Uma Abordagem de Acompanhamento da Evolução de Planos de Gestão de Dados Ativos

Annatercia Gomes Pinheiro[1], Renato Cerceau[1], Maria Luiza Machado Campos[1], Sérgio Manuel Serra da Cruz[1,2]

[1]Programa de Pós-Graduação em Informática (PPGI/UFRRJ) - Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brasil

[2]Programa de Pós-Graduação em Humanidades Digitais (PPGIHD/UFRRJ) - Universidade Federal Rural do Rio de Janeiro (UFRRJ) – Nova Iguaçú, RJ – Brasil

{annatercia, mluiza, serra}@ufrj.br, cerceau@gmail.com

Abstract. *Scientific data governance has become a fundamental part of research projects based on Open Science, however it is not always properly executed. This article presents Active Plans & Provenance (APProve), an approach to support the management and evolution of Active Data Management Plans (PGDA) based on data provenance and aligned with FAIR principles. The proof of concept is based on experiments of a real case of versioning of a VODAN-Br project plan considering the Horizon 2020 funding determinations. The tool was able to capture the versioning and origin metadata of the operations, indicating the feasibility of the approach.*

Resumo. *A governança de dados científicos tornou-se peça fundamental para projetos de pesquisa baseados em Ciência Aberta, no entanto nem sempre é executada de modo adequado. Este artigo apresenta o Active Plans & Provenance (APProve), uma abordagem de apoio à gestão e evolução de Planos de Gestão de Dados Ativos (PGDA) baseados em proveniência de dados e alinhados aos princípios FAIR. A prova de conceito é baseada em experimentos de um caso real de versionamentos de um plano do projeto VODAN-Br considerando as determinações do financiamento Horizon 2020. A ferramenta foi capaz de capturar o versionamento e os metadados de proveniência das operações, indicando a viabilidade da abordagem.*

1. Introdução

A governança de dados científicos vem progressivamente se tornando peça fundamental para projetos baseados em Ciência Aberta (Souza et al., 2020). Não fortuitamente, os Planos de Gestão de Dados (PGD) passaram a ser exigidos por órgãos de fomento nacionais ou internacionais. Por exemplo, a FAPESP desde 2017 exige que todo projeto submetido precisa incluir um PGD[1]. Recentemente, o CNPq passou a requerer PGD nas novas submissões de projetos de pesquisa.

Um PGD é um documento formal de um projeto de pesquisa que contém perguntas orientadoras que estimulam o pesquisador a planejar e descrever,

[1] https://www.abcd.usp.br/noticias/fapesp-comeca-exigir-plano-de-gestao-de-dados/

detalhadamente sua pesquisa, como os dados serão coletados ou gerados; quais as metodologias e padrões serão utilizados; se, como e sob que condições esses dados serão compartilhados ou abertos para a comunidade de pesquisa; e como eles serão curados e preservados (Simms at al., 2017 e Cardoso et al, 2019).

Do ponto de vista dos financiadores, um dos intuitos do PGD é assegurar maior transparência e viabilizar que os *datasets* e artefatos digitais poderão ser mais facilmente localizados e reutilizados por outros pesquisadores, projetos e a própria sociedade (Veiga et al., 2019). No entanto, atualmente persistem dois problemas principais relacionados a gestão e acompanhamento de PGD.

Primeiro, muitos pesquisadores ainda desconhecem o que é um PGD ou que é parte integrante de solicitação de subsídios, alguns não sabem como produzi-los ou gerenciá-los ao longo do ciclo de vida da pesquisa (Wittenburg et al., 2019). O outro problema diz respeito a limitações funcionais das atuais ferramentas geradoras de PGD. Em geral, a maioria das ferramentas produz PGDs estáticos, geralmente criados antes do início de um projeto, que não são capazes de correlacionar as evoluções e versionamentos de um plano ao longo das mudanças que ocorrem durante uma pesquisa (Henning et al., 2021). Esses problemas levam pesquisadores a perceber que os PGDs estáticos não refletem adequadamente seus esforços, métodos de pesquisa ou mesmo *datasets* utilizados ou produzidos durante a pesquisa.

Com o intuito de mitigar esses problemas concebemos o *Active Plans & Provenance* (*APProve*). Trata-se de uma ferramenta que visa apoiar gestores, pesquisadores ou instituições e mostrar de forma simples o versionamento de um PGD Ativo (PGDA) usando recursos tradicionais da área de proveniência (Moreau, 2010) e Princípios FAIR (Wilkinson et al., 2016). Os princípios FAIR (um acrônimo para *Findable, Accessible, Interoperable e Reusable*) são aplicados na gestão de objetos digitais, em especial em dados científicos. Quando aplicados na governança de dados ele podem melhorar a qualidade dos dados e, consequentemente, sua capacidade de reuso para novas pesquisas, respeitando restrições éticas, legais ou contratuais.

Este artigo tem como contribuição uma ferramenta que pode ser usada como um artefato externo complementar às ferramentas geradoras de PGD que até o momento não contemplam o versionamento desses planos ao longo do ciclo de vida de projetos baseados em Ciência Aberta. Nossa prova de conceito envolve experimentos que abordam um caso real da área da Saúde onde as alterações do PGDA do projeto *Virus Outbreak Data Network Brazil* (VODAN-Br) passaram a levar em consideração as determinações dos fundos do Horizon 2020[2] (Campos et al., 2020).

Este artigo está organizado da seguinte forma, a Seção 2 compara as principais ferramentas produtoras de PGDs à luz dos temas curadoria, Princípios FAIR e proveniência de dados. A Seção 3 apresenta os materiais e métodos. A Seção 4 apresenta a arquitetura *APProve*. A Seção 5 apresenta os primeiros resultados experimentais e discussão considerando o projeto VODAN-Br. Por fim, a Seção 6 apresenta as considerações finais e trabalhos futuros.

[2] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en

2. Trabalhos Relacionados

Os experimentos científicos mediados por computador manipulam crescentes volumes de dados, exigindo novas habilidades como, por exemplo, a curadoria digital. Ela é parte ciclo de vida da pesquisa, envolve a preservação de dados e de outros ativos digitais a longo prazo, propondo-se a assegurar reusabilidade, reprodutibilidade, entre outros (Mattoso et al., 2010 e Poole, 2015).

No entanto, apesar da curadoria e proveniência de dados serem tradicionais em Ciência Aberta e terem farta literatura, até o momento não se identificam estratégias produtoras de PGDA com suporte a proveniência e princípios FAIR que permitam que pesquisadores acompanhem o versionamento dos planos (Cardoso et al., 2019).

Um PGDA é documento formal, acionável por máquina e alinhado aos Princípios FAIR, que permite a troca de dados entre várias entidades, em particular ao longo de todo o ciclo de vida da pesquisa. Realizamos análises exploratórias e comparativos das funcionalidades dessas ferramentas baseadas no método proposto por Jones et al. (2020). Dentre as ferramentas mais utilizadas pela comunidade científica, destacamos, DMPOnline[3], DMPTool[4], Data Stewardship Wizard[5] e ARGOS[6] (Tabela 1). Verificamos que elas produzem planos estáticos em formato PDF sem indicação de versionamentos. Adicionalmente, constatamos na época das análises elas não correlacionavam a proveniência das alterações de um plano e nem ofereciam recursos de visualização das alterações dos PGD aos pesquisadores.

Tabela 1: Quadro comparativo das ferramentas lançadas até abril de 2022.

Características	DMPOnline	DMPTool	DS-Wizard	ARGOS
Permite construção de novos *templates* de planos	Parcialmente	Não	Sim	Sim
Permite compartilhamento de planos em formatos RDF, JSON, XML	Parcialmente	Não	Parcialmente	Sim
Simplicidade/Facilidade de uso	Parcialmente	Sim	Sim	Sim
Oferece suporte ao usuário	Sim	Sim	Sim	Sim
É acionável por máquina	Não	Não	Sim	Sim
Adota os Princípios FAIR	Sim	Sim	Sim	Sim
Possui código aberto	Sim	Sim	Sim	Sim
Permite edição de *datasets*	Não	Não	Não	Sim

Nossas análises apontaram que a ferramenta que mais se aproximava das necessidades de criação de um PGDA era o ARGOS. Trata-se de uma ferramenta *open source* e extensível que oferece serviços de criação, validação, monitoramento e publicação de PGD textuais exportáveis em um formato de JSON (padrão, interoperável em máquinas distintas) mas, diferentemente das ferramentas supracitadas, permite editar *datasets*. Durante as análises, verificamos que tais características permitem classificá-lo, dentro do escopo do trabalho, como uma das melhores opções para apoiar todo o ciclo de vida de pesquisas em Ciência Aberta tais como o VODAN-Br (Tabela 1).

[3] https://dmponline.dcc.ac.uk/
[4] https://dmptool.org/
[5] https://ds-wizard.org/
[6] https://argos.openaire.eu/home

O ARGOS permite que se criem planos correlacionados com os *datasets* associados às pesquisas. Por padrão, seus planos possuem cinco seções (*project, grant, authors, description* e *licence*) que são criadas de forma privada e posteriormente, podem ser compartilhados com os times de trabalho do projeto se necessário. Os planos e seus *datasets* são inicialmente criados como *drafts* em formato JSON e tratados como documentos atualizáveis sendo por fim publicados. Adicionalmente, os planos podem ser identificados e acessados através de ORCID e DOI. Apesar dessas vantagens, ele não permite acompanhar a evolução das versões de um plano, não é possível verificar pela própria ferramenta o que foi alterado ou mesmo quem e quando o plano foi alterado. Ou seja, ele sobrepõe um PGD, não correlaciona suas versões e nem agrega metadados de proveniência aos planos.

Conceitualmente, as etapas de elaboração de um PGD no ARGOS são: *draft*, validação, finalização e publicação; finalizado com a preservação no repositório Zenodo[7] que permite que os planos possam ser compartilhados diretamente em ambiente aberto de curadoria de dados. Outra observação importante é sobre os *templates* de PGD do ARGOS, eles são bastante extensos, um pouco cansativos de serem preenchidos, mas em contrapartida são bem completos e personalizáveis. Visando apoiar os pesquisadores, propomos uma abordagem para facilitar a gestão de um PGDA que varie no tempo.

3. Materiais e Métodos

Essa pesquisa é caracterizada por ser de natureza teórico-aplicada. Após revisão da literatura, adotamos o método proposto por Jones et al. (2020) para comparação das funcionalidades das ferramentas produtoras de PGD (Tabela 1). Como prova de conceito avaliamos o PGD estático do projeto VODAN-Br. Resumidamente, o projeto estabelece uma infraestrutura de dados federada alinhada aos princípios FAIR e que apoie a coleta de dados de prontuários de pacientes infectados por vírus SARS-CoV-2.

Os materiais utilizados da construção da *APProve* são as linguagens PHP, HTML5, Javascript, Bibliotecas Python para geração e visualização de grafos de proveniência e manipulação de arquivos JSON. O *schema* de dados foi persistindo localmente em uma instância do SGBD PostgreSQL v. 13, ele foi modelado tomando como base o *schema* relacional do próprio ARGOS para armazenar os PGDA, suas versões, dados do projeto e seus metadados de proveniência. Os códigos-fonte da ferramenta e seus *datasets* estão disponíveis em https://github.com/annatercia/Approve.

4. *Arquitetura APProve*

APProve é uma ferramenta *open source* e modular que foi concebida para ser compatível com o *data layer* do ARGOS. Ela oferece às agências de fomento, instituições de pesquisa, pesquisadores e gestores uma visão geral dos projetos e dos PGDAs que ela armazena, por meio do acompanhamento simples e rápido da variação das versões de um plano ao longo do ciclo de vida do projeto. Na Figura 1 verificam-se os fluxos de 1 até 5 internos que indicam o caminho de um PGD no ARGOS e integração com a ferramenta proposta. As linhas pontilhadas os fluxos de dados entre os módulos do *APProve*.

[7] Repositório aberto de uso geral desenvolvido sob o programa OpenAIRE e operado pelo CERN. https://zenodo.org/

Dentre as funcionalidades essenciais do *APProve* estão as do módulo de ETL, capaz de fazer a importação e tratamentos do arquivo JSON de um PGD do ARGOS e persistência na camada de dados (BD *APProve*). Também existem as funções de edição na Web UI e gestão usuários, permissões e administração de versões no *backend*. Além disso, o *APProve* coleta de metadados de proveniência retrospectiva sobre cada operação realizada sobre as versões do PGDA é armazenada localmente no BD. Esses metadados de proveniência são persistidos na camada de dados da ferramenta e conectados com as seções e versões de um PGDA.

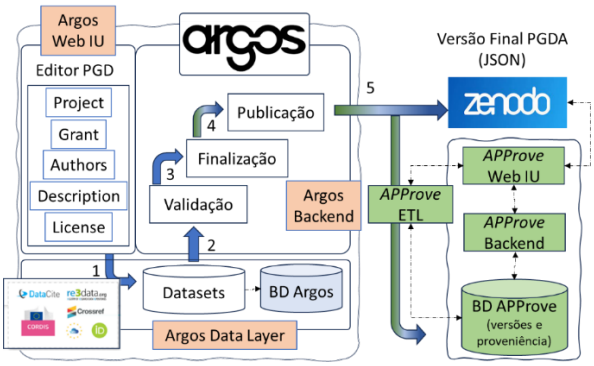

Figura 1: Representação conceitual da integração Argos-*APProve*.

5. Resultados Experimentais e Discussão

Os resultados dos primeiros experimentos computacionais visaram avaliar as funcionalidades da *APProve* usando como o PGD do projeto VODAN-Br que foi criado a partir da ferramenta ARGOS utilizando os *templates* do Horizon 2020. Os experimentos consideram as seções padrão de um PGD como ponto de partida. A seguir, derivamos duas novas versões do plano para avaliar as funcionalidades da ferramenta. Adotamos como condição do experimento que uma nova versão seria gerada se ocorresse qualquer alteração após finalização do preenchimento da versão anterior do PGD.

APProve oferece uma visualização *top-down* dos projetos de pesquisa previamente cadastrados e as informações sobre os pesquisadores do grupo de trabalho. A Figura 2 identifica os principais atributos descritivos do projeto VODAN-Br com as duas opções de consultas de "*Plans*" e "*Researchers*" associados.

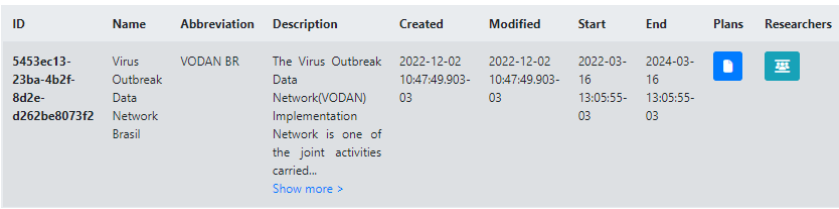

Figura 2: Fragmento de tela do *APProve* com visualização dos descritores do PGDA do projeto VODAN-Br.

Ao navegar pela interface "*Plans"* um usuário poderá comparar as três versões de um PGDA do VODAN-Br e identificar as variabilidades e os porcentuais de variação das seções de um plano (Figura 3). A interface "*Researchers"* exibe todos os pesquisadores cadastrados no PGDA. *APProve* faz comparação automática das versões do plano e exibe

as seções e os percentuais de variabilidade em relatórios dinâmicos e *online* de rastreamento de variabilidade do PGDA.

Version Number	Version 0	Version 1	Version 2
DMP ID	d381e0be-3900-433a-8d79-86cd53ee8647	f866937c-47fe-4a51-8637-9112b500b32b	9492013b-be34-42b7-b3ba-9291c877ee2c
Title	DMPVodan Brazil0	DMPVodan Brazil0	DMP Vodan Brazil [Modified to V2] Variability found ! 60% similarity !
Description	This PGD is to manager the Virus Outbreak Data Network(VODAN) Implementation Network is one of the j... Show more >	The Virus Outbreak Data Network(VODAN) Implementation Network is one of the joint activities carried... Show more > Variability found ! 96.6% similarity !	This PGD is to manager the Virus Outbreak Data Network(VODAN) Implementation Network is one of the j... Show more > Variability found ! 99.9% similarity !
Creation Date	2021-09-22 13:16:22-03	2022-04-12 06:30:46-03	2023-03-29 04:07:45-03
Modified Date	2022-12-07 14:07:23.328875-03	2022-04-12 17:37:35-03	2023-03-30 03:18:47-03

Figura 3: Fragmento de tela do *APProve* com comparativa de três versões um mesmo PGDA do projeto VODAN-Br.

Através dos links "*Show more* >" (embutidos nos relatórios dinâmicos), é possível visualizar individualmente as seções alteradas de um PGDA. Por exemplo, na coluna "Version 0" da Figura 3, o usuário confere quem foi o responsável pela criação da primeiro versão do PGDA e as alterações subsequentes daquela versão, quando e quais as ações ocorreram estão ilustrados na Figura 4. Além disso, o usuário pode conferir a estrutura textual da versão do PGDA ao selecionar a interface "*JSON Structure*". Também, é possível visualizar o grafo de proveniência das alterações ao selecionar "*Provenance Graph*".

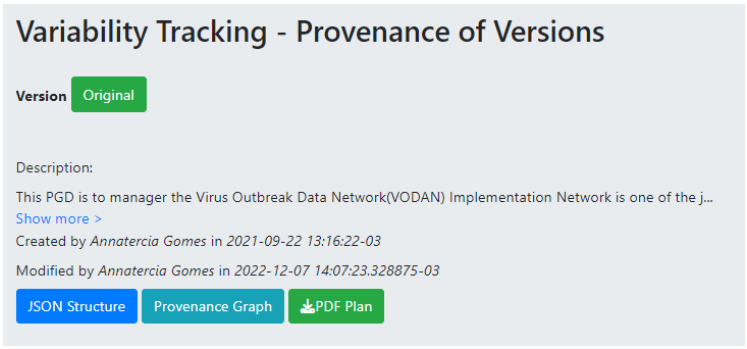

Figura 4: Fragmento de tela do *APProve* com exemplo de rastreamento da variabilidade na seção *Description* do PGDA do projeto VODAN-Br.

Ao comparar as versões do plano (Figura 3), além da ilustrar, de modo simplificado, quais seções foram alteradas, também exibe o percentual de variação daquela seção em relação à versão anterior. Através dos links "*Show more* >" é possível conferir as mudanças que ocorreram na versão do PGDA modificada. Por exemplo, na Figura 5 é possível identificar qual o foi o agente responsável pelas mudanças, *timestamp* em que elas ocorreram e onde elas foram aplicadas. Adicionalmente, é possível fazer *downloads* das versões intermediárias ou finais de um PGDA.

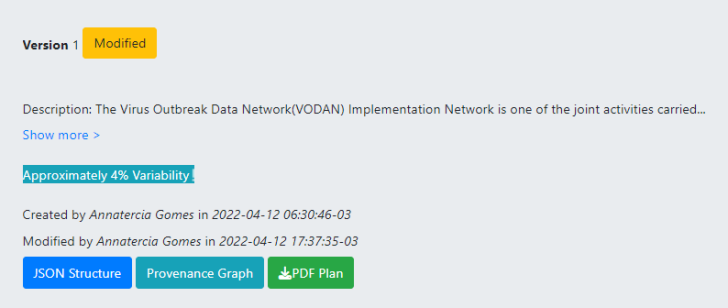

Figura 5: Visualização detalhada das versões.

Ao selecionar a interface "*Provenance Graph*" o usuário pode baixar o grafo de proveniência retrospectiva produzido pelo *APProve*. Ele é compatível com o padrão W3C PROV e reflete as alterações do PGDA. A Figura 6 ilustra um fragmento de proveniência que registra que o agente "`Researcher:AnnaterciaGomes`" que realizou a operação de alteração da descrição da entidade "`Vodan-Br:PGD-v0.json`", resultando em nova versão `Vodan-Br:PGD-v1.json` que utiliza um novo *dataset* denominado "`Sirio_Libanes001.zip`".

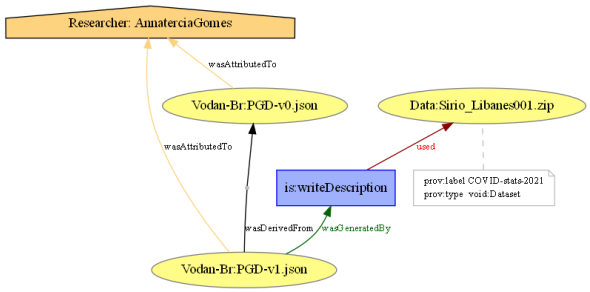

Figura 6: Fragmento de grafo de proveniência de uma alteração do PGDA.

Como resultado principal, observa-se que o *APProve* permite a comparação automática das versões de um PGDA de forma rápida. A ferramenta também captura as variações das versões do plano, através da visualização da proveniência das alterações de um plano. Ao comparar as versões do plano, além da ferramenta ilustrar rapidamente onde e o quanto uma versão do plano variou da versão anterior, é possível conferir o agente responsável pelas mudanças, as datas em que elas ocorreram e onde elas foram aplicadas através da visualização do grafo de proveniência.

6. Considerações Finais

Diversos órgãos de fomento reconhecem a importância da governança dos dados como sendo parte essencial das boas práticas de pesquisa. Para tanto, estimulam que os dados de projetos de pesquisa sejam compartilhados, visando o maior benefício possível para o avanço científico, tecnológico, socioeconômico e cultural de uma nação.

APProve é uma das primeiras ferramentas de apoio aos gestores, pesquisadores ou instituições que permite gerenciar o versionamento de um PGDA sistemas, comparar visualmente as diferentes versões e registar a proveniência das operações.

Nossos experimentos indicam que o *APProve* pode ser acoplado ao ARGOS. Essa abordagem automatiza tarefas típicas de gerenciamento de dados, contribui para a redução da carga de trabalho imposta às partes interessadas. Como trabalhos futuros

pretendem agregar novas funcionalidade: emissão de relatórios dos PGDA em formato PDF, incorporação de novos grafos indicando novas formas de visualização das alterações e testes com um maior número de PGDA (de domínios distintos) e mais usuários em ambientes distribuídos na nuvem da AWS.

Agradecimentos

Os autores agradecem à CAPES (código 001), ao CNPq (processos 306115/2021-2 e 400044/2023-4) e ao Fundo Nacional para o Desenvolvimento da Educação/PET-SI-UFRRJ.

Referências

Campos, M. L. M. et al. (2020). *VODAN BRAZIL - The Brazilian experience at the Virus Outbreak Data Network*. https://doi.org/10.5281/ZENODO.4291112

Cardoso, J., Miksa, T., & Borbinha, J. (2019). Debunking Active Data Management Plans. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 5308–5310. https://doi.org/10.1109/BigData.2018.8621860

Henning, P., et al. (2021). The FAIRness of data management plans: an assessment of some European DMPs. *Revista Eletrônica de Comunicação, Informação e Inovação Em Saúde*, *15*(3). https://doi.org/10.29397/reciis.v15i3.2270

Jones, S., et al. (2020). Data Management Planning: How Requirements and Solutions are Beginning to Converge. *Data Intelligence*, *2*(1–2). https://doi.org/10.1162/dint_a_00043

Mattoso, M. et al. (2010). Towards supporting the life cycle of large scale scientific experiments. Int. J. Bus. Process. Integr. Manag. 5(1): 79-92

Moreau, L. (2010). The foundations for provenance on the Web. *Foundations and Trends in Web Science*. https://doi.org/10.1561/1800000010

Poole, A. H. (2015). How has your science data grown? Digital curation and the human factor: a critical literature review. *Archival Science*, *15*(2). https://doi.org/10.1007/s10502-014-9236-y

Simms, S., et al. (2017). Machine-actionable data management plans (maDMPs). *Research Ideas and Outcomes*, *3*, e13086. https://doi.org/10.3897/rio.3.e13086

Souza, D. L. et al. (2020). A perspectiva dos pesquisadores sobre os desafios da pesquisa no Brasil. *Educação e Pesquisa*, *46*. https://doi.org/10.1590/s1678-4634202046221628

Veiga, V. S. de O., et al. (2019). Plano de gestão de dados fair: uma proposta para a Fiocruz | Fair data management plan: a proposal for Fiocruz. *Liinc Em Revista*. https://doi.org/10.18617/liinc.v15i2.5030

Wilkinson, M. D. et al. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. https://doi.org/10.1038/sdata.2016.18

Wittenburg, P., et al. (2019). The FAIR Funder pilot programme to make it easy for funders to require and for grantees to produce FAIR Data. https://arxiv.org/abs/1902.11162v2.