DOI: 10.5433/1981-8920.2024v29n1p68

FERRAMENTAS DE GERAÇÃO AUTOMÁTICA E SEMIAUTOMÁTICA DE METADADOS: UMA REFLEXÃO ENTRE OS ANOS DE 2010 A 2020

AUTOMATIC AND SEMI-AUTOMATIC METADATA GENERATION TOOLS: A REFLECTION BETWEEN THE YEARS 2010 AND 2020

Jean Carlos Borges Brito[a]
Dalton Lopes Martins[b]

RESUMO
Objetivo: Identificar as possibilidades e limitações para utilização das ferramentas analisadas. **Metodologia:** Trata-se de uma investigação exploratória, executando uma revisão bibliográfica e cuja busca foi realizada nas bases de dados Scopus, Web Of Science, ISTA, LISTA e LISA. Utilizou-se método misto na análise dos dados, com abordagens quantitativas e qualitativas. Foram encontrados 49 artigos científicos e após a aplicação dos critérios adotados, apenas 12 foram selecionados para a síntese. **Resultados:** Os resultados demonstraram diversas ferramentas e soluções para geração de metadados, utilizando variadas técnicas, métodos e funções, abordando sua implementação e uso. Nesse contexto, identificou-se as possibilidades e limitações dessas soluções, com a finalidade de contribuir para sua aplicação e aprimoramento em pesquisas futuras. **Conclusões:** Conclui-se que as ferramentas de geração automática e semiautomática de metadados são instrumentos que podem auxiliar os profissionais da informação na gestão organizada e eficiente dos acervos digitais, melhorando a recuperação da informação, o que reforça a contribuição dessa pesquisa no meio acadêmico-científico na área da Ciência da Informação.

Descritores: Revisão bibliográfica. Metadados. Geração automática e semiautomática. Possibilidades. Limitações.

1 INTRODUÇÃO

A transformação digital tem levado a uma produção massiva de dados e informações que são armazenadas em repositórios digitais em todo o mundo. Com a ascensão da tecnologia da informação e da comunicação (TIC), cada vez

[a] Doutor em Ciência da Informação pela Universidade de Brasília (UnB), Brasília, Brasil. E-mail: zuluaer@gmail.com

[b] Doutor em Ciência da Informação pela Universidade de São Paulo (USP). Docente na Universidade de Brasília (UnB), Brasília, Brasil. E-mail: daltonmartins@unb.br

mais dados são gerados por meio de dispositivos digitais, como *smartphones*, *laptops*, *tablets* e outros dispositivos inteligentes conectados à internet.

A importância dos dados digitais tem levado muitos governos, empresas e organizações a investir em tecnologia para gerenciar, armazenar e analisar grandes quantidades de informações. Bancos de dados, sistemas de gerenciamento de conteúdo e armazenamento em nuvem são algumas das tecnologias mais utilizadas para gerenciar esses dados.

No entanto, a quantidade de dados gerados diariamente pode ser esmagadora e, sem os sistemas certos para auxiliar nesta gestão, as informações podem se tornar inúteis ou fornecerem compreensões equivocadas.

Estudo publicado por Reinsel, Gantz e Rydning (2018) para o *International Data Corporation* (IDC), prevê que o crescimento de dados aumentará de 45 *zettabytes* em 2019 para cerca de 175 *zettabytes* até 2025. Essa informação demonstra que em cinco anos, 6 bilhões de pessoas ou 75% da população mundial interagirão com dados todos os dias e cada pessoa conectada terá pelo menos uma interação com dados a cada 18 segundos. Conforme o IDC, grande parte da economia atual depende de dados e essa dependência só aumentará no futuro à medida que as entidades capturam, catalogam, gerenciam e analisam os seus dados a partir de processos e tecnologias que aumentam a qualidade dos dados e permitam melhor exploração de seu valor agregado.

A utilização de abordagens de processos automatizados pode melhorar a eficiência das atividades de classificação, catalogação e indexação de assuntos com o uso de metadados nos repositórios digitais, disponibilizando em tempo real aquilo que é feito tradicionalmente de forma manual.

Com a quantidade crescente de dados gerados diariamente, tornam-se cada vez mais complexas e demoradas a sua classificação e organização. A utilização de sistemas automatizados pode agilizar esses processos e reduzir a margem de erro, permitindo que as informações sejam disponibilizadas em tempo real. Uma das abordagens mais comuns para automatizar esses processos é o uso de metadados.

Ulrich *et al*. (2022), afirmam que os metadados são informações criadas para descrever de forma detalhada e exclusiva os dados correspondentes,

servindo vários casos de uso em diferentes áreas de pesquisa, tais como: identificação de dados, classificação, recuperação e validação de um conjunto de dados.

De acordo com Mooers (1951), um usuário potencial da informação é capaz de converter sua necessidade de informações em uma lista de referências (metadados) para documentos armazenados e que contém informações úteis, auxiliando na sua obtenção e recuperação.

Crystal e Land (2003) discorrem que para criar metadados para um milhão de documentos deveriam ser alocados 60 empregados/ano para realizar essa tarefa. É considerado um trabalho árduo, lento e caro se executado manualmente, continuam esses autores. Além disso, considerando que o conceito de documento e suas possibilidades de expressão midiática se expandem de forma significativa na era da web, torna-se proibitivo imaginar que a catalogação dos documentos seguirá continuamente sendo realizada apenas de forma manual.

Polfreman, Broughton e Wilson (2008) elenca seis técnicas para geração de metadados, posteriormente investigadas por Park e Brenza (2015), conforme Quadro 1:

Quadro 1 – Técnicas empregadas para geração de metadados

Técnica Empregada	Descrição
Colheita de *Metatags*	Definida como processo de computação em que os valores para os campos de metadados são identificados e preenchidos por meio de um exame de *tags* de metadados em um documento ou anexado a ele, sendo a forma mais comum e simples de coleta por meio de *metatags* existentes em HTML e *tags* <meta>, onde várias soluções utilizam essa abordagem. A limitação se restringe na qualidade das *metatags* do documento, o que impacta sua efetividade. Esse tipo de ferramenta não é útil para gerar valores automaticamente de metadados para propriedades que ainda não foram descritas, devendo recorrer a outras soluções para atender essa necessidade.
Extração de Conteúdo	Técnica que aborda a captura de palavras e frases do corpo de um recurso de informação e as utilizam para fornecimento de metadados estruturados (rótulos) com a finalidade de representar o objeto. Ferramentas desenvolvidas para fazer uso dessa abordagem, utilizam uma mistura de técnicas baseadas em regras e estatísticas. A principal vantagem desse tipo de técnica é que a extração de metadados pode ser feita independentemente da qualidade dos metadados associados a

	qualquer recurso de informação.
Indexação ou Classificação Automática	Envolve o uso de aprendizado de máquina e algoritmos baseados em regras para extrair valores de metadados dos próprios recursos de informação, em vez de depender do conteúdo das *metatags* aplicadas aos recursos. Essa técnica também abrange o mapeamento de termos de metadados extraídos para vocabulários controlados.
Mineração de textos e dados	Faz uso de aprendizado de máquina, análise estatística, técnicas de modelagem e tecnologia de banco de dados, para processar grandes quantidades de dados e identificar padrões recorrentes. Esta é uma técnica complexa de implementação porque depende da qualidade e quantidade dos dados para desenvolver um modelo e usá-lo para treinar o sistema. Devido a essas características, poucas ferramentas foram totalmente desenvolvidas para aplicação em repositórios digitais.
Folksonomia ou marcação social	Confiam nas etiquetas geradas pelo autor e pelo usuário (na verdade, nos termos do assunto) para classificar os recursos. À medida que os recursos são acessados e compartilhados por outras pessoas, o vocabulário é usado e adicionado de maneira colaborativa. Podemos verificar algumas aplicações atuais utilizando essa técnica em ferramentas de redes sociais, tais como *Facebook* e *Instagram*. Apesar do seu valor do ponto de vista do usuário, é improvável que a *folksonomias* substitua inteiramente os vocabulários controlados por causa de sua falta de precisão ou autoridade.
Geração automática de metadados extrínsecos	Processo de extrair metadados sobre um recurso de informação que não está contido no próprio recurso, ou seja, um exemplo seria extrair metadados técnicos, como o formato e tamanho do arquivo, mas também pode incluir a extração de recursos mais complicados, como o nível de nota de um recurso educacional ou o público-alvo de um documento

Fonte: Polfreman, Broughton e Wilson (2008) e Park e Brenza (2015).

Nesse cenário, termos como geração automática e semiautomática de metadados têm sido utilizados com frequência na comunidade científica, propiciando novos desafios para a atuação dos gestores de repositórios digitas no panorama de uso dessas ferramentas para execução de tarefas de catalogação, classificação e indexação.

Diante dos fatos apresentados, este artigo tem como objetivo realizar uma reflexão sobre a geração de metadados. Realizaram-se buscas nas bases de dados *Scopus*, *Web of Science*, ISTA, LISTA e LISA, com a finalidade de subsidiar uma revisão de literatura sobre as ferramentas de geração automática e semiautomática de metadados, identificando suas técnicas, funções, e por fim, suas possibilidades e limitações. Tal resultado propiciará identificar oportunidades de uso e pontos de melhorias dessas soluções com o objetivo de

apoiar e fornecer melhor eficiência no trabalho dos profissionais da informação.

2 METODOLOGIA

O propósito desta pesquisa foi alcançar o objetivo específico de realizar uma reflexão sobre a geração de metadados, por meio do levantamento e análise das ferramentas automáticas e semiautomáticas, identificando suas técnicas, funções, considerando suas possibilidades e limitações.

Nesse contexto, o método adotado foi o de pesquisa bibliográfica, executando a exploração de estudos previamente conduzidos e reconhecidos por sua relevância, utilizando como fontes de conhecimento: artigos, periódicos e revistas científicas, a fim de obter informações atuais e pertinentes relacionadas ao tópico em questão (Marconi; Lakatos, 2003).

2.1 PLANEJAMENTO DA REVISÃO DE BIBLIOGRÁFICA

Inicialmente, elencaram-se questões de *background* para fornecer a compreensão básica e conceitual do tema. Em seguida, para melhor definição do escopo, estabeleceu-se questões de *foreground*, conforme Quadro 2:

Quadro 2 – Questões de background e foreground

Questões de *Background*	Questões de *Foreground*
O que são metadados? Para que servem? O que é geração automática e semiautomática de metadados?	Quais técnicas, características, funções, ferramentas e aplicações da geração automática e semiautomática de metadados? Suas possibilidades e limitações?

Fonte: Elaborado pelos autores.

Delimitou-se a seguinte questão de pesquisa: “Como as ferramentas de geração automática e semiautomática de metadados podem auxiliar os gestores de repositórios digitais em suas atividades, considerando as possibilidades e limitações dessas ferramentas?”.

Após a delimitação da questão de pesquisa, definiu-se as bases de dados que foram consultadas para obtenção de artigos. São apresentadas a seguir: Base de Dados Referenciais de Artigos de Periódicos em Ciência da

Informação (BRAPCI); *Library and Information Science Abstract* (*LISA*); *Library, Information Science & Technology Abstracts* (*LISTA*); *Emerald Publishing Limited; Information Science and Technology Abstracts* (*ISTA*)*; Wiley Online Library; Web of Science* e *Scopus*.

Elaborou-se as seguintes terminologias para compor as *strings* de busca, utilizando-se termos em português e inglês: “Geração automática de metadados”, “*Automatic metadata generation*”, “Geração semiautomática de metadados”, “*Semiautomatic metadata generation*”, “Ferramentas”, “*Tools*”, “Técnicas”, “*Techniques*”, “Características”, “*Features*”, “Funções”, “*Functions*”, “Aplicações” e “*Applications*”. Sabe-se que outros descritores poderiam trazer trabalhos eventualmente relacionados com o tema, como "indexação automática", por exemplo. No entanto, optou-se nessa pesquisa por analisar a literatura científica com foco no problema mais amplo, ou seja, a geração de metadados sejam eles descritivos ou temáticos. Em próximos trabalhos, pretende-se aprofundar em termos derivados.

Próximo passo, com o auxílio de um bibliotecário, foi a construção das sentenças de buscas utilizando os operadores lógicos, sendo definidas:

- Sentença em Português: ((("Geração automática de metadados") *OR* ("Geração semiautomática de metadados")) *AND* (Ferramentas *OR* Técnicas *OR* Características *OR* Funções *OR* Aplicações));
- Sentença em Inglês: ((("*Automatic metadata generation*") *OR* ("*Semiautomatic metadata generation*")) *AND* (*Tools OR Techniques OR Features OR Functions OR Applications*)).

Realizou-se atividades de pré-testes nas bases científicas para verificar se as sentenças deveriam passar por um processo de readequação. Concluído o pré-teste, finalizou-se a fase de planejamento e após a aprovação por especialista desta etapa, passou-se a execução da pesquisa.

2.2 Execução Da Revisão De Bibliográfica

Executou-se buscas nas bases científicas definidas na etapa de planejamento, realizando o preenchimento dos filtros de busca de forma a

delimitar a recuperação da informação. Aplicou-se os seguintes critérios de inclusão e exclusão, definidos no Quadro 3, para obter as publicações:

Quadro 3 – Critério de inclusão e exclusão

Critérios de Inclusão	Critérios de Exclusão
Trabalhos científicos publicados entre os anos de 2010 e 2020	Trabalhos científicos publicados antes do ano de 2010 e depois de 2020
Descrição de estudo de caso, experimentos ou *survey, outras RSL*	Título do trabalho não condizente com a proposta do projeto
Resumo ou *abstract* condizente com a proposta de pesquisa	Resumo ou *abstract* com fuga ao tema proposto na pesquisa
Artigos e periódicos	Publicações sem cunho científico
Publicações em inglês e português com disponibilidade completa e suporte em meio eletrônico	Disponibilização de partes da pesquisa, textos incompletos

Fonte: Elaborado pelos autores.

Duas bases de dados não retornaram resultados com as sentenças definidas na etapa de planejamento, sendo elas a BRAPCI e *Wiley Online Library*. Resolveu-se desmembrar as sentenças compostas em termos simples de busca, mas obteve-se zero resultado na recuperação de informação nesses dois repositórios com os termos definidos. As demais bases de dados pesquisadas retornaram o total de 49 trabalhos científicos, utilizando as sentenças definidas no planejamento.

A atividade seguinte consistiu na leitura do título e descrição dos registros por meio da busca nas bases de dados, identificando 15 artigos duplicados.

Posteriormente, passou-se a leitura do resumo e palavras-chave, excluindo 16 artigos por não possuírem relação com a temática da pesquisa. O total de 18 estudos foram selecionados para execução da análise. Deste total, 6 artigos completos foram excluídos após uma leitura preliminar, pois o conteúdo dos trabalhos não correspondia ao objeto da pesquisa, conforme os critérios de inclusão/exclusão estabelecidos. Por fim, 12 estudos foram incluídos na síntese.

A documentação deste processo pode ser acessada por meio dos Critérios de Qualidade e categorias de análise[1]. Todas as tabulações das publicações foram realizadas utilizando planilha eletrônica Microsoft Excel® para

[1] Disponível em: https://cutt.ly/tWiDg3k

apoiar a organização deste estudo. Executou-se *download* do arquivo e armazenado em repositório para fichamento e síntese, além de demonstrar no Quadro 4.

Quadro 4 – Artigos selecionados para realização de fichamento e síntese

ANO	AUTORES	TÍTULO	INSTITUIÇÃO / LOCAL / PAÍS	CATEGORIA
2011	Kovaevic et alAUDICHYA.	*Automatic extraction of metadata from scientific publications for CRIS systems*	Novi Sad *University* / Novi Sad / Sérvia	Ferramentas de geração automática de metadados
2012	Maratea A; Petrosino A; Manzo, M	*Automatic Generation of SCORM Compliant Metadata for Portable Document Format Files*	*Parthenope University* /Nápoles / Itália	
2012	Verborgh et al.	*Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform*	*Ghent University* / Ghent / Bélgica	
2012	Sah, M; Wade, V	*Automatic metadata mining from multilingual enterprise content*	Trinity *College* Dublin / Dublin / Irlanda	
2013	Costa *et al.*	*EURAC SDI: A Near Real Time and Offline Automatic Metadata Generation Processing Chain*	*Eurac Research* / Bozen / Itália	
2013	Vlachidis *et al.*	*Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature*	*University College* London / Londres / *United Kingdom*	
2015	Rafferty, J; Nugent, C; Liu, J	*Automatic Metadata Generation Through Analysis of Narration Within In structional Videos*	Ulster University / Belfast / Irlanda do Norte	
2018	Gonzalo *et al.*	*ScienceSearch: Enabling Search through Automatic Metadata Generation*	Berkeley *University* / Califórnia / EUA	
2018	Yang, G; Park, J	*Automatic Extraction of Metadata*	Drexel *University* e	

		Information for Library Collections	Mokpo *University* / Filadélfia e Mokpo / EUA e Coréia do Sul	
2019	Audichya, M, K; Saini J, R	*Computational linguistic prosody rule-based unified technique for automatic metadata generation for Hindi poetry*	Gujarat *Technological University* / Gujarat / Índia	
2020	Morris, V	*Automated Language Identification of Bibliographic Resources*	*British Library / Wetherby / United Kingdom*	
2015	Park, J; Brenza, A	*Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art*	Drexel *University* / Filadélfia / EUA	Ferramentas de geração semiautomática de metadados

Fonte: Dados da pesquisa.

Os artigos elencados no Quadro 4 demonstram que o tema pesquisado dentro do recorte temporal de 2010 a 2020 foi relevante e de interesse ao redor do mundo, pois se verificam investigações em diferentes universidades localizadas nos Estados Unidos, Europa e Ásia.

3 ANÁLISES E RESULTADOS

Para delinear melhor a compreensão do leitor, a síntese foi desenvolvida inicialmente abordando as ferramentas de geração automática de metadados, categorizando pelo tipo de técnica empregada, conforme Polfreman, Broughton e Wilson (2008). Posteriormente, abordaram-se os assuntos relativos à ferramenta de geração semiautomática de metadados, as possibilidades e limitações apresentadas.

3.1 FERRAMENTAS DE GERAÇÃO AUTOMÁTICA DE METADADOS – GAM

Após a introdução de documentos digitais na segunda metade do século XX, a área de Biblioteconomia tem presenciado o potencial uso da geração automática de metadados (GAM) como meio de simplificar o processo de descrição dos recursos de informação (Kleppe *et al.*, 2019).

As ferramentas de geração automática de metadados não necessitam de intervenção humana, pois os algoritmos se encarregam de realizar a ação de geração de metadados automaticamente, conforme as regras de negócio implementadas no *software* com o uso de inteligência artificial e técnicas de aprendizagem de máquina. Serão apresentadas as análises das ferramentas GAM, conforme as técnicas empregadas.

3.1.1 Colheita de metatags

Sah e Wade (2012) investigam a utilização de ferramentas de geração automática de metadados para fornecer informações avançadas de conteúdos acessados pelos usuários, fomentando aspectos de personalização do cliente, fazendo com que eles permaneçam mais tempo no site, incentivando-os a retornar ao provedor de serviços. Os autores desenvolveram uma ontologia *DocBook* e ontologia de tipo de recurso para extrair metadados estruturais e descritivos dos documentos *DocBook* no formato RDF.

3.1.2 Extração de conteúdo

Kovacevic *et al.* (2011) apresenta um método para a extração automática de metadados de artigos científicos em formato PDF, que é projetado como parte integrante do sistema de informação para monitorar as atividades de pesquisa. O método é implementado como um complemento à entrada manual de metadados, no sentido de que os resultados da extração são oferecidos ao curador para inspecionar e corrigir antes de armazená-los no repositório. O sistema é baseado no método de aprendizado de máquina denominado classificação e obteve-se melhor resultado com o uso do *modelo Support Vector*

Machines.

Vlachidis *et al.* (2013) realizam investigações sobre bibliotecas digitais, em especial a Europeana, tendo como objetivo a execução da geração automática de metadados com enriquecimento semântico significativo para seus objetos digitais vinculados, por meio do Europeana *Data Model* (EDM), que resume o *Cidoc Conceptual Reference Model* (CRM).

A pesquisa de Rafferty, Nugent e Liu (2015) apresentam um mecanismo de geração de metadados a partir de análises de clipes de vídeo, sendo que esses metadados devem ser usados no suporte ao fornecimento de instruções dinâmicas dentro de um paradigma *Smart Home*. Os autores utilizaram um método de anotação capaz de gerar metadados enriquecidos para videoclipes, sendo criado e implementado dentro de uma plataforma de avaliação chamada *Audio BaSEd Instruction ProfiLer* (ABSEIL). Esta plataforma destina-se a trabalhar em conjunto com o repositório de vídeo gerado pelo projeto *Personal IADL Assistant* (PIA). O objetivo do projeto PIA é ajudar os idosos, oferecendo orientação com atividades instrumentais da vida diária, tais como: preparação de refeições, como utilizar o controle remoto de uma TV, como se barbear, limpar e manter uma casa etc.

Os estudos de Yang e Park (2018) têm como objetivo apresentar um mecanismo de extração automática para atenuar problemas relacionados à aplicação inconsistente de metadados e à interoperabilidade semântica entre as coleções digitais. Eles sugerem a construção de gráficos conceituais, pois eles têm um bom potencial para facilitar a interpretação adequada dos conceitos de metadados e o uso preciso e consistente dos elementos de dados. Os autores demonstram um mecanismo de extração automática para coleções de bibliotecas chamado *ExMETA* que foi projetado utilizando gráficos conceituais como representação interna. A ferramenta é capaz de analisar sentenças em linguagem natural e gerar metadados descritivos e estruturais, permitindo a eliminação de intervenções humanas (ou seja, catalogadoras) que geralmente causam a atribuição incorreta e o mapeamento inconsistente de metadados no processo de produção padrão, como o *Dublin Core*, a partir de dados brutos de coleções digitais.

3.1.3 Indexação ou classificação automática

Gonzalo *et al.* (2018) apresenta um estudo sobre o *ScienceSearch*, uma infraestrutura de pesquisa escalável generalizada que utiliza o aprendizado de máquina para capturar metadados de dados, contexto e artefatos circundantes. A implementação se concentrou no conjunto de dados do Centro Nacional de Microscopia Eletrônica, unidade do Departamento de Energia do Laboratório Nacional *Lawrence Berkeley*. Esses dados possuem milhões de micrografias produzidas por centenas de cientistas. A problemática identificada foi que os arquivos de micrografia gerados, raramente incluem metadados além das configurações de captura do microscópio (por exemplo: exposição, contraste, tensão do sinal). A avaliação de desempenho mostrou que o *ScienceSearch* é capaz de executar consultas simples em um único nó, em mais de 11 milhões de *tags* de metadados em menos de cinco segundos.

A pesquisa de Audichya e Asini (2019) teve como objetivo estruturar e padronizar adequadamente o conhecimento disperso sobre a prosódia, denominada de *Hindi Poetries*, disponível de maneira deficiente ou contraditória em diferentes fontes de informação. Foi utilizada a técnica de linguística computacional e o trabalho de pesquisa também se concentrou em moldar, um conjunto de regras padronizadas, para a geração automática de metadados. Os autores testaram um gerador de metadados em 3.026 entradas que incluem diferentes poemas e parte de poemas que cobriam mais de 30 *Chhands*[2]. O resultado do trabalho foi suficiente para provar a robustez da metodologia e do mecanismo técnico do gerador de metadados, que alcançou 98,09% de taxa de sucesso, juntamente com 1,91% de falha devido a erros de formatação e uso irregular de delimitadores.

3.1.4 Mineração de textos e dados

Maratea, Petrosino e Manzo (2012) utilizaram algoritmos com técnicas de

[2] Joshi e Kushwah (2018). *Chhand* é uma estrutura textual em que os poemas eram escritos nas tradições poéticas da Índia antiga.

processamento de linguagem natural para geração automática de metadados para conteúdo de aprendizagem. Aplicou-se como padrão uma coleção de especificações para o *e-learning* baseado na *web* amplamente adotado, denominado Modelo de Referência de Objeto Compartilhável para Conteúdo (*SCORM*). O objetivo deste modelo é permitir a interoperabilidade, fácil acesso e reutilização de unidades de aprendizagem baseadas na *web* para indústria, governo e universidade.

Costa *et al.* (2013) apresentam em seu estudo a abordagem para geração automática de metadados por meio de um método baseado em regras, codificado manualmente, implementados como plugins que geram metadados em formato padronizado extraído de um conjunto heterogêneo de dados geoespaciais. Os autores discorrem que a estação receptora *EURAC* recebe diariamente dados brutos das missões da NASA: Aqua, Terra e Suomi NPP. A pesquisa do instituto lida com muitos dados de satélite diferentes: *LANDSAT*[3], *RapidEye*[4], *ENVISAT*[5] e *Quickbird*[6]. Eles executaram a ingestão e manuseio de dados, por meio de um aplicativo geral multitarefa centralizado, denominado *Data Exchange Server* (DES).

3.1.5 *Folksonomia* ou marcação social

Verborgh *et al.* (2012) apresentam uma plataforma genérica de solução de problemas semânticos, que combina automaticamente os serviços da *web* para realizar uma tarefa predefinida, usando a websemântica como fonte de conhecimento para iniciar e manter ativamente o contexto da tarefa. Os autores executaram a aplicação por meio de um caso de uso de anotação de imagem.

[3] *Landsat* são satélites para observação de recursos naturais da terra.

[4] *RapidEye* é uma constelação de 5 microsatélites que produzem um conjunto de imagens de qualquer ponto da Terra.

[5] *Envisat* é um satélite lançado em 2002 pela Agência Espacial Européia (ESA) e seu objetivo foi fornecer dados da atmosfera, do oceano, da terra e do gelo, visando o monitoramento do aquecimento global, do grau de contaminação atmosférica e dos riscos de desastres naturais.

[6] *Quickbird* é um conjunto de satélites mantido pela empresa *DigitalGlobe*, fornecendo imagens comerciais de alta resolução e são aplicadas na área de mapeamentos urbanos e rurais que requerem alta precisão dos dados (cadastro, redes, planejamento, telecomunicações, saneamento, transportes), além de aplicações voltadas à área ambiental, dinâmica de uso e cobertura das terras, agricultura e recursos florestais. (Embrapa, 2024).

Obtiveram resultados satisfatórios quanto à eficiência e escalabilidade, sendo utilizada a ferramenta de raciocínio *Eye* para resolução de problemas semânticos e capaz de criar composições holísticas.

3.1.6 Geração automática de metadados extrínsecos

Morris (2020) investigou se os códigos de idioma podem ser atribuídos automaticamente aos registros do *Machine Readable Catalog* – MARC, mecanismo criado para possibilitar a padronização de catalogação de registros bibliográficos em todos os sistemas virtuais de publicações. O objetivo foi avaliar a precisão de qualquer método viável de fazê-lo. A autora enfatiza que nos registros bibliográficos da versão MARC21, o conteúdo de idioma de um recurso de informação é registrado nas posições do campo 008 35–37, usando um código de três posições de uma lista controlada. Entretanto, uma análise do catálogo da *British Library* em outubro de 2018 revelou que esse código de idioma não era preenchido em quase 4,7 milhões de registros. Desses, 78% também não possuíam um código para o local de publicação no campo 008 posições 15–17. O uso de ferramentas de geração automática de metadados auxilia como solução para a problemática descrita por Morris.

3.2 Ferramentas De Geração Semiautomática De Metadados – GSAM

O artigo de Park e Brenza (2015) merece um destaque especial, pois essa publicação foi a mais indexada nas bases pesquisadas. Eles examinam uma variedade de ferramentas de geração semiautomática de metadados (N=39), analisando suas técnicas, recursos e funções. Esses autores desenvolveram uma matriz caracterizando cada ferramenta de geração semiautomática de metadados analisada: nome, local online, técnicas usadas para geração de metadados (Greenberg, 2004; Polfreman; Broughton; Wilson, 2008), além de breve descrição das funções e recursos da ferramenta, conforme Quadro 5:

Quadro 5 – Ferramentas de geração semiautomática de metadados

Ferramenta	Técnicas Empregadas	Características e Funções

Módulo ANVL Perl https://metacpan.org/pod/File::ANVL	Colheita de metatags.	Utilitário para ler e escrever arquivos no formato *Attribute-Name-Value List* (ANVL), que é um formato de dados simples e estruturado. Ele facilita a manipulação de dados baseados em pares de "atributo-nome-valor", sendo ideal para tarefas de configuração e armazenamento de dados de forma legível e organizada.
Apache POI – Text Extractor http://poi.apache.org/download.html	Extração de conteúdo; Colheita de *metatags;* Geração automática extrínseca.	Solução disponibilizada para extrair texto e dados de arquivos do *Microsoft Office*, como documentos *Word* e planilhas *Excel*. A ferramenta facilita a leitura e manipulação de conteúdo desses arquivos em projetos de *software*, sem precisar do *Microsoft Office* instalado.
Apache Stanbol (movido para *Apache Attic* em 2020) https://attic.apache.org/projects/stanbol.html	Extração de conteúdo; Indexação automática.	Ferramenta para enriquecimento semântico e gerenciamento de metadados de conteúdo. Ele fornecia funcionalidades para adicionar e manipular metadados semânticos em documentos, facilitando a integração e a análise de informações de forma mais inteligente. Como foi movido para o *Apache Attic*, o projeto está desativado e não recebe mais suporte ativo.
Apache Tika http://tika.apache.org/	Extração de conteúdo; Colheita de metatags; Geração automática extrínseca.	Utilitário que detecta e extrai texto e metadados de diversos tipos de documentos, como PDFs, arquivos do *Microsoft Office* e outros formatos. Ele facilita a análise e o processamento de conteúdo de arquivos ao converter e extrair informações de forma consistente.
Ariadne *Knowledge Pool System* https://sourceforge.net/projects/ariadnekps/	Colheita de metatags.	Plataforma de código aberto para gerenciar e compartilhar conhecimento. Ela fornece funcionalidades para armazenar, organizar e acessar informações de forma colaborativa, facilitando a gestão e a disseminação de conhecimento dentro de uma organização.
Bibframe *Tools* https://www.loc.gov/bibframe/implementation/	Colheita de metatags.	Conjunto de ferramentas projetadas para ajudar na implementação e conversão de dados bibliográficos para o formato BIBFRAME (*Bibliographic Framework*). Facilita a transição de registros bibliográficos para uma estrutura baseada em *Linked Data*, melhorando a interoperabilidade e a busca de informações bibliográficas.
Biblio *Citation Parser* https://metacpan.org/rele	Extração de conteúdo.	Conjunto de módulos que analisa e extrai informações de citações bibliográficas.

ase/MJEWELL/Biblio-Citation-Parser-1.10		Facilita a interpretação e o processamento de referências em diferentes formatos, permitindo a extração de dados como autores, títulos e datas de publicação.
CatMDEdit https://catmdedit.sourceforge.io/	Extração de conteúdo.	Ferramenta de edição de metadados projetada para documentação de recursos, com foco especial em informação geográfica. Ela oferece uma interface intuitiva para criar e gerenciar metadados e dentre suas funcionalidades estão a edição de registros de metadados, validação de conformidade com padrões geoespaciais e a capacidade de importar e exportar dados em diversos formatos.
CrossRef https://doi.crossref.org/simpleTextQuery	Extração de conteúdo.	Ferramenta que permite a busca e recuperação de informações bibliográficas a partir de textos simples. Ele usa o sistema *Digital Object Identifier* (DOI) para identificar e acessar metadados relacionados a publicações acadêmicas, como artigos e livros, facilitando a integração e o gerenciamento de referências em diversos contextos de pesquisa.
Data Fountains https://www.onworks.net/software/linux/app-data-fountains	Extração de conteúdo; Indexador automático; Colheita de metatag; Geração automática extrínseca.	Ferramenta projetada para facilitar a coleta, integração e análise de dados de várias fontes. Ela permite que usuários criem fluxos de dados, transformando informações de diferentes formatos em um formato unificado para análise.
Digital Record Object Identification **(DROID)** https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/	Geração automática extrínseca.	Ferramenta que auxilia na identificação e catalogação dos arquivos digitais, ajudando na gestão e preservação ao fornecer informações detalhadas sobre os tipos de arquivos e suas versões. A solução ajuda ainda a garantir a integridade e a acessibilidade dos arquivos ao longo do tempo.
Dublin Core Meta Toolkit https://sourceforge.net/projects/dcmetatoolkit/	Colheita de metatags.	Ferramenta para criar, gerenciar e validar metadados no formato *Dublin Core* (DC). Facilita a aplicação dos padrões DC para a descrição de recursos digitais, permitindo a organização e a busca de informações de forma estruturada e padronizada. Inclui ainda, funcionalidades para a edição, validação e exportação de metadados,

		ajudando na interoperabilidade e na gestão eficiente de recursos digitais.
Dspace https://duraspace.org/dspace/	Colheita de metatags; Geração automática extrínseca; Marcação Social.	Plataforma de código aberto para a gestão e preservação de repositórios digitais. Permite a criação, organização e disseminação de coleções digitais, como artigos acadêmicos, teses e outros materiais de pesquisa. O *DSpace* oferece funcionalidades para armazenar, catalogar e acessar documentos digitais de maneira eficiente, promovendo a preservação a longo prazo e o acesso aberto ao conhecimento.
Editor-Converter Dublin Core Metadata https://old.library.kr.ua/dc/dcedituni.html	Colheita de metatags; Geração automática extrínseca.	Ferramenta projetada para criar, editar e converter metadados no formato *Dublin Core*. A ferramenta permite a manipulação de metadados, onde o usuário pode gerenciar e integrar informações de forma padronizada e eficiente.
Embedded Metadata Extraction Tool (EMET) https://sourceforge.net/projects/emet/	Extração de conteúdo; Colheita de metatag; Geração automática extrínseca.	Projetada para extrair metadados incorporados em imagens digitais. A ferramenta analisa documentos, imagens e outros tipos de arquivos para identificar e recuperar informações ocultas, como autor, data de criação e outros detalhes relevantes que estão embutidos no próprio arquivo.
Firefox Dublin Core Viewer Extension https://www.splintered.co.uk/experiments/73/ (Plugin descontinuado)	Colheita de metatag; Geração automática extrínseca.	Era um *plugin* para o navegador Firefox que permitia visualizar metadados no formato *Dublin Core* (DC) diretamente nas páginas da *web*. Quando ativado, o *plugin* exibiria informações descritivas sobre o conteúdo de uma página, como título, autor e data de criação, de acordo com os padrões *DC*. Entretanto, o plugin foi descontinuado e não está mais disponível.
FreeCite (descontinuado) Foi substituído pela ferramenta: ***AnyStyle*** https://anystyle.io/	Extração de conteúdo.	*AnyStyle* oferece funcionalidades aprimoradas para a extração, formatação e gerenciamento de citações bibliográficas. Ela permite a importação de referências de diversos formatos e gera citações e bibliografias de acordo com diferentes estilos, facilitando a organização e o uso de referências em trabalhos acadêmicos e pesquisas.
General Architecture for Text Engineering **(GATE)** https://gate.ac.uk/overview.html	Extração de conteúdo; Indexação automática.	Plataforma de código aberto para processamento e análise de texto. Fornece um ambiente robusto para a criação de aplicações de processamento de linguagem natural (PLN) e engenharia

		de texto. O GATE permite a extração de informações, a análise de sentimentos, a mineração de texto e outras tarefas relacionadas ao tratamento de dados textuais. A ferramenta inclui uma série de componentes e módulos que facilitam o desenvolvimento e a implementação de soluções para diversos problemas de processamento de texto.
JHove https://jhove.openpreservation.org/	Geração automática extrínseca.	Ferramenta que extrai metadados sobre o formato e tamanho do arquivo e também executa a validação de variados tipos de formatos digitais, auxiliando na preservação digital.
Kea *http://community.nzdl.org/kea/*	Extração de conteúdo; Indexação automática.	Ferramenta que analisa textos completos e executa a extração de frases-chave. As frases-chave também podem ser mapeadas para ontologias personalizadas ou vocabulários controlados, atribuindo termos do assunto.
MarcEdit *https://marcedit.reeset.net/*	Colheita de *metatag*.	Ferramenta projetada para edição e manipulação de dados bibliográficos em formato MARC. Ela permite a edição em massa, a conversão entre diferentes formatos de dados e a validação de registros MARC. A solução também oferece um conjunto de utilitários, como um editor de registros, um gerador de relatórios e recursos para importação e exportação.
MetaGen https://www.codeproject.com/Articles/41910/MetaGen-A-Project-metadata-Generator-for-Visual-St	Extração de conteúdo; Indexação automática.	Solução desenvolvida para gerar metadados de projetos no *Visual Studio* (ambiente de desenvolvimento integrado da *Microsoft*). A ferramenta automatiza a criação de arquivos de metadados, facilitando a documentação e a gestão de informações do projeto; e permite a personalização dos dados gerados, mantendo os metadados organizados de forma eficiente.
SEOGenerator https://extensions.joomla.org/extension/site-management/seo-a-metadata/seo-generator/	Extração de conteúdo.	Um *plug-in*/extensão para *Joomla* (ferramenta de gestão de conteúdo web), que automatiza a geração de títulos, descrições e palavras-chave de páginas, melhorando a visibilidade nos motores de busca.
Meta Tag Extractor https://meta-tag-extractor.soft112.com/	Colheita de *metatag*.	Ferramenta que permite extrair metadados de páginas da web de forma simples, analisando o Localizador Uniforme de Recursos (URL) e gerando relatório com informações sobre as

		metatags (elementos utilizados em documentos HTML e XHTML), tais como: título, descrição e palavras-chave.
My Meta Maker https://uol.de/f/5/inst/physik/ag/ehemalige/hilf/vortraege/twente99/twente99-mmm.html	Colheita de *metatag*.	Ferramenta online para criação e edição de metadados para publicações *online*. A saída do *script* produz um código-fonte HTML para uma página de índice que descreve o documento.
Metadata Extraction Tool http://meta-extractor.sourceforge.net/	Geração automática extrínseca.	Ferramenta de código aberto que permite extrair metadados de diversos tipos de arquivos, como documentos, imagens e vídeos. Ela suporta uma variedade de formatos e fornece informações detalhadas sobre os metadados contidos nos arquivos.
Omeka http://omeka.org/	Geração automática extrínseca; Marcação Social.	Plataforma de *software* livre projetada para a gestão e exibição de coleções digitais. Ideal para bibliotecas, arquivos e museus. Permite a criação de exposições online e a catalogação de itens com metadados enriquecidos. A ferramenta é personalizável, suportando extensões e temas. Facilita a colaboração e a publicação de conteúdo.
Ont-O-Mat (descontinuado)	Extração de conteúdo.	Ferramenta destinada à construção e edição de ontologias. Ela facilita a modelagem de dados e a definição de relações entre conceitos. Muitos usuários estão migrando para outras ferramentas de modelagem de ontologias, como a *Protégé* <https://protege.stanford.edu/>.
Open Text Summarizer https://open-text-summarizer.soft112.com/	Extração de conteúdo.	Ferramenta que gera resumos automáticos a partir de textos longos, ajudando os usuários a extrair as principais ideias de forma rápida. Ela analisa o conteúdo e produz um resumo conciso, facilitando a compreensão e revisão de informações.
ParsCit https://github.com/knmnyn/ParsCit	Extração de conteúdo.	Ferramenta que extrai citações e metadados de artigos acadêmicos. Desenvolvida como um projeto de código aberto, ela analisa documentos em formato de texto e identifica as referências bibliográficas, facilitando a organização e a recuperação da informação.
Photo RDF-Gen http://www.webposible.com/utilidades/photo_rdf_generator_en.html	Colheita de *metatag*.	Ferramenta que gera metadados no formato RDF para imagens, permitindo a criação de descrições semânticas para arquivos fotográficos. Ela facilita a inclusão de informações relevantes, como autor, data e descrição, de forma

		que as imagens possam ser melhor indexadas e recuperadas na *web*.
PyMarc https://pypi.org/project/pymarc/	Colheita de *metatag*.	Biblioteca Python para manipulação de dados no formato *Machine-Readable Cataloging* (MARC). Ela permite a leitura, escrita e edição de registros MARC, facilitando a integração de bibliotecas e sistemas de gerenciamento de dados bibliográficos.
RepoMMan http://www.ukoln.ac.uk/repositories/digirep/index/RepoMMan (Não disponível em 23/07/2020)	Extração de conteúdo; Colheita de *metatag*; Geração automática extrínseca.	Ferramenta foi projetada para gerenciar e registrar metadados de repositórios digitais. Ela oferecia funcionalidades para a criação, edição e validação de metadados, facilitando a interoperabilidade entre diferentes sistemas. A ferramenta não recebe atualizações recentes e não é possível encontrar suporte ou documentação atualizada.
Sherpa/RoMEO https://v2.sherpa.ac.uk/api/	Colheita de *metatag*.	É um recurso *online* que agrega e analisa as políticas de acesso aberto de editoras do mundo todo e fornece resumos de direitos autorais de editoras e políticas de arquivamento de acesso aberto para cada periódico.
***Simple Automatic Metadata Generation Interface* (SamgI)** Não está disponível para acesso em 23/07/2020.	Extração de conteúdo; Geração automática extrínseca.	Era uma ferramenta projetada para gerar metadados automaticamente, facilitando a catalogação de recursos digitais. A interface visava simplificar o processo de criação de metadados, permitindo que usuários pudessem integrar facilmente informações em seus sistemas.
URL and Metatag Extractor https://metatagsextractor.com/	Colheita de *metatag*.	Ferramenta online que permite extrair metadados de URLs, como títulos, descrições e palavras-chave, ajudando a analisar como as páginas da web são indexadas. Ela oferece uma interface simples, onde os usuários podem inserir *links* e obter informações relevantes rapidamente.
Termine http://www.nactem.ac.uk/software/termine/	Extração de conteúdo.	Ferramenta de extração de termos e metadados de textos, desenvolvida para facilitar a identificação e organização de termos relevantes. Ela é útil para pesquisadores que trabalham com análise de linguagem natural, pois fornece funcionalidades para a extração e análise de termos, permitindo uma melhor compreensão e categorização de dados textuais.
Yahoo Content Analysis API	Extração de conteúdo;	Era uma ferramenta que oferecia análise de conteúdo para desenvolvedores,

(descontinuado em 30/06/2020)	Indexação automática.	permitindo a extração e análise de informações de texto. Detectava entidades/conceitos, categorias e relacionamentos dentro de conteúdo não estruturado.

Fonte: Adaptado e atualizado de Park e Brenza (2015).

Realizou-se acesso às ferramentas descritas no Quadro 5 e identificou-se *links* que estavam quebrados – sendo todos atualizados. Algumas ferramentas foram descontinuadas, outras possuem *link* para a documentação, mas não disponibilizam acesso ao projeto, código ou *software* para *download*.

Park e Brenza (2015) discorrem que apesar das ferramentas de geração semiautomática de metadados oferecerem muitos benefícios, especialmente no que se refere à racionalização do processo de criação de metadados, existem barreiras significativas à adoção e implementação generalizadas delas. Um fator é que muitas são desenvolvidas localmente para atender as necessidades específicas de um determinado projeto ou como parte da pesquisa acadêmica. Esse ambiente altamente focado para um contexto específico significa que a aplicabilidade geral das ferramentas é potencialmente diminuída.

Outro fator, discorrido pelos autores, é o alto nível de conhecimento técnico exigido para seu desenvolvimento e sua incorporação aos fluxos de trabalho diário no processo de criação de metadados. Por fim, ferramentas de geração semiautomáticas de metadados não são testadas em cenários do mundo real, tais como: pequenos tamanhos de amostra, escopo restrito de domínios de projeto e experimentos que não têm objetividade.

3.3 Possibilidades e Limitações das Ferramentas Analisadas

No Quadro 6 são descritas as possibilidades e limitações das ferramentas analisadas:

Quadro 6 – Possibilidades e Limitações das Ferramentas

Autores	Possibilidades	Limitações
Kovaevic *et al*. (2011)	Atribuir automaticamente os metadados em um recurso de informação e nos repositórios digitais.	Indicação de que, em vários casos, os metadados extraídos automaticamente não podem ser inseridos na base de dados, pois dependem de controle da curadoria.
Maratea, Petrosino e Manzo (2012)	Executar a produção de metadados para conteúdo de aprendizagem com o objetivo de facilitar sua busca e recuperação nos acervos digitais, além de permitir a interoperabilidade.	Dependência da seleção do vocabulário controlado, da qualidade dos campos de metadados extraídos de *strings* de referência e de um conjunto bem estruturado dos documentos PDF.
Sah e Wade (2012)	Fornecer informações avançadas para personalização de um sistema para o cliente, ampliando sua capacidade de uso e melhorando a experiência do usuário.	A mineração automática de metadados cognitivos é um desafio, uma vez que é muito difícil compreender automaticamente o conhecimento intelectual subjacente sobre o documento.
Verborgh *et al*. (2012)	Melhorar o significado dos metadados de objetos vinculados em um repositório por meio da anotação multimídia, aperfeiçoando a recuperação da informação.	O alto grau de especialização dos algoritmos para extração de metadados faz com que eles desconheçam o contexto em que operam, inclusive os que contém informações valiosas e muitas vezes necessárias. É necessário aperfeiçoar o tratamento de informações imperfeitas, tais como a incerteza e a incompletude.
Costa *et al*. (2013)	Disponibilizar sistema de catalogação espacial e ferramenta de geração automática de metadados geoespaciais para tomada de decisão e compartilhamento de dados entre instituições. A solução possui possibilidade de ser aplicada em diversas áreas, tais como: previsão do tempo para a agricultura, desastres naturais, desmatamento florestal, defesa e segurança pública, inteligência, comunicação, aquecimento e efeito estufa, saúde, monitoramento dos recursos hídricos e ambientais, tráfego urbano, aéreo e rodovias, etc;	Desafio em gerenciar grandes volumes de dados espaciais heterogêneos, assim como melhorar sua organização, busca e prevenir a duplicidades.
Vlachidis *et al.* (2013)	Resolver problemas semânticos nos repositórios, por meio do uso das ferramentas para geração automática de metadados,	Necessidade de avaliação da ferramenta em larga escala para analisar o desempenho de extrações de metadados,

	propiciando a execução de seu enriquecimento semântico.	considerando ambiguidades lexicais e a precisão de anotação do termo.
Rafferty *et al*. (2015)	Fornece instruções dinâmicas a partir de clipes de vídeo, facilitando a compreensão e acessibilidade à informação do usuário ao recurso de informação.	Ocorrência de falsos-positivos na geração automática; palavras semanticamente compatíveis não foram descobertas no processamento, necessitando fontes adicionais de sinônimos.
Park e Brenza, (2015)	Conversão de dados, documentos e arquivos; extração de metadados de textos e arquivos de imagens; aplicação de metadados em ontologias; análise de citações; criação de metadados para coleções; indexação de documentos etc.	As ferramentas são desenvolvidas localmente para atender necessidades específicas, altamente focadas para um contexto particular. Além disso, exigem alto grau de habilidade técnica para sua implementação.
Gonzalo *et al*. (2018)	Infraestrutura escalável generalizada para fornecer pesquisa em conjuntos de dados científicos, usando métodos diferentes para extração de metadados.	Escalabilidade e desempenhos impactados por variáveis externas: *hardware*, *software*, sistema operacional, ambiente/plataforma computacional, funcionalidades na própria ferramenta ativadas ou não.
Yang e Park (2018)	Auxiliar na atenuação de problemas de interoperabilidade semântica entre as coleções digitais.	Ausência de geração de meta-conhecimento ou mapa de conhecimento estruturado, como forma de explicitação do conhecimento por meio da documentação, impactando o repasse e o compartilhamento de uso da ferramenta.
Audichya e Asini (2019)	Propor e sugerir metadados de escritos antigos e apoiar sua catalogação, classificação e indexação, objetivando a preservação histórica e cultural.	Ferramenta depende totalmente da entrada correta e completa dos dados, de acordo com as regras gramaticais, tais como o uso de acentos, de caracteres comuns e especiais, além dos sinais de pontuação.
Morris (2020)	Gerar e incluir metadados ausentes em repositórios com grandes volumes de dados, objetivando a completude dos dados armazenados.	Limitação à descrição correta e completa de *metatags*, registros *DC* e *MARC* para extração e transformação.

Fonte: Elaborado pelos autores.

3.4 ABORDAGENS DE SOLUÇÕES TECNOLÓGICAS À POSTERIORI

Nas seções anteriores identificou-se um conjunto de técnicas e ferramentas específicas para geração automática e semiautomática de

metadados. Entretanto, não foi possível selecionar dentre as soluções tecnológicas relatadas, aquelas passíveis de serem aplicadas a dados reais de um repositório digital, considerando as limitações apresentadas na seção anterior. Realizou-se identificação à posteriori das seguintes soluções:

- **Ferramenta semiautomática de metadados** denominada ColetadorOAI[7], desenvolvida pelo Laboratório de Inteligência de Redes da Universidade de Brasília (UnB) em parceria com o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Essa solução é de código aberto, desenvolvida em *Python* e de fácil assimilação e implementação. Foi utilizada em projeto do IBICT com a Fundação Nacional de Artes (FUNARTE), no intuito de coleta, busca e recuperação da informação da produção científica implementando um repositório digital real sobre o mundo das artes no Brasil e experimentação de modelos de agregação de acervos digitais de artes de instituições de cultura ligadas ao Governo Federal;
- Após buscas com o termo “indexação automática”, outras **ferramentas para geração automática de metadados** em evidência que estão sendo utilizadas nos acervos digitais de bibliotecas nos últimos anos. Pode-se citar a compilação de 10 sistemas de indexação automática executada por Corrêa e Lapa (2015), cuja pesquisa evidenciou as seguintes ferramentas com maior destaque: BIB/DIÁLOGO, SISA, PRECIS, OGMA; e outras soluções que foram utilizadas em estudos com menor frequência: SINTAGMED, ZSTATION, SRIAC, SPIRIT, MISTRAL e SIRILICO.

Silva, Correia e Gil-Leiva (2020) realizam uma análise comparativa entre os sistemas SISA e MAUI, obtendo bons resultados de precisão na indexação. Na literatura acadêmica é notório o discurso sobre a necessidade de ferramentas para indexação automática de metadados, visto o volume e variedade de dados produzidos ao longo dos anos sem a devida classificação e organização sobre seu contexto, o que dificulta sua indexação, busca e recuperação.

Foi realizada busca exploratória nos sites de várias bibliotecas que

[7] Laboratório de Inteligência de Redes (2022).

contemplassem investigações sobre a geração automática de metadados, sendo encontrado um relatório descrito por membros da Biblioteca Nacional da Holanda, intitulado "*Exploration possibilities Automated generation of metadata*"[8] e publicado em 2019.

Kleppe *et al.* (2019) descreve no relatório a dificuldade da Biblioteca da Nacional da Holanda em realizar a atribuição da descrição dos recursos de informação (conhecido como "geração de metadados" ou "criação de registros bibliográficos"). Essa problemática se deu em parte ao grande crescimento de material eletrônico gerado e armazenado, culminando na necessidade de otimizar a descrição dos recursos de informação realizados até então manualmente. Desta forma, esses autores demonstram o uso de tecnologias inteligentes para analisar e descrever fontes como artigos de notícias, livros, transmissões de televisão, fotografias com o uso de geração automática de metadados. Aprofundaram a investigação em duas soluções:

- ANNIF, ferramenta desenvolvida pela Biblioteca Nacional da Finlândia. Essa solução oferece módulos existentes para processamento de linguagem natural e aprendizagem de máquina, podendo ser combinado de diversas maneiras, além de ser *open source*. Pode utilizar o classificador *FastText* como um *backend* alternativo.
- ARIADNE, ferramenta desenvolvida pela *Online Computer Library Center* (OCLC), alcançou boas pontuações, mas os pesquisadores não sabiam o bastante sobre a metodologia utilizada no *background* e nem o material para treinar o sistema e isso se deve ao fato de a ferramenta não ser *open source*.

O relatório forneceu indicadores interessantes sobre a solução ANNIF, *open source*, escrito em Python e todo o seu código está disponível no GitHUB, além de exemplos práticos de uso, comunidade de discussão ativa[9], artigos publicados discorrendo sobre a sua aplicação em grande conjunto de acervos digitais.

[8] Kleppe *et al*. (2019).

[9] Disponível em: annif-users@googlegroups.com. Acesso em: 21 jan. 2022.

Os testes executados por Kleppe *et al*. (2019) demonstram que a ferramenta ANNIF é robusta, possui nível de qualidade adequada e pode ser customizada para diversas necessidades e aplicações.

4 CONCLUSÕES

Conclui-se que as ferramentas de geração automática e semiautomática de metadados descritas possuem diversas possibilidades de aplicação com intuito de melhorar a compreensão dos dados, facilitar a busca e recuperação da informação nos repositórios digitais.

As ferramentas foram desenvolvidas para fornecer solução em diversos contextos, sendo algumas utilizadas apenas em ambiente de testes, faltando a sua aplicação em cenários reais. A maioria das soluções possui forte dependência da entrada correta, completa e acurada de dados.

Os gestores dos repositórios digitais são responsáveis por uma quantidade enorme de metadados relacionados a diferentes tipos de documentos que geralmente são indexados por títulos, assuntos e descritores para que possam ser recuperados posteriormente. Entretanto, nem todos os usuários desses sistemas e repositórios digitais executam a entrada correta e completa de metadados, o que dificulta a recuperação do objeto de pesquisa. O processo de indexação manual de documentos é um trabalho árduo, oneroso e dependendo do volume de dados é humanamente impossível de ser realizado. Essa problemática pode ser mitigada com o desenvolvimento ou adequação de ferramentas que utilizem técnicas de aprendizagem de máquina para realizar a geração automática e semiautomática dos metadados.

Há lacuna existente na literatura acadêmica no recorte temporal adotado sobre a criação de um modelo de referência para geração de metadados de uso geral. Observou-se na pesquisa que as ferramentas de geração automática e semiautomática de metadados são desenvolvidas para atender necessidades muito específicas.

Sugere-se pesquisas que mesclem funcionalidades de ferramentas de geração automática e semiautomáticas com intuito de verificar a complementação das soluções para se atingir um resultado com maior eficiência

na geração dos metadados e diminuir o tempo de implementação, auxiliando o processo de indexação, classificação e suporte para os gestores de repositórios digitais.

Segundo, é importante haver uma ferramenta multilíngue, que suporte diferentes algoritmos de forma flexível e adaptável a diferentes situações e que seja possível o uso de qualquer vocabulário controlado.

Terceiro, sugere-se testar diversos algoritmos separados e em conjunto para treinamento de um modelo de aprendizagem de máquina, com intuito de obter resultados mais satisfatórios no processo de geração automática de metadados.

Quarto, desenvolver pesquisas para uso de ferramentas de geração automática e semiautomática para sugestão de assuntos e indexar automaticamente em um repositório digital com a possibilidade de fornecer ao usuário a recuperação do recurso de informação por meio de buscas facetadas.

Por último, enfatizamos que o recorte temporal do presente artigo foi entre o decênio de 2010 a 2020 e por isso não se analisou ferramentas desenvolvidas e liberadas posteriormente.

REFERÊNCIAS

AUDICHYA, M, K; SAINI J, R. Computational linguistic prosody rule-based unified technique for automatic metadata generation for Hindi poetry. *In:* INTERNATIONAL CONFERENCE ON ADVANCES IN INFORMATION TECHNOLOGY, 1., 2019, Chikmagalur. **Proceedings** […] Chikmagalur: IEEE, 2019. p. 436-442. Disponível em: https://www.semanticscholar.org/paper/Computational-linguistic-prosody-rule-based-unified-Audichya-Saini/dd6be4b8ee154249df08dbeb3f1115611bc977ad. Acesso em: 28 nov. 2024.

COSTA, A; FIRDAUSY, T, P; INNEREBNER, M; MONSORNO, R. EURAC SDI: a near real time and offline automatic metadata generation processing chain. GI Forum, 1., 2013, [*S. l.*], **Proceedings** […] Berlim: VDE VERLAG GMBH, 2013. Disponível em: https://encurtador.com.br/XMa7f. Acesso em: 28 nov. 2024.

CRYSTAL, A; LAND, P. **Metadata and Search: Global Corporate Circle DCMI 2003 Workshop.** 2003. Disponível em: http://www.dublincore.org/groups/corporate/Seattle/. Acesso em: 07 mar. 2023.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA (EMBRAPA). Satélites de monitoramento. Campinas: Embrapa, 2024. Disponível em: https://www.embrapa.br/satelites-de-monitoramento/satelites. Acesso em: 22 set. 2024.

GONZALO, P, R; MATT, H; GUNTHER, H, W; COLIN, O; KATIE, A; LAVANYA, R. Science Search: enabling search through automatic metadata generation. *In:* INTERNATIONAL CONFERENCE ON E-SCIENCE, 14., 2018, Amsterdã, **Proceedings** […] Amsterdã: IEEE, 2018. Disponível em: https://www.osti.gov/biblio/1602828. Acesso em: 28 nov. 2024.

GREENBERG, J. Metadata Extraction and Harvesting: a comparison of two automatic metadata generation applications. **Journal of Internet Cataloging**, *[S. l.],* v. 6, n. 4, 2003. Disponível em: https://researchdiscovery.drexel.edu/esploro/outputs/journalArticle/Metadata-Extraction-and-Harvesting-A-Comparison/991014878230804721. Acesso em: 28 nov. 2024.

JOSHI, B. K; KUSHWAH, K. K. A Novel approach to automatic detection of Chaupai Chhand in Hindi Poems. *In:* INTERNATIONAL CONFERENCE ON COMPUTING, POWER AND COMMUNICATION TECHNOLOGIES (GUCON), 1., 2018, Greater Noida, **Proceedings** […] Greater Noida: IEEE, 2018. p. 223-228. DOI: 10.1109/GUCON.2018.8675052.

KLEPPE, M; VELDHOEN, S; WAAL-GENTENAAR, M. V. D; OUDSTEN, B. D; HAAGSMA, D. **Exploration possibilities Automated Generation of Metadata.** 2019. Disponível em: https://doi.org/10.5281/zenodo.3375192. Acesso em: 07 mar. 2023.

KOVACEVIC, A.; IVANOVIC, D.; MILOSAVLJEVIC, B.; KONJOVIC, Z. Automatic extraction of metadata from scientific publications for CRIS systems. **Program: Electronic Library and Information Systems**, *[S. l.],* v. 45, n. 4, p. 376-396, 2011. Disponível em: https://www.researchgate.net/publication/216592386_Automatic_extraction_of_metadata_from_scientific_publications_for_CRIS_systems. Acesso em: 28 nov. 2024.

LABORATÓRIO DE INTELIGÊNCIA DE REDES. ColetadorOAI-sickle.py. Brasília: UnB; IBICT, 2022. Disponível em: https://github.com/tainacan/data_science/blob/master/FUNARTE/BIBLIOTECA_DIGITAL/ColetadorOAI-sickle.py. Acesso em: 20 set. 2024.

MARATEA, A.; PETROSINO, A.; MANZO, M. Automatic Generation of SCORM Compliant Metadata for Portable Document Format Files. *In:* INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND TECHNOLOGIES – COMPSYSTECH, 13., Nova Iorque, 2012, **Proceedings** […] Nova Iorque: Association for Computing Machinery, 2012. Disponível em: https://www.researchgate.net/publication/262277720_Automatic_generation_of

_SCORM_compliant_metadata_for_portable_document_format_files. Acesso em: 28 nov. 2024.

MARCONI, M. A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

MOOERS, C. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, *[S. l.],* v.2, n.1, p. 20-32, 1951. Disponível em: https://courses.grainger.illinois.edu/cs473/fa2013/misc/zatocoding.pdf. Acesso em: 28 nov. 2024.

MORRIS, V. Automated Language Identification of Bibliographic Resources. **Cataloging & Classification Quarterly**, *[S. l.],* v. 58, n. 1, p. 1-27, 2020. Disponível em: https://bl.iro.bl.uk/concern/articles/6c99ffcb-0003-477d-8a58-64cf8c45ecf5?locale=en. Acesso em: 28 nov. 2024.

PARK, J.; BRENZA, A. Evaluation of Semi-Automatic Metadata Generation Tools: a survey of the current state of the art. **Information Technology and Libraries**, Chicago, v. 34, ed. 3, p. 22-42, 2015. Disponível em: https://ital.corejournals.org/index.php/ital/article/view/5889. Acesso em: 28 nov. 2024.

POLFREMAN, M.; BROUGHTON, V.; WILSON, A. **Metadata Generation for Resource Discovery.** JISC, 2008. Disponível em: http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/autometgen.aspx. Acesso em: 07 mar. 2023

RAFFERTY, J.; NUGENT, C.; LIU, J. Automatic Metadata Generation Through Analysis of Narration Within Instructional Videos. **Transaction Processing Systems**, J Med Syst, *[S. l.],* v. 94, n. 39, 2015. Disponível em: https://pubmed.ncbi.nlm.nih.gov/26254252/. Acesso em: 28 nov. 2024.

REINSEL, D.; GANTZ, J.; RYDNING, J. The Digitization of the World: From Edge to Core. **Data Age 2025**, *[S. l.],* nov., 2018. Disponível em: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Acesso em: 28 nov. 2024.

SAH, M; WADE, V. Automatic metadata mining from multilingual enterprise content. **Web semantics:** Science, services and agents on the world wide web, *[S. l.],* v. 11, p. 41-62, 2012. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3198936. Acesso em: 28 nov. 2024.

SILVA, S. R. de B; CORREA, R. F; GIL-LEIVA, I. Avaliação Direta e Conjunta de Sistemas de Indexação Automática por Atribuição**. Inf. & Soc.:**Est., João Pessoa, v.30, n.4, p. 1-27, out./dez. 2020. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/57259. Acesso em: 28 nov. 2024.

ULRICH, H; KOCK-SCHOPPENHAUER, A; DEPPENWIESE, N; GÖTT, R; KERN, J; LABLANS, M; MAJEED, R. W; STÖHR, M. R; STAUSBERG, J; VARGHESE, J; DUGAS, M; INGENERF, J. Understanding the Nature of Metadata: Systematic Review. **J Med Internet Res**, *[S. l.],* v. 24, p. 1, 2022. Disponível em: https://pubmed.ncbi.nlm.nih.gov/35014967/. Acesso em: 28 nov. 2024.

VERBORGH, R; VAN DEURSEN, D; MANNENS, E; POPPE, C; WALLE, R, V. Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform. **Multimed Tools Appl,** *[S. l.],* v. 61, p. 105–129, 2012. Disponível em: https://link.springer.com/article/10.1007/s11042-010-0709-6. Acesso em: 28 nov. 2024.

VLACHIDIS A., BINDING C., MAY K., TUDHOPE D. Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature. *In:* PRZEPIÓRKOWSKI A., PIASECKI M., JASSEM K., FUGLEWICZ P. (ed.). **Computational Linguistics. Studies in Computational Intelligence**, Springer; Berlin; Heidelberg, v. 458, 2013. Disponível em: https://pure.southwales.ac.uk/en/publications/automatic-metadata-generation-in-an-archaeological-digital-librar. Acesso em: 28 nov. 2024.

YANG, G; PARK, J. Automatic Extraction of Metadata Information for Library Collections. **International Journal of Advanced Culture Technology**, *[S. l.],* v. 6, n. 2, p. 117-122, 2018. Disponível em: https://koreascience.kr/article/JAKO201820540196117.page. Acesso em: 28 nov. 2024.

AUTOMATIC AND SEMI-AUTOMATIC METADATA GENERATION TOOLS: A REFLECTION BETWEEN THE YEARS 2010 AND 2020

ABSTRACT

Objective: Identify the possibilities and limitations for using the analyzed tools. **Methodology**: This is an exploratory investigation, carrying out a bibliographical review and the search was carried out in the Scopus, Web Of Science, ISTA, LISTA and LISA databases. A mixed method was used in data analysis, with quantitative and qualitative approaches. 49 scientific articles were found and after applying the adopted criteria, only 12 were selected for synthesis. **Results**: The results demonstrated several tools and solutions for generating metadata, using various techniques, methods and functions, addressing their implementation and use. In this context, the possibilities and limitations of these solutions were identified, with the aim of contributing to their application and improvement in future research. **Conclusions**: It is concluded that automatic and semi-automatic metadata generation tools are instruments that can assist information professionals in the organized and efficient management of digital collections, improving information retrieval, which reinforces the contribution of this research in the academic world- scientific in the area of Information Science.

Descriptors: Bibliographic review. Metadata. Automatic and semi-automatic generation. Possibilities. Limitations.

HERRAMIENTAS DE GENERACIÓN DE METADATOS AUTOMÁTICAS Y SEMIAUTOMÁTICAS: UNA REFLEXIÓN ENTRE LOS AÑOS 2010 AL 2020

RESUMEN

Objetivo: Identificar las posibilidades y limitaciones de uso de las herramientas analizadas. **Metodología**: Se trata de una investigación exploratoria, realizándose una revisión bibliográfica y la búsqueda se realizó en las bases de datos Scopus, Web Of Science, ISTA, LISTA y LISA. En el análisis de los datos se utilizó un método mixto, con enfoques cuantitativos y cualitativos. Se encontraron 49 artículos científicos y luego de aplicar los criterios adoptados, sólo 12 fueron seleccionados para su síntesis. **Resultados**: Los resultados demostraron varias herramientas y soluciones para generar metadatos, utilizando diversas técnicas, métodos y funciones, abordando su implementación y uso. En este contexto, se identificaron las posibilidades y limitaciones de estas soluciones, con el objetivo de contribuir a su aplicación y mejora en futuras investigaciones. **Conclusiones**: Se concluye que las herramientas de generación automática y semiautomática de metadatos son instrumentos que pueden ayudar a los profesionales de la información en la gestión organizada y eficiente de las colecciones digitales, mejorando la recuperación de información, lo que refuerza el aporte de esta investigación en el mundo académico-científico en el área de Ciencias de la Información.

Descriptores: Revisión bibliográfica. Metadatos. Generación automática y semiautomática. Posibilidades. Limitaciones.

Recebido em: 22.01.2023
Aceito em: 09.10.2023