MOUSEION
Revista do Museu e Arquivo Histórico La Salle
ISSN 1981-7207
Editora Unilasalle

MOUSEION

Canoas, n. 43, 2022.

http://dx.doi.org/10.18316/mouseion.vi43.11090

Agregador de Repositórios Científicos em Artes - coleta de dados e interoperabilidade entre repositórios

Luis Felipe Rosa de Oliveira[1]

Érika Demachki[2]

Dalton Lopes Martins[3]

Resumo: O projeto de pesquisa de implementação da Biblioteca Digital Científica de Artes, realizado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) em parceria com a Fundação Nacional de Artes (Funarte), objetiva implementar um acervo digital de publicações científicas na área de artes no Brasil, e é uma iniciativa importante no contexto da recuperação de publicações científicas desta temática, que é parte do contexto da cultura nacional. O objetivo deste trabalho foi descrever as etapas iniciais de implementação deste projeto de pesquisa, como a definição dos critérios de seleção dos acervos digitais, que serão a fonte de informação científica da área de artes, além de descrever quais os resultados encontrados ao recuperar os acervos digitais encontrados através dos critérios de seleção. Como metodologia, é utilizada uma abordagem quantitativa descritiva com definição dos critérios de seleção das fontes de informação, diagnóstico das fontes de informação disponíveis para coleta, e análise dos produtos científicos obtidos das fontes de informação. Os resultados apontam para um contexto de mais de 60% dos acervos digitais disponíveis com condições de interoperabilidade através do protocolo OAI-PMH.

Palavras-chave: Interoperabilidade; Artes; Repositórios Digitais; Repositórios Institucionais.

Aggregator of Scientific Repositories in Arts - data collection and interoperability between repositories

Abstract: The research project to implement the Scientific Digital Library of Arts, carried out by the Brazilian Institute of Information in Science and Technology (IBICT) in partnership with the National Arts Foundation (Funarte), aims

1 Doutor em Ciência da Informação (FCI-UnB), Mestre em Comunicação Social (FIC-UFG), Bacharel em Gestão da Informação (FIC-UFG). Bolsista CAPES. Pesquisador do projeto Tainacan. Realizando pesquisa na área de web semântica e enriquecimento de dados, utilizando a linguagem de programação Python e outras ferramentas de análise para pesquisa em Humanidades Digitais.

2 Mestranda em Ciência da Informação (FCI-UnB). Bacharel em Ciência da Computação (INF-UFG) e especialista em Gestão e Avaliação da Informação (UFG). Atualmente é servidora efetiva assistente em administração na Universidade Federal de Goiás e membro da equipe administradora do Portal de Periódicos da UFG. Atua principalmente nos seguintes temas: altmetria, indicadores de produção científica, comunicação científica, editoração científica, ciência de dados, ciência da informação.

3 Possui graduação em Engenharia Elétrica pela Universidade Estadual de Campinas e mestrado em Engenharia da Computação pela Universidade Estadual de Campinas. Doutor em Ciências da Informação pela ECA-USP. Coordena o projeto de pesquisa Tainacan - software livre para a construção social de repositórios digitais. Professor no curso de Biblioteconomia e do Programa de Pós-graduação em Ciência da Informação PGGCinf da Faculdade de Ciência da Informação (FCI) na Universidade de Brasília (UnB). E-mail: dmartins@gmail.com

to implement a digital collection of scientific publications in the arts area. in Brazil, and it is an important initiative in the context of the recovery of scientific publications on this subject, which is part of the context of national culture. The objective of this work was to describe the initial stages of implementation of this research project, such as the definition of selection criteria for digital collections, which will be the source of scientific information in the arts area, in addition to describing the results found when recovering the collections. found through the selection criteria. As a methodology, a descriptive quantitative approach is used, defining the criteria for selecting information sources, diagnosing the information sources available for collection, and analyzing the scientific products obtained from the information sources. The results point to a context of more than 60% of available digital collections with conditions of interoperability through the OAI-PMH protocol.

Keywords: Interoperability; Arts. Digital Repositories; Institutional Repositories.

Introdução

Repositórios digitais científicos são uma fonte indispensável para se recuperar informações, tanto na disponibilização do acesso à sociedade, quanto no contexto da pesquisa acadêmica. Livros, artigos, teses e dissertações, são algumas produções do contexto científico que se enquadram no escopo de repositórios institucionais de universidades e programas de pós-graduação.

O conceito de repositórios institucionais (RI) que orienta este artigo, compartilha do conceito defendido por (CROW, 2002), indicando que os RIs podem ser compreendidos como acervos digitais compostos da produção intelectual de comunidades do contexto universitário, tendo como objetivos promover base de dados para crítica e renovação científica, bem como a capacidade de avaliar métricas de produção científica no contexto universitário. Seguindo a mesma linha, Araújo (2019, p. 34), informa que os repositórios institucionais estão "fundamentados em difundir o conhecimento científico, com parâmetros estabelecidos pelo movimento de acesso aberto à informação, propondo uma nova visão para as instituições e um novo comportamento para os pesquisadores".

Porém, uma grande variedade de repositórios institucionais dificulta a recuperação de informação, uma vez que seria necessário acessar cada repositório e realizar, individualmente, a busca do material desejado. Para isso, iniciativas de agregação do conteúdo desses repositórios, tais como o *Google Scholar*, ou o *Microsoft Academic*, são amplamente utilizadas. A título de exemplo no contexto nacional, é possível mencionar a Biblioteca Digital Brasileira de Teses e Dissertações – BDTD, desenvolvida pelo Instituto Brasileiro de Informação em Ciência e Tecnologia – IBICT, que agrega as teses e dissertações de repositórios de instituições de ensino brasileiras.

Essas iniciativas são denominadas provedores de serviços, pois não concentram somente a produção de uma instituição, mas sim, agregam o conteúdo publicado em diferentes repositórios e os disponibilizam em um único ponto de acesso e busca (WEITZEL, 2006). Para tanto, é preciso que os repositórios institucionais a serem agregados, estejam organizados a partir de modelo de interoperabilidade comum, compartilhando o mesmo formato de coleta e padrão de metadados.

Um dos modelos de interoperabilidade, mais utilizados atualmente, é o OAI-PHM (O*pen Archives Initiative Protocol for Metadata Harvesting*), cuja premissa é a da possibilidade de comunicação entre sistemas que adotem o mesmo modelo de compartilhamento de dados, tanto no formato do arquivo (XML - *eXtensible Markup Language*), quanto no padrão de metadados (*Dublin Core*) (OLIVEIRA; CARVALHO, 2009).

Percebendo a lacuna de uma fonte de informação científica agregada na área de artes no Brasil, o CEDOC da Funarte propôs a implementação de um serviço de agregação de publicações científicas brasileiras em artes. O CEDOC já conta com um acervo institucional vasto, de cunho arquivístico e bibliográfico, e pretende com esse novo serviço, atender à comunidade de pesquisadores e a sociedade de forma geral, que já utilizam o portal digital da Funarte para pesquisa e exploração.

O projeto ainda está em fase de implementação, mas a primeira etapa já foi concluída e será apresentada neste artigo. Assim, o objetivo é descrever o processo de definição dos critérios de seleção dos acervos digitais que serão a fonte de informação científica da área de artes, além de apresentar quais os resultados descobertos nas primeiras tentativas de recuperar os acervos digitais, que foram encontrados através dos critérios de seleção.

Metodologia

Como metodologia, utilizamos uma abordagem quantitativa descritiva, a qual conta com duas etapas, sendo elas: definição dos critérios de seleção das fontes de informação; e diagnóstico das fontes de informação disponíveis para coleta, seguida pela análise dos produtos científicos obtidos das fontes de informação.

A etapa de definição dos critérios da seleção das fontes de informação foi realizada para definir quais os tipos de publicação científica seriam agregados e qual o escopo. A definição desses critérios foi avaliada e confirmada pelos colaboradores e gestores do CEDOC/Funarte, de acordo com as expectativas do projeto.

Dessa forma, foram definidos como tipos de publicações científicas: livros produzidos no contexto de programas de pós-graduação na área de artes, artigos científicos classificados na área de artes, e teses e dissertações defendidas no âmbito de programas de pós-graduação em artes. Outros materiais podem ser incluídos neste contexto posteriormente, mas, para a definição prévia do material a ser agregado, foram selecionados estes quatro tipos de conteúdo. Além disso, como escopo regional, todas as publicações devem ser nacionais, e como contexto técnico, os repositórios institucionais devem estar sob o modelo de interoperabilidade OAI-PMH, para possibilitar o serviço de agregação.

Para orientar a busca por esses repositórios institucionais selecionados, foram utilizadas informações obtidas da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), a instituição brasileira que orienta e avalia a qualidade dos cursos de pós-graduação no Brasil e que segmenta em áreas os cursos de pós-graduação brasileiros e periódicos nacionais e internacionais.

Assim, através da plataforma Sucupira[4], foi possível acessar os serviços de informação da CAPES. Para identificar os cursos de pós-graduação na área de artes, foi utilizada a base de "Cursos avaliados e Reconhecidos"[5], que indica o nome das universidades e o nome dos cursos; e para os periódicos científicos na área de artes, foi acessada a lista de periódicos obtida do "Qualis"[6].

4 Plataforma Sucupira (CAPES) - https://sucupira.capes.gov.br/sucupira/

5 Cursos Avaliados e Reconhecidos (CAPES) - https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/programa/quantitativos/quantitativoAreaConhecimento.jsf?areaAvaliacao=11

6 Qualis periódicos (CAPES) - https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf

A partir dos dados obtidos da CAPES sobre as fontes de informação na área científica de artes, foram criadas duas planilhas para controle, registro e validação das fontes de informação: uma para as Bibliotecas Digitais de Teses e Dissertações, e outra para os periódicos científicos.

Para o diagnóstico das fontes de informação de Bibliotecas Digitais de Teses e Dissertações (BDTDs), a partir das 36 instituições de ensino com pelo menos um programa de pós-graduação na área de conhecimento de Artes selecionadas, iniciamos a busca manual pelos *links* que apontassem para o *site* de cada repositório institucional que abriga a BDTD da instituição, utilizando a ferramenta de busca *online* do *Google*. Uma vez que a produção científica e acadêmica de pós-graduação das instituições é armazenada em um repositório institucional, o critério de busca foi baseado em pesquisar a expressão "sigla da instituição" ou "nome da instituição" + "repositório institucional". Em relação a análise das fontes retornadas na busca, esta foi realizada a partir de metodologia qualitativa, que buscou reconhecer se o *site* se tratava do repositório institucional desejado através dos seguintes critérios: descrição da página; domínio do site; e seu conteúdo.

Após a identificação dos *links* dos repositórios institucionais, foi feita a identificação do *link* OAI-PMH para coletar os registros de metadados de cada repositório. Essa identificação baseou-se em inserir a expressão "/oai/request?verb=Identify" como sufixo em cada URL de repositório encontrado, sendo essa a requisição padrão do protocolo OAI-PMH para recuperar informações de repositórios com suporte ao mesmo (LAGOZE *et al.*, 2015). Em seguida, fizemos a validação manual dos *links* OAI-PMH, acessando cada um deles para verificar quais davam acesso às informações de metadados dos repositórios.

O próximo passo na coleta de informações das Bibliotecas Digitais de Teses e Dissertações de Artes foi identificar os conjuntos e comunidades referentes aos programas de pós-graduação em Artes. O levantamento dos conjuntos e comunidades foi feito por um *script* na linguagem *Python*, que utilizou como fonte de acesso aos repositórios a planilha utilizada nas etapas anteriores. A avaliação dos conjuntos retornados pelo *script* foi feita de forma manual e qualitativa, visando a identificar quais conjuntos representavam os programas de pós-graduação em Artes. De 9.588 conjuntos, foram selecionados 75. Finalmente, após a seleção dos conjuntos e comunidades de cada BDTD dos programas de pós-graduação em Artes, automatizamos a coleta dos metadados através do *script* em *Python*.

Para o diagnóstico das fontes de informação de periódicos científicos, a partir dos 961 periódicos listados, fizemos uma filtragem e eliminação de duplicatas, chegando a 895 periódicos. Iniciamos a busca manual pelos *links* que apontassem para o *site* de cada periódico utilizando a ferramenta de busca do *Google*, pesquisando variações da expressão "número do ISSN da revista" + "título da revista". Através de uma análise qualitativa, classificamos os periódicos entre nacionais e internacionais, chegando a 635 periódicos nacionais a serem utilizados como fonte de informação para a coleta de metadados.

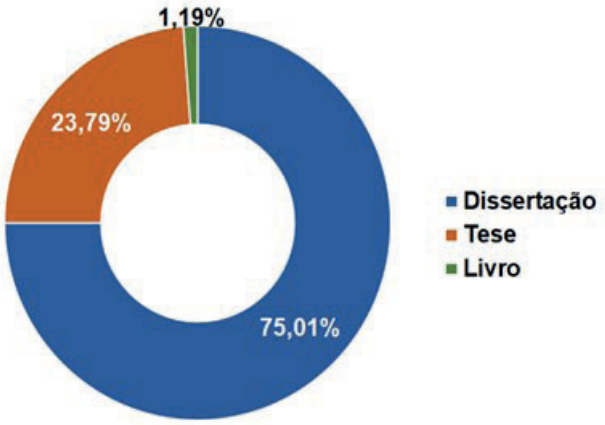
Em seguida, foi feita a identificação e validação do *link* OAI-PMH de cada periódico, seguindo o mesmo processo feito com as Bibliotecas Digitais de Teses e Dissertações. A coleta dos metadados foi automatizada utilizando *script* em *Python*, que utilizou como fonte de informação os *links* registrados e validados na planilha utilizada na etapa anterior. O *script* armazena os metadados coletados em outra planilha, e os erros encontrados na coleta em outra planilha. A seguir, descrevemos os resultados obtidos bem como questões identificadas no processo.

Resultados

Quanto às bibliotecas de teses e dissertações institucionais, de acordo com os critérios da seleção das fontes de informação orientados pela área de Artes da CAPES, são listados 70 cursos de pós-graduação de 35 instituições de ensino superior diferentes, sendo que destas, 30 possuem repositórios institucionais que abrigam Bibliotecas Digitais de Teses e Dissertações, porém, somente 25 com possibilidade de interoperabilidade, ou seja, que possuem o protocolo OAI-PMH. Destes 25 repositórios institucionais, foi possível coletar com sucesso metadados de 17 deles, o que equivale a 48,57% do total de universidades com programas de pós-graduação em Artes listados pela CAPES e 68% do total de repositórios com possibilidade de interoperabilidade (Figura 1).

Observa-se que, por mais que 25 repositórios institucionais possuam OAI-PMH, não foi possível realizar a coleta de oito deles. Não se pode identificar com clareza o motivo da falha ao acessar esses repositórios, porém, fatores como a instabilidade do serviço oferecido pela instituição, configuração com erros e demais problemas de manutenção do sistema podem ser os causadores da inacessibilidade.

Figura 1- Tipos de produção científica coletada das BDTDs



Fonte: Dados da pesquisa.

Foi coletado um total de 6.720 registros dos 17 repositórios institucionais, e os tipos de registros são ilustrados na Figura 1, sendo 5.041 dissertações, 1.599 teses e 80 livros, o que equivale a 75,01%, 23,79% e 1,19% dos registros, respectivamente. O Quadro 1 a seguir, mostra as instituições ranqueadas pelo número de registros coletados. A instituição de ensino superior com maior número de registros coletados foi a UNICAMP (Universidade Estadual de Campinas), com 2019 registros de publicações coletadas, o que equivale a 30,04% do total, seguido pela UFRGS (Universidade Federal do Rio Grande do Sul) e UNESP (Universidade Estadual Paulista) com 863 e 835 publicações coletadas, respectivamente. Juntas, somente essas três universidades reúnem 55,31% das publicações, mais da metade do total coletado.

Quadro 1 – Ranking de instituições de ensino superior por número de publicações coletadas

Posição	**IES**	**Cobertura**	**Número de Publicações**
1	UNICAMP	30,04%	2019
2	UFRGS	12,84%	863
3	UNESP	12,43%	835
4	UNB	9,76%	656
5	UNIRIO	4,33%	291
6	UFRN	4,30%	289
7	UFPA	3,60%	242
8	UFG	3,50%	235
9	UFPB	3,47%	233
10	UERJ	3,32%	223
11	UFPR	2,83%	190
12	UFES	2,53%	170
13	UFPEL	2,14%	144
14	UFSM	1,53%	103
15	UFJF	1,32%	89
16	UFRJ	1,32%	89
17	UFPE	0,73%	49

Nota: IES = Instituições de Ensino Superior.

Fonte: Dados da pesquisa.

Em relação aos periódicos científicos, a partir do total de 894 periódicos únicos que tiveram seu *qualis* avaliado a partir do quadriênio 2013-2016 na área de Artes listados pela CAPES, foram identificados os periódicos nacionais, que totalizaram 635. Destes, somente 440 possuem OAI-PMH permitindo interoperabilidade e foram utilizados como fonte de informação para a coleta de metadados, o que equivale a 49,21% do total de periódicos em Artes e 69,29% dos periódicos nacionais, respectivamente. Dos 440 periódicos selecionados para a coleta dos metadados, 344 tiveram suas publicações coletadas com sucesso, chegando a um total de 164.644 registros de publicações coletadas, e 96 periódicos com OAI-PMH dos quais não foi possível realizar a coleta.

Diferente das Bibliotecas Digitais de Teses e Dissertações, os periódicos científicos possuem uma cobertura de fontes mais bem distribuída, sendo o periódico responsável pelo maior número de publicações coletadas a Revista USP, com 4.974 registros que equivalem a apenas 3,02% do total coletado. Os demais possuem cobertura variando de 2,77% a 0,01%.

Ao analisarmos as publicações coletadas por *qualis* dos periódicos, identificamos que a maior cobertura provém de revistas com *qualis* B1 com 24,84%, o que equivale a 40.904 publicações coletadas, seguido por publicações com *qualis* C, com 15,44% de cobertura das publicações totais. 2,8% das publicações provêm de revistas com *qualis* A1, conforme mostra o Quadro 2.

Quadro 2 - Ranking de instituições de ensino superior por número de publicações coletadas.

Qualis	Número de publicações	Cobertura
A1	4603	2,80%
A2	16501	10,02%
B1	40904	24,84%
B2	18706	11,36%
B3	22341	13,57%
B4	22373	13,59%
B5	13795	8,38%
C	25421	15,44%
TOTAL	**164.644**	**100%**

Fonte: Dados da pesquisa.

Considerações Finais

Mesmo com resultados iniciais, nessa primeira etapa de definição dos critérios de seleção das fontes de informação e diagnóstico preliminar do quantitativo de publicações, foi possível entender que existe viabilidade no desenvolvimento e implementação do serviço de agregação científica na área de artes. Além disso, foi possível elencar alguns desafios que deverão ser enfrentados para a manutenção da plataforma.

Grande parte dos periódicos científicos classificados na área de artes possuem conteúdos também relacionados a outras áreas científicas. Isso se dá porque a classificação da CAPES considera a publicação a partir da formação dos autores, ou seja, não necessariamente em um periódico classificado na área de artes apresentará somente artigos nessa temática. Isso significa que deverá ser pensada uma estratégia para filtragem qualitativa destes artigos, de forma que o serviço de agregação preze pela qualidade temática das publicações agregadas.

Já no caso das bibliotecas de teses e dissertações institucionais, não foram identificados grandes desafios para a agregação futura das publicações. Destaca-se que houve uma grande complexidade no mapeamento das coleções de material específico dos programas de pós-graduação em artes, visto que os repositórios são institucionais e apresentam todo o acervo científico da universidade. Porém, após a definição de quais coleções serão coletadas, elas serão configuradas para agregação permanente no serviço, sendo que, para inclusão ou remoção de novas coleções, o mapeamento deve ser reconfigurado diretamente no serviço.

Vale ressaltar o importante papel de repositórios digitais que implementem um modelo de interoperabilidade, como o OAI-PMH. Independentemente do escopo do repositório, seja arquivístico, museológico e bibliográfico, permitir a agregação de objetos digitais entre os acervos possibilita a construção de uma rede de acervos a partir de um serviço de recuperação da informação, potencializando tanto o acesso ao conteúdo, quanto a visibilidade da sua instituição de origem.

Referências

ARAÚJO, A. A. **S Políticas de Funcionamento em Repositórios Institucionais: perspectivas abrangendo as editoras universitárias**. 2019. Dissertação (Mestrado) – Centro de Educação, Universidade Federal da Paraíba. João Pessoa, p. 142. 2019.

CROW, R. et al. The case for institutional repositories: a SPARC position paper. **ARL Bimonthly Report 223.** 2002.

LAGOZE, C.; SOMPEL, H. V.; NELSON, M.; WARNER, S. **The Open Archives Initiative Protocol for Metadata Harvesting.** 2015. Disponível em: <https://www.openarchives.org/OAI/openarchivesprotocol.html>. Data do acesso: 26 09 2021.

OLIVEIRA, R. R.; CARVALHO, C. L. **Implementação de Interoperabilidade entre Repositórios Digitais por meio do Protocolo OAI-PMH**. Goiás: Universidade Federal de Goiás. Relatório técnico RT-INF_003-09, 2009. Disponível em: <https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_003-09.pdf>Data do acesso: 23 09 2021.

WEITZEL, S. R. O papel dos repositórios institucionais e temáticos na estrutura da produção científica. **Em Questão**, v. 12, n. 1, p. 51-71, 2006.