve desafios específicos, dado o volume e a diversidade dos dados produzidos. Mais do que manter arquivos acessíveis, trata-se de garantir sua inteligibilidade no futuro, mesmo diante da rápida obsolescência tecnológica (Conway, 2010; Day, 2008; Hedstrom, 1998). Isso exige estratégias que assegurem o valor científico contínuo dos dados, permitindo sua reinterpretação em novos contextos e ampliando sua utilidade para as próximas gerações (Rieger, 2018).

A preservação digital é crucial nas geociências, pois muitos dados resultam de coletas únicas e irreproduzíveis em campo. Sem documentação adequada, torna-se inviável o reuso

científico, comprometendo a continuidade do conhecimento (Wallis; Rolando; Borgman, 2013). Essa limitação tem sido considerada um dos grandes desafios da era digital, demandando ações constantes para assegurar o acesso futuro a esses registros (Day, 2008).

A preservação de dados geocientíficos deve considerar a complexidade e o contexto da produção científica. Como destaca Lavoie (2014), não basta conservar arquivos digitais; é necessário preservar as relações e significados dos dados. Yakel (2007) define a curadoria digital como um processo contínuo que assegura autenticidade, contexto e usabilidade ao longo do ciclo de vida dos dados. Essa abordagem é essencial para compreender como os resultados foram produzidos e garantir sua reutilização em novos contextos.

Modelos digitais tridimensionais de afloramentos exemplificam os benefícios da preservação em geociências, ao gerar dados 3D reutilizáveis para arquivamento e interpretação de longo prazo (Burnham *et al*., 2022). Essas representações digitais ampliam o acesso global a estruturas geológicas e promovem os princípios FAIR, maximizando o valor científico dos dados (Wilkinson *et al*., 2016). Tal abordagem também reforça os fundamentos da ciência aberta, ao favorecer o compartilhamento e o reuso colaborativo dos dados geocientíficos (Ma *et al*., 2017).

A pesquisa em geociências computacionais ainda enfrenta desafios quanto à definição do que preservar, às diferenças entre projetos e às barreiras culturais e técnicas (Mullendore *et al*., 2021). Soluções como DFDL e Defuddle oferecem recursos promissores para preservação de dados com menor dependência de software, mas sua efetividade depende de documentação robusta e metadados bem estruturados, conforme destacam Mons *et al*. (2017).

As coleções científicas em geociências integram registros físicos e digitais que preservam informações valiosas sobre os processos da Terra (Wildman *et al*., 2022). Essa longevidade reforça a importância de curadoria contínua, especialmente em áreas como a paleontologia, que depende dessas coleções para compreender as mudanças climáticas ao longo do tempo (Hedstrom, 1998; Yakel, 2007).

Investimentos relevantes têm sido realizados na digitalização de coleções científicas, com destaque para plataformas como o *Global Biodiversity Information Facility* (GBIF) e o *Integrated Digitized Biocollections* (iDigBio). No entanto, essas iniciativas concentram-se sobretudo em dados contemporâneos de biodiversidade, o que dificulta a integração de acervos paleontológicos. Entre os principais desafios estão a ausência de padronizações

consolidadas, a diversidade de formatos históricos e a escassez de metadados estruturados, limitando a interoperabilidade desses dados em ambientes digitais (Little *et al*., 2023).

Para aumentar a utilidade das coleções científicas, é essencial que os metadados associados sejam bem estruturados, acessíveis e padronizados. Pesquisas e debates envolvendo gestores de coleções e pesquisadores destacam a necessidade de práticas unificadas de dados para garantir que os metadados sejam localizáveis, interoperáveis e reutilizáveis (Leonelli, 2016). No entanto, ainda há uma lacuna na adoção generalizada desses padrões entre as instituições, evidenciando a importância de portais de dados mais integrados, que atendam às múltiplas áreas das geociências. Além disso, a implementação conjunta dos princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) e CARE (*Collective Benefit, Authority to Control, Responsibility, Ethics*) oferece uma estrutura robusta para a gestão dessas coleções, assegurando que beneficiem e respeitem as partes interessadas passadas, presentes e futuras (Carroll *et al*., 2020; Little *et al*., 2023; Wilkinson *et al*., 2016).

A comunidade de geociências tem estado na vanguarda das iniciativas de ciência aberta, adotando dados abertos, código aberto, acesso aberto e coleções abertas como componentes fundamentais da prática científica moderna, exemplificados por iniciativas como o *EarthChem* (para compartilhamento de dados geoquímicos) e o RRUFF *Project* (para dados abertos de mineralogia) (Ma *et al*., 2017). Essa abertura opera em dois níveis. O nível básico envolve tornar os recursos livremente acessíveis online, com permissão para reutilização e modificação. Já um nível mais avançado se concentra em anotar e conectar esses recursos para estabelecer redes para pesquisa colaborativa (Ma *et al*., 2017). A evolução dessa cultura de ciência aberta foi facilitada por avanços tecnológicos na coleta de dados, recursos de armazenamento e conectividade à internet, permitindo que pesquisadores compartilhem conjuntos de dados e colaborem com eficiência em ambientes de campo e laboratório por meio de plataformas como EarthCube, GeoSciCloud e ferramentas como Jupyter Notebooks e GitHub (Ramdeen *et al*., 2016).

Um progresso significativo em dados abertos foi alcançado nas geociências, com inúmeros serviços de dados estabelecidos por agências federais como NASA, USGS e NOAA, juntamente com portais de dados criados pela comunidade, como *OneGeology, EarthChem*, RRUFF, PANGAEA e PaleoBioDB (Ma, 2018). Iniciativas como a *Global Digital Heritage* (GDH) exemplificam a democratização da ciência ao fornecer acesso gratuito a dados digitais e modelos 3D do patrimônio cultural e natural para governos, instituições e o público. Esses

modelos têm sido utilizados em contextos educacionais e científicos, como em aulas interativas sobre arqueologia, simulações de restauração patrimonial e análises comparativas em pesquisas de campo. Essa abordagem permite a exploração virtual de coleções inteiras, museus e paisagens arqueológicas, disponibilizando recursos antes inacessíveis para estudantes, cientistas e entusiastas em todo o mundo (Global Digital Heritage, 2025).

A implementação dos Princípios de Dados FAIR tornou-se essencial para garantir que os dados permaneçam utilizáveis por humanos e máquinas ao longo do tempo. No entanto, a manutenção de dados FAIR requer repositórios digitais confiáveis, como o PANGAEA e o EarthChem, que possuam governança sustentável, infraestrutura robusta e políticas abrangentes que apoiem práticas acordadas pela comunidade. Esses repositórios desempenham um papel crucial na preservação ativa de dados, adaptando-se às evoluções tecnológicas e às necessidades das partes interessadas, conforme delineado pelos Princípios TRUST (Lin *et al*., 2020).

Práticas de ciência aberta em projetos de geociências do tipo *Big Science,* aqueles que envolvem grandes colaborações, instalações especializadas, investimentos multinacionais e volumes massivos de dados, apresentam desafios específicos (Wallis; Rolando; Borgman, 2013). Nessas disciplinas, a gestão de dados não é determinada apenas pela escala, mas também por características culturais, organizacionais e técnicas próprias, que exigem sistemas personalizados de gerenciamento e estratégias específicas de preservação digital. Esses fatores influenciam diretamente a forma como os dados são compartilhados, reutilizados e integrados em novos contextos científicos.

Apesar dos avanços recentes, ainda persistem desafios relevantes para alinhar soluções de preservação digital aos fluxos de trabalho científicos contemporâneos e às práticas institucionais vigentes (Lavoie, 2014). Com o crescimento contínuo no volume de dados digitais, há um risco real de perda de registros, o que torna essencial que a comunidade acadêmica enfrente esses obstáculos para garantir a longevidade e a reutilização dos dados científicos. As ciências de campo, como a geologia, historicamente ficaram atrás das ciências laboratoriais na disponibilização de dados e amostras, ultrapassando apenas recentemente a prática de indicar que os dados estão “disponíveis mediante solicitação” (McNutt *et al.*, 2016; Ma, 2018). Superar barreiras culturais, financeiras e técnicas exige ações coordenadas entre financiadores, editores e sociedades científicas, promovendo maior transparência e reprodutibilidade na pesquisa. Recentemente, agências como a NSF e a NASA passaram a

exigir planos de gestão de dados como critério obrigatório para financiamento, reforçando esse movimento.

A intersecção entre preservação digital, coleções científicas e práticas de ciência aberta nas geociências estabelece uma estrutura abrangente para o gerenciamento de dados geocientíficos ao longo de seu ciclo de vida. A curadoria digital, definida como o conjunto de atividades que envolvem o gerenciamento de dados desde o planejamento de sua criação até a garantia de sua descoberta e reutilização futuras, desempenha um papel fundamental nessa estrutura (Kim; Warga; Moen, 2013).

A gestão de dados de pesquisa (GDR) é essencial em todas as disciplinas, inclusive nas geociências, pois reforça a importância de preservar e compartilhar dados de maneira eficiente para atender às exigências de periódicos e agências de fomento (Borgman, 2015). A integração de estratégias de preservação digital às práticas de GDR assegura que os dados permaneçam acessíveis, compreensíveis e reutilizáveis ao longo do tempo, especialmente quando alinhadas aos princípios FAIR (Wilkinson *et al*., 2016). Além disso, a preservação digital sustenta a gestão contínua de acervos patrimoniais, garantindo que dados e recursos informacionais permaneçam viáveis e úteis para pesquisas futuras (Soave; Lemos, 2022).

Além disso, o compartilhamento do conhecimento científico é fundamental para que os pesquisadores possam refletir sobre descobertas anteriores e disseminar recursos com comunidades mais amplas. No entanto, esse processo exige investimentos significativos e tecnologias de documentação eficientes, que contribuam para reduzir a carga de trabalho dos cientistas e incentivar sua participação (Feger *et al*., 2020).

De modo geral, a integração dessas práticas apoia a evolução da investigação científica na comunidade de geociências, contribuindo para a preservação e compartilhamento de dados valiosos, aprimoramento da colaboração e promoção de maior confiança pública nos esforços científicos por meio da transparência e da replicabilidade (Tanlongo *et al*., 2025).

3 DESCRIÇÃO METODOLÓGICA

A metodologia foi desenhada para cumprir funções OAIS (ingestão, armazenamento arquivístico, gestão de dados, planejamento de preservação e acesso) e materializar princípios FAIR/CARE/TRUST no acervo geocientífico histórico da Litoteca CECO, sujeito a OCR ruidoso, heterogeneidade de formatos e demanda por interoperabilidade com o Tainacan.

No âmbito do projeto de preservação digital e interação acadêmica da Litoteca CECO, desenvolvido em parceria com o Laboratório de Dados, Métricas Institucionais e Reprodutibilidade Científica (DataLab), ligado ao Programa de Pós-Graduação em Ciência da Informação (PPGCIN) da UFRGS, articulam-se saberes interdisciplinares e tecnologias emergentes para garantir a conservação e a reutilização de um acervo científico de alto valor histórico e acadêmico. O projeto foi concebido a partir da necessidade de digitalizar, organizar e disponibilizar publicamente o conteúdo da Litoteca (biblioteca de testemunhos coletados pelo CECO), promovendo sua integração à Ciência Aberta e contribuindo para a reprodutibilidade científica.

A estrutura metodológica do projeto, em desenvolvimento, foi dividida em cinco etapas principais: diagnóstico técnico, digitalização e extração de dados, consolidação e organização dos dados, produção científica e, por fim, preservação e gestão digital.

A primeira etapa consistiu no diagnóstico técnico do acervo, realizado por meio de visitas ao espaço físico da Litoteca no CECO, no Campus do Vale da UFRGS. Essa análise inicial teve como propósito compreender as características gerais do acervo, identificar os objetos a serem digitalizados e delimitar o escopo da intervenção. A equipe técnica avaliou o estado de conservação dos documentos (relatórios técnicos impressos que apresentam os resultados das análises conduzidas com os testemunhos), o tipo de informação contida nas pastas e os desafios logísticos relacionados ao transporte e manuseio das amostras. Também foram coletadas informações sobre o histórico da coleção, fundamentais para garantir a contextualização dos dados no momento da curadoria digital. Esse levantamento permitiu a criação de um plano de ação com base em critérios técnicos e científicos, priorizando tanto a preservação da integridade física do material quanto a fidelidade das futuras cópias digitais.

Na segunda etapa, a qual o projeto se encontra, foram iniciadas as ações práticas de digitalização e extração de dados. O processo de escaneamento está sendo conduzido nas instalações da Faculdade de Biblioteconomia e Comunicação (FABICO), especificamente no Centro de Documentação de Acervo Digital da Pesquisa (CEDAP), que dispõe da infraestrutura necessária (*hardware e software*) para a digitalização de documentos históricos com alta resolução. A digitalização está sendo realizada com scanners profissionais, seguindo padrões internacionais de preservação digital, como a utilização de arquivos em formato *Tagged Image File Format* (TIFF) e *Portable Document Format/Archive* (PDF/A), garantindo a legibilidade e a durabilidade das cópias digitais. A resolução mínima adotada foi de 300 dpi,

assegurando a qualidade necessária para futuras análises científicas. Os documentos digitalizados incluíram mapas, descrições técnicas em forma de relatório de amostras, tabelas geológicas e outros registros de campo relevantes associados aos testemunhos.

Em paralelo ao escaneamento, está sendo aplicada a tecnologia de reconhecimento óptico de caracteres (OCR) para converter as imagens digitalizadas em textos editáveis e pesquisáveis. Essa automação é importante para possibilitar o tratamento posterior das informações por meio de ferramentas de Inteligência Artificial (IA). São utilizados softwares especializados como o *Tesseract* OCR, em conjunto com rotinas de verificação manual para corrigir possíveis erros de leitura, especialmente em documentos manuscritos ou datilografados com baixa qualidade de impressão. Os textos extraídos são organizados em planilhas temporárias, nas quais se iniciou a etapa de categorização e padronização dos dados.

Implementamos um pipeline de PLN que combina OCR com extração de entidades (topônimos, coordenadas, profundidade, período geológico, tipo de rocha e identificadores de amostra). As saídas alimentam dossiês digitais por testemunho. A qualidade é monitorada por amostragem estratificada, com medição de F1 por entidade e reingestão de correções para ajuste fino dos modelos. Essa automação acelera a curadoria sem substituir a validação de especialistas.

A terceira etapa da metodologia envolve a consolidação e organização dos dados extraídos. Esse momento é importante para integrar os diferentes tipos de informações, textuais, visuais, numéricas, em arquivos digitais coesos, estruturados e navegáveis. Para isso, adotou-se uma lógica de hipertexto, na qual textos, imagens e metadados são vinculados entre si de forma contextualizada, garantindo a preservação da narrativa científica original dos documentos. Cada item do acervo, dessa forma, compõe um dossiê digital completo, reunindo informações descritivas, técnicas e contextuais. Os arquivos gerados obedecem a padrões de interoperabilidade, facilitando a futura integração com plataformas de repositórios e sistemas de gestão de acervos digitais.

Para aumentar a eficiência da organização e sugerir classificações automatizadas, a equipe faz uso de ferramentas de inteligência artificial voltadas ao processamento de linguagem natural. Modelos supervisionados foram treinados com *corpus* acadêmico das áreas de Geociências e Ciência da Informação, com o objetivo de reconhecer padrões textuais, extrair entidades nomeadas, sugerir descritores e realizar agrupamentos semânticos dos conteúdos digitalizados. Essa camada de automação não substitui o trabalho humano, mas

serve como suporte para agilizar a curadoria e aumentar a coerência da organização dos dados, respeitando os critérios de consistência, completude e relevância científica.

A quarta etapa do projeto envolve a transformação dos dados, organizados, em produtos de comunicação científica, passíveis de serem publicados. À altura da escrita deste texto, esta etapa encontra-se planejada. A principal estratégia a ser adotada consiste na produção de *data papers*, um tipo de artigo acadêmico voltado à descrição detalhada de conjuntos de dados, suas origens, métodos de coleta, formatos e possibilidades de reutilização. Cada *data paper* elaborado documenta um subconjunto do acervo digitalizado, como uma série de amostras geológicas de uma mesma região ou um grupo de registros de uma expedição científica específica. Esses artigos incluem a descrição técnica dos processos de digitalização, os metadados dos arquivos gerados, a contextualização histórica e científica dos documentos e orientações para uso em pesquisas futuras. A adoção do formato *data paper* visa garantir a citação adequada do trabalho de digitalização e aumentar a visibilidade e o impacto dos dados preservados. Além disso, esses documentos científicos contribuem para o reconhecimento da importância dos acervos como fontes primárias de pesquisa.

Por fim, a quinta etapa concentra-se na incorporação do acervo digital ao sistema Tainacan, uma plataforma livre de gestão de acervos digitais, mantida pelo Instituto Brasileiro de Museus (IBRAM) e por universidades brasileiras. A adoção do Tainacan foi estratégica, por sua compatibilidade com padrões internacionais de metadados, sua interface amigável e sua capacidade de integração com outros sistemas de informação científica. Durante a fase de planejamento e preparação, a equipe do projeto iniciou a configuração do Tainacan, definindo um conjunto específico de metadados adaptados às particularidades da Litoteca. Essa estruturação visa permitir a organização dos itens com base em critérios como localidade geográfica, período geológico, tipo de material, entre outras categorias relevantes para as Geociências.

A seguir, apresenta-se a quadro 1, que sintetiza o mapeamento de metadados definido para a incorporação do acervo ao Tainacan.

Quadro 1 - Mapeamento para Tainacan

Campo	Tipo	Obrigatório	Padrão/controle
ID do testemunho	Texto curto	Sim	Regex `CECO-[A-Z0-9-]+`

Coordenadas	Ponto geográfico	Sim	WGS84 (decimal)
Profundidade (m)	Numérico	Sim	≥ 0
Período geológico	Taxonomia	Sim	Vocabulário controlado
Tipo de rocha	Taxonomia	Opc.	Vocabulário controlado
Data da coleta	Data	Sim	ISO 8601
Equipe	Texto controlado	Opc.	Autoridades locais
Dossiê digital (URI)	Link	Sim	Persistente
Licença	Lista	Sim	CC BY 4.0 (padrão)

Fonte: elaborado pelos autores (2025)

Além da organização, o sistema foi estruturado para atender aos principais requisitos da preservação digital: integridade, autenticidade, proveniência e acessibilidade, conforme definidos pelo modelo OAIS (*CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS*, 2012). A integridade dos documentos está sendo assegurada por meio da criação de cópias de segurança em múltiplos dispositivos e da adoção de formatos digitais duráveis. A autenticidade está sendo garantida com o registro de informações detalhadas sobre autoria, datas de criação, modificações e vínculos institucionais. Com base nas informações coletadas na etapa inicial do projeto, está sendo feita a documentação da proveniência, a fim de estabelecer o contexto de produção e circulação dos documentos. Já a promoção da acessibilidade ocorre por meio da aplicação de licenças *Creative Commons* e da indexação dos dados em sistemas de busca e repositórios acadêmicos, permitindo que o acervo seja encontrado e utilizado por pesquisadores de diferentes áreas e instituições.

4 RESULTADOS PARCIAIS

Até o momento, o projeto de preservação digital da Litoteca do CECO/UFRGS alcançou resultados parciais nas duas primeiras etapas metodológicas: o diagnóstico técnico do acervo e o início do processo de digitalização e extração de dados. A análise inicial permitiu mapear o estado de conservação dos documentos, identificar os itens prioritários e planejar a

intervenção com base em critérios técnicos, científicos e de necessidade. Esse levantamento resultou na criação de um plano de ação que fundamentou decisões como a seleção dos materiais, os critérios e o planejamento logístico para digitalização, considerando a variedade de formatos do acervo, que exige estratégias específicas e equipamentos distintos.

Durante a digitalização, estão sendo aplicadas tecnologias compatíveis com padrões internacionais de preservação, como o uso de formatos duráveis (TIFF e PDF/A), resolução mínima de 300 dpi e protocolos padronizados de nomeação. Também está em andamento a aplicação de ferramentas de OCR para converter os documentos em textos pesquisáveis, com resultados promissores, embora ainda exijam revisão manual para garantir a acurácia. Cada item está sendo convertido em arquivos PDF individualizados com OCR incorporado, permitindo a recuperação textual e a extração automatizada de dados. Essa extração vem sendo potencializada por tecnologias de inteligência artificial, como o ChatGPT, com o reconhecimento de padrões estruturais dos documentos, viabilizando a organização das informações e o desenvolvimento de técnicas que estão sendo aprimoradas para as próximas etapas do projeto.

Como resultado prático, diversos relatórios técnicos e tabelas geológicas já foram digitalizados e organizados em dossiês digitais estruturados, atualmente em preparação para futura indexação e descrição no sistema Tainacan. A categorização inicial dos dados permitiu identificar padrões descritivos recorrentes, facilitando a padronização de metadados e a interoperabilidade com sistemas de repositórios acadêmicos. A atuação conjunta entre especialistas em Geociências e Ciência da Informação tem possibilitado ajustes metodológicos contínuos, respeitando as especificidades do acervo.

Além disso, foram realizados testes exploratórios com ferramentas de inteligência artificial voltadas à extração e organização automática de informações, como nomes geográficos, períodos geológicos e tipos de rocha. Embora ainda em fase experimental, esses testes indicam o potencial de aplicar técnicas avançadas de processamento de linguagem natural para apoiar a curadoria digital em acervos científicos complexos.

5 CONSIDERAÇÕES FINAIS

A metodologia desenvolvida no projeto de preservação digital da Litoteca do CECO/UFRGS representa uma contribuição importante para o campo da curadoria digital de dados, documentos e informação aplicada às Geociências e às Ciências Naturais. O trabalho

interdisciplinar entre especialistas em Geociências e Ciência da Informação permitiu a construção de um fluxo de trabalho potente, escalável e replicável em outras instituições com acervo similar.

Com resultados parciais já mensurados, constata-se que ao transformar documentos físicos em dados estruturados, abertos e reutilizáveis, o projeto, além de salvaguardar o patrimônio científico da universidade, reforça os compromissos institucionais do CECO/IGEO e do DataLab/PPGCIN com a Ciência Aberta, a Reprodutibilidade e a sustentabilidade do conhecimento científico.

Ademais, os dados e informações do acervo composto por testemunhos geológicos sobre a costa e o fundo marinho do sul do Brasil do CECO, podem subsidiar e instigar novas pesquisas no âmbito das geociências, oceanografia e áreas interdisciplinares.

REFERÊNCIAS

BORGMAN, Christine L. **Big data, little data, no data**: scholarship in the networked world. Cambridge: MIT Press, 2015.

BURNHAM, Brian; BOND, Clare; FLAIG, Peter P.; VAN DER KOLK, Dolores A.; HODGETTS, David. Outcrop conservation: Promoting accessibility, inclusivity, and reproducibility through digital preservation. **The Sedimentary Record**, [*s. l.*], v. 20, n. 1, p. 5–14, 2022. Disponível em: https://doi.org/10.2110/sedred.2022.1.2. Acesso em: 25 maio 2025.

CARROLL, Stephanie Russo *et al.* The CARE Principles for Indigenous Data Governance. **Data Science Journal**, [*s. l.*], v. 19,
p. 43, 2020. Disponível em: https://doi.org/10.5334/dsj-2020-043. Acesso em: 25 maio 2025.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (CCSDS). **Reference Model for an Open Archival Information System (OAIS)**. Magenta Book CCSDS 650.0-M-2. Washington, D.C.: CCSDS Secretariat, 2012. Disponível em: https://public.ccsds.org/pubs/650x0m2.pdf. Acesso em: 25 maio 2025.

CONWAY, Paul. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. **The Library Quarterly**, v. 80, n. 1, 2010. DOI: 10.1086/648463. Disponível em: https://www.journals.uchicago.edu/doi/10.1086/648463. Acesso em: 25 maio 2025.

DAY, Michael. Toward Distributed Infrastructures for Digital Preservation. **International Journal of Digital Curation**, v. 3, n. 1, 2008. DOI: 10.2218/ijdc.v3i1.39. Disponível em: https://www.ijdc.net/index.php/ijdc/article/view/39. Acesso em: 24 maio 2025.

FEGER, Andreas; GRIMMER, Jannis; KAPELLER, Jakob; THEIN, Henrik. Scaling up open science: Data sharing as a function of research complexity. **Science and Public Policy**, v. 47, n. 5, p.

620–629, 2020. Disponível em: https://doi.org/10.1093/scipol/scaa014. Acesso em: 25 maio 2025.

GLOBAL DIGITAL HERITAGE. **Digital Heritage Documentation for the World**. [*S. l.: s. n*.], 2025. Disponível em: https://globaldigitalheritage.org/. Acesso em: 25 maio 2025.

HEDSTROM, Margaret. Digital preservation: a time bomb for digital libraries. **Computers and the Humanities**, [*s. l.*], v. 31, n. 3, p. 189–202, 1998. Disponível em: https://deepblue.lib.umich.edu/handle/2027.42/42573. Acesso em: 25 maio 2025.

LAVOIE, Brian F. **The Open Archival Information System (OAIS) Reference Model**: Introductory Guide. 2. ed. York: Digital Preservation Coalition, 2014. Disponível em: https://www.dpconline.org/docs/technology-watch-reports/1359-twr14-02/file. Acesso em: 25 maio 2025.

LEONELLI, Sabina. **Data-centric biology**: A philosophical study. Chicago: University of Chicago Press, 2016.

LIN, Dawei; HODSON, Simon; PATERSON, Sally; TEPER, Jennifer; LUSSIER, Yves; WYATT, Sally; PETERS, Deborah. The TRUST Principles for digital repositories. **Scientific Data**, [*s. l.*], v. 7, n. 144, 2020. Disponível em: https://doi.org/10.1038/s41597-020-0486-7. Acesso em: 25 maio 2025.

LITTLE, Holly; KARIM, Talia; KRIMMEL, Erica; WALKER, Lindsay J. A community-driven strategy for addressing fossil taxonomy challenges. **Biodiversity Information Science and Standards**, [*s. l.*], v. 7, e111507, 2023. Disponível em: https://doi.org/10.3897/biss.7.111507. Acesso em: 25 maio 2025.

KIM, Jeonghyun; WARGA, Edward; MOEN, William E. Competencies required for digital curation: an analysis of job advertisements. **International Journal of Digital Curation**, [*s. l*.], v. 8, n. 1, p. 66–83, 2013. Disponível em: https://doi.org/10.2218/ijdc.v8i1.242. Acesso em: 25 maio 2025.

MA, Xiaogang. Data Science for Geoscience: Leveraging Mathematical Geosciences with Semantics and Open Data. *In*: SAGAR, B.S. Daya; CHENG, Qiuming; AGTERBERG, Frits (orgs.). **Handbook of Mathematical Geosciences**: Fifty Years of IAMG. Cham: Springer, 2018. p. 687–702. Disponível em: https://doi.org/10.1007/978-3-319-78999-6_34. Acesso em: 25 maio 2025.

MA, Xiaogang *et al*. Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. **ISPRS International Journal of Geo-Information**, v. 6, n. 11, p. 368, 2017. Disponível em: https://doi.org/10.3390/ijgi6110368. Acesso em: 24 abr. 2025.

McNUTT, Marcia *et al*. Liberating field science samples and data. **Science**, [*s. l.*], v. 351, n. 6277, p. 1024–1026, 2016. Disponível em: https://doi.org/10.1126/science.aad7048. Acesso em: 25 maio 2025.

MONS, Barend *et al*. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. **Information Services & Use**, v. 37, n. 1, p. 49–56, 2017. Disponível em: https://doi.org/10.3233/ISU-170824. Acesso em: 24 abr. 2025.

MULLENDORE, Gretchen L. *et al*. What About Model Data? Best Practices for Preservation and Replicability. **Frontiers in Climate**, Lausanne, v. 3, 2021. Disponível em: https://www.frontiersin.org/articles/10.3389/fclim.2021.763420/full. Acesso em: 25 maio 2025.

RAMDEEN, Sarah; GOLDSTEIN, Elizabeth B.; HORSBURGH, Jeffrey S. EarthCube: Developing a community-driven data and knowledge infrastructure for the geosciences. **Eos, Transactions American Geophysical Union**, [*s. l*.], v. 97, 2016. Disponível em: https://doi.org/10.1029/2016EO056769. Acesso em: 25 maio 2025.

RIEGER, Oya Y. **The state of digital preservation in 2018**: a snapshot of challenges and gaps. Ithaka S+R, 2018. Disponível em: https://doi.org/10.18665/sr.310626. Acesso em: 24 abr. 2025.

SOAVE, Maycon; LEMOS, Daniela Lucas da Silva. Curadoria digital em acervos do patrimônio cultural digital: aspectos teóricos e práticos no âmbito da Ciência da Informação. **Brazilian Journal of Information Science**: Research Trends, v. 16, 2022. Disponível em: https://dialnet.unirioja.es/descarga/articulo/8501735.pdf. Acesso em: 25 maio 2025.

TANLONGO, Federica; SCHIRRU, Luca; FREDELLA, Maria Incoronata; MERCURIO, Daniela; FREDA, Carmela. The ethical dimension of sharing solid Earth Science data. **Journal of Geoethics and Social Geosciences**, v. 2, n. Especial, p. 1–30, 2025. Disponível em: https://www.journalofgeoethics.eu/index.php/jgsg/article/view/64. Acesso em: 25 maio 2025.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. CENTRO DE ESTUDOS DE GEOLOGIA COSTEIRA E OCEÂNICA. **Linhas de pesquisa**. Porto Alegre, 2025. Disponível em: https://www.ufrgs.br/ceco/?page_id=37. Acesso em: 25 abr. 2025.

WALLIS, Jillian C.; ROLANDO, Elizabeth; BORGMAN,Christine L. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. **PLoS ONE**, v. 8, n. 7, p. e67332, 2013. DOI: 10.1371/journal.pone.0067332. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332. Acesso em: 25 maio 2025.

WILDMAN, Jamie P. *et al*. The value of museum and other uncollated data in reconstructing the decline of the chequered skipper butterfly *Carterocephalus palaemon* (Pallas, 1771). **Journal of Natural Science Collections**, [*s. l.*], v. 10, p. 31–44, 2022. Disponível em: https://www.natsca.org/sites/default/files/publications-full/JoNSC-Volume-10.pdf. Acesso em: 25 maio 2025.

WILKINSON, Mark D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, [*s. l.*], v. 3, n. 160018, p. 1–9, 2016. Disponível em: https://doi.org/10.1038/sdata.2016.18. Acesso em: 25 maio 2025.

YAKEL, Elizabeth. Digital curation. **OCLC Systems & Services**, [*s. l.*], v. 23, n. 4, p. 335–340, 2007. Disponível em: https://www.researchgate.net/publication/220418492_Digital_curation. Acesso em: 25 maio 2025.