DADOS DE PESQUISA: contribuição para o estabelecimento de um modelo de curadoria digital para o país

Luís Fernando SAYÃO
Doutor em Ciência da Informação pelo Convênio IBICT – UFRJ
Tecnologista do Centro de Informações Nucleares da Comissão Nacional de Energia Nuclear
lsayao@cnen.gov.br

Luana Farias SALES
Doutoranda em Ciência da Informação pelo Convênio IBICT – UFRJ
Analista em Ciência e Tecnologia do Instituto de Engenharia Nuclear da Comissão Nacional de Energia Nuclear
lsales@ien.gov.br

Resumo
A atividade de pesquisa científica, no contexto da eScience, produz e utiliza uma quantidade extraordinária de dados de pesquisa. Com a proliferação dos dados, se destaca a preocupação de como essas coleções de dados podem ser preservados para uso e reuso no futuro. O desafio da curadoria digital de dados científicos está na necessidade de preservar não somente a coleção de dados, mas também a sua capacidade de transmitir conhecimento para usuários futuros, permitindo-os reanalisar os dados em novos contextos. A gestão de dados de pesquisa é considerada essencial para condução da pesquisa científica no século XXI, mas os dados só podem ser gerenciados e preservados ao longo do tempo e do espaço por meio de compromissos institucionais sustentáveis. Com o objetivo de oferecer uma contribuição, o presente estudo apresenta uma análise multifacetada dos elementos necessários para a definição de um modelo de curadoria digital para o país. Foi tomado como principal recurso metodológico o exame de três relatórios considerados fundamentais no endereçamento de questões de curadoria de dados de pesquisa: os relatórios da *National Science Foundation* (NSF), do *Digital Data Curation* (DCC) e da Organização para a Cooperação e Desenvolvimento Econômico (OCDE). Como resultado, o estudo apresenta uma apreciação dos seguintes pontos: aspectos políticos, infraestrutura organizacional e tecnológica, pesquisa em curadoria digital, desenvolvimento de coleções, formação de especialistas, sustentabilidade econômica, implicações sociais, éticas e legais e oferecimento de serviços.
Palavras-chave: Dados de pesquisa. Curadoria digital. Preservação digital. eScience.

RESEARCH DATA: CONTRIBUTION TO THE STABLICHMENT OF A DIGITAL SCIENTIFIC DATA CURATION MODEL TO THE COUNTRY

Abstract
The research activity in the context of eScience yields and uses an extraordinary volume of digital research data. Along the proliferation of data has raised the concern for how those data collections are to be preserved for future use and reuse. The challenge of digital scientific data curation lies in the need to preserve not only data collections but also the ability they have to provide knowledge to future users, allowing them to reanalyze the data within new contexts. The management of research data is considered essential to the conduct of scientific research in 21st century, but digital research data can only be managed and preserved along the time and space through a sustained institutional commitment. To address this problem the present study presents a multi-faceted analysis on the elements necessary to defining national model of digital data curation. It was taken as the main

methodological resource the examination of three reports regarded as fundamental in addressing issues of research data curation: reports from the National Science Foundation (NSF), Digital Data Curation (DCC) and Organization for Economic Cooperation and Development (OECD). As a result, the study presents an assessment of the following issues: political aspects, requiring technologies, development of data collections, organizational infrastructure, research on data curation, professional skills, economic sustainability, social, legal and ethical aspects and provisions of services.
Key-words: Research data. Digital data curation. Digital preservation. eScience.

1 INTRODUÇÃO

A atividade de pesquisa cientifica do século XXI produz uma quantidade extraordinária de dados, principalmente em formatos digitais. Isto acontece essencialmente porque a tecnologia digital se torna cada vez mais um elemento onipresente nos processos da construção do conhecimento científico, seja por aumentar a capacidade dos instrumentos científicos, seja por reconstruir realidades por meio de simulação, ou ainda inaugurando formas inéditas de colaboração e compartilhamento de dados e informações. A capacidade dos computadores, que não para de crescer, aliada às possibilidades abertas pela Internet abriram novas aplicações para as fontes básicas de pesquisa – os dados de pesquisa – dando um novo impulso ao trabalho científico.

Em um contexto mais amplo, o fenômeno chamado *Big Data* descreve “um conjunto de problemas e suas soluções tecnológicas em computação aplicada com características que tornam seus dados difíceis de tratar” (XEXÉO, 2013, p.19). Existe um consenso de que as principais características deste fenômeno são marcadas por três “Vs” (volume, velocidade e variedade); apesar do termo ter sido criado para designar o impacto mercadológico da grande geração de dados é possível observar suas marcas também no fazer científico, tornando a problemática da variedade e volume de dados gerados em grande velocidade uma preocupação necessária no domínio da pesquisa científica.

Dessa forma, nas esferas científicas, as expectativas em torno de um mundo rico em dados são imensas e incluem desde descobertas de novas drogas, passando por um entendimento melhor sobre as mudanças climáticas e sobre a origem do universo, até metodologias mais apuradas para examinar a história e a cultura. A relevância dos dados no contexto das “big sciences”, como Astronomia, Física e Biologia, conduziu não somente ao surgimento de novos modelos de ciência – coletivamente chamados de “Quarto paradigma científico” ou “eScience” – mas possibilitou a emergência de novos campos de estudo como a Astroinformática e a Biologia Computacional (BORGMAN, 2010). Nessa direção,

pesquisadores em áreas específicas e cientistas da computação trabalham colaborativamente em muitos campos, definindo novos domínios de conhecimento e redesenhando os contornos disciplinares da ciência.

Assim, como desdobramento desse fato, torna-se essencial, para que a ciência contemporânea enfrente os desafios globais de sua agenda crítica, o desenvolvimento de metodologias - tecnológicas e gerenciais – que orientem a geração de dados, o desenvolvimento de coleções de dados, o armazenamento e análise e interpretação desses dados numa grande diversidade de contextos disciplinares.

Os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a compreender que estes dados, se devidamente tratados, preservados e gerenciados, podem constituir uma fonte inestimável de recursos informacionais. Os repositórios de dados se incorporam rapidamente à infraestrutura mundial de informação científica, e dessa forma os acervos de dados podem ser usados, reusados e compartilhados. Potencialmente, estes dados podem capacitar os pesquisadores a formularem novos tipos de indagações, hipóteses e a usarem métodos analíticos inovadores no estudo de questões críticas para a ciência e para a sociedade (MAYERMIK, 2012).

O acesso efetivo a dados de pesquisa, de uma forma responsável e eficiente, consubstanciado por tecnologias de informação e comunicação, se torna uma condição crítica para as políticas nacionais de ciência e tecnologia. O Relatório da Organização para Cooperação e Desenvolvimento Econômico - OCDE (2007) enfatiza essa condição, alinhando, entre tantas outras possibilidades, algumas situações em que os dados de pesquisa se tornam um fator imprescindível: na cadeia de inovação, na cooperação internacional, na promoção de novas pesquisas e testes de hipóteses novas ou alternativas, na diversidade de estudos e opiniões; na formação de novos pesquisadores, na exploração de tópicos não idealizados originalmente, na geração de novos conjuntos de dados a partir de dados de múltiplas fontes e, sobretudo, na promoção de uma atividade científica mais aberta e mais transparente, que tenha como princípio produzir conhecimento publicamente acessível.

Neste cenário de rápidas transições, estão sendo criadas infraestruturas nacionais e regionais que permitem que sejam exploradas as potencialidades dos dados de pesquisa. Isto decorre da constatação de que as coleções de dados de pesquisa em formatos digitais só podem ser gerenciadas e preservadas para acesso, reuso e compartilhamento, se estiverem apoiadas por compromissos institucionais de longo prazo plenamente

sustentáveis. Alia-se a isto, a preocupação das agências financiadoras em garantir o retorno dos investimentos públicos, não só do ponto de vista econômico, mas principalmente do ponto de vista da qualidade dos resultados das pesquisas; soma-se agora a preocupação com a capacidade de reuso dos dados em outros domínios disciplinares diferentes daqueles para os quais eles foram originalmente gerados.

Compreende-se, portanto, que há um reordenamento nos processos científicos, trazido pela gestão e compartilhamento de dados. A prática de boa gestão dos dados abre a possibilidade de verificação confiável dos resultados e permite pesquisas transversais e inovadoras desenvolvidas sobre informações já existentes, encurtando o ciclo clássico de comunicação científica e abrindo novas formas de interlocução e de otimização de recursos financeiros.

Nessa direção, novos padrões, procedimentos técnicos e gerenciais estão surgindo; esquemas de representação e de arquivamento, recuperação e disseminação estão sendo reconfigurados, tendo como pano de fundo o movimento de acesso livre às informações científicas, que estende rapidamente a sua pauta de interesse para as questões de acesso aos dados de pesquisa, enquanto parte essencial da memória científica mundial.

Pelo lado mais pragmático e operacional, um conjunto de atividades gerenciais, técnicas e informacionais fortemente padronizadas - chamado coletivamente de curadoria de dados de pesquisa -, permite que os dados possam ser tratados, arquivados em ambientes digitais confiáveis, preservados e reconfigurados de forma que possam ser aplicados em novos contextos científicos; sirvam de base para novas pesquisas; sejam aproveitados para fins educacionais; e, sobretudo, colaborem para minimizar a duplicação de esforços nas estratégias de criação de dados.

Enquanto no seio da Comunidade Europeia e nos Estados Unidos frutificam os empreendimentos que sustentam infraestruturas organicamente integradas que dão suporte aos processos de curadoria digital de dados de pesquisa, no Brasil, ainda são poucas e fragmentadas as ações em torno do tema, agravadas, ainda, pela incompreensão de suas potencialidades e pela falta de visão de futuro. No intuito de contribuir para a reflexão sobre um modelo de curadoria digital para o país, o presente estudo alinha rapidamente alguns elementos necessários para a construção de uma política nacional que se enquadre nos princípios universais da informação aberta e do livre intercâmbio de ideias, informação e conhecimento.

2 DADOS DE PESQUISA

> Os dados que coletamos hoje podem ser usados no futuro de forma que ainda não conseguimos imaginar. Os exploradores de antigamente que coletavam espécimes de plantas e animais não sabiam nada sobre DNA e hoje as amostras são submetidas a esse tipo de investigação. Quando você coleta os seus dados, reúne informações que, no futuro, poderão ser analisadas de formas muito diferentes. São coisas que terão um valor enorme para cientistas que nem nasceram. (POLIAKOFF, 2013, p. 1).

"Dados são fatos, números, letras e símbolos que descrevem um objeto, uma ideia, condição, situação ou outros fatores" (NATIONAL RESEARCH COUNCIL, 1999, p.15). A definição registrada pelo Relatório do NRC ainda em 1999, sugere a complexidade do conceito e a diversidade de formas que os dados podem tomar. Enfocando mais especificamente em "dados de pesquisa" o Relatório da OCDE, citado anteriormente, descreve o termo como "registros factuais usados como fonte primária para a pesquisa científica e que são comumente aceitos pelos pesquisadores como necessários para validar os resultados do trabalho científico" (OECD, 2007, p.13).

O que se observa é que a noção de dados pode variar consideravelmente entre pesquisadores e ainda mais entre áreas do conhecimento. A constatação que os dados são gerados para diferentes propósitos, por diferentes comunidades acadêmicas e científicas e por meio de diferentes processos, intensifica ainda mais essa percepção de diversidade. Tipos de dados podem incluir, por exemplo, números, imagens, textos, vídeos, ou áudio, *software*, algoritmos, equações, animações, modelos, simulações. Alguns tipos de dados têm valor imediato e duradouro, enquanto outros adquirem valor ao longo do tempo; alguns dados são capturados num momento específico e irrecuperável, enquanto outros são fáceis de se recriar (BORGMAN, 2010).

Essa heterogeneidade intrínseca aos dados de pesquisa implica que é necessário formular políticas de amplo espectro, que não só identifiquem, mas efetivamente sustentem os vários tipos de dados e a sua natureza díspar. O reconhecimento dessa idiossincrasia torna-se crucial quando se estabelecem as opções gerenciais e tecnológicas para o arquivamento persistente e para a curadoria digital.

O *National Science Board* da *National Science Foundation* (NSF) adota uma lógica de categorização - que já se tornou clássica - que considera as seguintes características: a natureza dos dados, sua reprodutibilidade, o nível de processamento ao qual eles foram submetidos. Cada uma dessas diferenças tem implicações importantes na formulação das

políticas de gestão de dados digitais de pesquisa e na forma como eles devem ser arquivados e preservados.

Seguindo a categorização proposta pelo NSF, os dados podem ser distinguidos pela sua natureza ou origem em: observacionais, computacionais e experimentais. Os dados observacionais são dados obtidos de observações diretas, que podem ser associadas a lugares e tempo específicos, como por exemplo, a erupção de determinado vulcão numa data específica, a fotografia de uma supernova, o levantamento das atitudes de uma comunidade. Os dados observacionais – por sua natureza instantânea - guardam uma importância crítica que os qualificam como registros históricos que não podem ser coletados uma segunda vez e, portanto, devem ser arquivados para sempre.

Os dados computacionais, por sua vez, são resultado~~s~~ da execução de modelos computacionais ou de simulações, seja, por exemplo, no domínio da Física Nuclear ou para a criação de ambientes virtuais culturais ou educacionais. Para esta categoria de dados a preservação por longo prazo pode não ser necessária, posto que os dados podem ser replicados ao longo do tempo. Entretanto, replicar o modelo ou simulação no futuro pode exigir um grande número de informações que incluem descrição das dependências de *hardware*, *software* e outras dependências técnicas e ainda os dados de entrada. Portanto para salvaguardar a capacidade de apresentação ou renderização – para se usar um termo que se torna comum – dos dados computacionais se torna crítico o arquivamento de um conjunto rico de metadados, dados de entrada além do modelo propriamente dito. É preciso notar que algumas vezes é mais conveniente preservar somente os dados de saída.

Os dados experimentais, por sua vez, são provenientes de situações controladas em bancadas de laboratórios, como por exemplo, medidas de uma reação química. Em tese, dados experimentais provenientes de experimentos que podem ser precisamente reproduzidos não precisam ser armazenados indefinidamente; porém, nem sempre é possível reproduzir precisamente todas as condições experimentais, “particularmente onde algumas variáveis experimentais podem não ser conhecidas e quando o custo de reproduzir o experimento é proibitivo” (NATIONAL SCIENCE BOARD, 2005, p.19). Dessa forma, a política de preservação de longo prazo de dados experimentais passa obrigatoriamente por um balanço entre custo da curadoria dos dados versus o custo da reprodutibilidade.

É necessário considerar também os registros do governo, de negócios, da vida publica e privada, entre outros, como fontes de dados úteis para a pesquisa científica, seja qual for a natureza do seu objeto: tecnológico, social ou humano (BORGMAN, 2010).

Conforme nos esclarece ainda o relatório publicado pela NSF, os dados brutos frequentemente são submetidos a sucessivos estágios de refinamentos e análises que são moldados pelos objetivos da pesquisa; além do mais, os dados brutos podem também gerar uma sucessão de versões e linhagens. Como consequência desses processos, há necessidade de caracterizar outro tipo de dado de pesquisa: os dados derivados de dados brutos por processamento ou pelas próprias atividades de curadoria. Enquanto os dados brutos podem constituir a forma mais completa de informação, os dados derivados podem estar mais convenientemente utilizáveis por outros pesquisadores. Dessa forma, a "preservação de dados de pesquisa em formas múltiplas pode ser justificável em muitas circunstâncias" (NATIONAL SCIENCE BOARD, 2005, p.19), principalmente quando os dados são voltados para o reuso.

Por fim, é necessário observar que o processo experimental estabelece outra distinção importante para definição de políticas de curadoria: dados intermediários - obtidos no processo preliminar da pesquisa - e os dados finais. Tipicamente, os dados finais são selecionados para serem incluídos nas bases de dados, e, com muita frequência, os dados intermediários não são arquivados ou permanecem inacessíveis para outros pesquisadores. Porém, conforme enfatiza o Relatório da NSF, há uma crescente compreensão por parte de todos os envolvidos no problema de que os dados intermediários têm potencial de reuso por outros pesquisadores. Nesse ponto, surgem questões de custo e valor em relação a este tipo de dados de pesquisa.

3 FORMAÇÃO DE COLEÇÃO DE DADOS DIGITAIS DE PESQUISA

Coleção de dados pressupõe – no âmbito do presente estudo - não somente uma base de dados ou um grupo de bases de dados, mas inclui também toda a infraestrutura tecnológica, organizacional e de gestão, bem como as políticas de sustentabilidade e os atores envolvidos. A compreensão dessa visão mais ampla é determinante para a formulação de políticas de formação de coleções de dados de pesquisa e para o investimento que se faz sobre elas, e nos fornece finalmente indícios dos elementos que devem compor um modelo sustentável de curadoria digital de dados de pesquisa.

Ainda de acordo com o esquema conceitual estabelecido pelo Relatório do *National Science Board* da NSF, as coleções de dados recaem em três categorias funcionais. "Cada uma dessas categorias levanta questões exclusivas para quem define as políticas da área" e dimensionam a intensidade dos processos de curadoria.

a) Coleções de dados de pesquisa que são produtos de um ou mais projetos específicos de pesquisa e tipicamente contêm dados que são objetos de processamento ou curadoria limitados; as coleções podem ou não estar em conformidade com os padrões adotados pela comunidade, tais como formatos de arquivo, estruturas de metadados e políticas de acesso; com alguma frequência não existem padrões aplicáveis ou eles são ainda incipientes, uma vez que os dados são inovadores e a comunidade usuária reduzida; pode ser que não haja intenção de preservar a coleção além do final do projeto. Este tipo de coleção geralmente conta com financiamentos de baixo montante e é vinculado a projetos de pesquisas específicos que podem requerer – até por medidas legais - políticas de seleção e retenção por parte dos autores, como assinala Pérez-González (2010);
b) Recurso ou coleção de dados que servem a uma dada comunidade científica ou tecnológica e estão vinculados a uma disciplina. Esse tipo de coleção digital comumente estabelece padrões que podem ser derivados de padrões já existentes ou do desenvolvimento de novos padrões, quando os disponíveis são inadequados ou simplesmente inexistentes. Esta ação é liderada pela própria comunidade envolvida. Os orçamentos vinculados a estes recursos são geralmente de volume intermediário e são financiados diretamente por agências governamentais. Tendo em vista as flutuações de prioridade típicas das agências de fomento– que têm origens políticas, comerciais e estratégicas, entre outras – torna-se difícil antecipar por quanto tempo este tipo de recurso informacional poderá ser mantido;
c) As coleções de dados referenciais, por sua vez, são projetadas para servir um espectro amplo de usuários provenientes de uma multiplicidade de comunidades científicas e educacionais – pesquisadores, estudantes e educadores; seu alcance extrapola os limites institucionais e nacionais e têm, geralmente, um forte rebatimento global. É essencial que as coleções desse tipo estejam em

conformidade com padrões robustos, bem estabelecidos e abrangentes; nessa circunstância, a seleção de padrões para estes recursos frequentemente tem o efeito da criação de um padrão de aplicação universal. Os orçamentos para o financiamento destas coleções são bastante elevados, refletindo o escopo e a amplitude do impacto desses recursos para a ciência. A expectativa é que essas coleções sejam mantidas indefinidamente e para tal os fundos provêm de diferentes fontes de financiamento e têm aplicação direta e compromissos de longo prazo.

4 PRINCÍPIOS E DIRETRIZES PARA O ACESSO A DADOS DE PESQUISA

O relatório da OCDE, publicado em 2007, intitulado "Princípios e diretrizes para acesso a dados de pesquisa financiados por recursos públicos", estabelece uma matriz de recomendações dirigidas aos setores governamentais responsáveis pelas políticas nacionais de C&T e aos respectivos órgãos financiadores das atividades de pesquisa. O objetivo final desses Princípios e Recomendações é aprimorar a eficiência e efetividade do sistema global de ciência. "Eles não pretendem impedir o seu desenvolvimento com obrigações e regulamentações onerosas ou impor novos custos aos sistemas nacionais de ciência" (OCDE, 2007, p.13), enfatiza o próprio documento. O Relatório estabelece balizamentos importantes na formação de sistemas nacionais de gestão de dados, principalmente os financiados pelo governo, e, portanto, é tomado como uma das referências-chave no presente estudo.

Resumidamente os princípios são os seguintes:

- **Acesso aberto** – significando acesso em termos igualitários para a comunidade internacional de pesquisa ao custo mais baixo possível. O acesso livre aos dados de pesquisa financiados por recursos públicos deve ser fácil, no tempo certo, amigável e preferencialmente baseado em sistemas web;
- **Flexibilidade** – exige que haja uma compreensão sobre o caráter transitório e, muitas vezes, imprevisível das tecnologias de informação, das especificidades de cada domínio de pesquisa e da diversidade de sistemas de pesquisa, sistemas legais, regulatórios e culturais de cada país quando da implementação dos princípios e diretrizes preconizados pelo documento da OCDE;
- **Transparência** – está relacionado à visibilidade dos dados, implicando que informações sobre dados de pesquisa ou sobre organizações produtoras de

dados, documentação sobre dados e especificações sobre condições de uso desses recursos devem estar disponíveis em escala internacional de forma transparente, preferencialmente via Internet. A falta de visibilidade de informações sobre dados de pesquisa já existentes ou de coleções que serão disponibilizadas no futuro, coloca um sério obstáculo para o acesso a esses dados e a sua reutilização;

- **Conformidade legal** – enfatiza que os acordos de acesso a dados de pesquisa devem respeitar os direitos legais e os interesses legítimos de todos os envolvidos nos empreendimentos públicos de pesquisa;
- **Proteção da propriedade intelectual** – recomenda que os acordos de acesso a dados de pesquisa devem considerar a aplicabilidade de lei de *copyright* ou de outras leis relacionadas à propriedade intelectual que podem ser relevantes para as bases de dados custeadas com recursos públicos;
- **Responsabilidade formal** – acordos de acesso devem promover práticas institucionais explícitas e formais, tais como regras e regulação que digam respeito às responsabilidades das várias partes envolvidas nas atividades relativas aos dados de pesquisa;
- **Profissionalismo** – acordos institucionais para a gestão de dados de pesquisa devem estar baseados em normas profissionais relevantes e valores que façam parte de códigos de conduta das comunidades científicas envolvidas;
- **Interoperabilidade** – as questões relacionadas à interoperabilidade tecnológica e semântica devem ser enfaticamente consideradas como fatores que possibilitam o acesso interdisciplinar e o uso de dados de pesquisa em escala internacional;
- **Qualidade** – o valor é a utilidade de dados de pesquisa, que depende fortemente da qualidade com que eles são planejados, produzidos, tratados e arquivados. Gestores de dados e organizações que desenvolvem coleções de dados devem estar atentos ao compromisso de se manter aderentes às normas explicitas de qualidade;
- **Segurança** – particular atenção deve ser dirigida para o uso de técnicas, metodologias e instrumentos que garantam a integridade e a segurança de dados de pesquisa;

- **Eficiência** – um dos objetivos centrais de se promover o acesso e o compartilhamento de dados de pesquisa é aprimorar a efetividade global das pesquisas financiadas com recursos públicos, no intuito de evitar a cara e desnecessária duplicação de esforços na formação de coleção de dados;
- **Responsabilização/prestação de contas** – a execução dos acordos de acesso aos dados de pesquisa deve ser objeto de avaliações periódicas por parte de grupos de usuários, instituições responsáveis e agências de fomento à pesquisa;
- **Sustentabilidade** – na qualidade de elemento imprescindível da infraestrutura de pesquisa atual, devidas considerações devem ser dadas à sustentabilidade do acesso aos dados de pesquisa custeados por fundos públicos. Isto significa, principalmente, assumir responsabilidades administrativas para garantir acesso permanente aos dados considerados de valor persistente e que requerem retenção de longo prazo.

5 UMA PROPOSTA PARA O PAÍS

Não obstante as tecnologias de informação e comunicação terem se tornado elementos essenciais em todas as disciplinas científicas, é necessário considerar ainda que o progresso científico não depende unicamente de tecnologias. Políticas voltadas para a pesquisa, fóruns apropriados, legislação específica, fundos para financiamento, valores culturais, ou seja, um espectro multidimensional de fatores afeta profundamente na natureza de novas descobertas, a velocidade com que elas são desenvolvidas e sua capacidade de se tornarem acessíveis e utilizadas efetivamente (OCDE, 2007).

Considerando como pressupostos básicos que os estoques de informações digitais são elementos fundamentais para o desenvolvimento da ciência e tecnologia, para os processos de inovação, para a educação e a cultura e para os empreendimentos governamentais e privados; fica claro que o futuro desses domínios e processos dependerá, em doses variadas, da competência das instituições responsáveis em prover acesso persistente a estes estoques, e que a capacidade de exploração, reutilização e transversalidade disciplinar desses recursos informacionais dependerá da sofisticação de tratamento e de gestão pelas quais eles tiverem passado desde seu planejamento.

Resta inconcluso estabelecer que informação deverá ser preservada, quem é o responsável pela preservação, que infraestrutura deverá ser desenvolvida, que controles

sociais, éticos e legais deverão ser aplicados, e, finalmente, quem pagará por tudo isso. As decisões são urgentes, pois o acesso aos dados no futuro vai depender de como vamos equacionar todas as variáveis que se sobrepõem.

Deslocando o olhar para os dados digitais, há um consenso nítido entre gestores de C&T, pesquisadores e profissionais das áreas de ciência da informação e de tecnologia da informação de que coleções digitais de dados pesquisa – principalmente em vista de sua complexidade, diversidade e fragilidade intrínseca – só podem ser preservados e gerenciados ao longo do tempo para acesso e reuso por meio de compromissos sustentáveis e duradouros que se entrelaçam em várias instâncias.

Com o objetivo de contribuir para a construção de um modelo multidimensional de curadoria digital de dados de pesquisa para o país, o presente estudo alinha as várias dimensões que devem dialogar para a composição de serviços sustentáveis de curadoria digital, de amplo alcance e cujas ações se desenrolem no ambiente de e-pesquisa. Nessa direção, são consideradas as seguintes instâncias: aspectos políticos, infraestrutura organizacional, desenvolvimento de coleções de dados, pesquisa, infraestrutura tecnológica e de padronização, formação de recursos humanos, sustentabilidade econômica, implicações sociais, legais e éticas e disponibilização de serviços.

A guisa de metodologia foram analisados três relatórios considerados primordiais no endereçamento do problema de curadoria de dados de pesquisa em âmbito nacional e internacional. São eles: **Long-lived digital data collections**: **enabling research and education in the 21st Century,** publicado pela agência norte-americana *National Science Foundation* (NSF) em 2005; **OECD Principles and guidelines for access to research data from public funding,** publicado pela *Organization for Economic Co-operation and Development* (OCDE) em 2007 e **A comparative study of the international approaches to enabling the sharing of research data**, publicado em 2008 pela *Digital Data Curation* (DCC). Complementarmente foram analisados trabalhos de autores que se destacaram na análise multidimensional do problema de curadoria digital de dados de pesquisa.

5.1 INSTÂNCIA POLÍTICA

Nos últimos anos, agências de financiamento de pesquisas de vários países e de alguns organismos internacionais vêm introduzindo a exigência de que a gestão de dados de pesquisa e um plano de compartilhamento de dados façam, obrigatoriamente, parte da

solicitação de auxílio para os projetos de pesquisa. Ações dessa natureza traduzem o reconhecimento, por parte dos formuladores de políticas de C&T, de que a preservação de dados de pesquisa traz benefícios perceptíveis para a sociedade. A partir dessa constatação é necessário, portanto, o estabelecimento de linhas de ações que assegurem a organização e a governança apropriadas para a atividade de preservação desses estoques informacionais. Além do mais é imprescindível garantir um fluxo contínuo de recursos destinados à sobrevivência por longo prazo das atividades de curadoria digital.

O que se observa é que a lacuna provocada pela inexistência de políticas coerentes, acessíveis e transparentes de arquivamento e acesso a dados de pesquisa revelam-se como barreiras para a pesquisa interdisciplinar e para a gestão efetiva de coleções de dados. Por outro lado, um esforço significativo está sendo dirigido, em escala mundial, no desenvolvimento de políticas e diretrizes que ordenem a gestão de dados de pesquisa. Estas iniciativas são levadas a cabo por um amplo espectro de instituições: pelos órgãos nacionais ligados à gestão de C&T, pelas agências de fomento à pesquisa, pelas instituições de pesquisa individualmente, como universidades e centros de pesquisa, e por outros parceiros-chave da comunidade internacional, como os organismos de padronização e organizações ligados ao movimento de livre acesso.

Mas um modelo neutro é um desafio inalcançável no contexto atual. Segundo o relatório do *Digital Data Curation* de autoria de Ruusalepp (2008), a ausência de um modelo universal voltado para o compartilhamento de dados de pesquisa é um desdobramento direto dos diferentes modelos de financiamento praticados pelos países individualmente. Os fóruns responsáveis pela formulação das políticas de gestão de dados devem atentar para um fato determinante destacado pelo relatório: “Por causa das diferenças na gestão, práticas e usos de coleções de dados em diferentes domínios da pesquisa, as políticas nacionais devem permanecem num patamar suficientemente geral para poderem ser efetivamente úteis na prática”.

A gestão para o acesso e reuso de coleções de dados de pesquisa, portanto, demanda uma infraestrutura de muitas faces, com muitos atores e costurada por compromissos políticos e financeiros duradouros. Subjacente a essa estrutura é necessário o desenvolvimento de um conjunto amplo de ações políticas de abrangência nacional, que estejam, porém, em harmonia com as políticas praticadas pelas principais iniciativas internacionais - incluindo o princípio de livre acesso aos dados de pesquisa e de

transparência pública e que considere as prioridades, idiossincrasias e as políticas das comunidades científicas e acadêmicas.

O presente estudo sinaliza que a instância política de um modelo de gestão e compartilhamento de dados de pesquisa para o país deve incluir:

- Fóruns para definição de políticas que tenham a participação de: gestores de C&T, agências financiadoras de pesquisa (CNPq, CAPES, FAPs), geradores de dados de pesquisa (universidades, centros e institutos de pesquisa), organizações com tradição na área de preservação digital, como o Arquivo Nacional, órgãos responsáveis por aumentar os conteúdos de valor na Internet, como o Comitê Gestor da Internet, Sociedades científicas, etc.
- Diretrizes e recomendações sobre padrões e tecnologias para a criação e implantação de rede de repositórios digitais de dados de pesquisa que sejam federados e interoperáveis.
- Linhas de financiamento de pesquisa em áreas de interesse como: preservação e curadoria digital, repositórios digitais, visualização de dados, ambientes colaborativos, metadados etc.
- Documentos estabelecendo diretrizes e estratégias para o desenvolvimento de uma ciberinfraestrutura nacional voltada para o arquivamento, acesso e reuso de dados de pesquisa.
- Exigências para depósito, gestão e disseminação de dados de pesquisa de projetos financiados com verba pública.
- Enquadramento da gestão de dados de pesquisa como elemento essencial na formulação de políticas de ciência, tecnologia e inovação.

5.2 INSTÂNCIA ORGANIZACIONAL

Por muitos séculos as bibliotecas e outras instituições de patrimônio intelectual armazenaram continuamente informações para uso corrente e futuro. Este fato moldou a forma como estas instituições foram organizadas e gerenciadas. Hoje, como afirma Pérez-Gonzáles (2010), se consolidou uma transformação qualitativa e irreversível. “A criação digital, as novas formas de comunicação em rede e os modelos de consumo da informação digital implica que autores, editores e instituições de pesquisa tenham que enfrentar novas estratégias, politicas e de infraestrutura, que permitam novas formas de gestão”. Esse

desafio é mais contundente quando se pensa em dados de pesquisa, caracterizados pela sua condição heterogênea, dinâmica e distribuída.

A trajetória de desenvolvimento da pesquisa científica, nas condições que hoje se apresenta, faz crer que as instituições acadêmicas precisarão de algum nível de curadoria de dados de pesquisa, entretanto é irreal se pensar que cada instituição individualmente poderá estabelecer capacidade local e própria de curadoria digital. Erway e Lavoie (2012) sustentam que a necessidade por especialização em cada área do conhecimento e a necessidade de um largo espectro de conhecimentos técnico e práticas em curadoria, aliadas aos riscos que devem ser assumidos e ao atingimento de uma economia de escala torna insensata a opção de replicar uma vasta gama de serviços de curadoria, infraestrutura, expertise em cada instituição de pesquisa.

Por outro lado, a diversidade de empreendimentos científicos sugere que uma pluralidade de modelos institucionais e de abordagens de gestão de dados específicos são mais efetivos em atender às necessidades dos usuários (OCDE, 2007), assegurar a qualidade dos dados e a agregação de usuários; entretanto, é necessário observar que a especialização em disciplinas pode levar a uma indesejável compartimentalização que anula um dos benefícios esperados com a curadoria digital que é encorajar a pesquisa interdisciplinar e a interpretação de dados em diversos contextos.

A abordagem nacional adotada por alguns países pode ser viável, dependendo da escala adotada. Por exemplo, a implantação de uma rede interoperável de repositórios de dados de pesquisa pode ajudar na descoberta de coleções relevantes de dados para reuso que podem facilitar a pesquisa multidisciplinar (ERWAY; LAVOIE, 2012). Esta abordagem pode ser aliada, primariamente, a ações colaborativas baseadas na criação de grupos de especialistas em assuntos, que recorrem à expertise de um *pool* de especialistas em vários aspectos técnicos de curadoria de dados. O trabalho colaborativo entre especialistas em assunto e em curadoria digital pode assistir a uma coletividade ampla de pesquisadores depositantes de grandes áreas de conhecimento, como Astronomia, Ecologia, Ciências Sociais, Saúde Pública etc. em âmbito nacional.

5.3 INSTÂNCIA DE DESENVOLVIMENTO DE COLEÇÕES DE DADOS DE PESQUISA

As bibliotecas de pesquisa e os repositórios digitais têm como desafio do nosso tempo a tarefa monumental de coletar uma quantidade extraordinária de dados digitais

gerados pela pesquisa contemporânea. Entretanto, o chamado “dilúvio de dados” que caracteriza a *Big Science*, aliado com a complexidade e o alto custo dos processos de curadoria e de preservação de dados, vão exigir que as organizações de pesquisa estabeleçam prioridades sobre o que eles vão finalmente coletar, mesmo diante das dificuldades teóricas e práticas de se operacionalizar conceitos tais como “avaliação de informação”, “valor da informação” e “necessidade de informação”.

Palmer e seus colaboradores (2011, p.1) enfatizam que a definição dos critérios de seleção de dados de pesquisa “é, num certo sentido, o que os desenvolvedores de coleções nas bibliotecas de pesquisa e nos arquivos sempre fizeram”. Mais explicitamente: julgar que fontes de informação têm valor suficiente para as suas comunidades-alvo para que se justifiquem os investimentos em formação de coleção, arquivamento, curadoria e preservação.

O potencial informacional crescente dos dados digitais distribuídos em rede de computadores transforma a visão que caracterizava dados de pesquisa, ainda registrados em mídia impressa, como simples subproduto dos processos de pesquisa. Nesse contexto, os dados só eram considerados na sua configuração final e, via de regra, eram descartados quando os projetos eram concluídos. A tecnologia digital interfere intensamente nas bases dessa ótica de avaliação: muitos tipos de dados científicos devem ser vistos hoje como componentes fundamentais da infraestrutura de sistemas modernos de pesquisa, cujo valor é expandido pelo acesso amplo, pelo seu potencial de reuso e, dessa forma, podem ter um longo ciclo de vida. “O valor do dado aumenta com o seu uso”, enfatiza Uhlir (2010).

Sob este ponto de vista, se destaca como maior desafio, quando do estabelecimento de políticas de desenvolvimento de coleções de dados de pesquisa, a definição de métricas e de modelos de avaliação que determinem - ou, de certa forma, predigam – as possibilidades de reuso de um particular conjunto de dados (PALMER et al., 2011), embora considerando as incertezas decorrentes desta qualificação.

Por outro lado, as ações para aquisição e retenção de dados de pesquisa - dependendo da área de conhecimento, natureza, formato, complexidade desses recursos, para citar algumas características - vão demandar estratégias de formação de coleções, infraestruturas tecnológicas e gerenciais e investimentos em curadoria digital em escalas bastante distintas. Além do mais, os dados necessários para dar apoio a pesquisas mais ativas, em termos da intensidade de uso e de geração de dados, como por exemplo, em

Astronomia, exigem coberturas mais seletivas e estratégicas, serviços de preservação e acesso, e, sobretudo, garantia de qualidade e de integridade.

Profissionais das áreas de Biblioteconomia e Ciência da Informação, cujos critérios e princípios de desenvolvimento de coleções são orientados pela avaliação de necessidades de comunidades de usuários, podem efetivamente adaptar suas práticas para a formação de coleções para repositórios de dados.

Fica patente, portanto, a necessidade do desenvolvimento e implantação de modelos teóricos e práticos de avaliação e de desenvolvimento de coleções de dados de pesquisa que venham ao encontro dos objetivos globais de formação de uma rede transversal, robusta, funcional e interoperável, que apoie os desafios da pesquisa científica contemporânea (PALMER et al., 2011).

Porém, no desenvolvimento de coleções de dados de pesquisa, outros problemas se interpõem. Um dos mais relevantes é assegurar que os dados possam manter a sua capacidade de transmitir informação e conhecimento ao longo do tempo e do espaço.

Disponibilizar os dados Internet é apenas uma das etapas de um ciclo complexo, e que isoladamente não garante que os dados possam ser acessados, reusados, e, sobretudo, terem seus significados e estruturas recompostos agora e no futuro. Tendo em vista que os bits não falam por si próprios e não impressionam nossos sentidos, para que eles possam manter a sua capacidade de serem interpretados em domínios distintos, transversalmente, é necessário que eles estejam suficientemente organizados e documentados. Dessa forma, torna-se imprescindível que informações contextuais – semânticas e estruturais – acompanhem os dados digitais de forma que eles estejam autodescritos. Isto é efetivado por meio de modelos conceituais de informação, expressos na prática por esquemas de metadados, que documentam, por exemplo, os elementos semânticos, as partes dos objetos e suas relações, as dependências técnicas, a proveniência, a identificação persistente, as restrições e direitos associados aos dados, as possíveis intervenções sofridas e seus efeitos. Ou seja, os metadados devem registrar idealmente tudo que deve ser de interesse do usuário, incluindo modelos de dados, equipamentos especiais, especificação da instrumentação, linhagem dos dados e muito mais.

Os metadados cumprem um papel de ponte para o futuro nas estratégias de preservação; além do mais ajudam na presunção de integridade e autenticidade dos dados digitais de pesquisa. A qualidade e precisão dos esquemas de metadados adotados e o rigor

da sua aplicação ~~e~~ são de crucial importância na garantia de que as coleções de dados possam ser acessadas, usados e reutilizados interdisciplinarmente pelo tempo que for necessário.

5.4 INSTÂNCIA DE PESQUISA

A inserção dos conhecimentos de curadoria na agenda de pesquisa de áreas de conhecimento como ciência da informação e ciência da computação torna-se essencial para a geração de um corpo consolidado de conhecimento que possa ser rebatido em todas as áreas que lidam com intensidade com informações e dados digitais. A fragmentação da pesquisa em curadoria digital, que caracteriza a área de estudo nos pais, se dá pela necessidade que alguns domínios de conhecimento, como Medicina e Ecologia, têm em gerir seus dados e extrair significado e viabilizar o reuso. Porém, permanece a necessidade de pesquisas coordenadas e de se criar linhas de investigação interdisciplinares, incentivadas por programas de fomento com perspectiva integradora, que possam gerar conhecimentos teóricos e práticos comuns e também específicos.

Essas ações de pesquisa, nos seus desdobramentos práticos, podem criar as bases para a produção de materiais de referência para a gestão de dados de pesquisa, como manuais, *guidelines*, cursos, normas e padrões, que têm, finalmente, importância crítica para as instâncias tecnológicas e de padronização e de formação de recursos humanos.

Alguns tópicos de uma possível agenda de pesquisa mostram a diversidade e interdisciplinaridade do problema:

- Dispositivos tecnológicos de visualização e compartilhamento de dados de pesquisa;
- Modelos e técnicas para processamento inteligente e de descoberta de dados por meio de taxonomias e ontologias; integração com os padrões da web semântica e do *linked data*;
- Concepção de novos tipos de publicação acadêmica que considerem vinculações semânticas entre dados e *e-prints*; impactos dessas publicações na comunicação científica;
- Metodologias de gestão de coleções de dados de pesquisa;
- Modelos de custo na implantação de sistemas de curadoria de dados de pesquisa;
- Interoperabilidade e integração de repositórios de dados de pesquisa;

- Impactos éticos e legais, propriedade intelectual, acesso aberto a dados de pesquisa versus privacidade;
- Desenvolvimento de esquemas de metadados voltados para a curadoria de dados de pesquisa.

5.5 INSTÂNCIA DE INFRAESTRUTURA TECNOLÓGICA E DE PADRONIZAÇÃO

O armazenamento seguro, a recuperação e o acesso a coleções de dados de pesquisa, além da exploração desses recursos por meio de serviços de informação e de aplicações computacionais – como, por exemplo, mineração e visualização de dados – exigem um conjunto de tecnologias e de padrões apropriados provenientes, em maior escala, da Tecnologia da Informação (TI) e da Ciência da Informação (CI). De igual importância são as normas e padrões que permeiam as ações de preservação e de curadoria digital e os vários níveis de interoperabilidade entre repositórios de dados e informações de pesquisa. Normas, padrões e protocolos, além de *hardware*, *software* e infraestrutura de rede se tornam essenciais na composição de ambientes de alta tecnologia conhecidos como "ciberinfraestrutura", que tem como objetivo mais geral a integração de serviços e recursos distribuídos para arquivamento, acesso e visualização.

Compreende-se por ciberinfraestrutura, como nos esclarece Pérez-González (2010, p. 3), "uma nova forma de cultura científica que se sustenta em uma robusta infraestrutura tecnológica de alto nível". Os dispositivos oferecidos por essa infraestrutura dão apoio a mecanismos inéditos de colaboração, baseados no acesso a uma quantidade extraordinária de dados, recursos de informacionais interpretados e reutilizados por potentes ferramentas de observação, visualização e simulação. Uma ciberinfraestrutura "é um meio que permite acesso e circulação de conhecimento distribuído, em que colaboram e se comunicam diferentes comunidades e disciplinas, rompendo fronteiras culturais, geográficas e temporais", complementa Pérez-González.

Em torno desta questão cabem algumas ações práticas na direção da formulação de uma política de gestão de dados de pesquisa:

- Definição de um elenco de normas, padrões e protocolos de especificações abertas, de aceitação internacional;
- Estabelecimento de ambientes de ciberinfraestrutura de abrangência nacional;

- Integração das ações já em andamento por instituições brasileiras vocacionadas para o problema.

5.6 FORMAÇÃO DE RECURSOS HUMANOS

"Sustentabilidade humana é crítica para assegurar continuidade e consistência ao longo do tempo de serviços de curadoria de dados de pesquisa", afirmam Mayernik e seus colaboradores (2012, p.12). Isto nos indica que estruturas educacionais e de recompensa apropriadas são componentes necessários para a promoção das práticas de acesso e compartilhamento de dados. Essas considerações se aplicam a quem financia, produz, gerencia e usa dados de pesquisa (OCDE, 2007).

O problema de coletar, organizar, indexar, arquivar e disseminar grandes coleções de dados – embora não seja um problema novo – é amplificado de forma extraordinária no ambiente da eScience. Curadores de dados provenientes das bibliotecas especializadas, dos arquivos e de setores da tecnologia da informação são capazes de gerir, inserir nos sistemas e preservar coleções de dados de pesquisa, entretanto os especialistas em assunto é que serão capazes de fazer as análises necessárias à reinterpretação e reuso dessas coleções. Isso significa que é necessário compor equipes de curadoria que conjuguem dinamicamente expertises de natureza distinta.

A necessidade de profissionais de informação multidisciplinares, que conjuguem conhecimento de áreas científicas e de engenharias, com conhecimento de biblioteconomia, ciência da informação e informática, delineia uma nova classe profissional, chamada por alguns autores de "profissional de eScience" (STANTON, 2011) ou ainda "cientista de dados", cuja missão é resolver problemas de gestão de informação em larga escala para pesquisadores com o uso de ferramentas inovadoras.

Considerando a extrema variação dos dados, os ambientes mais efetivos de gestão de curadoria são aqueles que permitem uma troca dinâmica de expertise, práticas e conhecimentos entre membros da equipe. "O compartilhamento de expertise desempenha um papel central nas operações em curso e no desenvolvimento de qualquer solução em curadoria de dados" (MAYERNIK et al., 2012, p. 12). Nessa direção, profissionais sofisticados de gestão de dados permitem que pesquisadores pratiquem uma ciência melhor, e ainda tornem possível que os profissionais de tecnologia da informação criem infraestruturas mais confiáveis, mais produtivas e mais eficazes, criando uma ponte entre os vários domínios. A

capacidade de traduzir as necessidades de informação do cientista em ferramentas da ciberinfraestrutura torna-se uma função essencial no fluxo gerido por este novo profissional de informação (STANTON, 2011, p. 91).

Como não há capacitação formal nessa área, os profissionais de gestão de dados terão que construir seus conhecimentos, ao longo do tempo, no trabalho cotidiano de curadoria e de articulação com áreas finalísticas. Dessa forma, se tornarão capazes de oferecer treinamento para novos usuários e novos profissionais de curadoria. Entretanto, é necessário estabelecer meios para acumulação, sistematização e disseminação desses novos conhecimentos, e também uma articulação direta com as instâncias preocupadas com a pesquisa na área de curadoria e preservação de dados. Parece bastante natural que as demandas por profissionais a eScience recebam acolhidas dos cursos tradicionalmente vocacionados para tal, como Biblioteconomia, Arquivologia e Ciência da Informação, ressaltando-se que a Ciência da Computação já deu passos importantes nessa área.

5.7 SUSTENTABILIDADE ECONÔMICA

Considerações sobre a persistência do acesso aos dados de pesquisa, na sua condição de elemento chave nas infraestruturas nacionais e internacionais de pesquisa – principalmente em relação aos dados financiados por recursos públicos - não podem ser encaradas como extensões ou algo acessório nos projetos e programas de pesquisa. A facilitação do acesso, a gestão e a preservação desses dados requerem planejamentos orçamentários específicos e um suporte financeiro apropriado. Essa constatação tem origem na própria natureza da curadoria digital que é um processo que se desenrola indefinidamente no tempo e no espaço; isto implica que o fluxo de fundos para a curadoria deve se compatibilizar com o ritmo dessa continuidade, o que parece óbvio, mas que na prática é frequentemente negligenciado.

Dessa forma, além da possível diversidade de arranjos dos vários atores envolvidos na pesquisa científica, persiste como condição crítica para um futuro de longo prazo para os dados de pesquisa, o reconhecimento de que a alocação contínua de recursos é um passo fundamental para os processos de curadoria. “Na ausência desse reconhecimento, o objetivo de manter por longo prazo o acesso a dados de pesquisa de qualidade não será alcançado” (ERWAY; LAVOIE, 2012, p.3).

Entretanto, assegurar a sustentabilidade econômica de conjunto de dados de pesquisa – e os serviços gerados a partir deles – ultrapassa a mera alocação de recursos. Na opinião de Erway e Lavoie (2012), o processo envolve a utilização eficiente destes recursos e a alavancagem de parcerias e colaboração no sentido de se alcançar uma economia de escala. Isto pode significar na prática que o estabelecimento de arranjos institucionais abrangentes e organicamente comprometidos sejam essenciais na sustentabilidade das coleções de dados de pesquisa de valor contínuo.

É necessário enfatizar ainda que modelos de custo sustentáveis para serviços de curadoria ou mesmo de preservação digital não são ainda bem entendidos, e não há na literatura da área formas e metodologias padronizadas para a condução dos processos de curadoria. Em termos mundiais, diferentes organizações adotam diferentes modelos financeiros.

Em termos práticos e mais imediatos, constata-se que o sucesso da implementação e operação de qualquer serviço de curadoria de dados digitais de pesquisa vai exigir uma análise minuciosa de todos os custos conhecidos e esperados para o futuro imediato, combinados com estratégias que assegurem a cobertura desses custos de forma contínua.

5.8 INSTÂNCIA SOCIAL, LEGAL E ÉTICA

Há um consenso nítido de que entre as principais barreiras sociais, éticas e legais interpostas entre as comunidades interessadas e o pleno acesso aos dados de pesquisa, está um quadro deficiente de proteção ao direito de propriedade intelectual, a dificuldade de documentar os dados para reuso e os problemas associados com a proteção da confidencialidade e privacidade. Há ainda uma tensão latente e não resolvida entre confidencialidade e abertura dos dados.

A legislação do país e os acordos internacionais, particularmente em áreas como direitos de propriedade intelectual e proteção da privacidade, afetam diretamente o acesso aos dados de pesquisa e as práticas de compartilhamento, e devem ser profundamente consideradas no projeto dos acordos de acesso de dados (OECD, 2007).

No ambiente acadêmico tipicamente não se reconhece completamente os diretos de propriedade intelectual relativos à produção e compartilhamento de dados. Faltam mecanismos de atribuição de crédito e de recompensa, de tal forma que o pesquisador que oferece abertamente seus dados para seus pares possa ser citado e reconhecido como autor

em qualquer situação e publicação que faça uso dos dados gerados por suas pesquisas. Na direção dessa demanda, o *Data Cite*[1] estabelece formas padronizadas de citação de dados e coleções de dados.

5.9 INSTÂNCIA DE SERVIÇOS

O acesso às coleções de dados de pesquisa, na forma de serviços convencionais e inovadores, dirigidos a segmentos variados de usuários, devem fazer parte das políticas de gestão de dados na qualidade de objetivo essencial. Além das facilidades tradicionais – como busca avançada, disseminação seletiva e *browsing* – os dados devem estar preparados para serem capturados por aplicações computacionais que proporcionem novas análises, estatísticas, indicadores e sirvam também de *input* para, por exemplo, sistemas de apoio à decisão e sistemas educacionais. É necessário ainda que as ciberinfraestruturas possam oferecer diferentes modalidades de interoperabilidade, como via OAI-PMH, OAI-ORE e *Linked data*. As interfaces para apresentação dos dados – preferencialmente via portais web - cumprem um papel importante na otimização do acesso, uso e reuso dos dados. Nessa direção as representações baseadas em tecnologias semânticas, taxonomias e ontologias tornam-se metodologias relevantes na descoberta de recursos.

O oferecimento de serviços baseados em coleções de dados de pesquisa amplia o escopo de atuação das bibliotecas de pesquisa e as recolocam no centro dos acontecimentos. Entretanto, esse novo papel impõe grandes desafios no delineamento de novos fluxos de trabalho e na implantação de infraestruturas tecnológica e gerencial para essas bibliotecas. Além do mais, um monitoramento regular se torna essencial, posto que novos conceitos de dispositivos informacionais para acesso e distribuição de informações de pesquisa estão permanentemente surgindo, um dos mais importantes atualmente é o CRIS – sigla para *Current Research Information System*. Um CRIS consiste basicamente num modelo de dados descrevendo objetos de interesse para as atividades de pesquisa e um conjunto de ferramentas para a gestão de dados. O objetivo do sistema é assistir o usuário em todos os processos de pesquisa, incluindo alocação de recursos, avaliação de projetos, identificação de novos mercados para produtos de pesquisa, análise de tendências e muito mais.

6 A GUISA DE CONCLUSÃO

[1] Disponível em: <www.datacite.org> Acesso em: 5 set. 2013.

Um dos maiores problemas da pesquisa científica atual não é, como se pode supor, a carência de dados e informações, mas o contrário, o seu excesso. O volume e a diversidade de dados desestruturados que o trabalho cientifico gera e tem que processar na procura de padrões, de comportamentos, ou seja, de novas descobertas, reordena de forma significativa os modelos de geração de conhecimento anteriores. O dilúvio de dados científicos – como é chamado pela comunidade científica – não é uma novidade, já estava nas previsões e se enquadra nos padrões mais genéricos caracterizados pelo fenômeno do "*big data*".

O pressuposto básico de dar significado prático e teórico a esse turbilhão de dados, na direção de novos desenvolvimentos - para as ciências da saúde, para o meio ambiente, para a compreensão do universo na sua imensidão cosmológica e no seu vazio infinitesimal - oferece um grande desafio para a própria ciência. É um enigma da esfinge se refletindo em si mesmo. A pesquisa multidisciplinar nessa área privilegia disciplinas como Ciência de Computação, Matemática, Ciência da Informação e também a Biblioteconomia, que por sua vez precisa realinhar os processos da biblioteca de pesquisa e definir as expertises e sistemas necessários para assumir as atividades de curadoria de dados de pesquisa.

Outro desafio que se interpõe entre a geração e o uso da informação é como tornar a ciência mais aberta, mais transparente, mais próxima de outros segmentos sociais, isto porque discutimos até aqui sobre o acesso aberto a dados e informações de pesquisa entre pesquisadores. Há, entretanto, uma demanda perceptível por dados científicos traduzidos e interpretados para segmentos "não científica" da sociedade: legisladores, formadores de opinião, políticos e o cidadão comum que procuram compreender os dilemas do nosso tempo, como as expectativas em torno das células-tronco e dos transgênicos e de outros problemas que mobilizam a opinião pública. Há ainda a discussão e os conflitos latentes sobre o acesso aos dados financiados por recursos públicos por parte das empresas privadas como fator de inovação e competitividade.

Mas colocar os dados disponíveis na web não significa que eles possam ser acessados, compartilhados, interpretados e reusados por todos os segmentos que os demandam. É preciso gerenciá-los de forma que eles possam estar arquivados de forma segura, sejam inteligíveis, possam ser avaliados quanto à qualidade, à veracidade e à integridade, e possam ser também retrabalhados para uso em outros contextos.

Nesse ponto, surge a necessidade de se criar infraestruturas de alcance nacional que ofereçam ambientes confiáveis e metodologias para captura, curadoria e disseminação adequada dos dados. Essas infraestruturas locais e nacionais devem se incorporar rapidamente aos sistemas globais de informação científica por meio de sistemas interoperáveis e abertos. Porém, muito além das ferramentas tecnológicas, é necessário também um conjunto de requisitos políticos, legais e éticos, econômicos e compromissos de longo prazo que se sobreponham sobre os ambientes tecnológicos. Nessa direção, o presente estudo pretendeu oferecer uma contribuição para o estabelecimento de uma política para o país de curadoria de dados de pesquisa.

Por fim é necessário reafirmar que o problema necessita de soluções imediatas, pois o acesso às coleções de dados de pesquisa no futuro vai depender da forma como equacionamos hoje todas as inúmeras faces e variáveis da curadoria e da preservação digital.

REFERÊNCIAS

BORGMAN, Cristine. Research data: who will share what, with whom, when, and why? In: CHINA-NORTH AMERICAN LIBRARY CONFERENCE, 5., 2010, Beijing. Disponível em: <http://works.bepress.com/borgman/238/>. Acesso em: 21 set. 2013.

ERWAY, Ricky; LAVOIE, Brian. **The economics of data integrity**. Ohio: OCLC, 2012. Disponível em: <http://www.webjunction.org/content/dam/research/publications/library/2012/erway-dataintegrity.pdf>. Acesso em: 21 set. 2013.

MAYERMIK, Matthew et al. The data conservancy instance infrastructure and organization service for research data curation. **D-Lib Magazine**, v. 18, n. 9/10, Sept./Oct. 2012.

NATIONAL RESEARCH COUNCIL. **A question of balance**: private rights and the public interest in scientific and technical databases. Washington, DC: National Academy Press, 1999. Disponível em: <http://www.nap.edu>. Acesso em: 19 maio 2013.

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: 1 fev. 2012.

OECD principles and guidelines for access to research data from public funding. Paris: Organization for Economic Co-operation and Development, 2007. Disponível em: <http://www.oecd.org/sti/sci-tech/38500813.pdf>. Acesso em: 20 set. 2013.

PALMER, Carole L.; WEBER, Nicholas M.; CRAGIN, Melissa M. The analytic potential of scientific data: understanding re-use value. **ASIST**, v. 9, n. 13, Oct. 2011.

PÉREZ-GONZÁLEZ, Lourdes. **Modelo/s de coste para la preservación de los datos científicos en la e-ciencia**. 2010. Disponível em: <http://eprints.rclis.org/8555/1/Perez.pdf>. Acesso em: 20 set. 2013.

POLIAKOFF, Martyn. [Depoimento]. Science as an open enterprise: open data for open science [Conferência]. São Paulo: Agência FAPESP, 28 fev. 2013. Participação em Conferência. In: JONES, Frances. **Editor-chefe da Nature fala sobre a abertura da ciência**. São Paulo: Agência FAPESP, 6 mar. 2013. Disponível em: <http://agencia.fapesp.br/16919>. Acesso em: 5 set. 2013.

RUUSALEPP, Raivo. **Infrastructure planning and data curation**: a comparative study of international approaches to enabling the sharing of research data. DCC Report comissioned by JISC, 2008. Disponível em: <http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/Data-SharingReport.pdf>. Acesso em: 21 set. 2013.

STANTON, Joffrey M. Education for eScience professionals: job analysis, curriculum guidance, and program consideration. **Journal of Education for Library and Information Science**, v.52, n.2, Apr. 2011.

UHLIR, Paul F. Information Gulags, intellectual straightjackets, and memory holes: three principles to guide the preservation of scientific data. **Data Science Journal**, v. 9, p. ES1-ES5, 2010. Disponível em: <https://www.jstage.jst.go.jp/article/dsj/9/0/9_Essay-001-Uhlir/_pdf>. Acesso em: 5 set. 2013.

XEXÉO, Geraldo. Big data: computação para uma sociedade conectada e digitalizada. **Ciência Hoje**, n.306, p.18-23 ago. 2013. Disponível em: <cienciahoje.uol.com.br/revista-ch/2013/306/pdf.../bigdata306.pdf/.../file>. Acesso em: 5 set. 2013.