CURADORIA DIGITAL:
um novo patamar para preservação de dados digitais de pesquisa

*Luis Fernando Sayão**
*Luana Farias Sales***

RESUMO: Uma parte considerável dos resultados das atividades de pesquisa está sendo criada em formatos digitais. Embora de grande valor, esses dados estão sob o risco de serem perdidos pela obsolescência tecnológica e pela fragilidade inerente das mídias digitais. Dessa forma, a gestão de dados de pesquisa num ambiente distribuído e em rede se torna um desafio crescente para o mundo da pesquisa e para a ciência da informação. Como resposta a esse desafio surge o conceito de curadoria digital, que envolve a gestão de dados de pesquisa desde o seu planejamento, assegurando a sua preservação por longo prazo, descoberta, interpretação e reuso. Nessa direção, o presente estudo analisa brevemente a importância dos dados de pesquisa e a idéia de curadoria digital e seus impactos na formulação de novos documentos e na comunicação científica.

Palavras-chave: Curadoria digital. Dados de pesquisa. eScience. Preservação digital.

* Doutor em Ciência da Informação pela Universidade Federal do Rio de Janeiro, Brasil. Tecnologista da Comissão Nacional de Energia Nuclear, Brasil. Docente do Programa de Pós-Graduação em Biblioteconomia da Universidade Federal do estado do Rio de Janeiro, Brasil.
E-mail: lsayao@cnen.gov.br.

** Mestre em Ciência da Informação pela Universidade Federal Fluminense, Brasil. Analista de Ciência e Tecnologia da Comissão nacional de Energia Nuclear, Brasil. Doutoranda no Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Rio de Janeiro, Brasil.
E-mail: lsales@ien.gov.br.

1 INTRODUÇÃO

No período de 1918 a 1919 a gripe espanhola se espalhou pelo mundo inteiro, matando de 20 a 80 milhões de pessoas. De origem viral, não havia tratamento conhecido. Como veio, se extinguiu. Com o intuito de pesquisar meios de evitar uma nova catástrofe, a comunidade internacional das áreas médica e de saúde pública procurou por décadas algum vestígio biológico do vírus causador dessa enfermidade. Só depois de muito tempo, foi encontrada uma amostra de tecido humano infectado pelo vírus num hospital militar da Inglaterra. A partir desses vestígios estão sendo desenvolvidas pesquisas para se descobrir vacinas e meios de tratamento da gripe espanhola. As pesquisas em torno da amostra só se tornaram possíveis graças à preservação dos arquivos científicos, datados de 1916, daquele hospital militar (DITADI, 2003).

Diante do fato de que alguns dados de pesquisa são únicos e não podem ser substituídos se forem destruídos ou perdidos, a questão crucial que se coloca é a seguinte: será que os atuais registros médicos e os demais registros de pesquisa que agora estão sendo documentados de forma digital ou já são gerados em formatos digitais estarão disponíveis para o acesso e para a reutilização em novas pesquisas daqui a alguns anos? Essa questão tem implicações mais amplas, posto que o volume de dados de pesquisa disponibilizados digitalmente está crescendo numa velocidade vertiginosa, engendrando concepções novas de documentos e redesenhando o ciclo tradicional de comunicação científica. É necessário ainda observar que, além de gerar novos dados digitais, os pesquisadores e os acadêmicos, já há algum tempo, começaram a creditar toda a confiança nos conteúdos digitais criados por outros cientistas para dar prosseguimento aos seus empreendimentos (ABBOUT, 2008), inaugurando um novo patamar de compartilhamento de dados e um diálogo transversal ao tempo e ao espaço.

O ato cotidiano das instituições de pesquisa de registrar nos sistemas formais de informação - tais como arquivos, bibliotecas, repositórios, bases de dados - os resultados de suas pesquisas na forma de documentos, parece não ser suficiente para salvaguardar os dados obtidos ao longo do trabalho de pesquisa. Quando, por exemplo, um estudante de doutorado conclui a sua pesquisa e esta é registrada na forma de um documento que conhecemos por tese, teremos aí somente um retrato parcial dos conteúdos intelectuais gerados no desenrolar de anos de trabalho. Geralmente os dados de pesquisa - que dão sustentação à tese e que serão analisados e discutidos pelo autor - adormecerão armazenados em computadores e mídias pessoais que inexoravelmente serão tragados pela obsolescência tecnológica, pela fragilidade das mídias e, sobretudo, pela falta de intencionalidade de preservá-los adequadamente de forma que sirvam de ponto de partida para novas pesquisas. Isto porque os objetos digitais nunca sobrevivem inercialmente como os seus equivalentes impressos.

O fato determinante é que as atividades pesquisa – como de resto, a maioria dos empreendimentos humanos – estão crescentemente dependentes de materiais digitais. Para que haja avanço do conhecimento científico com um nível mais aceitável de duplicação de esforços, é necessário o estabelecimento de metodologias e compromissos de longo prazo que garantam a capacidade dos dados em formatos digitais, que estão sendo gerados agora, de serem acessados, interpretados e reutilizados com a tecnologia corrente à época do acesso. Portanto, o arquivamento persistente, a preservação digital e o estabelecimento de modelos de informação para a preservação de registros científicos estão se tornando questões-chave para as áreas de pesquisa.

Dados e informações digitais gerados pelas atividades de pesquisa necessitam de cuidados específicos, tornando-se necessário a criação de novos modelos de custódia e de gestão de conteúdos científicos digitais que incluam ações de arquivamento seguro, preservação, formas de acrescentar valor a esses conteúdos e de otimização da sua capacidade de reuso. No intuito de por em prática soluções para o problema, observa-se, no âmbito de várias disciplinas, um esforço em torno do desenvolvimento de repositórios digitais orientados especialmente para uma gestão ativa de dados de pesquisa. É nesse ambiente que surge o conceito de curadoria digital de dados científicos, cujo principal desafio recai na necessidade de se preservar não somente o conjunto de dados, mas de preservar, sobretudo, a capacidade que ele possui de transmitir conhecimento para uso futuro das comunidades interessadas. Isto significa que os ativos genuínos da pesquisa cientifica devem permitir que futuros usuários reanalisem os dados dentro de novos contextos. Porém, para que ocorra um processo de preservação em que os significados dos dados possam atravessar a barreira do tempo, é necessário assegurar que os usuários no futuro estejam instrumentados com as informações essenciais para o efetivo reuso dos dados (CONWAY, 2011). É de se esperar, portanto, que essas informações estejam estruturadas por modelos de informação e traduzidas por esquemas de metadados.

O objetivo desse estudo é analisar as questões envolvidas na curadoria digital de dados de pesquisa, para tal, discutiremos brevemente a importância dos dados científicos nos padrões atuais de pesquisa; o conceito de curadoria digital e o seu ciclo de vida; a ideia de documentos ampliados que podem vincular publicações acadêmicas, como são as teses, a dados científicos; e para finalizar, uma pequena reflexão a cerca dos impactos da curadoria digital sobre o ciclo tradicional de comunicação científica; à guisa de conclusão ousamos sugerir novas questões para serem investigadas no domínio da ciência da informação.

2 A CIÊNCIA ORIENTADA POR DADOS

A Declaração de Berlin sobre o Acesso Aberto ao Conhecimento em Ciências e Humanidades, publicada em 2003, amplia o escopo do que se entende por acesso livre ao definir que as "contribuições de acesso livre incluem resultados de pesquisas científicas originais, dados não processados e metadados, fontes originais, representações digitais de materiais pictóricos e gráficos e materiais acadêmicos multimídia."

A expansão do conceito de acesso livre, incorporando agora coleções de dados de pesquisa, vem se consolidando amparada por várias ações cultivadas no próprio seio das comunidades científicas, que reconhecem esses estoques de informação como uma parte do patrimônio da ciência universal e um pilar imprescindível para o seu avanço. O acesso aos dados de pesquisa torna-se, portanto, um imperativo para a ciência com reflexos globais, dado que os pesquisadores trabalham em cooperação internacional e os dados são criados, compartilhados e acessados globalmente; mas que têm um rebatimento nos planos locais e nacionais visto que esses mesmos pesquisadores estão, tipicamente, inseridos em estruturas de financiamento de pesquisa, políticas e organizações acadêmicas de âmbito nacional (BRASE; FARQUHAR, 2011).

"É extremamente raro que novas abordagens fundamentais para pesquisa e educação surjam. A Tecnologia da Informação abriu caminho para essas mudanças cruciais e as coleções digitais estão no cerne dessas transformações". Assim começa o relatório publicado nos Estados Unidos pela National Science Board (2005, p.9), cujo titulo "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century", expressa o reconhecimento da National Science Foudation (NSF)[1] para: a importância crescente das coleções digitais para a pesquisa e para a educação; a rápida multiplicação de coleções de dados com o potencial de serem gerenciadas por década através de processos de curadoria; e para necessidade de investimentos na criação e manutenção de coleções de dados. O Relatório justifica a importância desses itens reafirmando que os dados de pesquisa viabilizam análises em níveis sem precedentes de precisão e sofisticação e oferecem novos conhecimentos por meio da integração inovadora de informações. Pelo grande volume e pela complexidade, as coleções digitais proporcionam novos fenômenos para estudo, ao mesmo tempo em que são forças poderosas para diversos segmentos da educação (NATIONAL SCIENCE BOARD, 2005).

Entretanto, o excesso de dados gerados e armazenados pela pesquisa moderna, numa escala sem precedentes, está muito além da capacidade humana de análise. (CESAR JÚNIOR, 2011). Esse verdadeiro dilúvio de dados vem sendo desencadeado principalmente pelo avanço extraordinário de instrumentos, sensores e escalas, que aumentaram exponencialmente a capacidade de obtenção de dados pela realização de observações e medições de fenômenos, somados às informações geradas artificialmente por simulações e por *software*.

Esse fato coloca na agenda crítica da ciência um problema novo que é a gestão de dados de pesquisa num mundo digital interligado por redes de computadores, onde há um fluxo intenso de dados, proveniente de diferentes fontes, sendo gerados, processados e compartilhados em ambientes multidisciplinares. A partir desse ponto, instala-se, então, um desafio importante do nosso tempo, que é ao mesmo tempo uma oportunidade extraordinária e absolutamente essencial para se conduzir a pesquisa científica neste século que velozmente se inicia (LANNOM, 2011). Esse desafio traz em si uma questão essencial para os atuais pesquisadores, que é quase o enigma da esfinge: como traduzir em significado e conhecimento a torrente de dados que caracteriza certos domínios da ciência contemporânea?

A resposta a essa questão passa pelo delineamento de um novo paradigma para a ciência, sobre o qual o fazer científico é reordenado pela intensificação do uso de redes e de computadores e pelo uso sem precedentes de conjuntos de dados distribuídos. É o quarto paradigma, ciência orientada por dados ou a *eScience*. Nesse ambiente, novos diálogos se estabelecem entre cientistas da computação e da informação e especialistas de diferentes domínios para o desenvolvimento de novos conceitos e teorias a partir da grande quantidade de dados disponibilizados por diferentes tecnologias (CESAR JÚNIOR, 2011).

Mas quais foram os paradigmas anteriores? No começo da longa aventura da evolução da ciência havia apenas a ciência experimental, em seguida veio a ciência teórica – que pode bem ser ilustrada pelas Leis de Kepler, as Leis de Newton e as equações de Maxwell, entre outras. Depois, como nos relata Gray (HEY, 2009 p. xviii), "os modelos teóricos tornaram-se muito complicados para serem resolvidos analiticamente e as pessoas começaram a simular. Essas simulações nos acompanharam durante

1 Disponível em: <http://www.nsf.gov/>

a maior parte da última metade do último milênio". No quarto paradigma temos a ciência unificando experimentos, teorias e simulações, através do uso intensivo de dados capturados por instrumentos cada vez mais sofisticados ou gerados por simulação, processados por *software* e armazenados em computadores na forma de bases de dados. Com a finalidade de se extrair entendimento e inferir significado, a partir desse último ponto os dados podem ser analisados por meio de metodologias de gerenciamento de dados, de soluções estatísticas e também por meio do uso de ferramentas de representação do conhecimento, como ontologias.

É nítida, portanto, a linha que separa os dados dos seus significados. Bell (2011) nos relembra que Keppler (1571-1630) – assistente do astrônomo dinamarquês Tycho Brahe (1546-1601) – foi quem pegou o caderno de observações astronômicas sistemáticas de Brahe e a partir daí formulou as leis do movimento planetário. Este fato estabeleceu uma divisão clara entre a mineração e a análise de dados experimentais. Por um lado, temos os dados coletados e cuidadosamente arquivados; por outro, a criação de teorias. Esta divisão é um dos aspectos determinante do quarto paradigma.

Nesse contexto de grandes mudanças, novos papeis e responsabilidades emergem como críticos para a gestão de conjuntos de dados de pesquisa, dentre eles está o "cientista de dados" que podem ser cientistas da computação ou cientistas da informação, engenheiros de software e de base de dados, especialistas em disciplinas, entre outros. Apesar de não ser ainda uma carreira de contornos bem definidos e de reconhecimento óbvio, a sua contribuição é fundamental para um diálogo bem sucedido entre todas as partes envolvidas.

Explicitada rapidamente a importância das coleções de dados de pesquisa para o avanço da ciência moderna, concluímos esta seção constatando que a ciência com uso intensivo de dados consiste de três atividades essenciais: captura, curadoria e análise. Tendo em vista esse fluxo, Bell (2011) argumenta que é preciso investir na criação de um conjunto de ferramentas genéricas que cubram todo o espectro de atividades – da captura e validação dos dados à curadoria, analise e, finalmente, arquivamento permanente. Em todo esse ciclo se interpõe o desafio de manter a capacidade de interpretação dos dados e o seu potencial de reuso em vários outros contextos. Prosseguindo no nosso estudo, convidamos o leitor a concentrar a atenção nas soluções e modelos propostos para enfrentar esses desafios.

3 A IMPORTÂNCIA DA GESTÃO DOS DADOS DE PESQUISA

Compreendendo a importância da gestão ativa de coleções de dados para a pesquisa do século XXI a D-Lib Magazine[2] – o periódico mais importante no universo das pesquisas em bibliotecas digitais – publicou no inicio do ano de 2011 um número especial sobre esse assunto. Nessa publicação estão endereçadas questões como acesso livre, curadoria digital, aquisição e gestão, qualidade e confiabilidade e as possíveis conexões entre dados de pesquisa e as publicações acadêmicas tradicionais, que oferecem oportunidades para o surgimento de concepções surpreendentes de documentos mais apropriados ao paradigma da ciência computacional e orientada por dados.

O problema da gestão de dados de pesquisa tem muitas faces que vão se revelando à medida que avançamos. No plano econômico, o custo-benefício de se manter o acesso e a capacidade de reuso aos dados de pesquisa é extremamente difícil de ser mensurado. O valor de um registro pode estar relacionado à possibilidade da reprodutibilidade de um determinado experimento aonde ele foi gerado ou capturado. Algumas pesquisas podem ser fáceis e baratas de se replicar; outras podem ser literalmente impossíveis de se reproduzir – como é a mensuração das características de uma particular erupção vulcânica – ou são repetíveis somente a custos e esforços inaceitáveis (JANSEN, 2006), como uma incursão na atmosfera de Marte. Nessa direção, o arquivamento eletrônico de dados começa a ser estimulado ativamente pelas agências de financiamento de pesquisa, que demandam mais e mais que os projetos científicos contemplem o arquivamento dos dados gerados no decorrer das pesquisas em repositórios de dados confiáveis[3]. O que nos indica que as agências que

2 Disponível em: < http://www.dlib.org/dlib/january11/brase/01brase.html>
3 Para uma análise sobre a confiabilidade dos repositórios digitais, consultar: SAYÃO, L.F. Repositórios digitais confiáveis para a preservação de periódicos

financiam ou que estabelecem as diretrizes para o setor de pesquisa começam a delinear políticas, estratégias e prioridades que considerem os dados de pesquisa de longa duração como um investimento importante que precisa ser protegido como tal.

O Relatório do Projeto Digital Repository Infrastructure Vision for European Research II (Driver II), desenvolvido sob os auspícios da Comunidade Europeia, justifica a preocupação das agências de fomento enfatizando que o acesso a dados de pesquisa proporciona uma série de vantagens, especialmente quando esses dados estão associados a manuscritos acadêmicos disponíveis online. Por exemplo: quando um pesquisador deposita seus dados brutos, ele abre a possibilidade dos seus pares replicá-los e, dessa forma, verificar o que está sendo defendido na publicação científica; isto possibilita também que outros pesquisadores reusem os dados, os comparem e os combinem com outros dados, de forma que novas pesquisas podem ser geradas. Outro benefício apontado pelo Relatório é que a curadoria dos dados torna possível traçar a linhagem dos vários produtos dos projetos de *eScience*, dado que esses projetos se desenvolvem por vários estágios, tais como captura de dados, processamento, modelagem e interpretação. "Se fosse possível destacar as inúmeras conexões entre os recursos que são produzidos durante os vários estágios do processo científico, isto poderia ser de grande utilidade" (VERHAAR, 2008, p.14), enfatiza o autor do Relatório.

Entretanto, para muitas comunidades acadêmicas a gestão e o acesso continuado a esta vasta quantidade de dados ainda é um problema distante de ser superado. Lamentavelmente, muitos dos dados que são produzidos, frequentemente a um custo alto para a sociedade como um todo, são irremediavelmente perdidos.

No curto período do que se convencionou chamar de era digital, algumas instituições científicas se comprometeram no desenvolvimento de atividades que pudessem salvaguardas os dados científicos digitais. Porém as poucas instituições engajadas nesse processo ainda não estabeleceram práticas e não garantiram os fluxos de recursos que assegurem o completo sucesso da gestão desses dados. O que se observa é que ainda persistem lacunas críticas e questões de pesquisas em aberto (LEE; TIBBO, 2007).

Mesmo assim, várias iniciativas importantes, lideradas pelas próprias comunidades científicas, já cumprem papel vital na garantia do acesso livre aos dados de pesquisa e no que se convencionou chamar de curadoria digital, como veremos a seguir.

Ancorado no lema "ajudando você a encontrar, acessar e reusar dados", foi fundada em Londres no ano de 2009 uma organização sem fins lucrativos, chamada de DataCite[4], cujos objetivos essenciais, desde então, são: estabelecer bases para o acesso mais fácil a dados de pesquisa na internet; aumentar o grau de aceitação dos dados de pesquisa como contribuições legítimas passíveis de serem citadas nos registros acadêmicos; dar sustentação ao arquivamento de dados de pesquisa de forma que seja possível que os resultados possam ser verificados e readaptados para futuros estudos.

A ideia central que alimenta as ações do DataCite é a citação de dados, significando que os dados de pesquisa devem ser citados da mesma forma como são citadas outras fontes de informação, tais como artigos e livros. O DataCite preconiza que a citação de dados permite o reuso e a verificação dos dados mais facilmente, possibilitando que o impacto dos dados possam ser rastreados, e que uma estrutura acadêmica que reconheça e recompense os produtores de dados possa ser, finalmente, criada.

Para cumprir seus objetivos o DataCite procura juntar as comunidades que lidam com conjunto de dados de pesquisa para que, de forma colaborativa, equacionem o desafio de tornar os dados de pesquisa visíveis e possíveis de serem acessados. Uma das iniciativas importantes nesse processo é o apoio aos centros de dado no assinalamento de identificadores persistentes e na definição de padrões para a publicação de dados; destaca-se também apoio aos editores científicos no sentido de os capacitarem a estabelecer *links* entre artigos e os dados subjacentes e eles. Para o usuário pesquisador, o DataCite oferece recursos e serviços que o ajudam a encontrar, identificar e citar conjunto de dados de forma confiável.

eletrônicos científicos. **RPA**, v.4, n.3, 2010. Disponível em < http://www.portalseer.ufba.br/index.php/revistaici/article/view/4709/3565>. Acesso em: 11 nov. 2011.

4 Disponível em: <http://datacite.org/>

Temos que considerar também o OpenAIRE[5] – Open Access Infrastructure for Research in Europe – que é um projeto de duração de três anos, iniciado em dezembro de 2009, cujo objetivo é apoiar a implementação do acesso aberto na Europa. Para isso vem estabelecendo uma ampla infraestrutura baseada numa rede distribuída de pontos de contato nacionais e regionais nos países europeus, que assegure o apoio localizado aos pesquisadores no seu próprio ambiente. O Projeto está focado em três principais objetivos, dentre eles está a gestão de dados científicos e a sua vinculação com publicações científicas, como explicitado na sua página web: "trabalhar com várias comunidades temáticas para explorar os requisitos, práticas, incentivos, fluxos de trabalho, modelos de dados e tecnologias para depósito, acesso e manipulação de *conjunto de dados de pesquisa* de várias formas em combinação com publicações de investigação científica."

Outra iniciativa essencial, porém com uma perspectiva voltada para a preservação e reuso de dados de pesquisa, é o Digital Curation Centre (DCC)[6], que é um ponto de disseminação de práticas e conhecimentos na área de curadoria digital. O lema que está estampado na sua *home page*, resume e justifica a importância das suas atividades: "porque boa pesquisa precisa de bons dados". O modelo de curadoria digital de dados científicos proposta pelo DCC será tratado com um grau a mais de detalhe nas seções seguintes.

4 AFINAL, O QUE É CURADORIA DIGITAL?

Os conhecimentos e as práticas acumulados na última década em preservação e acesso a recursos digitais resultaram num conjunto de estratégias, abordagens tecnológicas e atividades que agora são coletivamente conhecidas como "curadoria digital". Ainda que seja um conceito em evolução, já está estabelecido que a curadoria digital envolve a gestão atuante e a preservação de recursos digitais durante todo o ciclo de vida de interesse do mundo acadêmico e científico, tendo como perspectiva o desafio temporal de atender a gerações atuais e futuras de usuários Torna-se claro, portanto, que subjacente às metodologias utilizadas pela curadoria digital estão os processos de arquivamento digital e de preservação digital; porém, inclui também as metodologias necessárias para a criação e gestão de dados de qualidade e a capacidade de adicionar valor a esses dados no sentido de gerar novas fontes de informação e de conhecimento (LEE; TIBBO, 2007).

O DCC, na sua visão fundacional, nos informa, na sua página *web*, que a curadoria digital "envolve a manutenção, a preservação e a agregação de valor a dados de pesquisa durante o seu ciclo de vida"; e que a gestão ativa sobre esses dados reduz as ameaças ao seu valor de longo prazo e minimiza os riscos da obsolescência digital. Além de reduzir a duplicação de esforços na criação de dados de pesquisa, a curadoria reforça o valor de longo prazo dos dados existentes quando os tornam disponíveis para a reutilização em novas pesquisas de qualidade.

Abbott (2008) amplia um pouco mais a ideia de curadoria digital definindo-a como todas as atividades envolvidas na gestão de dados, desde o planejamento da sua criação – quando os sistemas são projetados -, passando pelas boas práticas na digitação, na seleção dos formatos e na documentação, e na garantia dele estar disponível e adequado para ser descoberto e reusado no futuro. A curadoria digital também inclui a gestão de grandes conjuntos de dados para uso diário, assegurando, por exemplo, que eles possam ser pesquisados e continuem viáveis, ou seja, capazes de serem lidos e interpretados continuamente. Nessa perspectiva, a ideia de curadoria digital estende-se além do controle do repositório que arquiva os recursos e envolve a atenção do criador do conteúdo e dos usuários futuros.

Portanto, verifica-se um deslocamento no padrão de arquivamento estático e inacessível promovido pelos *dark archives*, repositórios de acesso restrito voltados para garantir integridade e autenticidade. O foco da curadoria digital está na gestão por todo o ciclo de vida do material digital, de forma que ela permaneça continuamente acessível e possa ser recuperado por quem dele precise. Ampliando a capacidade dos dados serem recuperados e acessados estão os modelos de informação, expressos por metadados; além do mais, os metadados são também ferramentas importantes para os procedimentos de controle de autenticação (HIGGINS, 2011).

5 Disponível em: <http://www.openaire.eu/>
6 Disponível em: < http://www.dcc.ac.uk/>

A curadoria digital, em resumo, assegura a sustentabilidade dos dados para o futuro, não deixando, entretanto, de conferir valor imediato a eles para os seus criadores e para os seus usuários. Os recursos estratégicos, metodológicos e as tecnologias envolvidas nas práticas da curadoria digital facilitam o acesso persistente a dados digitais confiáveis por meio da melhoria da qualidade desses dados, do seu contexto de pesquisa e da checagem de autenticidade. Dessa forma, a curadoria contribui para assegurar a esses dados validade como registros arquivísticos, significando que eles podem ser usados no futuro como evidência legal. O uso de padrões comuns entre diferentes conjuntos de dados, proporcionado pela curadoria digital, cria mais oportunidades de buscas transversais e de colaboração. Na ótica financeira, o compartilhamento, o reuso dos dados e as oportunidades de novas análises, além de outros benefícios, valorizam e protegem o investimento inicial na obtenção dos dados.

A curadoria digital emerge como uma nova área de práticas e de pesquisa de espectro amplo que dialoga com várias disciplinas e muitos gêneros de profissionais.

5 CICLO DE VIDA DA CURADORIA DE DADOS CIENTÍFICOS

O DCC oferece um Modelo do Ciclo de Vida da Curadoria, expresso graficamente[7], que reflete uma visão de alto nível dos estágios necessários para o sucesso do processo de curadoria e de preservação de dados de pesquisa, que se inicia no estágio de conceitualização ou de recebimento do dado no repositório. O modelo proposto pelo DCC está orientado para o planejamento das atividades de curadoria nas organizações ou consórcios ajudando a garantir que todos os passos do ciclo serão cumpridos. Entretanto, isto não implica que todas as organizações devam cumprir o ciclo desde o primeiro estágio, na realidade, a operacionalização dos estágios dependerá das necessidades reais de cada organização.

Os elementos-chaves do modelo são: **dado**, **objetos digitais** e **bases de dados**. No centro do ciclo de vida da curadoria está o dado, que é qualquer informação codificada em formato binário. A ideia de dado inclui: os objetos digitais simples, que são aqueles compostos por um único arquivo, identificador e metadados, e os objetos digitais complexos, que por sua vez são formados pela combinação de outros objetos digitais formando uma unidade discreta, como é, por exemplo, uma página *web*. Nesse contexto, base de dados é definida como coleções estruturadas de registros ou de dados armazenados em sistemas de computadores.

O outro elemento básico do modelo são as **ações** que devem ser tomadas no decorrer do processo de curadoria. O modelo do DCC classifica as ações em três tipos: **ações para todo o ciclo de vida**; **ações sequenciais** e **ações ocasionais**.

As ações para todo o ciclo de vida são assim chamadas por compreenderem atividades que permeiam todo o ciclo de vida da curadoria digital. Para transmitir essa ideia de presença contínua, essas ações estão representadas graficamente como anéis concêntricos envolvendo os objetos de dados que estão no centro do modelo. As ações são as seguintes:

- D**escrição e a representação da informação** - é efetivada pela atribuição de metadados administrativos, técnicos, estruturais e de representação de acordo com os padrões apropriados; visa assegurar a descrição adequada e o controle de longo prazo; compreende também a coleta e a atribuição de informações de representação necessárias para o entendimento do dado e para a sua apresentação (ou renderização).
- **Planejamento da Preservação –** é necessária a definição de um plano de preservação cujo espectro englobe todo o ciclo de vida da curadoria do material digital, incluindo gestão, administração, politicas, e tecnologias.
- **Participação e monitoramento –** enfatiza a necessidade de atenção para as atividades que se desenrolam no âmbito das comunidades envolvidas com o problema de curadoria, bem como a necessidade de participação no desenvolvimento de padrões, de ferramentas e de *software* adequados ao problema e que possam também serem compartilhados;

7 Disponível em: <http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf >

- **Curadoria e preservação** - estar continuamente alerta e empreender as ações administrativas e gerenciais planejadas para a curadoria e preservação por todo o ciclo de vida da curadoria.

As **Ações Sequenciais**, por sua vez, são etapas que devem ser cumpridas repetidamente para assegurar que o dado permaneça em contínuo processo de curadoria de acordo com as melhores práticas. Essa sequência não é para ser cumprida meramente uma vez do começo ao fim; na realidade ela forma as bases da cadeia de curadoria e continua ciclicamente todo o tempo que o dado estiver sob curadoria.

A sequência de ações do modelo de ciclo de vida da curadoria digital proposto pelo DCC tem os seguintes estágios:

- **Conceitualização** – conceber e planejar a criação do dado, incluindo os métodos de captura e as opções de armazenamento; questões tais como propriedade intelectual, embargos e restrições, financiamento, responsabilidades, objetivos específicos da pesquisa, ferramentas de captura e calibração devem ser registradas.
- **Criação e/ou Recebimento** – compreende a criação do dado incluindo o elenco de metadados necessários à sua gestão e compreensão, ou seja, metadados administrativos, descritivos, estruturais e técnicos (os metadados de preservação também podem ser incluídos no momento da criação do dado). Nem sempre os dados são arquivados por quem os gerou, dessa forma, esse estágio inclui também a recepção dos dados segundo políticas bem documentadas, sejam dos seus criadores, de outros arquivos, de repositórios ou centro de dados; quando necessário, assinalar metadados apropriados para a curadoria e a preservação dos dados recepcionados.
- **Avaliação e seleção** – avaliar o dado e selecionar o que será objeto dos processos de curadoria e de preservação por longo prazo; manter-se aderente tanto às boas práticas quanto às políticas pertinentes e também às exigências legais.
- **Arquivamento** – transferir o dado para um arquivo, repositório, centro de dados ou outro custodiante apropriado.
- **Ações 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 adição de metadados de preservação e de informação de representação e a garantia de estruturas de dados ou formatos de arquivos aceitáveis.
- **Armazenamento –** armazenar o dado de forma segura mantendo a aderência aos padrões relevantes.
- **Acesso, uso e reuso –** garantir que o dado possa ser cotidianamente acessado tanto pela sua comunidade-alvo, quanto pelos demais usuários interessados no reuso do dado; isto pode ser realizado por meio de publicação disponível para as comunidades interessadas; porém, controle de acesso e procedimentos de autenticação podem ser aplicados.
- **Transformação –** compreende a criação de 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.

Por fim, o Modelo estabelece também os estágios que são aplicados eventualmente, chamados de ações ocasionais. Essas ações interrompem ou reordenam as ações sequenciais como desdobramento de uma decisão. Por exemplo, após uma avaliação pode ser decidido que o dado em questão não se enquadra no escopo de um repositório digital, isso implica que o dado deve ser transferido para outro custodiante. Em outra situação, o dado deve ser destruído, possivelmente por motivações legais.

- **Eliminação –** eliminar os dado que não foram selecionados para curadoria e preservação de longo prazo de acordo com políticas documentadas, diretrizes ou exigências legais.
- **Reavaliação –** retornar ao dado cujos procedimentos de avaliação foram falhos

para nova avaliação e possível seleção para curadoria.

- **Migração –** migrar os dados para um formato diferente; isto pode ser feito no sentido de compatibilizá-lo com o ambiente de armazenamento ou para assegurar a imunidade do dado contra a obsolescência de *hardware* e de *software*.

O modelo desenhado pelo DCC permite uma visão coletiva sobre o conjunto de funções necessárias à curadoria e à preservação de dados de pesquisa. Além de definir papeis, responsabilidades e conceitos, ele explicita a infraestrutura de padronização e as tecnologias que devem ser implementadas.

6 JUNTANDO DADOS E PUBLICAÇÕES: documentos ampliados

Não obstante todas as transformações comportamentais e sociais decorrentes do aparato tecnológico que permeia e dinamiza as atividades de pesquisa, a infraestrutura atual de comunicação científica ainda está fortemente centrada no armazenamento e na disseminação de recursos informacionais individuais. Partindo dos modelos de publicação na *web* e voltando aos sistemas formais de informação acadêmica, como as bibliotecas de pesquisa, verifica-se que eles entregam ao usuário basicamente um artigo ou uma monografia. "Muitos editores acadêmicos não aceitam outro produto de projetos de e-pesquisa, tais como base de dados, gravação de vídeos e serviços web" (VERHAAR, 2008, p.9). O que parece cada vez mais claro é que a heterogeneidade e a complexidade dos registros de resultados de pesquisa não podem mais ser expressas por documentos convencionais únicos, impressos ou mesmo digitais.

Recentemente vários estudos se concentraram na possibilidade de se entrelaçar produtos de e-pesquisa que se encontram distribuídos, gerando novas modalidades de documentos científicos. Por exemplo, Hunter (2007) visualizou um "pacote de publicações científicas" que encapsula e relaciona, na forma de objetos compostos, dados brutos com os seus subprodutos, publicações e metadados contextuais, de proveniência e administrativos. Enquanto Gray (HEY, 2009), no contexto de sua proposta de um método científico transformado, conceitualiza os "documentos sobrepostos" - *overlay documents*, no original em inglês -, que interligam artigos de periódicos revisados por pares, dados, anotações e comentários.

Nessa mesma direção, o Open Archive Initiative (OAI) define uma norma para descrição e intercâmbio de agregação de recursos *web* chamada de Object Reuse and Exchange (OAI-ORE). "Esta agregação, algumas vezes chamada de objetos digitais compostos, pode combinar recursos distribuídos com tipos múltiplos de mídia, incluindo texto, imagens, dado e vídeo. O objetivo da norma é expor o conteúdo rico dessa agregação para aplicações que suportem sistemas de autoria, depósito, intercâmbio, visualização, reuso e preservação", conforme explicitado na página *web* do OAI-ORE. A norma equaciona o problema básico que é a ausência de forma padronizada para descrever os elementos constituintes do objeto digital composto e os limites de uma agregação (LAGOZE; SOMPEL, 2008).

O Projeto DRIVER II tem como alvo investigar as formas pela qual a disponibilidade de dados de pesquisa pode ser usada para ampliar as publicações acadêmicas tradicionais. O documento abstrato que combina texto e dados de pesquisa é chamado de *enhanced publication* – termo ainda sem tradução para o português, mas que poderíamos, traduzi-lo por documento ampliado –, emerge da compreensão de que as publicações tradicionais são limitadas na sua capacidade de incorporar resultados de todo o ciclo do processo de investigação científica. Isso acontece especialmente quando grandes conjuntos de dados são gerados. Nesse momento fica evidente que os textos acadêmicos só podem apresentar os dados de pesquisa de forma condensada.

É um fato promissor observar que crescentemente os dados de pesquisa estão sendo armazenados em repositórios de dados confiáveis, onde, gerenciados sob os princípios da curadoria digital são preservados e mantêm a sua capacidade de reuso. Entretanto, na atual infraestrutura de comunicação científica estes conjuntos de dados não são conectados às publicações onde eles são discutidos e

analisados. A ideia que está por traz das publicações ampliadas é precisamente criar pontes que liguem os conteúdos dos repositórios institucionais, ou seja, publicações científicas, com os conteúdos dos repositórios de dados (VERHAAR, 2008).

Dessa forma, a publicação ampliada ou o documento ampliado é pensado como uma forma de objeto digital complexo que combina vários recursos heterogêneos, que são, porém, relacionados. A base para esse tipo de objeto ainda é a publicação acadêmica tradicional, por exemplo, uma tese e os seus conjuntos de dados gerados, somada também com os metadados necessários.

7 OS DADOS CIENTÍFICOS E A COMUNICAÇÃO CIENTÍFICA

De uma forma definitiva a ciência orientada por dados cria um ponto de inflexão no ciclo tradicional da comunicação científica. Disciplinas como física das partículas, química, astronomia, geologia, dependem de forma absoluta do uso intensivo de ambientes de rede altamente distribuídos, instrumentos automatizados, técnicas de captura de imagens e programas de simulação. Esse aparato tecnológico tem impactado ampla e profundamente a forma como os cientistas podem conduzir e disseminar as suas pesquisas (VERHAAR, 2008), desenhando novos fluxos de cooperação e compartilhamento e definindo conceitos inéditos para a comunicação e para o registro científico, que merecem estudos partindo de muitos olhares.

No domínio específico da curadoria digital, são inúmeras as reflexões que se podem fazer face aos impactos do reuso de dados de pesquisa, da publicação e da citação de coleções de dados e a partir do estabelecimento de novos conceitos de publicações acadêmicas - mais complexas e mais heterogêneas - sobre o ritual de comunicação científica. De uma forma geral, a curadoria de dados científicos adiciona velocidade ao ciclo da comunicação científica na medida em que oferece aos pesquisadores dados prontos para o reuso, ou seja, dados tratados, acompanhados por metadados semânticos e estruturais – que asseguram a fidedignidade de seu significado e a reconstrução correta de sua apresentação, somados a metadados que asseguram a integridade, precisão e autenticidade. Dessa forma, novas pesquisas de qualidade podem ser desenvolvidas, com a segurança necessária, a partir desses dados, que estão instrumentalizados para serem transportados para novos domínios. Pode-se observar que uma nova relação se estabelece entre os pesquisadores na medida em que um pesquisador, para desenvolver seus projetos, pode depositar toda a confiança nos dados levantados por outro, distante no tempo e no espaço.

Assim como se debate hoje fortemente a questão do acesso livre aos periódicos acadêmicos, criando-se novos modelos de disseminação de resultado de pesquisa - mais ágeis e mais dinâmicos e organicamente mais próximos das comunidades científicas -, hoje fica claro que é preciso estender o movimento de livre acesso também aos dados científicos, posto que esses recursos constituem uma parte imprescindível do estoque de conhecimento acumulado pelo trabalho acadêmico e de pesquisa, e que são financiados, na maioria das vezes, pelo dinheiro público. As facilidades propostas pelas organizações que lidam com dados de pesquisa para encontrar, identificar, arquivar, adicionar valor e reusar esses dados criam um novo canal de diálogo entre os acadêmicos e pesquisadores, que se reflete nos modelos de socialização acadêmica e de comunicação científica.

No novo ambiente de pesquisa redesenhado pelas práticas da *eScience*, o ciclo de vida da curadoria digital incorpora-se como uma peça-chave no fluxo tradicional de comunicação científica baseado tradicionalmente em artigos de periódicos. A curadoria digital, no momento em que gerencia e preserva os dados de pesquisa para que sejam acessados e compreendidos por outros pesquisadores estabelecendo um dialogo com o futuro, cria a possibilidade de se criar conceitos inovadores de documentos de registros de pesquisa, rompendo com o paradigma unidimensional e absoluto do artigo de periódico.

8 À GUISA DE CONCLUSÃO

A tecnologia digital nos coloca diante de um dos dilemas mais críticos do nosso tempo:

por um lado ela nos permite criar, manipular, armazenar e tornar disponível uma quantidade impressionante de informações; por outro lado, esta mesma tecnologia fugidia coloca em perigo a longevidade dos objetos informacionais por ela engendrada, colocando a humanidade – que depende cada vez mais dos estoques informacionais digitais – face a face com o perigo de uma amnésia digital. Isto porque os objetos digitais requerem metodologias de gestão que são muito diferentes das que são utilizadas no universo da impressão tradicional.

Uma das atividades humanas em que mais se gera e se manipula materiais digitais é precisamente o trabalho de pesquisa científica. Em alguns nichos específicos, a totalidade das atividades que se desenrolam nos laboratórios distribuídos está centrada num intenso fluxo de dados, nos mais diversos formatos digitais. Era de se esperar, portanto, que surgissem iniciativas que pudessem tornar os dados científicos digitais mais visíveis e sempre possíveis de serem acessados, mantendo a sua integridade, fidedignidade e o seu papel de evidência.

Nessa direção, a curadoria digital emerge como uma nova área de práticas e de pesquisa de espectro amplo que dialoga com várias disciplinas e muitos gêneros de profissionais. Ela une as tecnologias e boas práticas do arquivamento e da preservação digital e dos repositórios digitais confiáveis com a gestão dos dados científicos, criando uma nova área de pesquisa cujos desdobramentos, de amplo espectro, ainda são imprevisíveis. Isto porque, como se trata de uma área que só recentemente despontou como crítica para a pesquisa, ainda restam muitas lacunas práticas e teóricas a serem equacionadas, orientadas, preferencialmente, por uma abordagem multidisciplinar.

A Biblioteconomia e a Arquivologia, que se renovam cotidianamente para enfrentar novos problemas, têm muito a contribuir para a curadoria digital com suas experiências em gestão de patrimônios intangíveis. Representação e organização do conhecimento, os novos conceitos de bibliotecas, repositórios e arquivos digitais, a integridade e autenticidade de materiais digitais e a recuperação da informação, para citar alguns itens, são imprescindíveis para a gestão de coleções de dados de pesquisa; a Museologia digital, por sua vez, pode trazer aportes importantes na questão dos objetos digitais complexos e multimidiáticos, cuja presença é comum na curadoria de exposições museológicas virtuais e pode ser interessante para renderização de estruturas cientificas mais sofisticadas.

Porém, para a Ciência da Informação, os impactos nos obrigam a repensar alguns pontos críticos, como no conceito ancestral de documento, no modelo tradicional de disseminação de resultados de pesquisa e na extensão dos formatos de metadados como instrumentos de recomposição de significados e estruturas.

Esses pontos nos inspiram a propor novos itens para uma agenda de pesquisa dentro do domínio interdisciplinar da Ciência da Informação:

a) em primeiro lugar, seria importante avaliar como o ciclo da comunicação científica se altera mediante as novas formas de colaboração, socialização e disseminação proporcionadas pelo reuso de dados científicos, especialmente em áreas de conhecimento com maiores interfaces com a *eScience*;
b) em segundo, seria interessante investigar as novas modalidade de publicação científica, cuja gênese está na vinculação entre as publicações tradicionais depositadas repositórios em digitais temáticos e institucionais com os dados gerenciados pelos centros de dados e de curadoria digital;
c) por fim, em terceiro mas não menos importante, está a concepção de modelos de informação que possam orientar a definição de conjunto de metadado capazes de garantir significado, estrutura, fidedignidade e autenticidades aos dados de pesquisa — pelo tempo que for necessário.

CURATORSHIP DIGITAL: a new platform for digital preservation of research data

ABSTRACT: *A considerable part of the results of research activities is being created in digital formats. Although valuable, these data are at risk of being lost by technological obsolescence and by the inherent fragility of digital media. Thus, the management of research data in a digital networked and distributed environment becomes an increasing challenge for the world of research and for the information science. In response to this challenge arises the concept of digital curation, which involves the management of research data from its planning, ensuring its long-term preservation, discovery, interpretation and reuse. In this sense, this study briefly examines the importance of research data and of the idea of digital curation and its impact on the formulation of new documents and scientific communication.*

Keywords*: Digital Curation. Research Data. eScience. Digital Preservation.*

Artigo recebido em 22/02/2012 e aceito para publicação em 22/02/2012

REFERÊNCIAS

ABBOUT, Daisy**. What is digital curation?** Edinburgh, UK : Digital Curation Centre, 2008. Disponível em: <http://www.era.lib.ed.ac.uk/bitstream/1842/3362/3/Abbott%20What%20is%20digital%20curation_%20_%20Digital%20Curation%20Centre.doc>. Acesso em: 20 dez. 2011.

BELL, Gordon. Prefácio. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da eScience. São Paulo : Oficina do Texto, 2011.

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Berlin, 2003. Disponível em : <http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf>. Acesso em: 20 dez. 2011

BRASE, Jan; FARQUHAR, Adam. Access to research data. **D-Lib Magazine**, v. 17, n. 1/2, Jan. / Feb. 2011.

CESAR JÚNIOR, Roberto Marcondes**. Do mundo aos dados e dos dados ao conhecimento**. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma: descobertas científicas na era da eScience**. São Paulo : Oficina do Texto, 2011.

CONWAY, Esther et al. Curating scientific research data for the long term: a preservation analysis method in context**. The International Journal of Digital Curation,** n. 2, v.6, 2011**.**

DITADI, Carlos**. Preservação de documentos eletrônicos.** Rio de Janeiro **:** Arquivo Nacional/CTDE, 2003.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. Jim Gray on eScience: A Transformed Scientific Method. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **The Fourth Paradigm**: Data-Intensive Scientific Discovery, 2009. Disponível em: <http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf>. Acesso em: 20 dez. 2011.

HIGGINS, Sarah. Digital curation: the emergence of a new discipline**. The International Journal of Digital Curation,** v.6, n. 2, 2011. Disponíel em: <http://www.ijdc.net/index.php/ijdc/article/view/184>. Acesso em: 20 dez. 2011.

HUNTER, Jane. Scientific publication packages - A selective approach to the communication and archival of scientific output. **The International Journal of Digital Curation**, v.1, n.1, 2006. Disponível em: <http://www.ijdc.net/index.php/ijdc/article/view/8/4>. Acesso em: 13 jan. 2012.

JANSEN, Hans. Permanent access to electronic journals. **Information Services & Use**, v. 26, 2006. Disponível em: <http://iospress.metapress.com/content/7drby91r8t4gf8ap/fulltext.pdf >. Acesso em: 10 nov. 2010.

LAGOZE, Carl; SOMPEL, Herbert Van de. **Ore user guide – primer**. Open Archive

Initiative, 2008. Disponível em: <http://www.openarchives.org/ore/1.0/primer.html>. Acesso em: 13 jan. 2010.

LANNOM, Laurence. Research Data**. D-Lib Magazine,** v. 17, n. 1/2, Jan. / Feb. 2011. Disponível em: <http://www.dlib.org/dlib/january11/01editorial.html 2011>. Acesso em: 20 dez. 2011.

LEE, Cristopher; TIBBO, Helen. Digital curation and trusted respositories: steps toward success**. Journal of Digital Information**, v. 8, n. 2, 2007. Disponível em: <http://journals.tdl.org/jodi/article/viewArticle/229/183>. Acesso em: 20 dez. 2011.

NATIONAL SCIENCE BOARD. **Long-lived digital data collections**: enabling research and education in the 21st century. National Science Foundation, sept. 2005. Disponível em: <http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf>. Acesso em: 01 fev. 2012.

VERHAAR, Peter**. Report on object models and functionalities.** DRIVER II, 2008. Disponível em: <https://openaccess.leidenuniv.nl/bitstream/handle/1887/16018/Report_on_Object_Models_and_Functionalities.pdf?sequence=2>. Acesso em: 20 dez. 2011.