Descrever para Preservar: Metadados como Ferramenta para Gestão de Dados de Pesquisa

Describe to Preserve: Metadata as a Tool to Data Research Management

Lucas de Lima Rocha (1), Luana Farias Sales (2), Luís Fernando Sayão (3)

(1) Instituto de Engenharia Nuclear, Rua Hélio de Almeida, 75 – Ilha do Fundão, Rio de Janeiro. lucasdlrocha@gmail.com (2) lsales@ien.gov.br (3) Comissão Nacional de Engenharia Nuclear, Rua General Severiano, 90 – Botafogo, Rio de Janeiro. lsayao@cnen.gov.br

Resumo

Dados de pesquisa são ativos essenciais para o andamento de pesquisas científicas. Atualmente, grande parte desses dados são produzidos através de softwares computacionais, em uma escala de formatos, tamanhos e complexidade que aumenta cada vez mais. Com isso, torna-se importante identificar formas de preservar e descrever esses dados para reutilizá-los, visando economia de recursos financeiros e de esforços por parte de grupos de pesquisa. O presente estudo apresenta a ideia de metadados para a descrição dos dados de pesquisa como ferramentas importantes para a contextualização dos dados e sua possibilidade de reutilização. Apresenta-se tipos de metadados, sua importância para o atual estágio da ciência, dando atenção especial ao modelo OAIS – que se caracteriza como um dos principais metamodelos para descrição de metadados – e iniciativas de padronização e disponibilização dos metadados, como os projetos Dataverse e Chronopolis, que buscam formas de descrever dados de forma criteriosa, tornando sua recuperação e reutilização mais fácil para os grupos de pesquisa.

Palavras-chave: Metadados; Dados de Pesquisa; Preservação de Dados; Gestão de Dados.

Abstract

Research data are essential assets to the following of a scientific research. Nowadays, a major part of these data is produced through computational softwares, in a growing variety of formats, sizes and complexity. Because of that, it is important to identify ways to preserve and describe this data to reutilize it, thinking on the economy of the financial resources and the growing efforts of the research groups. The following study shows the idea of metadata to describe research data as important tools to the contextualization of the data and the possibility of reuse. The study presents types of metadata and their importance to the present stage of science, giving special attention to the OAIS model–characterized as one of the most important metamodels to describe metadata–and metadata standardization and availability initiatives, as the Dataverse and Chronopolis projects, that seek ways to describe data in a criterions form, making easier for research groups to retrieve and reuse data.

Keywords: Metadata; Research Data; Data Preservation; Data Management.

1 Introdução

Os metadados possuem papel central nas pesquisas científicas: são eles que dão sentido ao conjunto de dados coletados em uma determinada pesquisa, deixando aos pesquisadores o encargo de interpretá-los. Através da análise dos dados coletados e organizados em esquemas de metadados, é possível estabelecer diferentes perspectivas sobre um mesmo conjunto de dados.

Em um estudo de ciências da natureza, por exemplo, a comparação de dados coletados em um mesmo local, mas em diferentes datas, pode trazer à luz questões ligadas ao impacto ambiental que uma região sofre ao longo do tempo. E esse mesmo conjunto de dados pode ser valioso não apenas para pesquisadores de ciências da natureza, mas também para pesquisadores em ciências sociais, por exemplo, que procuram entender de que forma uma dada comunidade se organiza e como utiliza os recursos naturais disponíveis.

As possibilidades são inúmeras, e a capacidade de esses dados serem reutilizados é uma de suas faces mais valiosas. No entanto, a prática de os pesquisadores tornarem os dados de pesquisa disponíveis ainda não é comum ou cultural (CHAO, 2015), já que muitos deles não possuem treinamento formal em gerenciamento de dados, fazendo com que se torne difícil que eles os disponibilizem de maneira a torná-los visíveis (FEDERER, 2013).

Diferente de publicações acadêmicas, os dados não falam por si mesmos e não têm seus conteúdos explícitos, necessitando de ações que vão desde o planejamento no momento de sua criação, passam pela organização em coleções com referências estáveis e padronizadas e culminam em um arquivamento de longo prazo dos dados de valor permanente (SAYÃO; SALES, 2016). Daí a necessidade de estabelecer sentido a esses dados através dos metadados, aqui entendidos de acordo com a definição da National Information Stardard Organization (NISO), como "informações estruturadas que descrevem, localizam ou possibilitam que um recurso informacional seja fácil de recuperar, usar ou gerenciar" (NISO, 2017).

É nesse momento que se colocam as seguintes questões: quais dados devem ser preservados para que outros pesquisadores possam reutilizá-los com a garantia de que o conjunto apresentado seja de fato confiável? De que forma esses dados podem ser melhor acondicionados para que a fragilidade intrínseca ao mundo digital não se torne um problema? Quais os critérios devem ser levados em consideração no momento em que esses dados são disponibilizados para outros pesquisadores?

Este estudo tem por objetivo descrever boas práticas para organizar e disponibilizar os metadados de dados de pesquisa, assegurando confiabilidade, estabilidade e acesso em sua recuperação e reutilização.

2 Procedimentos metodológicos

Este estudo é dividido em dois momentos: no primeiro, o objetivo central é o de apresentar uma discussão teórica sobre os conceitos referentes aos dados e aos metadados de pesquisa, buscando conceitos de organizações que têm como foco central apresentar soluções para a apresentação consistente desses recursos informacionais; e, no segundo momento, ilustrar com exemplos práticos as experiências na utilização de descrições de metadados por diferentes iniciativas, nacionais e internacionais, que têm por objetivo disponibilizar dados de pesquisa para reuso.

Como ponto de partida, este trabalho utilizou pesquisas feitas pelo grupo de trabalho de Gestão do Conhecimento Nuclear no Instituto de Engenharia Nuclear (IEN), que vem efetuando esforços no sentido de solucionar problemas referentes à disponibilização de dados e de metadados de pesquisa das mais diferentes naturezas dentro do domínio nuclear, uma vez que a natureza dos dados gerados pelos pesquisadores do Instituto é heterogênea em formatos, extensões e tamanhos. Através de investigações na literatura, efetuou-se um levantamento bibliográfico de caráter exploratório, seguido de análise e síntese do conteúdo levantado. E, com o intuito de exemplificar de que forma a teoria está sendo atualmente colocada em prática, foram apresentados projetos que têm por objetivo disponibilizar os dados de pesquisa através de uma descrição consistente de seus metadados.

3 Metadados para dados de pesquisa

O termo ‘metadados’ foi utilizado primeiramente no contexto dos sistemas de bancos de dados para descrever e controlar a gestão e o uso dos dados, mas sua ideia remonta outros tempos, tendo suas raízes na catalogação realizada pelas bibliotecas e organizações similares, com função inicial de facilitar a descoberta de informações relevantes (SAYÃO, 2010). A catalogação tradicional é uma forma de atribuição de metadados: o Anglo-American Cataloguing Rules (AACR2) é um código de catalogação que utiliza padrões no contexto bibliográfico para descrever objetos informacionais, e o MARC 21 tem por função descrever, representar, intercambiar e gerenciar os dados bibliográficos e catalográficos, e são essas descrições que geram as informações estruturadas sobre os recursos informacionais – os metadados.

Atualmente, dentro do contexto de informações digitais, a função dos metadados se torna ainda mais ampla. De acordo com Sayão (2010), essas informações podem incluir inúmeras funções, tais como: controle de direitos, intercâmbio, comércio eletrônico, interoperabilidade técnica e semântica, reuso da informação e curadoria digital.

Um dos metadados mais importantes no contexto atual é o de identificação. Através de identificadores persistentes, conhecidos como Digital Object Identifiers (DOI) – que nasceram como uma estrutura genérica para gerir a identificação de conteúdos de redes digitais (DOI, 2017) – é possível garantir a autenticidade e a persistência dos dados no ambiente digital. Um dos maiores provedores de DOI é o DataCite, uma organização sem fins lucrativos que auxilia a comunidade científica a localizar, identificar e citar dados de pesquisa com confiança.

Para alcançar esse objetivo, diversas estratégias são adotadas pela organização, tais como: dar suporte à criação e acondicionamento dos DOI e dos metadados que os acompanham; oferecer serviços que deem suporte à busca especializada dos conteúdos de pesquisas; promover citação de dados e apoiar os esforços da comunidade, a comunicação responsável e a recuperação de materiais.

Os serviços oferecidos pelo DataCite incluem: suporte aos pesquisadores em seus esforços para encontrar, identificar e citar dados de pesquisa e outros objetos de pesquisa; dar suporte aos centros de dados no provimento de DOI aos conjunto de dados, fluxos de trabalho e padrões; dar suporte aos editores de periódicos ao permitir que os artigos científicos estejam ligados aos dados/objetos digitais citados; dar suporte às agências de fomento ao ajudá-los a entender o alcance e o impacto de seus apoios financeiros (DATACITE, 2017).

Além disso, o DataCite também apresenta um esquema geral para a padronização de metadados, contendo dezoito propriedades gerais, que variam entre Obrigatórias (Ob), Recomendadas (R) e Opcionais (Op). A tabela 1 (Tabela 1., em apêndice) descreve cada uma das propriedades gerais – algumas delas se desdobram em subpropriedades, e para ter uma visão geral de todo o conteúdo, consulte (DATACITE, 2015).

Além do DataCite, existem outras iniciativas que procuram estabelecer metadados para dados de pesquisa de forma padronizada e consistente. Entre os principais, podemos citar: Common European

Research Information Format (Cerif), Dublin Core e Data Catalog Vocabulary (DCAT).

De acordo com o Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores, o Cerif é um "padrão recomendado pela União Europeia para registrar informações sobre atividade de pesquisa"; o Dublin Core é um padrão neutro de metadados que pode ser aplicado a várias disciplinas e recursos, permitindo a composição de perfis de aplicação para áreas específicas; e o DCAT é "um vocabulário RDF projetado para facilitar a interoperabilidade entre catálogos de dados publicados na web" (SAYÃO; SALES, 2015, p. 86).

Além dos padrões gerais para metadados, há também os metadados qualificados, que necessitam de uma descrição específica de acordo com o domínio do conhecimento em que estão inseridos. Os metadados e a documentação que os acompanham asseguram que os dados possam ser corretamente interpretados para as suas comunidades e que possam ser utilizados de maneira eficiente ao longo do tempo, uma vez que são os metadados que portam informações de representação dos dados, o que possibilita uma estrutura para as coleções, além da explicitação de sua semântica.

Diferentes áreas do conhecimento possuem diferentes nomenclaturas e necessidades de descrição e, sem uma documentação associada aos dados, eles não passam de cadeias numéricas, variáveis, fragmentos de texto, áudio e vídeo. Sem um software que os decodifique, os objetos digitais são apenas cadeias de bits (BORGMAN, 2007). Para entender os dados, os usuários precisarão de metadados que explicitem como os instrumentos usados foram projetados e construídos; quando, onde e como os dados foram gerados ou coletados; e uma cuidadosa descrição dos estágios de processamento que geraram os produtos derivados dos dados, que são tipicamente usados para análise científica posteriores (GRAY et al., 2002).

Os metadados qualificados existem de acordo com seus domínios de conhecimento. Um mapeamento feito por Sayão e Sales (2015) identificou padrões de metadados qualificados para as biociências, ciências da terra, ciências exatas e ciências sociais & humanidades.

Os padrões identificados para as biociências são: Access to Biological Collection Data (ABCD); Darwin Core; Ecological Metadata Language (EML); e Genome Metadata. Para as ciências da terra, foram mapeados: Agricultural Metadata Element Set (AGMES); Astronomy Visualization Metadata (AVM); e Commom Information Model (CIM). Para as ciências exatas, foram mapeados: Crystallographic Information Framework (CIF); Flexible Image Transport Sytem (FITS); e Stardard for Documentation of Astronomical Catalogues (SDAC). E, nas ciências sociais & humanidades, foram mapeados: Data Documentation Initiative (DDI); Qualitative Data Exchange Format (QuDEx); e Statistical Data and Metadata Exchange (SDMX). Todos os conjuntos de padrões levam em conta as particularidades dos domínios que descrevem, auxiliando no arquivamento e posterior recuperação e acesso por outros pesquisadores (SAYÃO; SALES, 2015).

Os metadados podem ser subdivididos em quatro categorias diferentes: descritivos, estruturais, administrativos e de preservação. A tabela abaixo, elaborada a partir de Sayão (2010), define cada uma dessas categorias:

Tabela 1. *Tipos de metadados*

Metadados descritivos	É a face mais conhecida dos metadados, são eles que descrevem um recurso com o propósito de descoberta e identificação; podem incluir elementos tais como título, autor, resumo, palavras-chave e identificador persistente.
Metadados estruturais	São informações que documentam como os recursos complexos, compostos por vários elementos, devem ser recompostos e ordenados. Por exemplo, como as páginas de um livro, digitalizadas separadamente, são vinculadas entre si e ordenadas para formar um capítulo.
Metadados administrativos	Fornecem informações que apoiam os processos de gestão do ciclo de vida dos recursos informacionais. Incluem, por exemplo, informações sobre como e quando o recurso foi criado e a razão de sua criação. Nessa categoria, estão metadados técnicos que explicitam as especificidades e dependências técnicas do recurso; inclui também os metadados voltados para apoio à gestão dos direitos relacionados ao recurso.
Metadados de preservação	Constituem uma parte essencial das estratégias de preservação digital. A síntese de sua importância pode ser expressa pelo fato deles permitirem que um objeto digital esteja autodocumentado ao longo do tempo e, portanto, posicionado para a preservação de longo prazo e para o acesso contínuo, apesar da sua propriedade, custódia, tecnologia, restrições legais, e mesmo da sua comunidade de usuários estar continuamente mudando.

No contexto da preservação de dados de pesquisa para reutilização por futuros pesquisadores, os metadados se tornam valiosos quando bem descritos e acondicionados, oferecendo sentido a um conjunto de dados que possibilitam sua recuperação rápida e eficaz. De acordo com Sayão (2010), existem cinco categorias fundamentais para a preservação digital, sendo elas: proveniência, autenticidade, atividades de preservação, ambiente técnico e gestão de direitos.

A proveniência afirma que os metadados devem registrar informações do objeto desde sua origem,

traçando sua cadeia de custódia e de propriedade; a autenticidade afirma que devem incluir informações suficientes para validar que o objeto não sofreu alterações, intencionais ou não, que não tenham sido documentadas; as atividades de preservação preconizam a documentação das ações tomadas ao longo do tempo para preservar o objeto digital e as consequências dessas ações tomadas ao longo do tempo para preservar o objeto digital; o ambiente técnico exige a descrição das dependências técnicas necessárias para a apresentação dos objetos digitais, ou seja, os hardwares, sistemas operacionais e softwares necessários para que eles possam ser lidos; e a gestão dos direitos devem registrar tudo o que esteja submetido a questões de propriedade intelectual que possam limitar as ações de preservação, disseminação ou uso dos dados (SAYÃO, 2010).

Através dessas categorias, é possível estabelecer rotinas bem fundamentadas para as etapas de preservação dos dados de pesquisa.

Além de como estabelecer uma rotina para preservação dos dados, também é importante descobrir quais informações precisam ser preservadas para futuras reutilizações. A possibilidade multidisciplinar de utilização torna difícil que um modelo ou uma série de modelos específicos para cada área seja elaborado – visto que cada área possui suas próprias nomenclaturas e especificidades, como pode ser visto através dos metadados qualificados –, mas a criação de um metamodelo com informações necessárias para preservação, independente de um domínio do conhecimento, é um instrumento importante para determinar um arcabouço mínimo de elementos para que uma descrição de metadados seja eficiente. Esse metamodelo se materializou através do Modelo de Referência OAIS (Open Archival Information System), apresentado a seguir.

4 Informações necessárias para a descrição de metadados: o modelo OAIS

O Modelo de Referência OAIS é um esquema conceitual que disciplina e orienta um sistema de arquivo dedicado à preservação e à manutenção do acesso às informações digitais de longo prazo, tendo como propósito mais importante facilitar a compreensão do que é necessário para preservar e acessar essas informações. Se trata de um modelo conceitual, que tem por objetivo aumentar o grau de consciência e compreensão dos conceitos relevantes para o arquivamento dos objetos digitais (SAYÃO, 2010).

Trata-se de um modelo genérico, aplicável a qualquer contexto de preservação digital, que garante que as informações digitais sejam abertas, interoperáveis e com garantias de confiabilidade. Sua infraestrutura abstrata divide-se em dois modelos: o funcional e o de informação.

O modelo funcional compreende o conjunto de atividades que devem ser desempenhadas por um repositório que esteja modelado dentro dos princípios do OAIS, e inclui admissão, armazenamento, gestão de dados, planejamento de preservação, administração e acesso. Já o modelo de informação define as informações, expressas por metadados, necessárias para a preservação de longo prazo e acesso aos objetos armazenados num sistema baseado no OAIS. Constitui uma conceitualização dos objetos de informação incorporados, armazenados e disseminados por um repositório digital orientado para a preservação (SAYÃO, 2010).

O Modelo OAIS também especifica diferentes papéis, como o de produtor – aqueles que fornecem as informações que devem ser preservadas –; consumidor – aqueles que interagem com o sistema para adquirir a informação preservada –; comunidade-alvo – consumidores que devem compreender as informações preservadas –; e administração – aqueles que estabelecem as políticas gerais do repositório.

O modelo preconiza a necessidade de dois componentes para qualquer objeto digital que se deseja preservar: o objeto que será preservado e as informações sobre esse objeto (metadados); as informações se dividem em estruturais – especificações como formato dos dados e descrição dos hardwares e softwares utilizados para a geração dos dados – e semânticas – que acrescentam significado aos dados.

As informações no modelo OAIS devem ser submetidas através de pacotes de informação, que se subdividem em pacotes de submissão (SIPs), formados pelos conteúdos e metadados submetidos por entidades externas – ou seja, produtores – ao repositório; pacotes de armazenamento (AIPs), formatos pelos conteúdos e metadados que são efetivamente armazenados e gerenciados pelo repositório por longo prazo; e pacotes de disseminação (DIPs), que são as informações entregues pelo repositório como resposta às requisições de consumidores.

Os AIPs são subdivididos em quatro tipos de objetos informacionais necessários para a preservação de longo prazo. São eles: informações de conteúdo, aquelas que o repositório tem obrigação de preservar, como informações necessárias à interpretação dos objetos armazenados; informações de descrição de preservação, aquelas que apoiam e documentam a preservação dos objetos; informações de empacotamento, aquelas que agregam todos os componentes de um pacote de informação em uma unidade lógica; e informações descritivas, aquelas que apoiam os consumidores na descoberta e recuperação da informação.

Além dos SIPs, AIPs e DIPs, o modelo OAIS também conta com as Informações Descritivas de Preservação (PDIs), que se relacionam com o estado atual e passado

das informações de conteúdo, para que haja a garantia de que elas estejam identificadas de forma única e que não tenham sofrido qualquer tipo de alterações não previstas. Os PDIs são subdivididos em quatro grupos: informações de referência, informações de contexto, informações de proveniência e informações de fixidade.

As informações de referência identificam e localizam um objeto ao longo do tempo, objetivando manter sua integridade; exemplos são o DOI e o ISBN (International Standard Book Number); as informações de contexto é que dão sentido aos dados, documentando seus relacionamentos com outros conteúdos, além de descrever os softwares e hardwares necessários para sua utilização e seu modo de distribuição; as informações de proveniência garantem a integridade do objeto, com informações sobre sua história e origem, além de registrar as ações de preservações executadas ao longo de sua vida; e as informações de fixidade diz respeito à autenticação do objeto, garantindo que ele não sofreu nenhuma alteração não prevista através de mecanismos de segurança, como a assinatura digital (SAYÃO, 2010).

5 Projetos de disponibilização de dados de pesquisa

Existe uma série de projetos que tem por objetivo tornar os dados de pesquisa melhor representados no mundo digital, garantindo seu acesso, reutilização e aspectos importantes como primariedade, autenticidade e legitimidade. A maior parte deles possui um conjunto de boas práticas para o depósito de dados e, analisando-os, é possível estabelecer rotinas importantes para que os dados não sejam silenciados ou tenham seu significado perdido. Através da atribuição de metadados bem elaborados, é possível que os dados se tornem altamente reutilizáveis, economizando tempo e recursos financeiros para produzi-los novamente.

Com o objetivo de exemplificar o campo prático de descrição dos dados de pesquisa, apresentam-se abaixo projetos que objetivam certificar boas práticas de descrição, acondicionamento e acesso para tais dados, aumentando suas possibilidades de recuperação e reuso. A seguir, são descritos os projetos Dataverse, Chronopolis e DSpace, sendo o último desdobrado em um relato de experiência para a curadoria de dados do Instituto de Engenharia Nuclear (IEN).

5.1 Dataverse

O projeto Dataverse foi criado pelo Instituto de Ciências Sociais Quantitativas (IQSS) da Universidade de Harvard, em um esforço coletivo de colaboradores e contribuidores ao redor do mundo. O projeto tem por objetivo funcionar como um centro de compartilhamento, preservação, citação, exploração e análise de dados de pesquisa. Ele visa facilitar a disponibilização de dados para outros pesquisadores, facilitando sua reutilização de maneira mais fácil. Pesquisadores, autores de dados, editores, distribuidores de dados e instituições afiliadas recebem créditos acadêmicos e visibilidade na web quando os seus dados são utilizados (DATAVERSE, 2017).

Ao depositar os dados de pesquisa no Dataverse, os pesquisadores podem cumprir as requisições estabelecidas pelas agências de fomento sobre planos de gestão de dados de pesquisa. A criação de um plano de gestão de dados é uma boa prática para os projetos de pesquisa, que envolve a captação e a disseminação dos dados, e auxiliam ao assegurar que a coleção de dados possua integridade, qualidade e a proveniência necessários para dar suporte ao projeto. Além disso, assegura que os dados necessários para a replicação externa por outros pesquisadores estejam disponíveis para a comunidade de pesquisa.

5.2 Chronopolis

O Projeto Chronopolis foi fundado pelo Programa Nacional de Preservação e Infraestrutura de Informação Digital da Biblioteca do Congresso Norte-Americano (NDIIPP), com o objetivo de promover o compartilhamento de coleções multidisciplinares para preservação de longo prazo, em uma parceria entre o Centro de Supercomputadores de San Diego (SDSC), as Bibliotecas da Universidade de San Diego e seus parceiros no Centro Nacional para Pesquisas Atmosféricas (NCAR) e o Instituto de Estudos Computacionais Avançados da Universidade de Maryland (UMIACS).

De acordo com Minor et al. (2009), seu modelo de trabalho é desenvolvido através de uma abordagem de fases que contempla: um sistema de produção para gestão e preservação das coleções estáveis, que pode evoluir com o uso e a tecnologia, e crescer com a expansão de coleções individuais ou agregadas; a integração suave de novas tecnologias à medida que são desenvolvidas e testadas, para aumentar a capacidade e funcionalidade sem deixar de prestar o serviço; a administração bem planejada das instalações, que incluem a integração de políticas e procedimentos que dizem respeito à disponibilidade dos dados, integridade, segurança, períodos de retenção, seleção da coleção e padrões de metadados; e a exploração de políticas e de modelos de custos para preservação a longo prazo que assegurem a proteção de coleção de dados críticos para além do tempo de vida dos projetos e esforços que os geraram, para assim providenciar um plano para futura manutenção, curadoria e uso.

Cada parceiro do projeto Chronopolis possui minimamente 50 TB de capacidade de estocagem para coleções digitais, e sua metodologia aplica um mínimo de três cópias geograficamente distribuídas das coleções de dados, ao mesmo tempo em que possibilita auditoria dos processos de curadoria e acesso aos clientes de preservação. A parceria também desenvolve

boas práticas para a comunidade NDIIPP para armazenagem e transmissão de dados entre sistemas de arquivos digitais heterogêneos.

Entre os serviços prestados pelo projeto, estão o de ingestão dos dados, em que há negociações entre a plataforma de acondicionamento de dados e os produtores desses dados, onde são definidos critérios para disponibilização desde os seus aspectos mais elementares, como nomes, tamanhos, extensões e formatos de dados e de metadados, até processos de protocolos de transferência para recuperação da informação; serviço de replicação de dados, que garante aos consumidores munidos de um usuário e senha a capacidade de acesso aos dados, verificadas suas permissões de apenas leitura ou possibilidade de cópia para reutilização, garantindo a possibilidade de que os dados estejam disponíveis em mais de um servidor dos parceiros do projeto; e serviço de auditoria dos dados, instalado em todos os parceiros do projeto de forma independente, que garantem a autenticidade, primariedade e integridade dos dados acondicionados. Cada parceiro possui seu próprio processo de auditoria, garantindo que uma mesma coleção de dados possa ser avaliada de diferentes formas.

A faceta mais interessante do Projeto Chronopolis, no entanto, é o seu modelo de metadados, desenvolvido por um grupo de trabalho que procura estabelecer critérios para que os dados oferecidos aos consumidores estejam padronizados. Entre as especificações do modelo, estão: replicar os ativos em localizações múltiplas e geograficamente dispersas; monitorar regularmente os ativos para identificar deterioração ou corrupção; e substituir os ativos ao provedor de dados quando requisitado. Também são obrigações dos parceiros do projeto Chronopolis: estar em conformidade com os padrões de metadados da comunidade; ser extensível para dar suporte ao desenvolvimento futuro de serviços do projeto e dos padrões de metadados da comunidade; e promover confiança entre os provedores de dados do projeto.

5.3 DSpace

O DSpace nasceu como um software de código aberto desenvolvido para a criação de repositórios digitais, permitindo o gerenciamento da produção científica em qualquer tipo de material digital, com as garantias de acesso de longo prazo. Outra de suas vantagens está na disseminação do conhecimento institucional, uma vez que seus conteúdos ficam disponíveis na web e acessíveis por qualquer computador com acesso à internet.

O software é considerado um exemplo de sucesso na criação de repositórios, possuindo mais de mil repositórios ao redor do mundo, entre institucionais, de áudio e vídeo, de imagens, de museologia e herança cultural, de arquivos governamentais, de ferramentas educacionais e repositórios federativos (DSPACE, 2017). Entre os exemplos de repositórios institucionais que utilizam o DSpace como ferramenta de software, podemos citar: Instituto de Tecnologia de Massachusetts, nos Estados Unidos; Universidade de Illinois, nos Estados Unidos; Universidade de Auckland, na Nova Zelândia; Universidade de Cambridge, no Reino Unido; Universidade de Tsukuba, no Japão; entre outras.

No Brasil, diversas instituições também utilizam o DSpace na construção de seus repositórios, como é o caso da Universidade de São Paulo, Fundação Oswaldo Cruz, Senado Federal, Instituto Antonio Carlos Jobim e o Instituto de Engenharia Nuclear. Sobre a última instituição, algumas considerações serão tecidas na próxima seção.

5.3.1 DSpace e a experiência do IEN: o Repositório Institucional CarpedIEN

O Instituto de Engenharia Nuclear (IEN), em vista da necessidade de reunir a memória institucional, de gerir e preservar as informações geradas no instituto e de disseminar esse conhecimento para além da instituição, adotou o repositório institucional CarpedIEN, através da plataforma DSpace, como uma alternativa para preservar o conhecimento produzido pelos seus pesquisadores.

Entre os objetivos do repositório CarpedIEN, estão: servir de instrumento de apoio para a gestão do conhecimento no IEN; preservar a memória técnico-científica; gerar indicadores de produção acadêmica; servir de apoio a tomada de decisão administrativa; mapear o conhecimento produzido pela instituição; inserir o IEN no fluxo internacional promovido pela interoperabilidade dos repositórios; dar maior visibilidade à produção acadêmica do IEN; aumentar a oferta de serviços de informações mais qualificada para tecnologistas, pesquisadores e alunos; criar um ambiente de interação e troca de ideias entre o corpo de pesquisadores; e organizar e aumentar o nível de disponibilidade e acesso das informações geradas pelo IEN.

No momento de sua criação, condicionou-se que o repositório abrigaria diferentes tipos de materiais, entre eles: publicações técnico-científicas (artigos de periódicos, apresentações em congressos, teses e dissertações), materiais didáticos, documentos de gestão, documentos históricos e dados de pesquisa. No entanto, ao longo do tempo verificou-se a natureza plural dos dados de pesquisa que, no domínio da Engenharia Nuclear, variam de formatos, tamanhos e extensões. Dados de uma determinada área de conhecimento podem ser traduzidos em textos e tabelas, enquanto em outros podem se tratar de arquivos de vídeos, fotografias ou simulações em animação 3D.

Dada a grande pluralidade dos dados de pesquisa produzidos no domínio da Engenharia Nuclear – e a natureza heterogênea de seus metadados –, optou-se pela busca de projetos que dissessem respeito diretamente a disponibilização de dados de pesquisa, que objetivasse uma preservação de longo prazo e que garantisse os princípios de autenticidade, primariedade e autenticidade, levando em contas as recomendações do modelo conceitual proposto pelo OAIS. Atualmente, o grupo de pesquisa trabalha para certificar que seus dados de pesquisa estejam bem descritos e possam ser disponibilizados com todas as garantias citadas.

6 Considerações finais

O encurtamento das distâncias proporcionado pela internet trouxe inúmeras possibilidades para as ciências, e o desenvolvimento de novos softwares e hardwares com capacidades de armazenagem, cálculos e transmissão de dados científicos cada vez maiores faz com que a ciência avance a passos cada vez mais largos. No entanto, de mãos dadas com esse avanço tecnológico, a tradição de pesquisas relegadas a pequenos grupos de cientistas também precisa avançar.

Os dados de pesquisa científica são valiosos não apenas pelos resultados que geram quando visualizadas por um viés, mas sim pela possibilidade de serem vistos e revistos sob diferentes prismas dos diferentes domínios de conhecimento. A possibilidade de um mesmo grupo de dados poder ser utilizado por grupos de pesquisa que investigam assuntos os mais variados possíveis é cada vez mais fácil e econômico; no entanto, medidas precisam ser tomadas para que esses dados sejam possíveis de ser interpretados e que sua integridade, legitimidade, primariedade e autenticidade sejam garantidas.

Dada a importância de reutilização dos dados de pesquisa, sobretudo em ambientes de pesquisa com financiamentos públicos, é necessário que os dados estejam disponibilizados de forma que possam ser facilmente encontrados, totalmente recuperados e que exista a possibilidade de interpretação em futuras pesquisas com o máximo de eficácia possível. Para isso, é necessário que seja estabelecida uma cultura que capacite os pesquisadores a tornar os seus dados encontráveis, ou de utilizar profissionais especialistas nesse exercício, como bibliotecários e profissionais da informação.

Além de treinamento e capacitação, a padronização dos dados também precisa ser levada em consideração, com o objetivo de hospedá-los em diretórios que sejam inteligíveis. A herança das práticas executadas por bibliotecários e arquivistas, nesse sentido, são de grande importância. Através da criação de metadados que deem suporte tanto à disponibilidade dos dados brutos quanto das descrições do que significam, o acesso, recuperação e interpretação dos dados se torna menos dispendioso aos pesquisadores e profissionais que necessitam desses dados para o andamento de suas pesquisas.

Referências

BORGMAN, C. L. Scholarship in the digital age: information, infrastructure and the internet. The MIT Press, 2007.

CHAO, T. C. Mapping methods metadata for research data. Int J of Dig Cur, v. 10, n. 1, p. 82-94, fev. 2015.

DATACITE. DataCite metadata schema for the publication and citation of research data, 2015. Disponível em: <https://schema.datacite.org/meta/kernel-3.1/doc/DataCite-MetadataKernel_v3.1.pdf>. Acesso em: 7 jul. 2017.

_____. Our mission. Disponível em: <https://www.datacite.org/mission.html>. Acesso em: 28 maio 2017.

DATAVERSE. About the project. Disponível em: <http://dataverse.org/about>. Acesso em: 23 maio 2017.

DOI. Handbook. Introduction. Disponível em: <https://www.doi.org/doi_handbook/1_Introduction.html>. Acesso em: 24 maio 2017.

DSPACE. About DSpace. Disponível em: <http://www.dspace.org/introducing>. Acesso em: 28 maio 2017.

FEDERER, L. The librarian as research informationist: a case study. J Med Lib Assoc, v. 101, n. 4, out. 2013.

GRAY, J. et al. Online scientific data curation, publication, and archiving. Redmond: Microsoft Research Corporation, 2002.

MINOR, D. et al. Chronopolis: Preserving our Digital Heritage. In: International Conference On Preservation of Digital Objects, 6. São Francisco, 2009. Anais… São Francisco: California Digital Library, 2009.

NISO. Home page. Disponível em: <http://www.niso.org/home/>. Acesso em: 26 maio 2017.

SAYÃO, L. F. Uma outra face dos metadados: informações para a gestão da preservação digital. Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf., Florianópolis, v. 15, n. 30, p. 1-31, 2010.

SAYÃO, L. F.; SALES, L. F. Algumas considerações sobre os repositórios digitais de dados de pesquisa. Inf. Inf., Londrina, v. 21, n. 2, p. 90-115, maio/ago. 2016.

_____. Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores. Rio de Janeiro: CNEN, 2015.

Apêndice A – Propriedades gerais para padronização de metadados recomendadas pelo DataCite

Tabela 1. *Propriedades gerais para padronização de metadados recomendadas pelo DataCite*

Propriedade	*Definição*	*Obrigação*	*Exemplo*
Identificador	Identificador único que identifica um recurso.	Ob	DOI
Criador	Os pesquisadores principais envolvidos na produção dos dados, ou os autores dos dados, em ordem de prioridade.	Ob	Podem ser nomes institucionais e/ou pessoais. O DataCite permite a citação de até 10 mil nomes
Título	O nome ou título pelo qual o recurso é conhecido.	Ob	Texto livre
Editora	O nome da entidade que guarda, arquiva, publica, distribui, divulga ou produz o recurso.	Ob	World Data Center for Climate (WDCC)
Ano de Publicação	O ano em que o dado se tornou ou se tornará disponível ao público.	Ob	AAAA
Assunto	Assunto, palavras-chave, código de classificação ou uma frase descrevendo o recurso.	R	Texto livre
Contribuidores	Instituição ou pessoas responsáveis por coletarem, gerenciarem, distribuírem ou de alguma forma contribuírem para o desenvolvimento do recurso.	R	Podem ser nomes institucionais e/ou pessoais. O DataCite permite a citação de até 10 mil nomes
Data	Diferentes datas relevantes para o trabalho.	R	AAAA-MM-DD
Idioma	O idioma primário do recurso.	Op	en, de, fr
Tipo de recurso	Uma descrição do recurso.	R	Texto livre
Identificador alternativo	Um ou mais identificadores diferentes do identificador principal aplicado ao recurso que foi registrado. Pode ser uma sequência alfanumérica que é única no domínio utilizado. Pode ser utilizado para identificadores locais.	Op	Texto livre
Identificador relacionado	Identificadores de recursos relacionados. Esses devem ser identificadores globalmente únicos.	Op	Texto livre
Tamanho	Informação não estruturada sobre o tamanho do recurso.	Op	15 páginas, 6 MB
Formato	Informações técnicas sobre o recurso.	Op	PDF, XML, MPG
Versão	O número da versão do recurso.	Op	
Direitos	Qualquer informações sobre direitos autorais para o recurso.	Op	Texto livre
Descrição	Todas as informações adicionais que não caibam em nenhuma das outras categorias. Pode ser utilizado para informações técnicas.	R	Texto livre
Geolocalização	Região espacial ou nome do local em que os dados foram reunidos ou sobre os quais os dados falam.	R	Essa propriedade pode ser repetida para indicar diferentes localizações