RELATÓRIO FINAL
PILOTO DE REPOSITÓRIO DE DADOS DE PESQUISA
FASE 1

Parceria para o Governo Aberto (OGP)

FIOCRUZ ibict RNP REDE NACIONAL DE ENSINO E PESQUISA

Maio de 2020

RELATÓRIO FINAL
PILOTO DE REPOSITÓRIO DE DADOS DE PESQUISA (FASE 1)

Coordenação do piloto na Fiocruz: Coordenação de Informação e Comunicação (CINCO) da Vice-presidência de Ensino, Informação e Comunicação da Fundação Oswaldo Cruz (VPEIC/Fiocruz)

Maio de 2020

SUMÁRIO

1. INTRODUÇÃO

Um importante capítulo da história da Fundação Oswaldo Cruz (Fiocruz), no âmbito da Ciência Aberta, começa em 2017, quando a instituição inicia seus esforços pela implementação de uma política de gestão, compartilhamento e abertura de dados para pesquisa, que envolve uma série de estratégias e ações para estruturar processos científicos mais colaborativos e transparentes. Como expressão concreta da relevância dessa iniciativa, que vem se consolidando ao longo dos anos, encontra-se a proposta de estabelecer uma infraestrutura tecnológica entre as estratégias planejadas. Assim, como desdobramento no contexto das estratégias de implementação, a Fiocruz tem se articulado com outras instituições e organizações nacionais e internacionais para desenvolver e estimular o avanço da Ciência Aberta, considerando a importância da temática no atual contexto da sociedade.

Nesta perspectiva, a Fiocruz está participando do Compromisso 3, que tem como proposta "estabelecer mecanismos de governança de dados científicos para o avanço da Ciência Aberta no Brasil", do 4º Plano de Ação Nacional para Governo Aberto[1], criado em 2018. Este Plano foi desenvolvido no âmbito da Parceria para o Governo Aberto (do inglês Open Government Partnership - OGP), e é composto por nove Marcos[2], com atuação de diversos atores nacionais, tendo a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) como Coordenadora deste Compromisso, e a participação do Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC), do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), da Rede Nacional de Ensino e Pesquisa (RNP), da Universidade de Brasília (UnB), da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Associação Brasileira de Editores Científicos (ABEC), do Scientific Electronic Library Online (SciELO), da Comissão Nacional de Energia Nuclear (CNEN) e da Open Knowledge Brasil (OKBR).

Como parte do Compromisso 3, no dia 24 de junho de 2019, a Fiocruz foi convidada pela RNP e pelo IBICT como um dos estudos de caso na implantação de uma infraestrutura

[1] 4º Plano de Ação Nacional para Governo Aberto 4º Plano de Ação Nacional para Governo Aberto - Disponível: <https://governoaberto.cgu.gov.br/a-ogp/planos-de-acao/4o-plano-de-acao-brasileiro/4o-plano-de-acao-nacional_portugues.pdf

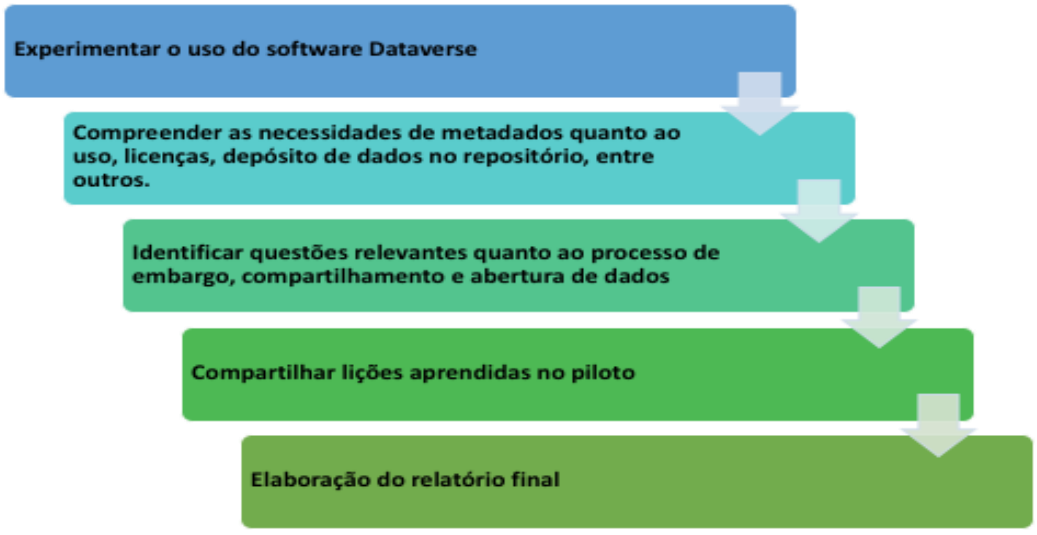
[2] Marcos do Compromisso 3: 01 - Implantação de uma rede interinstitucional pela Ciência Aberta; 02 - Realização de diagnóstico nacional e internacional da Ciência Aberta; 03 - Definição de diretrizes e princípios para políticas institucionais de apoio à Ciência Aberta; 04 - Promoção de ações de sensibilização, participação e capacitação em Ciência Aberta, 05 - Articulação com agências de fomento para a implantação de ações de apoio à Ciência Aberta; 06 - Articulação com editores científicos para a implantação de ações em apoio à Ciência Aberta; 07 - Implantação de infraestrutura federada piloto de repositórios de dados de pesquisa; 08 - Proposição de padrões de interoperabilidade para repositórios de dados de pesquisa; 09 - Proposição de conjunto de indicadores para aferição da maturidade em Ciência Aberta.

federada piloto de repositórios de dados de pesquisa no Brasil, aliado à proposição de desenvolver um modelo de padrões de interoperabilidade entre repositórios de dados de pesquisa. O software escolhido para o piloto de repositórios de dados de pesquisa foi o Dataverse[3], após estudo realizado pelo IBICT e RNP, em parceria com o Grupo de Trabalho em Rede de Dados de Pesquisa (GT-RDP). No estudo intitulado "Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas para compartilhamento de dados no Brasil: relatório"[4], investigou-se o uso de Dataverse e DSpace no desenvolvimento de repositório de acesso aberto a dados de pesquisa, através de 56 critérios, classificados em: ambiente do repositório (6), conjuntos de dados (6), descrição e documentação (11), produção de conjuntos de dados (7), armazenamento a longo prazo (5), acesso e uso (15) e desenvolvimento e manutenção do software (6). Esses critérios foram estruturados com base no modelo OAIS.

Neste contexto, entendemos que a cooperação técnica estabelecida é uma oportunidade que permite acelerar processos de troca do conhecimento graças ao compartilhamento de experiências e ao enriquecimento feito pelos atores que participam das ações. Além disso, fortalece processos internos nas instituições, proporcionando, assim, aportes de coletivos organizacionais.

Na execução do piloto FIOCRUZ-RNP-IBICT, foi desenvolvido um Plano de Trabalho com objetivos específicos e estratégicos, assim como as iniciativas mapeadas em diversos fóruns de engajamento de estudo. A figura 1 a seguir ilustra 05 (cinco) etapas do plano de trabalho.

Figura 1 – Etapas do Plano de Trabalho



[3] The Dataverse Project – Disponível em: <https://dataverse.org>

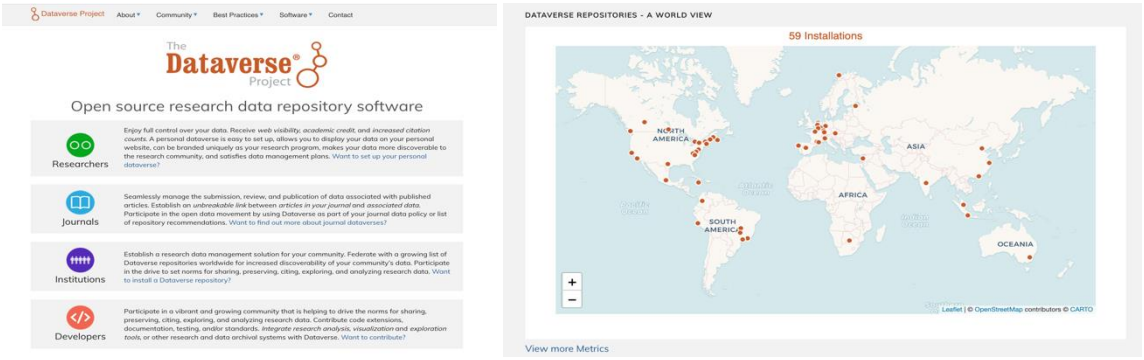
[4] ROCHA, Rafael Port et al. Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas: relatório 2018. Disponível em: <http://hdl.handle.net/10183/185126>.

As ações e testes propostos para alcance dos objetivos foram desenvolvidos por equipe interna da Fiocruz, com o apoio de profissionais do IBICT e RNP (item 3.1 deste relatório) entre os meses de agosto de 2019 e março de 2020. As experiências foram registradas e analisadas, possibilitando a elaboração deste relatório, que está organizado em: detalhamento da plataforma metodologia do trabalho, descrição dos testes, avaliação, recomendações e desafios para Fase 2 do piloto. Considera-se que os resultados apresentados no relatório podem servir de fonte para tomada de decisão, bem como memória do projeto.

2. DETALHAMENTO DA PLATAFORMA DATAVERSE

O Dataverse é um software de código aberto, desenvolvido pelo Instituto de Ciências Sociais Quantitativas de Harvard (IQSS), com colaboradores em todo o mundo (Figura 2). Permite armazenar, compartilhar, publicar, citar, explorar e analisar dados de pesquisa. Facilita a disponibilização de dados para outras pessoas e permite replicar o trabalho de outras pessoas com mais facilidade. Pesquisadores, periódicos, autores de dados, editores, distribuidores de dados e instituições afiliadas recebem crédito acadêmico e visibilidade na web. Um repositório do Dataverse é a instalação do software, que hospeda vários arquivos virtuais chamados *dataverses*. Cada *dataverse* contém conjuntos de *datasets*, e cada *dataset* contém metadados e arquivos de dados descritivos (incluindo documentação e código que acompanham os dados). Como método de organização, os *dataverses* também podem conter outros *dataverses*.

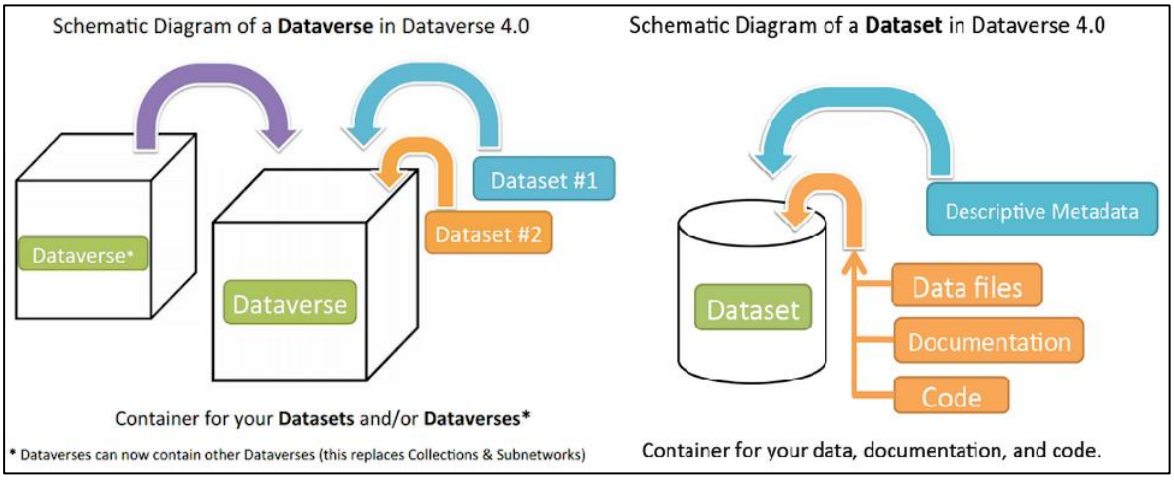
Figura 2 – The Dataverse Project



O Harvard Dataverse é um repositório de dados de pesquisa que contempla 59 instalações do software em diversos países do mundo[5]. Existem inúmeras possibilidades de uso do Dataverse como o uso individual, por um pesquisador, ou usos coletivos, em projetos e grupos de pesquisa, departamentos, periódicos ou organizações.

[5] Dados disponíveis no site do Dataverse em 10 de maio de 2020.

Figura 3 - Diagrama semântico do Dataverse e Dataset no Dataverse 4.0



Fonte: Adaptado de User guide do Dataverse (http://guides.dataverse.org/en/latest/user/index.html)

3. METODOLOGIA DO TRABALHO

Em 14 de agosto de 2019, no Rio de Janeiro, foi realizada a 1ª (primeira) Reunião do Piloto de Repositório de Dados de Pesquisa (Fase1) a fim de apresentar os objetivos propostos e discutir a metodologia a ser utilizada para atingir as metas propostas. Como coordenadora-geral do piloto, a Vice-presidente de Educação, Comunicação e Informação da Fiocruz (VPEIC/Fiocruz), Drª Cristiani Machado, conduziu um intercâmbio prolífero na discussão das principais dimensões da proposta. Também se pôde assinalar a pertinência do seguimento pontual das distintas etapas do piloto, e, em especial, para a realização de reuniões de acompanhamento periódicas com grupo organizado de acordo com competências técnicas específicas.

O objetivo de uma 2ª (segunda) reunião, ocorrida no dia 13 de setembro de 2019, consistiu na definição da responsabilidade da Coordenação-Geral de Gestão de Tecnologia de Informação (COGETIC/Fiocruz) na instalação do software Dataverse e a constituição do Grupo de Trabalho (GT), através da Portaria da Presidência da Fiocruz n° 6399/2019[6], de 11 de novembro de 2018. Apontou-se, ademais, a relevância em elaborar e executar o Plano de Trabalho, observando o escopo, objetivos, principais atividades, prazos e responsáveis na Fiocruz.

Dando continuidade às ações, em reunião específica com os 02 (dois) representantes dos pesquisadores selecionados para o piloto, foi realizada uma avaliação final sobre o uso dos dados das pesquisas coordenadas por eles, e concluiu-se que não seria indicada e apropriada a utilização dos dados, já que os mesmos ainda estavam sendo coletados e seriam analisados nas diversas pesquisas em curso. Visando, assim, garantir a proteção destas pesquisas, decidiu-se pela busca de dados de pesquisas desenvolvidas pela Fiocruz, publicados e disponíveis, com licenças favoráveis ao uso por qualquer pessoa interessada, em 02 (dois) repositórios multidisciplinares, o Zenodo e o Figshare.

Após esta decisão, foram realizadas reuniões entre Fiocruz, RNP e IBICT, que aconteceram nos dias 22 de outubro, 12 de novembro, 5 de dezembro de 2019, e em 2020, nos dias 6 de fevereiro, 09 e 16 de março. Nas reuniões discutiu-se, com os membros do GT, o andamento dos testes e se havia a troca de informações e experiências entre os profissionais, com encaminhamentos acordados entre os participantes. As reuniões aconteceram nas

[6] Portaria da Presidência da Fiocruz n° 6399/2019. Disponível em: <http://www.portaria.Fiocruz.br/Doc/P6399_2019.pdf>
Portaria instituiu o Grupo de Trabalho para participar de projeto piloto de uma infraestrutura tecnológica federada de repositórios de dados de pesquisa no Brasil, para o desenvolvimento de um modelo de padrões de interoperabilidade entre repositórios de dados de pesquisa, em parceria com Rede Nacional de Ensino e Pesquisa (RNP) e o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT).

instalações da Fiocruz, em Manguinhos, com endereço em sala virtual[7], para participação do IBICT e outros especialistas do GT-RDP.

As questões tratadas durante as reuniões do piloto serviram como referência na elaboração do "Manual do Dataverse", produzido pela equipe do IBICT[8].

3.1 Grupo de trabalho

O GT do Piloto OGP foi formado considerando as seguintes premissas:

1. O papel institucional da VPEIC/Fiocruz na condução das estratégias para gestão, compartilhamento e abertura de dados para pesquisa da Fiocruz e a representação institucional no Compromisso 3 do 4º Plano Nacional de Governo Aberto;
2. O papel institucional da COGETIC/Fiocruz e a responsabilidade de prover serviços de infraestrutura tecnológica e segurança da informação para sistemas e rede Fiocruz;
3. O papel institucional do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) e sua experiência na coordenação da Rede de Bibliotecas da Fiocruz e do Repositório Institucional Arca[9];
4. O papel institucional da Casa de Oswaldo Cruz (COC) e sua experiência na coordenação do Sistema de Gestão de Documentos Arquivísticos (SIGDA) e do Preservo – Complexo de Acervos da Fiocruz e no desenvolvimento do Programa de Preservação Digital de Acervos Científicos e Culturais da Fiocruz;
5. Pesquisadores da Fiocruz selecionados pela experiência vasta em compartilhamento e abertura de dados de pesquisa mapeada em visitas às unidades Fiocruz no período de mobilização da comunidade Fiocruz para discussão do "Termo de referência: gestão e abertura de dados para pesquisa na Fiocruz"[10];
6. Representantes institucionais do IBICT e RNP, responsáveis por conduzir as ações do estudo de caso Fiocruz, com o objetivo de implantação de uma infraestrutura federada piloto de repositórios de dados de pesquisa no Brasil, vinculado à proposição de desenvolver um modelo de padrões de interoperabilidade entre repositórios de dados de pesquisa no âmbito da Parceria para Governo Aberto.

[7] Sala de Reuniões: GT em Rede de Dados de Pesquisa Brasileira. Disponível em: <https://conferenciaweb.rnp.br/webconf/gt-rdp>

[8] PAGANINI, N. Lucas; BARRETO NETO, Vanderlino. **Manual do Dataverse**. Brasília: IBICT, 2019.

[9] O Repositório Institucional Arca foi lançado em abril de 2011. Reúne e disponibiliza a produção intelectual da Fiocruz de forma ampla, em consonância com o movimento de acesso aberto à informação científica. Disponível em: < https://www.arca.Fiocruz.br>

[10] **Termo de referência: gestão e abertura de dados para pesquisa na Fiocruz**. Disponível em: <https://www.arca.Fiocruz.br/handle/icict/26803>

Os nomes dos componentes do GT do Piloto OGP - profissionais indicados pelos diretores das instituições e unidades para participar das atividades - estão descritos no quadro a seguir:

Quadro 1 – Componentes do GT

Nomes indicados	Instituição/Lotação
Carolina Felicissimo	DPD/RNP
Claudete Fernandes	Fiocruz/Icict
Erick Penedo	Fiocruz /Icict
Fátima Martins	Fiocruz /VPEIC
Hataânderson Santos	Fiocruz /VPEIC
Ivone Sá	Fiocruz /COC
Karina Veras	Fiocruz /COC
Marcus Vinicius Pereira da Silva	Fiocruz /COC
Misael Araujo	Fiocruz /Cogetic
Simone Dib	Fiocruz /Icict
Thiago Carelli	Fiocruz /Cogetic
Vanderlino Barreto Coelho Neto	IBICT/CNEN
Vanessa Jorge	Fiocruz/VPEIC
Vinicius Belchior	Fiocruz/Icict
Viviane Veiga	Fiocruz/Icict
Washington Segundo	IBICT

Como descrito no item 3, em setembro de 2019, decidiu-se não utilizar os dados dos pesquisadores convidados para o piloto OGP, já que estes dados ainda não estavam públicos.

4. DESCRIÇÃO DOS TESTES

Durante o piloto, foram realizados diversos testes no software Dataverse. Para realizar o trabalho, a COGETIC instalou o software na rede interna da Fiocruz e foram configurados perfil e senha para testes dos participantes. O endereço para os testes foram os seguintes:

Endereço: https://dadosdepesquisa-beta.Fiocruz.br/

Vale ressaltar que o acesso ao repositório foi restrito à rede interna da Fiocruz, e participantes externos tiveram credenciais de acesso via rede privada virtual (Virtual Private Network - VPN), da Fiocruz. A partir destas configurações iniciais de disponibilização do ambiente de software para a realização dos testes, iniciaram-se os trabalhos, como descrito a seguir.

4.1 Teste de cadastro e importação dos dados

Essa etapa foi iniciada após a instalação do software Dataverse. Posteriormente, elencou-se 05 (cinco) ações, a saber:

a) Criação da comunidade (dataverse) raiz denominada Fiocruz;

b) Criação de sub-comunidades (*dataverses* dentro do Dataverse raiz Fiocruz), conforme as unidades organizacionais;

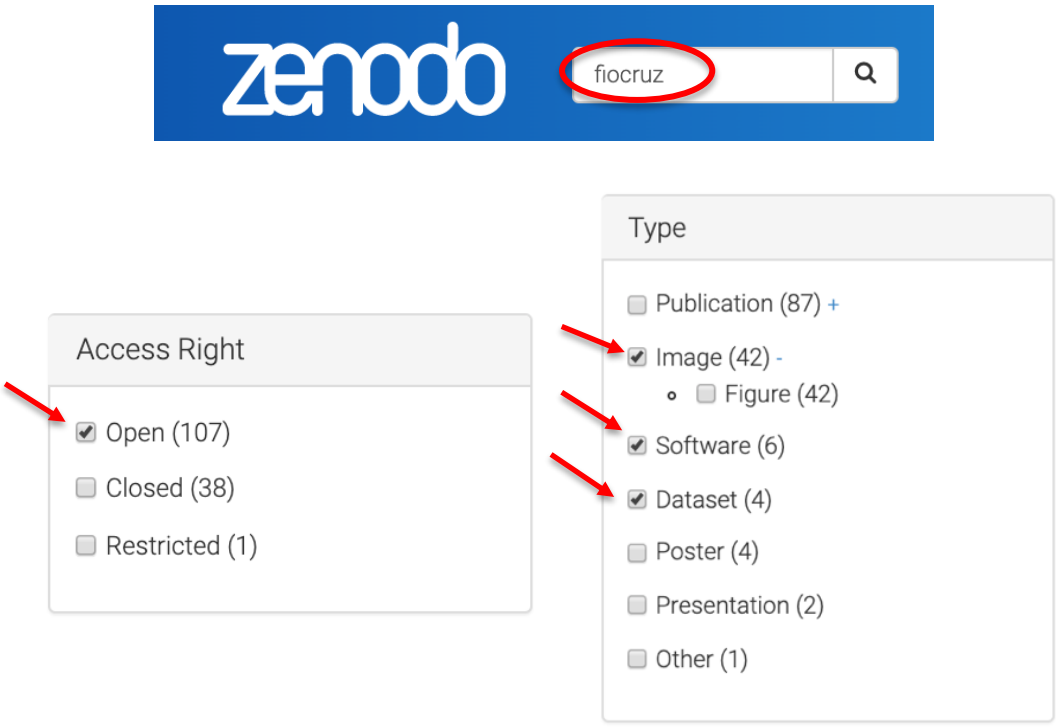
c) Migração do conjunto de dados que estavam armazenados no Repositório Institucional Arca;

d) Cadastro dos dados publicados por pesquisadores da Fiocruz, disponibilizados nos repositórios de dados: Zenodo (indicado pelo programa H2020 da Europa) e Figshare;

e) Realização de testes de inclusão, edição e exclusão de dataverses, *datasets* e arquivos, o que permitiu a análise do software de forma sistêmica.

Cabe destacar que, no repositório de dados Zenodo na elaboração da estratégia de busca, foram aplicados 02 (dois) filtros principais: "Fiocruz" e "oswaldo cruz". No filtro "Fiocruz", seguiu-se o protocolo descrito abaixo.

Figura 4 - Exemplo realizado para o filtro "Fiocruz" no Zenodo



Conforme imagem acima, na guia [*Access Right*], selecionamos a opção "*Open*"; na guia [*Type*] "*image*", "*software*", e, por último, "*dataset*", resultando em 31 arquivos que foram posteriormente importados.

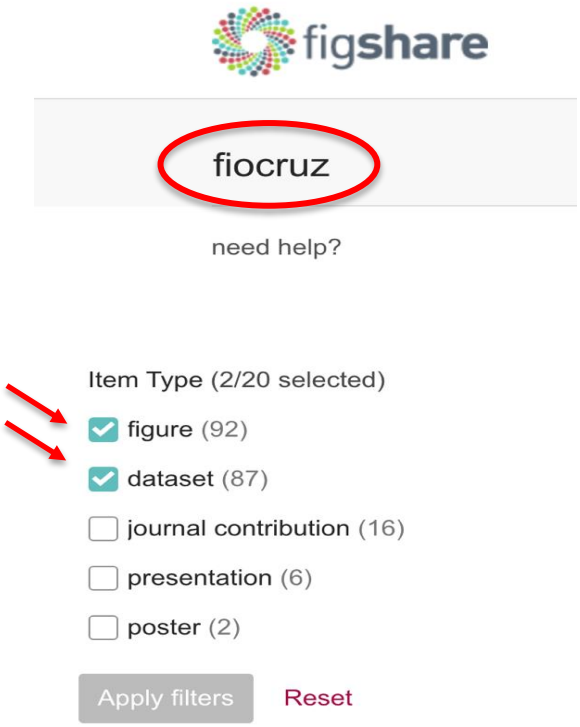
Para utilizar o filtro "oswaldo cruz", foram aplicados os mesmos filtros. Dessa forma, selecionamos na guia [*Access Right*], a opção "*Open*", e na guia [*Type*] selecionamos "*image*", e, posteriormente, "*dataset*", resultando na extração de 54 *datasets*. O volume final de *datasets* cadastrados totalizou 85 itens (Quadro 2).

Quadro 2 – Itens recuperados e adicionados do Zenodo

Filtro	Guia (Type)	Guia (Access Right)	Total adicionados no Dataverse
“Fiocruz”	*“Image”*	*“Open”*	26
“Fiocruz”	*“Software”*	*“Open”*	02
“Fiocruz”	*“Dataset”*	*“Open”*	03
“oswaldo cruz”	*“Image”*	*“Open”*	45
“oswaldo cruz”	*“Dataset”*	*“Open”*	09

Do repositório de dados Figshare foram cadastrados 152 *datasets* no Dataverse, utilizando também 02 (dois) filtros principais “Fiocruz” e “oswaldo cruz”, da seguinte forma:

Figura 5 - Exemplo realizado para o filtro “Fiocruz” no FigsharE



Quadro 3 – Itens recuperados e adicionados do Fighshare

Filtro	Guia (Type)	Guia (Licence)	Total	Total adicionados no Dataverse*
“Fiocruz”	*“Figure”*	CC BY 4.0.	84	35
“Fiocruz”	*“Dataset”*	CC BY 4.0.	85	59
“oswaldo cruz”	*“Figure”*	CC BY 4.0.	50	31
“oswaldo cruz”	*“Dataset”*	CC BY 4.0.	55	27

Nos casos em que não havia o registro de afiliação na Fiocruz, os arquivos de dados foram descartados. Também não foram importados os resultados de busca em que os dados já haviam sido adicionados através de outro filtro, como, por exemplo, quando o mesmo arquivo aparece tanto no filtro “Fiocruz” quanto no filtro “oswaldo cruz”.

A etapa de testes de cadastros auxiliou na identificação das funcionalidades do Dataverse como um todo. Destaca-se a importância desses testes de cadastro para a análise de metadados genéricos e específicos disponíveis; para o estudo dos perfis e das licenças apresentadas no software; para a compreensão do passo a passo necessário para a realização das tarefas no Dataverse (como a definição de restrição de acesso aos conjuntos de dados); para entender possíveis interações entre o software e outros repositórios de dados, entre outros.

A expertise adquirida com esta ação possibilitou reconhecer vantagens e desvantagens do Dataverse, além de elaborar os fluxos de trabalho que auxiliarão o pesquisador na utilização do software e de construir uma base sólida de conhecimentos que será importante para a implementação da fase 2 do piloto.

Outro importante destaque sobre os testes de cadastro realizados foi o entendimento quanto aos tipos de informações que serão solicitadas aos pesquisadores no momento do cadastro, possibilitando, assim, uma visão prática dos processos de trabalho pelos profissionais que serão responsáveis pela capacitação e divulgação do repositório.

Sobre a importação de *datasets* de outros repositórios para o Dataverse, verificou-se através de documentação do software que é possível realizar a importação de *datasets* por meio de arquivo nos formatos JSON e DDI via API. Cada arquivo possui sua estrutura própria, exigindo, com isto, as informações sobre os metadados e os arquivos para a criação dos *datasets*. Também pela API, é possível publicar *datasets*, bem como exportar os metadados em formatos como DDI, JSON, OAI_DDI, DCTERMS, Datacite, entre outros.

Também se verificou que é possível mover um *dataset* para outro *subdataverse,* por meio de uma interface Web, disponibilizada pelo *dashboard* do super-administrador (TI) do Dataverse, sem a necessidade de recorrer à programação via APIs.

4.2 Teste de link de datasets

Para a execução dos objetivos, foram realizados também os testes de links – referências – dos *datasets*, em *dataverses* distintos. Após cadastro e o salvamento do *dataset*, clica-se no botão "link", em seguida, na janela "link *dataset*", para inserção do nome de um outro *dataverse*. Repetimos esse procedimento e salvamos, tornando esse *dataset* com links em vários *dataverses*.

Instituído o fluxo acima, configurou-se algumas observações. Destacamos:

a) Não visualizamos a quantidade de links feitos, bem como outros *dataverses* foram gerados a partir do *dataset* original (DESVANTAGEM);

b) O *dataset* com links em outros *dataverses* é modificado quando é feita atualização (VANTAGEM);

c) O Dataverse não permite que o próprio usuário remova ou altere o link de um dataset. Somente o administrador do sistema pode realizar essa ação internamente.

4.3 Teste de metadados

Sobre a avaliação dos metadados que figuram na instalação, foram analisados os tipos disponíveis no padrão do Dataverse. O padrão de metadados utilizado no Dataverse é o Data Documentation Initiative (DDI). Foram identificados diversos formatos de citação, campos de ajuda, termos de uso; metadados autoria, *subject*, *keywords*, *mopic classification*, *related publication*, entre outros. Constatamos que o Dataverse permite a utilização de metadados específicos, conforme demanda de cada área do conhecimento e/ou necessidade institucional, sendo necessário solicitar inclusão desses metadados específicos ao administrador do sistema.

4.4 Teste de tamanho de arquivos permitidos

Conforme os testes relacionados sobre o tamanho de arquivos permitidos, identificou-se que o Dataverse permite, em sua instalação padrão, o armazenamento de arquivos de até 120mb. Arquivos superiores a esse tamanho apresentam erro no momento do *upload*.

Segundo a documentação do Dataverse, o limite de tamanho de arquivos é configurado pelo administrador do sistema.

4.5 Teste de perfis

O perfil é o papel que o usuário irá exercer em determinado *dataset* ou *dataverse*. A permissão que cada perfil possui estabelece diferentes níveis de autorização que o usuário recebe para realizar determinadas ações. Cada permissão está relacionada com um perfil específico.

Identificamos que o usuário pode ter um determinado perfil relacionado ao Dataverse e outro perfil para *subdataverses* ou *datasets,* com permissões diferentes.

Ao se atribuir um perfil a um usuário com permissão para determinado dataverse, isto não significa que este usuário irá receber automaticamente a mesma permissão para os *subdataverses*, sendo necessário atribuir as permissões de perfil para cada dataverse e para cada usuário.

Tipos de perfis existentes no Dataverse e ações permitidas:

a) Admin: adicionar *dataverse*, adicionar *dataset*, ver *dataverse* não publicado, ver *dataset* não publicado, baixar arquivo, editar *dataverse*, editar *dataset*, gerenciar permissões do *dataverse*, gerenciar permissões do *dataset*, publicar *dataverse*, publicar *dataset*, deletar *dataverse* e deletar rascunho de *dataset*.

b) Contribuidor: ver *dataset* não publicado, baixar arquivo, editar *dataset* e deletar rascunho do *dataset*. Esse perfil não pode alterar *dataset* já publicado.

c) Curador: adicionar *dataverse*, adicionar *dataset*, ver *dataverse* não publicado, ver *dataset* não publicado, baixar arquivo, editar *dataset*, gerenciar permissões do *dataset*, publicar *dataset* e deletar rascunho do *dataset*.

d) Criador de *dataset*: adicionar *dataset*.

e) Criador de *dataverse* + *dataset*: adicionar *dataverse* e adicionar *dataset*.

f) Criador de *dataverse*: adicionar *dataverse*.

g) Download de Arquivo: transferência de arquivo.

h) Membro: ver *dataverse* não publicado, ver *dataset* não publicado e baixar arquivo.

4.6 Teste de publicação e embargo

Os testes de publicação foram realizados através da publicação de *datasets* e modificando metadados e/ou arquivos que geraram um versionamento. Para publicar, o usuário deve cadastrar metadados e importar arquivos, se necessário. Na maioria dos casos, a publicação pode ser compreendida como um processo que disponibiliza o acesso aos metadados e os arquivos de uma pesquisa. Este processo exige um conjunto de identificadores persistentes digitais (Digital Object Identifier - DOI), que durante os testes do piloto ainda não tinham sido adquiridos. Para testar a publicação de dados, o IBICT disponibilizou um gatilho temporário que possibilitava a publicação, sem DOI, de forma a permitir os testes planejados. Em uma nova fase de testes, é necessário realizar o teste com identificadores persistentes como DOI's válidos.

Ressalta-se que, antes da publicação, o pesquisador pode definir possíveis restrições aos metadados, arquivos e/ou *datasets*, como descrito no item a seguir (4.7). Também existe a possibilidade de embargo total dos dados, logo antes da publicação. Em um *dataset* publicado, mesmo com restrições de acesso, os metadados ficam disponíveis e apenas os arquivos ficam restritos. É importante destacar que a publicação dos metadados é indicada como padrão, pois esta ação pode resguardar os arquivos e/ou *datasets* que por motivos legítimos necessitem de proteção, dando visibilidade à existência dos dados para possíveis colaborações e reúso. O fluxo dos processos de depósito e publicação dos dados deve prever esta ação.

4.7 Teste de restrição de arquivos (acesso restrito)

Também foram realizados testes de restrição de acesso a metadados e arquivos. É possível restringir o acesso marcando quais arquivos serão restritos.

Após solicitar acesso, o responsável recebe uma notificação e e-mail de solicitação de acesso, podendo ser autorizado ou negado. O solicitante receberá uma notificação e um e-mail indicando se o acesso foi liberado ou rejeitado.

Outra forma de restringir acesso ao *dataset* ou aos dados é por meio de um Livro de Visitas (*guestbook*), onde aquele que desejar sua utilização terá que responder questões definidas. O livro de visitas pode ser usado com a solicitação de acesso aos arquivos.

4.8 Teste de remoção de dataset

Outro teste realizado está relacionado à publicação e remoção de um *dataset*. Um *dataset* removido tem seu *status* alterado para "*Deaccessioned*". Quando é realizado o "desacesso", o Dataverse informa que o *dataset* não será público, e solicita o motivo da remoção. O

responsável pelo *dataset* continuará tendo acesso a ele, e, caso necessário, poderá republicá-lo.

4.9 Teste de segurança do Dataverse

O teste de segurança é responsável por verificar se todos os recursos de defesa do sistema realmente impedem os possíveis acessos indevidos, assim como mapear e solucionar problemas. O seu objetivo é garantir que o sistema continue funcionando normalmente, sob tentativas de acessos não permitidos.

Dentre outros objetivos, foram realizados diversos testes de segurança pela COGETIC em conjunto com uma consultoria externa especializada contratada. Desta forma, foi gerado um relatório completo dos testes que apontam, principalmente, para 08 (oito) itens críticos, sendo 07 (sete) de criticidade alta e 01 (um) de criticidade média. O relatório completo foi disponibilizado para todos os integrantes do GT do Piloto OGP.

Quadro 4 – Resultado do teste de segurança

Vulnerabilidade	Nível de impacto
Ausência de controle de segurança possibilita ataques de força bruta	Alto
Possibilidade de contorno do mecanismo de bloqueio a ataques automatizados	Alto
Versão desatualizada do apache detectada	Alto
Ausência de controle de segurança possibilita ataque password spraying	Alto
Aplicação possui política frágil de senhas	Alto
Atributo de segurança do *cookie* de sessão desabilitado	Alto
Servidor não força o uso de SSL/TLS	Alto
Ausência de controle de segurança possibilita ataques de roubo de cliques do usuário	Médio

Posteriormente, foi realizada reunião técnica entre COGETIC, RNP, IBICT e membros do GT-RDP, em 07/04/2020, para avaliação das falhas apresentadas no relatório do teste de invasão e discussão de possíveis alternativas para correção das falhas.

Algumas das questões críticas seriam resolvidas por soluções já pensadas e testadas no piloto, como a autenticação institucional no Dataverse com *Shibboleth*.

4.10 Teste de autenticação institucional via Shibbolet

Em 05/12/2019, foi realizada a primeira *webconferência,* em resposta ao ticket de pedido de suporte da COGETIC, na implantação da autenticação *Shiboleth* no Dataverse da FIOCRUZ. A equipe GIdLab[11], especializada em gestão de identidade da RNP, realizou o atendimento. Na primeira reunião, foram apresentados conceitos básicos de Gestão de Identidades, SAML, Frameworks, GIdLab e Dataverse. Após a reunião, a equipe do GIdLab enviou o passo-a-passo para implementar a autenticação no Dataverse, utilizando *Shibboleth,* e ficou à disposição para responder eventuais dúvidas. A próxima reunião de acompanhamento ficou pré-agendada para o dia 19/12/2019, para serem tratados os aspectos de autorização, sendo que essa reunião nunca aconteceu. Depois da interação inicial, foi percebida uma falta de priorização da FIOCRUZ para o atendimento. Foram feitas tentativas para remarcação de uma nova reunião. As próximas datas foram 07 e 16 de janeiro de 2020, mas em nenhuma delas o desenvolvedor escalado para a implementação na FIOCRUZ apareceu, sendo as reuniões desmarcadas nos dias previstos. Em fevereiro de 2020, as equipes do GIdLab e RNP foram informadas pela COGETIC que o desenvolvedor da FIOCRUZ tinha deixado a instituição.

Em 20 de março de 2020, foi realizada uma conversa da equipe do GIdLab com a nova equipe da FIOCRUZ escalada para o atendimento. Nessa conversa, foram tiradas dúvidas básicas que a equipe da FIOCRUZ estava com o passo-a-passo do *Shibboleth,* e a equipe do GIdLab realizou as alterações necessárias para a configuração do ambiente. Novamente, a webconferência para apresentação dos conceitos básicos em Gestão de Identidades, SAML, Frameworks, GIdLab e Dataverse para a nova equipe escalada da FIOCRUZ foi marcada e realizada no dia 25 de março de 2020, de forma a passar o conhecimento para a FIOCRUZ do que foi implementado pela equipe do GIdLab. Nessa reunião, foi demonstrada a relação de confiança estabelecida entre o Dataverse da FIOCRUZ com o ambiente de experimentação em Gestão de Identidade do GIdLab, denominado CAFeExpresso, comprovando a implementação do *Shibboleth* no Dataverse da FIOCRUZ.

Em 08 de abril de 2020, um novo ticket foi submetido pela equipe da FIOCRUZ, pedindo suporte do GIdLab na verificação do IdP da FIOCRUZ para a categoria R&S (*Research and Scholarschip*), necessário pelo Dataverse. Tal atendimento foi realizado e o IdP da FIOCRUZ deve liberar os atributos da categoria R&S. Depois de aprovado, a autenticação deve funcionar com as contas institucionais.

4.11 Teste de interoperabilidade com Archivematica

A ação de preservação digital de dados para pesquisa deve ter como objetivo tanto a garantia de acesso a longo prazo desses objetos digitais como a manutenção de seu significado, estrutura e autenticidade. Conforme definido pelo Programa de Preservação

[11] https://gidlab.rnp.br/

Digital dos Acervos Científicos e Culturais da Fiocruz em desenvolvimento, as estratégias preservacionistas adotadas devem seguir o modelo de referência OAIS (*Open Archival Information System*), aprovado como padrão ISO 14721:2003, que descreve um esquema conceitual para repositório digital genérico, aberto a todas as comunidades, com garantias de confiabilidade, que disciplina e orienta um sistema de arquivo dedicado a preservar e manter o acesso à informação digital por longo prazo.

A análise das funcionalidades do Dataverse, no que tange à preservação dos dados de pesquisa, identificou limitações técnicas que inviabilizam o uso desse software como ferramenta operacional para a execução integral da preservação digital dos dados que nele são inseridos. Isto porque, muito embora esse sistema tenha sido desenvolvido com base no OAIS, suas funcionalidades estão voltadas ao atendimento das responsabilidades inerentes ao produtor.

A clareza das limitações técnicas do Dataverse, quanto à preservação dos dados de pesquisa, levou ao desenvolvimento de sua integração nativa com um software de preservação digital também desenvolvido com base no modelo de referência OAIS, o Archivematica, a partir da versão 1.8, em novembro de 2018.

Aponta-se que o Archivematica é o software gratuito de código aberto adotado pelo Preservo – Complexo de Acervos da Fiocruz, em 2018, como solução tecnológica para preservação digital do patrimônio científico e cultural desta Fundação. Assim sendo, recomenda-se que a preservação digital dos dados de pesquisa da Fiocruz seja realizada com uso do software Archivematica, em interoperabilização direta com o Dataverse.

Entretanto, destaca-se que, até o término deste piloto, os metadados que foram inseridos no pacote SIP para o Archivematica (criado e enviado automaticamente para o software de preservação, visto à interoperabilidade) são, exclusivamente, os descritivos cadastrados no Dataverse. Verifica-se a necessidade de cadastrar novos metadados que atendam à presunção de autenticidade (pelo controle dos processos de gestão) no Dataverse, e avaliar se eles seriam exportados junto aos pacotes SIP que submeteremos à admissão no Archivematica. Essa fase ainda não foi testada porque ainda não foram definidos quais seriam os metadados adicionais inseridos no Dataverse.

4.12 Teste de coleta de dados para OASISBR

Um dos propósitos da infraestrutura federada de repositório de dados de pesquisa é a coleta de metadados dos repositórios institucionais para o metabuscador brasileiro, coordenado pelo IBICT, o OASISBR[12], que passará a coletar, também, para os metadados dos dados de pesquisa.

[12] O Portal brasileiro de publicações científicas em acesso aberto - oasisbr é um mecanismo de busca multidisciplinar que permite o acesso gratuito à produção científica de autores vinculados a universidades e

Neste contexto, o IBICT realizou testes de coleta de metadados do repositório Dataverse e, também, o arquivo de transformação de metadados, para atender o OpenAire 4.

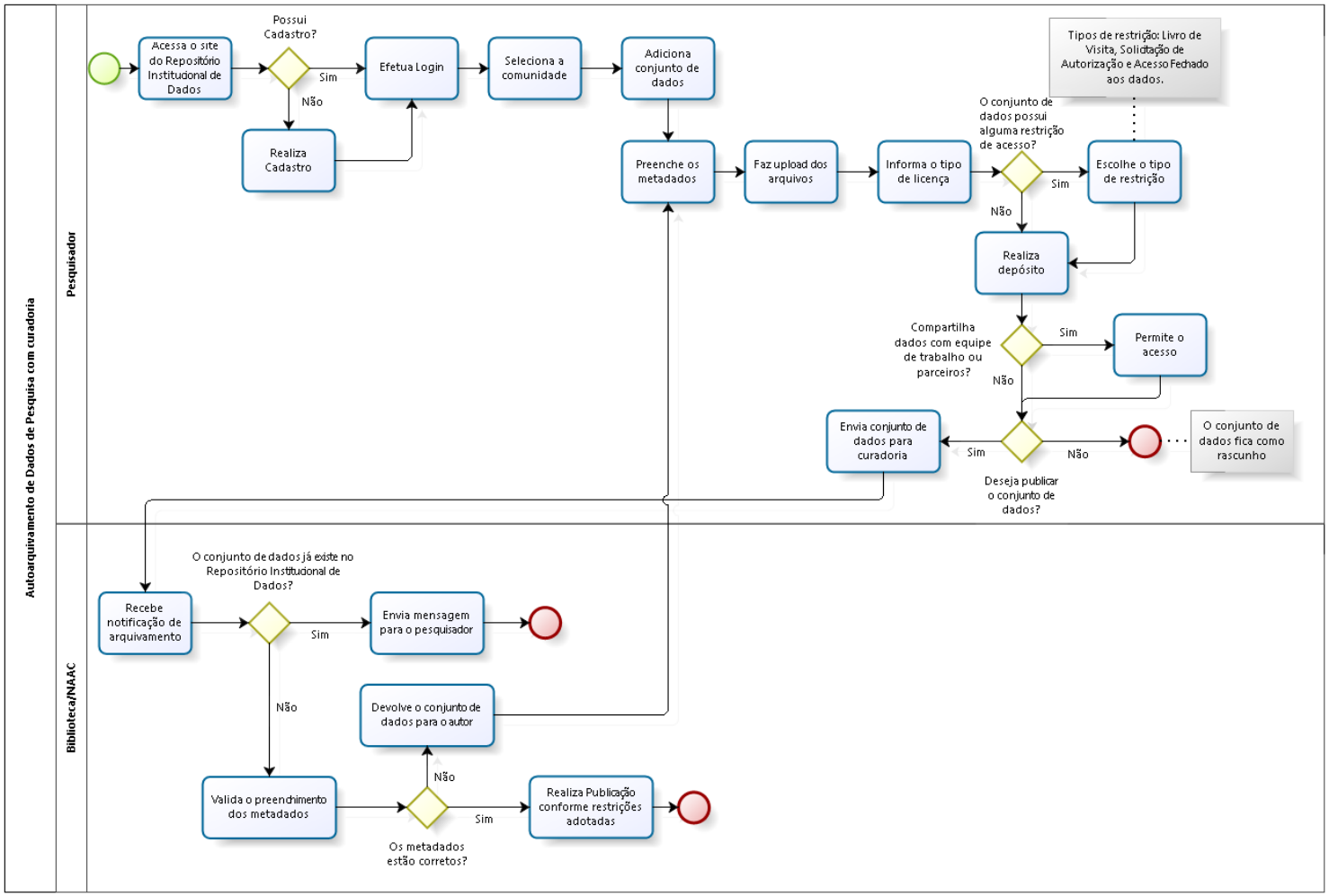
institutos de pesquisa brasileiros. Por meio do oasisbr é possível também realizar buscas em fontes de informação portuguesas.

5. FLUXOS DE DEPÓSITO E PUBLICAÇÃO

Há diversos fluxos de trabalho possíveis para depósito e publicação de dados em repositórios. Os fluxos podem variar, principalmente, em função da política de funcionamento estabelecida para o repositório e software escolhidos.

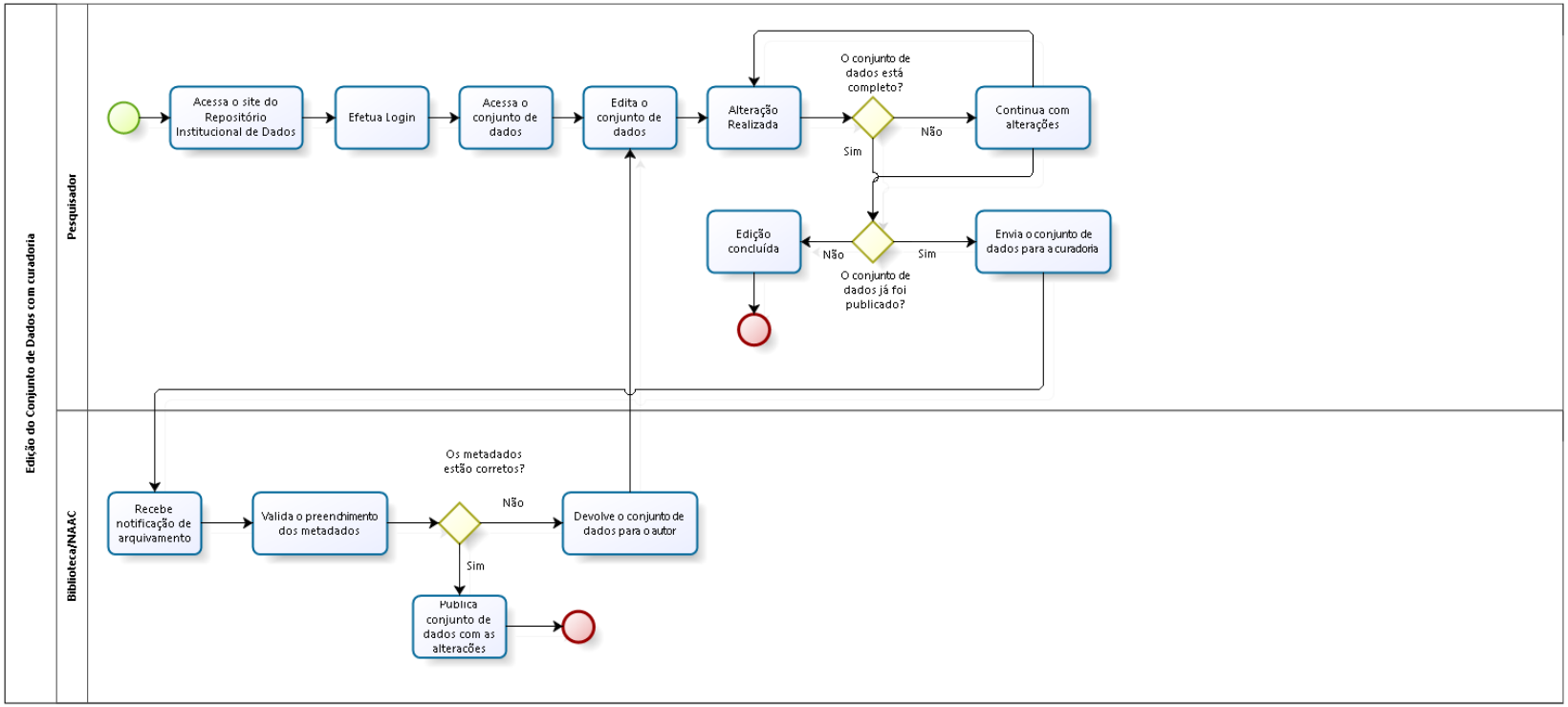
A partir dos testes realizados e registrados acima, foram elaboradas 3 (três) propostas pela equipe representada pelo ICICT, para os fluxos de autoarquivamento, edição, replicação e desacesso, como apresentados abaixo (Figuras: 2, 3 e 4):

Figura 2 - Autoarquivamento de Dados de Pesquisa com curadoria



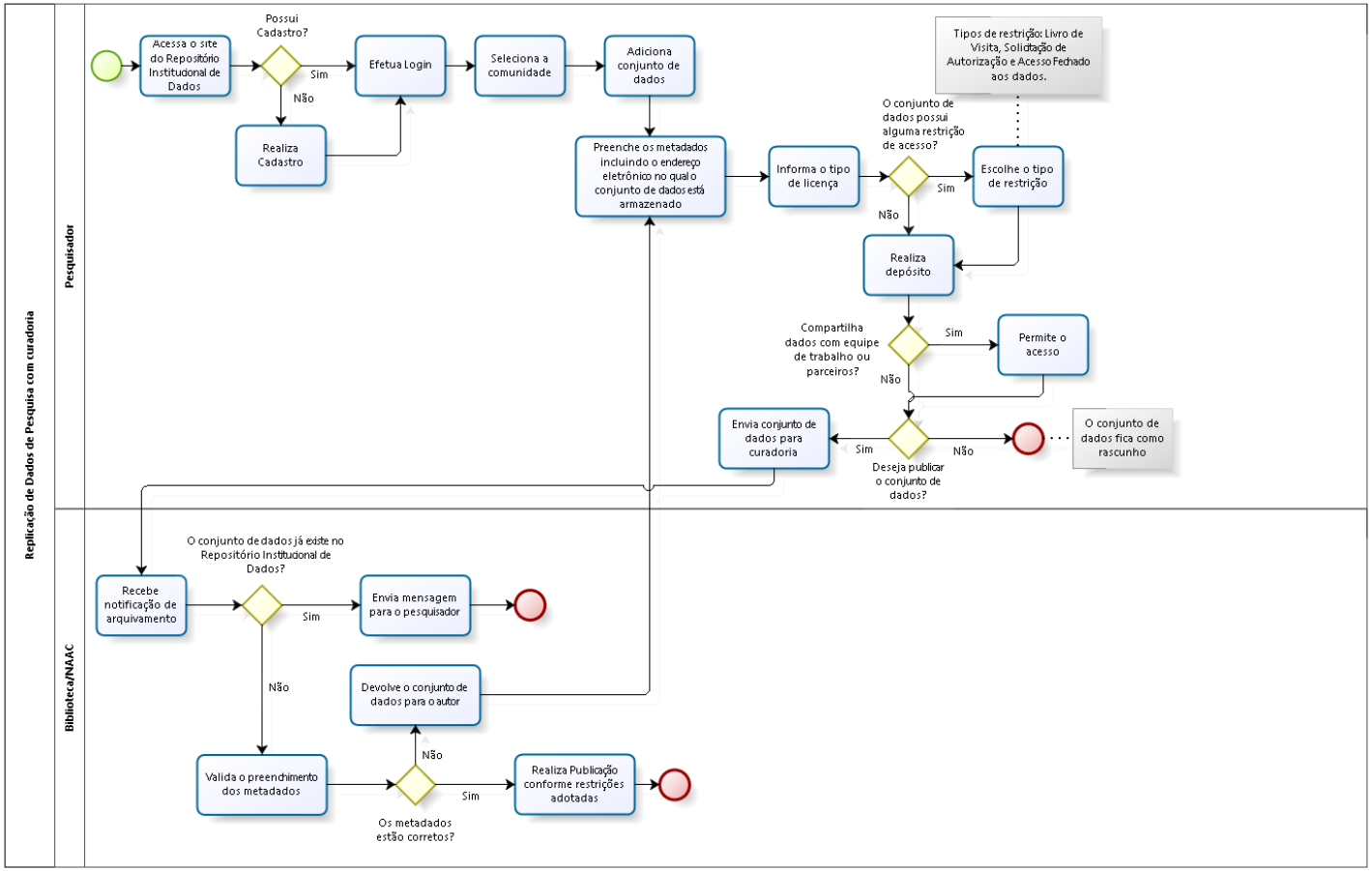
Fonte: Elaborado pela equipe

Figura 3 - Edição do Conjunto de Dados com curadoria



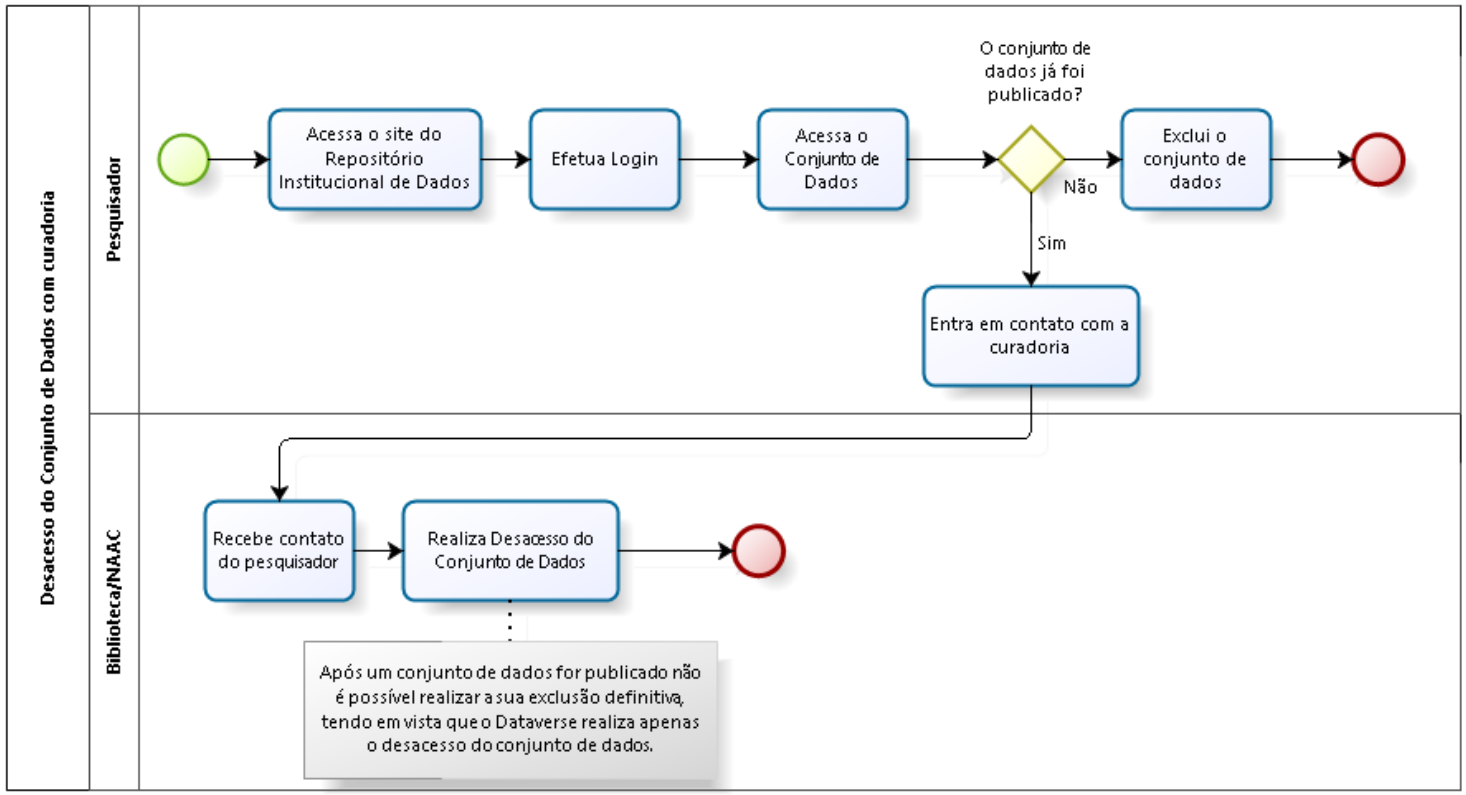
Fonte: Elaborado pela equipe

Figura 4 - Replicação de Dados de Pesquisa com curadoria



Fonte: Elaborado pela equipe

Figura 1 - Desacesso do Conjunto de Dados com curadoria



Fonte: Elaborado pela equipe

6. AVALIAÇÃO DO DATAVERSE

Os integrantes do GT do piloto OGP consideram que o software Dataverse respondeu positivamente aos testes realizados, podendo ser considerada uma ferramenta apta para testes mais detalhados como repositório institucional de dados de pesquisa. Destaca-se a necessidade de mais testes específicos e a solução de algumas questões levantadas durante o piloto OGP, listadas na seção 7 deste relatório, para sua adoção pela Fiocruz.

Para ações de melhoria do software, foi criada, pelo IBICT, uma lista de discussão nacional sobre o Dataverse, que já conta com mais de 140 membros. Ela será usada como canal central de discussões entre usuários do Dataverse no Brasil, da mesma forma como funciona a lista de discussão do DSpace (software para repositório de publicações científicas). Além da comunidade nacional criada, a GDCC (Global Dataverse Community Consortium)[13] é uma comunidade global do Dataverse, criada com o intuito de suportar e organizar os esforços de repositórios Dataverse espalhados pelo mundo. A Fiocruz deverá participar de comunidades nacionais e internacionais de desenvolvedores do Dataverse, seja para propor soluções tecnológicas para fragilidades percebidas no uso do sistema, seja para propor melhorias, sugestões e reclamações.

[13] http://dataversecommunity.global/

7. RECOMENDAÇÕES

A partir do trabalho realizado, o GT indica as seguintes recomendações para continuidade do processo de implementação do Dataverse Fiocruz:

1. Investir no elenco de profissionais (técnicos) que irão atuar na implementação do Repositório de Dados, com necessidade de:
    a) Designar uma equipe técnica com dedicação garantida para o desenvolvimento das atividades do repositório;
    b) Definir um profissional ou equipe de TI dedicada a customização das APIs da plataforma Dataverse Fiocruz;
    c) Alocar recursos financeiros para a capacitação da equipe técnica nas demandas de conhecimento;
    d) Disponibilizar a participação de equipe jurídica em nova fase do piloto.

2. Promover a integração da equipe técnica com pesquisadores que irão realizar o depósito de seus dados para testes;

3. Realizar novos testes com o identificador persistente DOI da FIOCRUZ, em funcionamento;

4. Verificar a possibilidade de tornar o campo licença editável, pois o programa tem como padrão o salvamento de licenças como CC0. O grupo considera importante uma preparação para esta etapa, que vai exigir um conhecimento das licenças existentes. Faz-se necessário esclarecer, junto a consultores jurídicos, todas as dúvidas sobre licenças e termos de uso;

5. Utilizar padrão do Arca (conforme seu manual) para forma de escrita da afiliação (com o acréscimo da sigla das instituições) e palavras chave;

6. Tornar o campo "*keyword*" obrigatório para otimizar o resultado das buscas;

7. Inserir no Dataverse textos padronizados pela Fiocruz, na parte de Ajuda e Dicas;

8. Seguir modelo Datacite para citação de dados;

9. Estabelecer estudo para definição de prazos de guarda dos dados, visando a preservação a longo prazo dos dados com valor de reúso e histórico para pesquisa;

10. Solicitar apoio consultivo técnico especializado do grupo de P&D da RNP, para tratativa das vulnerabilidades apresentadas no relatório Clavis;

11. Repetir teste de invasão, assim que a implementação da autenticação sugerida com o uso do *Shibboleth* e contas institucionais da FIOCRUZ for realizada e estiver operacionalizada;

12. Definir estratégias de capacitação e divulgação do repositório para os pesquisadores e profissionais da Fiocruz.

8. DESAFIOS PARA FASE 2 DO PILOTO

Como apoio para o planejamento para a Fase 2 do Repositório de Dados de Pesquisa, elencamos alguns dos principais desafios, são eles:

a) Adotar as recomendações listadas no item 7 deste relatório;

b) Estabelecer governança do repositório de dados para definição dos papéis e responsabilidades, assim como avaliar se o modelo de governança testado conseguirá dar conta da complexidade dos processos de implementação do repositório. É importante uma interação contínua entre as equipes que compõem a governança, pesquisadores e usuários dos dados publicados para avançar na maturidade de oferta do serviço de repositório de dados de pesquisa;

c) Identificar informações sobre restrições ao acesso, tipologia dos dados, necessidade de metadados específicos das pesquisas que farão parte da fase 2;

d) Testar fluxos desenhados no piloto OGP;

e) Executar testes de compartilhamento entre instituições;

f) Analisar se é possível e indicada a migração dos dados depositados no piloto para o modo produção;

g) Avançar em estudos e testes de soluções para armazenamento elástico. Como é esperado o aumento de depósitos de dados de pesquisa, recomenda-se uma solução baseada em um armazenamento elástico, sob risco de não funcionar devido a grandes volumes de dados;

h) Definir prazos de guarda dos dados dos projetos de pesquisa que participarão da fase 2, visando analisar o investimento humano e tecnológico necessário para curadoria e preservação digital dos dados a longo prazo;

i) Explicitar essas definições no “Termo de Uso” do serviço.

9. CONSIDERAÇÕES FINAIS

O piloto na fase atual teve duas metas a ser atingida: avaliar a solução tecnológica do Dataverse e o seu suporte para o armazenamento, compartilhamento e publicação de dados abertos de pesquisa. Concluímos que o nível de confiabilidade é atendido, e que o sistema está pronto para entrar em uso. Com certeza, deverá ser aperfeiçoado a partir de iniciativas identificadas no relatório, que objetivam tornar o software mais eficiente para os fins de uso da FIOCRUZ.