CAPÍTULO 5

PROPOSTA DE AUTOMAÇÃO DA CURADORIA DE METADADOS COM MACHINE LEARNING

Karolayne Costa Rodrigues de Lima[1]
Marcos Sfair Sunye[2]

[1] Universidade Federal do Paraná (UFPR). ORCID: https://orcid.org/0000-0002-6311-8482.
[2] Universidade Federal do Paraná (UFPR). ORCID: https://orcid.org/0000-0002-2568-5697.

5.1 INTRODUÇÃO

No cenário contemporâneo de *big data*, a gestão eficaz de grandes volumes de dados tornou-se um desafio central para diversas disciplinas, desde as ciências ambientais, agrárias, engenharias, saúde, até ciência da computação e informação. A gestão de dados científicos é crítica não apenas para o armazenamento de dados, mas também para sua recuperação, análise e reutilização. Acreditamos que o cerne da gestão reside primordialmente nos esquemas de metadados, que fornecem as estruturas necessárias para a organização, o acesso e o uso dos dados. No entanto, questões de interoperabilidade, qualidade dos metadados e adoção dos princípios FAIR (*Findable, Accessible, Interoperable and Reusable)* continuam a representar desafios significativos, especialmente em ambientes heterogêneos que cruzam fronteiras disciplinares e institucionais. Isso suscita uma questão investigativa premente: quais estratégias sobre a gestão de metadados podem ser desenvolvidas para aprimorar a interoperabilidade e a qualidade de dados científicos de forma eficaz e robusta em contextos multidisciplinares?

Uma abordagem promissora envolve aprimorar a curadoria digital e dados, especificamente através de uma curadoria dedicada a metadados. Isso visa minimizar erros descritivos, melhorar a completude dos metadados e elevar a qualidade do processo de representação descritiva. A adoção da Inteligência Artificial (IA) e *Machine Learning* (ML) na curadoria e análise de metadados tem o potencial de transformar significativamente a gestão de dados científicos. A eficácia na gestão de metadados torna-se relevante para a acessibilidade, reutilização e integração de dados em repositórios abertos.

O objetivo deste capítulo é, portanto, apresentar uma definição sobre o processo de curadoria de metadados e indicar uma possibilidade de automação do processo de extração de metadados descritivos utilizando IA e ML.

5.2 METADADOS E PROCESSO DE CURADORIA

No contexto da gestão de dados científicos, os metadados desempenham um papel fundamental ao servir como a espinha dorsal para a organização, acesso, preservação e compreensão dos dados. Conceitualmente, “metadados são uma declaração sobre algo potencialmente informativo” na definição de Pomerantz

(2015, p. 26, tradução nossa). Para Riley (2017, p. 1, tradução nossa[1]), metadados são "as informações que criamos, armazenamos e compartilhamos para descrever coisas"; ou ainda, como informações que detalham a origem, natureza e condições de uso de dados, facilitando sua recuperação e compreensão.

A partir disso, derivamos que metadados são informações estruturadas ou semi-estruturadas que descrevem, explicam, localizam ou, de alguma forma, facilitam o acesso, gestão e uso de recursos de informação. Essas informações são fundamentais para pesquisadores e gestores de dados ao navegarem por repositórios de dados, garantindo que os recursos sejam acessíveis e úteis ao longo do tempo.

Os metadados servem a vários propósitos essenciais na gestão de dados científicos. Como apontam Lee, Tibbo e Schaefer (2007), metadados permitem a descoberta de dados ao fornecer palavras-chave ou termos descritivos que facilitam as buscas em bases de dados. Os metadados garantem a integridade dos dados científicos ao manter um registro de sua proveniência, o que é vital para validar a pesquisa e suas conclusões. Eles também promovem a interoperabilidade entre diferentes sistemas e disciplinas científicas ao usar normas e padrões comuns, facilitando a colaboração e o compartilhamento de dados (Renear; Palmer, 2009).

Os metadados funcionam através de esquemas que estruturam a informação de modo que possa ser facilmente acessada, entendida e utilizada tanto por humanos quanto por máquinas (Chan; Zeng, 2006). Os metadados são organizados em elementos descritivos que podem incluir título, autor, data de criação, localização geográfica, e padrões de acesso (Gilliland, 2016; Kitchin, 2014). Esses elementos são frequentemente padronizados conforme normas internacionais, como as do *Dublin Core*, para garantir a consistência e facilitar a interoperabilidade entre diferentes plataformas e repositórios.

Existem diversos tipos de metadados, cada um servindo a diferentes necessidades dentro da gestão de dados científicos (Higgins, 2007; Gililand, 2016; Formenton *et al*., 2017; Silva, 2019):

- Metadados Descritivos: como o nome sugere, esses metadados descrevem os dados para facilitar sua identificação e descoberta. Incluem informações como título, resumo, autor e palavras-chave. Bates (2006) destaca a importância dos metadados descritivos na facilitação do acesso e recuperação de dados;

- Metadados Estruturais: referem-se à organização e ao formato dos dados,

[1] Trecho original: *the information we create, store, and share to describe things.*

explicando como os dados estão arranjados ou se relacionam. Os metadados estruturais podem detalhar como os dados estão formatados, quais tabelas estão relacionadas, ou como os arquivos multimídia estão codificados (Caplan, 2009);

- Metadados Administrativos: relacionados à gestão e preservação dos dados, esses metadados incluem informações sobre direitos autorais, restrições de acesso e registros de alterações. Metadados administrativos são relevantes para administrar a segurança e os direitos de uso dos dados;

- Metadados de Proveniência: informam sobre a origem dos dados, quem os coletou, com qual metodologia e em quais circunstâncias. Estes são essenciais para estabelecer a autenticidade e a qualidade dos dados (Moreau; Missier, 2013);

- Metadados de Contexto Integrado: formado por metadados que definem os aspectos "científicos" dos dados, incluindo propriedades e atributos que descrevem o tema tanto em dimensões qualitativas quanto quantitativas, a essência dos dados e seus elementos contextuais. Esta categoria se diferencia da categoria descritiva, pois os metadados científicos concentram-se em aspectos específicos da metodologia e da ciência que fundamentam os dados.

Apresentamos neste capítulo a categoria de “Metadados de Contexto Integrado”, que engloba os aspectos científicos e concentra os metadados descritores do fenômeno (um evento, fato, ocorrência ou objeto) e métodos de pesquisa (criação, coleta, tratamento e análise). A justificativa para esta nova categoria baseia-se em nossa observação de que os metadados relacionados às propriedades científicas dos dados estão frequentemente dispersos nas diversas categorias dentro dos esquemas descritivos, o que compromete a atenção devida durante o preenchimento dos metadados em repositórios. Essa dispersão é evidente em padrões de metadados multidisciplinares. Por outro lado, em padrões disciplinares, esses metadados tendem a ser mais eficientemente organizados em subgrupos específicos, como método, taxonomia, cobertura (geográfica e temporal) e anotações semânticas.

A categorização dos metadados em diferentes tipos — descritivos, estruturais, administrativos, proveniência e também de contexto integrado — atende a objetivos específicos dentro da gestão de dados, melhorando a eficácia da organização, utilização, e preservação dos dados (Gilliland, 2016). Essa divisão facilita a implementação de práticas detalhadas e orientadas para a gestão de dados, atendendo às

diversas necessidades e exigências de pesquisadores, bibliotecários, arquivistas e tecnologias de informação.

A tipologia de metadados dentro dos esquemas tem como base: a **interoperabilidade**, no sentido de facilitar a comunicação e o intercâmbio de dados entre diferentes plataformas e sistemas. Por exemplo, os metadados descritivos padronizados permitem que bibliotecas digitais, repositórios de dados e outras plataformas compartilhem e acessem dados de forma eficiente; a **automatização**, dado que os tipos de metadados permitem o desenvolvimento de ferramentas automatizadas que podem manipular especificamente metadados descritivos, estruturais, administrativos ou de proveniência, dependendo das necessidades do usuário ou do sistema; a **customização e flexibilidade**, pois os esquemas de metadados podem ser adaptados para enfatizar tipos de metadados mais relevantes para determinadas disciplinas ou tipos de dados, oferecendo flexibilidade para atender às necessidades específicas de diferentes comunidades científicas ou projetos de pesquisa; **gerenciamento de recursos digitais**, no sentido de que a tipologia ajuda na criação de políticas robustas de gestão de dados, onde cada tipo de metadado é gerenciado de acordo com seu papel específico, garantindo a longevidade, a acessibilidade e a utilidade dos recursos digitais (Zeng; Qin, 2016).

Essas tipologias estão representadas nos agrupamentos de metadados dentro dos esquemas e padrões. Um esquema de metadados é uma estrutura organizada que define tipos específicos de metadados para um conjunto de dados, especificando quais informações são necessárias para cada registro dentro de um sistema de informação (Zeng; Qin, 2016). Esquemas de metadados são projetados para serem ferramentas que facilitam a ordem, o controle e a descrição dos dados (Gilliland, 2016). Eles incluem definições de campos de metadados, especificações de conteúdo e regras de formatação, fornecendo um guia sobre como os metadados devem ser registrados e mantidos.

Um padrão de metadados, por outro lado, é um conjunto de diretrizes adotado amplamente que determina um formato consistente para a descrição de dados. Esses padrões são essenciais para garantir a interoperabilidade entre diferentes sistemas e plataformas. Padrões de metadados são fundamentais para facilitar a comunicação de informações de metadados de maneira uniforme, o que é particularmente útil em ambientes colaborativos e interdisciplinares (Pomerantz, 2015). Exemplos incluem o *Dublin Core*, um padrão simples e flexível usado globalmente para descrever uma ampla gama de recursos de rede, ou o *Machine-Readable Cataloging* (MARC), que é especificamente orientado para as necessidades das bibliotecas.

A tipologia de metadados está intrinsecamente ligada aos padrões de metadados através do modo como as informações são categorizadas e utilizadas. Cada padrão de metadados pode suportar diferentes tipos de metadados, como descritivos, estruturais, administrativos e de proveniência. Por exemplo, o padrão *Dublin Core* suporta metadados descritivos com seus 15 elementos fundamentais e qualificadores, que incluem título, criador e assunto, facilitando a descoberta e organização de recursos digitais (Dublin Core Metadata Initiative, 2020). Outro exemplo com maior complexidade é o padrão ISO 19115:2014 (*Geographic Information – Metadata*), norma internacional que define o esquema necessário para descrever informações e serviços geográficos por meio de metadados para diversos países com cerca de 300 elementos divididos em classes e subclasses (Loti *et al.,* 2019).

Os esquemas de metadados são caracterizados como um conjunto de elementos que, por sua vez, fornecem a infraestrutura necessária para implementar esses padrões de forma eficaz dentro de um ambiente particular. Eles detalham como os metadados devem ser capturados, interpretados e geridos de acordo com os padrões estabelecidos, assegurando que as práticas de documentação estejam alinhadas com as necessidades organizacionais e os requisitos tecnológicos (Zeng; Qin, 2016).

A gestão de dados científicos, portanto, envolve uma série de práticas necessárias para o melhor funcionamento dos repositórios e/ou outros sistemas que armazenam dados científicos. A curadoria de metadados é uma dessas práticas e é um conceito fundamental na gestão e preservação de informações digitais, que envolve processos meticulosos para a organização, manutenção e atualização de metadados ao longo do tempo. Embora a curadoria digital seja um termo amplo para designar diversos tipos de processos curatoriais (Triques; Arakaki; Castro, 2020), este ensaio tem foco na curadoria de metadados.

A curadoria de metadados abrange várias disciplinas, incluindo a Biblioteconomia, a Ciência da Informação, a Ciência da Computação e áreas específicas do conhecimento onde dados são intensivamente utilizados, como genômica, pesquisa climática e ciências sociais. Estas áreas dependem de metadados precisos e bem gerenciados para a análise de dados, colaboração entre pesquisadores e publicação de resultados (Palmer *et al.,* 2007).

Conceitualmente, curadoria de metadados é a prática de tratar detalhadamente os metadados para garantir que os recursos de dados sejam preservados, acessíveis e compreensíveis (Gilliland, 2016). O foco da curadoria de metadados recai no gerenciamento dos metadados associados aos dados científicos por meio de

um processo de criação, manutenção e gestão de metadados para garantir que os dados sejam facilmente localizados, acessados e compreendidos. A curadoria assegura que cada conjunto de dados possa ser autenticado e utilizado de acordo com as diretrizes estabelecidas, promovendo assim a integridade e confiabilidade da informação em repositórios.

Curadoria de metadados e curadoria de dados são processos complementares da gestão de dados, porém distintos. A curadoria de dados foca na sustentabilidade de longo prazo dos próprios dados, enquanto a curadoria de metadados se concentra em garantir que os dados sejam utilizáveis e compreensíveis para os usuários finais. Juntos, esses processos ajudam a construir repositórios de dados robustos que são essenciais para o avanço do conhecimento científico e acadêmico.

Os processos envolvidos na curadoria de metadados compreendem procedimentos tradicionais, tais como coleta e entrada, validação, classificação e categorização, além de anotação e enriquecimento. Estes procedimentos são intrinsecamente descritivos e são compreendidos dentre os 15 processos e três ações delineados no modelo do Ciclo de Vida da Curadoria Digital, da Digital Curation Centre (DCC) proposto por Higgins (2008). Os processos de curadoria de metadados consistem em procedimentos executados tanto de maneira repetitiva quanto ocasional, à medida que os conjuntos de dados são expandidos, como demonstrado no Quadro 5.1.

Quadro 5.1 – Processos da Curadoria de Metadados.

Processo	Descrição
Coleta e entrada	Inserir manualmente informações detalhadas sobre os dados, como autor, data, localização e palavras-chave. Este processo depende do conhecimento e da atenção do curador para evitar erros e garantir a precisão.
Validação	Revisar os metadados gerados para verificar sua precisão e completude. Isso pode incluir a verificação manual de datas, nomes de autores, e *links* para garantir que tudo esteja correto e atualizado.
Classificação e categorização	Classificar e categorizar dados com base em seu conteúdo ou propósito. Curadores humanos podem aplicar seu entendimento contextual para organizar dados de maneira que máquinas não conseguiriam replicar sem instruções específicas.

Anotação e enriquecimento	Adicionar anotações ou informações contextuais que ajudem outros usuários a entender e utilizar os dados de maneira mais eficaz. Isso pode incluir resumos, comentários sobre a qualidade dos dados, ou detalhes sobre metodologias de coleta de dados.

Fonte: Os autores (2024).

Os processos descritos no Quadro 5.1 são iniciados quando um usuário acessa o repositório e submete um conjunto de dados. Durante a submissão, o usuário, seja um gestor ou pesquisador, preenche um formulário com metadados que descrevem o recurso em aspectos administrativos, descritivos, de proveniência e contexto integrado. Após inserir as informações, o usuário realiza o *upload* do conjunto de dados, atribui uma licença ao recurso e concorda com os termos do repositório. Dada a extensão dos metadados necessários para conjuntos de dados maiores, o processo pode ser complexo e consumir tempo considerável, o que, segundo Curty *et al.* (2017), influencia a disposição dos pesquisadores para compartilhar e reutilizar dados. A automação desse processo poderia economizar tempo e esforço.

5.3 A NECESSIDADE DE AVANÇOS NA CURADORIA DE METADADOS

Considerando o crescimento volumétrico, a diversificação e a complexidade dos dados científicos, os tradicionais métodos manuais e heurísticas simples já não são suficientes para garantir a gestão eficiente de metadados. Tradicionalmente, a curadoria de metadados envolve processos manuais que são insustentáveis na era do *big data*.

Nos 21 repositórios de dados brasileiros consultados no Re3data[2], através da consulta às diretrizes de submissão de dados, observamos que a submissão e o preenchimento dos metadados ainda é realizado manualmente, com os processos de curadoria conduzidos por gestores e/ou diretamente pelos pesquisadores. Em ambos os casos, o procedimento é executado por agente humano. Embora existam vantagens na representação descritiva realizada por humanos, sobretudo devido à capacidade de inferência e raciocínio contextual aos dados — especialmente quanto a descrição de metadados de contexto integrado —, acreditamos que a

[2] Disponível em: https://www.re3data.org/search?query=&countries[]=BRA. Acesso em: 13 abr. 2024.

extração automática de metadados, sejam eles descritivos, administrativos ou de proveniência, poderia ser realizada com maior precisão e uniformidade através de um processo de automação, sobretudo nos que possuem grandes volumes de dados.

Dada a justificativa para automação dos metadados, a implementação de Inteligência Artificial (IA) e *Machine Learning* (ML) surge como um avanço promissor. Estas tecnologias oferecem a capacidade de automatizar a geração e curadoria de metadados, o que pode significativamente aprimorar a eficiência dos processos envolvidos. Algoritmos de ML podem ser aplicados para analisar e extrair automaticamente informações relevantes de grandes conjuntos de dados estruturados ou semi-estruturados, utilizando técnicas como o Processamento de Linguagem Natural (PLN) para gerar metadados descritivos e precisos (Bizer; Heath; Berners-Lee, 2009).

A Inteligência Artificial (IA) é o campo de estudo que busca emular capacidades humanas através de sistemas computacionais. Isso inclui raciocínio, aprendizado, percepção visual e linguagem natural. A IA é definida como o estudo de agentes que recebem percepções do ambiente e realizam ações que maximizam suas chances de sucesso em algum objetivo ou tarefa (Russell; Norvig, 2013).

Já o *Machine Learning* (ML) é um subcampo da IA que foca especificamente no desenvolvimento de algoritmos que permitem que computadores aprendam a partir de dados e façam previsões ou tomem decisões baseadas nesse aprendizado sem serem explicitamente programados para cada tarefa. Ainda, o ML envolve a construção de modelos que podem inferir padrões a partir de dados complexos e fazer previsões precisas (Emygdio, 2021). Enquanto IA abrange uma gama mais ampla de capacidades cognitivas simuladas, ML é focado em modelos e algoritmos que aprendem e mudam suas estruturas de processamento com base na experiência (dados) que recebem.

Nessa perspectiva, considerando a rotina de submissão de um conjunto de dados em repositórios, é viável a integração de IA e ML na curadoria de metadados de forma a permitir que o repositório aprenda continuamente com os diversos padrões de dados, melhorando assim a precisão e a utilidade dos metadados gerados. Por exemplo, sistemas de aprendizado supervisionado podem ser treinados para identificar padrões comuns e anomalias nos dados, sugerindo categorizações e *tags* de metadados mais adequadas para cada conjunto de dados (Halevy; Norvig; Pereira, 2009).

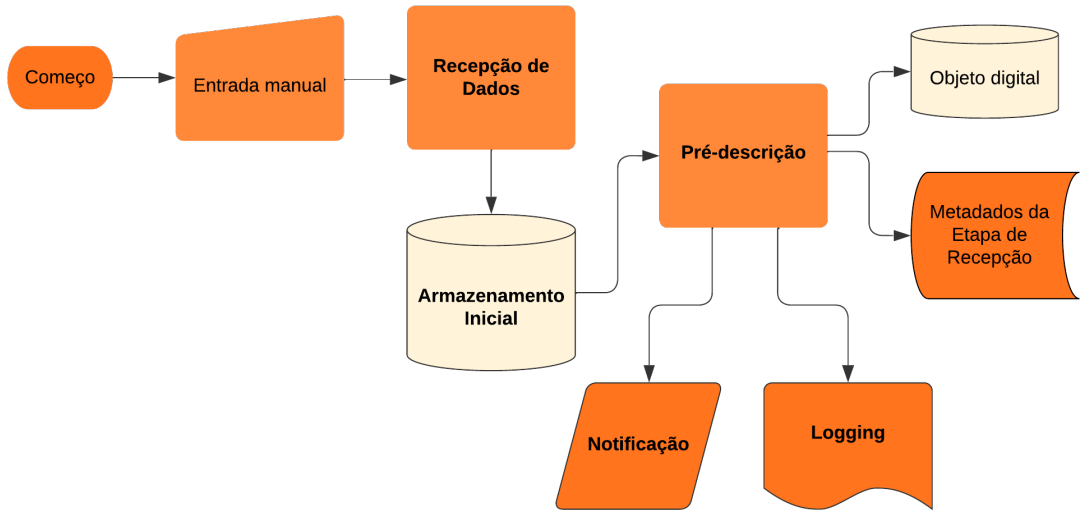
Algoritmos de ML já são utilizados para classificação automática dos metadados "*description*" e "*material type*" em Recursos Educacionais Abertos (Segarra-Faggioni; Romero-Pelaez, 2022). Em outra aplicação, Yuan *et al.* (2009) utilizam algoritmos para realizar a extração automática de metadados a partir da cadeia de citações nas referências citadas.

5.3.1 ETAPAS PARA A AUTOMATIZAÇÃO DE METADADOS DESCRITIVOS PARA REPOSITÓRIO DE DADOS

A aplicação de IA e ML na Ciência da Computação tem evoluído rapidamente, fornecendo ferramentas capazes de processar e analisar grandes volumes de dados com precisão e eficiência. Algoritmos de aprendizado de máquina, como redes neurais e algoritmos de *clustering*, são particularmente eficazes na identificação de padrões e na classificação automática de grandes conjuntos de dados (Jordan; Mitchell, 2015).

Neste capítulo propomos a automação da curadoria de metadados como um processo constituído por duas etapas e ações sequenciais. Cada uma dessas etapas e ações requer a implementação de ferramentas específicas e algoritmos para ser executada. A indicação de ferramentas e algoritmos são apenas sugestões observadas na literatura e naturalmente que outras ferramentas podem ser utilizadas e adaptadas para a automação. A Etapa 1 é a Etapa de *Recepção e Ingestão de Dados* que consiste em uma etapa preliminar de descrição que fará a identificação e o registro da entrada de dados, coletando informações como a origem, tipo do arquivo, o formato e a data de recebimento. O fluxo de trabalho dessa etapa ocorre conforme ilustrado na Figura 5.1.

Figura 5.1 – Etapa 1: Recepção e ingestão de dados



Fonte: Os autores (2024).

A Etapa 1 - *Recepção e Ingestão de Dados* é composta por quatro ações que visam identificar, registrar a entrada do conjunto de dados no repositório, assegurando neste momento o preenchimento automático de metadados de proveniência. É a etapa que antecede a automação da extração dos metadados descritivos.

A ação de coleta de dados pode ser realizada a partir de várias fontes, como *upload* de usuários, transferências via API – *Application Programming Interface* – ou importações de outros sistemas. Para isso, é necessário configurar pontos de entrada e formulários de *upload*, bem como *endpoints* da API, garantindo a validação da integridade e do formato dos arquivos recebidos.

Uma vez coletados, os dados devem ser armazenados em um local temporário para processamento subsequente. Este *Armazenamento Inicial* envolve organizar os dados em diretórios estruturados por data de recepção e tipo de dado, assegurando que sejam mantidos em um formato seguro e acessível.

A ação de *Pré-descrição* envolve registrar metadados sobre cada conjunto de dados, incluindo informações como nome do arquivo, tamanho, data de recebimento e origem. Durante esta fase, é criado um registro inicial no repositório e são atribuídos identificadores únicos (DOI, *Handle* ou outros) aos conjuntos de dados, garantindo sua rastreabilidade e organização.

Por último, a ação de *Notificação* e *Logging* serve para informar as partes interessadas (usuário depositante e gestores) sobre a recepção dos dados e manter

registros detalhados das atividades de ingestão. Isso inclui o envio de notificações automáticas e a manutenção de *logs* detalhados das operações de ingestão para fins de auditoria e rastreamento, assegurando a transparência e a conformidade com as diretrizes estabelecidas pela política de gestão do repositório.

Ao final da Etapa 1, o repositório abrigará uma tabela com o registro do objeto digital e seus atributos (metadados). Na Etapa 2, *Automação de Metadados Descritivos*, o repositório fará um novo processamento do registro (objeto digital e seus metadados) automatizando a extração dos metadados descritivos.

O Quadro 5.2 indica um uma proposta para a automação de metadados descritivos dentro de um repositório de dados. Para cada fase do processo uma ação correspondente é delineada, juntamente com recomendações de ferramentas de código aberto e algoritmos sugeridos na literatura para conduzir essas operações.

Quadro 5.2 – Etapa 2: Automação de metadados descritivos.

Fase	Ação	Ferramentas / Algoritmos[3]	Descrição
Pré-processamento dos dados	Coleta e normalização de dados	Apache Tika PDFMiner	Extrair texto de formatos como PDF ou HTML e converter para um formato de texto padronizado.
Aplicação de PLN	Extração de informações e análise semântica	NLTK spaCy BERT (Hugging Face's Transformers)	Usar PLN para identificar e extrair informações chave (título, autores, palavras-chave); aplicar BERT para análise semântica profunda.
Geração de metadados	Classificação de informações	scikit-learn (SVM, árvores de decisão)	Classificar as informações extraídas em categorias específicas de metadados.
Validação e enriquecimento	Validar e enriquecer os metadados	Talend scikit-learn (k-means, PCA)	Validar metadados com scripts customizados e Talend; enriquecer metadados utilizando técnicas de aprendizado não supervisionado.
Integração e atualização	Armazenamento e atualização contínua	Elasticsearch Apache Solr	Integrar e atualizar metadados no repositório, usando sistemas de gerenciamento eficientes.

Fonte: Os autores (2024).

A primeira fase do processo de automação envolve o pré-processamento dos dados, onde documentos são coletados de diversas fontes e formatos, como *Portable Document Format* (PDF) e *HyperText Markup Language* (HTML). É possível utilizar ferramentas como *Apache Tika* e *PDFMiner* para extrair e normalizar textos, assegurando uniformidade essencial para análises subsequentes.

[3] Todas as ferramentas e algoritmos apresentados são de código aberto.

Após a normalização, inicia-se a aplicação de Processamento de Linguagem Natural (PLN) para extrair e analisar informações contidas nos textos. Ferramentas como NLTK e SpaCy são aplicadas para tarefas de identificação e extração de entidades nomeadas, como nomes de autores e títulos. Adicionalmente, o modelo de linguagem profunda BERT é empregado para uma análise semântica mais robusta, permitindo identificar relações complexas e inferências contextuais nos dados (Jurafsky; Martin, 2021).

Seguindo a extração e análise, a terceira fase é a *Geração de Metadados*, onde as informações são classificadas em categorias predefinidas conforme o esquema de metadados. Algoritmos de aprendizado de máquina, como *Support Vector Machines* (SVM) e árvores de decisão do pacote *scikit-learn*, organizam esses dados em metadados estruturados, transformando dados brutos em informação organizada e pronta para uso (Géron, 2019). Essa tecnologia é valiosa para a automação de metadados descritivos, onde a precisão e a organização dos dados são imperativas.

A fase de *Validação e Enriquecimento* é importante para garantir a precisão dos metadados gerados. Utiliza-se Talend para validar os metadados conforme regras estabelecidas e técnicas de aprendizado não supervisionado, como k-means e análise de componentes principais (PCA), para detectar padrões e sugerir possíveis enriquecimentos. De acordo com Loshin (2013), esse processo ajuda a preencher lacunas e aprimorar a qualidade e completude dos metadados.

A *Integração e Atualização* contínua dos metadados são realizadas usando sistemas de gerenciamento como Elasticsearch e Apache Solr. Esses sistemas facilitam a indexação e o acesso rápido aos metadados e permitem atualizações regulares à medida que novos dados são adicionados ao repositório (Manning; Raghavan; Schütze, 2008). Essa última fase assegura que o repositório se mantenha relevante, atualizado e eficaz para pesquisa e recuperação de informações.

Ao longo desse processo, a sinergia entre várias tecnologias e métodos especializados é essencial para transformar efetivamente grandes volumes de dados brutos em metadados precisos e úteis, garantindo que os repositórios de dados sejam uma fonte confiável e valiosa de informação científica. A automatização da curadoria de metadados descritivos em repositórios digitais pode ser uma alternativa viável para otimizar a acessibilidade, gerenciamento, interoperabilidade e preservação de grandes volumes de dados.

A proposta de automação de metadados acima é simplificada e de cunho ensaístico e focou na descrição de metadados de proveniência e descrição. Todavia, como

trabalho futuro, é intuito desenvolver a proposta de automação da extração de forma que a automação seja validada e expandida para as outras tipologias. Naturalmente que essa proposta não é exaustiva e indica apenas um caminho para beneficiar a extração automática de metadados em repositórios. Outros repositórios de códigos como o GitHub, por exemplo, não entraram no escopo da pesquisa, mas constituem-se como fontes relevantes para a pesquisa de ferramentas para os processos de automação.

5.4 DESAFIOS E LIMITAÇÕES NA AUTOMAÇÃO

Os desafios e limitações da automação de metadados podem ser categorizados em várias áreas, incluindo técnicos, éticos, organizacionais e relacionados à qualidade dos dados (Quadro 5.3).

Quadro 5.3 – Desafios técnicos na automação.

Desafios Técnicos	
Complexidade dos dados	Variedade de Formatos: dados em diferentes formatos (texto, imagem, vídeo) apresentam desafios na extração e padronização de metadados.
	Dados Não Estruturados: a extração de metadados de dados não estruturados, como documentos de texto e multimídia, pode ser difícil e imprecisa.
Precisão e confiabilidade	Erros de Extração: algoritmos de IA e ML podem cometer erros ao identificar e extrair metadados, resultando em dados incorretos ou incompletos.
	Contexto Semântico: capturar o contexto correto dos dados para gerar metadados precisos pode ser complicado, especialmente em áreas especializadas.
Integração de sistemas	Compatibilidade: garantir que os sistemas automatizados sejam compatíveis com os repositórios de dados existentes e com outros sistemas de informação.
	Interoperabilidade: facilitar a interoperabilidade entre diferentes plataformas e sistemas é um desafio técnico significativo.

Desafios Éticos e de Privacidade	
Privacidade	Dados Sensíveis: a automação pode inadvertidamente expor dados sensíveis ou pessoais durante o processo de extração e curadoria.
	Consentimento e Conformidade: garantir que a automação esteja em conformidade com leis de privacidade e que o consentimento apropriado seja obtido para o uso dos dados.
Viés algorítmico	Discriminação: algoritmos de IA podem herdar ou amplificar vieses presentes nos dados de treinamento, levando a discriminação e desigualdades.
	Transparência e Explicabilidade: tornar os processos algorítmicos transparentes e explicáveis para evitar vieses e garantir a confiança dos usuários.
Desafios Organizacionais	
Resistência à Mudança	Adoção de Novas Tecnologias: organizações podem ser resistentes a adotar novas tecnologias de automação devido à falta de familiaridade ou medo de mudanças.
	Treinamento e Capacitação: necessidade de treinar a equipe existente para trabalhar com novas ferramentas e sistemas automatizados.
Recursos e custos	Investimento Inicial: a implementação de sistemas de automação pode requerer um investimento significativo em infraestrutura e software.
	Manutenção Contínua: manter e atualizar sistemas de automação pode ser custoso e requerer recursos especializados
Desafios Relacionados à Qualidade dos Dados	
Qualidade e integridade dos dados	Dados Incompletos ou Imprecisos: sistemas automatizados dependem da qualidade dos dados de entrada; dados incompletos ou imprecisos podem comprometer a qualidade dos metadados gerados.
	Atualização dos Dados: manter os metadados atualizados conforme os dados subjacentes mudam ao longo do tempo.

Padronização e normatização	Padrões Diversos: a existência de múltiplos padrões de metadados pode complicar a automação, exigindo a harmonização de diferentes esquemas.
	Consistência: garantir a consistência dos metadados gerados por diferentes algoritmos e sistemas automatizados.

Fonte: Os autores (2024).

Apesar das limitações destacadas no Quadro 5.3, há diversas soluções para enfrentar os desafios técnicos e garantir a precisão e a confiabilidade dos metadados gerados. Essas soluções podem envolver a utilização de algoritmos adicionais que revisem os metadados extraídos, identificando e corrigindo possíveis erros antes que os dados sejam integrados aos repositórios. Ferramentas de verificação automática são úteis para detectar inconsistências e lacunas, assegurando que os metadados atendam aos padrões estabelecidos de qualidade e completude.

Os desafios éticos e de privacidade devem ser seriamente tratadas por meio do desenvolvimento de políticas e diretrizes claras para garantir que a automação da curadoria de metadados esteja em conformidade com as leis de privacidade e segurança de dados. Essas políticas devem proteger informações sensíveis contra exposições inadvertidas. As políticas e diretrizes devem adotar práticas transparentes e explicáveis na aplicação de algoritmos de IA e ML, evitando vieses e garantindo a confiança dos usuários. Implementar *frameworks* éticos e realizar auditorias regulares pode ajudar a manter a conformidade e a ética no uso dessas tecnologias.

A capacitação e o treinamento contínuo da equipe são determinantes para superar a resistência organizacional à mudança. Investir em programas de formação e capacitação que familiarizem funcionários e servidores com novas ferramentas e sistemas automatizados pode facilitar a transição e aumentar a aceitação interna. Um programa de capacitação que mantenha uma alta curva de aprendizado por meio de *workshops*, cursos de atualização e oficinas práticas são estratégias eficazes para preparar a equipe a utilizar essas tecnologias de forma eficiente.

Os desafios relacionados aos recursos e custos também merecem atenção. A adoção de soluções de código aberto e ferramentas modulares pode ser uma abordagem estratégica para reduzir o investimento inicial e permitir que as instituições escalem suas operações de forma gradual, distribuindo os custos ao longo do tempo. Afinal, investir em soluções de código aberto é uma das abordagens da

Ciência Aberta tanto na economia de recursos quanto na ampliação da colaboração científica entre pessoas e instituições. Há exemplos consolidados de sucesso no uso de *software* livre em todo o ciclo de vida da atividade científica, em especial nos repositórios digitais.

Integrar sistemas de monitoramento e atualização contínua mantém a relevância e a eficácia dos metadados ao longo do tempo. Utilizar sistemas de gerenciamento eficientes, como Elasticsearch e Apache Solr, facilita a indexação e o acesso rápido aos metadados, além de permitir atualizações regulares à medida que novos dados são adicionados aos repositórios. Essas ferramentas não apenas melhoram a acessibilidade e a interoperabilidade dos dados, mas também garantem que os repositórios permaneçam atualizados e alinhados com as necessidades dos usuários e as melhores práticas de gestão de dados.

5.5 CONCLUSÃO

Este capítulo enfatizou a importância crítica e as vantagens de implementar processos de automação na curadoria de metadados dentro de repositórios científicos, com especial enfoque no Aprendizado de Máquina (*Machine Learning*). Esta abordagem não apenas aprimora a precisão e a eficiência na gestão de metadados de proveniência e descritivos, mas também representa um avanço significativo na capacidade de repositórios de dados para suportar a pesquisa científica e a colaboração interdisciplinar.

No que tange à gestão de repositórios, a automação da curadoria de metadados representa uma estratégia possível para a otimização da gestão de repositórios de dados em universidades, especialmente em contextos onde as equipes são reduzidas e os servidores não possuem dedicação integral a essa função. A implementação de processos automatizados pode significativamente reduzir a demanda por recursos financeiros e humanos, aliviando a carga de trabalho dos gestores de repositórios e permitindo que tarefas repetitivas sejam executadas de forma eficiente e consistente. Ações de automação “poupam” o tempo dos gestores, liberando-os para atividades estratégicas e de maior valor agregado, além de reinvestido na capacitação contínua nas demais práticas de curadoria digital. Dessa forma, as universidades podem assegurar uma gestão mais eficaz e sustentável de seus repositórios de dados, mantendo a qualidade e a integridade das informações armazenadas, mesmo com recursos limitados.

A aplicação de IA e ML na curadoria de metadados destaca-se por sua habilidade de processar volumes grandes e complexos de dados de forma mais eficiente do que os métodos manuais tradicionais, garantindo a qualidade e a utilidade dos metadados gerados. Os algoritmos de PLN, por exemplo, permitem uma análise semântica detalhada que enriquece os metadados com contextos que podem ser essenciais para pesquisadores que buscam dados relevantes dentro de um vasto repositório.

A introdução de métodos automáticos para a geração, classificação, validação e enriquecimento de metadados facilita a interoperabilidade e a acessibilidade dos dados, características estas básicas e fundamentais para o apoio a iniciativas de Ciência Aberta (princípios FAIR e CARE) e para a promoção de um ambiente acadêmico mais colaborativo e transparente.

Enquanto este capítulo se concentrou na automação de metadados de proveniência e descritivos, os métodos discutidos têm aplicabilidade potencial para outras categorias de metadados, como os administrativos, e para a nova classe, os metadados científicos, sugerindo um campo para futuras pesquisas e desenvolvimentos. A contínua evolução das tecnologias de IA e ML provavelmente trará ainda mais ferramentas avançadas para a curadoria de metadados, reforçando a importância de adaptações contínuas e atualizações das práticas de gestão de dados científicos para maximizar seu valor e acessibilidade. Em nossa perspectiva, os metadados não devem ser vistos apenas como dados sobre dados, mas também com uma camada crítica de informação que proporciona visibilidade, acessibilidade e interoperabilidade aos conjuntos de dados científicos.

REFERÊNCIAS

BATES, M. J. Fundamental forms of information. **Journal of the American Society for Information Science and Technology**, [*s. l.*], v. 57, n. 8, p. 1033-1045, June 2006. DOI: https://doi.org/10.1002/asi.20369. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20369. Acesso em: 15 jun. 2025.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: the story so far. **International Journal on Semantic Web and Information Systems**, [*s. l.*], v. 5, n. 3, p. 1-22, 2009. DOI: http://doi.org/10.1145/3591366.3591378. Disponível em: https://dl.acm.org/doi/10.1145/3591366.3591378. Acesso em: 15 jun. 2025.

CAPLAN, P. **Understanding PREMIS**. Washington, DC: Library of Congress, Feb. 2009. Disponível em: http://www.loc.gov/standards/premis/understanding-premis.pdf. Acesso em: 27 fev. 2024.

CHAN, L. M.; ZENG, M. L. Metadata interoperability and standardization: a study of methodology part I: achieving interoperability at the schema level. **D-Lib Magazine**, [*s. l.*], v. 12, n. 6, June 2006. DOI: http://doi.org/10.1045/june2006-chan. Disponível em: https://www.dlib.org/dlib/june06/chan/06chan.html. Acesso em: 27 fev. 2024.

CURTY, R. G. *et al*. Attitudes and norms affecting scientists' data reuse. **Plos ONE**, [*s. l.*], v. 12, n. 12, e0189288, 2017. DOI: http://doi.org/10.1371/journal.pone.0189288. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0189288. Acesso em: 15 jun. 2025.

DUBLIN CORE METADATA INITIATIVE. **Dublin Core**. [*s. l.*]: DCMI, [2020]. Disponível em: https://www.dublincore.org/specifications/dublin-core/. Acesso em: 29 mar. 2024.

EMYGDIO, J. L. Inteligência Artificial da perspectiva da Ciência da Informação: onde estamos em termos de raciocínio computacional. **Fronteiras da Representação do Conhecimento**, Belo Horizonte, v. 1, n. 2, p. 171-193, dez. 2021. Disponível em: https://periodicos.ufmg.br/index.php/fronteiras-rc/article/view/37518/29324. Acesso em: 30 abr. 2024.

FORMENTON, D. *et al*. Os padrões de metadados como recursos tecnológicos para a garantia da preservação digital. **Biblios**: Revista de Biblioteconomía e Ciência da Informação, [*s. l.*], n. 68, p. 82-95, 2017. DOI: https://doi.org/10.5195/biblios.2017.414. Disponível em: https://biblios.pitt.edu/ojs/biblios/article/view/414. Acesso em: 15 jun. 2025.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**: concepts, tools, and techniques to build intelligent systems. 2nd ed. California: O'Reilly Media, 2019.

GILLILAND, A. J. Setting the stage. *In*: BACA, M. (ed.). **Introduction to metadata**. 3rd ed. Los Angeles: Getty Publications, 2016. Disponível em: http://www.getty.edu/publications/intrometadata/setting-the-stage/. Acesso em: 27 fev. 2024.

HALEVY, A.; NORVIG, P.; PEREIRA, F. The unreasonable effectiveness of data. **IEEE Intelligent Systems**, v. 24, n. 2, p. 8-12, 2009. DOI: http://doi.org/10.1109/MIS.2009.36. Disponível em: https://ieeexplore.ieee.org/document/4804817. Acesso em: 15 jun. 2025.

HIGGINS, S. The DCC Curation Lifecycle model. **International Journal of Digital Curation**, [*s. l.*], v. 3, n. 1, p. 134-140, 2008. DOI: http://doi.org/10.2218/ijdc.v3i1.48. Disponível em: https://ijdc.net/index.php/ijdc/article/view/48. Acesso em: 15 jun. 2025.

HIGGINS, S. **What are metadata standards**. Edinburgh: Digital Curation Centre, Feb. 2007. Disponível em: http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards. Acesso em: 29 mar. 2024.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. **Science**, [*s. l.*], v. 349, n. 6245, p. 255-260, 17 July 2015. DOI: http://doi.org/10.1126/science.aaa8415. Disponível em: https://www.science.org/doi/10.1126/science.aaa8415. Acesso em: 15 jun. 2025.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3rd ed. Hoboken, New Jersey: Prentice Hall, 2021 *Online manuscript released Jan. 2025*. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 1 mar. 2024.

KITCHIN, R. Big Data, new epistemologies and paradigm shifts. **Big Data & Society**, [*s. l.*], v. 1, n. 1, Apr.-June 2014. DOI: http://doi.org/10.1177/2053951714528481. Disponível em: https://journals.sagepub.com/doi/10.1177/2053951714528481. Acesso em: 15 jun. 2025.

LEE, C. A.; TIBBO, H. R.; SCHAEFER, J. C. Defining what digital curators do and what they need to know: the DigCCurr Project. *In*: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 7th, 2007, Vancouver. **Proceedings** [...]. New York: ACM, 2007. p. 49-50. DOI: https://doi.org/10.1145/1255175.1255183. Disponível em: https://dl.acm.org/doi/10.1145/1255175.1255183. Acesso em: 23 maio 2025.

LOSHIN, D. **Big Data Analytics**: from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph. Amsterdam: Elsevier, 2013.

LOTI, L. B. S. *et al*. Atualização da norma ISO 19115 e os impactos no Perfil de Metadados Geoespaciais do Brasil. *In*: SIMPÓSIO BRASILEIRO DE GEOINFORMÁ-

TICA, 20., 2019, São José dos Campos. **Anais** [...]. São José dos Campos: [*s. n.*], 2019. p. 218-223. Disponível em: http://mtc-m16d.sid.inpe.br/col/sid.inpe.br/mtc-m16d/2019/11.27.18.28/doc/218-223.pdf. Acesso em: 1 mar. 2024.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008. Disponível em: https://nlp.stanford.edu/IR-book/information-retrieval-book.html. Acesso em: 1 mar. 2024.

MOREAU, L.; MISSIER P. (ed.).**PROV-DM**: the PROV Data Model. [*s. l.*]: W3C Recommendation, 30 Apr. 2013. Disponível em: https://www.w3.org/TR/2013/REC-prov-dm-20130430/. Acesso em: 27 fev. 2024.

PALMER, C. L. *et al*. Data curation for the long tail of science: the case of environmental sciences. *In*: INTERNATIONAL DIGITAL CURATION CONFERENCE, 3., 2007, Washington. **Anais** [...]. Washington: [s.n.], 2007. p. 1-5.

POMERANTZ, J. **Metadata**. Cambridge: The MIT Press, 2015. (The MIT Press essential knowledge Series).

POMERANTZ, J.; PEEK, R. Fifty shades of open. **First Monday**, [*s. l.*], v. 21, n. 5, 2 May 2016. DOI: http://doi.org/10.5210/fm.v21i5.6360. Disponível em: https://firstmonday.org/ojs/index.php/fm/article/view/6360. Acesso em: 15 jun. 2025.

RENEAR, A. H.; PALMER, C. L. Strategic reading, ontologies, and the future of scientific publishing. **Science**, [*s. l.*], v. 325, n. 5942, p. 828-832, Aug. 2009. DOI: http://doi.org/10.1126/science.1157784. Disponível em: https://www.science.org/doi/10.1126/science.1157784. Acesso em: 15 jun. 2025.

RILEY, J. **Understanding metadata**: what is metadata, and what is it for? Baltimore, MD: NISO, 2017. Disponível em: https://digital.library.unt.edu/ark:/67531/metadc990983/m2/1/high_res_d/understanding_metadata.pdf. Acesso em: 27 fev. 2024.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. tradução Regina Célia Smille. 3. ed. Rio de Janeiro: Elsevier: Campus, 2013.

SEGARRA-FAGGIONI, V.; ROMERO-PELAEZ, A. Automatic classification of OER for metadata quality assessment. *In*: INTERNATIONAL CONFERENCE ON ADVANCED LEARNING TECHNOLOGIES, 2022, Bucharest. **Anais** [...]. Bucharest: IEEE, 2022. p. 16-18. DOI: http://doi.org/10.1109/ICALT55010.2022.00011. Disponível em: https://ieeexplore.ieee.org/document/9853751. Acesso em: 15 jun. 2025.

SILVA, F. C. C. **Gestão de dados científicos**. Rio de Janeiro: Interciência, 2019.

TRIQUES, M. L.; ARAKAKI, A. C. S.; CASTRO, F. F. Aspectos da representação da

informação na curadoria digital. **Encontros Bibli**: revista eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 25, p. 1-21, 2020. DOI: https://doi.org/10.5007/1518-2924.2020.e69898. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e69898. Acesso em: 15 jun. 2025.

YUAN, S. *et al*. Automatic Metadata Extraction for Educational Resources Based on Citation Chain. *In*: WU, Y. (ed.). **Advanced Technology in Teaching:** proceedings of the 2009 3rd International Conference on Teaching and Computational Science (WTCS 2009). Berlin: Springer, 2009. (Advances in Intelligent and Soft Computing, v. 116). DOI: http://doi.org/10.1007/978-3-642-11276-8_14. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-11276-8_14. Acesso em: 15 jun. 2025.

ZENG, M. L.; QIN, J. **Metadata**. 2nd ed. Chicago: Neal-Schuman Publishers, 2016.

Como citar este capítulo:

LIMA, Karolayne Costa Rodrigues de; SUNYE, Marcos Sfair. Proposta de automação da curadoria de metadados com machine learning. *In*: ARAÚJO, Paula Carina de; LIMA, Karolayne Costa Rodrigues de (org.). **Práticas de ciência aberta.** Brasília, DF: Editora Ibict, 2025. Cap. 5, p. 96-117. DOI: 10.22477/9788570131966.cap5.