RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

Gestão da Preservação digital em repositórios de dados de pesquisa

Henrique Denes Hilgenberg Fernandes
Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, DF, Brasil
denes@ibict.br

Alexandre Faria de Oliveira
Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, DF, Brasil
alexandreoliveira@ibict.br

DOI: https://doi.org/10.26512/rici.v11.n1.2018.8541

Recebido/Recibido/Received: 2017-11-17
Aceitado/Aceptado/Accepted: 2017-12-08

ARTIGOS

Resumo: Este artigo propõe o uso do modelo Open Archival Information System (OAIS) com a adequada curadoria digital, empregando metadados PREMIS adaptados, para assegurar a preservação digital de dados de pesquisa. O artigo traz também uma revisão dos conceitos de dados e repositórios de pesquisa, preservação digital, incluindo o modelo OAIS e curadoria digital, com especial destaque aos metadados de preservação. Propõe adaptações na Informação de Descrição de Preservação, em particular nos descritores de contexto, proveniência e fixidade.

Palavras-chave: Curadoria de dados; Gestão de dados de pesquisa; Metadados; Open Archival Information System; Preservação digital.

Managing Long-term Digital Preservation in Research Data Repositories

Abstract: This article proposes the Open Archival Information System (OAIS) Reference Model with an adequate digital curation, using adapted PREMIS metadata, to ensure the digital preservation of research data. This paper also presents a concepts' review of research data and repositories, digital preservation, including the OAIS Model and digital curation, especially outstanding preservation metadata. It proposes adaptations in the Preservation Description Information (PSI), particularly in the descriptors of context, provenance and fixity.

Keywords: Data curation; Digital preservation; Metadata; Open Archival Information System; Research data management.

Gestión de la preservación digital a largo plazo en repositorios de datos de investigación

Resumen: Este trabajo propone el uso del modelo Open Archival Information System (OAIS) con la adecuada curaduría digital, empleando metadatos PREMIS adaptados, para asegurar la preservación digital de datos de investigación. El artículo trae también una revisión de los conceptos de datos y repositorio de investigación, preservación digital, incluyendo el modelo OAIS y la curaduría digital,

destacando la función de los metadatos de preservación. Propone adaptaciones en la Información de Descripción de Preservación, en particular en los descriptores de contexto, procedencia y fijeza.

Palabras-clave: Curaduría de datos; Gestión de datos de investigación; Metadatos; Open Archival Information System; Preservación digital.

1 Introdução

Na atual sociedade pós-industrial a pesquisa alcançou um *status* nunca antes observado, tornado possível pelas ferramentas de colaboração entre pesquisadores, pela velocidade da disseminação de publicações científicas e de inovação e pela alta tecnologia que dotou os laboratórios experimentais com aparatos de grande precisão. Simulações computacionais, redes de sensores, instrumentos científicos de última geração produzem enormes quantidades de dados precisos e qualificados que documentam ou descrevem fenômenos investigados pela ciência.

Passa a existir uma ciência conduzida por dados, que somada aos mecanismos de colaboração em escala global induzem a um novo modelo de ciência que pode ser chamado de "quarto paradigma científico" ou *e-Science* (SAYÃO; FARIAS SALES, 2014). E a pesquisa moderna está cada vez mais dependente desses dados, que se tornaram imprescindíveis para a correta interpretação e o entendimento dos resultados.

Diante disso, a comunicação científica tem apresentado várias iniciativas para disponibilização dos dados de pesquisa que estão sendo armazenados em repositórios confiáveis e gerenciados sob os princípios da curadoria digital. Esses dados são preservados e mantêm sua capacidade de reuso (FARIAS SALES; SAYÃO, 2012), podendo ainda ser compartilhados.

Tendo em vista que o uso, reuso e compartilhamento dos dados de pesquisa está condicionado à sua preservação, a curadoria digital vem desenvolvendo estratégias, tecnologias e atividades que tornam essa realidade possível. A curadoria é efetivada através de metadados que documentam o conteúdo, dependências técnicas, proveniência, identificação persistente, ações de preservação sofridas e restrições de acesso da informação preservada.

O presente trabalho apresenta como uma efetiva gestão da preservação digital em repositórios de dados de pesquisa pode ser alcançada através da correta implementação do modelo OAIS (Open Archival Information System) (CCSDS, 2002) e da adequada curadoria digital com o uso dos metadados apropriados. Também propõe a adaptação dos metadados da Informação de Descrição de Preservação e outras entidades, especializando-os para tratar situações específicas no contexto de dados de pesquisa.

Esse trabalho está organizado de forma que a seção 2 revisa os conceitos de dados de pesquisa e repositórios de dados de pesquisa. A seção 3 descreve os fundamentos da preservação e o Modelo OAIS. A curadoria digital, introduzindo um modelo proposto pelo DCC (Digital Curation Centre) é apresentada na seção 4, que também explica a evolução dos metadados de preservação a partir daqueles propostos pelo modelo OAIS. A seção 5 apresenta uma proposta para uma gestão da preservação e da curadoria de dados de pesquisa e a seção 6 conclui o artigo.

2 Dados de Pesquisa

Os avanços tecnológicos vêm contribuindo de forma significativa na qualidade das pesquisas científicas, bem como na melhoria da qualidade dos dados gerados durante e após o processo de coleta e análise de informações. Machado (2015) afirma que, ao longo dos anos, vários métodos estão sendo utilizados para armazenar os resultados de pesquisas. Porém, esses métodos não utilizaram uma padronização de recuperação e disponibilização dos dados para a comunidade científica.

Para Costa (2017), as iniciativas para a comunicação de dados de pesquisa, iniciaram-se na década de 1950, mas somente a partir dos anos 2000 a ciência da informação ampliou a discussão no sentido de que os resultados de pesquisas científicas devem ser de amplo acesso a fim de garantir os atributos da ciência moderna.

No panorama informacional recente, observa-se uma visão que caracteriza os dados de pesquisa não apenas alicerces de pesquisas realizadas, mas como subsídios essenciais para outros pesquisadores, instituições acadêmicas e agências de fomento. Esses dados eram descartados ou até mesmo armazenados de forma inapropriada e sem nenhuma gestão, sendo quase sempre guardados com os próprios geradores da informação, perdendo-se no tempo e no espaço.

2.1 Definindo Dados de Pesquisa

Segundo relatório da Organisation for Economic Co-operation and Development (OECD), os dados de pesquisa são registros factuais usados como fonte primária para a pesquisa científica e são comumente aceitos pelos pesquisadores como necessários para validar os resultados dos seus trabalhos (OECD, 2007). Para Machado (2015), dados de pesquisa representam a base a partir da qual os cientistas constroem os seus conhecimentos, sendo necessários não somente para validação, mas também para o desenvolvimento de novos saberes. Segundo Sayao e Farias Sales (2014), dados de pesquisa possuem uma

amplitude bastante complexa e podem se manifestar em uma multiplicidade de formas, variando consideravelmente entre pesquisadores e as diversas áreas do conhecimento.

A National Science Foundation (NSF) em seu documento intitulado *Proposal Award Policies & Procedures Guide* (NATIONAL SCIENCE FOUNDATION, 2014), classifica dados de pesquisa de acordo com duas características principais:

> i. A primeira, baseada na natureza dos dados, podendo ser experimentais, observacionais ou computacionais. Os dados experimentais são oriundos de experimentos em ambientes controlados, portanto são passíveis de reprodução. Já os dados observacionais são obtidos através do registro de observações diretas em um dado momento e algumas vezes considerados como registros históricos. Os dados computacionais são resultantes de uma execução de codificações computacionais e simulações.
>
> ii. A segunda característica é baseada no nível de tratamento que os dados recebem, podendo ser classificadas em duas categorias. A primeira é chamada de dados brutos *(raw data*) ou formato primário. Nessa categoria, os dados são passíveis de diversos processamentos e podem ser reutilizados em outras pesquisas. A segunda incorpora os dados que servem a uma comunidade científica específica baseando-se em coleções de uma área do conhecimento, podendo ser objetos de normalização por padrões já existentes.

É interessante destacar que a gestão e o compartilhamento dos dados de pesquisa exigem um trabalho em parceria entre pesquisadores e prestadores de serviços de curadoria, muitas vezes responsáveis por um repositório de dados de pesquisa. Para tal, é necessário que a equipe esteja devidamente capacitada, garantindo o reuso e o acesso à informação.

Dessa forma, os repositórios de dados de pesquisa são recorrentemente apontados como estratégicos para a efetivação da comunicação dos dados e como instrumento de gestão do conhecimento científico.

2.2 Repositórios de Dados de Pesquisa

Para Costa (2017), repositórios de dados de pesquisa são bases digitais onde são armazenados, disseminados e salvaguardados os dados de pesquisa. Para a autora, uma das principais fontes de informações sobre repositórios de dados de pesquisa está no Registry of Research Data Repositories (re3data.org)[1].

O Registry of Research Data Repositories é um registro global de repositórios de dados de pesquisa que abrange repositórios de diferentes disciplinas. Ele apresenta repositórios para armazenamento permanente e acesso a conjuntos de dados para pesquisadores, órgãos de

[1] https://www.re3data.org/about

financiamento, editores e instituições acadêmicas**.** Assim, o registro das instituições provedoras de dados de pesquisa é muito importante para promover uma cultura de compartilhamento, maior acesso e melhor visibilidade dos dados de pesquisa no mundo.

Nesse sentido, ao analisar os registros de repositórios de dados de pesquisas brasileiros no Registry of Research Data Repositories foi possível encontrar seis participantes. São eles:

> a) Banco de Dados de Exploração e Produção (BDEP), que armazena, organiza e disponibiliza informações geofísicas, geológicas e geoquímicas brasileiras;
> b) WorldClim - Global Climate Data: um conjunto de camadas climáticas globais (grades climáticas) com uma resolução espacial de cerca de um quilômetro quadrado;
> c) Global Collaboration Engine: um ambiente colaborativo *online* que permite a pesquisadores compartilhar, comparar e integrar estudos locais e regionais com dados globais para avaliar a relevância global de seu trabalho;
> d) Instituto Brasileiro de Informação em Ciência e Tecnologia Dataverse Network: fornece um repositório de dados de pesquisa que cuida de boas práticas de preservação e arquivamento de longo prazo, para que os pesquisadores possam compartilhar, manter o controle e obter o reconhecimento de seus dados;
> e) International Ocean Discovery Program: uma colaboração de pesquisa marítima internacional que explora a história e a dinâmica da Terra, usando plataformas de pesquisa oceânica para recuperar os dados registrados em sedimentos e rochas do fundo do mar e para monitorar ambientes subsequentes; e
> f) Repositório de Dados de Levantamentos Biológicos, do Programa de Pesquisa em Biodiversidade (PPBio Data Repository): criado em 2004 com o objetivo de promover estudos de biodiversidade no Brasil, descentralizando a produção científica de centros acadêmicos já desenvolvidos, integrando atividades de pesquisa e disseminando resultados para diversos propósitos, incluindo gerenciamento ambiental e educação.

Dessa forma, fica claro que no Brasil ainda existem poucos repositórios de dados de pesquisas registrados. Destaca-se, portanto, a importância de se utilizar repositórios de dados de pesquisa para armazenar, disseminar e aumentar a visibilidade das pesquisas que são realizadas no país.

3 Preservação Digital

Tendo em vista o reuso, a salvaguarda de dados obtidos a partir de fenômenos singulares e a análise e comparação de dados obtidos em diferentes pesquisas e experimentos, faz-se necessária a preservação cumulativa e perpétua dessas informações, de forma a atender as gerações atuais e futuras de pesquisadores e usuários.

A preservação digital é uma atividade bastante recente e que ainda está em fase de desenvolvimento. Ela é tão necessária que já vem sendo aplicada há pelo menos três décadas, de quando datam os primeiros esforços conjuntos do CCSDS (Consultative Committee for

Space Data Systems) e da ISO (International Organization for Standardization) para a padronização de procedimentos (THOMAZ e SOARES, 2004).

Mesmo com trinta anos de existência, é muito pouco tempo quando se pensa em preservação digital de longo prazo e as expectativas de preservação por longos períodos são, por ora, apenas hipóteses. As atuais técnicas de arquivamento de longo prazo utilizam-se de dispositivos cuja vida útil é de aproximadamente cinco anos (CASTRO *et al*, 2009), obrigando os sistemas de preservação a realizar migrações periódicas de dispositivos e *softwares*. No longo prazo, serão inúmeras migrações.

Sem estudos de caso reais disponíveis até o momento, recorre-se à probabilidade e à simulação computacional. Em Gomes (2010), há uma simulação de um sistema de preservação digital considerando-se o período de cem anos. O autor realizou um experimento com uma ferramenta para preservação digital, utilizando o simulador de redes Peersim[2] e simulando adversidades, através do Processo de Poisson, para avaliar o arquivamento de longo prazo no período acima citado. Num experimento inicial, sem um processo de auditoria de réplicas e considerando uma rede de 2040 repositórios e com 250 objetos em cada um, em pouco mais de 37 anos todos os objetos foram perdidos. Um segundo experimento, dessa vez com utilização de auditoria, avaliou uma rede de 512 repositórios, com 256 objetos, em uma configuração reduzida devido ao custo computacional do algoritmo de auditoria. Nesse segundo caso, em 100 anos foram perdidos 36 objetos, restando preservados 99,888 % dos objetos inicialmente inseridos no sistema.

Além das questões relativas à tecnologia da informação que vêm sendo pesquisadas e testadas nessa fase de incipiência, as bibliotecas e arquivos também vêm pesquisando e testando a migração da enorme quantidade de formatos de dados existentes para formatos padronizados (HEDSTROM, 1997/1998) e, dessa forma, essas especialidades atuam em conjunto na elaboração de estratégias, modelos e ferramentas para a preservação digital.

As instituições interessadas na preservação digital precisam planejar, investir e organizar-se, adotando padrões, manuais de referência, metadados e uma infraestrutura tecnológica para a preservação e formar uma rede distribuída de relacionamentos com outras instituições que também preservem acervos digitais como estratégias estruturantes. Como estratégias operacionais, Lee *et al* (2002) afirmam que a preservação pode ser obtida a partir da emulação, da migração e do encapsulamento. Essas estratégias procuram responder a um

[2] http://peersim.sourceforge.net/

dos problemas da preservação que é o fato que com o passar do tempo os formatos em que os arquivos são gravados deixam de ser suportados pela infraestrutura de *software* e *hardware* existentes, tornando-se obsoletos. A emulação é uma estratégia pela qual um *software* ou *hardware* atual imita um ambiente antigo e obsoleto, oferendo suporte aos arquivos antigos. A migração consiste em auditar periodicamente os formatos de arquivos que estão prestes a se tornar obsoletos e migrá-los para outras formas de suporte, compatíveis com tecnologias atuais. Por fim, o encapsulamento coloca num mesmo pacote de dados o arquivo e todas as informações técnicas necessárias para que ele possa ser aberto no futuro.

Além dessas três estratégias, Thibodeau (2002) menciona a preservação de tecnologias que se tornaram obsoletas, por meio de museus tecnológicos, tornando possível oferecer suporte a arquivos antigos. Thomaz e Soares (2004) sugerem ainda a impressão em papel ou microfilme, o que consideram uma estratégia híbrida e *low tech*, mas que é uma alternativa como "substituto arquivístico" e que pode resultar num produto com uma expectativa de vida de centenas de anos.

3.1 Um Modelo para a Preservação Digital

Apesar da preservação digital ser uma área nova e em constante evolução, a premente necessidade de se preservar acervos digitais antes que eles sejam definitivamente perdidos levou o Consultative Comittee for Space Data Systems (CCSDS), em conjunto com a ISO (International Organization for Standardization) a um esforço conjunto que resultou no Modelo OAIS (Open Archival Information System) (CCSDS, 2002). O objetivo do modelo é suportar e garantir que as características mais relevantes do arquivamento digital como a disponibilidade, durabilidade e confiabilidade dos dados sejam mantidas e consideradas pelos sistemas de preservação digital, por meio de padrões que permitam a manutenção, o compartilhamento e a distribuição do material preservado (CASTRO *et al*, 2009). A disponibilidade está relacionada ao acesso, permissões de acesso e *copyright*. Durabilidade e confiabilidade visam garantir que o material permaneça estático ao longo do tempo, com auditorias frequentes que identifiquem arquivos obsoletos ou corrompidos, assim como mudanças não autorizadas.

Uma implementação do OAIS consiste numa organização de pessoas e sistemas que assumem a responsabilidade de preservar informações e torná-las acessíveis a uma classe de usuários definida como comunidade-alvo (THOMAZ e SOARES, 2004).

O gerenciamento efetivo de todas as formas de preservação digital é viabilizado pela criação e manutenção dos metadados vinculados ao arquivo. Eles documentam processos

técnicos associados à preservação, especificam direitos de acesso, e asseguram a autenticidade do conteúdo digital. Registram também a cadeia de custódia de um objeto digital. Dessa forma, o OAIS inclui um modelo de informação para metadados de preservação, como ilustrado na Figura 1, extraída de Márdero Arellano (2004).

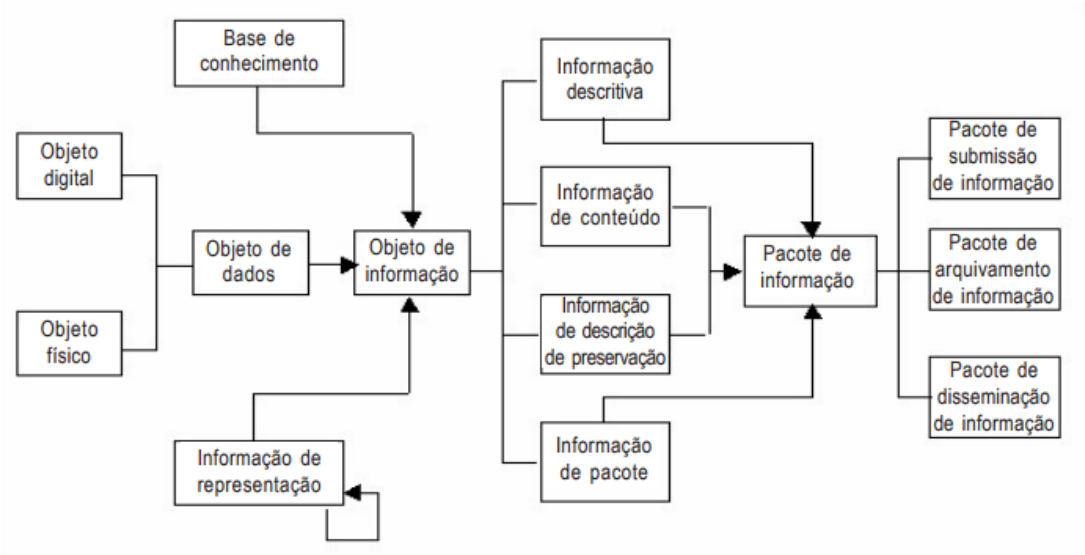

Figura 1 - Modelo de informação OAIS

Fonte: Márdero Arellano (2004)

Um objeto digital, segundo Hunter e Choudhury (2004) e Lorie (2002) é definido como todo e qualquer objeto de informação que possa ser representado através de uma sequência de dígitos binários. São exemplos de objetos digitais os documentos de texto, fotografias, bases de dados.

Pelo modelo, um objeto de dados pode ser tanto um objeto digital ou um objeto físico, como um documento em papel, uma amostra ou evidência física de um fato. Um objeto de dados passa a ser um objeto de informação a partir do momento em que ele acrescenta algum significado de relevância para a comunidade alvo. A comunidade-alvo, por sua vez, possui uma base de conhecimento própria, por exemplo, saberes de uma determinada área para um grupo de pesquisadores nesse campo, que é usada para dar significado ao objeto de informação.

A informação de representação é a informação complementar, necessária para o entendimento do objeto de informação pela comunidade alvo, sempre que a base de conhecimento dessa comunidade não seja suficiente. Observe-se que há uma auto referência na entidade informação de representação no modelo. Isso se deve ao fato de que a

informação de representação pode necessitar de mais informação de representação para possibilitar o entendimento e significância do objeto de informação.

O objeto de informação pode ser de quatro tipos: informação descritiva, informação de conteúdo, informação de descrição de preservação e informação de pacote. A informação de conteúdo é a informação principal e objeto da preservação, que não deve ser interpretado como uma sequência de *bits*, pois uma migração de formato fatalmente altera essa sequência. No modelo, ela já está associada à respectiva informação de representação.

A informação de descrição de preservação contém os detalhes necessários para a adequada preservação da informação de conteúdo associada, podendo ainda descrever referência, contexto, proveniência e fixidade. A informação de referência identifica inequivocamente objetos internos ou externos ao repositório ao longo do tempo, mantendo a sua integridade. A informação de contexto define como o objeto interage com o ambiente digital exterior (*hardware*, *software* e dependências diversas, como *links*, por exemplo). As informações de proveniência documentam a história do objeto armazenado, sua fonte ou origem, cadeia de custódia e ações de preservação sofridas, como migrações de suporte. Informações de fixidade são a forma como o conteúdo é mantido estático ao longo do tempo, com mecanismos para prevenir e detectar mudanças, podendo compreender assinaturas digitais ou *checksuns*, que garantem que o objeto não sofreu nenhuma alteração não documentada.

Ainda sobre os tipos de objeto de informação, a informação de pacote reúne a informação de conteúdo e a informação de descrição de preservação num mesmo pacote de informação e a informação descritiva permite a busca e recuperação da informação de pacote.

O modelo admite três tipos de pacotes de informação: Pacote de Submissão de Informação (PSI), Pacote de Arquivamento de Informação (PAI) e Pacote de Disseminação de Informação (PDI). O Pacote de Submissão de Informação (PSI) é a pacote enviado do produtor da informação para o arquivo. O Pacote de Arquivamento de Informação (PAI) é o pacote efetivamente armazenado no arquivo e o Pacote de Disseminação da Informação (PDI) é o pacote transferido do arquivo para o consumidor, como recuperação da informação, em resposta a uma solicitação.

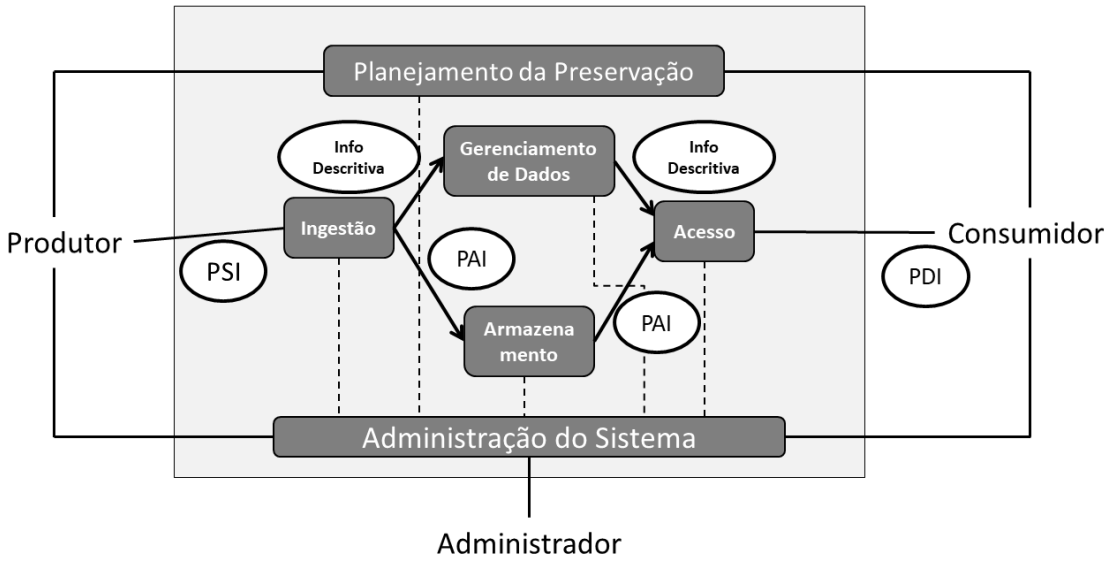

Figura 2 - Modelo funcional OAIS

Fonte: Imagem dos autores

O modelo funcional do OAIS é ilustrado na figura 2, onde podem ser observadas as principais entidades: Ingestão, Armazenamento, Gerenciamento de Dados, Acesso, Administração do Sistema e Planejamento da Preservação.

A Ingestão recebe dos produtores de informação os Pacotes de Submissão de Informação (PSI) e prepara o conteúdo para armazenamento e gerenciamento dentro do arquivo. Verifica a qualidade dos PSI's, gera o Pacote de Armazenamento da Informação (PAI), que é encaminhado ao Armazenamento e a informação descritiva, que é enviada para o Gerenciamento de Dados.

O Armazenamento guarda, recupera e mantém os PAI, gerenciando as transações de armazenamento e recuperação de dados, mantém as mídias atualizadas, verifica erros, recupera-se de falhas e fornece cópias dos PAI ao Acesso. O Gerenciamento de Dados administra a base de dados do arquivo.

A Administração do Sistema negocia acordos de submissão com produtores de informação, realiza auditorias das submissões, gerencia as migrações e configurações de *hardware* e *software*, inovação tecnológica e melhorias das operações em geral. Também é responsável por garantir o enquadramento do sistema às normas e padrões e atender solicitações diversas e outras rotinas de caráter administrativo.

O Planejamento de Preservação monitora o ambiente OAIS e fornece recomendações para que a informação preservada permaneça acessível por longo prazo. Recomenda e planeja migrações e faz adaptações nos metadados para o atendimento a situações específicas.

Por fim, o Acesso permite a realização de busca e recuperação da informação, aplicando os controles de acesso às informações protegidas ou restritas, gerando respostas e entregando-as aos consumidores de informação.

Existem diversas ferramentas de preservação que implementam o OAIS, cada uma delas tendo suas peculiaridades e arquiteturas próprias, mas o núcleo do modelo segue sempre as recomendações do OAIS. Entres essas iniciativas se destacam o Lockss (Lots of Copies Keep Stuff Safe) (REICH, 2009; MANIATIS *et al*, 2005), da Stanford University, que procura garantir a integridade das publicações eletrônicas pela manutenção de cópias em vários nós de uma grande rede, conferindo periodicamente essas cópias para verificar a congruência informacional. Outro exemplo é o Farsite (CASTRO *et al*, 2009), desenvolvido pela Microsoft Research, que visa construir um sistema de arquivos distribuídos de alta disponibilidade. Foi projetado para grandes corporações ou universidades, numa rede de até 105 *peers* e ainda utiliza capacidade de processamento para proteger o conteúdo com criptografia.

4 Curadoria Digital de Dados de Pesquisa

O Digital Curation Centre (DCC) define em seu *website* que a curadoria digital envolve a manutenção, preservação e agregação de valor aos dados de pesquisa digitais durante o seu ciclo de vida e que uma gestão ativa dos dados de pesquisa reduz as ameaças ao seu valor de pesquisa no longo prazo e mitiga o risco de obsolescência digital. Uma ideia ampliada da curadoria digital (ABBOTT, 2008) a define como o conjunto de atividades envolvidas na gestão dos dados, desde planejar as condições da sua geração, boas práticas na digitalização, escolha dos formatos, documentação e na garantia da disponibilidade e adequação para serem descobertos e reusados na atualidade e também no futuro.

O Digital Curation Centre fornece, ainda, um Ciclo de Vida da Curadoria, que é um dos diversos modelos propostos para a curadoria, sendo bastante utilizado. Trata-se de um modelo gráfico, com uma perspectiva de alto nível dos estágios necessários para a curadoria e preservação bem-sucedidas dos dados. O modelo é apresentado na Figura 3, do original em inglês, extraída do *website* do DCC.

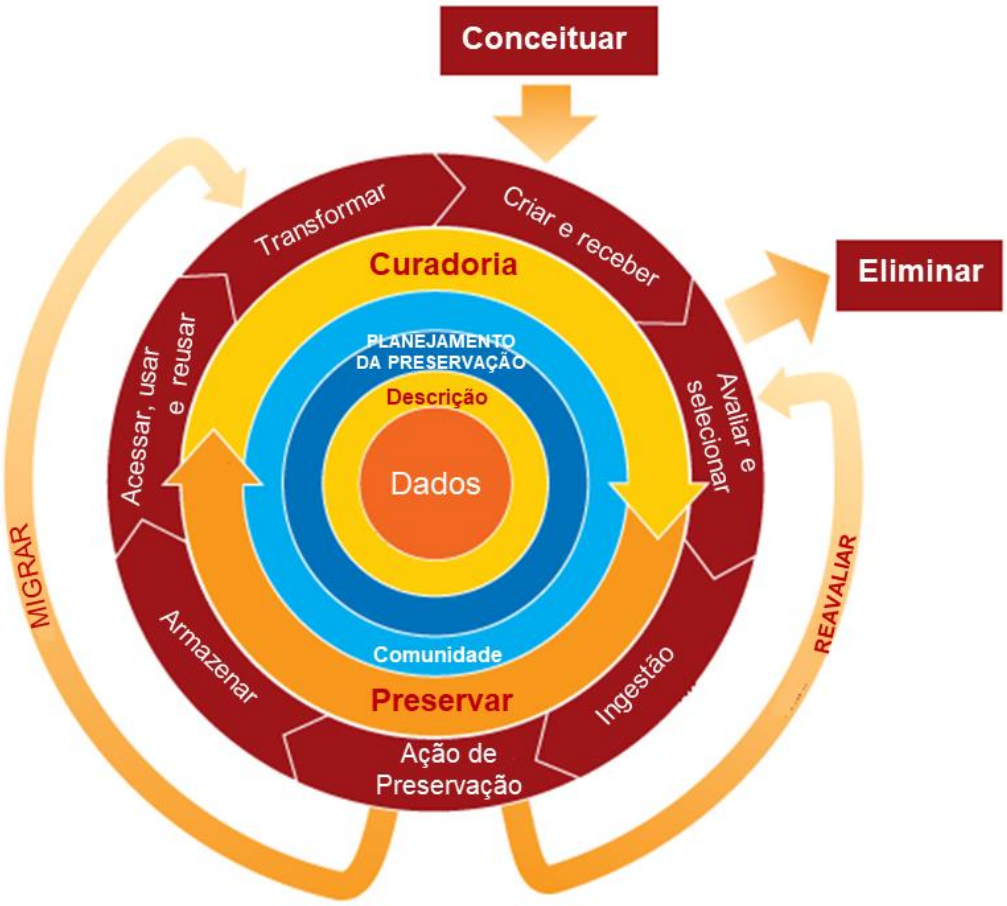

Figura 3 - Ciclo de Vida da Curadoria

Fonte: Digital Curation Centre, 2017.

No modelo proposto, o insumo a ser trabalhado é o dado (*Data*), no centro do modelo, que pode ser objetos digitais (*Digital Objects*) ou bases de dados (*Databases*). Os objetos digitais podem ser objetos simples ou complexos. Os objetos digitais simples são aqueles compostos por um único arquivo, identificador e metadados. Já os objetos complexos são formados pela combinação de outros objetos digitais, formando uma unidade discreta, como uma página *web*. As bases de dados são definidas como coleções estruturadas de registros ou de dados armazenados em sistemas de computadores (FARIAS SALES; SAYÃO, 2012).

As sequências de ações do modelo de ciclo de vida proposto são representadas pelos seguintes estágios, conforme descrito a seguir (DIGITAL CURATION CENTRE, 2017):

> i. *Conceptualise* ou "Conceituar": conceber e planejar a criação do dado, incluindo os métodos de captura e as opções de armazenamento.

ii. *Create or Receive* ou “Criar e receber”: criar o dado incluindo o elenco de metadados necessários à sua gestão e compreensão, ou seja, metadados administrativos, descritivos, estruturais e técnicos. Os metadados de preservação podem ser também incluídos no momento da criação do dado.

iii. *Appraise & Select* ou “Avaliar e selecionar”: avaliar o dado e selecionar o que será objeto dos processos de curadoria e de preservação por longo prazo. Manter-se aderente às práticas, às políticas pertinentes e às exigências legais.

iv. *Ingest* ou “Ingestão”: transferir o dado para um arquivo, repositório, *data center* ou outro custodiante apropriado.

v. *Preservation Action* ou “Ação de preservação”: promover ações para assegurar a preservação de longo prazo e a retenção do dado de natureza oficial. As ações de preservação devem assegurar que o dado permaneça autêntico, confiável e capaz de ser usado enquanto mantém sua integridade. Essas ações de preservação incluem: a limpeza do dado e a sua validação, a inclusão de metadados de preservação, de informação de representação e a garantia de estruturas de dados ou formatos de arquivos aceitáveis.

vi. *Store* ou “Armazenar”: armazenar o dado de forma segura, mantendo a aderência aos padrões relevantes.

vii. *Access, Use & Reuse* ou “Acessar, usar e reusar”: assegurar que o dado pode ser cotidianamente acessado tanto pela sua comunidade alvo quanto pelos demais usuários interessados no reuso do dado. Pode ser realizado na forma de informação disponível publicamente. Um controle de acesso robusto e procedimentos de autenticação podem ser aplicados.

viii. *Transform* ou “Transformar”: criar novos dados a partir do original, por exemplo, pelo processo de migração para diferentes formatos ou pela criação de subconjuntos, realizada por meio de seleção ou formulação de consultas, derivando novos resultados que podem ser publicados.

O modelo estabelece também alguns estágios que são aplicados ocasionalmente:

ix. *Dispose* ou “Eliminar”: eliminar o dado que não foi selecionado para curadoria e preservação de longo prazo de acordo com políticas documentadas, diretrizes e exigências legais.

x. *Reappraise* ou “Reavaliar”:

Retornar ao dado cujos procedimentos de avaliação foram falhos para nova avaliação e possível seleção.

xi. *Migrate* ou “Migrar”: migrar os dados para um formato diferente. Isso pode ser feito no sentido de compatibilizá-lo com o ambiente de armazenamento ou para assegurar a imunidade do dado em relação à obsolescência de *hardware* e de *software*.

4.1 Metadados para a Preservação: Origem e Evolução

Os metadados para a preservação digital são informações que visam à manutenção da fixidade, viabilidade, rendimento, entendimento e autenticidade de materiais digitais no contexto da preservação digital (CAPLAN, 2006). Possuem elementos administrativos,

estruturais e técnicos (acerca dos formatos e codificações dos arquivos). Podem ainda incluir permissões e restrições de uso e de propriedade intelectual. Os outros metadados descritivos do conteúdo não são considerados metadados de preservação.

Segundo Caplan (2006), as bibliotecas e arquivos entendem metadados de preservação sob diferentes perspectivas. Na perspectiva das bibliotecas, a evolução dos metadados de preservação começou com um relatório (WATERS; GARRETT, 1996), onde o objetivo era preservar a integridade da informação, com especial atenção a conteúdo, fixidade, referência, proveniência e contexto, que hoje representam as Informações de Descrição da Preservação, descritas no modelo OAIS.

Os trabalhos de pesquisa e desenvolvimento prosseguiram, culminando com a elaboração do modelo OAIS e sua aprovação como norma ISO 14721 em 2003. A partir do modelo, surgiram diversas iniciativas para padronizar metadados. A National Library of Australia (NLA) foi uma das primeiras instituições a realmente implementar um arquivo digital, com o projeto PANDORA[3] e esteve na vanguarda da preservação de longo prazo dos arquivos australianos (NATIONAL LIBRARY OF AUSTRALIA, 1999). Seguiram-se, então, duas importantes especificações, oriundas dos projetos CEDARS (CURL Examples in Digital Archiving) (RUSSELL *et al*, 2000) e NEDLIB (Networked European Deposit Library) (LUPOVICI; MASANÈS, 2000). O CEDARS foi uma iniciativa britânica do U.K. Consortium of University Research Libraries e objetivou uma especificação detalhada dos metadados de preservação previstos no OAIS. Já o NEDLIB foi a iniciativa europeia, conduzida pela Koninklijke Bibliotheek (Biblioteca Nacional Holandesa) e especificou os metadados de preservação do OAIS com foco específico no problema da obsolescência.

Em 2003, o Online Computer Library Center (OCLC) Digital Archive e o consórcio de bibliotecas RLG (Research Libraries Group) criaram o grupo de trabalho PREMIS (Preservation Metadata: Implementation Strategies), composto por representantes de instituições que pesquisavam e utilizavam a preservação digital. O PREMIS passou a atuar com um notável rigor na aplicação dos metadados de preservação, desenvolvendo um conjunto de elementos de metadados altamente refinados, que potencialmente serviam de fundamento para possíveis implementações, materializado em (ONLINE COMPUTER LIBRARY CENTER/ RESEARCH LIBRARIES GROUP 2005).

[3] http://pandora.nla.gov.au/

Entre suas características, o PREMIS possui um dicionário de dados com entidades de dados diferentes daquelas do modelo OAIS, mas existe uma correspondência entre elas, o que quer dizer que as entidades de um modelo podem ser mapeadas para as respectivas entidades do outro.

Apesar de muitos arquivistas terem participado do grupo de trabalho inicial do PREMIS, e ainda participarem do PREMIS Implementers' Group (PIG), a perspectiva da arquivística para metadados de preservação tem seguido um caminho diverso (CAPLAN, 2017), que teve origem no projeto de pesquisa "Functional Requirements for Evidence in Recordkeeping", da University of Pittsburgh School of Information Science, que culminou com uma especificação de metadados (UNIVERSITY OF PITTSBURGH SCHOOL OF INFORMATION SCIENCE, 1996).

Outra iniciativa da área de arquivística foi a da University of British Columbia (UBC), no Canadá, que desenvolveu um projeto que resultou num conjunto de oito modelos de metadados, dando origem ao InterPARES (International Research on Permanent Authentic Records in Electronic Systems)[4].

Essas duas iniciativas da arquivística fundaram escolas que persistem e coexistem até hoje.

5 Gestão da Preservação e da Curadoria de Dados de Pesquisa

Sayão e Farias Sales (2014) descrevem a importância dos dados de pesquisa para a ciência aberta, o impacto desse novo ambiente orientado por dados na comunicação científica, infraestruturas existentes para o tratamento dessa informação e propõem um modelo de curadoria digital de dados de pesquisa para o país, onde são consideradas as instâncias de Sustentabilidade Econômica, Aspectos Sociais, Éticos e Legais, Pesquisa, Política, Recursos Humanos, Desenvolvimento de Coleções de Dados, Serviços, Tecnologias e Padrões e Configuração Organizacional.

Em outro trabalho (FARIAS SALES; SAYÃO, 2012), os mesmos autores destacam o papel da curadoria digital e a importância dos metadados para a preservação digital de dados de pesquisa. E são exatamente a curadoria digital e os metadados de preservação que aqui se ramificam, especializando-se no caso particular dos dados de pesquisa.

[4] http://interpares.org/

Considerando que os dados de pesquisa não são necessariamente publicados, que possuem uma grande variedade de formatos (vídeos, sons, arquivos gerados por *softwares* especializados) e que podem possuir restrições de acesso éticas, legais ou referentes a *copyright,* podendo tratar-se também de informações sensíveis em outros âmbitos, o modelo OAIS, que é recomendado para a preservação de quaisquer documentos em formato digital, é totalmente compatível com esses dados. No que diz respeito aos processos do OAIS e ao seu modelo de informação, os dados de pesquisa em muito se assemelham aos outros tipos de registros, com algumas características que irão requerer um tratamento especializado nas ações de curadoria digital, podendo exigir metadados de preservação adaptados para esse fim.

Devido à maior variedade de formatos de arquivo que os dados de pesquisa apresentam, mais informação de representação pode ser necessária. No caso das publicações ampliadas (*enhanced publication*), em que os dados de pesquisa são ligados às pesquisas já publicadas, - por exemplo, Object Reuse and Exchange (ORE) (OAI, 2014) – pode ser necessário um nível maior de detalhamento da informação descritiva para busca e recuperação.

Acerca da Informação de Descrição de Preservação, os descritores de referência podem requerer metadados mais especializados, como é o caso da informação de representação. Por poderem se apresentar em formatos muito específicos como arquivos de um determinado *software* de engenharia ou de outros aplicativos científicos, que são muito suscetíveis a atualizações de formatos e descontinuidade de suporte, os dados de pesquisa irão requerer metadados mais detalhados para os descritores de contexto, proveniência e fixidade.

Havendo a possibilidade de migração de suporte em caso de obsolescência ou descontinuidade dos formatos de arquivo, possivelmente haverá muitas migrações no longo prazo e os descritores de proveniência deverão registrar adequadamente essas alterações, assim como os descritores de fixidade deverão garantir que não houveram migrações desautorizadas e não registradas.

Por fim, na hipótese de não ser possível a migração para um suporte compatível, sugere-se o encapsulamento das informações de acesso nos descritores de contexto, com o maior nível de detalhe possível acerca dos formatos de arquivos, podendo ainda incluir nesses descritores o próprio *software* capaz de processar a referida codificação.

Todas as adaptações propostas são suportadas pelo modelo OAIS e pelos metadados de preservação PREMIS, não sendo necessárias alterações nos seus padrões.

6 Conclusão

O presente trabalho trouxe uma revisão dos conceitos de dados de pesquisa e seu contexto, assim como uma resenha acerca da preservação e curadoria digitais, trazendo os modelos OAIS para preservação e o DCC para curadoria, com especial destaque aos metadados. Propõe, ainda, que os metadados sejam especializados para lidar com situações especiais de dados de pesquisa, em particular os descritores de contexto, proveniência e fixidade, da Informação de Descrição de Preservação. Recomenda um maior detalhamento das migrações nos descritores de proveniência e, para os casos em que a migração não seja possível, o encapsulamento das informações de formato de arquivos e até mesmo do próprio *software* que suporta a codificação nos descritores de contexto.

Dessa forma, é possível garantir a preservação digital de longo prazo de dados de pesquisa, a partir da correta implementação do modelo OAIS, com a adequada curadoria digital e fazendo uso de metadados apropriados e adaptados para este novo cenário.

A título de trabalhos futuros, propõe-se uma especificação completa de descritores de contexto e proveniência da Informação de Descrição de Preservação adaptados para o panorama em voga e um estudo de caso de um ambiente onde se preserve dados de pesquisa.

Referências

ABBOT, D. **What is digital curation?** DCC Briefing Papers: Introduction to Curation, Digital Curation Centre, 2008. Disponível em: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation Acesso em: 1 nov. 2017.

CAPLAN, P. Instalment on "Preservation Metadata". In: DIGITAL CURATION CENTRE. **Curation Reference Manual.** Edinburgh: DCC, 2017. Disponível em: http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/preservation-metadata/preservation-metadata.pdf Acesso em: 5 nov. 2017.

CASTRO, C. Y. H. de; SUNYE, M. S.; BONA, L. C. E. de; CASTILHO, M. A. Repositórios Institucionais Confiáveis: Repositório institucional como ferramenta para a preservação Digital. In: SAYÃO, L.; TOUTAIN, L. B.; ROSA, F. G.; MARCONDES, C. H. (Org.). **Implantação e gestão de repositórios institucionais: políticas, memória, livre acesso e preservação**. Salvador: EDUFBA, 2009.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (CCSDS). **Reference Model for an Open Archival Information System (OAIS). Magenta Book (CCSDS 650.0-M-2)**. Washington, DC: CCSDS, 2012. Disponível em: https://public.ccsds.org/Pubs/650x0m2.pdf Acesso em: 03 nov. 2017.

COSTA, M. P. **Fatores que Influenciam a Comunicação de Dados de Pesquisa sobre o Vírus da Zika na Perspectiva de Pesquisadores**. 2017. Tese (doutorado) - Universidade de Brasília, Brasília. Disponível em: http://www.acervodigital.ufpr.br/bitstream/handle/1884/24882/dissertacao.pdf

DIGITAL CURATION CENTRE. **DCC Curation Lifecycle Model**. Edinburgh: DCC, 2017. Disponível em: http://www.dcc.ac.uk/resources/curation-lifecycle-model Acesso em: 02 nov. 2017.

FARIAS SALES, L.; SAYÃO, L. F. O Impacto da curadoria digital dos dados de pesquisa na comunicação científica. **Encontros Bibli: Revista eletrônica de biblioteconomia e ciência da informação**, v. 17, n. 2, p. 118-135, 2012.

GOMES, E. **HIDRA: Arquivamento Digital De Alta-Confiabilidade Utilizando Auditoria Em Redes Peer-to-peer**. 2010. Dissertação (mestrado) – Universidade Federal do Paraná, Curitiba. Disponível em: http://www.acervodigital.ufpr.br/bitstream/handle/1884/24882/dissertacao.pdf

HEDSTROM, M. Digital preservation: a time bomb for digital libraries. **Computer and the Humanities**, v. 31, n. 3, p. 189-202, 1997/1998.

HUNTER, J.; CHOUDHURY, S. A semi-automated digital preservation system based on semantic web services. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, 4, 2004. Tucson, Arizona. **Proceedings.** Tucson: ACM/IEEE-CS, 2004.

LEE, K.; SLATERY, R. L.; McCARY, Y. The state of the art and pratice in digital preservation. **Journal of Research of the National Institute of Standards na Technology**, v. 107, n. 1, p. 93-106, 2002.

LORIE, R. A methodology and system for preserving digital data. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, 2., 2002 Portland, Oregon. **Proceedings**. Portland, ACM/IEEE-CS, 2002.

LUPOVICI, C.; MASANÈS, J. **Metadata for Long-term Preservation**, 2000. Disponível em: https://www.kb.nl/sites/default/files/docs/preservationmetadata.pdf Acesso em: 5 nov. 2017.

MACHADO, D. R. **Dados de pesquisas em repositório institucional: O caso do Edinburgh DataShare**. 2015. Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: http://hdl.handle.net/10183/119157

MANIATIS, P.; ROUSSOPOULOS, M.; GIULI, T.; ROSENTHAL, D.; BAKER, M. The LOCKSS peer-to-peer digital preservation system, **ACM Transactions on Computer Systems**, v. 23, n. 1, p. 2-50, 2005.

MÁRDERO ARELLANO, M. A. Preservação de documentos digitais. **Ciência da Informação**, v. 33, n. 2, p. 15-27, 2004.

NATIONAL LIBRARY OF AUSTRALIA (NLA). **Preservation Metadata for Digital Collections**, Canberra: NLA, 1999. Disponível em http://pandora.nla.gov.au/pan/25498/20020625-0000/www.nla.gov.au/preserve/pmeta.html Acesso em: 5 nov. 2017.

NATIONAL SCIENCE FOUNDATION (NSF). **Proposal Award Policies & Procedures Guide (PAPPG)**. Arlington, VA: NSF, Jan 30, 2017. Disponível em: https://www.nsf.gov/pubs/policydocs/pappg17_1/nsf17_1.pdf Acesso em: 16 nov. 2017.

OPEN ARCHVIVES INITIATIVE. **Object Reuse and Exchange**. Disponível em: http://www.openarchives.org/ore Acesso em:5 nov. 2017.

ONLINE COMPUTER LIBRARY CENTER. RESEARCH LIBRARY GROUP. **Data dictionary for preservation metadata: final report of the PREMIS Working Group**. 2005. Disponível em: http://www.oclc.org/content/dam/research/activities/pmwg/premis-report.pdf Acesso em: 5 nov. 2017.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). **OECD Principles and Guidelines for Access to Research Data from Public Funding**. Paris: OECD, 2007. Disponível em: https://www.oecd.org/sti/sci-tech/38500813.pdf Acesso em: 16 nov. 2017.

REICH, V. Distributed Digital Preservation. In: INDO-US WORKSHOP ON INTERNATIONAL TRENDS IN DIGITAL PRESERVATION, 2009. Pune, India. **Proceedings**. Pune, 2009.

RUSSELL, K.; ERGEANT, D.; STONE, A.; WEINBERGER, E.; DAY, M. **Metadata for Digital Preservation: The CEDARS Project Outline Specification Draft for Public Consultation**, 2000. Disponível em: https://pdfs.semanticscholar.org/ed88/a4807c0d9b30090b5919a076c0c8dfb83e5a.pdf Acesso em: 5 nov. 2017.

SAYÃO, L. F.; FARIAS SALES, L. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde,** Rio de Janeiro, v. 8, n. 2, p. 76-92, 2014. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/611

THIBODEAU, K. **Overview of technological approaches to digital preservation and challenges in coming years**. Washington DC: CLIR, 2002. Disponível em: https://www.clir.org/pubs/reports/pub107/thibodeau.html Acesso em: 1 nov. 2017.

THOMAZ, K. P.; SOARES, A. J. A preservação digital e o modelo de referência Open Archival Information System (OAIS). **DataGramaZero: Revista de Ciência da Informação,** v. 5, n. 1, 2004.

UNIVERSITY OF PITTSBURGH. SCHOOL OF INFORMATION SCIENCE (UPSIS). **Metadata Specifications Derived from the Fundamental Requirements: A Reference Model for Business Acceptable Communications**, 1996. Disponível em: http://web.archive.org/web/20000302194819/www.sis.pitt.edu/~nhprc/meta96.html Acesso em: 5 nov. 2017.

WATERS, D.; GARRETT, J. **Preserving Digital Information: Final Report of the Task Force on Archiving of Digital Information.** Disponível em: https://www.clir.org/pubs/reports/pub63watersgarrett.pdf Acesso em: 5 nov. 2017.

Recebido/Recibido/Received: 2017-11-17
Aceitado/Aceptado/Accepted: 2017-12-08