XXI ENANCIB
Encontro Nacional de Pesquisa em Ciência da Informação
50 anos de Ciência da Informação no Brasil: diversidade, saberes e transformação social
Rio de Janeiro • 25 a 29 de outubro de 2021

XXI Encontro Nacional de Pesquisa em Ciência da Informação – XXI ENANCIB

GT 8 – Informação e Tecnologia

GERAÇÃO AUTOMÁTICA E SEMI-AUTOMÁTICA DE METADADOS: REVISÃO BIBLIOGRÁFICA

AUTOMATIC AND SEMI-AUTOMATIC METADATA GENERATION: BIBLIOGRAPHIC REVIEW

Jean Carlos Borges Brito – Universidade de Brasília (UNB)
Dalton Martins – Universidade de Brasília (UNB)

Modalidade: Trabalho Completo

Resumo: O artigo tem como objetivo apresentar e contextualizar o tema geração automática e semi-automática de metadados, suas ferramentas, técnicas, características e funções por meio de uma revisão bibliográfica. Foi realizada pesquisa exploratória em bases de dados científicas da Ciência da Informação, selecionando periódicos específicos para a avaliação. Utilizou-se método misto na análise dos dados, com abordagens quantitativas e qualitativas. Esta pesquisa visa contribuir com o meio acadêmico-científico ao demonstrar a evolução do tema. Foram encontrados 49 artigos nas bases de dados e após a aplicação dos critérios de inclusão/exclusão e qualidade, apenas 12 foram selecionados para a síntese qualitativa. Identificaram-se pesquisas em 04 categorias de análise, demonstrando 32 ferramentas de geração semi-automática de metadados, 08 de geração automática de metadados, 02 de extração automática e 01 de catalogação com ferramentas automatizadas. O estudo sugere o desenvolvimento de um modelo de referência para geração de metadados de uso geral e o uso complementar de ferramentas de geração automática e semi-automática para auxiliar os gestores de repositórios digitais.

Palavras-Chave: Revisão Bibliográfica; Metadados; Geração Automática e Semi-automática de Metadados.

Abstract: *The article aims to present and contextualize the theme of automatic and semi-automatic generation of metadata, its tools, techniques, characteristics, and functions through a literature review. Exploratory research was carried out in scientific databases of Information Science, selecting specific journals for evaluation. A mixed method was used in data analysis, with quantitative and qualitative approaches. This research aims to contribute to the academic-scientific environment by demonstrating the evolution of the subject. Forty-nine articles were found in the databases and after applying the inclusion/exclusion and quality criteria. From these articles, only 12 were selected for the qualitative synthesis. Research was identified in 04 categories of analysis, demonstrating 32 semi-automatic metadata generation tools, 08 automatic metadata generation, 02 automatic extraction, and 01 cataloging with automated tools. The study suggests the development of a reference model for general-purpose metadata generation and the complementary use of automatic and semi-automatic generation tools to help digital repository managers.*

Keywords: *Bibliographic Review; Metadata; Automatic and Semi-automatic Metadata Generation.*

1 INTRODUÇÃO

A humanidade produz informações em volumes e variedades exponenciais diariamente que são armazenadas principalmente em repositórios digitais. Conteúdos físicos legados estão sendo convertidos e mantidos em bancos de dados e unidades de armazenamento lógico, acompanhando a transformação digital atual por qual passa vários países do mundo, abrangendo governos, empresas e sociedade mundial.

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. De acordo com esses autores, grande parte da economia atual depende de dados e a confiança 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 catalogação de dados nesse ambiente digital, disponibilizando em tempo real aquilo que é feito tradicionalmente de forma manual. A catalogação é compreendida como:

> [...] o estudo, a preparação e a organização de mensagens, com base em registros do conhecimento, reais ou ciberespaciais, existentes ou passíveis de inclusão em um ou vários acervos, visando a criar conteúdos comunicativos que permitam a interseção entre as mensagens contidas nestes registros do conhecimento e as mensagens internas dos usuários (MEY; SILVEIRA, 2020, p. 126).

Cunha e Cavalcanti (2008, p. 70) ampliam essa interpretação discorrendo que "[...] a catalogação abrange não somente a descrição bibliográfica, mas também a análise temática com seus produtos, entre eles a identificação temática". Um usuário potencial da informação é capaz de converter sua necessidade de informações em uma lista de referências para documentos armazenados e que contém informações úteis (MOOERS, 1951). Neste contexto, os metadados desempenham um papel essencial, pois descrevem os dados, facilita sua compreensão e corrobora na eficácia da catalogação e na sua obtenção.

De acordo com Pomerantiz (2015), metadado indica algo que está além dos dados, sendo uma declaração sobre esses dados. Haynes (2018) corrobora com o entendimento de que o metadado tem a função de facilitar o entendimento dos relacionamentos e evidenciar a utilidade das informações obtida dos dados. 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.

Conforme Greenberg (2003), o incremento de metadados com qualidade fornece valor agregado ao conjunto de dados, além de melhorar sua classificação e busca. Os pesquisadores necessitam identificar métodos de produção de metadados mais eficientes e menos dispendiosos. Devido ao alto custo da inserção manual de metadados, é mister o fomento e incentivo em desenvolvimento de ferramentas que possam auxiliar a geração automática ou semi-automática de metadados, melhorando sua escalabilidade.

De acordo com Maratea, Petrosino e Manzo (2012), a geração automática de metadados (GAM) iniciou-se com a introdução de documentos digitais desde os anos de 1950 e diz respeito à sua indexação, abstração e classificação de forma automática.

Park e Brenza (2015) apresentam em seus estudos, ferramentas de geração semi-automática de metadados (GSAM), que dizem respeito ao uso de *software* para criação de registros de metadados com graus variados de supervisão por um especialista humano.

De acordo com Greenberg (2003), a geração automática de metadados, baseado no conhecimento sobre indexação automática – associação de termos a documentos –, é mais eficiente, possui menor custo e é mais consistente do que processos executados por seres humanos. Com a ascensão da internet, outro técnica muito importante para esse ambiente é a extração de metadados, pois é um método de geração automática e ocorre quando um algoritmo automaticamente extrai metadados do conteúdo de um recurso de informação exibido através de um navegador *web*. Mesmo com todas essas vantagens da GAM, a autora enfatiza que segundo alguns pesquisadores, os meios mais efetivos para criar metadados é integrar métodos automáticos e semi-automáticos, um complementando o outro.

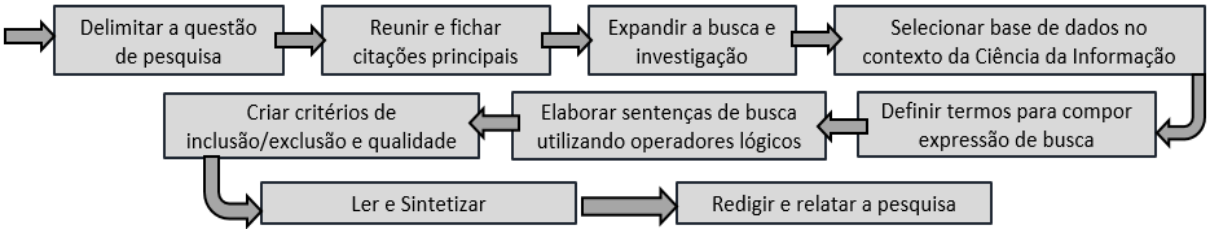
Diante dos argumentos apresentados, pesquisas voltadas para criação e integração de ferramentas de geração automática e semi-automática de metadados são muito importantes para fornecer auxílio às pessoas e profissionais em gerenciar quantidades e tipos cada vez maiores de dados e metadados dos recursos de informação.

2 DESENVOLVIMENTO

O método de pesquisa será bibliográfico, pois compreende a investigação em estudos já realizados, revestidos de importância, nas seguintes fontes de conhecimento: artigos, periódicos e *journals*, fornecendo dados atuais e relevantes relacionados com o tema (MARCONI; LAKATOS, 2003, p. 158).

Para se atingir os resultados esperados através da revisão bibliográfica, construiu-se as seguintes etapas com a finalidade de facilitar a recuperação dos dados, conforme Figura 1.

Figura 1 – Etapas da revisão bibliográfica.



Fonte: Elaborado pelo autor (2021).

2.1 PLANEJAMENTO DA REVISÃO BIBLIOGRÁFICA

Será realizada a delimitação da questão de pesquisa, definindo a população ou o problema, a intervenção, comparação e o resultado. Inicialmente, elencaram-se questões de *background* para fornecer a compreensão básica e conceitual do tema. Em seguida, para melhor definição do escopo, estabeleceram-se questões de *foreground.*

Tabela 1 - Questões de *background* e *foreground*.

QUESTÕES DE *BACKGROUND*	QUESTÕES DE *FOREGROUND*
O que são metadados? Para que servem? O que é geração automática e semi-automática de metadados?	Quais técnicas, características, funções e ferramentas de geração automática e semi-automática de metadados?

Fonte: Elaborado pelo autor (2021).

Com o entendimento dos conceitos e da abrangência, formulou-se a seguinte questão de pesquisa para a investigação: "**Quais as aplicabilidades das ferramentas de geração automática e semi-automática de metadados para o gestor de repositórios digitais?**". Da questão elaborada, podem-se evidenciar os seguintes componentes:

Tabela 2 – Descrição e componentes da pergunta.

DESCRIÇÃO	COMPONENTES DA PERGUNTA
População	Gestor de repositórios digitais

Intervenção	Geração automática e semi-automática de metadados
Comparação	Ferramentas, técnicas, características e funções.
Desfecho	Aplicabilidades atuais das ferramentas

Fonte: Elaborado pelo autor (2021).

Após a delimitação da questão de pesquisa, consultou-se profissional especializado da área de biblioteconomia da Universidade de Brasília para sugestão de bases consolidadas no contexto da Ciência da Informação, sendo elencadas: 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.*

Foram definidos os seguintes termos para compor a expressão de busca em português e inglês, no singular e plural: Geração automática de metadado, *Automatic metadata generation*, Geração semi-automática de metadado, *Semi-automatic metadata generation*, Ferramenta, *Tool*, Técnica, *Technique*, Característica, *Feature*, Função, *Function*, Aplicação e *Application*.

Elaborou-se as sentenças de buscas utilizando os seguintes operadores lógicos:

- Sentença em Português: ((("Geração automática de metadado") *OR* ("Geração semi-automática de metadado")) *AND* (Ferramenta *OR* Técnica *OR* Característica *OR* Função *OR* Aplicação));
- Sentença em Inglês: ((("*Automatic metadata generation*") *OR* ("*Semi-automatic metadata generation*")) *AND* (*Tool OR Technique OR Feature OR Function OR Applications*)).

Realizaram-se atividades de pré-testes nas bases científicas para verificar se as sentenças deveriam passar por um processo de readequação, o que foi confirmado. Algumas bases não retornaram informações, sendo executadas alterações nas sentenças elaboradas, conforme orientação de busca/ajuda do próprio periódico e revista. Essa documentação pode ser consultada em <https://cutt.ly/pWiqjrN>. Concluído o pré-teste, finalizou-se a fase de planejamento e após a aprovação por especialista, passou-se a etapa de execução.

2.2 EXECUÇÃO

Nesta etapa executou-se buscas nas bases científicas utilizando os termos, sentenças e os operadores lógicos definidos na etapa de planejamento. Ao acessar a página de cada

periódico, realizou-se o preenchimento dos filtros de busca de forma a delimitar a recuperação da informação. Aplicou-se os seguintes critérios de inclusão e exclusão para obter as publicações:

Tabela 3 - Critérios 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
Descrição de estudo de caso, experimentos ou *survey*	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

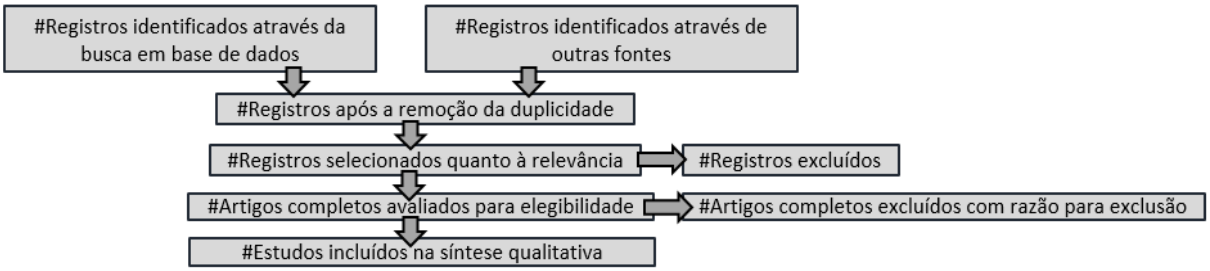
Fonte: Elaborado pelo autor (2021).

Duas bases de dados não retornaram resultado 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. Interessante comentar que a BRAPCI por ser uma base referencial em Ciência da Informação com artigos indexados desde 1972, não possui investigações relacionadas com o tema geração automática e semi-automática de metadados. Foi constatado que os filtros do formulário dessa plataforma apresentavam falhas. Ao utilizar qualquer termo de busca como teste, o critério de inclusão denominado: Trabalhos científicos publicados entre os anos de 2010 e 2020; retornava consultas desde 1972 e não no período selecionado. Vale registrar que essa atividade foi realizada no período de 6 a 16 de maio de 2020.

As outras bases de dados pesquisadas retornaram o total de 49 trabalhos científicos, utilizando as sentenças definidas no planejamento.

Para analisar os critérios mínimos de qualidade será utilizado um conjunto de itens com base em evidências, baseado em um *framework* elaborado por Moher *et al.* (2009) denominado *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA), adaptado para esta revisão bibliográfica, obedecendo o seguinte fluxograma:

Figura 2 – Fluxo de seleção de artigos, baseado no *framework* PRISMA.

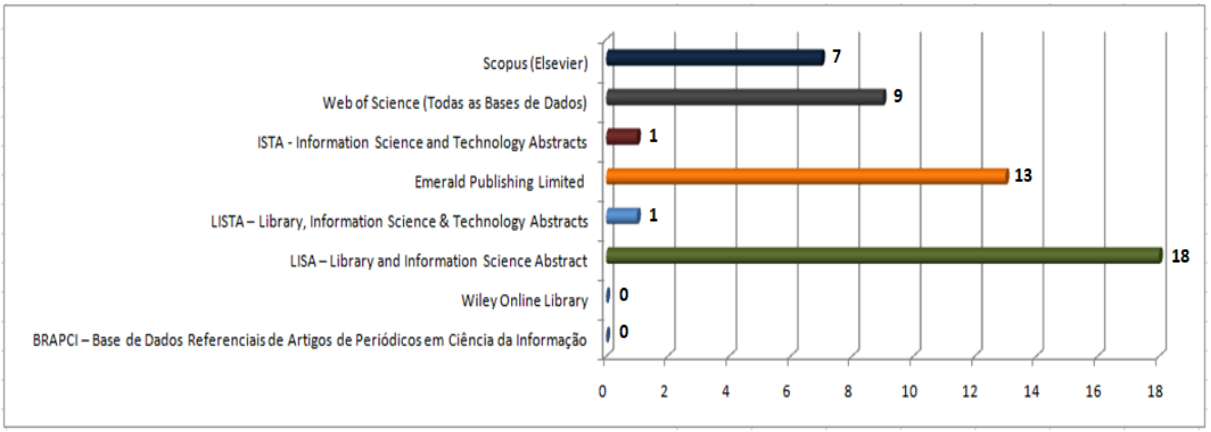


Fonte: Adaptado de Moher *et al.* (2009).

A primeira atividade desta etapa consistiu na leitura do título e descrição dos registros identificados através da busca nas bases de dados. Foram identificados 15 artigos duplicados, ou seja, na busca executada retornaram resultados em mais de uma base de dados para o mesmo artigo.

A segunda atividade foi a leitura de resumo, *abstract* e palavras-chave, sendo excluídos 16 artigos devido não serem úteis para continuidade da pesquisa. O total de 18 estudos foi selecionado para realização de análise. Entretanto, 6 artigos completos foram excluídos após uma leitura preliminar, pois o conteúdo dos trabalhos não correspondia ao objeto desta revisão bibliográfica. Por fim, 12 estudos foram incluídos na síntese qualitativa.

Figura 3 – Quantidade de artigos nas bases pesquisadas (N = 49).



Fonte: Elaborado pelo autor (2021).

Estabeleceu-se 04 (quatro) categorias de análise para facilitar a extração de dados dos estudos e corroborar com a execução da síntese: ferramentas de geração semi-automática de metadados; ferramentas de geração automática de metadados; extração automática de metadados e catalogação com ferramentas automatizadas. A documentação deste processo pode ser acessada através do endereço <https://cutt.ly/tWiDg3k>. Todas as tabulações das

publicações foi realizada utilizando planilha eletrônica Microsoft Excel® para apoiar a organização deste estudo.

Ao final desta etapa de execução, elaborou-se a Tabela 4, ordenando os registros pela data de publicação daqueles artigos selecionados que serão objeto de realização do fichamento e da síntese qualitativa.

Tabela 4 – Artigos selecionados para realização de fichamento e síntese.

DATA DE PUBLICAÇÃO	AUTORES	TÍTULO	NOTAS ADICIONAIS	LOCAL	PAÍS	INSTITUIÇÃO DE ORIGEM	CATEGORIA DE ANÁLISE
2011	Kovaevic *et al*.	*Automatic extraction of metadata from scientific publications for CRIS systems*	*Program: Electronic Library and Information Systems Vol. 45 No. 4, pp. 376-396*	Novi Sad	Sérvia	*Novi Sad University*	Extração automática de metadados
2012	Maratea A; Petrosino A; Manzo, M	*Automatic Generation of SCORM Compliant Metadata for Portable Document Format Files*	*International Conference on Computer Systems and Technologies – CompSysTech*	Nápoles	Itália	*Parthenope University*	Ferramenta de geração automática de metadados
2012	Verborgh *et al.*	*Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform*	*Multimed Tools Appl 61, 105–129*	Ghent	Bélgica	*Ghent University*	Ferramenta de geração automática de metadados
2012	Sah, M; Wade, V	*Automatic metadata mining from multilingual enterprise content*	*Web semantics: Science, services and agents on the world wide web, Vol 11, p. 41-62*	Dublin	Irlanda	*Trinity College Dublin*	Extração automática de metadados
2013	Costa *et al.*	*EURAC SDI: A Near Real Time and Offline Automatic Metadata Generation Processing Chain*	*GI Forum, Conference Proceedings, volume 1*	Bozen	Itália	*Eurac Research*	Ferramenta de geração automática de metadados
2013	Vlachidis *et al.*	*Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature*	*In: Przepiórkowski A., Piasecki M., Jassem K., Fuglewicz P. (eds) Computational Linguistics. Studies in Computational Intelligence, vol 458. Springer, Berlin, Heidelberg*	Londres	*United Kingdom*	*University College London*	Ferramenta de geração automática de metadados
2015	Park, J; Brenza, A	*Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art*	*Information Technology and Libraries, Volume 34, Ed. 3, p. 22-42*	Filadélfia	EUA	*Drexel University*	Ferramenta de geração semi-automática de metadados
2015	Rafferty, J; Nugent, C; Liu, J	*Automatic Metadata Generation Through Analysis of Narration Within In structional Videos*	*Transaction Processing Systems, J MedSyst nº 39, 94*	Belfast	Irlanda do Norte	*Ulster University*	Ferramenta de geração automática de metadados

2018	Gonzalo *et al.*	*ScienceSearch: Enabling Search through Automatic Metadata Generation*	*Conferência: 14th IEEE International Conference on E-Science (E-Science), p. 93-104*	Califórnia	EUA	*Berkeley University*	Ferramenta de geração automática de metadados
2018	Yang, G; Park, J	*Automatic Extraction of Metadata Information for Library Collections*	*International Journal of Advanced Culture Technology, Vol.6, nº 2, p. 117-122*	Filadélfia e Mokpo	EUA e Coréia do Sul	*Drexel University e Mokpo University*	Ferramenta de geração automática de metadados
2019	Audichya, M, K; Saini J, R	*Computational linguistic prosody rule-based unified technique for automatic metadata generation for Hindi poetry*	*1st International Conferenceon Advances in Information Technology*	Gujarat	Índia	*Gujarat Technological University*	Ferramenta de geração automática de metadados
2020	Morris, V.	*Automated Language Identification of Bibliographic Resources*	*Cataloging & Classification Quarterly, 58:1, 1-27*	Wetherby	*United Kingdom*	*British Library*	Catalogação com ferramentas automatizadas

Fonte: Elaborado pelo autor (2021).

Observando os artigos relacionados, nota-se que o assunto pesquisado é relevante e de interesse ao redor do mundo, pois se verifica investigações sobre o tema por diversos pesquisadores de várias universidades localizadas nos Estados Unidos, Europa e Ásia. A lacuna que se percebeu nesse levantamento foram pesquisas realizadas no contexto da Ciência da Informação no Brasil e na América Latina.

2.2 REALIZAÇÃO DA SÍNTESE

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 nos métodos de aprendizado de máquina, ou seja, classificação e obteve-se melhor resultado com o uso do modelo *Support Vector Machines*.

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

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 e usa a *web*semântica como fonte de conhecimento para iniciar e manter ativamente o contexto da tarefa. Os autores realizaram a aplicação através de um caso de uso de anotação de imagem. 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.

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.

Costa *et al* (2013) apresentam em seu estudo a abordagem para geração automática de metadados através 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*, *RapidEye*, *ENVISAT* e *Quickbird*. Eles executaram a ingestão e manuseio de dados, através de um aplicativo geral multitarefa centralizado, denominado *Data Exchange Server* (DES).

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

outros modelos de metadados. O enriquecimento semântico se caracteriza pelo processo de fornecimento de maior significado aos dados e metadados, auxiliando a integração, a compreensão e o processamento. Foi empregando o kit de ferramentas de Arquitetura Geral para Engenharia de Texto (GATE).

O artigo de Park e Brenza (2015) merece um destaque, pois a publicação foi a mais indexada nas bases pesquisadas. Eles examinam uma variedade de ferramentas de geração semi-automáticas de metadados (N=39), analisando suas técnicas, recursos e funções. Esses autores desenvolveram uma matriz caracterizando cada ferramenta de geração semi-automática de metadados, descrevendo: nome, local *online*, técnicas usadas para geração de metadados, além de breve exposição das funções e recursos da ferramenta. Realizou-se acesso a 32 aplicações no período de 11/07/2020 a 26/07/2020, sendo que algumas delas possuem *link* para a documentação, mas não ao projeto e ao *software* para *download*: *Apache Stanbol* foi movido para o *Apache Attic; Ont-O-Mat* foi descontinuado; *FreeCite* foi substituído por *AnyStyle*; *RepoMMan* não estava disponível, *Sherpa/Romeu* passou para o projeto *Sherpa* na versão 2; *Simple Automatic Metadata Generation Interface* (SAMGL) não estava disponível; *Yahoo Content Analisys* API e o *plugin* do *Firefox Dublin Core Viewer Extension* foram descontinuados.

A pesquisa de Rafferty *et al.* (2015) apresenta 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.

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.

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.

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*[1]. 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.

Morris (2020) investigou se os códigos de idioma podem ser atribuídos automaticamente aos registros do MARC e avaliar a precisão de qualquer método viável de fazê-lo. A autora enfatiza que nos registros bibliográficos do MARC21, o conteúdo de idioma de um recurso de informação é registrado nas posições do campo 008 35–37, usando um

[1]Quadra/estrofe usada nas tradições poéticas do Norte da Índia e Paquistão

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.

3 CONSIDERAÇÕES FINAIS

Conclui-se que as ferramentas de geração automática e semi-automática de metadados descritas na síntese possuem diversas aplicabilidades na qual citamos: produção de metadados para conteúdos de aprendizagem, resolução de problemas semânticos, fornecimento de informações avançadas para personalização de um sistema para o cliente, disponibilização de metadados geoespaciais para tomada de decisão, melhorar o significado dos metadados de objetos vinculados em uma biblioteca digital, fornecer instruções dinâmicas a partir de clipes de vídeo, geração e inclusão de metadados faltantes em repositórios com grandes volumes de dados, auxiliar na atenuação de problemas de interoperabilidade semântica, gerar metadados de escritos antigos e apoiar a catalogação, atribuição automática de registros em um recurso de informação, dentre outras.

Esta pesquisa possui limitação identificada a posteriori que foi a não inserção do termo de busca “indexação automática” que é um método usado em ferramentas de geração automática de metadados, o que corroborou para ausência neste trabalho de estudos desenvolvidos no Brasil, especificamente na BRAPCI, fato reconhecido por estes pesquisadores ao final desta investigação.

Há lacuna existente na literatura acadêmica de criação de um modelo de referência para geração de metadados para uso geral. Observou-se na pesquisa que as ferramentas de geração automática e semi-automática de metadados são desenvolvidas para atender necessidades muito específicas.

Sugere-se pesquisas que mesclem funcionalidades de ferramentas de geração automática e semi-automáticas 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 catalogação e suporte para os gestores de repositórios digitais em suas tarefas.

REFERÊNCIAS

AUDICHYA, Milind; SAINI Jatinderkumar R. **Computational linguistic prosody rule-based unified technique for automatic metadata generation for Hindi poetry**. *In*: International Conference on Advances in Information Technology (ICAIT), 2019, Chikmagalur. p. 436-442.

COSTA, Armin; FIRDAUSY, Tania P; INNEREBNER, Markus; MONSORNO, Roberto. **EURAC SDI:** A Near Real Time and Offline Automatic Metadata Generation Processing Chain. *In*: Conference Proceedings, 2013, Salzburg, v. 1, p. 2-5.

CRYSTAL, Abe; LAND, Paula. **Metadata and Search: Global Corporate Circle DCMI 2003 Workshop**. 2003. Disponível em http://www.dublincore.org/groups/corporate/Seattle/. Acesso em: 24 jul. 2020.

CUNHA, Murilo Bastos da; CAVALCANTI, Cordélia Robalinho de Oliveira. **Dicionário de biblioteconomia e arquivologia**. Brasília: Briquet de Lemos, xvi, 451 p. 2008.

GONZALO, P. Rodrigo; MATT, Henderson; GUNTHER, H. Weber; COLIN, Ophus; KATIE, Antypas; LAVANYA, Ramakrishnan. **ScienceSearch:** Enabling Search through Automatic Metadata Generation**.** *In:* IEEE 14th International Conference on e-Science, 2018, Amsterdam, p. 93-104.

GREENBERG, Jane. **Metadata Extraction an Harvesting:** a comparison of two automatic metadata generation applications. Journal of Internet Cataloging, v. 6, (4), 2003.

HAYNES, David. **Metadata for Information Management and Retrieval**: understanding metadata and its use. 2ª Edição, *London*: *Facet Publishing*, 2018.

KOVACEVIC, Aleksandar; IVANOVIC, Dragan; MILOSAVLJEVIC, Branko; KONJOVIC, Zora. **Automatic extraction of metadata from scientific publications for CRIS systems.** Program: *Electronic Library and Information Systems*, v. 45, nº. 4, p. 376-396, 2011.

MARATEA, Antonio; PETROSINO, Alfredo; MANZO, Mario. **Automatic Generation of SCORM Compliant Metadata for Portable Document Format Files**. *In*: Proceedings of the 13th International Conference on Computer Systems and Technologies, CompSysTech, p. 360-367, 2012.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 5ª ed., São Paulo: Atlas, 2003.

MEY, Eliane Serrão Alves; SILVEIRA, Naira Christofoletti. **Considerações teóricas aligeiradas sobre a catalogação e sua aplicação**. *InCID*: R. Ci. Inf. e Doc., Ribeirão Preto, v. 1, n.1, p. 125-137, 2010.

MOHER, David; LIBERATI, Alessandro; TETZLAFF, Jennifer; ALTMAN, Douglas G. **Preferred Reporting Items for Systematic Reviews and Meta Analyses:** The PRISMA Statement. The PRISMA Group. PLoSMed 6(7): e1000097. doi:10.1371/journal.pmed.1000097, 2009.

MOOERS, Calvin N. **Zatocoding applied to mechanical organization of knowledge**. American Documentation, v. 2, n. 1, p. 20-32, 1951.

MORRIS, Victoria. **Automated Language Identification of Bibliographic Resources**. *Cataloging & Classification Quarterly*, v. 58, issue 1, p. 1-27, 2020.

PARK, Jung-ran; BRENZA, Andrew. **Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art**. Information Technology and Libraries, v. 34, ed. 3, p. 22-42, Chicago, 2015.

POMERANTZ, Jeffrey. **Metadata**. Cambridge, MA: The MIT Press, 2015.

RAFFERTY, Joseph; NUGENT, Chris; LIU, Jun; CHEN, Liming. **Automatic Metadata Generation Through Analysis of Narration Within Instructional Videos.** Transaction Processing Systems**,** Jornal of Medical System, v. 39, nº 94, 2015.

REINSEL, David; GANTZ, John; RYDNING, John. **The Digitization of the World:** From Edge to Core. Data Age 2025, An IDC White Paper, November, 2018.

SAH, Melike; WADE, Vincent. **Automatic metadata mining from multilingual enterprise content**. Journal of Web Semantics First Look, v. 11, p. 41-62, 2012.

VERBORGH, Ruben; VAN DEURSEN, Davy; MANNENS, Erik; POPPE, Chris; WALLE, Rik Van de. **Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform**. Multimed Tools Appl, v. 61, p. 105–129, 2012.

VLACHIDIS, Andreas; BINDING, Ceri; TUDHOPE, Douglas; MAY, Keith. **Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature**. *In*: Przepiórkowski A., Piasecki M., Jassem K., Fuglewicz P. (eds) Computational Linguistics. Studies in Computational Intelligence, v. 458. Springer, Berlin, Heidelberg, 2013.

YANG, Gi-Chul; PARK, Jeong-Ran. **Automatic Extraction of Metadata Information for Library Collections.** International Journal of Advanced Culture Technology, v. 6, nº 2, p. 117-122, 2018.