

Indicadores persistentes via Crossref para teses e dissertações: relato de experiência da UNICAMP

Persistent indicators via Crossref for theses and dissertations: UNICAMP experience report

Alessandra Karyne Neves, Universidade Estadual de Campinas - akaryne@unicamp.br

Gabriel Souza Almeida, Universidade Estadual de Campinas
g216195@dac.unicamp.br

José Maria Lopes, Universidade Estadual de Campinas - jml65@unicamp.br Keite Aparecida Duarte, Universidade Estadual de Campinas - keite@unicamp.br

Eixo 6 - O mundo digital: apropriação e desafios

1 INTRODUÇÃO

Os repositórios institucionais são uma realidade nas instituições de todo o mundo como meio para disseminar e preservar a produção da memória institucional, principalmente nas universidades. No decorrer dos anos, vimos que não bastava reunir os conteúdos em um mesmo lugar, mas sim, oferecer um instrumento de qualidade em que os usuários possam extrair as informações, as métricas e resultados confiáveis. Partindo desse princípio, a Unicamp planejou qual seria a melhor forma de tornar seu repositório eficaz.

Primeiramente, o Sistema de Bibliotecas da Unicamp (SBU) apresentou um projeto à Coordenação Geral da Universidade com várias propostas de ações e, desde 2018, o Repositório Institucional da Produção Científica e Intelectual da Unicamp (RI), integra o Projeto Estratégico da universidade. Tal integração possibilitou investimento financeiro no RI, visando torná-lo "instrumento oficial incumbido de armazenar a produção dos docentes, alunos de graduação e pósgraduação e servidores técnicos e administrativos vinculados à Unicamp" (Gestão estratégica, 2017). Muitas ações foram realizadas para que esse objetivo fosse atingido, uma delas refere-se aos identificadores persistentes. Este trabalho tem como objetivo relatar a experiência da Unicamp para inserir os identificadores persistentes em todos os trabalhos de teses e dissertações produzidos na Unicamp desde 1959.





2 FUNDAMENTAÇÃO TEÓRICA

No ano de 2004, teve início um trabalho retrospectivo de digitalização e depósito de todas as publicações de teses e dissertações na Biblioteca Digital da Unicamp, o qual foi concluído em 2010. Como atualmente todas as teses e dissertações da Unicamp nascem em formato digital, tornou-se rotina na universidade o fluxo no Sistema de Gestão Acadêmico (SIGA), em que o aluno de pós-graduação deve publicar o trabalho no SIGA para ser homologado pela Pró-Reitoria de Pós-Graduação, que posteriormente segue para publicação no RI. Nesse cenário, iniciou-se o estudo sobre os identificadores persistentes para essa fonte de informação. Segundo Gabriel Junior, et al. 2020, p.3

O termo identificador persistente caracteriza um objeto digital acessível pela Web a fim de identificá-lo de forma única e independentemente da localização de sua representação. Este identificador possibilita que o recurso seja corretamente citado ao longo do tempo, permitindo também que várias cópias estejam disponíveis em locais diferentes.

É sabido que há vários tipos de identificadores, porém naquela fase dos estudos, a Unicamp utilizava dois deles nas plataformas do SBU, são eles: o *Digital Object Identifier* (DOI) no Repositório de Dados e nos artigos do Portal de Periódicos Eletrônicos Científicos; e o Handle na plataforma Dspace, usada pelo RI. O desafio foi definir qual identificador persistente usar para as teses e dissertações. Baseado no estudo de Santos (2019), cujo fez uma comparação entre estes dois identificadores e descreveu alguns pontos importantes: conceito, composição, sintaxe, tipo de identificador, finalidade, função, benefício, normas, entre outros.

O DOI provê uma infraestrutura administrativa e workflows para a gerência e identificação persistentes de objetos digitais que possuam atributos de propriedade intelectual no ambiente Internet. Vai além de sérum simples identificador, abrangendo também os desafios de gerenciar o comércio eletrônico, as questões ligadas ao copyright para objetos digitais publicados na Internet, controle de transações e ainda facilitar a comunicação entre os editores e seus clientes. Nesse sentido, o DOI se constitui numa resposta à crescente preocupação dos produtores de conteúdo quanto à proteção de direitos em redes eletrônicas abertas. O Handle é um sistema distribuído de grande alcance, voltado para identificação persistente de objetos digitais e outros recursos na Internet independente de localização - desenvolvido para aplicações no domínio das bibliotecas digitais. Ele foi concebido para assinalar, administrar e resolver identificadores persistentes conhecidos como



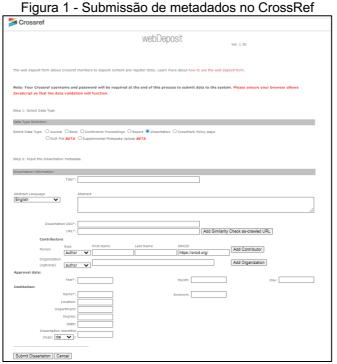


"handles", que podem ser usados como URN –Uniform Resource Names. (SAYÃO, 2007)

3 MÉTODOS DA PESQUISA

Partindo desse estudo o SBU decidiu adquirir o DOI e realizou um convênio com a Associação Brasileira de Editores Científicos (ABEC), pois "Em 2014, a ABEC firmou, junto à CrossRef, um acordo para representação no qual a aquisição do DOI é facilitada. Com isso, viabiliza-se a obtenção desse identificador de forma mais econômica" (Brito, et al, 2016, p. 7). Após essa ação criou-se o prefixo junto ao CrossRef 10.47749 e o sufixo T/UNICAMP.2020.1126282 composto pelo tipo de material, ano e código de título.

O próximo passo foi realizar a submissão de todos os metadados das teses e dissertações no CrossRef, visto que até o momento do estudo a Unicamp estava com 57 mil trabalhos para serem submetidos. A opção para esse tipo de material era acessar o formulário web da CrossRef e indicar o tipo de publicação em que será ofertado o DOI, neste caso as teses/dissertações, e preencher os metadados de forma manual, conforme a imagem abaixo.



Fonte: Web deposit, 2022





Partindo do princípio a quantidade de material que a Unicamp possui, foram realizados vários testes para submeter os metadados, buscando um processo que submetesse em lote, sem um número limitado de títulos. Depois de muito estudo e contato com a CrossRef, concluímos que a única maneira de submeter os metadados das teses e dissertações em lote seria criar um *script*. Então, a Divisão de Tratamento da Informação (DTRI) criou uma metodologia para, através de 4 procedimentos, submeter os metadados em grande quantidade:

- Extrair um relatório de todas as teses e dissertações catalogadas no Acervus da Unicamp;
- Fazer a qualidade dos metadados que foram submetidos no CrossRef;
- Transformar o arquivo .xls em .xml;
- Depósito no CrossRef usando a ferramenta crossref-upload-tool.

4 RESULTADOS

O primeiro procedimento extrair os metadados dos registros de teses e dissertações que foram catalogados no gerenciador Sophia Software. Para isso, foi usado como condição todos os materiais cadastrados, tais como Teses, Teses Digitais, Dissertações e Dissertações Digitais. O relatório recebido do suporte Sophia veio na extensão Excel com 69 258 trabalhos acadêmicos, após verificar-se se todos os trabalhos pertenciam à Unicamp, utilizando o filtro Faculdades\Institutos a qual o registro pertence, o número reduziu para 58 664. O segundo passo foi separar a planilha pelo ano da publicação para trabalhar com a qualidade dos dados. O terceiro passo, foi identificar os metadados que seriam trabalhados.

O segundo procedimento foi realizar a qualidade das informações nas planilhas, pois verificou-se alguns dados ausentes e outros com informações incorretas. Esta fase do trabalho envolveu 15 pessoas na execução dos serviços, sendo 2 funcionários administrativos e 13 alunos bolsistas que trabalham nos projetos da DTRI. Foi desenvolvido um manual de procedimento, contendo a descrição de cada coluna e como deveriam ser analisadas e/ou preenchidas, pois nenhuma coluna poderia ficar em branco. As ferramentas usadas para esse trabalho foram a Base acervus da Unicamp e o Repositório Institucional.

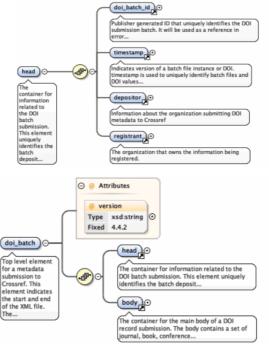




O terceiro procedimento foi criar um tutorial arquivo XLS para XML – DTRI. A demanda deste trabalho surgiu para vincular um código DOI às teses e dissertações da Unicamp e submeter os metadados no CrossRef. O objetivo principal deste pequeno tutorial é transformar um arquivo .xls (um arquivo XLS é aquele que contém uma planilha criada com o Excel) em .xml, que é o arquivo que será adicionado na plataforma do CrossRef.

Para isso, se faz necessário um arquivo denominado schema, que "é um arquivo codificado em linguagem baseada em padrão XML que contém a definição da estrutura de um documento XML, as definições de tipo, tamanho, ocorrência e regras de preenchimento dos elementos que compõe documento XML." (FlexDocs Documentos Eletrônicos). Para a criação desse arquivo schema, pode-se consultar disponíveis site do CrossRef: estruturas as no https://data.crossref.org/schemas/crossref4.4.2.xsd https://data.crossref.org/reports/help/schema_doc/4.4.2/index.html. O intuito como dito, é referente ao DOI, portanto, seguimos a sugestão de estrutura do "doi batch", que pode ser consultado nas seguintes imagens:

Figura 2 - Estrutura geral do doi batch com head e body e exemplo de preenchimento



Fonte: XML Schema documentation generated, 2022





Enfim, toda a estrutura principal e todas as subestruturas podem ser encontradas com todos os detalhes nos sites e links do CrossRef que foram apresentados acima. Lembrando que algumas subestruturas não são obrigatórias e, portanto, cabe ao interessado, dependendo do seu objetivo e dados disponíveis, escolher como será o arquivo schema. Por exemplo, a informação "degree" (grau) não era obrigatória, contudo, foi escolhido enviar junto aos outros dados. O arquivo schema desenvolvido para a Unicamp tem a seguinte estrutura:

Figura 3: Arquivo schema

```
<?xml version="1.0" encoding="utf-8"?>
       <doi batch
                                                       xmlns="http://www.crossref.org/schema/4.3.6"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:jats="http://www.ncbi.nlm.nih.gov/JATS1"
xmlns:ai="http://www.crossref.org/AccessIndicators.xsd"
xsi:schemaLocation="http://www.crossref.org/schema/4.3.6
https://www.crossref.org/schemas/crossref4.3.6.xsd">
                    "http://www.crossref.org/schen
         <doi_batch_id></doi_batch_id>
         <timestamp>
                              </timestamp
         <depositor>
          <depositor_name>
                                     </depositor_name>
          <email address>
                                     r</email_address>
         </depositor>
         <registrant>
        </head>
        <body>
          <dissertation language="pt" publication_type="full_text" reference_distribution_opts="none">
          <person name contributor role="author" sequence="first">
              <given name>
                                                     </given name>
           </person_name>
          <titles>
              <subtitle>
                                     </subtitle>
           </titles>
          <approval date>
         <month>
                              </month>
         <day>
                              </day>
         <year>
                              </year>
    </approval_date>
    <institution>
         <institution name>
                                                  </institution name>
         <institution_acronym>
                                                  </institution_acronym>
         <institution_place>
                                                  </institution_place>
         <institution_department>
                                                  </institution_department>
    </institution>
    <degree>
                              </degree>
    <publisher item>
        <identifier id_type="">
                                                  </identifier>
    </publisher_item>
    <doi_data>
         <doi>
                              </doi>
         <resource>
                                        </resource>
    </doi data>
   </dissertation>
 </body>
</doi_batch>
```

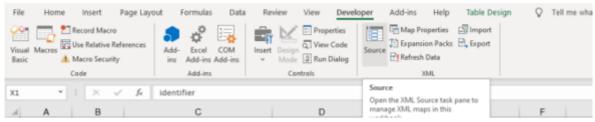
Fonte: Própria, 2022





Novamente, a descrição detalhada do que deve incluir em cada um desses itens também pode ser encontrada nos links apresentados acima. Por exemplo, a subestrutura "institution_place" deve ser preenchida com "a cidade primária da Instituição". Neste caso, foi utilizada "Campinas". Utilizaremos como exemplo uma planilha de Dissertações de 2018. No Excel, o primeiro passo é abrir a Aba "Developer" (Desenvolvedor, na versão em português) e depois clicar em "Source" (Código-fonte, em português).

Figura 4 - Aba "Desenvolvedor"



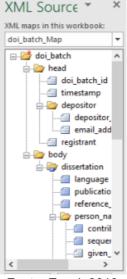
Fonte: Excel, 2019

Abrirá uma janela ao lado direito do monitor (XML Source), clique em "XML Maps" (Mapas XML). Após isso, outra caixa aparecerá e então deve-se clicar em "Add" (Adicionar) e selecionar o arquivo schema que acabou de ser mencionado, depois clique em "ok". Desta forma, poderá visualizar todas as estruturas e subestruturas que estão presentes no arquivo schema, a imagem a seguir demonstra um Mapa XML.





Figura 5 - Janela "XML Source" com o schema adicionado



Fonte: Excel, 2019

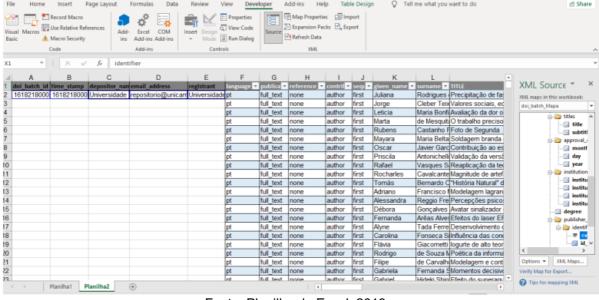
Alguns detalhes importantes antes de prosseguir: lembre-se de formatar a planilha e estruturar as colunas de acordo com o mapa XML, por exemplo, a primeira coluna será referente ao valor do *doi_batch*, depois *timestamp*, depositor e assim por diante. Outra curiosidade, os valores do *doi_batch* e *timestamp* são os mesmos e pode-se gerar o seu valor do *timestamp* no link *epochconverter*, de acordo com o horário e data em que se estiver efetuando este trabalho.

O próximo passo é arrastar cada uma das estruturas do mapa XML da direita para cima das colunas da planilha. Por exemplo, o mapa "doi_batch_id" será arrastado para cima do número 1618218000, o mesmo para "time_stamp" e assim por diante. Neste caso, o mapa "language" será arrastado para cima da coluna "language". Após completar isso, sua planilha se parecerá com a seguinte imagem:





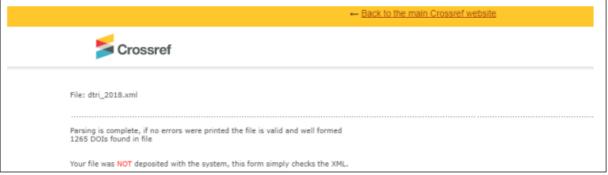
Figura 7: Planilha do Excel após mapear todas as colunas



Fonte: Planilha do Excel, 2019

Ao encerrar todo este mapeamento, pode-se exportar o arquivo XML. Novamente na aba "Developer", ao lado de "Source", encontra-se a opção "Export", clique nela e salve o arquivo com o nome que desejar. Por fim, há a possibilidade de se testar o arquivo XML no site do CrossRef e encontrar eventuais falhas. Através do link https://www.crossref.org/02publishers/parser.html, clique em "Choose File", escolha o arquivo XML e depois clique em "upload", então, em caso de sucesso, poderá visualizar a quantidade de "DOI's" detectados e a seguinte mensagem:

Figura 8: Teste do Arquivo XML no site do CrossRef



Fonte: METADATA quality check, 2022





No caso do exemplo relatado, 1 265 DOIs foram detectados e não houve nenhuma mensagem de erro. Vale destacar que esta ação não submete o arquivo e os dados ao sistema ainda, esta é apenas uma ferramenta de testes.

O quarto procedimento foi efetuar o depósito na CrossRef. Dentre as ferramentas oferecidas pela CrossRef para o depósito de DOI, a que melhor atendeu foi a ferramenta *crossref-upload-tool*. Pois ela permite realizar o upload de vários arquivos XML de uma única vez. Para isto, utilizamos um computador com o sistema operacional Linux – Centos 7.x, que possuía o JAVA 1.8 instalado. Nesse computador, realizamos o download da ferramenta *crossref-upload-tool.jar* e copiamos para pasta /usr/local/lib. Também criamos a pasta /crossref, onde foram salvos os arquivos XML para serem depositados na CrossRef.

Figura 9 - Pasta crossref e os arquivos XML

```
/crossref > ls -la
ot@tortuga
otal 11132
wxr-xr-x
            2 root root
                               310 Sep 10 10:58
                              4096 Dec 10
               root root
                           835266 Aug 15 18:06 dissertacao_1979.xml
237769 Aug 15 18:13 dissertacao_1984.xml
               root root
               root root
                          1066958 Aug 15 18:19 dissertacao_2000.xml
                          1320680 Aug
                                       15 18:25 dissertacao 2001.xml
                          1517267 Aug
                                        15 18:34 dissertacao_2013.xml
                                                 dissertacao_2018.xml
                                   Aug
                            23326 Aug
                                           14:46 dissertacao semdata.xml
                                   Aug
                          1214312 Aug
                                   Aug
               root root
                           948192 Aug
                           907832 Aug
               root root
                                           18:39
             /crossref
```

Fonte: Própria, 2022

Todas as vezes que os arquivos XML são gerados eles são copiados para a pasta do /crossref. E o comando é executado (*java -jar /usr/local/lib/crossref-upload-tool.jar* —user [nome do usuário] [senha] --metadata /crossref)



Figura 10 - Execução da ferramenta da Crossref

Fonte: Própria, 2022

Após a ferramenta finalizar o upload dos arquivos XML para o servidor da CrossRef, esses arquivos entram em uma fila de submissão (figura 11) para serem processados e assim o depósito dos identificadores únicos (DOI) de cada tese e/ou dissertações é realizado.

Figura 11 – Imagem da fila de submissão com os arquivos XML para ser processados



Fonte: Própria, 2022

Ao término do processamento, um relatório é enviado por e-mail informando se o depósito foi bem-sucedido e fornece o motivo das falhas. Após o recebimento do relatório, verifica-se ocorrência de erro ou conflito. No final do relatório é





apresentando um resumo com a quantidade de total DOI, com e sem erros (figura 12)

Fonte: Própria

Caso o relatório apresente falhas ou advertências, o próximo passo é localizar no relatório quais os DOIs que apresentaram conflitos e/ou falhas. Mas, quando o erro não aparece junto com o DOI, é preciso procurar pela falha no arquivo XML que foi enviado à CrossRef, que relata apenas o erro de que a data é inválida.

Após as falhas e conflitos identificados, eles são reportados a área responsável para a realização das correções. O arquivo XML corregido é reenviado para ser depositado novamente. Por fim, é enviado um relatório contendo apenas os DOIs que foram depositados.

5 CONSIDERAÇÕES FINAIS

A equipe envolvida com o Repositório Institucional da Unicamp, conseguiu atender o objetivo proposto nessa fase do planejamento estratégico, seguindo os passos explanado nesse relato de experiencia, é possível submeter no CrossRef vários conjuntos de metadados e obter os DOIs com apenas uma planilha. O mais importante para esse processo ser bem-sucedido é o trabalho de qualidade no momento da catalogação e ter uma boa extração de relatório.





REFERÊNCIAS

BRITO, Ronnie Fagundes de; et.al. **Guia do usuário do digital object identifier**. Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia, 2015. Disponível em: https://www.abecbrasil.org.br/arquivos/Guia_usuario_DOI-online3.pdf

COUTO, G. C. N. do; et.al. Atribuição do DOI (Digital Object Indentifier) em Teses e Dissertações da UFU: relato de experiência. **Ciência Da Informação**, v. 48, n. 3, 2020. Disponível em: http://revista.ibict.br/ciinf/article/view/4854.

DIRETORIA acadêmica da Unicamp. **Fluxo de teses**. Campinas: DAC, 2022. Disponível em: https://www.dac.unicamp.br/portal/processos/proc-negocios-dac/fluxo-tese

FlexDocs Documentos eletrônicos. Validação de schema XML, 2008. Disponível em:

https://flexdocs.net/guiaCTe/funcao.XML.schemas.html

GEPlanes: planejamento estratégico da Unicamp. Campinas: Gestão estratégica, 2017. Disponível em:

https://www.geplanes.cgu.unicamp.br/geplanes/projetos.html?tipoConsulta=ESTR ATEGICO.

GABRIEL JUNIOR, Rene Faustino; et al. **Acesso aberto a dados de pesquisa no Brasil**: identificadores persistentes para dados de pesquisa. 2020. Disponível em: http://hdl.handle.net/20.500.11959/1258

METADATA quality check. CrossRef, 2022. Disponível em: https://www.crossref.org/02publishers/parser.html

SANTOS, G. C. Estudo comparativo das vantagens da interoperabilidade entre o DOI e HANDLE para sistemas de informação. **Boletim Técnico do PPEC**, Campinas, SP, v. 4, 2019. Disponível em:

https://econtents.bc.unicamp.br/boletins/index.php/ppec/article/view/9169.

SAYAO, L. F. Interoperabilidade das bibliotecas digitais: o papel dos sistemas de identificadores persistentes -URN, PURL, DOI, Handle System, CrossRef e OpenURL.Transinformação,Campinas ,v. 19,n. 1,p. 65-82,abr. 2007 Disponível em: https://www.scielo.br/j/tinf/a/NTr5XbPG7LG5pWH876MmWVN

WEB deposit, versão 1.56. CrossRef, 2022. Disponível em: https://apps.crossref.org/webDeposit/

XML Schema documentation generated. CrossRef, 2022 Disponível em: https://data.crossref.org/reports/help/schema doc/4.4.2/index.html

