RECIIS – Rev. Eletron. de Comun. Inf. Inov. Saúde. 2014 jun.; 8(2) – p.76-92
[www.reciis.icict.fiocruz.br] e-ISSN 1981-6278 | DOI:10.3395/reciis.v8i2.934.pt

Artigo original

Dados abertos de pesquisa: ampliando o conceito de acesso livre

Open research data: extending the concept of free access

Datos abiertos de investigación: ampliando el concepto de acceso libre

Luís Fernando Sayão[i]
Luana Farias Sales[ii]

RESUMO

O desenvolvimento de uma nova geração de experimentos, sensores, instrumentos e software de simulação faz com que a pesquisa científica contemporânea produza e utilize uma quantidade extraordinária de dados. Esse fato caracteriza o conceito emergente de e-Science, que oferece um conjunto de ferramentas tecnológicas para a coleta e análise de dados de pesquisa e possibilita que novos enfoques, aplicações, inovações e serviços sejam oferecidos pela ciência moderna. Porém, para que os dados sejam preservados eles precisam passar por processos de curadoria digital, cuja principal metodologia é atribuir a eles metadados estruturais e semânticos que garantam transmissão de conhecimento para o futuro. Por outro lado, no âmbito da ciência aberta, existe uma tendência mundial para dar acesso livre aos periódicos científicos e essa demanda se estende agora para o acesso livre e inteligível dos dados gerados pela pesquisa científica. Este trabalho discute brevemente a importância dos dados científicos abertos e os seus impactos nos atuais sistemas de informação para a pesquisa e, finalmente, propõe elementos para a composição de um modelo de curadoria digital de dados de pesquisa para o país.

Palavras-chave: Dados de Pesquisa; Curadoria Digital; e-Science; Ciência aberta, Análise de dados

ABSTRACT

The development of a new generation of experiments, sensors, instruments and simulation software makes the contemporary scientific research to produce and use an extraordinary amount of data. This fact characterizes the emerging concept of e-Science, which offers a set of technological tools for the collection and analysis of research data, and enables the modern science to offer new approaches, applications, innovations and services. However, for the data to be preserved they must pass through digital curation processes, whose main methodology is to assign to them structural and semantic metadata to ensure the transmission of knowledge for the future. On the other hand, in the context of open science, there is a global tendency to give free access to scientific journals and this demand extends now for open and intelligible access of data generated by scientific research. This paper briefly discusses the importance of open scientific data and its impacts for today's information systems research and, finally, proposes elements for building a digital curation model of research data for the country.

Keywords: Research Data; Digital Curation; e-Science; Open Science, Data Collection

RESUMEN

El desarrollo de una nueva generación de experimentos, sensores, instrumentos y software de simulación permite que la investigación científica contemporánea produzca y utilice una cantidad extraordinaria de datos. Este hecho caracteriza el concepto emergente de e-Science, que ofrece un conjunto de herramientas tecnológicas para la colecta y análisis de datos de

[i]Comissão Nacional de Energia Nuclear, Centro de Informações Nucleares, Rio de Janeiro, Brasil. sayao@cnen.gov.br
[ii]Comissão Nacional de Energia Nuclear, Instituto de Engenharia Nuclear, Rio de Janeiro, Brasil. sales@ien.gov.br

investigación y posibilita que nuevos enfoques, aplicaciones, innovaciones y servicios sean ofrecidos por la ciencia moderna. Sin embargo, para que los datos sean preservados necesitan pasar por procesos de tutoría digital, cuya principal metodología es atribuir a los metadados estructurales y semánticos que garanticen transmisión de conocimiento para el futuro. Por otro lado, en el ámbito de la ciencia abierta existe una tendencia mundial para dar el acceso libre e inteligible de los datos generados por la investigación científica. Este trabajo discute brevemente la importancia de los datos científicos abiertos y su impacto en los actuales sistemas de información para la investigación y, finalmente, propone elementos para la composición de un modelo de tutoría digital de datos de investigación para el país.

Palabras clave: Datos de Investigación; Tutoría Digital; e-Science; Ciencia abierta; Análisis de datos.

Submetido:30/Mar/2014
Aceito: 20/Mai/2014
Conflitos de interesse: Não há conflitos de interesse a declarar.
Fontes de financiamento: Não houve

Introdução

A Declaração de Berlin sobre o Acesso Aberto ao Conhecimento em Ciências e Humanidades[1], publicada em 2003, amplia as fronteiras do movimento de livre acesso ao explicitar o que se compreende por contribuições de acesso livre. O documento declara que essas contribuições incluem "resultados de pesquisas científicas originais, dados brutos [dados não processados] e metadados, fontes originais, representações digitais de materiais pictóricos e gráficos além de material acadêmico multimídia".

A expansão do conceito de acesso livre – um pilar de importância crítica para a prática de uma ciência mais aberta – não está circunscrita somente às publicações acadêmicas tradicionais, como são os artigos de periódicos; suas demandas avançam para outros conteúdos que incluem, de forma privilegiada, a disponibilização aberta e de forma inteligível de dados de pesquisa. Os pressupostos de uma pesquisa aberta incluem também na sua agenda de preocupações itens como nível de abertura das ferramentas, instrumentos, dispositivos laboratoriais, software e formatos usados em experimentos científicos, como fatores inibidores da interoperabilidade e do compartilhamento.

De uma maneira direta, essas demandas renovadas da comunidade científica estão localizadas no escopo da chamada "ciência aberta" cuja preocupação primordial é tornar a atividade de pesquisa mais transparente, mais colaborativa e mais eficiente. A ideia de ciência aberta tem muitas faces e muitos significados, porém, o mais eloquente deles é o que reconhece, primordialmente, que o conhecimento científico é um patrimônio da humanidade e, que, portanto, deve estar disponível livremente para que as pessoas – cientistas ou não – possam usá-lo, reusá-lo e distribuí-lo sem constrangimentos tecnológicos, econômicos, sociais ou legais.

O crescimento contínuo da quantidade de dados produzidos pelos diversos segmentos da sociedade – agências governamentais, instituições de pesquisa, indústria – confere a esses recursos a condição de componente fundamental para a pesquisa científica moderna e os identifica também como parte dos fenômenos relacionados ao chamado big data, tão comentado ultimamente. 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.

No contexto da pesquisa científica atual, há uma compreensão de que uma nova ordem se sobrepõe ao que se convencionou considerar como *output* dos processos de investigação científica. "Os editores [científicos] reconhecem que em muitas disciplinas os dados, em várias formas, são agora o principal produto de pesquisa" conforme enfatiza Murray-Rust[2]. Uma sequência genômica, a velocidade de partículas subatômicas, as respostas de levantamento social, a frequência de substantivos num corpus de textos, as imagens de satélites de outros planetas, todos

esses recursos informacionais são como dados de pesquisa. Todos esses dados, praticamente, podem ser descritos e armazenados em bases de dados e usados para propósitos de pesquisa[3]. Todos eles são resultados que precisam ser considerados como parte da infraestrutura mundial de informação científica.

Quando consideramos os dados de pesquisa, as condições e delineamentos preconizados pelos movimentos em prol da prática de uma ciência aberta se consolidam como consensual entre cientistas, organizações de fomento, editores científicos e outros atores envolvidos no mundo científico. Várias ações e movimentos cultivados no próprio seio das comunidades científicas partem do pressuposto de que esses estoques de informação configuram um recurso imprescindível para o avanço da ciência. O acesso aos dados de pesquisa torna-se, portanto, um imperativo para a ciência com reflexos globais, considerando que os pesquisadores trabalham em cooperação internacional e os dados são criados, compartilhados e acessados em escala planetária; mas que têm, entretanto, um rebatimento nos planos locais e nacionais, visto que esses mesmos pesquisadores estão, tipicamente, inseridos em estruturas organizacionais locais e submetidos a políticas de fomento a pesquisa de âmbito nacional[4].

Como um fenômeno do nosso tempo, entende-se que há um reordenamento nos processos científicos trazido pela gestão e compartilhamento de dados de pesquisa. A prática de boa gestão desses recursos abre a possibilidade de verificação confiável dos resultados dos experimentos e permite pesquisas transversais e inovadoras desenvolvidas sobre informações já existentes.Dessa forma, encurta o ciclo clássico de comunicação científica e abre novas formas de interlocução e de socialização no mundo científico, além de contribuir para a racionalização dos recursos financeiros públicos aplicados na pesquisa científica.

O presente estudo pretende analisar os diversos aspectos mais intensamente relacionados à gestão, preservação, compartilhamento e acesso aos dados de pesquisa no ambiente de pesquisa identificado como e-Science ou quarto paradigma, tendo como pano de fundo o papel desses recursos informacionais para a ciência aberta.

Nesta direção, o artigo se organiza da seguinte maneira: primeiramente, é feita uma contextualização do tema que traz à discussão o que vem a ser uma ciência conduzida por dados e qual a importância dos dados de pesquisa para o que vem sendo chamado atualmente de “ciência aberta”; em seguida, é colocado em pauta o conceito de dados de pesquisa – objeto principal deste artigo - , bem como a classificação de seus tipos, os processos, as técnicas e ferramentas que o envolve; em um terceiro momento, o artigo traz à tona os impactos deste novo ambiente orientado por dados na comunicação científica e as infraestruturas existentes para tratamentos desses dados, chegando finalmente à proposta dos elementos que devem ser considerados para a composição de um modelo de curadoria digital de dados de pesquisa para o país.

Uma ciência conduzida por dados

O reconhecimento do potencial informacional dos dados digitais, distribuídos em rede de computadores, para a ciência contemporânea transforma a visão que caracterizava dados de pesquisa, registrados em mídia impressa ou mesmo em formatos digitais, como simples subprodutos dos processos de pesquisa. Nesse contexto, os dados eram considerados somente na sua configuração final, sem considerar os seus ciclos de vida, versões e linhagens e, via de regra, eram descartados ou armazenados em mídias ou servidores sem a devida gestão quando os projetos eram concluídos. Quase sempre eram tragados silenciosamente pelo tempo: pela obsolescência tecnológica, pela efemeridade dos formatos e pela fragilidade das mídias digitais.

As tecnologias digitais aliadas aos tentáculos planetários das redes de computadores têm transformado de maneira vertiginosa a forma como os dados de pesquisa podem ser produzidos, disseminados, gerenciados, compartilhados e usados, tanto na ciência como em outros empreendimentos da sociedade, como nas esferas governamentais e nos negócios. Uma nova geração de sensores, instrumentos científicos avançados, software de simulação, colaboratórios, escalas mais precisas produzem em ritmo exponencial quantidades imensas e diversificadas de dados de pesquisa brutos ou não processados.

A relevância dos dados no contexto da "Big Science" como o da astronomia, da física e da biologia, somada aos mecanismos de colaboração em escala global, induziram não só ao surgimento de novos modelos de ciência – coletivamente chamados de "quarto paradigma científico" ou "e-Science" – mas possibilitaram a emergência de novos campos de estudo como a astroinformática e a bioinformática[5]. Existem, hoje, disciplinas científicas que são totalmente – em todos os seus ciclos – orientadas por dados. 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.

A tecnologia digital, como ferramenta fundamental da e-Science, interfere intensamente na forma como os dados se inserem nesses novos processos de geração de conhecimento: 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 aberto e pela ampliação – via processos de curadoria digital - do seu potencial de reuso. Dessa forma, as coletas de dados de pesquisa podem ter um longo ciclo de vida e se integrar aos sistemas tradicionais de informação para a pesquisa na forma de bases de dados armazenados em repositórios de dados e de vinculações às publicações acadêmicas tradicionais, como os artigos de periódicos e teses.

Esse fenômeno contemporâneo cria oportunidades sem precedentes para acelerar a pesquisa científica e gerar riquezas com base na exploração desse acúmulo de dados; abre a possibilidade de que a imensidão de dados gerados pela pesquisa científica contemporânea possa ser coletada, comparada e analisada, engendrando novos conhecimentos e novas questões de pesquisas. Ferramentas avançadas de software e de mineração de dados ajudam a interpretar e transformar os dados brutos em configurações ilimitadas de informação e conhecimento. Perguntas instigantes e recursivas colocadas perante os vários segmentos científicos podem agora ser endereçadas, pela combinação de múltiplas fontes de dados provenientes de domínios diferentes, através da aplicação de modelos complexos e de métodos inéditos de análise. "A capacidade dos cientistas de compartilhar e combinar importantes conjuntos de dados é o fundamento a partir do qual novos enfoques para resolução de problemas podem ser desenvolvidos"[6]. O compartilhamento e o intercâmbio permitem descobrir conexões no que estava antes desconectado, concluem Berman e seus colaboradores[6]. Dessa forma, como ressaltam Uhlir e Schröder[3], "a produção de conjuntos de dados constitui o primeiro estágio para aprimorar o conhecimento de partes da natureza e da sociedade, engendrando novas pesquisas e inovação".

Entretanto, uma vez que diferentes áreas científicas possuem padrões, práticas e políticas distintas em relação aos seus dados de pesquisa, torna-se essencial para o efetivo uso e reuso desses recursos o estabelecimento de infraestruturas técnicas, gerenciais e sociais que facilitem a integração dos conjuntos de dados de diferentes domínios e a criação de canais de colaboração entre as diferentes comunidades. Em muitas áreas, como a da física de partículas, ciberinfraestruturas baseadas na computação em grade – em que as tarefas estão divididas em várias máquinas - são implementadas para dar suporte tecnológico à colaboração global.

Os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a compreender que esses 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, esses dados podem capacitar os pesquisadores a formular novos tipos de indagações, hipóteses e a usar métodos analíticos inovadores no estudo de questões críticas para a ciência e para a sociedade[7].

A importância dos dados de pesquisa para uma ciência aberta

Assim como se debate hoje, fortemente, a questão do acesso livre aos periódicos acadêmicos, criando-se novos modelos de disseminação de resultados de pesquisas - mais ágeis e mais dinâmicos e organicamente mais próximos das comunidades científicas -, 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 pesquisas 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. Isto porque grande parte da ciência contemporânea é construída com base em dados digitais de pesquisas, num ciclo que inclui a sua coleta, análise, publicação, reanálise, crítica e reuso[8]. Os dados e conjuntos de dados de pesquisas providenciam as evidências necessárias para conferir veracidade, autenticidade e capacidade de reprodutibilidade ao corpo de conhecimento publicado nos periódicos, o que parece ser fundamental para o progresso científico. Portanto, quanto maior a capacidade dos sistemas de informação de oferecer dados de pesquisas livremente e que sejam tratados por metadados, de forma que possam ser interpretados e reutilizados pelo maior número possível de pesquisadores de diversas áreas, maior será o grau de transparência, de reprodutibilidade e de eficiência do processo de geração de conhecimento científico, e maior será a amplitude de aplicação dos projetos de pesquisa para a sociedade.

É perfeitamente compreensível que o acesso aberto inteligível à coleta de dados de pesquisas seja uma etapa crucial para os pressupostos de ciência aberta. Porém, as suas atribuições vão mais além do que somente a reprodutibilidade e a verificação do que está registrado na literatura acadêmica. O potencial cognitivo dos dados redesenha, através do reuso, os fluxos tradicionais de comunicação científica, estabelecendo novos padrões de socialização e de trabalho cooperativo independente de barreiras geográficas e disciplinares. O valor do dado de pesquisa está diretamente relacionado à possibilidade de uso e ao seu potencial de ser reinterpretado em outras áreas e contextos diferentes da que originalmente o gerou.

Uhlir e Schoröeder[3] vão mais adiante na análise dos benefícios científicos e socioeconômicos de uma ciência mais aberta, tendo como ponto central o papel dos dados de pesquisas financiadas por recursos públicos. Esses autores alinham algumas das muitas razões para o desenvolvimento de regimes de acesso mais abrangentes nas esferas institucionais, nacionais e internacionais, tendo o acesso livre como uma regra predominante.

- Reforça a pesquisa científica aberta;
- Incentiva a diversidade de análise e de opiniões;
- Promove novos tipos de pesquisa;
- Possibilita a aplicação de ferramentas automatizadas online de descoberta de conhecimento;
- Permite a verificação de resultados prévios;
- Torna possível o teste de hipóteses e de métodos novos ou alternativos de análise;
- Dá suporte a estudos sobre métodos de coleta de dados e de mensuração;
- Facilita a formação de novos pesquisadores;
- Possibilita a exploração, por outros pesquisadores, de tópicos não previstos pelos pesquisadores iniciais;
- Permite a criação de novos conjuntos de dados, de informações e de conhecimentos quando os dados de múltiplas fontes são combinados;
- Ajuda a transferir informação factual para países em desenvolvimento, promovendo a capacitação de pesquisadores nesses países;
- Promove a pesquisa interdisciplinar, intersetorial, interinstitucional e internacional;

Geralmente, ajuda a maximizar o potencial de pesquisa das novas tecnologias digitais e das redes de computadores, proporcionando um retorno maior para os investimentos públicos em pesquisa.

Nessa perspectiva, faz-se necessário compreender melhor o que é dado de pesquisa. A seção seguinte vai nessa direção.

Afinal, o que é dado de pesquisa?

O Relatório da OECD[iii] – sigla em inglês para Organização para a Cooperação e Desenvolvimento Econômico - descreve a expressão "dados de pesquisas" 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"[9]. Entretanto, o que se observa é que a amplitude do que se entende por dados de pesquisa sugere um conceito complexo que pode se manifestar numa multiplicidade de formas.

A noção de dados pode variar consideravelmente entre pesquisadores e, ainda mais, entre áreas do conhecimento. A constatação de 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, 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 passíveis de se reproduzir[5]. 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)[iv] adota uma lógica de categorização 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 pesquisas e na forma como eles devem ser arquivados e preservados.

Seguindo a categorização proposta pelo National Science Board[10], os dados podem ser distinguidos pela sua natureza ou origem em: observacionais, computacionais e experimentais.

- Dados observacionais – são obtidos por meio 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[v], o levantamento das atitudes de uma comunidade. Os dados observacionais – por sua natureza instantânea – guardam uma importância crítica que os qualifica como registros históricos que não podem ser coletados uma segunda vez e, portanto, devem ser submetidos a processos de curadoria que os preserve para sempre.
- Dados computacionais – são resultados da execução de modelos computacionais ou de simulações, seja, por exemplo, no domínio da física 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 a 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. É preciso notar que algumas vezes é mais conveniente preservar somente os dados de saída.
- Dados experimentais – 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 necessitam ser armazenados indefinidamente; porém, na prática, nem sem-

[iii] <http://www.oecd.org/>

[iv] <http://www.nsf.gov/nsb/>

[v] Estrela em uma fase na qual ela passa por explosões, aumentando sua luminosidade e, posteriormente, diminuindo-a aos poucos[11].

pre é possível reproduzir precisamente todas as condições experimentais, particularmente onde algumas variáveis experimentais não podem ser conhecidas e quando os custos de reprodução do experimento são proibitivos"[10].

É necessário considerar também os registros do governo, de negócios, da vida pública 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[5].

Como se observa nas definições da NSF, o potencial de replicação das pesquisas é um aspecto fundamental a ser considerado na tomada de decisão sobre como gerenciar os dados de pesquisa. A possibilidade ou não de se obter esses dados novamente ou ainda a possibilidade de reaproveitar esses dados para a realização de uma nova pesquisa agregam à curadoria de dados de pesquisa um valor ainda maior.

A partir daí, pode-se pressupor que o reuso e o compartilhamento de dados e informações, num ambiente de pesquisa caracterizado pela pluralidade de visão sobre esses recursos, abrem a possibilidade de se conceituar formas inéditas de agregações abstratas de produtos de pesquisa que sejam portadores de interpretações específicas, criando, dessa forma, novos constructos intelectuais que possuam os atributos mínimos dos recursos informacionais, ou seja, possam ser identificados e tenham sua autoria reconhecida. Esses novos constructos podem constituir formas de expressão que portem novas unidades de pensamento, conceitos, opiniões etc.

Assim, o reuso e a interpretação de dados de pesquisa em diferentes contextos é um desafio importante na área de curadoria digital de dados de pesquisas e para a e-Science que tem que lidar com os enigmas colocados pela grande quantidade de dados produzidos pelas disciplinas científicas que se enquadram no quarto paradigma, constituindo-se para ambas as áreas objetos essenciais de pesquisa. A subseção a seguir discutirá brevemente a questão do reuso de dados de pesquisa.

Reuso de dados de pesquisas

> "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 ainda nem nasceram" [12].

A ciência como um todo avança com maior qualidade, menor custo e mais eficiência quando abre a possibilidade para que o maior número possível de pesquisadores disponha de vias de acesso aos dados acumulados por seus antecessores e contemporâneos. Isso evita, objetivamente, o custo da duplicação de esforços e permite novas interpretações em diferentes contextos científicos para esses dados e, além do mais, permite que eles sejam integrados e retrabalhados de forma mais criativa, descortinando horizontes para novas pesquisas.

Nessa perspectiva, o conceito de "reuso" torna-se de fundamental importância para a ciência aberta, sendo compreendido de maneira ampla como o uso de dados - normalmente sem explícita permissão - para estudos, previstos ou não pelo autor original dos dados, por outros pesquisadores. O reuso inclui processos de agregação em base de dados, parâmetros em simulação e combinação de dados de diferentes fontes gerando novos *insights*[2].

Ainda segundo Murray-Rust[2], a prática de publicar e reusar dados de pesquisa varia enormemente entre diferentes disciplinas. Algumas áreas, como a de biociências, têm uma longa tradição de exigir que os dados sejam publicados e, a partir desse ponto, de agregá-los em bancos de dados financiados publicamente. As disciplinas classificadas como pertencentes à Big Science – como a astronomia e a física de partículas - já estabeleceram uma política bem consolidada de reuso de dados de pesquisas que torna mandatório que dados de telescópios, satélites, acelera-

dores de partículas, fontes de nêutrons, entre outros, sejam universalmente disponíveis para reuso. Para tal, oferecem ciberinfraestruturas sofisticadas para o compartilhamento dos dados gerados por seus aparatos.

Fica claro, então, que o reuso dos dados de pesquisas está sujeito à sua preservação e gestão, que podem ser feitas por meio das técnicas de curadoria digital de dados de pesquisas que serão comentadas a seguir.

Curadoria de dados de pesquisas

Disponibilizar as coletas de dados de pesquisas na Web é apenas uma das etapas de um ciclo complexo, e que isoladamente não garante que esses recursos possam ser acessados, reusados, e, sobretudo, ter seus significados e estruturas recompostos agora e no futuro. Nos processos de desenvolvimento de coletas de dados, muitos problemas técnicos e gerenciais se interpõem; porém, o mais relevante deles é assegurar que um conjunto de dados de pesquisas possa manter a sua capacidade de transmitir informação e conhecimento ao longo do tempo e do espaço e que, dessa forma, possa ser reusado enquanto persistir o seu valor informacional.

Entretanto, os bits - que compõem a maioria dos dados de pesquisa - não falam por si próprios e não impressionam nossos sentidos. Para que eles possam manter a sua capacidade de ser 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 os direitos associados aos dados, as possíveis intervenções sofridas e seus efeitos. Ou seja, os metadados devem registrar idealmente tudo o que deve ser de interesse do usuário, incluindo modelos de dados, equipamentos especiais, especificação da instrumentação, linhagem dos dados e muito mais. Essas informações têm um forte impacto na capacidade dos dados de transmitir conhecimentos e poder ser interpretados e reusados.

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.

Portanto, torna-se claro que, subjacentes às metodologias utilizadas pela curadoria digital, estão os processos de arquivamento digital e de preservação digital; porém, ela 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 informações e de conhecimentos[13]. As tecnologias e os modelos de gestão para a preservação de longo prazo, definidos pelos repositórios digitais confiáveis, cumprem um papel importante no âmbito da curadoria digital de dados de pesquisas.

O Data Curator Center (DCC)[vi]– cujo lema é “porque a boa pesquisa precisa de bons dados” - informa em uma página da Web que a curadoria digital “envolve a manutenção, a preservação e a agregação de valor a dados de pesquisas digitais 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 pesquisas, a curadoria reforça o valor de longo prazo dos dados existentes quando os torna disponíveis para a reutilização em novas pesquisas de qualidade.

Daisy Abbott[14] 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 pela definição de boas práticas na digitação, na seleção dos formatos e na documentação, de modo a se tornarem dispo-

[vi]<http://www.dcc.ac.uk/ >

níveis e adequados para serem descobertos e reusados 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 e vai além do controle do repositório que arquiva os recursos e envolve a atenção do criador do conteúdo e dos usuários futuros.

A importância dos dados para a ciência contemporânea está tornando a curadoria digital, o arquivamento persistente, a preservação digital e o estabelecimento de modelos de informação para a preservação de registros científicos questões-chave para as áreas de pesquisa. Para atender esta necessidade de estabelecimento de novos modelos de informação que aliem preservação de dados científicos digitais à disseminação e acesso aos registros, novos formatos de publicação vêm surgindo, como é o caso da publicação ampliada. Este novo modelo de publicação em que os dados são ligados aos resultados de pesquisas e publicados em um e-print tradicional é possibilitado pelas novas tecnologias de informação e comunicação (TICs) e merece atenção especial.

Publicações ampliadas: juntando dados e e-prints

Não obstante todas as transformações comportamentais e sociais decorrentes do aparato tecnológico que permeia 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 pesquisas, verifica-se que eles entregam ao usuário basicamente um artigo ou uma monografia. O que parece cada vez mais claro é que a heterogeneidade e a complexidade dos registros de resultados de pesquisas 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. Nessa direção, a Open Archive Initiative (OAI)[vii] define normas para descrição e intercâmbio de agregações de recursos da Web em seu projeto denominado Object Reuse and Exchange (OAI-ORE). Conforme explicitado numa página desse projeto na Web[22],

> Essas agregações, algumas vezes chamadas de objetos digitais compostos, podem combinar recursos distribuídos com múltiplos tipos de mídia, incluindo texto, imagens, dados e vídeo. O objetivo dessas normas é expor o rico conteúdo nessas agregações para aplicações que mantêm sistemas de autoria, depósito, intercâmbio, visualização, reuso e preservação[22].

As normas equacionam 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[15].

O Projeto DRIVER II – sigla para Digital Repository Infrastructure Vision for European Reearch[viii] - tem como alvo investigar as formas pelas quais a disponibilidade de dados de pesquisas podem ser usadas para enriquecer as publicações acadêmicas tradicionais. O documento abstrato que combina e-prints e dados de pesquisas - chamado de “publicação ampliada” - emerge da compreensão de que as publicações tradicionais são limitadas na sua capacidade de incorporar os resultados de todo o ciclo de geração de conhecimentos da ciência contemporânea. Isso acontece especialmente quando grandes conjuntos de dados são gerados. Nesse momento, fica evidente que os textos acadêmicos só podem apresentar os dados de pesquisas de forma condensada.

A valorização dos dados de pesquisas como recursos relevantes para uma ciência aberta tem reflexo na implantação de infraestruturas gerenciais e tecnológicas para o arquivamento desses dados. É um fato promissor observar que crescentemente os dados de pesquisas estão sendo armazenados em repositórios de dados confiáveis, onde, ge-

[vii] <http://www.openarchives.org/>
[viii] <http://www.driver-support.eu/>

renciados sob os princípios da curadoria digital são preservados e mantêm a sua capacidade de reuso. Entretanto, na atual infraestrutura de comunicação científica, esses conjuntos de dados não estão conectados às publicações onde são discutidos e analisados. A ideia que está por trás das publicações ampliadas é precisamente criar pontes que liguem os conteúdos dos repositórios institucionais, ou seja, que liguem publicações científicas e conteúdos dos repositórios de dados[16].

Dessa forma, a publicação ampliada é pensada como uma forma de objeto digital complexo que combina vários recursos heterogêneos que, porém, são relacionados. A base para esse tipo de objeto ainda é a publicação acadêmica tradicional, por exemplo, uma tese e o conjunto de dados que dá sustentação às suas análises e argumentações, somada também com os metadados necessários para manter a semântica, estrutura e gestão dessa nova publicação. Naturalmente, um artigo de periódico oferece uma "visão" dos significados e interpretação dos dados – e apresentações de congressos e trocas informais podem oferecer outras "visões" – mas o dado em si é cada vez mais um importante recurso para a comunidade científica e essa é a principal justificativa para acoplar ao artigo de periódico os dados da pesquisa que o embasa, conforme enfatiza Vehaar[16].

Nesta perspectiva, a publicação ampliada é uma ferramenta útil para a abertura e disseminação dos dados de pesquisas de forma integrada, garantindo aos dados sua significação original e a identificação de sua autoria. Além disso, ao unir os dados de uma pesquisa ao seu resultado final publicado em um e-print, a publicação ampliada permite a preservação da memória da pesquisa científica realizada, consentindo a replicação da pesquisa para fins de validação ou ainda para agregar valor a uma nova pesquisa. Sendo assim, a publicação ampliada pode ser considerada um veículo de comunicação científica de grande importância para a comunidade de pesquisadores.

É interessante observar que a abertura dos dados, sua curadoria, bem como os novos veículos de comunicação que utilizam a publicação ampliada podem impactar fortemente o processo de comunicação científica, alterando o seu ciclo uma vez que uma nova relação se estabelece entre os pesquisadores quando um pesquisador, para desenvolver seus projetos, passa a depositar toda a confiança nos dados levantados por outro, distante no tempo e no espaço. A próxima seção comentará esses impactos.

Comunicação científica num ambiente orientado por dados

De uma forma definitiva, a ciência orientada por dados e pelas tecnologias digitais criam 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, desenhando novos fluxos de cooperação e compartilhamento e definindo conceitos inéditos para a comunicação e para o registro científico.

Tomando como referência os princípios da curadoria digital, são inúmeras as reflexões que se podem fazer face aos impactos do reuso de dados de pesquisas, da publicação e da citação de coletas 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 e reusados sob novos propósitos.

No novo ambiente de pesquisa redesenhado pelas práticas da e-Science, 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 diálogo com o futuro, cria a possibilidade de se criar conceitos inovadores de documentos de registros de pesquisas, rompendo com o paradigma unidimensional e absoluto do artigo de periódico.

O acesso efetivo a dados de pesquisas, 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 pesquisas 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.

Nessa direção, infraestruturas para gerenciamento de dados de pesquisa vêm sendo criadas mundialmente com a finalidade de reunir, preservar, dar acesso e auxiliar os pesquisadores na gestão de seus dados de pesquisas. A seção a seguir apresenta um conceito de infraestrutura de tratamento de dados de pesquisas de origem europeia e o padrão adotado pela comunidade para tornar suas informações interoperáveis.

CRIS (Current research information systems e CERIF (Common European Research Information Format): infraestruturas de tratamento de dados de pesquisa

A crescente complexidade das atividades de pesquisa, a imensa geração de dados e informações e a necessidade de gerenciar processos propiciou o surgimento de infraestruturas tecnológicas com vistas ao tratamento e à recuperação dessas informações. Essas infraestruturas vêm sendo criadas não apenas para o armazenamento de dados, mas principalmente para gerenciar os processos e as etapas das atividades de pesquisa. Os benefícios são vistos não apenas pelos pesquisadores, mas pelos gestores, pelas agências de fomento, pelas empresas, bem como pelo público em geral. Essas infraestruturas permitem a contextualização das atividades científicas, otimizam os fluxos de trabalho, tornando a produção mais transparente, além de padronizá-las e permitir sua avaliação e reavaliação para o bom andamento das pesquisas, bem como para o reuso de dados e para a viabilização de novas descobertas.

Um exemplo de infraestrutura nesses moldes é o Current Research Information System (CRIS), que consiste em um modelo de dados que descreve um conjunto de objetos de interesse para as atividades de pesquisa e uma série de ferramentas que possibilitam ao usuário (pesquisador, gestor etc.) a gestão de seus dados de pesquisa em todos os processos, incluindo alocação de recursos, avaliação de projetos, identificação de novos mercados para produtos de pesquisa, análise de tendências entre outros serviços.

Em geral, o CRIS é construído para uma dada comunidade, como por exemplo, o United States Data Agriculture (USDACRIS[ix]), que fornece documentação e relatórios para as atividades agrícolas, ciência dos alimentos, nutrição humana, e silvicultura.

No entanto, a ideia do CRIS não é nova. Há aproximadamente 40 anos, diversos sistemas nos moldes do padrão CRIS vêm sendo desenvolvidos pelo mundo, muitas vezes com outros nomes, mas sempre como mecanismo de apoio à organização e recuperação de informações relevantes para a comunidade científica.

[ix]< http://cris.nifa.usda.gov/>

Normalmente, o CRIS tem informações sobre os projetos, pessoas, unidades organizacionais, programas de financiamento, resultados de pesquisas (produtos, patentes e publicações), instalações e equipamentos, e eventos, ou seja, todo tipo de informação que, de alguma forma, pode dar apoio às atividades de Pesquisa & Desenvolvimento (P&D) seja para um financiador, para uma instituição de pesquisa, para o pesquisador, para o público ou para os meios de comunicação.

São exemplos de informações constantes nos CRIS o currículo dos pesquisadores e suas páginas, portfólios de projetos de pesquisa, bibliografias, instituições com pesquisas correlatas, informações sobre oportunidades de inovação, informações sobre instalações e equipamentos, eventos etc.

O sucesso do CRIS, a riqueza informacional da Web e a proliferação de uma grande variedade de sistemas voltados para as comunidades científicas tornaram as informações para a pesquisa heterogêneas e distribuídas. Como consequência, a busca por esse tipo de informação transformou-se numa tarefa árdua para os usuários. Dito de outra maneira, a informação agora armazenada e tratada estava distribuída em sistemas diversos fazendo com que o usuário gastasse muito tempo navegando separadamente em cada um deles.

Lopatenko[17] mostra esse problema no seu artigo sobre recuperação de informações no CRIS. Segundo ele, normalmente pesquisadores ou gestores de informações em políticas de pesquisa não se limitam apenas à informação armazenada em um dos sistemas existentes. Ao contrário, informações de pesquisas em qualquer área da ciência e tecnologia estão espalhadas por uma variedade de sistemas de informações heterogêneos e, por isso, há uma forte necessidade de reunir todas as informações possíveis ou, de pelo menos, o sistema indicar onde essas informações podem ser encontradas. Lopatenko enfatiza a importância de saber se as informações reunidas na pesquisa são efetivas e completas. No entanto, segundo ele, pesquisas anteriores revelaram que a integração de dados de instituições de pesquisas não resolve o problema, especialmente se as instituições forem regidas por órgãos diferentes ou se não usufruírem de benefícios diretos de participação em tais redes de informações. Assim, o autor reafirma a necessidade de encontrar uma solução para o problema de integração dos dados, solução esta que será o compartilhamento de um padrão com três características essenciais: 1) fácil de implementar para qualquer participante, 2) flexível o suficiente para abraçar a diversidade, a estrutura e o significado dos dados em diferentes estados, organizações, ou áreas da ciência e 3) poderoso para fornecer serviços sofisticados de recuperação de informações. Para isso, sugere o uso de ontologia e de padrões sugeridos pelo W3C Consortium[x].

Nessa direção, a Comunidade Europeia criou o European CRIS (EUROCRIS)[xi], uma organização sem fins lucrativos, voltada para o desenvolvimento de sistemas de informações de pesquisas e para a interoperabilidade entre esses sistemas.

A ideia de fazer esses sistemas interoperarem é permitir que o usuário final possa acessar as informações, disponibilizadas no CRIS distribuídos e heterogêneos, bem como em repositórios, em um local único. Para isso, o EuroCRIS vem adotando uma série de estratégias, como: troca de experiência entre os membros em geral; criação do DRIS (diretório de CRIS); estudo e desenvolvimento de atividades conjuntas de P&D; conferência bienal sobre CRIS; reuniões semestrais com os membros, seminário estratégico anual, workshops, ligações com parceiros estratégicos, desenvolvimento de estratégia e infraestrutura, e o mais importante deles, o desenvolvimento do Common European Research Information Format (CERIF), um padrão recomendado aos estados-membros da comunidade europeia, inicialmente com a finalidade de facilitar o intercâmbio de informações entre bases de dados de projetos de pesquisa.

Criado em 1991, o CERIF, com o passar do tempo, precisou ser revisto e, assim, foi também estendido a outros tipos de informações, além daquelas dos projetos de pesquisa. Nessa direção, a versão CERIF2000 apresentou diretrizes para um modelo de dados CRIS mais completo e um núcleo base que permitia a troca de informações de ma-

[x] <http://www.w3.org/>
[xi] <http://www.eurocris.org/>

neira flexível, possibilitando que a maioria dos CRIS existentes pudessem manter suas características próprias, e ainda assim interoperar com os demais CRIS existentes na comunidade.

O CERIF2008 – última versão disponível – descreve um modelo de dados formal – que permite a interoperabilidade entre os sistemas de gestão da investigação, a partir de informações sobre pessoas, projetos, organizações, publicações, patentes, eventos, prêmios, equipamentos etc., um modelo de dados físico[18,19] e um formato de troca de dados em XML[19].

Além disso, de acordo com Ivanovic, Surla e Rackovic[20], o modelo de dados CERIF tem uma camada semântica que permite a classificação de entidades e suas relações de acordo com algum esquema de classificação. Outras "entidades" do modelo de dados CERIF estão ligadas à camada semântica através da "entidade" <cfClass>, que descreve o papel da pessoa na criação do resultado (autor da publicação, editor da publicação, presidente do conselho de eventos, gerente de projetos etc), a classificação do resultado da pessoa (ex: monografia, revista de papel etc), a classificação das publicações em que o resultado é publicado (ex: principal revista de importância internacional, revista nacional etc), a classificação do evento onde o resultado é apresentado (conferência de importância internacional, conferência de importância nacional etc) e a classificação do prêmio que é dado à pessoa (excelente prêmio internacional, prêmio internacional, prêmio nacional etc.).

Complementarmente, de acordo com a página mantida pelo grupo gestor, essa versão incluiu a recomendação de um tesauro multilíngue denominado Ortelius, que padronizou a indexação de assunto e os códigos utilizados para as áreas de atividades econômicas e produtos e ainda uma lista controlada de valores e atributos de determinados elementos (por exemplo, o papel de uma pessoa no projeto).

Em suma, a inovação apresentada pelo CERIF está na sua estrutura de dados formais, garantindo a integridade dos dados e evitando múltiplas instâncias dos mesmos valores de atributos; no uso de relações *n:n* permitindo declarar o papel e a duração temporal dos projetos; na preservação das características individuais de cada sistema e em sua essência multilíngue. Interessante observar que, assim como essa pesquisa, o modelo CERIF está preocupado não apenas em identificar as "entidades" a serem descritas, mas também as relações que elas possuem umas com as outras, o que propicia a formação de uma rede interligada de informações.

No Brasil, as iniciativas semelhantes ao CRIS são raras, e o que se encontrou mais próximo foi a Plataforma Lattes[xii]. Entretanto, o sistema CRIS conforme concebido na Europa considera não apenas informações sobre pessoas e instituições, como é o caso do Lattes, mas seu primeiro e principal objeto são os projetos de pesquisa e, mais recentemente, os dados não processados gerados por esses projetos, o que não se encontra em nenhuma das agências brasileiras de financiamento, que seriam as principais interessadas. O que se observa, portanto, é que no Brasil ainda não há um sistema avançado de gerenciamento, acesso e compartilhamento da produção científica nacional, como é o EuroCRIS.

Considerando que as atividades de pesquisa atuais geram grande quantidade de dados de pesquisa, e que esses dados devem ser preservados e compartilhados para novos usos e reusos - principalmente porque grande parte dessas atividades é financiada com verba pública e porque é preciso conferir agilidade ao desenvolvimento e à geração de novos resultados – verifica-se que o desafio está no estabelecimento de uma política nacional que possa ser apoiada pelas instituições de pesquisa. Assim, como fruto de investigações já realizadas nesta direção, a seção a seguir apresenta uma proposta de modelo de curadoria digital de dados de pesquisa para o Brasil.

Subsídios para um modelo de curadoria digital

Não obstante as tecnologias de informação e comunicação terem se tornado elementos essenciais para a grande maioria das 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 fi-

[xii] <http://lattes.cnpq.br/>

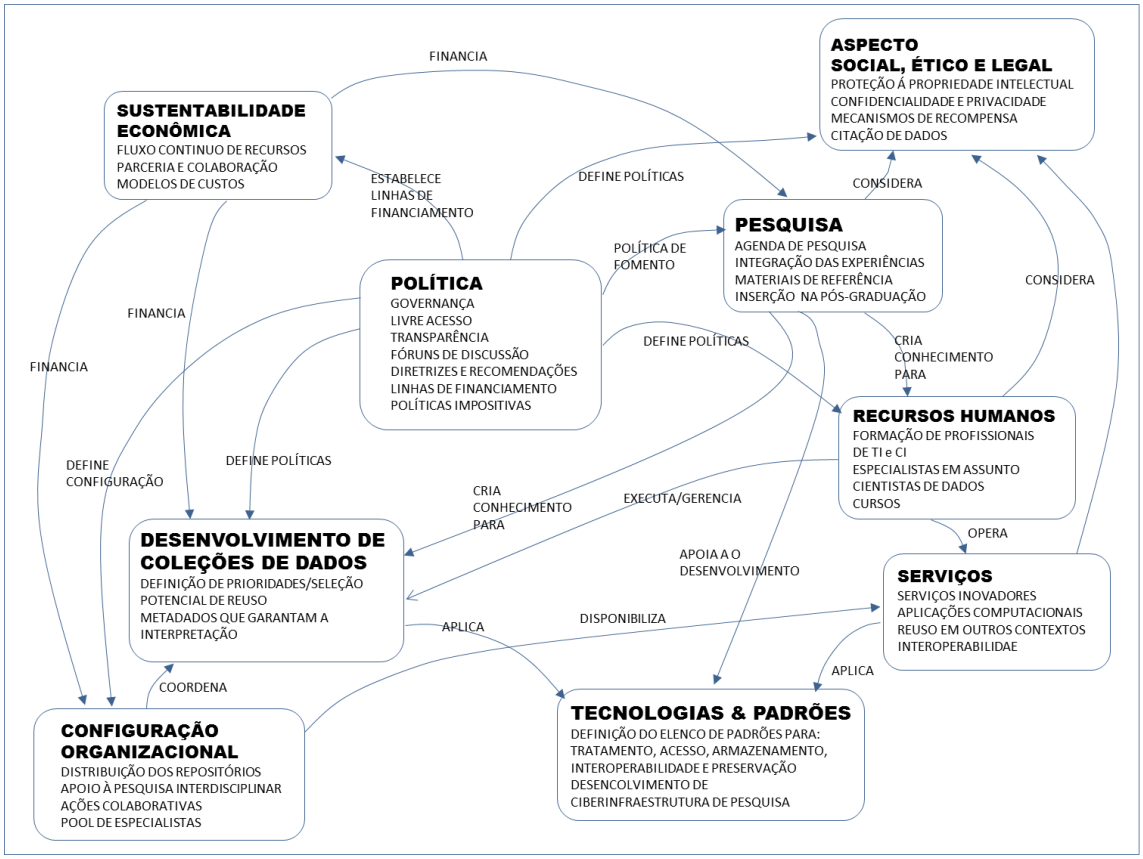
nanciamento, valores culturais, ou seja, um espectro multidimensional de fatores afeta profundamente a natureza de novas descobertas, a velocidade com que elas são desenvolvidas e a sua capacidade de se tornarem acessíveis e utilizadas efetivamente[9].

Em relação aos dados de pesquisas, 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 coletas de dados de pesquisas – principalmente tendo em vista a natureza complexa, heterogênea, distribuída e a fragilidade intrínseca desses artefatos digitais – só podem ser preservados e gerenciados ao longo do tempo, para acesso e reuso, por meio de compromissos sustentáveis e duradouros.

A gestão dinâmica de dados de pesquisas, voltada para uma ciência mais aberta, tem muitas faces e muitos atores, porém, nenhum deles isoladamente é capaz de garantir a capacidade dos dados transmitirem informação e conhecimento aos pesquisadores de hoje e do futuro. Portanto, um modelo de curadoria digital de dados de pesquisa de âmbito nacional deve alinhar as várias dimensões do problema e definir as interlocuções necessárias para a composição de serviços sustentáveis de curadoria digital de amplo alcance e cujas ações se desenrolem em ambientes de e-pesquisa.

Nessa direção, Sayão e Sales[21] propõem um modelo no qual são consideradas as seguintes instâncias: política, organizacional, desenvolvimento de coleções de dados, pesquisa, infraestrutura tecnológica e de padronização, formação de recursos humanos, sustentabilidade econômica, serviços e implicações sociais, legais e éticas. Essas instâncias são resumidas a seguir, e as relações entre elas são representadas na Figura 1.

Figura 1. Elementos para composição de um modelo de curadoria digital e suas relações



- Instância política – define políticas, diretrizes, recomendações e estratégias, além de financiamento contínuo, para o desenvolvimento de uma ciberinfraestrutura nacional voltada para o arquivamento, acesso e reuso de dados de pesquisa.
- Instância organizacional – estabelece as configurações organizacionais necessárias para a implantação de repositórios digitais de dados de pesquisa no país.
- Desenvolvimento de coleções de dados – cria os critérios de seleção e as métricas para a avaliação de qualidade, alcance e potencial de reuso dos dados, além dos parâmetros de tratamento técnicos, sobretudo em relação aos metadados, a que os dados devem ser submetidos.
- Instância de pesquisa – preocupa-se com a inserção dos conhecimentos de curadoria digital na agenda de pesquisa de áreas de conhecimento, como a de ciência da informação e ciência da computação, no sentido de se criar um corpo consolidado de conhecimento que possa ser rebatido em todas as áreas que lidam com intensidade com informações e dados digitais.
- Instância de infraestrutura tecnológica e de padronização – estabelece a ciberinfraestrutura necessária para o armazenamento seguro, a recuperação, o acesso a coleções de dados de pesquisas, o planejamento de serviços inovadores; estabelece também as normas e os protocolos que permeiam as ações de preservação e de curadoria digital e os vários níveis de interoperabilidade entre repositórios de dados e informações de pesquisa.
- Instância de formação de recursos humanos – trata da sustentabilidade humana crítica para assegurar continuidade e consistência, ao longo do tempo, de serviços de curadoria de dados de pesquisas, cujas considerações se aplicam a quem financia, produz, gerencia e usa dados de pesquisas.
- Instância de sustentabilidade econômica – define modelos que garantam a sustentabilidade econômica das estruturas de curadoria, posto que 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 ser compatibilizado com o ritmo dessa continuidade.
- Instância social, legal e ética – preocupa-se com as barreiras sociais, éticas e legais interpostas entre as comunidades interessadas e o pleno acesso aos dados de pesquisas, tendo em vista o 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.
- Instância de serviços – delineia o acesso às coletas de dados de pesquisas, na forma de serviços convencionais e inovadores, dirigidos a segmentos variados de usuários; além das facilidades tradicionais – como busca avançada, disseminação seletiva e *browsing* – os dados devem estar preparados para serem capturados por aplicações computacionais, como *data mining*, que proporcionem novas análises, estatísticas, indicadores e sirvam também de *input* para, por exemplo, sistemas de apoio à decisão e sistemas educacionais.

À guisa de conclusão

Parece não haver dúvidas de que o chamado dilúvio de dados que caracteriza o ambiente da e-Science terá um profundo efeito sobre a atual infraestrutura de pesquisa mundial. Esses efeitos já estão presentes nos novos ambientes de gestão de pesquisa, como o definido pelos padrões e recomendações CRIS e nas superestruturas de cooperação e compartilhamento providas pela computação em grade.

Os próprios sistemas de informação para a pesquisa terão que sofrer mudanças profundas em algumas dos seus fundamentos mais tradicionais, como é, por exemplo, o periódico científico, que entrega aos usuários-pesquisadores no final de projeto de pesquisa, um objeto textual impresso ou digital único que está longe de poder conter a riqueza, diversidade e complexidade dos reais produtos de pesquisa da ciência contemporânea.

Portanto, desafio que se interpõe para os profissionais de informação e de computação é integrar os sistemas e serviços de informação orientados para documentos, como são os catálogos online (OPACs) das bibliotecas de pesquisas e os repositórios institucionais e temáticos de hoje, com os sistemas de informação orientados para dados, como são, por exemplo, os repositórios de dados de pesquisas e os bancos de dados científicos.

Mas é muito importante que esse novo regime de informação definido pela e-Science, enquanto uma expressão da ciência aberta, possa se disseminar para todos os segmentos da sociedade. Posto que, na maior parte das vezes, a discussão que se instala sobre as questões críticas que permeiam as utopias de uma ciência aberta transcorrem todas em função da própria ciência e de sua execução transparente para os próprios pesquisadores.

Considerando essa reflexão, um novo desafio se coloca entre a geração e o uso da informação científica e que também tem desdobramentos sobre os sistemas de disseminação de informações: como tornar a informação científica mais aberta, mais transparente e mais próxima de outros segmentos sociais? É preciso estar claro que há uma demanda perceptível por dados científicos decodificados e recondicionados para os "não-pesquisadores": legisladores, formadores de opinião, políticos, professores e o cidadão comum, que precisam conhecer os enigmas científicos do nosso tempo, como as expectativas em torno da célula-tronco, dos alimentos transgênicos, e das mudanças climáticas e de outros problemas científicos que mobilizam a opinião pública, para tomar decisões, emitir sentenças, elaborar leis, transmitir para seus alunos, ou mesmo só para entender o que se faz nos laboratórios com o dinheiro público. Uma ciência aberta pode ser também uma ciência inteligível por todos.

Referências

1. 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.
2. Murray-Rust P. Open data in science. Serials Review 2008; 34(1): 52-64.
3. Uhlir P, Schröder P. Open Data for Global Science. Data Science Journal Jun 2007; 6 (Open Data Issue). Disponível em: <http://www.spatial.maine.edu/icfs/Uhlir-SchroederPaper.pdf>. Acesso em: 04 abril 2014.
4. Brase J, Farquhar A. Access to research data. D-Lib Magazine Jan/Feb 17(1/2). Disponível em: <http://www.dlib.org/dlib/january11/brase/01brase.html>. Acesso em: 30 mar. 2014.
5. Borgman C. Research data: who will share what, with whom, when, and why? In: China-North American Library Conference, 5.Beijing, 17 aug. 2010. Disponível em: <http://works.bepress.com/borgman/238/>. Acesso em: 21 mar. 2014.
6. Berman F, Wilkinson R, Wood J. Buiding Global Infrastructure for data sharing and exchange through the Research Data Alliance. D-Lib Magazine Jan/Feb 2014; 20 (1/2). Disponível em: <http://www.dlib.org/dlib/january14/01guest_editorial.html>. Acesso em: 04 abr. 2014.
7. Mayernik M et al. The data conservancy instance infrastructure and organization service for research data curation. D-Lib Magazine Sep/Oct 2012; 18(9/10). 2012. Disponível em: <http://www.dlib.org/dlib/september12/mayernik/09mayernik.html>. Acesso em: 01 fev. 2014.
8. Molloy J. The Open Knowledge Foundation: open data means better science. PLoS Biology Dec 2011; 9(12). Disponível em: <http://www.plosbiology.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pbio.1001195&representation=PDF>. Acesso em 01 fev. 2014.
9. 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: 31 mar. 2014.

10. 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. 2014.
11. Aulete C. Dicionário escolar da língua portuguesa. Rio de Janeiro: Lexicon; 2012.
12. Poliakoff M. [Depoimento]. In: Jones F. Editor-chefe da Nature fala sobre a abertura da ciência. Agência FAPESP, São Paulo, 06 mar. 2013. Disponível em: <http://agencia.fapesp.br/16919>. Acesso em: 01mar. 2014.
13. Lee C, Tibbo H. Digital curation and trusted respositories: steps toward success. Journal of Digital Information 2007; 8(2). Disponível em:<http://journals.tdl.org/jodi/index.php/jodi/article/view/229/18>. Acesso em: 20 mar. 2014.
14. Abbott D. What is digital curation? Edinburgh, UK : Digital Curation Centre, 2008. Disponível em: <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation>. Acesso em: 20 mar. 2014.
15. Lagoze C, Sompel HV. ORE user guide – primer. Open Archive Initiative, 2008. Disponível em: <http://www.openarchives.org/ore/1.0/primer.html>. Acesso em: 30 mar. 2014.
16. Verhaar P. 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 >
17. Lopatenko AS. Information retrieval in current research information systems. arXiv preprint cs/0110026, 2001. Disponível em: <http://arxiv.org/ftp/cs/papers/0110/0110026.pdf>. Acesso em: 30 mar. 2014.
18. Jorg B et al. CERIF 2008 - 1.0 Full Data Model (FDM) Introduction and Specification. 2009a. 43p. Disponível em: <http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/CERIF2008_1.0_FDM.pdf> Acesso em: 04 abr. 2014.
19. Jorg B et al. CERIF 2008—1.0 XML Data Exchange Format Specification. 33p. 2009b. Disponível em: <http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/CERIF2008_1.0_XML.pdf>. Acesso em: 16 fev. 2010.
20. Vanović D, Surla D, Racković M. A CERIF data model extension for evaluation and quantitative expression of scientific research results. Scientometrics 2011; 86(1): 155-172.
21. Sayão LF, Sales LF. Dados de Pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. Tendências da Pesquisa Brasileira em Ciência da Informação 2013; 6(1). Disponível em:<http://inseer.ibict.br/ancib/index.php/tpbci/article/view/102/146>. Acesso em: 04 abril 2014.
22. Open Archvives Initiative. Objective Reuse and Exchenge. Disponível em: < http://www.openarchives.org/ore>. Acesso em: 21 mar. 2014.