INDEXAÇÃO AUTOMÁTICA: DESENVOLVIMENTO DE UMA PLATAFORMA DE INDEXAÇÃO DE ACERVOS TEXTUAIS DIGITAIS PARA REPOSITÓRIOS INSTITUCIONAIS[1]

Gabriela Garibaldi da Cruz [2], Divino Ignácio Ribeiro Junior[3], Gabrieli Fonseca[4]

[1] Vinculado ao projeto "Indexação Automática: Desenvolvimento De Uma Plataforma De Indexação De Acervos Textuais Digitais Para Repositórios Institucionais"
[2]Acadêmica do Curso de Biblioteconomia com Habilitação em Gestão da Informação– FAED – Bolsista PROPPG/UDESC.
[3] Orientador, Departamento de Biblioteconomia – FAED – divino.ribeiro@udesc.br.
[4] Acadêmica do Curso de Biblioteconomia com Habilitação em Gestão da Informação– FAED.

A presente pesquisa se classifica como exploratória e bibliográfica, tendo por objetivo inicial levantar e analisar leituras e documentos sobre tecnologias e softwares usados na realização da indexação automática (em Repositórios Digitais), assim como estudar essas referências e filtrá-las. E a partir dessa fundamentação, realizar testes de colheita de dados e registrar os resultados ou processos dos testes, na forma de relatório. Em seguida, fazer uma avaliação da qualidade de indexação dentro dos repositórios digitais e as tecnologias disponíveis para tal, e então, fazer a apresentação dos resultados finais dos estudos dentro do repositório experimental.

A etapa realizada e relatada neste resumo consistiu no levantamento das publicações relacionadas à plataformas de repositórios digitais, com objetivo de identificar pesquisas semelhantes ao projeto em questão.

O método aplicado no desenvolver da pesquisa foi o estudo comparativo do conteúdo das publicações científicas pesquisadas. Assim sendo, o levantamento bibliográfico desenvolveu-se por diversas bases/bancos de dados, como: a Base de Dados e Ciência da Informação (BRAPCI), o Repositório Digital de Teses e Dissertações da USP, o Repositório Institucional da Universidade Federal do Ceará, o Repositório de Documentos WSL, o Repositório Institucional da UFVJM, o Sistema de Publicação Eletrônica de Teses e Dissertações (TEDE) – nesse sistema estão integradas várias bibliotecas digitais universitárias -, o Repositório Digital da UFRGS (LUME), Scientific Eletronic Library Online (SCIELO), entre outros portais de periódicos e acervos digitais. Além disso, as buscas foram feitas com auxílio dos operadores booleanos (AND, OR, AND NOT) e mecanismos de pesquisas avançadas disponíveis nas próprias bases.

Para uma melhor estruturação dos dados coletas, utilizamos uma planilha do Excel para separar as referências por linha e suas especificações em três colunas (palavras-chave, resumo e link de acesso). Facilitando a filtragem e verificação de cada elemento.

De todos os registros selecionados, o mais antigo foi publicado em 2003 ("O protocolo oai-pmh para interoperabilidade em bibliotecas digitais") e os mais recentes são de 2020 ("Implementação de aspectos de acessibilidade em biblioteca digital desenvolvida com o dspace" e "Sistemas de Indexação automática por atribuição: uma análise comparativa").

Das 40 referências levantadas e registradas na planilha:

- 30 tinham "repositório" ou seu equivalente em inglês "repository" dentre as palavras-chave. Dentro desse segmento, 6 registros formavam o termo composto "repositório digital";
- 20 tinham "informação" ou seu equivalente em inglês "information" nas palavras-chave;
- 20 tinham "Dspace" como uma das palavras-chave;
- 13 tinham "digital" na composição de suas palavras-chave;

- 6 tinham “software” como palavra-chave;
- 2 tinham o termo “ciência da informação” como palavra-chave;
- 2 tinham “tecnologia” ou “tecnologias” como uma das palavras-chave;
- E 2 tinham “indexação” nas palavras-chave.

A maior parte dos documentos estudados estão escritos na língua portuguesa (Brasil), mas encontramos também em inglês e espanhol.

Dentre os autores pesquisados, os mais frequentes na listagem de referências foram: Milton Shintaku, que aparece em 4 registros diferentes, publicados respectivamente nos anos: 2005, 2010, 2018 e 2020; Cibele Araujo Camargo Marques dos Santos e Renato Machado Sobral, presentes em 2 registros distintos, publicados em 2017; Fernando Luiz Vechiato, aparece 2 vezes, ambas publicações do ano de 2018 e por último Renato Fernandes Correa, aparece também 2 vezes, com publicações mais recentes: 2019 e 2020.

Feita essa análise, vale destacar que os autores mais frequentes estão juntos, ou seja, seus nomes aparecem praticamente na mesma frequência pois têm coautoria no mesmo trabalho. Outro ponto importante analisado nos registros, é o fluxo de publicações em cada ano:

- De 2003 a 2011: houve no máximo uma publicação a cada ano, voltada para os temas: Repositórios, Indexação, DSpace, Software ou Tecnologia;
- Em 2012 houve um crescimento de publicações, registramos cinco delas, todas sobre: Repositório, Dspace, Software, Sistema de informações ou Repositório Institucional;
- No ano de 2013, registramos duas publicações, ambas sobre repositório (digital, institucional e informacional);
- Nos anos 2014 e 2015, registramos só duas publicações em cada ano, todas sobre repositórios digitais;
- Dos anos de 2016 e 2017 recuperamos três registros em cada ano, mas os dois anos tiveram como assuntos predominantes: Dspace e Indexação;
- No ano de 2018 ocorreu o maior número de publicações, registramos dez publicações, com assuntos como: metadados, repositórios, repositório institucional, Dspace, acesso aberto e customização de repositórios e de metadados;
- Em 2019 foram nove publicações, voltadas para: indexação automática, ciência ou tecnologia da informação, Dspace, repositório, software (livre), preservação digital ou bibliotecas digitais;
- Por último o ano de 2020, com quatro publicações com foco em: indexação, Dspace, indexação automática, recuperação da informação e software.

Por fim, constatamos uma presença muito pequena na literatura sobre a temática do projeto. Especificamente sobre o desenvolvimento de ferramentas para indexação automática verificamos apenas que software Dspace possui um recurso de indexação ‘fult-text’ para busca em texto completo dos arquivos pdf armazenados pela plataforma, e esse recurso é nativo, ou seja, está disponível no seu código-fonte e não é resultado de alguma aplicação de pesquisa.

Tal constatação corrobora a necessidade da continuidade dos trabalhos de pesquisa no sentido de se alcançar o desenvolvimento de ferramentas que realizem a indexação automática de conteúdo de um repositório digital.

Palavras-chave: Indexação. Recursos e ferramentas tecnológicas. Repositório.