ISSN: 2237-826X
ENTREVISTA | INTERVIEW

Curadoria digital e dados de pesquisa

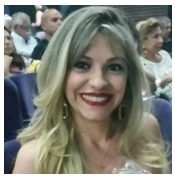
Digital curation and research data

Luís Fernando Sayão[1], Luana Farias Sales[1]

[1]*Comissão Nacional de Energia Nuclear - CNEN, Rio de Janeiro, RJ, Brasil*

Autor para correspondência/*Mail to:* Luana Ferreira Sales (lsales@ien.gov.br)

Luís Fernando Sayão possui graduação em Física pela Universidade Federal do Rio de Janeiro (1978), mestrado em Ciência da Informação pela Universidade Federal do Rio de Janeiro/Instituto Brasileiro de Informação em Ciência e Tecnologia (UFRJ/IBICT) e doutorado em Ciência da Informação pela UFRJ/IBICT (1994). Trabalha desde 1980 na Comissão Nacional de Energia Nuclear onde já exerceu os cargos de: chefe do Centro de Informações Nucleares (CIN); chefe da Divisão de Tecnologia da Informação; representante do Brasil no INIS - International Nuclear Information System (AIEA/ONU); coordenador-geral da RRIAN - Red Regional de Información en el Área Nuclear. É conselheiro do CONARQ - Conselho Nacional de Arquivos, membro do Câmara Técnica de Documentos Eletrônicos do CONARQ; docente do Programa de Pós-Graduação em Biblioteconomia da UNIRIO - Universidade Federal do Estado do Rio de Janeiro e do Programa de Pós-Graduação em Memória e Acervos da Fundação Casa de Rui Barbosa. Foi membro do Comitê Técnico-Científico do IBICT e da Comissão de Ensino da CNEN. É coautor do livro "Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores" e organizador/autor dos livros "Bibliotecas digitais: saberes e práticas" e "Implantação e gestão de repositórios institucionais". Tem como áreas de interesse: bibliotecas e arquivos digitais, publicações eletrônicas, interoperabilidade de sistemas de informação, acervos culturais digitais, curadoria de dados de pesquisa e preservação digital.



Luana Farias Sales possui Graduação em Biblioteconomia e Documentação pela Universidade Federal Fluminense (2003). Mestre em Ciência da Informação pelo convênio UFF/IBICT (2004-2006). Doutora em Ciência da Informação pelo Programa de Pós-Graduação do IBICT/UFRJ (2011-2014)., Atualmente é Analista em C & T da CNEN, atuando como pesquisadora do Instituto de Engenharia Nuclear na linha de Gestão do Conhecimento Nuclear e coordenando o Repositório Carpe dIEN : repositório de dados e informações em Energia Nuclear. Atua ainda como professora da Universidade Federal do Estado do Rio de Janeiro – UNIRIO, ministrando disciplinas relacionadas à Organização do Conhecimento. Tem experiência na área de Ciência da Informação, com ênfase em Organização e Representação do Conhecimento e Recuperação de Informação. Possui interesse em tópicos ligados à Comunicação Científica, Tecnologia de Informação e Gestão do Conhecimento, desenvolvendo pesquisas especificamente nas temáticas de e-Science, curadoria digital de dados de pesquisa, biblioteca digital, metadados, repositórios institucionais, repositórios de dados, Sistemas CRIS, objetos digitais e publicações ampliadas, sendo coautora de diversos artigos sobre estas temáticas e ainda do "Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores".

Resumo

Os Doutores Luis Fernando Sayão e Luana Farias Sales apresentam as condições que sustentam o desenvolvimento da área de curadoria digital e os emergentes repositórios de dados de pesquisa, assim como as controvérsias, desafios e responsabilidades éticas para os depositantes, gestores dos repositórios, usuários e profissionais da informação.

Palavras-chave: Dados de pesquisa; Curadoria digital; e-Science; Ciência aberta; Repositórios científicos.

Abstract

Luis Fernando Sayão, PhD and Luana Farias Sales, PhD present the conditions underpinning the development of the digital curation area and the emerging research data repositories, as well as the controversies, challenges and ethical responsibilities for depositants, repository managers, users and information professionals.

Keywords: *Research data; Digital curation; E-Science; Open science; Scientific repositories.*

1. Que elementos/condições caracterizam a emergência do desenvolvimento da área de curadoria digital?

A sociedade contemporânea – apoiada pelo poder pervasivo das tecnologias digitais e da *web* – produz e consome um volume extraordinário de informações em formatos digitais. Esses registros digitais são criados e aplicados em todo espectro social, mudando comportamento, negócios, formas de governar, de ensinar, inaugurando padrões inéditos de socialização e dando margem ao surgimento de novos fenômenos como é o *Big Data*.

Na esfera científica, as transformações são mais contundentes: o desenvolvimento de novos aparatos científicos, instrumentos, sensores, escalas e o uso intensivo de modelos de simulação da natureza geram uma quantidade imensa de dados, delineiam as fronteiras de um novo paradigma científico – conhecida como *eScience* -, criam novas metodologias de produção de conhecimento científico, formas de compartilhamento e de socialização entre os pesquisadores e alteram o fluxo tradicional de comunicação científica e de revisão por pares. Aliado a isso, os pressupostos da ciência aberta, que considera o conhecimento científico como um bem da humanidade, demandam que metodologias, instrumentos, softwares e dados estejam abertos para garantir o princípio da

reprodutibilidade e de autocorreção da ciência e a transparência de seus fluxos e de sua trajetória de erros e acertos.

Neste cenário de mudanças, os dados deixam de ser um mero subproduto da atividade de pesquisa e se tornam protagonistas na geração de novos conhecimentos e descobertas. Para tal, precisam estar disponíveis e interpretáveis para que possam transmitir conhecimento no tempo e no espaço e para que sejam reusados em diversos contextos alimentando a pesquisa interdisciplinar.

O uso e a geração intensiva de dados pelas atividades acadêmicas e de pesquisa criam a necessidade urgente de infraestruturas gerencias e tecnológicas que tratem de forma dinâmica o ciclo de vida dos dados – do seu planejamento até a seu arquivamento confiável – e insira esses ativos na infraestrutura mundial de informação para a pesquisa. Neste ponto que os procedimentos práticos e teóricos, que coletivamente chamamos de curadoria digital, se inserem como essencial para a captura, tratamento, preservação, arquivamento e acesso, pelo tempo que for preciso, dos dados de pesquisa considerados de valor continuo.

2. Caso existam, quais são as correntes/motivações contrárias ao depósito e compartilhamento de dados brutos associados aos manuscritos de pesquisa?

Muitos pesquisadores ainda estão presos na ideia de que os dados gerados por suas pesquisas são propriedades deles e não precisam ser compartilhados ou só precisam ser compartilhados de forma restrita entre seus colegas mais próximos. Porém, grande parte desses dados é produzida ou coletada com financiamento público e podem ser consideradas, portanto, um bem público. Nessa perspectiva, esses dados precisam que o seu potencial de reuso se realize em novas pesquisas e que se amplie seu alcance como inspirador de pesquisas interdisciplinares; aliados a isso é preciso também otimizar os recursos investidos na pesquisa pelas agências de fomento e minimizar a duplicação de esforços dispendidos em ciclos de geração de dados. Subjacente a essa discussão está o pressuposto fundamental de que a ciência avança mais rapidamente quando os seus estoques informacionais são compartilhados.

Porém, o padrão de compartilhamento varia enormemente no mundo heterogêneo e complexo da ciência. Há domínios científicos em que o compartilhamento de dados é determinante para o avanço da área e faz parte das suas metodologias, das suas formas de socialização e da sua cultura, como acontece, por exemplo, na Astronomia. Outros domínios disciplinares são tradicionalmente mais fechados e precisam de mudanças comportamentais e políticas mandatórias para disseminarem abertamente os seus dados.

Contudo, é preciso observar que uma parcela das coleções de dados é proveniente de pesquisas que precisam que seus produtos de pesquisa sejam arquivados e acessados de forma restrita. Isto acontece por vários motivos, por exemplo, devido à possibilidade de patentes, por interesses comerciais, por segurança ou por se tratarem de dados sensíveis que precisam de tratamentos específicos, como processos de anonimização. No entanto, as políticas de gestão de dados, preveem um intervalo de tempo para que esses dados possam ser utilizados de forma privilegiada por seus autores – conhecido como tempo de embargo – e dispõe de instrumentos para tratamento de dados que apresentem elementos mais críticos. Portanto, mesmo esses dados que não podem ser compartilhados em um primeiro momento precisam ser preservados e geridos para reuso futuro. A própria seleção do que vai ser ou não preservado e quando vai ser compartilhado já é uma das etapas do processo de curadoria digital de dados de pesquisa.

É preciso lembrar que um fator determinante para o crescimento da cultura de compartilhamento de dados é a formalização dos mecanismos de autoria e de recompensa por parte das instituições de pesquisa, agências de fomento e editores científicos. Os pesquisadores que geram, coletam e organizam as coleções de dados demandam que a autoria sobre esses dados seja identificada e reconhecida, e a partir daí possam ser citados, avaliados por pares e recompensados por seu trabalho, como são por um artigo de periódico.

3. No caso de um pesquisador desejar ter seus dados de pesquisa amplamente disponíveis e acessíveis a longo prazo, quais as alternativas disponíveis?

Existem várias opções, como disponibilizar no *website* do projeto ou da página pessoal, mas publicar as coleções de dados na *web* é só uma das etapas de um fluxo mais longo e complexo. Os dados para transmitirem conhecimento precisam de tratamentos específicos, de catalogação e de documentação que inclui metadados descritivos e disciplinares e documentos que garantam sua interpretação, como cadernos de laboratório, roteiro de entrevistas, manuais, dicionário de dados etc.; além do mais, os dados precisam ser identificados por meio de identificadores persistentes como o DOI (Digital Object Indentifier). Nessa direção, a melhor opção é depositar os dados em repositórios de dados, sejam institucionais, disciplinares, multidisciplinares, orientados para projetos ou os ligados aos periódicos científicos. Há centenas de repositórios com tecnologias, gestão, licenças, fluxos, submissão, custos, níveis de certificação e estratégias de preservação que variam muito. A maneira mais fácil de localizá-los e de identificar as suas características é consultando os diretórios de repositório como o re3data.org (www.re3data.org). Além do mais, os repositórios de dados dão maior visibilidade aos dados, o que implica em mais citações para estes e para os artigos correspondentes; permitem um maior grau de reuso e

compartilhamento; oferecem estruturas de preservação – como migração - e armazenamento seguro; oferecem ainda possibilidade de interação entre os pesquisadores, anotações e administração de licenças e de tempo de embargo. Outro fator de grande importância é o estabelecimento de *links* entre dados e artigos de periódicos que possibilita a formulação de novas concepções de publicações científicas e de um alto grau de contextualização para os dados, artigos e projetos.

4. O que é uma "publicação ampliada"/*enhanced publication*? Já existem boas práticas nesse sentido no Brasil? E, especificamente, no campo da Ciência da Informação?

Uma publicação ampliada é uma instância de um objeto digital complexo que se caracteriza pela vinculação de artigos aos dados que subsidiaram a sua criação, bem como à outras informações que auxiliem na contextualização da pesquisa, por exemplo, informações sobre os autores, sobre os equipamentos utilizados, projetos, etc. Essas vinculações podem ter valores semânticos e estarem associados à diversas ontologias. No Brasil, no campo da Ciência da Informação, a revista Informação & Tecnologia, do GT-8 da ANCIB, se configura como uma tentativa de publicação ampliada, no entanto, muitos artigos ainda aparecem no modelo tradicional porque os dados das pesquisas não foram cedidos pelos seus autores.

5. Quais os procedimentos éticos utilizados na criação, disponibilidade e armazenamento dos *datasets*, em relação a dados sensíveis (área de Saúde, segmento de negócios, por exemplo)? Qual o papel de cada ator neste sentido? (depositante, repositório, usuário, por exemplo)?

As infraestruturas de gestão de coleções de dados de pesquisa precisam estar ancoradas em uma política formal que defina todos os procedimentos do fluxo de curadoria: procedimentos práticos, formatos, metadados, arquivamento, preservação e segurança, licenças e, sobretudo, o tratamento que deve ser conferido ao compartilhamento de dados sensíveis, pessoais, confidenciais, de informações de interesse comercial ou que serão importantes nos processos de patenteamento, além das questões associadas à proteção de direitos de propriedade intelectual.

O pesquisador tem um papel importante nessas questões, pois é cada vez mais exigido pelas instituições de pesquisa e pelas agências de fomento que o próprio pesquisador formalize em um documento - conhecido como "Plano de Gestão de Dados de Pesquisa" - entre outras coisas, seu compromisso sobre como as questões éticas e de privacidade e os procedimentos para garantir a segurança dos dados, especialmente dos dados sensíveis, na fase de compartilhamento serão endereçados. Por exemplo: as estratégias de anonimização e de criptografia, no caso de transmissões por rede dos *datasets* das áreas de saúde que identificam pacientes; e o cuidado com o compartilhamento de dados que envolvam a localização de espécimes em riscos e de habitats sensíveis.

Pelo lado do usuário, os termos de uso e as licenças associadas aos dados – ambas estabelecidas ou adotadas pelas políticas de cada repositório – definem as responsabilidades e os limites de uso que os usuários devem respeitar, incluindo nesse escopo o reconhecimento da autoria dos dados por meio de citação padronizada.

6. Sob sua perspectiva, em que aspectos as discussões sobre curadoria digital - notadamente orientadas para a comunicação científica - extrapolam este ambiente e se dirigem ao "mercado" (por exemplo, a já reconhecida importância da inserção do cientista de dados em *fintechs* (ex: Nubank) e alguns meios de comunicação (ex.: NexoJornal)?

O conceito de reuso é determinante nessa questão. A possibilidade de as coleções de dados de pesquisa serem analisadas e reinterpretadas em outros contextos fora dos limites do domínio científico, ou de suas metodologias serem usadas em empreendimentos comerciais ou culturais e artísticos ou ainda em processos de inovação é plenamente viável. Isto é especialmente importante e tecnologicamente factível quando caminhamos de uma *web* de documentos para uma *web* de dados, em que as tecnologias digitais se associam às tecnologias semânticas e criam ambientes de informação mais inteligentes, contextualizados e convergentes, baseados em sistemas de informação interoperáveis

No âmbito de dados culturais, a ideia de reuso e reinterpretação de conteúdos culturais digitais começa a se ampliar e se tornar também um novo nicho de negócios para a indústria de conteúdos.

Por exemplo, a Europeana, apoiada na ontologia Europeana Data Model (EDM), disponibiliza os recursos que ela agrega, para reuso em diversos setores. O Projeto Europeana Space (http://www.europeana-space.eu), cujo lema é "um espaço de possibilidades para o reuso criativo de conteúdo cultural" ilustra bem esse novo conceito de reinterpretação de informações culturais. O objetivo do Projeto – conforme informa seu *website* - é a criação de novas oportunidades de emprego e de crescimento econômico no setor das indústrias criativas europeias com base nos recursos culturais digitais. Como resultado final, o projeto espera gerar produtos e serviços inéditos prontos e testados para serem distribuídos no mercado.

Esse tipo de reuso implica numa curadoria mais sofisticada e com perspectivas mais abertas. Porém, antes de tudo é preciso compreender que os dados de pesquisa são, em grande parte, gerados/coletados com recursos

públicos; são disseminados baseados em licenças que não permitem o uso comercial; e existe uma preocupação forte com questões éticas, de privacidade e de propriedade intelectual. Isto significa que teremos pela frente debates interessantes ao redor dessa questão.

7. Quais competências devem ser desenvolvidas por profissionais para o trabalho com curadoria digital?

O ponto crucial da curadoria de dados de pesquisa é que ela precisa estar inserida organicamente nos fluxos de geração de conhecimento científico para ter validade. De uma forma prática, isto significa que os profissionais de informação e as infraestruturas informacionais e tecnológicas subjacentes às bibliotecas de pesquisa devem estar imbricadas nas atividades dos laboratórios e das outras atividades acadêmicas e de pesquisas da instituição. Isto demanda uma reformulação nos perfis profissionais dos bibliotecários e arquivistas que agora se tornam profissionais de dados. Além disso, a interlocução necessária à curadoria exige ainda novos profissionais provenientes dos domínios específicos – como pesquisadores - e profissionais da área de computação. Vejamos:

A biblioteca de pesquisa – alinhadas a ciência voltada para produtos finais: teses, artigos, inventos, patentes - sempre se concentrou na pós-publicação, na custodia e disseminação de artefatos informacionais acabados, como periódicos e livros; o ciclo de vida dos dados de pesquisa, entretanto, é mais complexo e pressupõe a captura de dados em diferentes estágios de seu processamento e ainda considera linhagens e versões que variam no tempo.

A curadoria se inicia ainda no planejamento dos dados e não se encerra com o fim dos projetos, pois os dados continuam a evoluir. Como desdobramento, as bibliotecas agora têm que se preocupar com os estágios de pré-publicação e com uma gestão contínua. O bibliotecário tem que se preocupar com a gênese dos dados e com toda a documentação necessária à interpretação e à contextualização dos dados ao longo do tempo. Isto implica que ele precisar conhecer as peculiaridades da área em que atua e os fluxos de trabalho dos laboratórios e seus produtos de pesquisa, precisa conhecer também todo o ciclo de vida dos dados de sua instituição e como isso se relaciona com a atividade de curadoria; os metadados gerais e os do domínio disciplinar; os padrões de catalogação; e as teorias que subsidiam a organização do conhecimento e seus instrumentos como taxonomias tesauro e ontologias.

Porém, a curadoria de dados de pesquisa exige uma equipe com muitas expertises que vão além da biblioteconomia de dados:

Para começar, os dados – especialmente os observacionais - precisam ser preservados e arquivados de modo que propriedades arquivísticas, como proveniência, confiabilidade e autenticidade sejam mantidas. Isto acontece porque os pesquisadores do futuro vão basear suas pesquisas na confiança que têm sobre esses dados, portanto eles devem estar hospedados em ambientes estáveis e confiáveis, pois a *web*, por si só, não tem memória permanente, não foi projetada para isso, e não garante que essas propriedades sejam asseguradas. Endereçar essas questões é o papel dos arquivistas de dados.

A curadoria implica, em grande escala, em adicionar valor aos dados. Nessa direção, os dados precisam ser avaliados, analisados, enriquecidos com anotações, ligados por *hiperlinks* com outros recursos, comentados, agregados, "regerados" e descartados. Este papel, que exige conhecimento profundo da área, é atribuído ao pesquisador e a outros especialistas da área.

Por fim, os dados precisam ser analisados, processados por programas e equipamentos específicos que identificam padrões que ajudem na formulação de hipóteses, como *data mining*; precisam ser visualizados e experimentados por sistemas interativos que exibem suas diversas faces. Isto fica por conta dos cientistas da computação, que no contexto da curadoria são chamados de cientistas de dados.

Como citar esta entrevista (APA):
Sayão, L. F. & Sales, L. F. (2016). Curadoria digital e dados de pesquisa. *AtoZ: novas práticas em informação e conhecimento, 5*(2), 67 – 71. Recuperado de: http://dx.doi.org/10.5380/atoz.v5i2.49708