A GESTÃO DE DADOS DE PESQUISA NO CONTEXTO DA E-SCIENCE: BENEFÍCIOS, DESAFIOS E OPORTUNIDADES PARA ORGANIZAÇÕES DE P&D

Patrícia Rocha Bello Bertin
Embrapa - Empresa Brasileira de Pesquisa Agropecuária
patricia.bertin@embrapa.br

Marcos Cezar Visoli
Embrapa - Empresa Brasileira de Pesquisa Agropecuária
marcos.visoli@embrapa.br

Debora Pignatari Drucker
Embrapa - Empresa Brasileira de Pesquisa Agropecuária
debora.drucker@embrapa.br

Resumo: O compartilhamento de recursos computacionais, o acesso distribuído a grandes conjuntos de dados e o uso de plataformas digitais para colaboração e comunicação são características que distinguem a e-Science do paradigma científico tradicional. Esse trabalho situa a Gestão de Dados de Pesquisa como elemento propulsor do avanço científico e tecnológico no paradigma emergente da e-Science, e alerta quanto aos desafios que se impõem às universidades e organizações atuantes no setor de Pesquisa e Desenvolvimento. Além de explorar os princípios e os fundamentos da Gestão de Dados de Pesquisa, o trabalho propõe, a partir de uma abordagem pragmática, um roteiro para a implantação de um programa institucional de Gestão Dados de Pesquisa.

Palavras-chave: Plano de gestão de dados. Ciclo de vida - dados de pesquisa. Validação. Descrição. Descoberta. Preservação. Integração. Análise.

RESEARCH DATA MANAGEMENT IN THE CONTEXT OF E-SCIENCE: BENEFITS, CHALLENGES AND OPPORTUNITIES FOR R&D ORGANIZATIONS

Abstract: The sharing of computational resources, with distributed access to large data sets and digital platforms for collaboration and communication are some of the characteristics that distinguish the emerging e-Science paradigm from the traditional scientific method.This study situates Research Data Management as a driving force for scientific and technological progress in the new paradigm of e-Science, alerting to the challenges faced by universities and Research and Development organizations. Besides exploring the principles and foundations of Research Data Management, this study takes a practical approach to the topic by proposing a roadmap for the implementation of an institutional Research Data Management program.

Keywords: Data management plan. Research data - life cycle. Collect. Assure. Describe. Discover. Preserve. Integrate. Analyze.

1 O NOVO PARADIGMA CIENTÍFICO E A GESTÃO DE DADOS DE PESQUFISA

O termo ‘e-Science’ foi primeiramente enunciado em estudos do Conselho Nacional de Pesquisa do Reino Unido, para designar o modo então emergente de produção do conhecimento científico, o qual se baseava no uso compartilhado de recursos e tecnologias da informação e da comunicação, e permearia diferentes disciplinas e domínios científicos (RESEARCH COUNCILS UK, 2010). O Reino Unido foi uma das primeiras nações a investir na construção de uma infraestrutura de apoio à pesquisa intensiva em informação, por meio da criação, em 2001, do programa e-Science no âmbito do Conselho Nacional de Pesquisa. (BORGMAN, 2007).

A e-Science[1] (GRAY, 2009) caracteriza-se por explorar, no próprio fazer científico, ferramentas computacionais avançadas, que possibilitam amplo acesso a recursos geograficamente dispersos, incluindo coleções de dados, instrumentos científicos e mecanismos de visualização de alto desempenho. Como movimento internacional de crescente interesse e debate, a meta essencial da e-Science é utilizar-se das Tecnologias da Informação e da Comunicação (TICs) como fundamento ‘e-infraestrutura’ para uma transformação profunda no método científico, de modo a extrair o maior proveito possível dos resultados de pesquisa. Do ponto de vista tecnológico, Wouters (2004, p. 2, tradução nossa) enumera em três as principais dimensões práticas da e-Science: “o compartilhamento de recursos computacionais, o acesso distribuído a grandes conjuntos de dados, e o uso de plataformas digitais para colaboração e comunicação”.

Sabe-se que, com os avanços em tecnologia da informação, a capacidade das organizações, grupos e indivíduos de gerar dados tem se tornado cada vez maior, trazendo oportunidades ímpares e, ao mesmo tempo, impondo dificuldades à organização e preservação desses grandes volumes de dados – fenômeno amplamente discutido na atualidade, sob a designação *Big Data* (CHEN et al., 2014). Tal desenvolvimento pode ser observado de maneira análoga no setor de Pesquisa e Desenvolvimento. Isto porque o desenvolvimento científico sempre esteve fundamentado na aquisição e análise de dados produzidos ou obtidos das mais variadas formas e por meio de instrumentos diversos. Com o avanço tecnológico recente na área de instrumentação científica, pesquisadores têm produzido uma quantidade de

[1] Designações correlatas em inglês são *cyberinfrastructure*, *e-Research* e *cyberscience*, entre outras.

dados sem precedentes, muitos dos quais são subutilizados ou pouco explorados em seu potencial para o avanço científico e tecnológico.

Frequentemente, e mesmo nas organizações de pesquisa mais modernas, os dados de pesquisa estão dispersos, mal documentados ou inacessíveis, o que impossibilita sua reutilização em novas análises e a obtenção de respostas a novas perguntas. Dentro de um mesmo grupo de pesquisa, por exemplo, não é incomum que uma nova questão científica provoque um movimento de 'garimpo' de dados, por meio de contatos com ex-alunos, buscas em computadores antigos de membros do projeto – em geral, máquinas encostadas nos cantos dos laboratórios, aguardando por doação – ou, até mesmo, em mídias diversas contendo backups, por vezes obsoletas ou inadequadas para a preservação de dados em longo prazo (DRUCKER, 2012).

1.1 O ESTADO DA ARTE DA GESTÃO DE DADOS DE PESQUISA E A PERSPECTIVA DOS PRINCIPAIS ATORES DO SISTEMA CIENTÍFICO

A gestão de dados de pesquisa tem assumido crescente relevância no cenário científico internacional, na medida em que proliferam estudos que demonstram os benefícios a ela associados, como o aumento de citações e a reutilização de dados arquivados. (PIWOWAR; DAY; FRIDSMA, 2007). O editorial *Data's Shameful Neglect*, publicado pela revista *Nature* em 2009, já destacava a gestão de dados de pesquisa como um dos principais alicerces da produção científica, argumentando que como tal, deveria integrar currículos acadêmicos em todas as áreas do conhecimento (NATURE, 2009).

Ao longo dos últimos anos, diversas agências internacionais de fomento[2] à pesquisa têm passado a requerer, como critério para a concessão de recursos, o comprometimento por parte dos pesquisadores de que os dados derivados dos estudos sejam propriamente

[2] A *National Science Foundation* (USA) requer, desde Janeiro de 2011, que propostas de projeto incluam um documento suplementar denominado 'Plano de Gestão de Dados' (<http://www.nsf.gov/bfa/dias/policy/dmp.jsp>), motivada pela iniciativa da britânica *Natural Environment Research Council* (NERC <http://www.nerc.ac.uk/research/sites/data/dmp/>), de 2010 (requerimentos específicos se aplicam a determinados diretórios e programas de pesquisa). Agências de fomento inglesas, como o *Medical Research Council* e a *Wellcome Trust* (UK) exigem, além de um plano de gestão, que os pesquisadores compartilhem seus dados com a comunidade científica "no momento certo e de modo responsável, assegurando assim que esses dados possam ser verificados e usados como base para avançar o conhecimento científico e suas aplicações na área de saúde" (WELLCOME TRUST UK, 2010 - tradução nossa). Ver também as Diretrizes em Gestão de Dados (*Guidelines on Data Management)* do programa Horizonte 2020 da Comissão Européia (EUROPEAN COMMISSION, 2013).

gerenciados e arquivados em repositórios temáticosou institucionais, a fim de garantir preservação em longo prazo e maior facilidade de compartilhamento.

A plataforma re3data.org (<http://www.re3data.org/about/>), lançada em 2012 pela *German Research Foundation* (DFG), cataloga repositórios de dados de pesquisa de diferentes disciplinas acadêmicas em todo o mundo. No Brasil, a plataforma reconhece apenas quatro repositórios de dados. São eles: o 'Banco de Dados de Exploração e Produção (BDEP)' da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (<http://www.bdep.gov.br/?lng=br>); o 'Repositório de Dados de Levantamentos Biológicos', do Centro de Estudos Integrados da Biodiversidade Amazônica (<https://ppbiodata.inpa.gov.br/metacatui/>); a plataforma de dados climáticos WorldClim (<http://worldclim.org/>); e o repositório do Instituto Brasileiro de Informação em Ciência e Tecnologia, IBICT Dataverse Network (<https://repositoriopesquisas.ibict.br/dvn/>)[3]. Para fins de comparação, o serviço re3data.org cataloga 884 repositórios de dados de pesquisa nos Estados Unidos da América, 238 no Reino Unido, 133 no Canadá, 30 na Índia e 25 na China[4]. Alguns repositórios internacionais de referência são: o holandês 3TU.Datacentrum (<http://datacentrum.3tu.nl/en/home/>), o americano BioLINCC (<https://biolincc.nhlbi.nih.gov/home/>), o UK Data Archive (<http://www.data-archive.ac.uk/home>), o GBIF (Global Biodiversity Information Facility <http://www.gbif.org/>) e o Dryad (<http://datadryad.org/>).

Analogamente, algumas revistas científicas passaram a requerer que os dados utilizados nos artigos por elas publicados sejam documentados em repositórios digitais com pré-condição para a publicação (BUTLIN, 2011; WHITLOCK et al., 2010).

Para alcançarem sustentabilidade e competitividade no sistema científico moderno, portanto, instituições e pesquisadores devem garantir o apropriado gerenciamento e preservação dos dados de pesquisa, de modo a possibilitar a verificação futura de resultados e a reutilização dos dados originais. Mais que isso, estudos têm apontado que dados de pesquisa bem-organizados, documentados, preservados, acessíveis e verificados quanto à sua acurácia e validade são mais

[3] Embora a plataforma re3data.org não os catalogue, no entanto, é possível identificar ao menos quatro repositórios de dados de pesquisa já bem estruturados no Brasil: o Repositório de Dados do PELD – Sítio Floresta Amazônica (Pesquisas Ecológicas de Longa Duração, <http://peld.inpa.gov.br/knb/style/skins/peld/>); o Sistema de Informação Ambiental do Programa Biota/Fapesp (SinBiota, <http://sinbiota.biota.org.br/>); o Sistema de Informação sobre a Biodiversidade Brasileira (<http://www.sibbr.gov.br/>), o repositório de dados do PPBio (Programa de Pesquisa em Biodiversidade Amazônia Ocidental <https://ppbio.inpa.gov.br/repositorio/dados>).

[4] Levantamento realizado em janeiro de 2017.

facilmente compartilháveis e reutilizáveis, ocasionando: a ampliação do impacto, da visibilidade e da credibilidade do pesquisador, da pesquisa e da instituição; a descoberta de novos usos e inovação; a prevenção de fraudes; e a redução da replicação de esforços e custos associados à pesquisa (APPEL; MACIEL; ALBAGLI, 2016; ARZBERGER et al., 2004; FIENBERG; MARTIN; STRAF, 1985; NATIONAL RESEARCH COUNCIL - US, 1999).

Como se pode depreender do parágrafo acima, para que sejam verdadeiramente úteis, os dados de pesquisa devem possuir estrutura e organização. É nesse sentido que se configura a necessidade de aprimoramento dos mecanismos e das práticas de gestão de dados de pesquisa em universidades e organizações de P&D.

Este trabalho apresenta os princípios e fundamentos da gestão de dados de pesquisa e, empregando uma abordagem prática, aponta os elementos essenciais para um programa institucional de Gestão de Dados de Pesquisa.

2 PRINCÍPIOS, NOÇÕES FUNDAMENTAIS E PLANEJAMENTO EM GESTÃO DE DADOS DE PESQUISA

Embora não exista uma definição consensual na literatura, pode-se afirmar que dados compreendem o elemento tangível, ou o registro do qual derivam a informação e o conhecimento. Dados de pesquisa, por sua vez – e também de modo simplificado –, são todo o tipo de registro produzido, compilado ou utilizado no decorrer da pesquisa. Uma definição popular na literatura, oferecida pela Organização para a Cooperação e Desenvolvimento Econômico (OECD), estabelece que:

> [Dados de pesquisa são os] registros factuais usados como fontes primárias na pesquisa científica, e que são geralmente aceitos na comunidade científica como sendo necessários para validar os resultados de pesquisa. Um conjunto de dados de pesquisa constitui uma representação parcial e sistemática do objeto de investigação. (OECD, 2007, p. 13)

Em termos práticos, no entanto, o que se entende por dado de pesquisa varia de acordo com a disciplina, a área do conhecimento, o contexto e até mesmo com sua destinação ou finalidade. A fotografia de uma construção municipal depositada em um arquivo histórico, por exemplo, pode não representar muito para um agrônomo; enquanto que, para um historiador, aquela fotografia torna-se um dado de pesquisa. Vale ressaltar que nem todos os dados derivados de um projeto de pesquisa são publicados em artigos científicos; e ainda, que

dados produzidos no âmbito de um determinado projeto podem ser utilizados em uma agenda de pesquisa completamente diferente, por outro grupo de pesquisadores. Ou seja, o 'ruído' observado em uma determinada pesquisa pode consistir em 'sinal' para outro pesquisador, fato que reforça a importância do gerenciamento, compartilhamento e preservação dos dados de pesquisa.

Gestão de dados de pesquisa (GDP) é, assim, uma expressão abrangente que envolve tanto os aspectos rotineiros de planejamento, aquisição, organização, estruturação, definição de fluxos analíticos e ferramenta computacional apropriada para o armazenamento de dados, quanto às questões relativas à preservação, à organização, ao compartilhamento, à proteção e à confidencialidade destes para a instituição que possui o direito sobre tais dados, bem como o acesso e disponibilização para a sociedade (COX; PINFIELD, 2013; WHYTE; TEDDS, 2011).

Pode-se destacar o 'planejamento' como sendo um dos princípios-chave da GDP: idealmente, deve-se refletir sobre a gestão de dados ainda na concepção do projeto[5]. Diversas questões devem ser consideradas, quando do planejamento da GDP, particularmente:

- Tipos, formatos e conjuntos de dados existentes;
- Métodos existentes para a coleta de dados;
- Questões legais, éticas e relacionadas à propriedade intelectual;
- Níveis de acesso;
- Formas de compartilhamento e reutilização dos dados;
- Gestão, curadoria de dados[6] e armazenamento de curto prazo;
- Depósito (arquivo) e preservação em longo prazo.

O **Quadro 1** apresenta questões que podem auxiliar o pesquisador na elaboração de um Plano de Gestão de Dados de Pesquisa.

Reflita sobre os itens a seguir, no momento de elaboração da sua proposta de pesquisa:

[5] Organizações internacionais como a britânica *Digital Curation Centre* e a *California Digital Library*, entre outras, oferecem ferramentas online que auxiliam o pesquisador no planejamento da gestão de dados – disponíveis em <https://dmponline.dcc.ac.uk/> e <https://dmp.cdlib.org/>, respectivamente.

[6] Sayão e Sales (2012, p. 184) definem 'curadoria' de dados de pesquisa como 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".

Quadro 1. Pontos de reflexão para elaboração de um Plano de Gestão de Dados de Pesquisa (elaboração própria).

1. Tipos de dados de pesquisa
- Defina os tipos de dados que serão coletados na pesquisa.
- Reflita sobre o(s) formato(s) no(s) qual(is) esses dados serão obtidos.
- Procure estimar o volume de dados que será coletado (número e tamanho de arquivos, objetos).

2. Formatos dos dados e padrões
- Defina ou produza tabelas de codificação, dicionários de dados ou outro tipo de documentação para explicar os termos, nomes das variáveis, códigos e abreviações utilizadas.
- Procure documentar a forma como os dados serão coletados ou criados.

3. Políticas de acesso
- Preocupe-se em remover informações pessoais ou confidenciais dos dados, para garantia de privacidade.
- Reflita sobre quem deterá os direitos autorais sobre os dados da sua pesquisa.

Defina e produza documentação sobre como o crédito institucional e pessoal deve ser reconhecido para os dados da pesquisa em questão.
- Assegure que os dados da pesquisa sejam rotulados e organizados logicamente, pela utilização de nomes de arquivos consistentes e fáceis de compreender.

4. Uso de dados e distribuição
- Reflita sobre o tempo de vida (validade) dos dados resultantes da pesquisa.
- Observe se existem razões para limitar ou restringir a reutilização ou redistribuição dos dados. Se sim, defina o período de embargo necessário.

5. Preservação de dados e arquivamento
- Identifique um repositório digital ou outra infraestrutura que fará com que seus dados de pesquisa estejam acessíveis para visualização e download, quando cabível.
- Reflita e organize a forma de armazenamento em longo prazo e preservação dos seus dados (tanto itens físicos quanto digitais).
- Considere os recursos necessários para custear o depósito e a preservação dos dados em longo prazo, caso estes sejam exigidos pelo repositório digital escolhido.
- Estabeleça planos de segurança de dados para garantir que sejam armazenados e que backups sejam realizados na frequência desejada.
- Procure utilizar formatos de dados e softwares que permitam o compartilhamento e garantam a validade dos dados em longo prazo, tais como softwares não proprietários ou baseados em padrões abertos.
- Nos processos de conversão de um formato para outro (caso previstos), assegure-se de que dados não foram perdidos ou modificados.

Cabe ressaltar que conjuntos de dados[7] devem estar acompanhados de informação que descreva a sua origem (tempo ou espaço, métodos, instrumentos de coleta, fluxos analíticos), âmbito, autoria, propriedade e condições de reutilização – ou seja, de 'metadados'. Em paralelo com a interoperabilidade tecnológica, a existência de metadados adequados e normalizados é um requisito essencial para garantir a fácil recuperação, o acesso e a reutilização dos dados de pesquisa (WILKINSON, 2016).

Cabe ressaltar, no entanto, que a GDP não se limita a dados estruturados – como aqueles contidos em uma tabela ou banco de dados –, mas considera uma variedade de formatos e de suportes.

[7] Em termos simples, conjuntos de dados (*datasets*) são coleções de fatos relacionados entre si e registrados em formato computacional comum.

2.1 FORMATOS USUAIS DOS DADOS DE PESQUISA

O dado de pesquisa pode nascer digital ou ser convertido para um formato digital, o que facilita o gerenciamento com o suporte de ferramentas tecnológicas. Alguns dos formatos mais comuns estão listados abaixo:

- Texto: arquivos textuais comuns, MS Word, Portable Document Format (PDF), Rich Text Format (RTF), Hyper-Text Markup Language (HTML), Extensible Markup Language (XML), etc.
- Numérico: SPSS, Stata, MS Excel, SAS, arquivos hierárquicos, etc.
- Multimídia: JPEG, TIFF, GIF, MPEG, Quicktime, Bitmap, PNG, etc.
- Modelo: 3D, estatístico, macroeconômico, causal, de similaridade, etc.
- Software: arquivos e códigos binários escritos em uma variedade de linguagens de programação (script), como Java, C, Perl, Python, Ruby, PHP.
- Disciplina-específico: ex. Crystallographic Information File (CIF), da química.
- Instrumento-específico: ex. Carl Zeiss Digital Microscopic Image Format (ZVI).
- Espaciais: Representações computacionais de dados geográficos, frequentemente como dados matriciais e vetoriais.

Sendo assim, os dados de pesquisa podem estar contidos em uma variedade de suportes:

- Documentos (texto, Word), planilhas.
- Atas de laboratório, cadernos de campo, diários de pesquisa.
- Questionários, transcrições, tabelas de codificação.
- Fitas/CDs/DVDs de áudio e vídeo.
- Fotografias, filmes.
- Resultados de ensaios.
- Slides, artefatos, espécimes, amostras.
- Coleção de objetos digitais adquiridos e produzidos durante o processo de pesquisa.
- Arquivos de dados estatísticos ou de outra natureza.
- Conteúdo de bancos de dados (vídeo, áudio, texto, imagens).
- Modelos, algoritmos, scripts.
- Conteúdo de uma aplicação (inputs, outputs, arquivos de log para análise de software, softwares de simulação, esquemas).
- Metodologias e *workflows*.
- Procedimentos operacionais padrão e protocolos.
- Mapas e arquivos de dados espaciais, como *shapefiles* e imagens de satélite.

Além dos itens explicitados acima, os seguintes artefatos, produzidos no âmbito dos projetos de pesquisa, podem ser de interesse gerencial, por conterem registros dos dados de pesquisa: comunicações formais e correspondências (e-mails e correspondências em papel); propostas de projeto e arquivos associados; relatórios técnicos e de pesquisa; publicações científicas e técnicas; comunicações em mídia social, como *blogs*, *wikis*, *tweets*, etc.; e arquivos digitais contendo dados coletados por meio de experimentos.

2.2 TIPOLOGIA DOS DADOS DE PESQUISA

Dados de natureza distinta demandam uma abordagem gerencial também diferenciada. Uma classificação útil para dados de pesquisa – dentre as diversas possíveis –, é oferecida a seguir:

- Observacionais: dados únicos e insubstituíveis, normalmente capturados em tempo real, tais como imagens estáticas (fotos, radiografias) e dinâmicas, ou ainda aqueles coletados por meio de questionários.
- Experimentais: dados de resultados experimentais, por exemplo, de equipamentos de laboratório, por vezes reprodutíveis, mas de alto custo, tais como cromatogramas e *microassays*.
- De simulação: dados produzidos para reproduzir o comportamento de sistemas a partir do uso de modelos, nos quais o próprio modelo e os metadados podem ser mais importantes do que os dados de saída, tais como modelos econômicos ou climáticos.
- Derivados ou compilados: resultantes da transformação ou combinação de dados brutos, normalmente reprodutíveis, mas de alto custo, como bancos de dados compilados, meta-análises, mineração de texto, e dados censitários agregados.
- De referência: um conglomerado (estático ou orgânico) ou coleção de conjuntos de dados menores (revisados por pares), muito provavelmente publicados e curados, como bancos de dados genéticos, censitários e bases de dados cristalográficos.
- Workflows analíticos: documentação das sequências de transformações, processamentos, algoritmos, análises e outras etapas utilizadas para alcançar determinado resultado, permitindo a reprodutibilidade e o reuso.

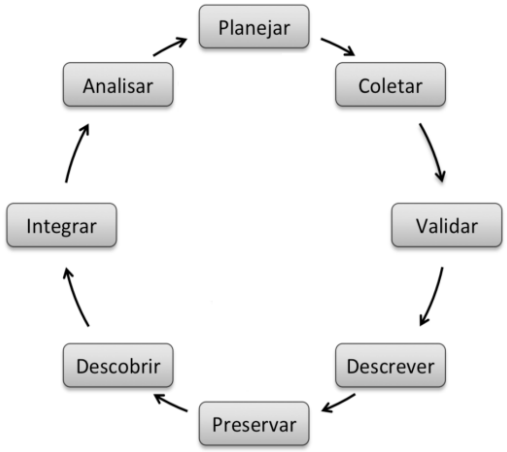
2.3 CICLO DE VIDA DOS DADOS DE PESQUISA

De modo geral, os dados de pesquisa têm uma vida útil maior do que os projetos que os originam: encerrado o projeto, pesquisadores podem continuar a trabalhar com os dados;

projetos de continuidade podem ser iniciados, a partir dos mesmos dados; ou estes podem, ainda, ser utilizados por outros pesquisadores. Ou seja, dados de pesquisa possuem um 'ciclo de vida': são gerados, processados, analisados e preservados, podendo ser novamente reutilizados, e cada um desses estágios demanda abordagens e ferramentas distintas de gerenciamento.

Podem-se encontrar, na literatura científica e na *World Wide Web*, diversos modelos ou formas de interpretação do 'ciclo de vida dos dados'. A Figura 1 destaca o modelo idealizado pela cooperação internacional *DataONE (Data Observation Network for Earth)*, apoiada pela *US National Science Foundation*:

Figura 1. Ciclo de vida dos dados,



Fonte: segundo a DataONE[8].

Embora a figura represente um ciclo lógico, os processos podem ocorrer simultaneamente ou se repetirem. A seguir, uma breve descrição das etapas que compõem o ciclo de vida dos dados proposto pela DataONE:

- Planejamento: Mapeamento de todos os processos e recursos para o ciclo completo de vida dos dados. Deve incluir os objetivos do projeto, impacto, resultados e produtos esperados, um plano de gestão de dados e a política de backup.
- Coleta: Definição da estratégia mais adequada para a coleta de dados e armazenamento em arquivos de dados legíveis e interoperáveis. O resultado desta etapa é um modelo de dados que descreve como os dados estão estruturados.
- Validação: Definição dos processos que assegurarão a qualidade dos dados, como protocolos a serem seguidos, procedimentos para calibração de equipamentos, técnicas para identificar erros e como tratá-los.

[8] Traduzido a partir de DataONE (<https://www.dataone.org/data-life-cycle>).

- Descrição: Documentação completa dos dados, com a descrição de quem, quando, o quê, porque, onde e como os dados foram obtidos, utilizando o conceito de metadados, ou dados sobre dados, para possibilitar o reuso e compartilhamento. Existem diversos padrões e ferramentas para auxiliar nesta etapa, e frequentemente metadados são documentados em linguagem XML, facilitando a interoperabilidade e a recuperação da informação.
- Preservação: Preservação dos dados em curto e longo prazo. Curto prazo para evitar perda de dados por acidentes, e longo prazo para que viabilize o acesso e uso dos dados no futuro. É definir o que será preservado, o local e a documentação que deve acompanhar os dados.
- Descoberta: Localização, obtenção ou recuperação dos dados, por meio de metadados, ontologias ou thesaurus, por exemplo, permitindo novos usos.
- Integração: Combinação de dados de diferentes fontes, internas ou externas ao projeto e instituição. A viabilidade e a qualidade da integração de dados dependerão das técnicas de gerenciamento de dados aplicadas durante o ciclo de vida.
- Análise: Dados são analisados com diferentes objetivos e por diferentes pesquisadores, gerando resultados que instigarão a novas perguntas, levando à formulação de um novo ciclo de planejamento e etapas seguintes.

3 CONSTRUINDO UM PROGRAMA INSTITUCIONAL DE GESTÃO DE DADOS DE PESQUISA

Para promover o conhecimento científico, a *e-Science* requer o estabelecimento de relações mutuamente benéficas entre pesquisadores e entre grupos de pesquisa, independentemente de barreiras geográficas, políticas e, muitas vezes, institucionais (GRAY, 2009). O novo paradigma científico demanda intrinsicamente, um esforço coletivo e interdisciplinar, com intercâmbio facilitado de dados e informações – afinal, a exploração das grandes quantidades de dados gerados ao longo das atividades de pesquisa é uma de suas características fundamentais.

A fim de inserir-se no contexto da e-Science, faz-se necessário que a gestão de dados de pesquisa seja incorporada à estratégia organizacional. O principal objetivo de um programa institucional de GDP deve ser o de oferecer, à comunidade de pesquisadores, as ferramentas, o treinamento, o apoio e a orientação que são necessários para o apropriado gerenciamento dos dados de pesquisa em todo o seu ciclo, assegurando assim o uso responsável desse valioso

recurso para a PD&I. Para alcançar esse objetivo, deve-se primeiramente obter um maior entendimento da real situação dos dados de pesquisa na instituição. Algumas das questões que precisam ser respondidas são:

- Com que tipos de dados de pesquisa a organização lida?
- Como esses dados são produzidos e validados?
- Onde são armazenados os dados e os conjuntos de dados de pesquisa, e por quanto tempo esses dados podem ser úteis?
- Como esses dados serão preservados, de modo possam ser acessíveis daqui a alguns anos?
- Qual a qualidade dos conjuntos de dados da organização?
- Quem são os autores dos dados?
- Quais dados de pesquisa proveem bases para inovação tecnológica e, por isto mesmo, necessitam ser protegidos? E quais dados podem ser disponibilizados abertamente?
- Quem pode reutilizar os dados de pesquisa produzidos pela organização? E de que modo os dados de pesquisa podem ser reutilizados?

Para o levantamento do estado da arte da gestão de dados de pesquisa na organização, recomenda-se a aplicação de um instrumento de diagnóstico (questionário eletrônico)[9] junto a representantes das unidades de pesquisa. A análise do panorama completo compreende o primeiro estágio para a implantação de uma governança de dados de pesquisa, ao qual se seguirão os seguintes estágios:

- Elaboração de uma política institucional e implantação da estratégia de gestão de dados de pesquisa. Como o custo de guarda e preservação de dados de pesquisa pode ser elevado, a organização deve estabelecer critérios e decidir quais dados precisam ser mantidos, quais devem ser protegidos e quais devem ser disseminados, dentro de um mesmo grupo de pesquisa, entre diferentes grupos, ou externamente à organização.
- Implantação de uma estrutura de governança de dados de pesquisa, com definição dos elementos essenciais, atores e responsabilidades.
- Desenvolvimento de ferramentas e serviços de suporte à gestão de dados de pesquisa, como por exemplo: (i) um ou mais repositórios de dados de pesquisa, que possibilitem o armazenamento, a referenciação, o acesso, a reprodutibilidade e a reutilização de

[9]A exemplo da ferramenta gratuita denominada CARDIO (*'Collaborative Assessmento of Research Data Infrastructure and Objectives'*), disponível em <http://cardio.dcc.ac.uk/>. Desenvolvida pela Universidade de Glasgow, a plataforma integra as melhores funcionalidades das ferramentas de curadoria de dados existentes, podendo ser aplicada no âmbito de departamentos ou grupos de pesquisa, ou ainda em avaliações múltiplas que, tomadas juntamente, constroem um panorama institucional realista.

dados em longo prazo; (ii) treinamentos sobre o ciclo de vida dos dados[10]; (iii) serviços de apoio e consultoria para o desenvolvimento de planos de gestão de dados; (iv) consultoria quanto a aspectos legais; (v) suporte para a seleção de ferramenta tecnológica apropriada e treinamentos para o uso daquelas disponibilizadas pela organização; (vi) e serviço de curadoria dos dados de pesquisa na instituição.

Cabe ressaltar, no entanto, que aspectos culturais podem representar uma barreira importante para o avanço da GDP na organização. Isto porque os dados gerados pela pesquisa são frequentemente vistos como propriedade individual e não organizacional, não raramente mantidos em suportes, sistemas, equipamentos e computadores particulares. Para superar as resistências culturais, faz-se necessária a implantação de um plano de comunicação arrojado – que instrua os pesquisadores quanto às vantagens associadas à gestão e ao compartilhamento dos dados de pesquisa e sobre o fato de que, embora o crédito moral pela descoberta científica seja do autor, o dado de pesquisa é propriedade da organização. Aliado a isto, é importante que os sistemas de avaliação de desempenho e recompensa incluam critérios que valorizem a organização, preservação, o compartilhamento e o reuso dos dados de pesquisa, sem os quais a e-Science não se estabelecerá prevalente.

4 REFLEXÕES FINAIS

Com uma gestão de dados de pesquisa pouco estruturada e, na maior parte das vezes, relegada à vontade individual de pesquisadores, estima-se que grande parte dos dados produzidos por organizações brasileiras de P&D possa ter-se perdido ou estar sob risco, em razão da fragilidade e obsolescência tecnológica dos suportes e mídias nos quais estão contidos, ou simplesmente devido ao não compartilhamento pelos responsáveis por sua produção ou obtenção.

Para que a instituição, o pesquisador e a sociedade colham os benefícios decorrentes da GDP, recai primeiramente sobre o pesquisador a responsabilidade de identificar as melhores práticas de gestão de dados em sua área de atuação, de modo que possam ser organizados, preservados e compartilhados responsavelmente. A alta direção da instituição,

[10] Há, na web, cursos gratuitos de orientação a pesquisadores em gestão de dados de pesquisa, como o MANTRA (*Research Data Management Training*, <http://datalib.edina.ac.uk/mantra/#sthash.z2cW1r1z.dpuf>), que oferece materiais específicos para determinados atores no processo de gestão de dados de pesquisa, como bibliotecários, por exemplo (<http://datalib.edina.ac.uk/mantra/libtraining.html>).

por sua vez, deve estimular e reforçar os benefícios do gerenciamento e do compartilhamento dos dados de pesquisa, oferecer as ferramentas e os serviços necessários à descoberta e reutilização dos dados, e trabalhar para garantir que os autores recebam o crédito pelos dados citados.

REFERÊNCIAS

APPEL, A. L.; MACIEL, M. L.; ALBAGLI, S. A e-Science e as novas práticas de produção colaborativa do conhecimento científico. **Revista Internacional de Ciencia y Sociedad**, v. 3, n. 1, p. 41–52, 2016.

ARZBERGER, P. et al. Promoting access to public research data for scientific, economic, and social development. **Data Science Journal**, v. 3, n. 29, p. 135–152, 2004.

BORGMAN, C. L. **Scholarship in the Digital Age:** Information, Infrastructure and Internet. Cambridge: MIT Press, 2007.

BUTLIN, R. Data archiving. **Heredity**, v. 106, n. 5, p. 709, maio 2011.

CHEN, M. et al. **Big Data:** Related Technologies, Challenges and Future Prospects. Vancouver: Springer International Publishing, 2014.

COX, A. M.; PINFIELD, S. Research data management and libraries: Current activities and future priorities. **Journal of Librarianship and Information Science**, v. 46, n. 4, p. 299–316, 2013.

DRUCKER, D. P. **A integração da informação sobre biodiversidade e ecossistemas para embasar políticas de conservação:** o projeto Biota Gradiente Funcional como estudo de caso. [s.l.] Universidade Estadual de Campinas, 2012.

EUROPEAN COMMISSION. **Guidelines on FAIR Data Management in Horizon 2020**, 2013. Disponível em: <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>. Acesso em: 2 fev. 2017.

FIENBERG, S. E.; MARTIN, M. E.; STRAF, M. L. (EDS.). **Sharing Research Data**. Washington: National Academy Press, 1985.

GRAY, J. Jim Gray on eScience: A transformed scientific method. In: HEY, T.; TANSLEY, S.; TOLLE, K. (Eds.). **The fourth paradigm:** Data-intensive scientific discovery. Redmond: Microsoft Research, 2009.

NATIONAL RESEARCH COUNCIL - US. **A question of balance:** Private rights and the public interest in scientific and technical databases. Washington: National Academy Press, 1999.

NATURE. Data's shameful neglect. **Nature - Editorial**, v. 461, n. 7261, p. 145, 10 set. 2009.

OECD, O. FOR E. C. AND D.-. **OECD Principles and Guidelines for Access to Research Data**. [s.l: s.n.]. Disponível em: <http://www.oecd.org/sti/sci-tech/38500813.pdf>. Acesso em: 2 fev. 2017.

PIWOWAR, H. A.; DAY, R. S.; FRIDSMA, D. B. Sharing detailed research data is associated with increased citation rate. **PloS one**, v. 2, n. 3, p. e308, 21 jan. 2007.

RESEARCH COUNCILS UK. **Delivering the UK's e-Infrastructure for Research and Innovation**. Swindon: RCUK, 2010. Disponível em: <http://www.rcuk.ac.uk/documents/research/esci/e-infrastructurereviewreport-pdf/>. Acesso em: 2 fev. 2017.

SAYÃO, L. F.; SALES, L. F. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informacao e Sociedade**, v. 22, n. 3, p. 179–191, 2012.

WELLCOME TRUST UK. **Policy on data management and sharing**. Disponível em: <www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm>. Acesso em: 18 abr. 2016.

WHITLOCK, M. C. et al. Data Archiving. **The American Naturalist**, v. 175, n. 2, p. 145–146, 2010.

WHYTE, A.; TEDDS, J. **Making the case for research data management**. (Digital Curation Centre Briefing Papers). Edinburgh: JISC, 2011. Disponível em: <http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm>. Acesso em: 2 fev. 2017.

WILKINSON, M. D. Comment: The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, p. 160018, 2016.

WOUTERS, P. **What is the matter with e-Science ?** Thinking aloud about informatisation in knowledge creation. 4S & EASST Conference: Public Proofs, Science, Technology and Democracy. Anais… Ecole des Mines, Paris, 2004.