SELEÇÃO E AVALIAÇÃO DE COLEÇÕES DE DADOS DIGITAIS DE PESQUISA: uma possível abordagem metodológica

Luana Farias Sales
Comissão Nacional de Energia Nuclear/Instituto de Engenharia Nuclear, Email: luanafsales@gmail.com
Márcia Teixeira Cavalcanti
Instituto de Engenharia Nuclear, Email: marciacavalcanti@gmail.com

Resumo
A área de Ciências Nucleares, assim como diversos outros domínios científicos, produz intensivamente uma diversidade de dados de pesquisa, que vão desde resultados de experimentos até dados gerados a partir de simulações, como, por exemplo, os originados de pesquisas em realidade virtual e inteligência artificial. Este fato vem sendo evidenciado no âmbito do Instituto de Engenharia Nuclear (IEN), uma unidade da Comissão Nacional de Energia Nuclear (CNEN), órgão vinculado ao Ministério da Ciência, Tecnologia e Inovação (MCTI) do Brasil. Apesar das especificidades da área Nuclear o problema de identificação, coleta e seleção dos dados de pesquisa é um desafio que se apresenta em qualquer domínio ou instituição que queira iniciar um projeto de curadoria de dados científicos. Neste sentido, uma questão que se coloca neste contexto é como identificar os dados de pesquisas produzidos dentro de uma instituição de pesquisa científica? Sendo assim, o artigo objetiva apresentar uma possível abordagem metodológica, aplicada no IEN, que se acredita ser possível replicar em qualquer instituição de pesquisa um caminho metodológico possível dentre outros, sendo indicada sua aplicação em instituições nas quais ainda não exista um plano de gestão de dados (PGD).
Palavras chave: Desenvolvimento de Coleção. Dados de Pesquisa. Seleção de Dados.

1 Introdução

O número crescente de dados digitais

A IBM divulgou recentemente que 90% dos dados que existem hoje no mundo foram gerados nos últimos quatro anos, sendo praticamente inconcebível a quantidade de dados que são gerados e armazenados em escala global. Em um curso sobre curadoria digital ministrado na Fundação Casa de Rui Barbosa no final de 2015 foi abordada a questão da geração de dados nos dias atuais. Segundo o palestrante, Aquiles Alencar Brayner, em um dia são gerados 7TB no Twitter e 10TB no Facebook, as duas redes sociais mais usadas, e até 2020 teremos aproximadamente 35 ZB (1.1 trilhão GB) de dados digitais disponíveis. Mas como garantir que estes dados fiquem acessíveis hoje e no futuro? Este é um dos desafios que a sociedade atual está enfrentando.

No campo científico o cenário é o mesmo, pois o aprimoramento tecnológico permitiu que os cientistas não apenas colocassem em prática o que antes ficava apenas no terreno teórico, como também fez multiplicar o número de dados gerados a partir de aparatos tecnológicos: satélites, microscópios, telescópios, dentre outros.

> Hoje, a descoberta científica não ocorre apenas através de um rigoroso, bem definido processo de testes de hipóteses. Os vastos volumes de dados, as complexas relações, difíceis de descobrir, os intensos e mutáveis tipos de colaboração entre as disciplinas e os novos tipos de publicação quase em tempo real estão acrescentando a descoberta de padrões e de regras ao

> método científico (ANDERSON, 2008 apud ABBOT, 2011, p.134).

Abbot (2011) vislumbra um cenário em que o alinhamento entre a oferta e a demanda da ciência dependerá da busca de novas relações que superem barreiras culturais e linguísticas para se permitir a colaboração, vendo este processo como algo muito mais comum com jogos em rede do que com o método científico tradicional. Essa nova forma de fazer ciência "incluirá redes informais de serviços potenciais, rápida inovação nas margens e uma parceria muito mais estreita entre aqueles que criam conhecimento e aqueles que o usam" (ABBOT, 2011, p. 134).

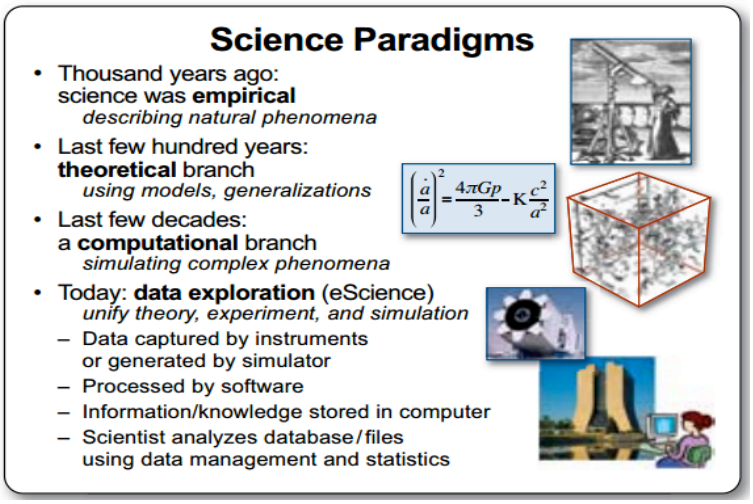
Diversos autores já nomeiam a época em que vivemos como a da pesquisa científica centrada em dados, mas Fox e Hendler (2011, p.159) apontam que "as tecnologias tradicionais não foram projetadas para lidar com a escala e a heterogeneidade de dados no mundo moderno". As bases de dados de pesquisa estão crescendo em número e volume, e os diferentes campos da ciência passam a ter acesso a um número mais abrangente de fontes de dados.

Os mesmos autores indicam problemas que surgem com a consolidação da *e-science* e a queda das barreiras ao acesso livre aos dados de pesquisas: qual a forma correta de uso dos dados que não foram gerados pelo próprio cientista, como usar tipos de dados nunca vistos antes, como usar dados de outras disciplinas sem entender seus termos. E apontam que muitos cientistas já usam a própria internet como seu principal computador em razão do volume crescente de dados que se encontram disponíveis (Fox; Hendler, 2011, p.159-160).

A *e-science* se configura como um novo paradigma científico orientado por e para os dados de pesquisa e intrinsecamente ligado à tecnologia. De acordo com o *National e-Science Centre*, futuramente a *e-science* será a referência científica, sendo realizada cada vez mais através da colaboração global entre os cientistas possibilitada pela internet, e isso fará com que ela exija o acesso a coleções de dados muito grandes, além de grandes recursos computacionais que irão demandar uma infraestrutura mais poderosa.

> Foi Johannes Kepler, assistente de Tycho **Brahe** [grifo do autor], quem pegou o catálogo de observações astronômicas sistemáticas de Brahe e formulou as leis do movimento planetário. Este fato estabeleceu a divisão entre a mineração e a análise de dados experimentais coletados e cuidadosamente arquivados, de um lado, e a criação de teorias, de outro. Esta divisão é um dos aspectos do Quarto Paradigma (BELL, 2011, p.11).

Essa nova ciência baseada em computação intensiva de dados, para Bell (2011), vai inaugurar um novo Quarto Paradigma científico, que Gray (2009) ilustra na imagem a seguir:



Fonte: GRAY, 2009, p.xviii

Embora as pesquisas baseadas em dados ocorram com mais frequência e tenham sua origem no campo das Ciências Naturais e Biológicas, no campo das Ciências Sociais e Humanas não está muito diferente.

> Os dados associados com esse tipo de investigação são provenientes de múltiplas fontes: experimentos científicos que investigam o comportamento do ambiente, medições que capturam diferentes aspectos dos modelos ou simulações contrastantes. Exemplos específicos de dados em ciências naturais são medições de precipitações de chuvas, as observações astronômicas, bases de dados de modelos genéticos ou estruturas cristalográficas. Nas ciências sociais os dados são gerados através de pesquisa de opinião ou mapas com censos de informações georreferenciadas. Em humanidades, podem incluir fotografias de antigas escrituras em pedra, e em medicina neuroimagens que captam a atividade do cérebro. (MARTÍNEZ-URIBE; MACDONALD, 2008, p.274 tradução nossa)

Jim Gray, em palestra proferida na Califórnia em 2007 (TOLLE; TANSLEY; HEY, 2011, p. 17), já vislumbrava a necessidade de se produzir ferramentas que atendessem ao ciclo completo da pesquisa: da captura dos dados e sua curadoria até a análise e visualização. Segundo ele [na época em que proferiu sua palestra] existem péssimas ferramentas de gerenciamento de dados para a maioria das disciplinas científicas que realmente ajudem os cientistas a capturar seus dados, cura-los, analisa-los, para poderem visualiza-los posteriormente. Isso faz com que a pesquisa científica acabe por gastar boa parte do financiamento na construção de softwares para poder lidar com a informação gerada pela própria pesquisa. Para ele, a necessidade de se criar novas ferramentas ocorre para que toda a literatura científica e todos os dados gerados nas pesquisas científicas estejam *online* e possam ser interoperáveis.

Se medidas apropriadas de preservação não forem colocadas em prática em breve, corremos o risco de continuar perdendo os dados digitais oriundos da pesquisa científica, como já vem acontecendo em virtude do acelerado avanço tecnológico, pois dados gerados em suportes que se tronaram obsoletos, como os disquetes, já não encontram equipamentos em que possam ser utilizados e nem programas. Assim, além da necessária criação de ferramentas, a curadoria digital de dados de pesquisa se mostra como o caminho complementar a ser seguido para se alcançar o objetivo da preservação e do compartilhamento seguros.

A área de Ciências Nucleares, assim como diversos outros domínios científicos, produz intensivamente uma diversidade de dados de pesquisa, que vão desde resultados de experimentos até dados gerados a partir de simulações, como, por exemplo, os originados de pesquisas em realidade virtual e inteligência artificial. Este fato vem sendo evidenciado no âmbito do Instituto de Engenharia Nuclear (IEN), uma unidade da Comissão Nacional de Energia Nuclear (CNEN), órgão vinculado ao Ministério da Ciência, Tecnologia e Inovação (MCTI) do Brasil.

Como desdobramento de suas atividades acadêmicas e de pesquisa e desenvolvimento o IEN vem produzindo grande quantidade de informações e dados de pesquisa em formatos digitais. Esses recursos informacionais são parte da memória institucional do IEN e se configuram como parte considerável do conhecimento produzido na instituição. Esse conhecimento é explicitado em forma de relatórios, teses e dissertações, artigos, multimídia, apresentações, aulas, programas de computador, simulações, ambientes virtuais e coleções de dados de pesquisa, que estão registrados nos mais diversos formatos e mídias, precisando ser preservado para as gerações futuras.

Com a finalidade de preservar, integrar, recuperar, compartilhar e reusar em âmbito nacional os dados e informações da área Nuclear, possibilitando que estes dados sirvam de base para novas atividades de pesquisa e cumpram o seu papel de memória, foi desenvolvido no IEN uma plataforma tecnológica[1] para apoiar a custódia e a preservação desses recursos informacionais, possibilitando a integração da produção técnico-científica e também dos dados digitais de pesquisas (ex: resultados de experimentos, medidas, resultados de levantamentos, gráficos, diagramas, tabelas ou modelos em 3D, imagens, vídeos e gravações em áudio, fórmulas matemáticas estruturas químicas legíveis por máquina, código fonte de software, entre outros).

No entanto, a criação de uma ferramenta não é suficiente para solucionar o problema da gestão e da preservação do conhecimento, sendo necessário o desenvolvimento de técnicas e metodologias para aquisição, seleção, tratamento e, consequentemente, preservação dos dados e informações digitais que serão depositados na plataforma. Neste sentido, a identificação dos dados produzidos pela comunidade de pesquisadores locais se torna relevante para o estabelecimento de uma política institucional para o desenvolvimento de uma coleção de dados a serem preservados.

Apesar das especificidades da área Nuclear o problema de identificação, coleta e seleção dos dados de pesquisa é um desafio que se apresenta em qualquer domínio ou instituição que queira iniciar um projeto de curadoria de dados científicos. Neste sentido, uma questão que se coloca neste contexto é como identificar os dados de pesquisas produzidos dentro de uma instituição de pesquisa científica? Sendo assim, o artigo objetiva apresentar uma possível abordagem metodológica, aplicada no IEN, que se acredita ser possível replicar em qualquer instituição de pesquisa como mecanismo auxiliar na identificação de dados científicos. Esclarecendo que este é apenas um caminho metodológico possível dentre outros, sendo indicada sua aplicação em instituições nas quais ainda não exista um plano de gestão de dados (PGD).

Curadoria digital de dados

Segundo o relatório elaborado pelo consórcio *Blue Ribbon Task Force on Sustainable Digital Preservation and Access/BRTF-SDPA*, um fórum de especialistas patrocinado pela Fundação Nacional de Ciências dos EUA e a Andrew W. Mellon Foundation em colaboração

[1] Essa plataforma, denominada Carpe dIEN, pode ser consultada em <http://carpedien.ien.gov.br>.

com diversas outras instituições de pesquisa, intitulado *Sustainable economics for a digital planet* (2010), e dentro de uma perspectiva econômica:

> A informação digital é um recurso vital em nossa economia do conhecimento, valiosa para a pesquisa e a educação, para a ciência e as humanidades, para as atividades criativas e culturais e para as políticas públicas. Mas a informação digital é inerentemente frágil e frequentemente corre o risco de se perder. O acesso a valiosos materiais digitais amanhã depende das ações de preservação tomadas hoje; e, ao longo do tempo, o acesso depende de alocação permanente e eficiente dos recursos para preservação (Sustainable Economics for..., 2010, p.1 tradução nossa).

A preservação da informação digital é um problema social urgente, segundo o relatório, por ser frágil e propensa à degradação. A preservação segura vai garantir a conservação da informação ao longo do tempo, mas para isso devem-se usar as melhores práticas em todo o ciclo de vida dos dados: criação, descrição e curadoria, além de seu armazenamento seguro, que garanta seu uso e reuso. O acesso à informação de amanhã depende de ações cautelares tomadas hoje, pois um fato fundamental da sustentabilidade digital é que, sem preservação, não há acesso. Neste cenário, vemos surgir um novo papel e campo de trabalho para os profissionais que lidam com informação: a curadoria digital.

O conceito de curadoria, originalmente, aparece associado a museus, mas por empréstimo linguístico é utilizado hoje para se referir ao tratamento empregado aos objetos digitais. Para Beagrie (2006), curadoria digital ainda é um termo muito novo e em evolução, passível de ser percebido de formas diferentes por diferentes indivíduos e disciplinas.

Segundo o *Digital Curation Centre*/DCC[2], a curadoria digital envolve "manter, preservar e adicionar valor ao conjunto de dados digitais de pesquisa através de seu ciclo de vida" (tradução nossa). Por ser um termo recente ainda é possível encontrarmos algumas divergências para defini-lo.

A curadoria digital pode ser entendida como uma prática que pretende preservar objetos digitais, obedecendo seu ciclo de vida e garantindo o seu acesso futuro, e será bem sucedida se conseguir manter essa informação disponível por tempo indeterminado, se sobrepondo até mesmo à obsolescência digital (de software e hardware).

O DCC tem, em sua página, um modelo gráfico de curadoria do ciclo de vida dos dados de pesquisa, onde apresenta as etapas necessárias para uma curadoria de sucesso. Este modelo, tido como referência pelos profissionais da área, define elementos chave e ações que devem ser tomadas em todo o processo. As ações do DCC que envolvem a curadoria de dados se dividem e se subdividem em:

a) Por todo o ciclo de vida dos dados: descrição e representação da informação; plano de preservação; monitoramento e participação; curadoria e preservação.

b) Sequenciais: conceitualização; criar ou receber; avaliar e selecionar; transferir os dados (para um arquivo, repositório etc.); ação de preservação; armazenar; acesso, uso e reuso; transformar (criar novos dados a partir dos originais).

c) Ocasionais: dispor dados que não passaram pelo processo de curadoria; reavaliar; migrar dados (para diferentes formatos).

Segundo Harvey (2010, p.7), ainda que nos perguntemos sobre o que é curadoria digital, é possível afirmarmos o que ela não é: ela não é arquivamento digital e ela não é preservação digital, embora ambos sejam aspectos importantes dela. Curadoria digital é um

[2] http://www.dcc.ac.uk/

conceito muito mais inclusivo que arquivamento e preservação digitais, e sua ênfase está no fato de que adiciona valor aos conjuntos de dados e objetos digitais.

> Para preservar os dados científicos para as gerações futuras é essencial que as comunidades de pesquisa (usuários e produtores de dados), serviços de informática (que sabem gerenciar a tecnologia nas organizações) e bibliotecas (com as suas competências na preservação e experiência em repositórios) trabalhem juntos (Lyon, 2007). A infraestrutura necessária para a realização deste objetivo não consiste simplesmente em soluções tecnológicas de armazenamento. É muito importante tomar as medidas necessárias desde o momento da criação destes valiosos recursos digitais (Doorn, Tjalsma, 2007). Todas as atividades de gestão de dados científicos se resumem no termo curadoria de dados. (MARTÍNEZ-URIBE; MACDONALD, 2008, p.276 tradução nossa).

Os pesquisadores vêm sendo, cada vez mais, solicitados a arquivarem seus dados brutos, ou dados primários, em repositórios de dados de acesso público. A criação de repositórios de dados é um dos caminhos para a preservação de dados gerados pela pesquisa científica e o exercício da curadoria digital:

> O armazenamento de dados científicos em repositórios e sua reutilização é uma das preocupações do recém-lançado Programa FAPESP de Pesquisa em eScience, expressão que resume o desafio de pesquisa para organizar, classificar e garantir acesso ao gigantesco volume de dados gerados continuamente em todos os campos de pesquisa, a fim de extrair novos conhecimentos e fazer análises abrangentes e originais. (MARQUES, 2014, p.56)

Os repositórios de dados científicos, ou repositórios de dados de pesquisa, em vista do material que nele é depositado, composto de conteúdos com características próprias e que necessitam de um tratamento diferenciado, incluindo também conjuntos específicos de metadados para organizar estes conjuntos de dados, acabam por se diferenciarem dos repositórios digitais de maneira geral.

Como está publicado em seu sítio[3], o Instituto de Engenharia Nuclear (IEN) é uma unidade da Comissão Nacional de Energia Nuclear (CNEN), e desde sua fundação, em 1962, vem contribuindo para o domínio nacional de tecnologias na área nuclear e correlatas. Suas atividades de pesquisa, desenvolvimento e inovação geram produtos e serviços como patentes, publicações, licenciamento de tecnologias, fornecimento de radiofármacos, ensaios e análises de materiais, recolhimento de rejeitos radioativos, consultorias e formação de recursos humanos.

O IEN busca reunir tanto sua produção técnico-científica quanto os dados oriundos das atividades de pesquisa, e para consolidar essa reunião criou um repositório institucional denominado Plataforma CarpedIEN, cujo objetivo é servir de ferramenta para auxiliar os pesquisadores na gestão e preservação dos dados oriundos de suas pesquisas e encoraja-los a explorar e adotar novas formas de comunicação científica em ambientes digitais e pelo compartilhamento de dados.

O repositório institucional (RI) desenvolvido no IEN utiliza o software livre DSpace, que é orientado para a criação de repositórios institucionais e à preservação digital. Podemos classificar um RI como parte de uma estrutura maior formada pelas bibliotecas digitais que tem como função a preservação dos objetos digitais (documentos, livros, relatórios, dados,

[3] http://www.ien.gov.br/oinstituto/historico.php

projetos) em longo prazo. No IEN o repositório institucional é voltado para a preservação de dados e informações em energia nuclear.

> A Plataforma CarpedIEN pode ser vista como uma estratégia para dar ordenação e visibilidade à informação científica produzida pela instituição e é definida como um "repositório voltado para o arquivamento, gestão, preservação e disseminação de dados e informações em formatos digitais gerados em decorrência das atividades de ensino e pesquisa do IEN" (SALES, 2013). A definição de sua política, bem como o modelo de metadados construído para a plataforma considerou em sua essência a preservação e o uso de padrões de tratamento para representação dos documentos técnico-científicos, bem como dos dados de pesquisa. Assim sendo, pode ser considerada uma primeira iniciativa rumo à curadoria digital de dados de pesquisa na Instituição. (SALES, 2014, p.154)

O RI do Instituto é organizado em comunidades que representam as áreas temáticas de pesquisa. No ano de 2013 o IEN, com o objetivo de organizar para poder apoiar de forma mais justa as pesquisas desenvolvidas, lança uma chamada interna para que seus pesquisadores, tecnologistas e analistas em C&T identificassem parceiros para poderem se organizar em áreas temáticas. Oito áreas temáticas foram aprovadas:

1) Engenharia e Tecnologia de Reatores Nucleares
2) Química Nuclear e Radioquímica
3) Desenvolvimento de Instrumentação Nuclear
4) Desenvolvimento de Tecnologia para Sistemas Complexos
5) Realidade Virtual Aplicada à Área Nuclear
6) Aplicação de Técnicas Nucleares na Indústria, Saúde e Meio-Ambiente
7) Gestão do Conhecimento Nuclear: Preservação, Disseminação e Compartilhamento do Conhecimento gerado no IEN
8) Desenvolvimento e caracterização de materiais funcionais e estruturais para o setor nuclear

O questionário elaborado e aplicado busca ter uma primeira amostragem sobre como os pesquisadores do Instituto de Engenharia Nuclear/IEN geram, armazenam e compartilham os dados originados por suas pesquisas, para identificar, dentro de cada projeto em andamento, quais são os tipos de dados gerados, como e se os dados estão sendo tratados, e estabelecer uma ordem de prioridades para dar início ao processo de desenvolvimento de coleção de dados de pesquisa.

Este questionário também faz parte de um projeto institucional que pretende alimentar o repositório de dados criado no IEN, e para isso está em curso um processo de desenvolvimento de protótipos metodológicos para a curadoria digital de dados de pesquisa na área nuclear. Sendo esta temática nova no Brasil, e na área de Ciências Nucleares, um método para seleção de dados precisa ser estabelecido.

O protótipo metodológico em que se está trabalhando agora, e que será descrito a seguir, parte da elaboração de critérios para a seleção de dados na área nuclear que passarão pelo processo de curadoria digital. No caso específico do IEN, *lócus* deste trabalho, como os dados gerados pelas pesquisas não passaram por um plano de gestão de dados – que deveria ter sido estabelecido antes do desenvolvimento das pesquisas - foi necessário fazer uma avaliação destes dados (*data appraisal*) para decidir quais dariam início ao processo de curadoria.

2. Procedimentos metodológicos

A pesquisa no Instituto de Engenharia Nuclear

Segundo a Apresentação do Progress Report:

> Pesquisa e Desenvolvimento em Engenharia Nuclear é o processo institucional que permite ao Instituto de Engenharia Nuclear/IEN desenvolver, renovar e criar o conhecimento necessário para atender as necessidades presentes e futuras de seus clientes. Em 2013 uma ampla revisão da organização dos campos de pesquisa e desenvolvimento foi realizada, com o objetivo de identificar e selecionar áreas de pesquisa promissoras de acordo com as estratégias adotadas pelo Instituto. Em decorrência disso, em 2015, foi organizada uma publicação, o *Progress Report*, de acordo com as contribuições destas áreas de pesquisa selecionadas, destacando as pesquisas realizadas pelo IEN nos anos de 2013-2014 (SAMPAIO, 2015, p.2 tradução nossa).

As análises apresentadas nas tabelas que compõem o próximo item tiveram origem no questionário aplicado aos autores que submeteram resumos de pesquisas em andamento e publicações para serem inseridos no *Progress Report*, um relatório de progresso de pesquisas, e que formam nossa população alvo, ou seja, o grupo a quem esse levantamento se aplica. Este questionário é uma das abordagens adotadas pelo IEN para identificar os dados de pesquisas institucionais. Ao todo foram cinquenta e um resumos, e o procedimento para submeter os resumos das publicações foi todo *onlin*e. Nesse processo se solicitou aos autores que respondessem, também *online*, o questionário elaborado. Para cada resumo apresentado e aceito para publicação temos um questionário correspondente respondido.

Além de mapear os dados que são gerados pelas pesquisas desenvolvidas no IEN, outro objetivo do questionário foi obter informações sobre as pesquisas para que futuramente possa ser criado um plano de gestão dos dados (PGD) e ter informações suficientes para auxiliar os pesquisadores na elaboração do PGD para seus projetos de pesquisa. A criação de um PGD é uma condição colocada pela maioria das agências de fomento internacionais, e, no caso do Brasil, atualmente a Fundação de Amparo à Pesquisa do Estado de São Paulo/FAPESP já vem seguindo este modelo.

> O PGD descreve o ciclo de vida de gestão para todos os dados que serão coletados, processados ou gerados por um projeto de pesquisa. De uma forma abreviada, ele se constitui em um documento formal que estabelece um compromisso de como esses dados serão tratados durante todo o desenvolvimento do projeto, e também após a sua conclusão (SAYÃO; SALES, 2015, p.15).

Os dados de pesquisa gerados por investigações desenvolvidas institucionalmente são de propriedade também da instituição (dado os investimentos desta em salário, infraestrutura e material para a realização dos estudos). Para que os dados de pesquisa possam ser compartilhados e reusados, seja pelos próprios pesquisadores que os originaram seja por grupos de pesquisas correlatos, eles precisam ser tratados, preservados e curados. No entanto, uma tarefa importante que procede todas elas é a tarefa de identificação dos dados que estão sendo produzidos. Mesmo quando institucionalmente existe um setor voltado para tal tarefa, é difícil cobrir todas as pesquisas que as áreas produzem, dado que cada área pode ter diversas linhas de investigação e inúmeras pesquisas em desenvolvimento. Aqui entra a importância de se implantar um plano de gestão de dados (PGD).

Os critérios nomeados para a seleção dos dados

Segundo Beagrie (2010), os dados se tornaram fundamentais em diferentes áreas de pesquisa porque possuem uma importância estratégica frente os desafios globais atuais, são centrais em projetos interdisciplinares e aumentam substancialmente em escala e complexidade. Mas tudo isso vem acompanhado de alguns desafios, se destacando principalmente o custo que a preservação destes dados em longo prazo gera. O compartilhamento e o reuso dos dados podem ser citados como um dos inúmeros benefícios que a preservação em longo prazo de conjuntos de dados de pesquisa traz, como também o aumento da transparência dos resultados das pesquisas.

Estabelecer o valor de conjuntos de dados de pesquisas não é uma tarefa simples. Para a "U.S. National Archives and Records Administration" (2007), exige conhecimento e sensibilidade para os interesses dos pesquisadores. Na elaboração do protótipo metodológico para a curadoria digital de dados de pesquisa na área nuclear, foi necessário pensar em uma forma eficaz de seleção e avaliação dos dados de pesquisa para o início do processo de curadoria.

No que tange às ações sequenciais do modelo de curadoria do ciclo de vida dos dados do DCC, o processo de desenvolvimento de coleções inclui as fases chamadas *appraise and select* (avaliar e selecionar), que são definidas como o processo de desenvolver critérios para determinar quais dados e objetos digitais serão mantidos em longo prazo e os que serão descartados (HARVEY, 2010).

Appraisal é um termo oriundo da arquivologia e, no contexto atual de objetos digitais e da curadoria digital de dados de pesquisa, uma definição para o termo seria "O processo que determina se os materiais possuem valor suficiente para serem acessados dentro de um repositório" (PIERCE-MOSES, 2005, p.22). S*election* se define como o que será adicionado às coleções das bibliotecas, e se constitui em um conjunto de critérios comuns e particulares a cada biblioteca que irá funcionar como um guia no momento da formação ou da ampliação da coleção.

Harvey (2007) apresenta dois caminhos opostos para a seleção de dados para preservação. O primeiro prevê um futuro tecnológico em que o computador terá seu poder de processamento aumentado além de diminuir os custos de arquivamento, o que permitirá manter todos os dados e evitará a necessidade de tomar decisões sobre o que se deve arquivar e manter para uso futuro. O segundo, no qual ele se baseará, faz algumas suposições importantes:

> Não é prático manter o acesso a todos os dados indefinidamente. Não é desejável manter o acesso a todos os dados indefinidamente. Avaliação e seleção de dados para a preservação se baseiam no seu significado e valor contínuo. Precisamos selecionar e preservar mais do que apenas os próprios dados, a fim de entendê-los no futuro. Uma avaliação ou política de seleção que defina claramente os processos e a base para a tomada das decisões de seleção é necessária. Decisões de avaliação e seleção devem ser aplicadas de forma consistente. Decisões de avaliação e seleção devem ser baseadas em uma compreensão clara dos objetivos da organização que aceita a responsabilidade de preservação (HARVEY, 2007, p.8 tradução nossa).

Ele levanta a questão de que as decisões sobre a seleção de materiais digitais têm necessariamente pesadas implicações de recursos em curso, ao contrário do material à base de papel, para o qual os custos contínuos de manutenção podem ser suspensos por períodos de tempo sem grande efeito prejudicial. E que os critérios de *appraisal* e *selections* usados para materiais tradicionais não se aplicam aos dados sem sofrerem modificações. Assim, elenca três diferenças entre essas práticas para objetos em papel e objetos digitais, no caso, dados de pesquisa: ênfase na capacidade técnica para se preservar dados que exigem ações contínuas; o

custo corrente na preservação de dados; tomada de decisão de preservação no início da existência dos dados para não se tornarem inacessíveis ou desaparecerem (HARVEY, 2007, p.10).

Não existe uma política mandatória para a preservação de dados gerados pelas pesquisas no IEN, portanto não houve uma tomada de decisão para preservação no início de sua geração, fazendo com que se torne urgente a criação de critérios para avaliar e selecionar a enorme quantidade de dados gerados pelas pesquisas, impedindo, assim, que estes se tornem inacessíveis ou se percam.

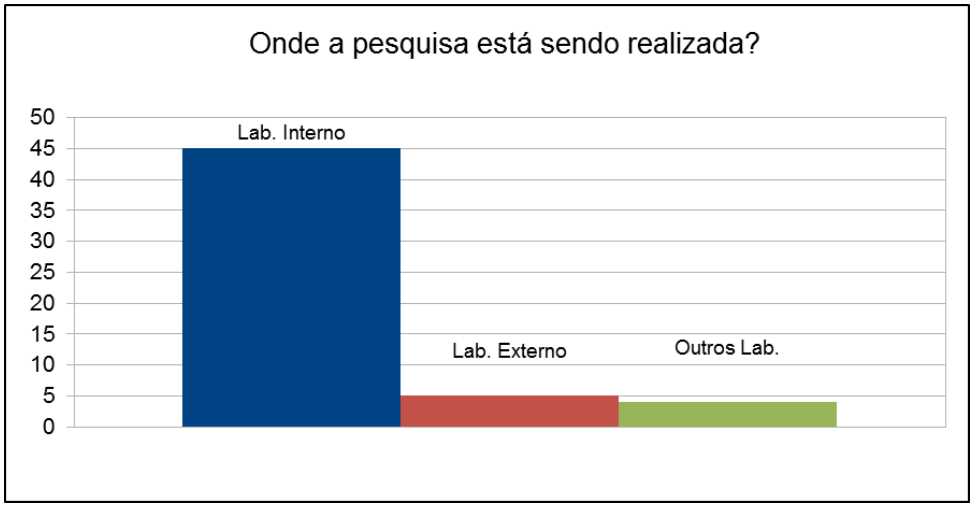
Após leituras e análises sobre os diferentes tipos de dados que são gerados pela pesquisa na área Nuclear, foi elaborado um conjunto de critérios para avaliar e selecionar quais os dados que darão início ao processo de curadoria digital no IEN. Estes critérios são:

1) Dificuldade de reprodução e fragilidade dos dados - dados gerados por experimentos de difícil reprodução ou com grande custo para serem reproduzidos.
2) Potencial de Reuso - se os dados possuem grande potencial de reuso devem ser prioritários
3) Formatos/Mídia - dados gerados por software que mudam com constância devem ser prioritários e preservados junto com a versão do software
4) Proveniência - dados que possuem proveniência significa que possuem uma história e que mudam com constância e merecem ser preservados
5) Embargos - dados que não podem ser disponibilizados devem ser preservados, mas tendo o processo de curadoria encurtado podem ficar para depois
6) Ética - dados que precisam ser anonimizados tem o processo de curadoria mais detalhado e devem ser deixados para depois

3. Resultados

O questionário

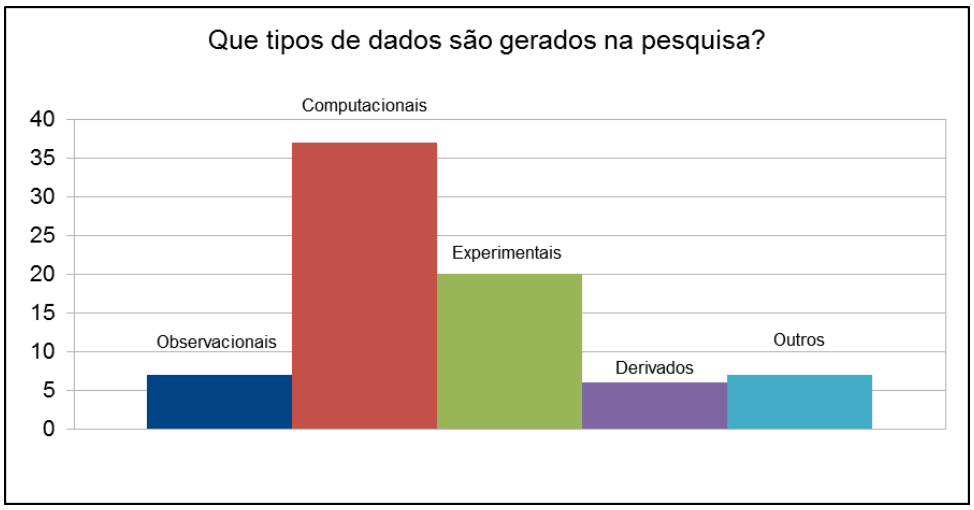
O questionário foi estruturado com perguntas fechadas de múltipla escolha que comportam mais de uma alternativa de resposta e com uma única questão composta de alternativas mutuamente exclusivas. As tabelas estão organizadas pelo número de respostas para cada questão.



Fonte: Dados gerados pela pesquisa (2015)

O objetivo desta pergunta era saber com qual intensidade as pesquisas desenvolvidas dentro do IEN se realizam em colaboração com outras instituições. Todos os cinco respondentes que marcaram a opção laboratório externo também marcaram laboratório

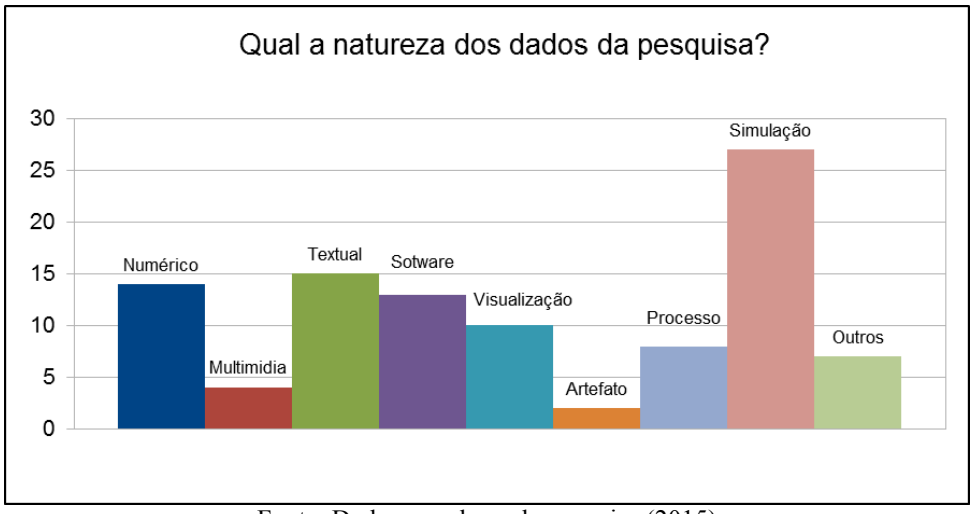
interno, o que demonstra um número relativamente baixo de colaboração externa entre os respondentes. Já a opção outros laboratórios obteve quatro respostas. Ambas as opções, laboratório externo e outros laboratórios, não solicitavam que estes fossem identificados, pois identificar o local onde as pesquisas são desenvolvidas é importante para sabermos se será possível ter acesso a todos os dados de pesquisas gerados no projeto, ou se as pesquisas possuem chances de serem reproduzidas.



Fonte: Dados gerados pela pesquisa (2015)

A questão buscou identificar os diferentes tipos de dados que são produzidos durante o período de desenvolvimento da pesquisa. Os dados computacionais são os mais frequentes, seguidos dos dados experimentais, justamente os de maior nível de fragilidade.

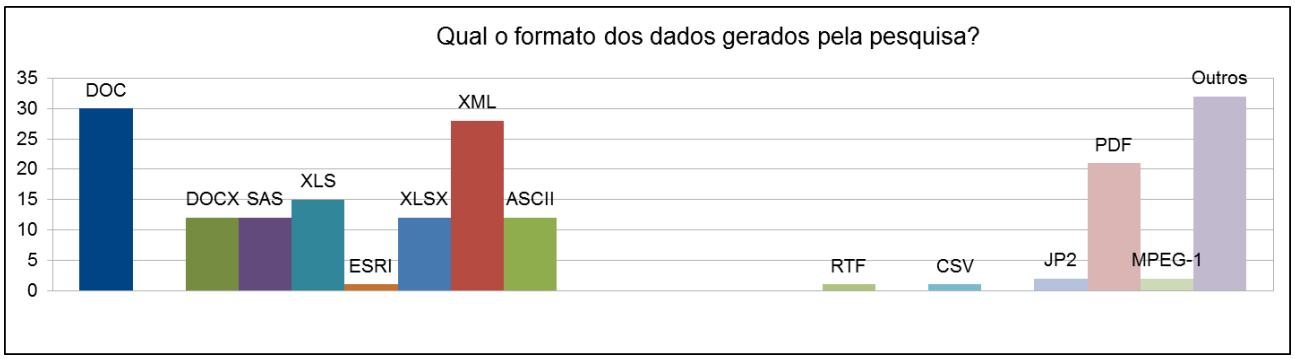
Dados computacionais são os dados gerados a partir da execução de modelos computacionais ou de simulações, já os dados experimentais são aqueles provenientes de situações controladas em bancadas de laboratórios. Sendo o IEN uma instituição de pesquisa no campo das ciências exatas, é fácil compreendermos que sejam esses tipos de dados os que mais sejam gerados. Além dos dados é necessário se preservar também outros itens utilizados nas pesquisas, como software, instrumentos, equipamentos.



Fonte: Dados gerados pela pesquisa (2015)

Os dados de pesquisa gerados pelo Instituto podem ser classificados de diferentes modos: numérico (medidas, resultados de levantamentos, resultados de experimentos, fórmulas, equações, algoritmos), multimídia (Imagens, vídeo, áudio, animações, filme,

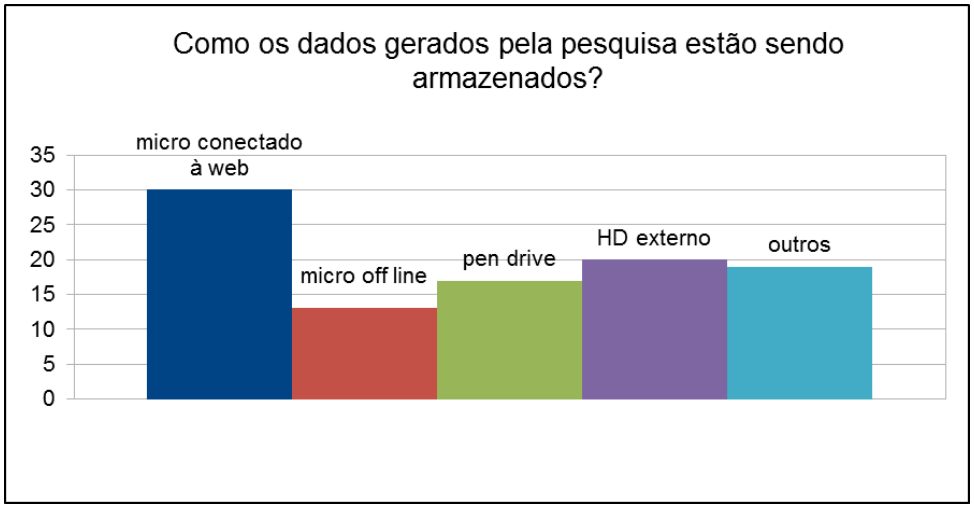
fotografia), textual (metadados, questionários, entrevistas, anotações, normas, padrões, certificados, caderno de laboratório, transcrição, correspondências, diário, caderno de campo), software (bases de dados, simulações, códigos nucleares), visualização (tabelas, gráficos, diagramas, modelos em 3D, modelos reduzidos, desenhos), artefato (espécimes, amostras, maquete), processo (procedimentos operacionais padronizados, workflows, protocolos, teste) e simulação, dentre outros. Os Dados que mais aparecem são os resultantes de simulação, seguidos pelos textuais. Essa informação é importante para nos auxiliar na escolha dos padrões de metadados a serem utilizados para descrever esses tipos de dados.



Fonte: Dados gerados pela pesquisa (2015)

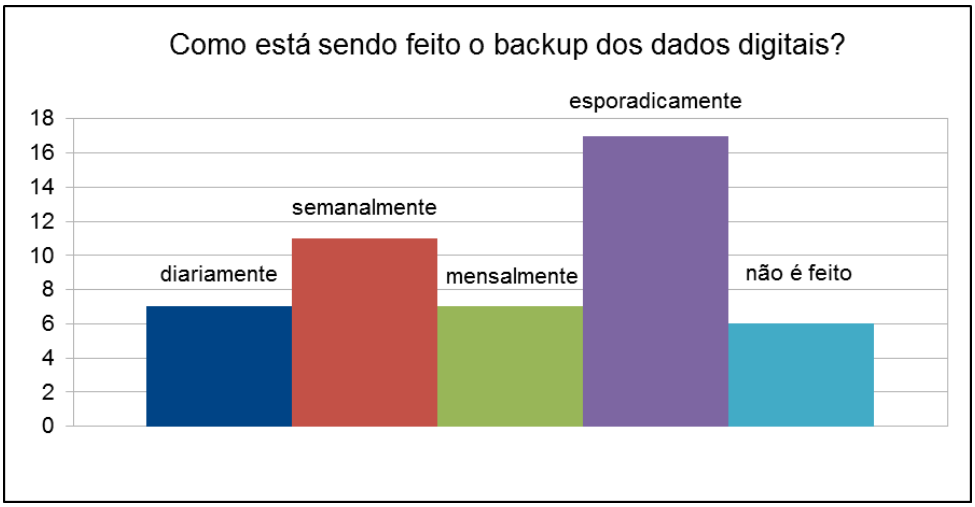
Nesta pergunta as opções correspondem ao formato de arquivo que contem os dados gerados. O formato de arquivo é a forma usada por cada software para reconhecer os dados gerados, sendo que cada um deles tem um formato específico para que possa tratar as informações contidas no arquivo gerado: doc (Documento de texto do Microsoft Word); docx (Documento de texto do Microsoft Word na versão a partir de 2007); sas (conjunto de dados organizados no formato de tabela); xls (um arquivo de planilhas do Microsoft Excel); xlsx (um arquivo de planilhas do Microsoft Excel a partir da versão 2010); esri (conjunto de dados geoespaciais); xml (arquivo de dados em que as marcações definem a estrutura dos dados); ascii (arquivo de texto); rtf (arquivo de texto); csv (arquivo de texto); jp2 (arquivo de imagem); pdf (arquivo que pode ser visualizado independente do programa que o criou); mpeg-1 (arquivos de áudio e vídeo compactados). Essa informação é importante para nos auxiliar na escolha dos padrões de metadados a serem utilizados para descrever esses tipos de dados.

De todos os formatos apresentados como opção de resposta, doc e xml são os que mais aparecem depois de outros, mas não foi pedido que os respondentes especificassem suas respostas.



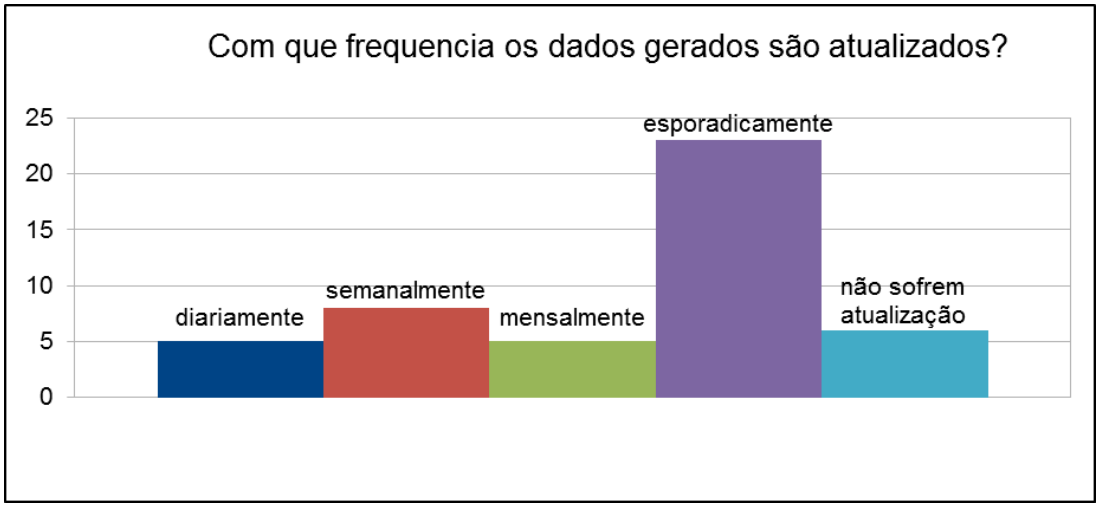
Fonte: Dados gerados pela pesquisa (2015)

O objetivo da questão era identificar de que forma os dados gerados pelos pesquisadores em suas pesquisas diárias estavam sendo preservados. Muitas pesquisas realizadas no IEN produzem dados que são sigilosos para acesso externo, o que significa que estes dados devam ser armazenados em lugares que não permitam qualquer possibilidade de acesso externo. Entretanto, a maioria dos respondentes armazenam seus dados em computadores conectados a web, e na proporção oposta vemos o armazenamento em computadores *off line*. Essa pergunta nos ajuda a identificar quem tem se preocupado de alguma forma com a preservação das informações geradas pela pesquisa.



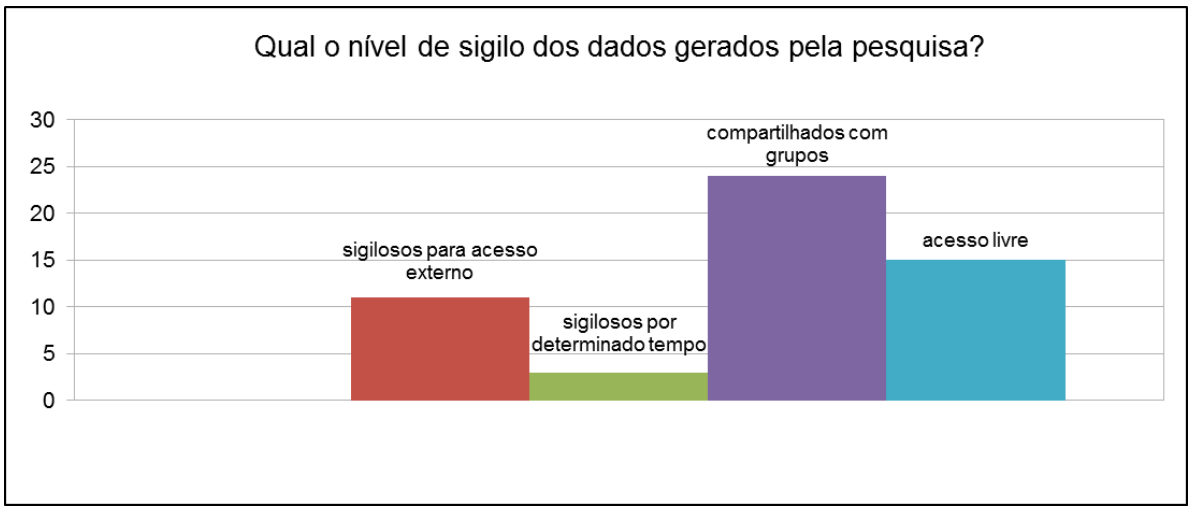
Fonte: Dados gerados pela pesquisa (2015)

Com que frequência o backup dos dados é realizado? Para a maioria dos respondentes, essa cópia de segurança é feita esporadicamente. Como a opção mensalmente existe, esporadicamente significa que essa cópia demora mais de um mês para ser feita. Mais preocupante ainda é não fazer, mesmo que o número de respondentes nesta opção seja pequeno. É preciso se criar a consciência de que a cópia de segurança é primordial para a pesquisa, pois é ela que permite a restauração dos dados em caso de perda dos originais, seja por apagamento ou por estarem corrompidos. Essa pergunta nos ajuda a identificar quem tem se preocupado com a preservação das informações geradas pela pesquisa.



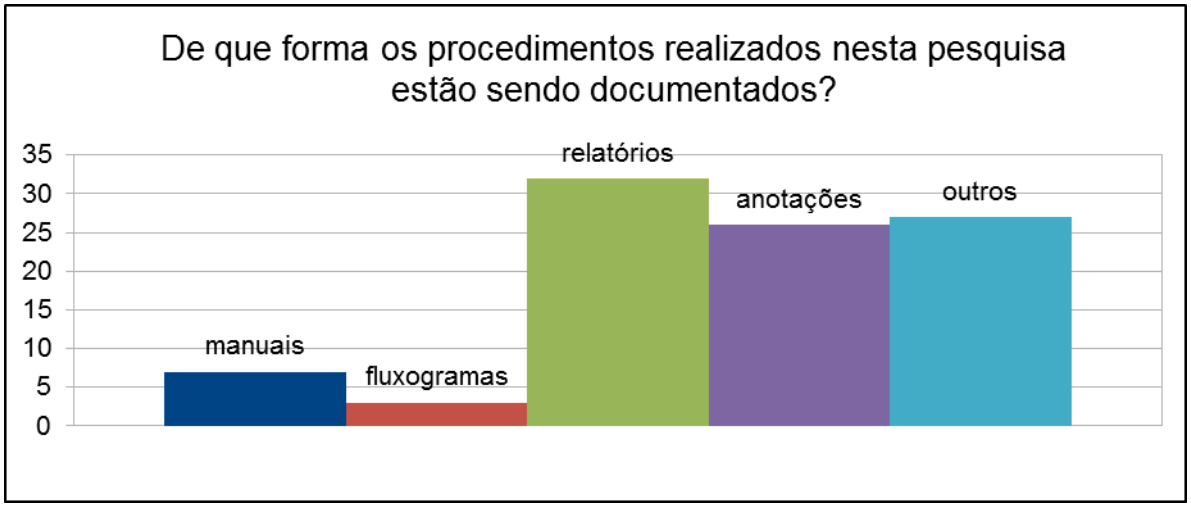
Fonte: Dados gerados pela pesquisa (2015)

Com relação à atualização dos dados gerados pela pesquisa encontramos o mesmo problema que aparece no backup, a atualização também é feita de forma esporádica, o que significa que demora mais de um mês para ser feita, pois as opções diariamente, semanalmente e mensalmente estão disponíveis no questionário. Também encontramos respondentes, em um número pequeno, que declaram não fazer a atualização dos dados. Não fica claro se essa atualização não é feita porque a pesquisa gera dados que não necessitam ser atualizados ou se isso ocorre como uma opção por parte do pesquisador.



Fonte: Dados gerados pela pesquisa (2015)

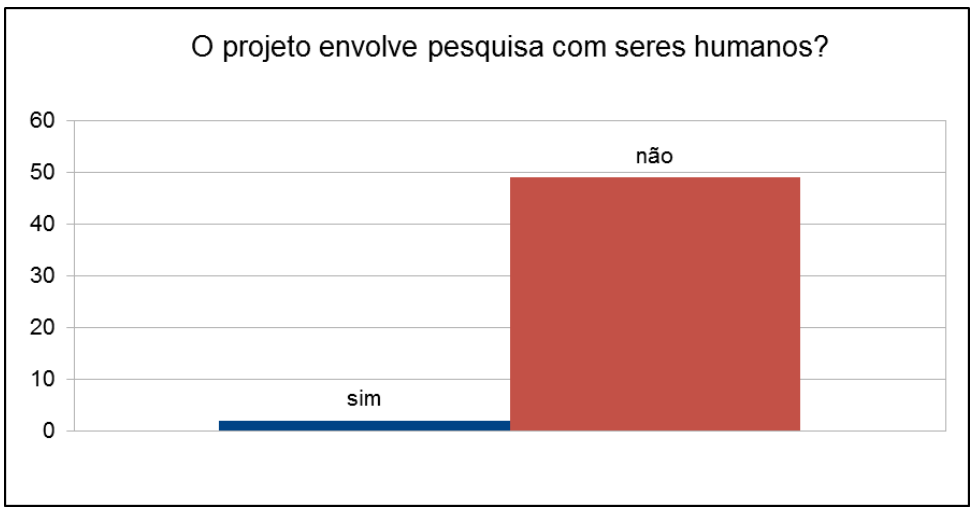
Com relação ao sigilo dos dados gerados pela pesquisa, a maioria dos respondentes informa que os dados são compartilhados internamente com os grupos, sendo que os integrantes destes grupos podem ser externos, pois existem pesquisas desenvolvidas com pesquisadores externos. Esse compartilhamento é feto informalmente, pois não há uma sistemática de tratamento desses dados. Possivelmente os dados sejam abertos a outros pesquisadores, mas muitos não devem conseguir reutilizá-los por causa da ausência de padrões de tratamento. Essa pergunta é importante para nos ajudar a determinar o tipo de preservação a ser feita e nos indicar onde devem ser armazenados esses dados.



Fonte: Dados gerados pela pesquisa (2015)

O ato de documentar consiste em registrar por meio de documentos, e a questão buscou identificar como os procedimentos realizados pelos pesquisadores estão sendo documentados por eles. A maioria dos respondentes declarou que o registro se dá na forma de relatórios, anotações e outros. De novo esclarecemos que não foi solicitado aos respondentes que identificassem o que eles classificam como outros. Mesmo que não tenha sido pedido para definir o tipo de relatório gerado, e sendo a pesquisa realizada a partir de fomentos públicos ou privados, creditamos o maior número de registros nesse formato por consistirem de relatórios que devem ser escritos para prestação de contas para a agência ou empresa que financiou a pesquisa.

Saber se as pesquisas estão sendo documentadas é muito importante porque essa documentação, provavelmente, mostra qual a fase em que os dados são gerados e se eles estão sendo preservados e ou compartilhados.



Fonte: Dados gerados pela pesquisa (2015)

Por fim, foi questionado se o projeto desenvolvido abarca pesquisa com seres humanos, e apenas dois responderam que sim. Mesmo com um número tão pequeno é importante que o pesquisador tenha ciência da necessidade de anonimizar os dados resultantes destas pesquisas.

A partir da elaboração do conjunto de critérios para avaliação e seleção de dados, e de posse das respostas obtidas com a aplicação do questionário, foi realizado um ranqueamento

para se chegar aos projetos escolhidos e que darão início ao processo de curadoria no IEN.

4 Considerações finais

Como o questionário estava ligado a cada pesquisa relatada, foi possível identificar o que cada investigação estava gerando em termos de dados e como esses dados estão sendo tratados. Como não existe política mandatória institucional para depósito de dados, o questionário forneceu subsídio para identificar quais as pesquisas prioritárias em termos de tratamento para preservação, reuso e compartilhamento. Para essa identificação foi criado um conjunto de critérios que permitiu ranquear os projetos prioritários a terem seus dados preservados. Esclarecendo que todos os dados de projetos originados no IEN serão preservados, mas por causa das características singulares de cada pesquisa, alguns dados se tornam mais suscetíveis de serem reproduzidos o que os tornam prioritários no processo de curadoria.

Apesar de a abordagem ter sido adotada com certo sucesso, alguns problemas foram identificados. Um problema já citado é a ausência de uma política institucional que torne mandatório o depósito dos dados de pesquisa no setor responsável pela gestão desses dados. Neste caso, apesar de serem conhecidos os tipos de dados que as pesquisas estão gerando, ainda não foi possível obtê-los para curadoria, o que ainda impede que os pesquisadores reusem e compartilhem seus dados em âmbito institucional e até mesmo externamente.

Buscando preencher essa lacuna, o Grupo de Pesquisa em Gestão do Conhecimento do IEN continua trabalhando no marketing formal e informal para conscientizar os pesquisadores sobre a importância de depositarem seus dados para que eles possam ser tratados, preservados e reusados em pesquisas futuras, agilizando, assim, o andamento das pesquisas.

Com relação ao questionário, foram detectadas algumas questões que precisam de uma reformulação de suas respostas. Por exemplo, ao dar a possibilidade do respondente escolher entre interno/externo/outros com relação ao laboratório onde a pesquisa está sendo desenvolvida, detectamos a necessidade de solicitar que ele identifique qual é esse laboratório externo/outros quando marcar esta opção. Estas falhas serão corrigidas para futuras aplicações do questionário.

Outra questão importante detectada está relacionada à necessidade das instituições de pesquisa, em especial as que financiam as pesquisas, começarem a solicitar um plano de gestão de dados de pesquisa a serem armazenados junto ao setor responsável com os projetos das pesquisas em andamento. Seria uma forma de incentivar a preservação do conhecimento que está sendo desenvolvido e financiado pelas instituições.

Este trabalho visou ser uma primeira contribuição a uma das etapas do processo de curadoria de dados de pesquisa, isto é, a etapa da seleção e da avaliação dos dados. Como o tema curadoria digital de dados de pesquisa ainda é incipiente no Brasil, acredita-se que muitas instituições que, neste momento, começam a lidar com o tema, enfrentarão essa problemática. Outros trabalhos que abordem outras fases da curadoria digital de dados de pesquisa ainda precisam ser desenvolvidos e aplicados futuramente.

O próximo passo será contatar os pesquisadores responsáveis e fazer um acompanhamento mais próximo da pesquisa, de suas etapas metodológicas, identificando a etapa de geração dos dados e auxiliando o pesquisador na elaboração de um plano para gestão de dados de pesquisa.

SELECTION AND EVALUATION OF DIGITAL RESEARCH DATA COLLECTIONS: a possible methodological approach

Abstract

The area of Nuclear Sciences, as well as many other scientific domains, intensively produces a plurality of search data, ranging from results of experiments to data generated from simulations, such as, for example, those arising in virtual reality and intelligence artificial research. This fact has been evidenced in the Instituto de Engenharia Nuclear (IEN), a unit of the Comissão Nacional de Energia Nuclear (CNEN), an agency under the Ministério da Ciência, Tecnologia e Inovação (MCTI) in Brazil. Although the specifics of nuclear area the problem of identification, collection and selection of research data is a challenge that presents itself in any field or institution that wants to start a scientific data curation project. In this sense, a question that arises in this context is how to identify the research data produced within a scientific research institution? The article presents a possible methodological approach applied to the IEN, which can be to replicate in any research institution as a mechanism assist in identifying scientific data. Clarifying that this is only a methodological way possible among others, and indicated its application in institutions in which there is still no data management plan (DMP).

Keywords: Data Collection. Research Data. Data Appraisal

Referências

ABBOT, Mark R. Um novo caminho para a ciência? In: HEY, Tony; TRANSLEY, Stewart; TOLLE, Kristin (orgs). **O quarto paradigma:** descobertas científicas na era da eScience. São Paulo, Oficina de Textos, 2011.

BEAGRIE, Neil. Digital Curation for Science, Digital Libraries, and Individuals. **The International Journal of Digital Curation**, n.1, v.1, 2006.

BEAGRIE, Neil; LAVOIE, Brian; WOOLLARD, Matthew. **Keeping Research Data Safe 2**. London: JISC, 2010. Disponível em: <http://www.webarchive.org.uk/wayback/archive/20140615221405/http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf>. Acesso em: 10 mar 2016.

BELL, Gordon. Prefácio. In: HEY, Tony; TRANSLEY, Stewart; TOLLE, Kristin (orgs). **O quarto paradigma:** descobertas científicas na era da eScience. São Paulo, Oficina de Textos, 2011.

DIGITAL Curation Centre (DCC). **What is digital curation?** Disponível em: <http://www.dcc.ac.uk/digital-curation/what-digital-curation> Acesso em: 10 mar 2016.

FOX, Peter; HENDLER, James. eScience semântica: o significado codificado na próxima geração de ciência digitalmente aprimorada. In: HEY, Tony; TRANSLEY, Stewart; TOLLE, Kristin (orgs). **O quarto paradigma:** descobertas científicas na era da eScience. São Paulo, Oficina de Textos, 2011.

GRAY, Jim. Jim Gray on eScience: A Transformed Scientific Method. Based on the transcript of a talk given by Jim Gray to the NRC-CSTB1 in Mountain View, CA, on January 11, 2007. In: HEY, Tony; TRANSLEY, Stewart; TOLLE, Kristin (orgs). **The Fourth Paradigm. Data-Intensive Scientific Discovery**. Redmond, WA: Microsoft Research, 2009. 284pp.

HARVEY, Ross. Selection and Appraisal. In: ROSS, S.; DAY, M. (ed.). **DCC Digital Curation Manual**, 2007. Disponível em: <http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection> Acesso em: 10 mar 2016.

HARVEY, Ross. **Digital Curation:** a how-to-do-it manual. New York: Neal-Schuman, 2010.

MARQUES, Fabrício. Ciência transparente. **Revista FAPESP**, abril, 2014.

MARTÍNEZ-URIBE, Luis; MACDONALD, Stuart. Un nuevo cometido para los bibliotecarios académicos: data curation. **El profesional de la información**, maio./jun., v.17, n.3, p.273-280, 2008.

PIERCE-MOSES, Richard. **A Glossary of Archival and Records Terminology**. Chicago: Society of American Archivists, 2005.

SALES, Luana Farias. **Integração semântica de publicações científicas e dados de pesquisa:** proposta de modelo de publicação ampliada para a área de ciências nucleares. Rio de Janeiro, 2014. Tese (Doutorado em Ciência da Informação). Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

SAMPAIO, Paulo B. Presentation. **IEN Progress Report**, 2013-2014, n.2, 2015.

SAYÃO, Luis Fernando, SALES, Luana Farias. **Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores**. Rio de Janeiro: CNEN/IEN, 2015.

Sustainable Economics for a Digital Planet. Ensuring Long-Term Access to Digital Information. **Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access**. February 2010.

TOLLE, Kristin; TANSLEY, Stewart; HEY, Tony (orgs). Jim Gray e a eScience: um método científico transformado. In: HEY, Tony; TRANSLEY, Stewart; TOLLE, Kristin (orgs). **O quarto paradigma:** descobertas científicas na era da eScience. São Paulo, Oficina de Textos, 2011.

U.S. National Archives and Records Administration. **Strategic Directions:** Appraisal Policy. 15 maio 2007. Disponível em: <http://www.archives.gov/records-mgmt/initiatives/appraisal.html> Acesso em: 10 mar 2016.