Avaliação de Ferramentas de Extração Automática de Metadados na Catalogação de Artigos Científicos do CONNEPI

Igor Matheus B. Moreira[1], Mônica Ximenes C. da Cunha[1]

[1]Departamento de Informática –
Instituto Federal de Educação, Ciência e Tecnologia de Alagoas (IFAL)

{igor.mbm, mxcc}@hotmail.com

Abstract. *This article describes the search and test steps of computational solutions for extracting and automatically cataloging scientific publication metadata for the CONNEPI event repository since its origin in 2006. A systematic literature review was performed to identify the most commonly used tools. Then, comparative tests were performed between three of them: Cermine, Grobid and pdfx. The results did not signal to a predominant tool, with a high percentage of correct answers in all the predefined metadata. Therefore, the next step of the research is to adapt an extraction tool to the reality of headings formats of the CONNEPI publications.*

Resumo. *Este artigo descreve as etapas de busca e teste de soluções computacionais para extração e catalogação automática de metadados de publicações científicas para o repositório do evento CONNEPI, desde a sua origem, em 2006. Foi realizada uma revisão sistemática de literatura para identificar as ferramentas mais utilizadas. Em seguida foram realizados testes comparativos entre três delas: Cermine, Grobid e pdfx. Os resultados não sinalizaram para uma ferramenta predominante, com alto percentual de acertos em todos os metadados predefinidos. Assim sendo, a próxima etapa da pesquisa está sendo adaptar uma ferramenta de extração para a realidade de formatos de cabeçalhos das publicações do CONNEPI.*

1. Introdução

Os Institutos Federais (IFs) da Rede Norte e Nordeste têm realizado anualmente, desde 2006, o Congresso Norte e Nordeste de Pesquisa e Inovação (CONNEPI). A cada edição deste evento, os anais têm sido disponibilizados de forma individualizada. Cada IF responsável gerou uma mídia digital, como foi o caso das primeiras edições do evento, ou um link avulso da internet, que muitas vezes se encontra indisponível para acesso, ou ainda organizou os anais em um link no site do evento. Um cenário que mostra o grau de fragmentação da informação, que acarreta em dificuldade de localização, acesso, disseminação das publicações e geração de estatísticas.

Moura e Santos (2016) construíram um repositório para agregar essas publicações do CONNEPI. O layout e a codificação foram concluídos. O repositório foi testado com o povoamento manual de artigos e respectivos metadados das edições 2006 a 2010. No entanto, devido à quantidade considerável de artigos das demais edições, enxergou-se que a extração manual dos metadados seria uma tarefa morosa e repetitiva.

Assim sendo, surgiu a necessidade de busca, teste e adaptação de uma ferramenta de extração e catalogação automática de dados para a conclusão do povoamento do repositório, que consiste em eventos que ocorreram entre 2011 e 2018. Este trabalho aborda as duas primeiras etapas supramencionadas.

A catalogação é considerada uma das mais importantes operações para o perfeito funcionamento de um sistema de recuperação de documentos. De acordo com Santiago (2004), a catalogação consiste em um processo no qual o documento é identificado por elementos bibliográficos, tais como nomes dos autores, título, palavras-chave, instituição de origem, fontes de publicação, dentre outros dados que se julgar necessários. Assunção (2005), por sua vez, apresentou a catalogação como "um processo através do qual se descreve formalmente um documento ou recurso e se estabelece um número variado e variável de pontos de acesso com o objetivo de proporcionar, ao usuário final, a possibilidade de encontrar, identificar, selecionar e obter o documento ou recurso descrito ou a informação nele contida".

Enfim, a catalogação consiste, portanto, em um caminho facilitador para a localização das informações desejadas. Além de assumir grande importância como atividade, que inclui não apenas a descrição bibliográfica do documento, mas toda sua representação descritiva. Mey (1995) descreveu as seguintes funções da catalogação: 1) Permitir ao usuário: a) localizar um item específico, b) escolher entre as várias manifestações de um item, c) escolher entre vários itens semelhantes, sobre os quais, inclusive, possa não ter conhecimento prévio algum, d) expressar, organizar ou alterar sua mensagem interna; 2) Permitir a um item encontrar seu usuário; 3) Permitir a outra biblioteca: a) localizar um item específico, b) saber quais os itens existentes em acervos que não o seu próprio.

O sucesso na pesquisa e localização de um documento, ou de um conjunto de documentos relevantes para um determinado utilizador, dependente muito da qualidade e da consistência da informação descritiva disponível para a pesquisa. Em outras palavras, encontrar e adaptar uma solução computacional que se encaixe na problemática abordada é imprescindível para o bom funcionamento da catalogação.

A informação descritiva de um documento é conhecida como metadados. A definição mais utilizada para a palavra metadados é dados sobre dados. Metadados consistem nas informações criadas, armazenadas e compartilhadas para descrever coisas e permitem a interação com essas coisas para obter o conhecimento que se deseja (RILEY, 2017). São dados estruturados que descrevem as características de uma fonte e que compartilham características muito similares para catalogar dados que são obtidos em bibliotecas, museus e arquivos (KOWATA, 211). Segundo Ikematu (2001) os metadados são dados que descrevem atributos de um recurso e suportam várias funções: localização, descoberta, documentação, avaliação, seleção, etc. Enfim, metadados são informações estruturadas que descrevem, explicam, localizam, ou seja, tornam fácil recuperar, usar ou gerenciar uma fonte de informações.

Os metadados podem ser classificados em três tipos (Dos Santos, 2011): (1) metadados descritivos descrevem uma fonte de informação para fins de identificação e recuperação utilizando elementos como título, autor, resumo e palavras-chave, (2) metadados estruturados descrevem a organização interna dos objetos e das relações entre eles; o exemplo mais comum é o esquema do banco de dados, e (3) metadados

administrativos apoiam as atividades de gerenciamento do acervo de recursos de informação como controle de permissões de acesso, localização de arquivos e critérios de avaliação da qualidade. No caso desta pesquisa, a busca é pelos metadados do tipo descrito.

É nessa vertente que este trabalho se posiciona, ao enxergar os entraves que ainda existem para a ampla visibilidade das informações disseminadas pelo CONNEPI e tentar solucioná-los por meio de pesquisas e adaptações de soluções computacionais que possam extrair metadados predefinidos e posteriormente realizar a devida catalogação.

Este artigo está organizado da seguinte forma: a seção 2 descreve o procedimento metodológico utilizado na pesquisa; a seção 3 descreve os resultados obtidos e a seção 4 sinaliza para as conclusões.

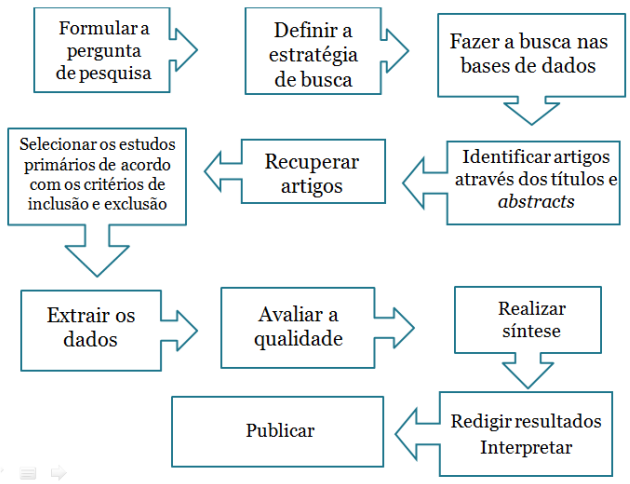
2. Metodologia

Esta pesquisa aplicada adota postura epistemológica positivista, com abordagem quantitativa. A pesquisa quantitativa está centrada na objetividade. Influenciada pelo positivismo, considera-se que a realidade só pode ser compreendida com base na análise dos dados brutos, recolhidos com o auxílio de instrumentos padronizados e neutros (FONSECA, 2002).

A primeira etapa da pesquisa consistiu na realização de uma revisão sistemática de literatura (RSL) para identificar ferramentas de extração já existentes, com código-fonte disponível. Em seguida, as ferramentas mais mencionadas nos artigos científicos, encontrados por meio da RSL, foram selecionadas. Após a seleção, foram realizados testes de extração com amostra de artigos de cinco edições do CONNEPI para averiguar qual a mais adequada ao problema desta pesquisa.

A figura 1 exibe todo o processo da revisão sistemática de literatura percorrido por esse projeto.

Figura 1. Processo da Revisão Sistemática de Literatura (RSL).



Fonte: Dados da Pesquisa

O primeiro passo da RSL foi definir as perguntas de pesquisa: 1) Quais ferramentas open source de extração de metadados existentes e testadas? e 2) Com quais tipos de documentos essas ferramentas trabalham?. Em seguida foram definidas as

strings de busca e as 04 bases de dados iniciais: Google Acadêmico, IEEE Digital Library, ACM Digital Library e Science Direct. Durante a aplicação das strings, três outras bases de dados foram adicionadas no escopo da pesquisa. A relação completa das bases e respectivas strings está disponível no protocolo da RSL, no link: https://bit.ly/2Z3bjem. O quadro 1 exibe as informações finais dessa etapa.

Quadro 1. Base de dados e respectivas strings de busca.

Base de Dados	String
Google Acadêmico (Português)	(extração de metadados or extração de dados or extração automática) and (ferramentas) and (documentos or pdf)
Google Acedêmico (Inglês)	(metadata extraction or data extraction or information extraction or automatic extraction) and (tools or system or algorithms or machine learning) and (open source) and (documents or pdf or scientific literature) and (abstract) filetype:pdf
Bibliioteca Digital UFMG	((Metadados) AND (extração automática) AND (Literatura científica))
Lume UFRGS	metadados) and (extração automática) and (ferramentas) and (literatura científica)
Arxiv	((abs:metadata AND abs:(automatic AND extraction)) AND abs:tools)
IEEE Digital LIbrary	(metadata) and (automatic extraction) and (tools) and (scientific documents or pdf)
ACM Digital Library	(metadata) AND (automatic extraction) AND (tools) AND (scientfic document or pdf)
Science Direct	(metadata extraction or information extraction or automatic extraction) and (tools) and (open source) and (documents or pdf or scientific literature)

Fonte: Dados da Pesquisa.

Com o objetivo de selecionar os melhores trabalhos científicos para serem adotados na filtragem do montante inicial, alguns critérios de seleção foram pensados. O quadro 2 exibe os critérios de inclusão e exclusão inseridos na atual pesquisa.

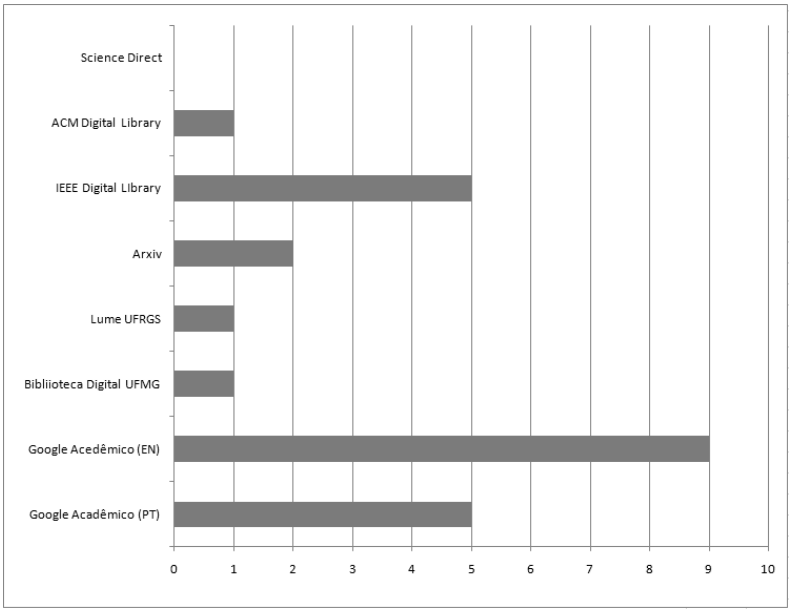
Quadro 2. Critérios de inclusão e exclusão de artigos.

Critérios de Inclusão	Critérios de Exclusão
Trabalhos científicos publicados entre os anos de 2007 e 2017	Trabalhos científicos publicados antes de 2006 e depois de 2017
Citação de uma ou mais métodos de extração de artigos científicos	Título do trabalho não condizente com a proposta do projeto
Possível disponibilização ou acesso do código da ferramenta	Resumo do trabalho científico com fuga do tema proposto pelo projeto
Apenas trabalhos científicos (artigo, monografia, dissertação, tese, etc)	Aborda apenas ferramentas de extração de citações bibliográfica
	Ferramentas de extração que necessitam de licença de uso

Fonte: Dados da Pesquisa

A aplicação das strings nas respectivas bases de dados retornou um total de 7117 trabalhos científicos. A primeira etapa de seleção consistiu em uma análise dos títulos de cada trabalho. Foram selecionados 160 trabalhos para a próxima etapa, que consistiu da leitura do resumo. Por fim, 30 estudos científicos foram classificados como úteis para a continuidade do projeto. Esse quantitativo final ainda sofreu uma redução para 24, pois foi identificado que 6 deles eram livros sobre o tema pesquisado. Porém, a atual pesquisa tem como foco apenas trabalhos científicos. O gráfico 1 mostra o quantitativo dos trabalhos resultantes, separados por base.

Gráfico 1. Número de artigos selecionados em cada base de dados.



Fonte: Dados da Pesquisa.

De acordo com o gráfico 1, é possível perceber que o Google Acadêmico (EN) teve o maior número de trabalhos científicos utilizados na pesquisa, totalizando-se 9. Seguido por IEEE e Google Acadêmico (PT) com 5 trabalhos cada. Arxi logo em seguida com 2, e por fim, com 1 trabalho cada, os repositórios da UFRGS e UFMG, além da biblioteca ACM. Vale ressaltar que 3 trabalhos do IEEE e 1 trabalho do Arxiv também foram encontrados no Google Acadêmico (EN), mas como o link de acesso era dos próprios repositórios citados, eles foram contabilizados apenas pelos mesmos e não pelo Google Acadêmico.

Após a leitura e coleta de informações dos estudos científicos selecionados por meio da RSL, ocorreu a segunda etapa da pesquisa, que consistiu dos testes das ferramentas mais mencionadas pelos autores. As três ferramentas selecionadas foram submetidas a testes de extração de uma amostra de 50 artigos do CONNEPI, entre as edições de 2011 e 2015. Inicialmente, os respectivos metadados foram extraídos manualmente, consistindo de uma base de resultados desejados, o que permitiu comparar com os resultados obtidos na extração automática por parte de cada uma das ferramentas.

3. Resultados

Com base na leitura dos artigos levantados pela RSL, foram identificadas várias ferramentas de extração de metadados. Além das ferramentas, propriamente ditas, que executam a atividade de forma automática, foram identificadas formas alternativas de extração, como o caso das bibliotecas open source citadas no trabalho de Barbosa (2016) que propõem um modelo para extrair conhecimento de artigos científicos, através de uma rede complexa formada pelos atributos e seus relacionamentos, e que utiliza a biblioteca Apache PDFBox para executar a etapa de extração. Também foi localizado o trabalho de Bodó et al. (2017), afirmando que as ferramentas mais

populares usadas para obter os fragmentos de texto de documentos PDF - para Java, C # e Python - são as bibliotecas PDFBox, iText e PDFMiner.

Outro método de extração estudado durante o projeto foi a técnica baseada em regras, citada por Guo et al. (2011), que apresenta uma estrutura baseada em regras para a extração automática de metadados. Esse trabalho avalia o desempenho do esquema proposto no SemreX, um sistema de compartilhamento de literatura semântica baseado em P2P que oferece serviços de compartilhamento de literatura entre pesquisadores de ciência da computação.

Como o objetivo da pesquisa era obter as ferramentas que estavam disponíveis e com o código aberto em repositório público, nos limitamos ao estudo aprofundado de algumas que atenderam tal critério. O trabalho de Grossi Junior (2016) comparou a capacidade de extração de metadados de algumas ferramentas pré-selecionadas - Cermine, CiteSeer, CrossRef e ParsCit - utilizando um experimento empírico com um conjunto de artigos. Tkaczyk et al. (2014), por sua vez, avaliaram o Cermine em comparação com Grobid, Pdfx, Parscit e Pdf-Extract, exibindo resultados satisfatórios para a mesma.

Williams et al. (2016) citaram algumas ferramentas open source: SVMHeaderParse, Grobid, parsCit. Singh at al. (2016) realizou a comparação dos resultados de um framework open source (OCR++) com os resultados do Grobid.

A partir da leitura das publicações científicas supracitadas foi possível responder a questão de pesquisa: Quais ferramentas open source de extração de metadados existentes e testadas?. A tabela 1 apresenta o número de citações de cada ferramenta nos trabalhos selecionados.

Tabela 1. Número de citações de cada ferramenta de extração automática.

Ferramentas de Extração	Quantidade de artigos que as citam
Grobid	5
SVMH	4
Cermine	3
ParsCit	3
PDFX	2
Apache PDFBox	2
TeamBeam	2
Itextpdf	2
Artic	2
Mendeley	2
Flux-cim	1
Font Box	1
Apache Lucene Library	1
OCR++	1
PDFExtract	1
Docear's PDF Inspector	1
PDFMeat	1
Sciplore Xtract	1
PDFSSA4MET	1
CiteULike	1
CiteSeer	1
CrossRef	1
PDFMiner	1

Fonte: Dados da Pesquisa.

Já a segunda questão de pesquisa definida para esta RSL: Quais documentos essas ferramentas trabalham?, pode ser facilmente respondida uma vez que todas as ferramentas testadas e identificadas durante a pesquisa utilizaram o documento PDF como o objeto da extração. O Portable Document Format (PDF) é um formato de arquivo que foi criado com o objetivo inicial de ser independente de aplicativo, hardware e sistema operacional (ROSENTHOL, 2013).

Bodó et al. (2016) afirmou que embora artigos acadêmicos possam ser encontrados em uma grande variedade de formatos na Internet, o Portable Document Format é sem dúvida o mais popular. Portanto, é suficiente considerar este formato para extração de metadados. Barbosa (2016), por sua vez, propôs um modelo cujo processo de captura consiste em extrair atributos de artigos acadêmicos, mais especificamente no formato PDF, que é o tipo mais comum e difundido na comunidade acadêmica.

De acordo com Bast et al. (2017), extrair o texto do corpo de um documento PDF é uma tarefa importante, mas surpreendentemente difícil, e que a razão para tal dificuldade é que o PDF é um formato baseado em layout que especifica as fontes e posições dos caracteres individuais em vez das unidades semânticas do texto (por exemplo, palavras ou parágrafos) e seu papel no documento (por exemplo, texto do corpo ou legenda). Para Souza (2014) o padrão PDF permite a geração de metadados incluídos no documento. No entanto, muitos autores não definem essas informações, tornando esse recurso pouco confiável ou incompleto. Este fato tem motivado pesquisas que visam extrair metadados automaticamente.

Após finalizar a etapa de leitura e coleta de informações dos estudos científicos, encontrados por meio da RSL, se seguiu para a etapa de teste das ferramentas, para isso foram escolhidas as mais mencionadas pelos autores: CERMINE, Grobid e pdfx. As selecionadas foram submetidas a testes de extração com uma amostra de 50 artigos do CONNEPI, sendo dez de cada uma das edições de 2011 a 2015.

Inicialmente, os respectivos metadados foram extraídos manualmente, o que permitiu comparar os resultados com os obtidos automaticamente por cada uma das ferramentas e assim atestar qual delas se mostrou mais adequada ao contexto do CONNEPI. Os 50 artigos foram testados em cada ferramenta. A tabela 2 apresenta os resultados agregados da extração, sinalizados por três categorias: sucesso, extração parcial e falha (na extração dos metadados).

Tabela 2. Resultados dos testes das ferramentas selecionadas.

	Cermine			Grobid			Pdfx		
	Sucesso	Parcial	Falha	Sucesso	Parcial	Falha	Sucesso	Parcial	Falha
Título	64%	10%	26%	84%	14%	2%	**86%**	0%	14%
Autor	**56%**	16%	28%	34%	52%	14%	0%	76%	24%
Instituição	16%	30%	54%	**22%**	48%	30%	2%	16%	82%
Resumo	**56%**	38%	6%	26%	64%	10%	36%	58%	6%
Palavra-chave	28%	6%	66%	30%	0%	70%	**42%**	4%	54%

Fonte: Dados da pesquisa.

Ao analisar os dados da tabela 2, é possível identificar que no teste de extração com os artigos da base do CONNEPI não houve uma ferramenta predominante. Os melhores resultados obtidos para cada metadado, a saber: título, autor, instituição, resumo e palavra-chave, que podem ser visualizados no campo “Sucesso”, se alternam entre as ferramentas.

Outro aspecto importante a ser ressaltado é a alta porcentagem dos resultados no campo “Parcial”. A parcialidade na extração foi considerada em dois casos. O primeiro, quando dois metadados são extraídos e alocados em apenas um campo, como exibido no quadro 3, no qual o metadado “Autor” foi extraído juntamente com o metadado “Instituição”. E também no quadro 4, quando os metadados “Resumo” e “Palavra-chave” foram extraídos em apenas um campo. Logo, a extração de “Autor” para quadro 3 e “Resumo” para o quadro 4 foram consideradas parciais.

Quadro 3. Situações de parcialidade 1.

	Extração Manual	PDFX
Título	A CONSTRUÇÃO CÊNICA POR MEIO DA MEMÓRIA CULTURAL	A CONSTRUÇÃO CÊNICA POR MEIO DA MEMÓRIA CULTURAL
Autor	A. C. M. FREITAS[1] e M. D. OLIVEIRA[2]	A. C. M. FREITAS 1 e M. D. OLIVEIRA 2 **Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte – Campus Caicó** 2
Instituição	Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte – Campus Caicó	X

X = Falha na extração

Fonte: Dados da Pesquisa.

Quadro 4. Situações de parcialidade 2.

	Extração Manual	Cermine
Resumo	Este estudo tem como foco a Educação Ambiental (EA) e sua relação com o ensino de Ciências. Apresenta como objetivo geral revelar como a Educação Ambiental está inserida dentro do conteúdo programático do componente curricular de Ciências do Ensino Fundamental, de uma escola privada no interior do Rio Grande do Norte (RN). Ao observar as mais diferentes definições, a Educação Ambiental consiste em um processo permanente de formação que busca desenvolver conhecimentos e habilidades na intenção de favorecer ao homem à aquisição de valores e o desenvolvimento de atitudes necessárias para lidar com as questões ambientais e consolidar respostas sustentáveis ao meio. Desta maneira, a partir dos estudos de Amaral (1995); Maknamara (2009) e Reigota (1999) foi possível compreender como se dá a interação da EA com a disciplina em questão, respectivamente pelas modalidades básicas, o seu desenvolvimento e as concepções empregadas pelos docentes. A pesquisa foi desenvolvida de acordo com a abordagem qualitativa, onde o cenário na qual desenvolveu-se foi a Escola Nossa Senhora das Graças, instituição privada, localizada na cidade de Florânia. Participaram da investigação 10 jovens do nono ano do ensino fundamental. O instrumento utilizado para a construção dos dados foi o questionário. A análise mostrou que 80% dos alunos participantes indicaram as Ciências Naturais como sendo uma das disciplinas pela qual eles mais se interessavam. Dos temas acima que envolvem a disciplina, com 80% de aceitação, “Meio Ambiente” foi o que os alunos disseram ter mais afinidade.	Este estudo tem como foco a Educação Ambiental (EA) e sua relação com o ensino de Ciências. Apresenta como objetivo geral revelar como a Educação Ambiental está inserida dentro do conteúdo programático do componente curricular de Ciências do Ensino Fundamental, de uma escola privada no interior do Rio Grande do Norte (RN). Ao observar as mais diferentes definições, a Educação Ambiental consiste em um processo permanente de formação que busca desenvolver conhecimentos e habilidades na intenção de favorecer ao homem à aquisição de valores e o desenvolvimento de atitudes necessárias para lidar com as questões ambientais e consolidar respostas sustentáveis ao meio. Desta maneira, a partir dos estudos de Amaral (1995); Maknamara (2009) e Reigota (1999) foi possível compreender como se dá a interação da EA com a disciplina em questão, respectivamente pelas modalidades básicas, o seu desenvolvimento e as concepções empregadas pelos docentes. A pesquisa foi desenvolvida de acordo com a abordagem qualitativa, onde o cenário na qual desenvolveu-se foi a Escola Nossa Senhora das Graças, instituição privada, localizada na cidade de Florânia. Participaram da investigação 10 jovens do nono ano do ensino fundamental. O instrumento utilizado para a construção dos dados foi o questionário. A análise mostrou que 80% dos alunos participantes indicaram as Ciências Naturais como sendo uma das disciplinas pela qual eles mais se interessavam. Dos temas acima que envolvem a disciplina, com 80% de aceitação, “Meio Ambiente” foi o que os alunos disseram ter mais afinidade. **Palavras-chave: educação ambiental, ensino de ciências, ensino fundamental.**
Palavra-chave	educação ambiental, ensino de ciências, ensino fundamental.	X

X = Falha na extração

Fonte: Dados da Pesquisa.

O segundo caso diz respeito, particularmente, à extração do metadado “Autor” com informações incorretas. Como é o caso do resultado exibido no quadro 5, em que um autor foi considerado pela ferramenta de extração como duas pessoas distintas.

Quadro 5. Situações de parcialidade 3.

	Extração Manual	Grobid
Autor	Francisco Camilo da Silva[1], Maria Icleide Viana da Silva[2], Lesso Benedito dos Santos[3]	<author> <persName <forename type="first">Maria</forename> <surname>Icleide Viana Da Silva</surname> </persName> </author> **<author> <persName <forename type="first">Lesso</forename> <surname>Benedito</surname> </persName> </author> <author> <persName <surname>Santos</surname> </persName> </author>**

Fonte: Dados da Pesquisa.

4. Conclusões

Com a proposta de solucionar os problemas de fragmentação enfrentado pelo CONNEPI, a presente pesquisa teve como objetivo utilizar ferramentas automáticas de extração de metadados para catalogar os artigos publicados nas edições de 2011 a 2015, e assim realizar o povoamento dos mesmos no repositório digital.

Inicialmente, foi adotada a revisão sistemática de literatura (RSL) como metodologia para a primeira etapa da pesquisa. A RSL teve o papel de permitir a evolução dos estudos de modo organizado, obtendo ao final do processo um conjunto seleto de trabalhos científicos mais condizentes com a temática do projeto.

Como resultado da revisão sistemática, foi possível ter em mãos estudos acadêmicos que apresentavam, testavam ou comparavam ferramentas open source de extração automática de metadados. As três ferramentas mais mencionadas pelos autores(as) foram selecionadas para a próxima etapa.

Após comparar os resultados que as ferramentas obtiveram na extração, utilizando artigos do CONNEPI como amostra, foi possível atestar que não houve uma ferramenta predominante, mas sim, resultados considerados satisfatórios, para cada metadado em questão, alternados entre elas.

Identificou-se também, que a parcialidade na extração ocorreu com bastante frequência. O que deixou como questionamento, como seria os resultados da extração se os problemas de parcialidade fossem solucionados? Assim sendo, um trabalho em andamento está sendo adaptar uma das ferramentas para obter melhores resultados. Além de realizar novos testes, separadamente, para cada edição do CONNEPI. Já que cada edição apresentada até o momento possui formatos de padronização dos artigos distintos.

5. Referências

ASSUNÇÃO, Maria Clara Rabanal da Silva. Catalogação de documentos musicais escritos: uma abordagem à luz da evolução normativa. 2005. 128f. Dissertação (Mestrado em Ciências documentais). Universidade de Évora, Évora, 2005.

BARBOSA, Leonardo Maia. Um modelo para extrair conhecimento de artigos científicos utilizando redes complexas. 2016. Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, 2016.

BAST, Hannah; KORZEN, Korzen. 2017. A Benchmark and Evaluation for Text Extraction from PDF. In Proceedings of Joint Conference On Digital Libraries, Toronto, Ontario, Canada, June 2017 (JCDL'17).

BODO, Zalan; CSATO, Lehel. A Hybrid Approach for Scholarly Information Extraction. Studia Universitatis Babeș-Bolyai Informatica, [S.l.], v. 62, n. 2, p. 5-16, dec. 2017.

DOS SANTOS, V. Uma arquitetura suportada por busca semântica para recuperação de fontes de informação em repositórios de metadados. Dissertação de Mestrado. Programa de Pós-Graduação em Informática, Universidade Federal do Estado do Rio de Janeiro, 2011.

FONSECA, João José Saraiva da. **Metodologia da pesquisa científica**. Ceará: Universidade Estadual do Ceará, 2002.

GROSSI JÚNIOR, José Alberto. Análise comparativa de ferramentas de extração de metadados em artigos científicos. 2016. 84f. Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação. Belo Horizonte, 2016.

GUO, Z.; Jin, H. Reference Metadata Extraction from Scientific Papers. 12th International Conference on Parallel and Distributed Computing, Applications and Technologies, PDCAT 2011, Gwangju, Korea, October 20-22, 2011.

IKEMATU, R. S. Gestão de metadados: sua evolução na tecnologia da informação. Data Grama Zero - Revista de Ciência da Informação, 2(6), 2001.

KOWATA, ELISABETE TOMOMI. Metadados de Bancos de Dados Relacionais: Extração e Exposição com o Protocolo OAI-PMH. Dissertação de Mestrado. 2011. 127 p. Programa de Pós– Graduação em Ciência da Computação - Instituto de Informática da Universidade Federal de Goiás.

MANICA, Edimar; CERVI, Cristiano Roberto; GALANTE, Renata de Matos. Um Processo Automático para Extração de Metadados de Documentos PDF Usando um Template XML. In: Escola Regional de Banco de Dados (ERBD 2008), 4, 2008. Anais... Disponível em http://download.docslide.net/documents/um-processo-automatico-para-extracao-de-metadados-dedocumentos-pdf-usando.html. Acesso em 18 de junho de 2017.

MEY, Eliane Serrão Alves. Introdução à catalogação. Brasília. Briquet de Lemos, 1995.

MOURA, F. R. E.; SANTOS, L. G. C. Desenvolvimento de um Repositório Digital para armazenar as Publicações Científicas do CONNEPI. In: Congresso Norte e Nordeste de Pesquisa e Inovação dos Institutos Federais (CONNEPI), 11, 2016. Anais... Maceió, 2016.

RILEY, Jenn. UNDERSTANDING METADATA - WHAT IS METADATA AND WHAT IS IT FOR? National Information Standards Organization (NISO), 2017. Disponível em http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf. Acesso em 18 de junho de 2017.

ROSENTHOL, l. Developing with PDF: dive into the portable document format. 1. ed. [S.1]: O´REILLY, 2013.

SANTIAGO, M. C. C. Metadados para recuperação da informação em ambiente virtual. Dissertação (Mestrado em Ciência da Informação) Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de Rio de Janeiro, RJ, 2004.

SOUZA, Alan Pinto. Metadata **extraction from Scientific Documents in PDF.** 59 f. Dissertação de Mestrado – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre: PPGC da UFRGS, 2014.

TKACZYK, Dominika; SZOSTEK, Pawel; DENDEK, Piotr Jan; FEDORYSZAK, Mateusz. CERMINE -- Automatic Extraction of Metadata and References from Scientific Literature. 11th IAPR International Workshop on Document Analysis Systems. 2014.

WILLIAMS, K.; WU, J. WU, Z; GILES, C. L. 2016. Information extraction for scholarly digital libraries. 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL). Newark, NJ, USA. 2016.