UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMUNICAÇÃO

MANUELA KLANOVICZ FERREIRA

DADOS DE PESQUISA
Contribuição da Infraestrutura para a Promoção da Reprodutibilidade e do Reúso

Porto Alegre
2024

MANUELA KLANOVICZ FERREIRA

DADOS DE PESQUISA
Contribuição da Infraestrutura para a Promoção da Reprodutibilidade e do Reúso

Tese apresentada como requisito parcial para a obtenção do Título de Doutora em Comunicação pelo Programa de Pós-graduação em Comunicação da Universidade Federal do Rio Grande do Sul.

Orientadora: Profª. Drª. Samile Andrea de Souza Vanz

Porto Alegre

2024

CIP - Catalogação na Publicação

Klanovicz Ferreira, Manuela
DADOS DE PESQUISA - Contribuição da Infraestrutura para a Promoção da Reprodutibilidade e do Reúso / Manuela Klanovicz Ferreira. -- 2024.
182 f.
Orientador: Samile Andrea de Souza Vanz.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Biblioteconomia e Comunicação, Programa de Pós-Graduação em Comunicação, Porto Alegre, BR-RS, 2024.

1. Reprodutibilidade de pesquisas. 2. Reúso de dados. 3. Gestão de dados de pesquisa. I. Andrea de Souza Vanz, Samile, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

MANUELA KLANOVICZ FERREIRA

DADOS DE PESQUISA
Contribuição da Infraestrutura para a Promoção da Reprodutibilidade e do Reúso

Tese apresentada ao Programa de Pós-Graduação em Comunicação da Universidade Federal do Rio Grande do Sul – UFRGS, como requisito parcial para obtenção do título de Doutora em Comunicação.

Aprovada em _____.

Banca Examinadora:

_____
Profª. Drª. Samile Andrea de Souza Vanz (ORIENTADORA)
Universidade Federal do Rio Grande do Sul

_____
Profª. Drª. Carolina Howard Felicissimo (EXAMINADORA)
Rede Nacional de Ensino e Pesquisa (RNP)

_____
Profª. Drª. Marcia Cristina Bernardes Barbosa (EXAMINADORA)
Universidade Federal do Rio Grande do Sul

_____
Profª. Drª. Caterina Marta Groposo Pavão (EXAMINADORA)
Universidade Federal do Rio Grande do Sul

_____
Profª. Drª. Sonia Elisa Caregnato (EXAMINADORA)
Universidade Federal do Rio Grande do Sul

_____
Prof. Dr. Rene Faustino Grabriel Junior (SUPLENTE)
Universidade Federal do Rio Grande do Sul

AGRADECIMENTOS

Eu agradeço a todos que contribuem para que a Ciência seja levada adiante e progrida neste mundo.

Especificamente no caso da minha pesquisa de doutorado, agradeço:

* à minha família, por entender todas as vezes em que não pude estar presente;

* à UFRGS, por ter, mais uma vez, propiciado uma educação pública e de qualidade para mim, entre tantos outros também beneficiados, que descobri o que era mestrado e doutorado só na quinta série, num dia que chovia muito e, para os três alunos molhados que compareceram na aula, a professora resolveu explicar estes conceitos;

* ao CPD da UFRGS, por ter me concedido afastamento nos dois últimos anos de doutorado, sendo que os dois primeiros foram durante o auge de uma pandemia;

* a todos que acreditam na Ciência e tomaram a vacina da COVID-19 e todas as demais vacinas para eles disponíveis;

* especialmente à minha orientadora, Samile Vanz, por sua compreensão das diferentes situações que atravessamos durante este período, por me incentivar a progredir, mostrando o caminho, me cobrar e me ouvir;

* ao grupo de Comunicação Científica que, com as experiências compartilhadas, proporcionou que eu aprendesse ainda mais sobre este assunto tão complexo.

I am among those who think that science has great beauty.
Marie Curie

RESUMO

Este trabalho analisa a contribuição das ferramentas que compõem a infraestrutura de suporte à pesquisa, particularmente as fornecidas pelo Laboratório Insterinstitucional de e-Astronomia (LIneA), para a reprodutibilidade das pesquisas o reúso dos dados ao longo do tempo. O LIneA foi escolhido por se tratar de uma instituição brasileira que oferece aos seus pesquisadores uma infraestrutura computacional com ferramentas necessárias para a manipulação da grande quantidade de dados em Astronomia, característica comum à e-Science. Este estudo adotou como procedimentos metodológicos a revisão bibliográfica para identificação das ferramentas de suporte à pesquisa que promovem a reprodutibilidade e reúso dos dados na ciência em geral; a pesquisa documental para identificar o funcionamento destas ferramentas; entrevistas semi-estruturadas com membros do LIneA e de suas colaborações que desempenham diferentes papéis a fim de identificar como as pesquisas aproveitam a infraestrutura oferecida pelo LIneA e qual a contribuição desta para a reprodutibilidade das pesquisa e o reúso dos dados; além do experimento de reprodução de pesquisas descritas em três artigos publicados por pesquisadores membros das colaborações apoiadas pelo LIneA. Durante este processo, foi identificada a utilização, no LIneA, de ferramentas dentre as quais destaca-se o versionamento de código-fonte de programas de análises através do Git Hub, a descrição interativa de fluxos de análises de dados utilizando Jupyter Hub e o encapsulamento do ambiente computacional por meio de containers Docker. Nas entrevistas, observou-se que os pesquisadores do LIneA utilizam a infraestrutura oferecida, por vezes solicitando auxílio de membros da equipe de TI, seja no uso do Git Hub ou Docker, para a disponibilização de artefatos de pesquisa ou na transformação de dados de terceiros para o reúso do pesquisador. Entretanto os pesquisadores relataram não ter conhecimento da reutilização dos artefatos produzidos por suas pesquisas, apesar de eles mesmo reusarem artefatos de terceiros. O experimento de avaliação de reprodutibilidade partiu da leitura dos três artigos selecionados e posterior coleta dos artefatos de pesquisa neles descritos. Os respectivos artefatos foram parcialmente recuperados, devido à falta de referência para o recorte dos dados de entrada utilizados nos artigos, ou de referência para a correta versão do código-fonte dos experimentos ou, também, pela falta de disponibilidade das dependências de software necessários para o ambiente computacional. Considera-se que o LIneA vem adotando sistematicamente práticas e ferramentas de suporte ao desenvolvimento e à documentação das pesquisas, as quais precisam alcançar um uso padronizado e combinado para atingir a reprodutilidade das pesquisas. Para este fim, com base nos trabalhos relacionados, sugere-se a criação de uma política de curadoria de dados, com o estabelecimento de um padrão de compartilhamento de artefatos para as pesquisas desenvolvidas pelos membros do LIneA, assim como a contratação de equipe responsável pela gestão dos dados de pesquisa para auxiliar tanto na elaboração desta política como na sua adoção pelos pesquisadores e membros do LIneA. Esta pesquisa evidencia a importância da curadoria digital se estender para além dos dados, abrangendo os programas e ambiente computacional utilizados, além de enumerar diversar ferramentas que podem ser empregadas com este propósito e investigar o seu uso pelo LineA.

Palavras chave: Reprodutibilidade de pesquisas. Reúso de dados. Gestão de dados de pesquisa.

ABSTRACT

This work analyzes the contribution of the tools that constitute the research support infrastructure, particularly those provided by the Interinstitutional Laboratory of e-Astronomy (LIneA), to the reproducibility of research and the reuse of data over time.. LIneA was chosen because it is a Brazilian institution that offers its researchers the necessary computational infrastructure with tools to handle the large amount of data in Astronomy, a common characteristic of e-Science. This study adopted methodological procedures including a literature review to identify research support tools that promote reproducibility and data reuse in science in general; documentary research to identify the functioning of these tools; semi-structured interviews with LIneA members and their collaborators who play different roles to identify how research benefits from the infrastructure offered by LIneA and how it contributes to the reproducibility of research and data reuse; as well as the experiment of reproducing research described in three articles published by researchers from collaborations supported by LIneA. During this process, the use of tools at LIneA was identified, with particular emphasis on versioning analysis program source code via GitHub, interactively describing data analysis workflows using JupyterHub, and encapsulating the computational environment with Docker containers. In the interviews, it was observed that LIneA researchers use the offered infrastructure, sometimes requesting assistance from IT team members, either in the use of GitHub or Docker for making research artifacts available or transforming third-party data for researcher's reuse. However, researchers reported not being aware of the reuse of artifacts produced by their research, despite themselves reusing third-party artifacts. The reproducibility evaluation experiment started with reading the three selected articles and subsequently collecting the research artifacts described in them. The respective artifacts were partially recovered due to a lack of reference for the entry data used in the articles, or reference for the correct version of the experiment source code, or also due to the unavailability of the necessary software dependencies for the computational environment. It is considered that LIneA has been systematically adopting practices and tools to support the development and documentation of research, which need to achieve standardized and combined use to attain research reproducibility. To this end, based on related works, it is suggested to create a data curation policy, establishing a standard for sharing artifacts for research developed by LIneA members, as well as hiring a team responsible for research data management to assist both in the creation of this policy and in its adoption by researchers and LIneA members.This research highlights the importance of digital curation extending beyond data to include the programs and computational environment used. It also lists various tools that can be employed for this purpose and investigates their use by LIneA.

Keywords: Research reproducibility. Data reuse. Research data management.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ACM	*Association for Computing Machinery*
AGN	*Active Galaxy Nuclei*
ALMA	*Atacama Large Millimeter Array*
API	Interface de Programação da Aplicação
CAAE	Certificado de Apresentação de Apreciação Ética
CAP	*CERN Analysis Preservation*
CEP	Comitê de Ética em Pesquisa
CERN	*European Organization for Nuclear Research*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COD	*CERN Open Data*
CPD	Centro de Processamento de Dados
CTIO	Observatório de Cerro Tololo
DES	*Dark Energy Survey*
DESI	*Dark Energy Spectroscopic Instrument*
DNA	Ácido desoxirribonucleico
DOI	*Digital Object Identifier*
EUA	Estados Unidos da América
FAIR	*Findable, Accessible, Interoperable, Reusable*
FAPERJ	Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FASEB	*Federation of American Societies for Experimental Biology*
FINEP	Financiadora de Estudos e Projetos
IDE	*Integrated Development Environment*
IEEE	*Institute of Electrical and Electronics Engineers*
INCT	Instituto Nacional de Tecnologia
JDDCP	*Joint Declaration of Data Citation Principles*
LCO	*Las Campanas Observatory*
LHC	*Large Hadron Collider*
LIneA	Laboratório Interinstitucional de e-Astronomia

LNCC	Laboratório Nacional de Computação Científica
LSBG	*Low-Surface-Brightness Galaxy*
LSST	*Legacy Survey of Space and Time*
MMV	Monitor de Máquinas Virtuais
MRF	*Multi-Resolution Filtering*
MV	Máquina Virtual
NASA	*National Aeronautics and Space Administration*
OAI-PMH	*Open Archives Initiative – Protocol for Metadata Harvesting*
ON	Observatório Nacional
ORCID	*Open Researcher and Contributor ID*
OSF	*Open Science Framework*
PDF	*Portable Document Format*
PGD	Plano de Gestão de Dados
PGDam	Plano de Gestão de Dados Acionável por Máquina
PHP	*Hypertext Preprocessor*
QLF	*Quick Look Framework*
RaaS	*Reproducibility as a Service*
RDP	Rede de Dados de Pesquisa
REANA	*Reusable Analyses Service*
RNP	Rede Nacional de Pesquisa
SaaS	*Software as a Service*
SDSS	*Sloan Digital Sky Survey*
SO	Sistema Operacional
SQL	*Structured Query Language*
SVN	*Subversion*
TI	Tecnologia da Informação
UFRGS	Universidade Federal do Rio Grande do Sul
UFRJ	Universidade Federal do Rio de Janeiro
URL	*Uniform Resource Locator*
WoS	*Web of Science*

SUMÁRIO

1 INTRODUÇÃO

Para ser válida, a pesquisa científica deve ser avaliada por outros pesquisadores, que não os próprios autores, a partir da análise dos artefatos produzidos durante o seu progresso. Para Meadows (1998), a comunicação da pesquisa é tão vital quanto o seu desenvolvimento, não cabendo a esta reivindicar com legitimidade este nome enquanto não houver sido analisada e aceita pelos pares, o que exige, necessariamente, que seja comunicada.

Quando apenas a comunicação impressa das pesquisas era possível de ser feita para um grande público, por meio de livros ou periódicos, o maior volume da comunicação científica consistia em descrever, textualmente e/ou com imagens, os procedimentos utilizados e os resultados alcançados. Por intermédio dos cada vez mais acessíveis computadores, não apenas a literatura pôde ser digitalizada, outros artefatos gerados também tornaram-se digitais em praticamente todos os campos do conhecimento.

Sem minimizar o desafio de condensar uma pesquisa científica em publicações tradicionais, o resultado deste processo, apesar de conteúdo complexo, possui um formato padrão, composto por textos e imagens, geralmente em arquivos PDF. Contudo, a disponibilidade digital de outros produtos de pesquisa pode possuir, além de conteúdo complexo com diversos padrões inteligíveis apenas a especialistas, diversos formatos de arquivo, passíveis de serem visualizados ou alterados, enfim usados, apenas com ferramentas e ambientes de software ou hardware específicos, o que aumenta a complexidade da tarefa de comunicar estas pesquisas.

Segundo Hey, Tansley e Tolle (2009, p. xvii), em texto baseado na transcrição de uma palestra de janeiro de 2007 de Jim Gray, estamos experimentando o quarto paradigma da ciência focado na exploração dos dados e que unifica os três paradigmas pregressos: o experimental, no qual eram realizados experimentos concretos, como deixar cair simultaneamente uma maçã e uma melancia para ver se ambos alcançavam o chão ao mesmo tempo; o teórico, onde criavam-se teorias para generalizar o que era observado nos experimentos, como a lei de inércia de Newton ou a lei das órbitas de planetas de Kepler; e o computacional, que permitiu realizar simulações numéricas para verificar as teorias relativas a problemas muito complexos para serem resolvidos analiticamente, como no caso dos sistemas de simulação para previsão do tempo diária. No quarto paradigma, também chamado de e-Science ou *big science*, os dados são muito volumosos e coletados por meio de instrumentos ou gerados artificialmente a partir de simulações. O pesquisador precisa de uma infraestrutura, ou seja, um conjunto de

ferramentas que inclui serviços de ferramentas computacionais e de manipulação e análise colaborativa e multidisciplinar de dados, além de códigos, softwares e ambiente de pesquisa virtuais (UNESCO, 2021) que permitam o processamento, a análise e a curadoria deste grande volume de dados, ou seja, na e-Science a quantidade de dados é tão grande, que o pesquisador não é mais capaz de visualizá-los e compreendê-los sem um pré-processamento. O termo, e-Science, foi cunhado em 2000 pelo então diretor do Conselho de Pesquisa do Reino Unido, John Taylor:

> Taylor reconheceu o papel cada vez mais importante que a TI (Tecnologia da Informação) deve desempenhar na pesquisa científica colaborativa, multidisciplinar e com uso intensivo de dados do século 21 e usou o termo e-Science para abranger a coleção de ferramentas e tecnologias necessárias para apoiar tal pesquisa (Hey; Tansley; Tolle, 2009, p. 227, tradução nossa).

Uma das características mais importantes da ciência é a confiabilidade alcançada através de sua avaliação (Mueller; Passos, 2000; Ziman, 2007), sendo a reprodutibilidade um meio para isso. Não bastava Kepler, Galileu ou Newton afirmarem suas descobertas em publicações científicas, era necessário que outros pudessem repetir seus experimentos por conta própria para que se convencessem. O principal conceito associado à reprodutibilidade é o de replicabilidade e, apesar de algumas divergências em relação ao significado desses conceitos (Barba, 2018), o mais usual é vincular a reprodutibilidade ao reúso dos dados da pesquisa e a replicabilidade ao uso de dados novos. Desta forma, o conceito de reprodutibilidade está relacionado ao de reúso de dados, pois é mediante o acesso aos artefatos de uma pesquisa que esta poderá ser reproduzida. Embora o fato de uma pesquisa ser reprodutível não garantir que suas análises estejam corretas, pois os procedimentos podem não estar corretamente modelados ou executados, reproduzir uma pesquisa permite afirmar que seus procedimentos e dados foram adequadamente documentados (Peng; Hicks, 2021) a ponto de praticar sua reprodução e, inclusive, identificar essas possíveis falhas nos procedimentos.

Com o aumento da complexidade das pesquisas na e-Science, tanto descrever, quanto reproduzir seus experimentos vêm tornando-se tarefas cada vez mais intrincadas, pois utilizam-se inúmeras ferramentas ao longo dos estudos, incluindo aparelhos para coleta de informações e execução de experimentos que, geralmente, dependem de ambientes computacionais específicos. Com a finalidade de proporcionar um maior grau de reprodutibilidade das pesquisas, existem iniciativas para encapsular estes ambientes computacionais de execução em diversas áreas do conhecimento (Boettiger, 2015; Heumüller; Nielobock; Küger; Ortmeier, 2020; Trisovic *et al.*, 2020). Descrever estas pesquisas de forma suficiente para que sejam

reprodutíveis por outros pesquisadores e, até mesmo, pelos próprios autores, aumenta o custo, seja monetário, seja em tempo (Baker, 2016a; Peng; Hicks, 2021). Entretanto, ter estas pesquisas bem descritas, documentadas e preservadas é essencial para permitir que elas sejam verificadas e promover o reúso efetivo de seus dados (Peng; Hicks, 2021; Moreau; Wiebles; Boettiger, 2023).

A Ciência Aberta é um movimento no qual um dos objetivos é o compartilhamento de forma irrestrita de todos os produtos das pesquisas. Foi precedida pelo movimento de Acesso Aberto, que focava em publicações finais de pesquisa, como artigos, incentivando que estes fossem compartilhados abertamente e não apenas em periódicos pagos de acesso restrito. Atualmente, com o advento do e-Science, a Ciência Aberta desponta olhando também para os demais produtos da pesquisa, como os dados. E, apesar da abertura dos dados de pesquisa não ser essencial para que esta seja reprodutível (Chen *et al.*, 2019), pois dados compartilhados de forma restrita também podem levar à reprodução, quanto mais aberto forem os dados, maiores as chances da pesquisa ser reproduzida com sucesso. Isso porque o acesso a dados abertos possui menos barreiras, sendo disponível a todos.

É importante salientar, todavia, que a publicação dos dados não promove apenas a reprodutibilidade das pesquisas, ela possibilita o seu reúso, onde o valor atribuído à pesquisa está diretamente relacionado ao potencial de seus dados serem reinterpretados em outras áreas e contextos diferentes daqueles que originalmente os gerou (Piwowar; Vision, 2013), estabelecendo novos padrões de socialização e de trabalho cooperativo independente de barreiras geográficas ou disciplinares (Sayão; Sales, 2014). Contudo, simplesmente abrir os dados não é o suficiente para garantir que eles sejam reusados (Chen *et al.*, 2019), pois a falta de documentação destes dados e de como eles foram utilizados no estudo original pode impedir sua reutilização.

Garantir que os dados sejam reusáveis exige uma ampla e adequada documentação, a qual tem o objetivo de descrever o contexto em que a pesquisa foi executada de forma que seja possível compreender como os resultados foram obtidos. Nesta documentação da pesquisa estão incluídas anotações de laboratório, procedimentos de coleta, limpeza e normalização dos dados, execução de experimentos, inclusive os que não obtiveram sucesso junto aos motivos do insucesso (Sayão; Sales, 2015; Munafò *et al.*, 2017). Tudo isso para levar os demais pesquisadores a compreenderem como e por que a pesquisa foi desenvolvida de determinada maneira. Esta documentação precisa conter metadados ricamente descritos com atributos precisos e relevantes, tais como a procedência detalhada dos dados (Wilkinson *et al.*, 2016),

sendo necessário documentar os procedimentos de obtenção e de tratamento dos dados de pesquisa desde a sua coleta (Chen *et al.*, 2019; Peng; Hicks, 2021).

Também é importante que todas as informações relativas aos dados de pesquisa sejam acessíveis por humanos e por máquinas, sendo mais sensato estabelecer padrões mínimos para a documentação dos dados, como metadados mínimos e em formatos abertos de arquivos e protocolos. Em repositórios de dados de propósito geral, torna-se ainda mais importante prover requisitos padronizados de documentação dos dados, pois um agente automático pode encontrar qualquer tipo de dado nestes repositórios e precisa compreendê-lo (Wilkinson *et al.*, 2016).

A e-Science está despontando, princípios e padrões estão sendo criados para aumentar a qualidade da comunicação das pesquisas e a tendência é que estratégias e ferramentas evoluam para facilitar que os pesquisadores criem e compartilhem dados que atendam a estes padrões. A pesquisa científica gera um grande volume de dados, neste sentido, a gestão e o compartilhamento devem ser executados de forma a não inviabilizar o reúso, permitindo, por exemplo, que em um grande conjunto de dados apenas o subconjunto utilizado seja referenciado inequivocamente, questão ainda em aberto, segundo Silvello (2018).

Quando se discute e-Science, também surge a dimensão da propriedade intelectual e direitos autorais. Neste cenário, todos os colaboradores que contribuíram para a coleta, criação, gestão e/ou curadoria dos dados devem ser elencados e creditados. O desafio está em repensar os direitos de propriedade intelectual, de um paradigma proprietário, para um espaço de processos colaborativos e compartilhados (Oliveira; Guimarães; Koshiyama, 2019). Utilizar ferramentas que já documentem as contribuições feitas por cada colaborador ao longo do processo da pesquisa e, no final, orientem para a escolha da licença mais adequada para os objetivos dos autores seria o ideal. Além disso, a organização de princípios para a citação de dados como o *Joint Declaration of Data Citation Principles* (JDDCP) (Martone, 2014), cujo um dos objetivos é o crédito e a atribuição, mostra como esta demanda por parte dos autores dos dados está sendo acolhida com o avanço dos estudos sobre gestão de dados. Entretanto ainda é necessário que os padrões de citação de dados sejam amplamente adotados, para permitir uma melhor análise de quais métricas poderiam ser mais adequadas em relação aos dados (Silveira; Barbosa; Ferreira; Caregnato, 2020).

Sendo o compartilhamento dos dados uma prática emergente na comunicação científica, é natural encontrar desafios diversos para que estes dados sejam identificados inequivocamente, disponibilizados em conjunto com sua descrição de procedência, com uma descrição do contexto em que foram coletados. Também deve-se considerar a necessidade de um licenciamento claro e a atribuição correta de crédito para cada colaborador da pesquisa.

Todas estas documentações com informações da pesquisa são importantes para permitir o reúso dos dados em outras pesquisas científicas e até mesmo no contexto de ensino.

O objeto de estudo deste projeto é a reprodutibilidade das pesquisas científicas no âmbito da e-Science e o reúso dos seus dados. Este tema abrange a qualidade da comunicação destas pesquisas e de seus dados por meio de uma documentação completa, planejada, atualizada e executada ao longo do desenvolvimento da pesquisa.

Uma das principais áreas representantes da e-Science é a Astronomia contemporânea, na qual, segundo Rosa (2019, p. 15), "[...] todo o astrônomo será (ou dependerá diretamente) de um cientista de dados para fazer descobertas importantes". As pessoas não olham mais através de um telescópio, em vez disso os dados são capturados com o uso de instrumentos ou gerados por meio de simulações antes de serem processados por programas cujos resultados serão armazenados em computadores para, só então, serem analisados pelos pesquisadores (Hey; Tansley; Tolle, 2009, p. xix). Observatórios astronômicos geram uma quantidade tão grande de dados que desafiam a capacidade dos astrônomos em manipular estes dados, como o telescópio Hubble[1] com cerca de 20GB por semana desde 1990, ou o ALMA[2] (*Atacama Large Millimeter Array*) no Chile que captura 2TB de dados todos os dias (Rosa, 2019). Sabendo que geralmente, cada geração de observatórios é dez vezes mais sensível que a anterior, certamente a necessidade de uma infraestrutura que permita aos pesquisadores manipular os dados Astronômicos só aumentará.

Este estudo visa analisar a contribuição da utilização das ferramentas que compõem a infraestrutura de suporte à pesquisa para a reprodutibilidade das pesquisas e o reúso de seus dados ao longo do tempo, particularmente em Astronomia. Esta área foi escolhida por manipular grande quantidade de dados, o que depende da utilização ferramentas de suporte. Almeja-se, com o estudo, obter uma perspectiva de como ferramentas de suporte à pesquisa podem ser utilizadas para automatizar sua documentação nas áreas de e-Science, especialmente na Astronomia.

Na seção seguinte são argumentadas as justificativas para o desenvolvimento desta pesquisa e, em seguida, são descritos o objetivo geral e os objetivos específicos a serem alcançados.

[1] Telescópio Espacial Hubble - https://pt.wikipedia.org/wiki/Telesc%C3%B3pio_espacial_Hubble
[2] ALMA (*Atacama Large Millimeter Array*) - https://pt.wikipedia.org/wiki/Atacama_Large_Millimeter_Array

1.1 JUSTIFICATIVA

Considerando o exposto, existem diversos desafios para a disponibilização dos produtos de uma pesquisa de forma a permitir a reprodutibilidade desta e o reúso de seus dados. Considera-se mais apropriado conduzir uma pesquisa bem documentada desde o princípio, pois pode não ser possível recuperar a documentação no final para tornar a pesquisa reprodutível (Peng; Hicks, 2021). Para cada desafio, surgem estratégias e ferramentas que visam facilitar e, desta forma, aumentar a qualidade da documentação das pesquisas e de seus dados durante o seu desenvolvimento, desde o planejamento da gestão dos dados, passando pelo encapsulamento do ambiente digital de execução dos experimentos, até um ambiente compartilhado de colaboração para a disponibilização dos dados. Apesar destas iniciativas para facilitar e incentivar o aumento da documentação das pesquisas, pode-se constatar que existe um baixo índice de reprodutibilidade das pesquisas, sendo as pesquisas compartilhadas com documentação insuficiente, ao mesmo tempo em que ocorre uma evolução das ferramentas para a documentação das pesquisas, conforme descrito a seguir.

O baixo índice de reprodutibilidade das pesquisas pode ser constatado em enquete realizada pela revista Nature em 2016 (Baker, 2016a), com 1.576 pesquisadores. O estudo questiona se existe uma "crise de reprodutibilidade" na ciência e conclui que mais de 70% dos respondentes falharam em reproduzir experimentos de outros cientistas e mais da metade falhou em reproduzir seus próprios experimentos. Entre os motivos apontados estavam: o relato seletivo dos resultados, mantendo ocultos resultados discrepantes que tornem as conclusões das pesquisas menos unânimes; a reexecução insuficiente no laboratório de origem da pesquisa; a metodologia, o código e os dados brutos não disponíveis. Nesta enquete, as respostas dos pesquisadores de diversas áreas do conhecimento confirmaram o que estudos anteriores em áreas específicas haviam apontado. Na área das Ciências Médicas, Scott *et al.* (2008) passaram cinco anos retestando mais de 70 drogas reportadas em estudos anteriores e não obtiveram sucesso em atestar suas eficácias para o tratamento de um tipo de esclerose. A iniciativa *Open Science Collaboration* (2015) tentou refazer 100 estudos experimentais e correlacionais na área da psicologia e obtiveram sucesso em apenas 36% dos casos. Em ambos os estudos, foi apontado como motivo para a baixa reprodutibilidade alcançada o viés da publicação de resultados positivos que gera uma significância estatística falsa.

Esta baixa reprodutibilidade possui uma causa multifatorial, podendo também ser justificada pelo fato dos artefatos das pesquisas serem compartilhados com documentação insuficiente, onde não ficam claros todos os passos realizados durante a pesquisa. Para Donoho

(2010) o compartilhamento dos trabalhos científicos deve ser feito em conjunto com os demais artefatos produzidos pela pesquisa, a fim de que os resultados sejam verificados. Segundo ele, os resultados das pesquisas são apresentados de forma pouco convincente em periódicos, conferências e livros devido, geralmente, às limitações impostas pelo formato dessas mídias.

Multiplicam-se ferramentas que permitem que dados de pesquisa sejam depositados de forma gratuita e em acesso aberto. Entretanto, estas ferramentas não primam pela descrição dos dados que disponibilizam o que, muitas vezes, pode comprometer o reúso e a reprodutibilidade da pesquisa por outros pesquisadores que não sejam os autores dos dados. Recentemente, Trisovic *et al.* (2020) não obtiveram sucesso em reexecutar nem sequer um terço dos experimentos em Python disponibilizados no repositório de Harvard[3], mesmo com a disponibilização dos arquivos executáveis, pois ocorriam erros pela falta da documentação das dependências de execução.

A computação está cada vez mais presente na rotina da sociedade, o que inclui a pesquisa, sendo um meio para que os experimentos sejam planejados, descritos, realizados e armazenados (Wallis; Rolando; Borgman, 2013; Peng; Hicks, 2021). A melhoria da interface entre os computadores e a sociedade tem como um dos recentes avanços o *smartphone* o qual permitiu que pessoas comuns portem computadores potentes em suas mãos e com as pontas dos dedos realizem atividades antes complexas, como enviar uma mensagem em tempo real, editar um vídeo ou monitorar suas atividades diárias e sinais vitais. A evolução da e-Science pode ser alavancada por avanço análogo, onde os pesquisadores tenham a tarefa de curadoria dos dados facilitada com a evolução das ferramentas para a documentação das pesquisas. Atualmente existem propostas de soluções para a coleta automática de informações sobre o ambiente computacional de execução da pesquisa (Boettiger, 2015; Brinckman *et al.*, 2019; Heumüller; Nielobock; Krüger; Ortmeier, 2020; Trisovic *et al.*, 2020; Cheifet, 2021; Hernández; Colom, 2023). Entre estas propostas, algumas são mais facilmente adotadas por pesquisadores com menor conhecimento da área de computação.

Conforme a definição de reprodutibilidade explicitada na seção 2.2 deste trabalho, reproduzir uma pesquisa implica em reusar seus dados. Desta forma, as qualidades exigidas dos dados para possibilitar a reprodutibilidade das pesquisas tendem a ser as mesmas exigidas para possibilitar seu reúso, pois apenas a partir de uma curadoria apropriada dos dados ambos são possíveis. A documentação automatizada da pesquisa poderia realizar-se de forma padronizada

[3] Harvard Dataverse - https://dataverse.harvard.edu/

e compreensível por cientistas, mas também por máquinas, permitindo a reprodução também automatizada, poupando recursos nas duas situações.

Há alguns estudos internacionais versando sobre ferramentas e estratégias para aumentar a qualidade da documentação em fases da pesquisa em áreas do conhecimento como física de altas energias (Chen *et al.*, 2019) e saúde pública (Peng; Hicks, 2021). Entretanto, na área de Astronomia não foram encontradas pesquisas que analisem as diversas iniciativas sendo utilizadas em conjunto, cada uma em uma fase do ciclo de vida dos dados da pesquisa, a fim de verificar a viabilidade de seu real uso pelos pesquisadores.

Ferramentas de Gestão de Dados de Pesquisa diferentes estão surgindo: repositórios especializados para a comunidade e-Science, também descritas como *big science*, cuja pouca variedade de formatos de dados e o grande volume de recursos de seus projetos permite a disponibilização de serviços customizados que automatizam o tratamento deste volume de dados e da documentação das pesquisas; e repositórios de propósito geral que disponibilizam dados de diversas áreas do conhecimento, havendo uma grande variedade de formatos e informações sobre estes dados, o que dificulta a disponibilização de serviços automatizados para o seu tratamento (Wallis; Rolando; Borgman, 2013). É preciso descortinar as diversas iniciativas existentes para possibilitar a esses pesquisadores, que não contam com repositórios especializados e são usuários de repositórios de propósito geral, acesso a meios de aumentar o reúso de seus dados a partir de uma documentação mais completa. Compreender como a e-Science está fazendo a gestão de seus dados pode ser um primeiro passo para descobrir como as demais áreas de pesquisa, cuja quantidade de dados é cada vez maior a ponto de também ser necessário o uso da computação como meio para alcançar seus resultados, podem usufruir destas mesmas estratégias e ferramentas para fazer a gestão eficiente de seus dados.

A investigação de como projetos brasileiros de pesquisa, com interesse na disponibilização de seus dados, adotam estratégias e utilizam uma infraestrutura para a documentação das pesquisas é importante para entender como se dá este processo e sua efetiva contribuição para a reprodutibilidade das pesquisas e o reúso dos dados. O estudo do Laboratório Interinstitucional de e-Astronomia (LIneA), referência no Brasil no fornecimento de uma infraestrutura para projetos na área de Astronomia, considerada uma área de e-Science, permitiu um aprofundamento da investigação de seus processos de documentação das pesquisas, incluindo seus dados. Esta delimitação de escopo tornou esta pesquisa de doutorado viável.

Também pretende-se contribuir para a continuidade de projeto no âmbito desta Universidade que já conta com um piloto de Repositório de Dados Científicos da UFRGS[4], cujo objetivo é oferecer infraestrutura aos seus pesquisadores que permita reunir, preservar, descrever e difundir o grande volume de dados gerados durante suas pesquisas. Atualmente este projeto encontra-se suspenso. Da mesma forma, o projeto citado contribui para a presente pesquisa.

Além disso, como oriunda da área de Tecnologia da Informação (TI), desde 2009 trabalho com a comunicação científica e convivo com sua ávida demanda por automatizar os seus processos sem, entretanto, deixar de primar pela qualidade, completude e veracidade das informações. Por muitas vezes, também, me deparei com a obsolescência dos artefatos digitais de uma pesquisa, inclusive da minha pesquisa de mestrado, que passaram a não ser mais acessíveis ou usáveis, quer porque o link de referência estivesse quebrado, quer porque o formato dos arquivos digitais não fosse mais suportado para acesso e, até mesmo, por falta de instruções do significado dos dados disponibilizados. Estas situações, idealmente, deveriam ter sido mitigadas pelos próprios pesquisadores durante a realização dos seus estudos, pois são difíceis de serem corrigidas, posteriormente, pela equipe responsável pela comunicação institucional das pesquisas. O pesquisador, entretanto, que exerce diversas funções dentro da Universidade, precisa que estas tarefas de documentação das pesquisas e compartilhamento de dados sejam facilitadas ao máximo. Na área da TI, verifico a utilização da tecnologia para automatizar diversos processos e acredito que é preciso aproximar estas ferramentas dos pesquisadores. Desta forma, o cenário da presente pesquisa compõe-se tanto com a motivação pessoal como profissional.

1.2 OBJETIVOS

A partir das justificativas levantadas acima, são apresentados os objetivos geral e específicos.

[4] Decisão n.º185/2017, de 04 de agosto de 2017 (CONSUN) - https://lume.ufrgs.br/handle/10183/165183

1.2.1 Objetivo geral

Analisar a contribuição das ferramentas que compõem a infraestrutura de suporte à pesquisa, particularmente as fornecidas pelo Laboratório Insterinstitucional de e-Astronomia (LIneA), para a reprodutibilidade das pesquisas e o reúso dos dados ao longo do tempo.

1.2.2 Objetivos específicos

Os objetivos específicos são:

a) elencar as ferramentas de suporte à pesquisa atualmente utilizadas e, entre estas, as fornecidas pelo LIneA aos pesquisadores para promover a documentação das pesquisas;

b) verificar como as ferramentas de suporte à pesquisa fornecidas pelo LIneA influenciam o padrão de reúso dos dados acessados por essas ferramentas, ao longo do tempo;

c) analisar os motivos que levaram os pesquisadores a utilizar as ferramentas fornecidas pelo LIneA;

d) identificar, a partir da visão dos membros do LIneA, possíveis dificuldades na utilização destas ferramentas;

e) experimentar a capacidade de reprodução retrospectiva das pesquisas realizadas, utilizando as ferramentas de suporte à pesquisa fornecidas pelo LIneA.

No capítulo seguinte, apresentam-se os conceitos de reprodutibilidade de pesquisa e reúso de dados, dentre outros conceitos relacionados à gestão de dados de pesquisa. Em seguida, é desenvolvido um abrangente levantamento das ferramentas disponíveis atualmente para a documentação das pesquisas, particularmente em e-Science, finalizando-se o capítulo com um caso bem documentado de utilização destas ferramentas, bem como uma revisão de trabalhos relacionados à utilização destas ferramentas. Posteriormente, abordam-se as questões metodológicas e o escopo da presente pesquisa, seguida da apresentação e análise dos resultados. Então, tem-se as considerações finais com as principais contribuições desta pesquisa.

5 CONSIDERAÇÕES FINAIS

Esta tese deteve-se em compreender como estão sendo utilizadas, atualmente, ferramentas de suporte à pesquisa científica e sua contribuição para que as pesquisas sejam reprodutíveis e para o aumento do reúso de seus dados. A investigação se deu por meio da apuração das ferramentas que compõem a infraestrutura de suporte à pesquisa, particularmente as fornecidas pelo LIneA (Laboratório Interinstitucional de e-Astronomia), o qual adota estratégia (de comunicação, suporte e reconhecimento de contribuição) e fornece infraestrutura (de acesso a dados, análise desses dados e bases de conhecimento) para que seus mais de uma centena de colaboradores consigam acessar dados astronômicos, realizar análises colaborativas e alcançar cada vez mais resultados.

Para atingir o primeiro objetivo específico, foi realizado inicialmente um abrangente levantamento, por meio da revisão bibliográfica, das ferramentas disponíveis atualmente para a documentação das pesquisas, particularmente em e-Science. As ferramentas mais citadas no levantamento bibliográfico tiveram seu funcionamento investigado em profundidade, com instalação ou criação de cadastro, conforme o caso, e elaboração de exemplos de teste simulando documentação de pesquisas, totalizando quinze ferramentas descritas na seção 2.3. Entre estas, as quatro mais citadas foram, por ordem, a ferramenta de containers Docker, a ferramenta de versionamento de código fonte Git Hub, a ferramenta de gestão de projetos de pesquisa *Open Science Framework* (OSF) e a ferramenta Jupyter Hub para a descrição interativa de programas de processamento de dados.

Posteriormente, foi realizado o levantamento das ferramentas fornecidas pelo LIneA para dar suporte à documentação das pesquisas das colaborações por ele apoiadas. A partir do "Manual de boas vindas", descrito na seção 4.2.4, disponível no Website do LIneA, foi possível descobrir os meios para acessar os principais instrumentos oferecidos pelo laboratório. Com o cadastro como pesquisadora, obteve-se acesso à infraestrutura do LIneA. Constatou-se a utilização de ferramentas para o acesso aos dados, implementação e execução de experimentos, diretamente relacionadas ao perfil de e-Science das pesquisas desenvolvidas pelo laboratório. Dentre estas, encontram-se ferramentas próprias, desenvolvidas pelos membros do laboratório, como o DES Science Portal e o DES Science Server, para acesso aos catálogos astronômicos, assim como ferramentas disponibilizadas por comunidades de software livre, algumas das quais são instaladas e mantidas dentro da infraestrutura do LIneA, como o Jupyter Hub, e outras que são utilizadas gratuitamente de forma hospedada por seus fornecedores, como o Git Hub. Foi considerado que estas ferramentas podem ter contribuição direta na reprodutibilidade e reúso,

suposição essa confirmada pelas entrevistas, nas quais os participantes relataram como a utilização da infraestrutura facilitava o acesso e manipulação dos dados, bem como a criação e execução de experimentos de forma colaborativa. Também verificou-se a utilização de ferramentas consideradas de gestão que oferecem apoio às colaborações, estando associadas à comunicação e ao treinamento entre os membros do LIneA e das colaborações por ele apoiadas, como os minicursos no Google Sala de Aula, o Slack como ferramenta de troca de mensagens interna e, até mesmo, o "Manual de boas vindas". Estas ferramentas são essenciais para a descoberta, pelos pesquisadores, de quais são e como podem ser utilizadas as ferramentas de suporte oferecidas pelo LIneA.

A análise das entrevistas realizadas com quatro membros do LIneA, com diferentes papéis dentro do laboratório, e dois pesquisadores membros das colaborações por ele apoiadas contribuiu para o alcance dos objetivos de levantamento das ferramentas, assim como os três objetivos específicos seguintes.

Ao verificar como as ferramentas de suporte à pesquisa fornecidas pelo LIneA influenciam o padrão de reúso ao longo do tempo dos dados acessados por essas ferramentas a partir das entrevistas, concluiu-se que as colaborações internacionais apoiadas pelo LIneA e seus membros fomentam uma infraestrutura para preservação e acesso aos dados da colaboração observados por telescópios. Sem esta infraestrutura, o acesso aos dados, que possuem um grande volume, seria bastante difícil. Entretanto, mesmo havendo esta iniciativa de compartilhamento dos dados, este compartilhamento ainda é feito sem um identificador único persistente, por exemplo, que poderia aumentar as chances destes mesmos dados serem localizados no futuro. Atualmente estes dados estão disponíveis como uma lista de arquivos em um Website e também acessíveis por meio de uma base de dados para a consulta SQL.

A análise dos motivos que levaram os pesquisadores a utilizar as ferramentas fornecidas pelo LIneA mostrou que o fato das ferramentas facilitarem a rotina dos pesquisadores por meio do acesso aos dados, no caso do DES Science Portal (descontinuado) e do ambiente do Jupyter Hub; por meio da disponibilização de um ambiente de processamento, novamente referente ao Jupyter Hub; e também de uma plataforma para documentação, preservação e compartilhamento do código-fonte dos experimentos, no caso do Git Hub, foram os principais motivadores para a utilização destes componentes da infraestrutura do LIneA pelos seus integrantes. Além disso, o Git Hub foi utilizado como opção para o cumprimento da exigência da agência de fomento, que financia a implementação dos produtos de software da equipe do LIneA, de que estes softwares sejam públicos.

Também foram identificados motivos para a não utilização desta infraestrutura, no

caso dos containers Docker, a falta de conhecimento de sua utilização, por parte dos pesquisadores. Outra dificuldade apontada foi o tempo despendido na aplicação do versionamento do Git Hub para todos os projetos, sendo relatado que, inicialmente, o pesquisador desenvolve o seu programa de forma independente e, quando alcança uma certa maturidade, este programa passa a ser tutorado também por uma pessoa da equipe de TI que auxilia e garante um certo padrão de qualidade da documentação do programa dentro do Git Hub. Para ambas as dificuldades, sugere-se a adoção de uma infraestrutura que facilite a utilização de containers, equivalente às plataformas de RaaS (*Reproducibility as a Service*) Code Ocean e Whole Tale, descritos na seção 2.3.3 e o acréscimo de um profissional especializado e dedicado à curadoria dos dados de pesquisa. O LIneA, inclusive, lançou, em janeiro de 2024, um edital[99] com bolsas para incluir diversos novos profissionais em sua equipe, atualmente de sete pessoas. Dentre estas vagas está prevista a contratação de um especialista em curadoria digital.

Outra dificuldade apontada pelos entrevistados, por exemplo, para o compartilhamento dos seus artefatos de pesquisa no Git Hub ou por meio do Website do LIneA, visto que ainda não foi criada a versão oficial de um Dataverse para conter os dados de saída, foi a competição pela publicação dos resultados, alegando-se que o compartilhamento dos dados ou programas poderia ocasionar a perda de alguma oportunidade de publicação, que poderia ser antecipada por outro pesquisador. Neste sentido, sugere-se a criação de uma política de curadoria de dados, esta política versaria sobre os dados intermediários das pesquisas, assim como seus produtos de software, pois para os dados de entrada, já existe uma política ditada pelas colaborações internacionais. Nesta política, deve estar previsto o nível, ou níveis, de compartilhamento autorizados pelas pesquisas desenvolvidas dentro do LIneA, com previsão de tempo de embargo, se for o caso. Além de previsão de um padrão para estes compartilhamentos, prevendo como a infraestrutura do LIneA deveria ser utilizada nestes casos. A instituição de um identificador único para estes artefatos, mesmo enquanto embargados, também é indicada a fim de permitir a sua localização no futuro e a citação destes por outros pesquisadores que venham a utilizá-los. Isso pode facilitar a garantia de reconhecimento para os autores originais dos artefatos.

Cabe destacar que a infraestrutura fornecida pelo LIneA influencia o desenvolvimento das pesquisas dentro do grupo, da mesma forma como foi previsto por Moreau, Wiebles e

[99] Edital Bolsas INCT do e-Universo 2024 - https://linea.org.br/wp-content/uploads/2024/01/Edital-bolsas-INCT-2024.pdf

Boettiger (2023) que as ferramentas de suporte poderiam impactar o ecossistema da pesquisa científica. Essa influência é verificada, por exemplo, no momento em que pela facilidade de acesso aos dados da colaboração DES proporcionada pelas ferramentas DES Science Server e Jupyter Hub, os pesquisadores também utilizam estas ferramentas para análise destes dados, e este uso acaba impondo um padrão de desenvolvimento destas pesquisas, delimitado por estas ferramentas. Neste sentido, a criação da política de curadoria de dados de pesquisa do LIneA também poderia ser influenciada pela infraestrutura disponível, conforme defendido por Pasquetto *et al.* (2016) quando argumenta que as políticas de dados também são moldadas pela infraestrutura disponível no que tange às características limitadas pelas ferramentas. Antes da criação da política, a partir da utilização das ferramentas são convencionados, mesmo que não de forma oficial, padrões de formato de dados, de programas de processamento, bem como de documentação dos mesmos. E a política, quando criada, tende a refletir esses padrões já pré-adotados por meio das ferramentas.

O objetivo de experimentar a capacidade de reprodução retrospectiva das pesquisas, utilizando as ferramentas de suporte à pesquisa fornecidas pelo LIneA, foi realizado por meio da tentativa de reprodução de três pesquisas descritas em artigos de pesquisadores membros das colaborações apoiadas pelo laboratório. Diversos fatores contribuíram para o insucesso na reprodução dos experimentos descritos nos artigos, destacando-se: a falta de referência direta aos dados de entrada, a não indicação de referência direta aos códigos-fonte dos experimentos implementados pelos artigos e indicação de suas versões e, por fim, a incompatibilidade com as dependências de software, que quando citadas também não tiveram suas versões apontadas, ou a indisponibilidade destas dependências. Para solucionar estes obstáculos sugeriu-se a criação de um formulário para padronizar o armazenamento das informações das pesquisas como localização de dados de entrada, código-fonte, imagens de container, quando houver, e publicações relacionadas. Conclui-se que, apesar das ferramentas adotadas pelo LIneA, mesmo assim a reprodutibilidade integral das pesquisas não pôde ser atingida.

Partindo das premissas de que: uma pesquisa para ser considerada válida precisa ser comunicada com qualidade, ou seja, a ponto de ser compreendida e avaliada por seus pares (Meadows, 1998); e de que o objetivo da ciência é o acúmulo de conhecimento e este ocorre quando novas pesquisas utilizam descobertas de pesquisas anteriores para avançar mais, é considerado curioso que os pesquisadores não consigam reproduzir as pesquisas de outros, nem sequer as próprias pesquisas passadas, assim como, tenham restrições e dificuldades em reutilizar dados de outros pesquisadores (Baker, 2016a). Investigando a situação atual no LIneA, percebe-se que a comunicação científica precisa evoluir. A Ciência precisa ir além de

compartilhar seus resultados por meio da literatura, é preciso compartilhar os dados utilizados nas pesquisas e esses dados devem ter sua compreensão facilitada por meio da adoção de padrões de formato e conteúdo, bem como de uma documentação completa e sem erros, essencial para o entendimento do significado de dados que se tornam cada vez mais complexos. Mas, além dos dados, as pesquisas têm muito mais artefatos para compartilhar: cadernos de laboratório, programas de automatização de experimentos, metodologias de experimentos que não obtiveram sucesso junto das respectivas decisões a que essas falhas levaram. Afinal, uma forma bastante didática de compreender um dado é conseguir vislumbrar como ele já foi utilizado em outra pesquisa, o que permite que se possa refinar experimentos anteriores com aquele dado, conforme defendido por Peng e Hicks (2021) que alegam que as reanálises permitem aprender muito com o processo, ou utilizar novos dados semelhantes para realizar os mesmos experimentos de uma pesquisa pregressa. Mas tudo isso parte de uma documentação apropriada.

No entanto, a documentação dos artefatos de pesquisa para sua preservação é uma tarefa dispendiosa. Para as áreas de e-Science, que precisam de uma infraestrutura computacional para processar suas grandes quantidades de dados, é possível encontrar ferramentas que podem ser utilizadas para facilitar esta documentação, como os Jupyter Notebook, que permitem a utilização, no mesmo documento, de textos formatados e imagens, bem como de trechos de código-fonte executáveis com um “play”; o Git Hub para desenvolvimento, gestão e compartilhamento do código-fonte de experimentos; ou os container Docker para a preservação do ambiente computacional de processamento dos experimentos, incluindo o código-fonte e as dependências com suas versões precisas, assim como os dados de entrada, ou uma amostra.

As informações coletadas por esta tese mostram que o LIneA, como um centro de e-Science, vem adotando sistematicamente práticas e ferramentas de suporte ao desenvolvimento e à documentação das pesquisas, as quais também fluem para a gestão de dados de pesquisa. A estratégia de adoção de práticas e ferramentas modulares que atendem a algumas etapas do ciclo de vida das pesquisas mostrou-se eficiente, no sentido que despertou nos pesquisadores a atenção à importância de documentar e preservar os artefatos de pesquisa. Conforme novas práticas e ferramentas forem adotadas pelo LIneA, é possível que novas versões dos artefatos de pesquisa sejam geradas, onde esses artefatos estarão mais bem documentados com seus ambientes de execução melhor preservados.

Atualmente, o LIneA utiliza este conjunto de ferramentas com o objetivo de dar acesso aos dados de colaborações como o DES, ou seja, o foco da infraestrutura atual está no uso e

reúso dos dados das colaborações, ao invés de na reprodutibilidade das pesquisas. Entretanto, sabendo que a reprodutibilidade depende do acesso e reúso dos dados, pode-se dizer que o LIneA já trilhou os primeiros passos rumo a reprodutibilidade de suas pesquisas. A utilização da infraestrutura atual, voltada ao reúso dos dados, poderia ser aproveitada com uma expansão de seu objetivo de acesso aos dados, a fim de abarcar, também, a reprodutibilidade das pesquisas. Para isso, sugere-se a criação de uma política de curadoria de dados, prevendo o nível de compartilhamento e qualidade de documentação exigidos para os artefatos produzidos pelo LIneA, assim como a contratação de um responsável pela curadoria digital dos artefatos de pesquisa que auxilie na implementação desta política.

É preciso considerar que a solução padronizada de preservação de ambiente computacional das pesquisas através de containers poderia também levar a uma obsolescência no momento em que a tecnologia de container utilizada, Docker, Singularity, Podman ou outro, deixasse de ser suportada por ter sido substituída por novas tecnologias. Entretanto, a partir da adoção de uma tecnologia padrão, seria possível promover a migração dos containers padronizados, possivelmente de forma automatizada, pois haveria apenas um problema: migrar container para uma nova tecnologia de preservação de ambiente computacional, ao invés do problema muito maior que seria migrar qualquer artefato digital de uma pesquisa para um formato suportado ao tempo em que seu acesso se faz necessário. O problema da migração de containers seria semelhante à migração do formato PDF adotado para praticamente toda a literatura científica. Mesmo hoje, quando ocorre a necessidade de transformá-lo para outros formatos como XML, uma linguagem de marcação que permite inserir *tags* nas informações a fim de que sejam mais facilmente identificáveis por máquina, mesmo neste contexto o problema não é migrar qualquer formato de arquivo de texto (DOC, DOCX, ODF, TXT, PPT, PPTX, EBOOK), o problema é migrar um PDF para XML.

Entre as limitações desta pesquisa, pode-se citar o fato da autora desta tese não possuir formação em Astronomia e possuir formação em Computação. Entretanto, conforme relatos nas entrevistas, há casos em que membros do LIneA que possuem formação em Astronomia também não obtiveram sucesso na reprodução total dos experimentos, na qual os mesmos artefatos (dados e código-fonte) são utilizados. Inclusive, para os artigos selecionados, foi solicitada a ajuda de membros do LIneA, que também não conseguiram resolver os problemas de não recuperação da versão correta de código-fonte ou de incompatibilidade com as dependências de software. Caso a replicação dos experimentos fosse almejada, o conhecimento em Astronomia seria necessário para selecionar novos dados de entrada e implementar novos experimentos.

Além disso, por ser composta por um estudo de caso, esta pesquisa não tem a capacidade de formular uma teoria abrangente sobre o uso de ferramentas de suporte à pesquisa e sua contribuição para o reúso e a reprodutibilidade. Ao invés disso, representa as interpretações e perspectivas dadas pela fonte do caso, o LIneA. Entretanto, as inferências e conclusões dessa análise podem ser extrapoladas para outras áreas de e-Science e para áreas pertencentes à cauda longa da ciência, onde também simplesmente depositar os arquivos em repositórios de dados de propósito geral, sem atentar para a qualidade da documentação destes dados, não é suficiente para a reprodutibilidade destas pesquisas e, muitas vezes, nem mesmo para o reúso destes dados. As instituições precisam estar cientes da dificuldade enfrentada pelos pesquisadores para documentar suas pesquisas a ponto destas serem reprodutíveis. Neste sentido, precisam criar políticas que, além de reivindicar novas obrigações de disponibilização de dados e implementação de pesquisas com níveis de qualidade reprodutíveis, também prevejam a adoção de ferramentas e práticas de suporte a estas tarefas, como a disponibilização de pessoas especialistas em curadoria de dados e de ferramentas de suporte que podem facilitar a adoção de um padrão de documentação para os dados e para as pesquisas. A ideia de haver níveis de qualidades de dados de pesquisa pode ser importante ao implementar um repositório de dados, por exemplo, onde seja possível atender rapidamente as necessidades de automatização e compartilhamento dos cientistas, muitas vezes devido a esta rapidez atendendo a níveis baixos de reprodutibilidade. Mas é possível também ver estes níveis como degraus que se pretende atingir para os dados de um repositório, podendo haver revisão periódica dos dados armazenados para garantir seu nível de reprodutibilidade e reúso ou mesmo para alcançar um nível maior gerando uma versão mais bem preservada e documentada. Apesar do LIneA ainda não ter uma política, ele já disponibiliza uma infraestrutura para a elaboração dos experimentos e está em vias de contratar um responsável pela curadoria dos dados.

Considera-se como principal contribuição desta pesquisa mostrar que, para além da preservação e da disponibilização dos dados para reúso, também é complexo executar no futuro experimentos de pesquisas e que esta reprodução depende de diversos fatores como a disponibilidade dos softwares e do ambiente computacional, além dos dados. Também foi possível encontrar propostas de soluções para estas dificuldades, dando luz às ferramentas que podem ser utilizadas durante o desenvolvimento das pesquisas para aumentar a sua documentação e promover a reprodutibilidade e reúso, como os containers Docker, o Jupyter Notebook e o Git Hub. E ainda como estas ferramentas estão sendo usadas nas pesquisas em e-Science e, especificamente, no LIneA, verificando-se que, além da disponibilidade das ferramentas, é necessário, para promover a reprodutibilidade e o reúso, que estas sejam

utilizadas em conjunto e colaborativamente, seguindo orientações de uma política de curadoria de dados. Ademais, constatou-se que estas ferramentas estão sendo cada vez mais citadas nas publicações científicas, o que leva a presumir que também estão sendo mais utilizadas nas pesquisas. Esse incremento do uso pode levar estas mesmas ferramentas a evoluírem de forma a possibilitar a sua utilização por mais áreas da ciência, tanto por meio do suporte a uma variedade maior de pesquisas, como pela simplificação de seu uso, permitindo o acesso a um número maior de pesquisadores.

Como trabalhos futuros, considera-se investigar outras áreas de e-Science a partir de outros centros de fornecimentos de infraestrutura e dados como, por exemplo, a Plataforma de Ciência de Dados Aplicados à Saúde (PCDaS/Icict)[100] que é uma iniciativa do Laboratório de Informação em Saúde (Lis), do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict), da Fundação Oswaldo Cruz (Fiocruz), e também em parceria com o Laboratório Nacional de Computação Científica (LNCC), cujo objetivo principal é disponibilizar serviços tecnológicos e de computação científica para armazenamento, gestão e análise de grandes quantidades de dados para pesquisadores, docentes e discentes de instituições de ensino e pesquisa, auxiliando a descoberta de informação útil a partir de grandes ou complexas bases de dados, bem como a tomada de decisão orientada por dados.

Como pesquisas futuras, também caberia investigar como a adoção de selos com níveis de reprodutibilidade de pesquisas atrelados à disponibilidade de seus artefatos, conforme sugerido por Feger *et al.* (2021) e Trisovic *et al.* (2020), poderiam influenciar a reutilização destes artefatos por outros pesquisadores. Assim como a adoção incremental destes selos, a partir do versionamento, também incremental, dos conjuntos de dados que seriam cada vez mais bem documentados a fim de ganharem novos níveis de reprodutibilidade atestados através destes selos. Este versionamento incremental dos artefatos poderia ocorrer, inclusive, a partir de outros pesquisadores que estivessem tentando reproduzir as pesquisas originais.

No âmbito do projeto de Repositório Digital de Dados da UFRGS, pretende-se adotar esta estratégia de versionamento com completude incremental dos artefatos de pesquisa a fim de permitir uma disponibilização inicial dos dados da pesquisa em um nível mais baixo de qualidade de documentação, mas de forma que este nível possa ser aumentado a ponto de tornar a pesquisa reprodutível.

[100] PCDaS/Icict - https://pcdas.icict.fiocruz.br/

REFERÊNCIAS

ABBOTT, Timothy M. C. *et al.* The dark energy survey: Data release 1. **The Astrophysical Journal Supplement Series**, v. 239, n. 2, p. 18, 2018. Disponível em: https://doi.org/10.3847/1538-4365/aae9f0. Acesso em: 27 jan. 2024.

ACM, Artifact Review and Badging, 2020. Disponível em: https://www.acm.org/publications/policies/artifact-review-and-badging-current. Acesso em 18 out. 2023.

AGUENA, Michel *et al.* The WaZP galaxy cluster sample of the dark energy survey year 1. **Monthly Notices of the Royal Astronomical Society**, v. 502, n. 3, p. 4435-4456, 2021. Disponível em: https://doi.org/10.1093/mnras/stab264. Acesso em: 11 out. 2023.

BAKER, Monya. 1,500 scientists lift the lid on reproducibility. **Nature**, v. 533, p. 452-454, maio 2016a. Disponível em: https://doi.org/10.1038/533452a. Acesso em: 29 abr. 2022.

BAKER, Monya. Muddled meanings hamper efforts to fix reproducibility crisis. **Nature**, junho 2016b. Disponível em: https://doi.org/10.1038/nature.2016.20076. Acesso em: 5 set. 2023.

BARBA, Lorena A. Terminologies for reproducible research. **arXiv preprint arXiv:1802.03311**, 2018. Disponível em: https://doi.org/10.48550/arXiv.1802.03311. Acesso em: 28 set. 2023.

BERTIN, Emmanuel. Automatic astrometric and photometric calibration with SCAMP. In: **Astronomical Data Analysis Software and Systems XV**. 2006. p. 112. Disponível em: http://aspbooks.org/custom/publications/paper/351-0112.html. Acesso em: 15 dez. 2023.

BLISCHAK, John D.; DAVENPORT, Emily R.; WILSON, Greg. A quick introduction to version control with Git and GitHub. **PLoS computational biology**, v. 12, n. 1, p. e1004668, 2016. Disponível em: https://doi.org/10.1371/journal.pcbi.1004668. Acesso em: 11 nov. 2023.

BOETTIGER, Carl. An introduction to Docker for reproducible research. **ACM SIGOPS Operating Systems Review,** v. 49, n. 1, p. 71-79, Jan. 2015. Disponível em: http://doi.acm.org/10.1145/2723872.2723882. Acesso em: 21 fev. 2022.

BORGMAN, Christine L. **Big Data, little Data, No Data: Scholarship in the Networked World**; MIT Press: Cambridge, MA, USA, 2015.

BORGMAN, Christine L. *et al*. Data management in the long tail: Science, software and service. **The International Journal of Digital Curation**, v. 11, n. 1, p. 128-149, 2016. Disponível em: dx.doi.org/10.2218/ijdc.v11i1.428. Acesso em:

BRINCKMAN, Adam. *et al.* Computing environments for reproducibility: Capturing the "Whole Tale". **Future Generation Computer Systems**, v. 94, p. 854-867, 2019. Disponível em: https://doi.org/10.1016/j.future.2017.12.029. Acesso em: 13 de jan. 2023.

CANALI, Stefano. Towards a contextual approach to data quality. **Data**, v. 5, n. 4, p. 90, 2020. Disponível em: https://doi.org/10.3390/data5040090. Acesso em: 20 fev. 2023.

CAREGNATO, Sonia E. *et al.* Reúso de dados: princípios FAIR e o ecossistema de pesquisa. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de da- dos de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 187 – 200. Disponível em: https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicados%20a%20gest%C3%A3o%20de%20dados%20de%20pesquisa_2021.pdf. Acesso em: 15 abr. 2023.

CHEIFET, Barbara. Promoting reproducibility with code ocean. **Genome Biology**, v. 22, n. 1, p. 1-2, 2021. Disponível em: https://doi.org/10.1186/s13059-021-02299-x. Acesso em: 13 de jan. 2023.

CHEN, Tracy X. *et al.* Best practices for data publication in the astronomical literature. **The Astrophysical Journal Supplement Series**, v. 260, n. 1, p. 5, 2022. Disponível em: https://doi.org/10.3847/1538-4365/ac6268. Acesso em 13 jan. 2024.

CHEN, Xiaoli *et al.* Open is not enough. **Nature Physics**, 2019. Disponível em: https://doi.org/10.1038/s41567-018-0342-2. Acesso em: 24 mai. 2022.

CITO, Jürgen; GALL, Harald C. Using Docker Containers to Improve Reproducibility in Software Engineering Research, 2016 IEEE/ACM **38th International Conference on Software Engineering Companion** (ICSE-C), Austin, TX, USA, 2016, pp. 906-907.

CLAERBOUT, Jon F.; KARRENBACH, Martin. Electronic documents give reproducible research a new meaning. In: **SEG technical program expanded abstracts 1992**. Society of Exploration Geophysicists, 1992. p. 601-604. Disponível em: https://doi.org/10.1190/1.1822162. Acesso em: 28 nov. 2023.

COSTA, Luiz A. N. LIneA: Um Centro de e-ciência na era de Big Data. **Palestra**, 2021. Instituto de Biofísica Carlos Chagas Filho (IBCCF UFRJ). Disponível em: https://www.youtube.com/watch?v=JGy1t_jUhPU. Acesso em: 12 jan. 2024.

COSTA, Luiz A. N. LIneA: Um Novo Modelo de Trabalho em Ciência. Entrevistador: Helio J. Rocha-Pinto. *In*: **Revista Brasileira de Astronomia**, Ano 5, Número 20, p. 20-25. Out.-dez. de 2023. Disponível em: https://sab-astro.org.br/wp-content/uploads/2023/12/RBA-20_online.pdf. Acesso em: 12 nov. 2023.

CRANMER, Kyle. *et al.* Analysis Preservation and Systematic Reinterpretation within the ATLAS experiment. *In*: **Journal of Physics: Conference Series**. IOP Publishing, 2018. Disponível em: https://doi.org/10.1088/1742-6596/1085/4/042011. Acesso em: 12 de set. 2022.

CURTY, Renata. Abordagens de reúso e a questão da reusabilidade dos dados científicos | Approaches for data reuse and the issue of scientific data reusability. **Liinc em Revista**, v. 15, n. 2, p. 177–193, 2019. Disponível em: https://doi.org/10.18617/liinc.v15i2.4777. Acesso em: 28 nov. 2022.

DONOHO, David. L. An invitation to reproducible computational research, **Biostatistics**, v. 3, n. 11, pp. 376-388, 2010. Disponível em: https://doi.org/10.1093/biostatistics/kxq028. Acesso em: 25 abr. 2022.

DU, Caifan *et al*. Understanding progress in software citation: a study of software citation in the CORD-19 corpus. **PeerJ Computer Science**, v. 8, p. e1022, 2022. Disponível em: https://doi.org/10.7717/peerj-cs.1022. Acesso em 15 mar. 2024.

DUARTE, Jorge; BARROS, Antonio. **Métodos e técnicas de pesquisa em comunicação**. Atlas, 2006.

FASEB. Enhancing research reproducibility. Federation of American Societies for Experimental Biology. Disponível em: https://www.faseb.org/FASEB/media/PDF/News/Washington%20Update/FASEB_Enhancing-Research-Reproducibility_1.pdf. Acesso em: 23 nov. 2023.

FEGER, Sebastian Stefan *et al.* Tailored science badges: Enabling new forms of research interaction. In: **Designing Interactive Systems Conference 2021**. 2021. p. 576-588. Disponível em: https://doi.org/10.1145/3461778.3462067. Acesso em: 03 jan. 2022.

FEITELSON, Dror. G. From Repeatability to Reproducibility and Corroboration. **ACM SIGOPS Operating Systems Review**, 49(1):3–11, 2015. ISSN 01635980. Disponível em: https://doi.org/10.1145/2723fe872.2723875. Acesso em: 3 set. 2021.

FIGUEIREDO, Dalson *et al.* Seven reasons why: a user's guide to transparency and reproducibility. **Brazilian Political Science Review**, v. 13, 2019. Disponível em: https://doi.org/10.1590/1981-3821201900020001. Acesso em: 20 nov. 2023.

FISCHER, Tobias *et al.* A robostack tutorial: Using the robot operating system alongside the conda and jupyter data science ecosystems. **IEEE Robotics & Automation Magazine**, v. 29, n. 2, p. 65-74, 2021. Disponível em: https://doi.org/10.1109/MRA.2021.3128367. Acesso em: 13 jan. 2024.

FOKIANOS, Pamfilos *et al.* CERN Analysis Preservation and Reuse Framework: FAIR research data services for LHC experiments. In: **EPJ Web of Conferences**. EDP Sciences, 2020. Disponível em: https://doi.org/10.1051/epjconf/202024506011. Acesso em 23 dez. 2023.

GABRIEL JUNIOR, Rene F. *et al.* Acesso aberto a dados de pesquisa no brasil: mapeamento de repositórios, práticas e percepções dos pesquisadores e tecnologias. **Ciência da Informação**, v. 48, n. 3, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/136453. Acesso em: 8 mai. 2023.

GOODMAN, Alyssa *et al.* Ten simple rules for the care and feeding of scientific data. **PLoS computational biology**, v. 10, n. 4, p. e1003542, 2014. Disponível em: https://doi.org/10.1371/journal.pcbi.1003542. Acesso em 18 out. 2023.

GOODMAN, Steven N.; FANELLI, Daniele; IOANNIDIS, John PA. What does research reproducibility mean? **Science translational medicine**, v. 8, n. 341, p. 341ps12-341ps12, 2016. Disponível em: https://doi.org/10.1126/scitranslmed.aaf5027. Acesso em 12 set. 2023.

GSCHWEND, Julia *et al.* DES science portal: Computing photometric redshifts. **Astronomy and Computing**, v. 25, p. 58-80, 2018. Disponível em: https://doi.org/10.1016/j.ascom.2018.08.008. Acesso em: 3 de maio de 2023.

HARZING, Anne-Wil. Publish or Perish (**software**), versão 8.8.4384, 2023. Disponível em: https://harzing.com/resources/publish-or-perish/command-line. Acesso em: 23 fev. 2023.

HERNÁNDEZ, José Armando; COLOM, Miguel. Repeatability, Reproducibility, Replicability, Reusability (4R) in Journals' Policies and Software/Data Management in Scientific Publications: A Survey, Discussion, and Perspectives. **arXiv preprint arXiv:2312.11028**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2312.11028. Acesso em: 13 mar. 2024.

HEUMÜLLER, Robert; NIELOBOCK, Sebastian; KRÜGER, Jacob; ORTMEIER, Frank. Publish or perish, but do not forget your software artifacts. **Empir Software Eng** 25, 4585–4616 (2020). Disponível em: https://doi.org/10.1007/s10664-020-09851-6. Acesso em: 24 abr. 2022.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (ed.). Jim Gray on eScience: a transformed scientific method. *In*: **The fourth paradigm**: data- intensive scientific discovery. Redmond: Microsoft Research, p. 17-31, 2009. Disponível em: http://digital.library.unt.edu/ark:/67531/metadc31516/. Acesso em: 29 abr. 2022.

HILDRETH, Michael D. *et al.* HEP Software Foundation Community White Paper Working Group-Data and Software Preservation to Enable Reuse, **arXiv preprint arXiv:1810.01191,** 2018. Disponível em: https://arxiv.org/pdf/1810.01191. Acesso em: 12 de set. 2022.

HOWISON, James; BULLARD, Julia. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. **Journal of the Association for Information Science and Technology**, v. 67, n. 9, p. 2137-2155, 2016. Disponível em: https://doi.org/10.1002/asi.23538. Acesso em 15 mar. 2024.

LABORATÓRIO INTERINSTITUCIONAL DE E-ASTRONOMIA. LIneA: um centro de e-ciência. Vídeo institucional, 2023. Disponível em: https://www.youtube.com/watch?v=4oiEKtzTkTA. Acesso em: 23 nov. 2023.

LARAWAY, Sean *et al.* An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis. **Perspectives on Behavior Science**, v. 42, p. 33-57, 2019. Disponível em: https://doi.org/10.1007/s40614-019-00193-3. Acesso em: 20 nov. 2023.

LUQUE, Elmer *et al.* Deep SOAR follow-up photometry of two Milky Way outer-halo companions discovered with Dark Energy Survey. **Monthly Notices of the Royal Astronomical Society**, v. 478, n. 2, p. 2006-2018, 2018. Disponível em: https://doi.org/10.1093/mnras/sty1039. Acesso em 13 out. 2023.

LUQUE, Elmer *et al.* Digging deeper into the Southern skies: a compact Milky Way companion discovered in first-year Dark Energy Survey data. **Monthly Notices of the Royal Astronomical Society**, v. 458, n. 1, p. 603-612, 2016. Disponível em: https://doi.org/10.1093/mnras/stw302. Acesso em: 12 dez. 2023.

MALIK, Tanu. Reproducible eScience: The Data Containerization Challenge. *In:* **2023 IEEE 19th International Conference on e-Science (e-Science)**. IEEE, 2023. p. 1-5. Disponível em: https://doi.org/10.1109/e-Science58273.2023.10254837. Acesso em: 9 out. 2023.

MARTONE, Maryann (ed.). **Data Citation Synthesis Group**: Joint Declaration of Data Citation Principles. San Diego, CA: FORCE11; 2014. Disponível em: https://doi.org/10.25490/a97f-egyk. Acesso em: 21 fev. 2022.

MEADOWS, Arthur J. **Communication research**. San Diego : Academic Press, 1998.

MIKSA, Tomasz; OBLASSER, Simon; RAUBER, Andreas. Automating Research Data Management Using Machine-Actionable Data Management Plans. **ACM Transactions on Management Information Systems**, v. 13, n. 2, article 18, Dec. 2021, 22 p. Disponível em: https://dl.acm.org/doi/full/10.1145/3490396. Acesso em: 10 abr. 2023.

MOREAU, David; WIEBELS, Kristina; BOETTIGER, Carl. Containers for computational reproducibility. **Nature Reviews Methods Primers**, v. 3, n. 1, p. 50, 2023. Disponível em: https://doi.org/10.1038/s43586-023-00236-9. Acesso em: 9 set. 2023.

MOUAT, Adrian. **Using Docker: Developing and Deploying Software with Containers**. O'Reilly Media, Inc, 2016. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=wpYpCwAAQBAJ&oi=fnd&pg=PP1&dq=container+history+docker&ots=QhL5yLnT9Q&sig=7PVD3sSBTX9ab0gIq5lrMIqkVps#v=onepage&q&f=false. Acesso em: 20 nov. 2022.

MUELLER, Suzana P. M.; PASSOS, Edilenice. As questões da comunicação científica e a ciência da informação. *In:* MUELLER, Suzana P. M.; PASSOS, Edilenice J. L. (Orgs.). **Comunicação científica**. Brasília: Ciência da Informação, 2000. p. 13-22. Disponível em: https://repositorio.unb.br/handle/10482/1444. Acesso em: 12 set. 2023.

MUNAFÒ, Marcus R. *et al.* A manifesto for reproducible science. **Nature human behaviour**, v. 1, n. 1, p. 1-9, 2017. Disponível em: http://dx.doi.org/10.1038/s41562-016-0021. Acesso em: 29 fev. 2024.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE *et al.* Reproducibility and replicability in science. 2019. Disponível em: https://doi.org/10.17226/25303. Acesso em: 23 nov. 2023.

OLIVEIRA, Adriana C. S. DE; GUIMARÃES, Patrícia B. V.; KOSHIYAMA, Débora C. A. D. G. a Ciência Aberta E Os Direitos De Propriedade Intelectual: Um Olhar a Partir Da Economia Criativa E Da Ciência Do Commons. **Revista de Direito da Cidade**, Rio de Janeiro, v. 11, n. 1, p. 663-681, 2019. Disponível em: https://doi.org/10.12957/rdc.2019.32031. Acesso em: 21 fev. 2022.

OPEN SCIENCE COLLABORATION. Estimating the reproducibility of psychological science. **Science**, v. 349, n. 6251, p. aac4716, 2015. Disponível em: http://dx.doi.org/10.1126/science.aac4716. Acesso em: 15 abr. 2023.

PASQUETTO, Irene V.; SANDS, Ashley E.; DARCH, Peter T.; BORGMAN, Christine L. Open data in scientific settings: From policy to practice. *In:* **Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems**. 2016. p. 1585-1596. Disponível em: https://doi.org/10.1145/2858036.2858543. Acesso em: 12 mar. 2022.

PASQUETTO, Irene V.; RANDLES, Bernadette M.; BORGMAN, Christine L. On the reuse of scientific data. 2017. Disponível em: https://doi.org/10.5334/dsj-2017-008. Acesso em: 15 abr. 2023.

PENG, Roger D.; DOMINICI, Francesca; ZEGER, Scott L. Reproducible epidemiologic research. **American journal of epidemiology**, v. 163, n. 9, p. 783-789, 2006. Disponível em: https://doi.org/10.1093/aje/kwj093. Acesso em: 20 nov. 2023.

PENG, Roger D. Reproducible research in computational science. Science, v. 334, n. 6060, p. 1226-1227, 2011. Disponível em: https://doi.org/10.1126%2Fscience.1213847. Acesso em: 5 out. 2023.

PENG, Roger D.; HICKS, Stephanie. C. Reproducible Research: a Retrospective. **Annual Review of Public Health**, 42:79-93, 2021 Disponível em: https://doi.org/10.1146/annurev-publhealth-012420-105110. Acesso em: 24 abr. 2022.

PENNOCK, Maureen. Digital Curation: a Life-Cycle Approach to Managing and Preserving Usable Digital Information, **Library & Archives**, v. 1, n. 1, p. 1-3, 2007. Disponível em: http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf. Acesso em: 10 dez. 2022.

PEREZ-RIVEROL, Yasset *et al*. Ten simple rules for taking advantage of Git and GitHub. **PLoS computational biology**, v. 12, n. 7, p. e1004947, 2016. Disponível em: https://doi.org/10.1371/journal.pcbi.1007142. Acesso em 15 nov. 2023.

PIMENTEL, João Felipe *et al*. A large-scale study about quality and reproducibility of jupyter notebooks. In: **2019 IEEE/ACM 16th international conference on mining software repositories (MSR)**. IEEE, 2019. p. 507-517. Disponível em: https://doi.org/10.1109/MSR.2019.00077. Acesso em: 20 fev. 2023.

PIWOWAR, Heather A.; VISION, Todd J. Data reuse and the open data citation advantage. **PeerJ**, v. 1, p. e175, 2013. Disponível em: https://doi.org/10.7717/peerj.175. Acesso em: 18 abr. 2021.

POTTERBUSCH, Megan; LOTRECCHIANO, G. R. Shifting paradigms in information flow: An open science framework (OSF) for knowledge sharing teams. **Informing Science**, v. 21, p. 179, 2018. Disponível em: https://doi.org/10.28945/4031. Acesso em: 14 jan. 2023.

PRÍNCIPE, Pedro. *et al.* Relatório técnico sobre ferramentas para a elaboração de Planos de Gestão de Dados. 2020. Disponível em: https://hdl.handle.net/1822/67530. Acesso em: 12 de jan. 2022.

ROKEM, Ariel; MARWICK, Ben.; STANEVA, Valentina. Assessing reproducibility. The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive

Sciences, p. 3-18, 2018. Disponível em: https://doi.org/10.31235/osf.io/gne3w. Acesso em: 5 out. 2023.

ROSA, Reinaldo, Astronomia do Futuro. *In*: **Revista Brasileira de Astronomia**, Ano 1, Número 4, p. 10-21. Out.-dez. 2019. Disponível em: https://sab-astro.org.br/wp-content/uploads/2022/01/RBA-4online.pdf. Acesso em: 12 nov. 2023.

ROSENFELD, Rogério; BERGMANN, Thaisa S. O Brasil no LSST. *In*: **Revista Brasileira de Astronomia**, Ano 5, Número 20, p. 11-19. Out.-dez. de 2023. Disponível em: https://sab-astro.org.br/wp-content/uploads/2023/12/RBA-20_online.pdf. Acesso em: 12 nov. 2023.

ROUGIER, Nicolas P. *et al.* Sustainable computational science: the ReScience initiative. **PeerJ Computer Science**, v. 3, p. e142, 2017. Disponível em: https://doi.org/10.7717/peerj-cs.142. Acesso em 23 nov. 2023.

SALES, Luana F.; SAYÃO, Luis F. A grande a pequena Ciência: análise das diferenças na gestão de dados de pesquisa. **Informação & Sociedade: Estudos**, 29(3), 2019. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/47615. Acesso em: 30 mai. 2022.

SAMUEL, Sheeba; KÖNIG-RIES, Birgitta. Understanding experiments and research practices for reproducibility: an exploratory study. **PeerJ**, v. 9, p. e11140, 2021. Disponível em: https://doi.org/10.7717/peerj.11140. Acesso em: 13 out. 2023.

SANTOS, Henrique M. da; FLORES, Daniel. Os impactos da obsolescência tecnológica frente à preservação de documentos digitais. **Brazilian Journal of Information Science: research trends**, v. 11, n. 2, 2017. Disponível em: https://doi.org/10.36311/1981-1640.2017.v11n2.04.p28. Acesso em: 18 dez. 2022.

SAYÃO, Luis F.; SALES, Luana F. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação & Sociedade: Estudos**, v. 22, n. 3, p. 179-191, 2012. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/12224. Acesso em: 18 dez. 2022.

SAYÃO, Luis F.; SALES, Luana F. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **RECIIS**, Rio de Janeiro, v. 8, n. 2, p. 76-92, jun. 2014. Disponível em: https://www.arca.fiocruz.br/handle/icict/17102. Acesso em: 21 fev. 2022.

SAYÃO, Luís F.; SALES, Luana F. Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores. 2015. Disponível em: https://www.aben.com.br/Arquivos/420/420.pdf. Acesso em: 11 mar. 2022.

SCHÖCH, Christof. Wiederholende Forschung in den digitalen Geisteswissenschaften. In: **DHd**. 2017. Disponível em: https://doi.org/10.5281/zenodo.277113. Acesso em: 6 mai. 2023.

SCOTT, Sean *et al.* Design, power, and interpretation of studies in the standard murine model of ALS. **Amyotrophic Lateral Sclerosis**, v. 9, n. 1, p. 4-15, 2008. Disponível em: https://doi.org/10.1080/17482960701856300. Acesso em: 10 out. 2023.

SILVEIRA, Lucia, BARBOSA, Amanda D., FERREIRA, Manuela K., CAREGNATO, Sonia. E. Citação de dados científicos: scoping review. **Encontros Bibli: revista eletrônica de biblioteconomia e da ciência a informação**, *25*, 01–31, 2020. Disponível em: https://doi.org/10.5007/1518-2924.2020.e72153. Acesso em: 9 de out. 2022.

SILVELLO, Gianmaria. Theory and practice of data citation. **Journal of the Association for Information Science and Technology**, v. 69, n. 1, p. 6-20, 2018. Disponível em: https://doi.org/10.48550/arXiv.1706.07976. Acesso em: 21 fev. 2022.

SIMMS, Stephanie. *et al.* Machine-actionable data management plans (maDMPs). **Research Ideas and Outcomes**, v. 3, e13086, abr. 2017. Disponível em: https://doi.org/10.3897/rio.3.e13086. Acesso em: 10 abr. 2023.

SPICHTINGER, Daniel; SIRE, Jarkko. The development of research data management policies in Horizon 2020. **Research Data Management-A European Perspective**, p. 11-23, 2017. Disponível em: https://doi.org/10.1515/9783110365634-002. Acesso em: 10 out. 2022.

TANOGLIDIS, Dimitrios *et al.* Shadows in the dark: Low-surface-brightness galaxies discovered in the dark energy survey. **The Astrophysical Journal Supplement Series**, v. 252, n. 2, p. 18, 2021. Disponível em: http://dx.doi.org/10.3847/1538-4365/abca89. Acesso em: 13 out. 2023.

TRISOVIC, Ana. **Data preservation and reproducibility at the LHCb experiment at CERN**. 2018. Tese de Doutorado. University of Cambridge. Disponível em: https://doi.org/10.17863/CAM.30973. Acesso em: 26 maio 2022.

TRISOVIC, Ana *et al.* Advancing Computational Reproducibility in the Dataverse Data Repository Platform. *In:* **Proceedings of the 3rd International Workshop on Practical Reproducible Evaluation of Computer Systems**, P-RECS '20, Stockholm, Sweden, 23 June 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 15–20. Disponível em: https://doi.org/10.1145/3391800.3398173. Acesso em: 8 jun. 2022.

TRISOVIC, Ana *et al.* Repository approaches to improving quality of shared data and code. **Data**, [s. l.], v. 6, n. 2, p. 1–12, 2021. Disponível em: https://doi.org/10.3390/data6020015. Acesso em: 21 fev. 2022.

TRISOVIC, Ana *et al.* A large-scale study on research code quality and execution. **Scientific Data**, v. 9, n. 1, p. 60, 2022. Disponível em: http://dx.doi.org/10.1038/s41597-022-01143-6. Acesso em: 27 nov. 2022.

UNESCO. Recommendation on Open Science. [S.l.: s.n.], 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en. Acesso em: 27 mai. 2024.

VAN DE SANDT, Stephanie *et al.* The definition of reuse. **Data Science Journal**, v. 18, n. 1, p. 1–19, 2019. Disponível em: http://dx.doi.org/10.5334/dsj-2019-022. Acesso em: 28 nov. 2022.

VAN GEND, Thijmen; ZUIDERWIJK, Anneke. Open research data: A case study into institutional and infrastructural arrangements to stimulate open research data sharing and

reuse. **Journal of Librarianship and Information Science**, v. 55, n. 3, p. 782-797, 2023. Disponível em: https://doi.org/10.1177/09610006221101200. Acesso em: 25 jul. 2023

VANZ, Samile A. *et al.* Acesso aberto a dados de pesquisa no Brasil: práticas e percepções dos pesquisadores: relatório 2018. 2018. Disponível em: http://hdl.handle.net/10183/185195. Acesso em: 17 set. 2021.

VANZ, Samile A. *et al.* Diretrizes para o estabelecimento de um checklist para curadoria de dados de pesquisa. **Informação em Pauta**, v. 6, n. 00, p. 1-18, 26 out. 2021. Disponível em: http://www.periodicos.ufc.br/informacaoempauta/article/view/68088/197501. Acesso em: 22 out. 2022.

VEIGA, Viviane *et al.* Plano de Gestão de Dados acionável por máquina, da teoria à prática: uma análise das ferramentas ARGOS e FioDMP. **BiblioCanto**, v. 9, n. 2, p. 16-29, 2023. Disponível em: https://doi.org/10.21680/2447-7842.2023v9n2ID33640. Acesso em: 20 dez. 2023.

VINES, Timothy H. *et al*. Mandated data archiving greatly improves access to research data. **arXiv preprint arXiv:1301.3744**, 2013. Disponível em: https://doi.org/10.1096/fj.12-218164. Acesso em: 13 mar. 2024.

WAARD, Anita de; COUSIJN, Helena; AALBERSBERG, Ijsbrand. 10 aspects of highly effective research data: Good research data management makes data reusable. December 2015. Disponível em: https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data. Acesso em: 26 mai. 2022.

WALLIS, Jillian C.; ROLANDO, Elizabeth; BORGMAN, Christine L. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. **PLoS ONE**, 8(7), 2013. ISSN 19326203. Disponível em: https://doi.org/10.1371/journal.pone.0067332. Acesso em: 26 mai. 2022.

WHYTE, Angus; TEDDS, Jonathan. Making the case for research data management. **Digital Curation Centre**, 2011. Disponível em: https://www.dcc.ac.uk/sites/default/files/documents/publications/Making%20the%20case.pdf. Acesso em: 30 abr. 2022.

WILKINSON, Mark D. *et al.* Comments: The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, Mar. 2016. Disponível em: https://doi.org/10.1038/sdata.2016.18. Acesso em: 21 fev. 2022.

WILLING, Carol. JupyterLab and JupyterHub - Perfect Together. **Palestra**, 2018. Bay Area Regional Python Conference 2018 (PyBay2018). Disponível em: https://www.youtube.com/watch?v=AXCo39qMn1E. Acesso em 12 nov. 2023.

YIN, Robert K. Case study research: Design and methods. Sage, 2002.

ZIMAN, John. Science in Civil Society. Imprint Academic : UK, 2007.

APÊNDICE A – Roteiro de entrevista

Muito obrigada por concordar em participar desta entrevista.

O objetivo deste estudo é identificar ferramentas e estratégias utilizadas pelo LIneA e analisar sua contribuição para a reprodutibilidade das pesquisas em Astronomia e reúso dos dados ao longo do tempo. Pretende-se, então, identificar práticas de documentação de pesquisas comuns em áreas distintas.

Esta entrevista será dividida em cinco blocos:

I. Identificação do participante da área de pesquisa
II. Caracterização dos artefatos de pesquisa (dados)
III. Práticas de documentação, compartilhamento, reúso
IV. Ferramentas utilizadas
V. Direitos e licenças

I. Identificação do participante e de área de pesquisa

1.1) Qual a sua área e nível (doutorado, doutorando(a), mestrado, mestrando(a)) de formação?
1.2) Há quanto tempo trabalha com pesquisa científica?
1.3) Qual a sua cidade/estado atual?
1.4) Qual a área/subárea/especialidade do participante? Por favor, fale sobre o perfil das pesquisas em que atua.

II. Caracterização dos artefatos de pesquisa (dados)

2.1) Como são produzidos os artefatos da pesquisa: procedimentos, entrevista, coleta de dados, monitoramento? Quem está envolvido?
2.2) Qual o volume médio dos artefatos de pesquisa produzidos em um ano ou ao final da mesma?
2.3) Os artefatos da pesquisa precisam ser atualizados com o tempo? Existe mais de uma versão destes artefatos? Versões anteriores são descartadas ou preservadas?

III. Práticas de documentação, compartilhamento, reúso

3.1) Existe uma política ou atividade de gerenciamento dos artefatos produzidos?

3.2) Existe planejamento para a gestão dos artefatos de uma pesquisa específica? É elaborado um Plano de Gestão de Dados?
3.3) Em qual fase da pesquisa é feita a documentação dos dados?
3.4) Já publicou artigo com exigência da revista de anexar os artefatos da pesquisa?
3.5) Como os artefatos da pesquisa são organizados (coleções), armazenados, documentados (existem metadados associados)? Considera que esta organização e documentação permite o entendimento da pesquisa por outro pesquisador para que os dados possam ser reusados?
3.6) Compartilha os artefatos da pesquisa com o grupo de pesquisa? E com outros grupos dentro e fora da instituição? E em acesso aberto? Quais artefatos são compartilhados: dados brutos, dados finais?
3.7) O que lhe motiva(ria) a compartilhar os dados de pesquisa?
3.8) Qual sua opinião sobre os processos de documentação das pesquisas e dos seus dados?
3.9) Você documenta sua pesquisa de forma a permitir o reúso dos dados e a reprodutibilidade?
3.10) Quais as vantagens e desvantagens de documentar uma pesquisa nos mínimos detalhes com o objetivo de compartilhamento, reúso e reprodutibilidade?
3.11) Já reutilizou artefatos de pesquisas de outros pesquisadores? Do seu grupo, de outro grupo? A documentação junto aos artefatos foi suficiente para que fossem reutilizados ou foi necessário contato com o pesquisador autor da pesquisa para entendimento dos dados?
3.12) Tem conhecimento da reutilização dos artefatos de sua pesquisa por outros pesquisadores? Do seu grupo ou fora dele? Foi necessário aos usuários questionar você sobre estes artefatos para sua reutilização? Em caso afirmativo, por quais motivos?
3.13) Ao citar um trabalho científico, daria preferência para um que disponibiliza os dados de pesquisa?

IV. Ferramentas utilizadas

4.1) No seu grupo de pesquisa são utilizadas algumas das ferramentas abaixo que otimizam a documentação das pesquisas?

Questionar se conhece ferramentas, caso não tenham sido citadas:

- Containers
- Jupyter Notebooks
- Jupyter Lab
- Jupyter Binder
- Google Colab
- Whole Tale

- Code Ocean
- Ferramenta de gerenciamento de projetos (de pesquisa) (Open Science Framework – osf.io)
- DMP Tool
- Ferramentas de disponibilização de dados (Dataverse, Figshare, Zenodo)

4.2) Para você, qual a contribuição das ferramentas de pesquisa utilizadas no seu grupo para a reprodutibilidade da sua pesquisa?

4.3) Para você, qual a contribuição das ferramentas de pesquisa utilizadas no seu grupo para o reúso dos dados de sua pesquisa?

V. Direitos e licenças

5.1) Os dados produzidos na sua pesquisa ou de seu grupo são sensíveis? Existem questões éticas e de privacidade?

5.2) A instituição à qual você pertence adota algum tipo de licença padrão para o compartilhamento de dados? Qual?

5.3) Se a instituição à qual você pertence não adota uma licença padrão para o compartilhamento de dados, quais os critérios que você utiliza para escolher a licença mais apropriada?

APÊNDICE B – E-mail Convite aos Candidatos a Participantes

De: Manuela Ferreira <manuelakf@cpd.ufrgs.br>
Assunto: CONVITE para participante do LIneA - Entrevista para TESE “Dados de Pesquisa - Estratégias para promover a reprodutibilidade e o reúso”
To: <e-mail do candidato>
Cc: Samile Andrea de Souza Vanz samile.vanz@ufrgs.br

Prezado <Nome do candidato>

Como vai?

Este é o convite para participar da etapa de entrevistas da pesquisa intitulada “Dados de Pesquisa - Estratégias para promover a reprodutibilidade e o reúso”, sob orientação da Professora Doutora Samile Vanz, cujo objetivo é analisar a contribuição da utilização de estratégias e ferramentas de suporte à pesquisa para a reprodutibilidade das pesquisas e o reúso dos seus dados.

Seu contato foi selecionado por você ser participante do LIneA.

Sua colaboração no presente estudo é muito importante, mas a decisão em participar é sua.

A coleta de dados ocorrerá por meio de entrevista semi-estruturada com roteiro prévio que observa variáveis que visam atingir o objetivo principal do estudo. As entrevistas serão aplicadas de modo síncrono em horário e data pré-agendados, por intermédio de plataforma de videoconferência (Meet ou MConf) com gravação de áudio e vídeo.

As entrevistas estão programadas para durar de 30 minutos a 1 hora.

Proposta de agendas:
dd/mm: turno da tarde (entre 14:30 e 17:30)
dd/mm: turno da tarde (entre 14:30 e 17:30)
dd/mm: turno da manhã (entre 8 e 11 horas)

Entre os turnos propostos, informe o horário que fica melhor para você.

Caso concorde em participar, solicitamos que preencha o Termo de Consentimento Livre e Esclarecido disponível em https://docs.google.com/forms/d/e/1FAIpQLSdCdiskibDhQ59uyVa4uPdOHxDVX2fBW8SL_KkIJ7q98hfvNQ/viewform.

Aguardamos seu retorno sobre o convite.
Desde já agradecemos a atenção.

Manuela Klanovicz Ferreira
Doutoranda em Comunicação
Mestra em Computação

APÊNDICE C – Termo de Consentimento Livre e Esclarecido

Você está sendo convidado a participar de pesquisa que analisar a contribuição da utilização das ferramentas de suporte à pesquisa fornecidas pelo Laboratório Interinstitucional de e-Astronomia (LIneA) para a reprodutibilidade das pesquisas em Astronomia e o reúso de seus dados ao longo do tempo. O LIneA é uma entidade brasileira que fornece uma infraestrutura tanto para o acesso a dados astronômicos como para manipulação e experimentação com estes dados. Não foram verificados estudos versando sobre a influência da aplicação de estratégias ou disponibilização de infraestrutura para a documentação das pesquisas para alcançar sua reprodutibilidade e reúso de seus dados. É preciso investigar como as estratégias e infraestrutura oferecidas pelo LIneA são utilizadas pelos seus membros e sua contribuição para a reprodutibilidade das pesquisas e o reúso dos dados.

A coleta de dados para este projeto ocorrerá por meio de entrevista semi-estruturada com roteiro prévio versando sobre sua experiência e conhecimento em relação ao tema da pesquisa. As entrevistas serão aplicadas através de reunião on-line com duração de 30 minutos a 1 hora que, com a sua autorização, será gravada. As respostas serão transcritas, organizadas e analisadas pela pesquisadora responsável. As entrevistas que serão realizadas possuem um risco intelectual mínimo, mas pode haver quebra de sigilo. Para minimizar este risco, garantimos o sigilo em relação às respostas dos participantes, sendo utilizadas como dados confidenciais (informações pessoais e identidade não serão reveladas) e utilizadas apenas para fins científicos como publicação de artigos, resumos em congressos e escrita da tese.

Ao participar desta pesquisa, você não terá nenhum benefício direto, nem qualquer tipo de despesa, bem como não receberá pagamento por sua participação. Entretanto, esperamos que futuramente os resultados deste estudo sejam usados para contribuir para a adoção de práticas de documentação que aumentem a reprodutibilidade das pesquisas e o uso dos dados.

Os procedimentos utilizados nesta pesquisa obedecem aos critérios de ética na Pesquisa de Seres Humanos conforme a Resolução nº 510, de 07 de abril de 2016, do Conselho Nacional de Saúde. Você pode desistir de participar desta pesquisa em qualquer momento. O consentimento à participação não retira o direito à indenização por eventuais danos causados pela pesquisa. Orientamos a guardar uma cópia deste termo de consentimento. Os dados coletados serão guardados por período mínimo de cinco anos e ficarão sob encargos da pesquisadora responsável. O projeto foi avaliado pelo CEP-UFRGS, órgão colegiado, de caráter consultivo, deliberativo e educativo, cuja finalidade é avaliar – emitir parecer e acompanhar os

projetos de pesquisa envolvendo seres humanos, em seus aspectos éticos e metodológicos, realizados no âmbito da instituição. Meios de contato com o CEP UFRGS: Av. Paulo Gama, 110, Sala 311, Prédio Anexo I da Reitoria - Campus Centro, Porto Alegre/RS - CEP: 90040-060. Fone: +55 51 3308 3787 E -mail: etica@propesq.ufrgs.br Horário de Funcionamento: de segunda a sexta, das 08:00 às 12:00 e das 13:30 às 17:30h.

Para concordar em participar, basta ler e preencher este termo. Você é livre para retirar seu consentimento em qualquer momento da pesquisa, sem prejuízo algum. Além disso, é garantido o acesso ao registro do seu consentimento sempre que solicitado e este formulário está configurado para enviar uma via do seu preenchimento deste termo para o endereço de e-mail preenchido. Os resultados desta pesquisa serão disponibilizados em repositório digital e encaminhados a todos os participantes.

Caso você possua perguntas sobre o estudo ou deseje retirar seu consentimento ou ter acesso ao registro do seu consentimento, você pode contatar a pesquisadora responsável Manuela Klanovicz Ferreira (PPGCOM/UFRGS) pelo e-mail manuelakf@cpd.ufrgs.br, bem como a orientadora Profa. Dra. Samile Andréa de Souza Vanz pelo e-mail samile.vanz@ufrgs.br.

Após este esclarecimento, solicitamos seu consentimento de forma livre e esclarecida para ser um(a) participante desta pesquisa. Para tanto, preencha o formulário abaixo. Desde já agradecemos a disponibilidade.

Pergunta 1 - *Você acha que está suficientemente informado(a) a respeito da pesquisa "Dados de Pesquisa – Estratégias de Reprodutibilidade de Reúso" e concorda de livre e espontânea vontade em participar como informante?*

( ) Sim ( ) Não

Pergunta 2 - *Nome completo do Participante da Pesquisa*: _____

Muito obrigada!

Link para Formulário do Termo de Consentimento Livre e Esclarecido - https://docs.google.com/forms/d/e/1FAIpQLSdCdiskibDhQ59uyVa4uPdOHxDVX2fBW8SL_KkIJ7q98hfvNQ/viewform