Interoperabilidade e integração entre DSpace e Archivematica

Laura Marques de Magalhães[a] *e Felipe Augusto Arakaki*[b]

1. Introdução

A Internet modificou a forma como a informação é gerada e transmitida entre os meios de comunicação. Os usuários encontram-se em um novo contexto de busca pela informação e o seu mediador, profissional da informação, deve avaliar suas necessidades de forma que o objetivo do fluxo informacional seja alcançado.

Araújo Junior (2007, p.65) afirma que no processo de busca e recuperação da informação, a motivação do usuário está relacionada a sua necessidade informacional, e a recuperação da informação deve estar próxima à expectativa ou demanda informacional do que foi demandado por esse usuário.

Tendo em vista a interação da informação com o meio digital, a *World Wide Web* – WWW (ROBREDO; BRÄSCHER, 2010), criada por Berners-Lee, inicia em 1989 uma inovação no compartilhamento automatizado de informações. Juntamente com a popularização da Web na década de 90, observou-se a preocupação em relação à localização e recuperação da

a Graduada em Arquivologia pela Universidade de Brasília – UnB. E-mail: lauramarques.m@gmail.com.

b Doutor em Ciência da Informação pela Universidade Estadual Paulista – UNESP. Professor na Universidade de Brasília – UnB. E-mail: fe.arakaki@gmail.com. ORCID: https://orcid.org/0000-0002-3983-2563.

informação e como alternativa para esse problema, foram adotados metadados para descrever e localizar recursos informacionais.

Os metadados são importantes para organizar os recursos eletrônicos, assim como para a necessidade de constante descobertas e informações a serem disponibilizadas na internet e nas intranets. A utilização de padrões de metadados proporciona aos sistemas de informação e de gestão do conhecimento, que recursos e aplicações sejam integrados e compartilhados (ALVES; SOUZA, 2007).

Os metadados são caracterizados como: estrutura padronizada de descrição (ALVES, 2010, p. 46-47); transformação dos dados em conhecimento (IKEMATU, 2001) e a representação, estruturação, gerenciamento, preservação, uso e reúso de informações (ARAKAKI, 2019).

Um dos padrões que é utilizado mundialmente é o padrão de metadados Dublin Core (DC), que se destina a descrever os objetos digitais, sejam eles textos, sons, imagens, vídeos ou sites da web. O DC, conhecido por ser um padrão de metadados para Web, focou em elementos que pudessem representar a informação de uma forma referencial entre diferentes instituições com vistas a recuperação da informação na internet (GRÁCIO, 2002, p. 31). O DC pode ser utilizado com diversas linguagens como *HyperText Markup Language* (HTML), *Extensible Markup Language* (XML), dentre outras.

Nesse âmbito do contexto dos metadados, apresentam-se sistemas de informação, como o DSpace e o Archivematica, que foram construídos, o primeiro software, com o objetivo de captura, armazenamento e disseminação do objeto digital e o segundo software, foi construído com o objetivo de gestão e acessibilidade do objeto digital.

Para Smith et al., (2003) o DSpace é um sistema de código aberto, disponível gratuitamente para instituições, o qual pode ser adaptado ou expandido conforme necessidade. O gerenciamento dos materiais de pesquisa e de publicações, a partir do momento que se utiliza do sistema, promove uma maior visibilidade e acessibilidade à comunidade e usuários.

O Archivematica, é um software livre assim como o DSpace, e tem o objetivo de auxiliar no processo de armazenamento de documentos em formato digital. Além disso, o Archivematica segue normas e padrões para preservação digital, como o padrão ISO-OAIS.

O Padrão ISO-OAIS é definido como um modelo de referência que a partir de um esquema lógico, orienta um sistema de arquivo dedicado à preservação e manutenção do acesso a informações digitais a longo prazo (CCSDS, 2012).

Os dois softwares livres, são grandes aliados para a área da Ciência da Informação, com funções ligadas ao compartilhamento de dados, preservação e garantia de acesso aos usuários.

O objetivo principal desta proposta é de apresentar a possível interoperabilidade entre os softwares DSpace e Archivematica, de forma que, as informações sejam de fácil acessibilidade para seus usuários e preservadas a longo prazo.

2. Procedimentos metodológicos

Os procedimentos metodológicos desta proposta abarcam buscas e leituras de artigos e textos científicos, por meio de pesquisa exploratória e qualitativa da literatura, a fim de compreender e apresentar conceitos e ligações entre os softwares DSpace e Archivematica. Os termos da estratégia de busca foram por meio de palavras-chave, como "Biblioteca Digital" "Preservação Digital" "Dublin Core" "DSpace" "Archivematica". Não obstante, foram utilizadas como base de dados: Scielo, Google Acadêmico, página Web das revistas científicas e de publicações e livros, escolhidas pela confiabilidade e acessibilidade, além de citações vindas de cientistas da informação. O recorte temporal compreende os anos entre 2002 e 2020.

3. Resultados e discussões

Nesta seção será apresentado os conceitos do padrão de metadados Dublin Core, dos sistemas de informação e a proposta de integração entre os dois softwares: Archivematica e DSpace.

3.1 Padrão de Metadados Dublin Core

A origem do padrão DC data da 2º Conferência Internacional *World Wide Web,* de 1994, onde ocorreram discussões sobre semântica e na descoberta de recursos de informação relacionados à Web. Em 1995, com a realização de um evento da OCLC (*Online Computer Library Center*) e o NCSA (*National Center for Supercomputing Applications*), as discussões foram pautadas num conjunto semântico da busca e recuperação de recursos da Web. O resultado dessas discussões culminou nos elementos de metadados intitulado DC cujo nome faz jus à localização do evento em Dublin, Ohio. O "*Core*" corresponde aos princípios do padrão, núcleo de informações, ou seja, núcleo base de informações para localização de documentos na Web. (DUBLIN CORE METADATA INICIATIVE, c2015; HARPER, 2010).

A instituição que administra o DC é a *Dublin Core Metadata Initiative* (DCMI), uma iniciativa que foi concebida em 1995, em Ohio, nos Estados Unidos, que visa o desenvolvimento de padrão de metadados para organização da informação. A Iniciativa de Metadados Dublin Core (*Dublin Core Metadata Initiative* – DCMI) é categorizada como uma organização que promove a adoção e interoperabilidade dos padrões de metadados.

O DC possui dois níveis de descrição. O primeiro, DC Simples, que possui 15 elementos descritivos; e, o outro, DC Qualificado, que possui elementos que qualificam o metadado,

caracterizando e atribuindo mais semântica ao metadado. Seus elementos e qualificadores auxiliam na descrição dos objetos digitais.

Os metadados são fundamentais para promover a interoperabilidade entre sistemas, como o DSpace e o Archivematica. Neste contexto, o DC é uma alternativa recomendada pela *Open Archives Initiative* (OAI) para interoperabilidade entre sistemas de informação e utilizado muito utilizado em Repositórios e Bibliotecas digitais. O DC possui ainda duas (2) ISO que regulamenta e estabelece sua estrutura e metadados como a ISO 15836-2017, "*Information and documentation – The Dublin Core™ metadata element set*", e a ISO 15836-2, "*Information and documentation – The Dublin Core™ metadata element set – Part 2: DCMI Properties and classes"*.

3.2 Sistemas de informação e preservação digital

Os sistemas de informação buscam de alguma forma, atender os requisitos de um ambiente digital ligado aos centros de informação como o arquivo, biblioteca e/ou museu. A crescente inserção da informação no meio digital levanta algumas questões pontuais, especialmente voltadas para a preservação digital.

Torino (2017) lembra que "[…] apenas a preservação digital poderá assegurar que o documento possa ser utilizado em longo prazo […]" e que além dos arquivos, os metadados, sejam eles descritivos ou de direito autoral ou técnicos, e a estrutura do documento devem ser considerados no âmbito da preservação.

Para Sayão (2010), a revolução digital transforma constantemente o modo de criação, comunicação e preservação do conhecimento científico por meio dos acadêmicos, de forma que, a preocupação maior é com a proteção dos conteúdos a longo prazo.

Na área da Biblioteconomia, por exemplo, em muitas instituições o DSpace é utilizado como o software que possibilita a um repositório digital o armazenamento, gerenciamento, preservação e o acesso à produção intelectual. Já a área da Arquivologia, possui normas e padrões internacionais que influenciam diretamente na preocupação com a preservação da informação digital, a obsolescência tecnológica e o armazenamento dos documentos.

As duas áreas de forma síncrona, preocupam-se em preservar o objeto digital, dar acesso aos usuários e aliar sempre que possível aos softwares do mercado para atingir seus objetivos relacionados à informação. Dessa forma, foram abordados dois sistemas de informação: o DSpace e o Archivematica.

O DSpace possui uma arquitetura simples e eficiente, utiliza-se de uma tecnologia voltada ao acesso aberto, com o propósito de atender demandas de Repositórios Institucionais para armazenamento, preservação e disseminação de registros e traduz a estrutura organizacional da institucional mediante seu sistema de comunidades/coleções (ROSA; MEIRELLES; PALACIOS, 2011).

As características mais evidenciadas do DSpace são (VIANA; MÁRDERO ARELLANO; SHINTAKU, 2005): é um software livre; a arquitetura de software é simples e eficiente; uso da tecnologia de ponta; direcionado para o acesso aberto; e implementado com o intuito de servir intencionalmente de repositório institucional.

A partir disso, pode-se inferir que o DSpace é um dos softwares livres mais utilizado em Repositórios e Bibliotecas digitais, levando em conta suas características de acesso aberto e fácil utilização.

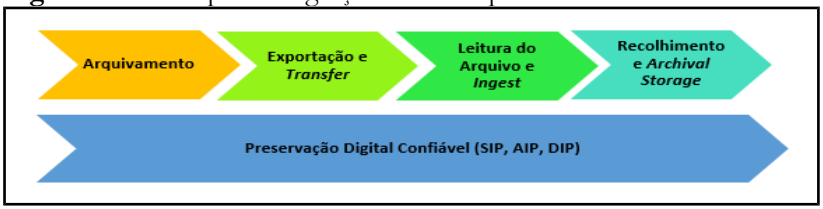
Já o Archivematica configura-se como um software essencial para a preservação e o acesso a longo prazo dos objetos digitais. Em conformidade com o modelo OAIS, este modelo apresenta três tipos de pacotes de informação, com seus devidos metadados específicos para cada pacote. Os pacotes se dividem em:

- SIP: Pacote de informação para submissão – está relacionado à admissão de documentos digitais e seus metadados associados;
- AIP: Pacote de informação para arquivamento – está relacionado ao acondicionamento e armazenamento de documentos digitais e seus metadados associados; e
- DIP: Pacote de informação para disseminação – está relacionado ao acesso aos documentos digitais e seus metadados associados (CONSELHO NACIONAL DE ARQUIVOS, 2015, p.3).

3.3 Proposta de integração entre o DSpace e o Archivematica

Após a análise da literatura sobre a temática de repositórios e bibliotecas digitais, preservação digital, metadados e sistemas de informação de código aberto, pensou-se em uma proposta de integração entre o DSpace e o Archivematica conforme apresentada na Figura 1.

Figura 1 – Fluxo para integração entre DSpace e o Archivematica



Fonte: Autores.

Na primeira seta, apresenta-se o processo de arquivamento. Nela, um item, uma coleção ou até mesmo uma comunidade do DSpace pode ser escolhida para a integração.

Na segunda seta, há a exportação desse item, coleção ou comunidade para o Archivematica. Um arquivo ZIP é gerado para ser feita a leitura pelo software. Em *transfer*, o Archivematica recebe o pacote do DSpace e o transforma em SIP. Lembrando que, um arquivo "mets.xml" é gerado e nele conterá os metadados Dublin Core.

Na terceira seta, após a leitura do arquivo pelo Archivematica, há uma etapa de *ingest*, na qual o pacote SIP é selecionado e após o pacote ser aceito, o SIP se transformará em

um AIP para armazenar o pacote de informação. Para que o diretório "item/coleção/comunidade" seja preservado e acessado futuramente, é obrigatório escolher a opção "*Normalize for preservation and access*" no Archivematica.

Na quarta seta, no processo de recolhimento pelo Archivematica, em *Archival Storage*, o pacote será processado (surgimento do DIP) e disponibilizado para download. Feito o download no computador, têm-se os dois pacotes DIP e AIP. O primeiro pacote de disseminação e o segundo, de arquivamento. No pacote de arquivamento têm-se os arquivos *bagit* que garantem que o arquivo está completo, contendo todos os metadados e demais informações do pacote. No pacote de disseminação, o intuito é o de acesso.

Com a integração entre os dois softwares, DSpace e Archivematica, a preservação digital e o acesso a longo prazo dos objetos digitais estão assegurados, conforme exemplificado na última seta com o SIP, AIP e DIP.

4. Considerações finais

A evolução da Web modificou o processo informacional em que, anteriormente, os dados eram apresentados sem uma relação intrínseca e semântica. Com a inserção dos metadados, utilizados para representar a informação, o processo de busca e recuperação da informação tem sido cada vez mais intuitivo e completo.

Para a área científica, a utilização dos metadados e dos sistemas de informação possibilitam que os princípios da preservação digital a longo prazo sejam cumpridos, e o usuário,

principal consumidor e beneficiário, tenha acesso independentemente de fatores como lugar ou tempo.

A Iniciativa Dublin Core dá origem ao padrão DC, que se estabelece como um conjunto de elementos de metadados (simples e qualificado), utilizado para facilitar a descrição dos recursos da Web. A iniciativa promove a adoção e interoperabilidade do padrão DC pela sua facilidade de uso, simplicidade e aplicabilidade.

A integração demonstra que, por meio da utilização dos dois softwares, DSpace e Archivematica, um item, uma coleção ou até mesmo uma comunidade podem ser preservados e garantido seu acesso futuro pelos usuários.

Do ponto de vista técnico, a integração efetua-se pela exportação dos dados em arquivo ZIP do software DSpace para o software Archivematica, sem a necessidade de outro sistema ou programa para integrar, no qual o próprio arquivo ZIP cumpre o plano de preservação do Archivematica pelo modelo OAIS, e este mesmo arquivo é processado em cada pacote SIP, AIP e DIP para assegurar a confiabilidade e acesso desse arquivo a longo prazo.

Observou-se que para uma interoperabilidade bem-sucedida entre o DSpace e o Archivematica é necessário a adoção de metadados e que a integração pode ser aplicada em qualquer instituição que utilize os dois softwares e queira preservar e dar acesso aos seus objetos digitais.

Algum dos casos que implementaram a integração é de uma tese do autor Francisco Lopes de Aguiar, da Universidade de São Paulo, no domínio da Sociedade Brasileira para o Progresso da Ciência – SBPC. Há a possibilidade de integração com um terceiro software chamado ArchivesSpace. Esta integração tripla, melhora a criação e reutilização de metadados e agiliza a ingestão e o depósito de arquivos digitais (ARCHIVESSPACE, 2021).

Agradecimentos

Agradeço o apoio técnico e computacional do Laboratório de Tecnologias da Tomada de Decisão – LATITUDE, da Universidade de Brasília, que conta com apoio do CNPq – Conselho Nacional de Pesquisa (Outorgas 312180/2019-5 PQ-2, BRICS2017-591 LargEWiN e 465741/2014-2 INCT em Cibersegurança), da CAPES – Coordenação de Aperfeiçoamento do Pessoal de Nível Superior (Outorgas 23038.007604/2014-69 FORTE e 88887.144009/2017-00 PROBRAL), da FAP-DF – Fundação de Amparo à Pesquisa do Distrito Federal (Outorgas 0193.001366/2016 UIoT e 0193.001365/2016 SSDDC), do Ministério da Economia (Outorgas 005/2016 DIPLA e 083/2016 ENAP), da Secretaria de Segurança Institucional da Presidência da República do Brasil (Outorga ABIN 002/2017), do Conselho Administrativo de Defesa Econômica (Outorga CADE 08700.000047/2019-14), da Advocacia-Geral da União (Outorga AGU 697.935/2019), do Ministério das Cidades, (Outorga MC 01/2019), do Ministério da Justiça e Segurança Pública, (Outorga

MJSP 01/2019) e dos Decanatos de Pesquisa e Inovação e de Pós-Graduação da Universidade de Brasília (DPI/DPG/UnB).

Referências

AGUIAR, F. L. de. **Dspace e Archivematica: concepção e criação de um protótipo de repositório digital aplicado no domínio da SBPC**: sob uma perspectiva interdisciplinar entre Arquivística e Organização e Representação do Conhecimento. Tese (Doutorado em Ciência da Informação – Escola de Comunicação e Artes. Universidade de São Paulo), São Paulo, 2018. 667f.

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. 132 f. Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2010. Disponível em: http://hdl.handle.net/11449/103361. Acesso em 24 mai. 2021.

ALVES, M. das D. R.; SOUZA, M. I. F. Estudo de correspondência de elementos metadados: DUBLIN CORE e MARC 21. RDBCI: **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, SP, v. 5, n. 1, p. 20-38, 2007.

ARAKAKI, F. A. Metadados e modelo PROV: perspectivas dos dados de proveniência em contextos digitais. **Informação & Informação**, Londrina, v. 25, n. 3, p. 187-211, jul./set. 2020. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/41869. Acesso em: 24 mai. 2021.

ARAÚJO JUNIOR, R. H. de. **Precisão no processo de busca e recuperação da informação**. Brasília: Thesaurus, 2007.

ARCHIVESSPACE. **Integrations with Archivesspace**. Disponível em: https://archivesspace.org/developers/integrations-with-archivesspace. Acesso em: 24 jun. 2021.

CCSDS. CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM. **Reference Model for an Open Archival Information System (OAIS)**. Blue book (CCSDS 650.0- M-2). Washington, DC, 2012. Disponível em: http://public.ccsds.org/publications/archive/650x0m2.pdf. Acesso em: 24 mai. 2021.

CONSELHO NACIONAL DE ARQUIVOS (BRASIL). CÂMARA TÉCNICA DE DOCUMENTOS ELETRÔNICOS. **Diretrizes para a implementação de repositórios arquivísticos digitais – RDC-Arq**. Rio de Janeiro: Arquivo Nacional, 2015. Disponível em: http://conarq.gov.br/images/publicacoes_textos/diretrizes_rdc_arq.pdf. Acesso em: 25 mai. 2021.

DUBLIN CORE METADATA INITIATIVE. **Dublin Core Metadata Element Set, Version1.1**: Reference Description. July 1999. Disponível em: https://www.dublincore.org/. Acesso em: 25 mai. 2021.

GRÁCIO, J. C. A. **Metadados para a descrição de recursos da Internet: o padrão Dublin Core, aplicações e a questão da interoperabilidade**. Marília: Unesp, 2002. 104f. Dissertação (Mestrado) - Programa de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2002.

IKEMATU, R. S. Gestão de metadados: sua evolução na tecnologia da informação. DataGramaZero, **Revista de Ciência da Informação,** v. 2 n. 6, dez, 2001. Disponível em: http://www.dgz.org.br/dez01/Art_02.htm. Acesso em: 24 mai. 2021.

ROBREDO, J.; BRÄSCHER, M. (Orgs.). **Passeios pelo bosque da informação**: estudos sobre a representação e organização da informação e do conhecimento – eroic. Brasília. DF: IBICT, 2010. 335 p. ISBN: 978-85-7013-072-3. Disponível em: http://www.ibict.br/publicacoes/eroic.pdf. Acesso em: 10 jul. 2021.

ROSA, F.; MEIRELLES, R. F.; PALACIOS, M. Repositório institucional da universidade Federal da Bahia. **Inf. & Soc.:Est.,** João Pessoa, v.21, n.1, p. 129-141, 2011.

SAYÃO, J. Repositórios digitais confiáveis para a preservação de periódicos eletrônicos científicos. **Ponto de Acesso**, Salvador, v. 4, n. 3, p. 68-94, dez. 2010. Disponível em: http://www.portalseer.ufba.br/index.php/revistaici/article/view/4709/3565. Acesso em: 25 mai. 2021.

SMITH, M. et al. Dspace: an open source dynamic digital repository. **D-Lib Magazine**, 9, 1 (2003). Disponível em: http://www.dlib.org/dlib/january03/smith/01smith.html. Acesso em: 24 mai. 2021

TORINO, E. Políticas em repositórios digitais: das diretrizes à implementação. In: VECHIATO, F. et al. (org.). **Repositórios digitais:** teoria e prática. Curitiba: EDUTFPR, 2017. p. 91-114.

VIANA, C. L. de M.; MÁRDERO ARELLANO, M. A.; SHINTAKU, M. **Repositórios institucionais em ciência e tecnologia uma experiência de customização do DSpace**. 2005. Disponível em: http://eprints.rclis.org/archive/00005563/01/viana358.pdf. Acesso em: 25 mai. 2021.