Dados escuros à luz do controle público

Alessandro Marinho de Albuquerque, Carina F. Dorneles

PPGCC – INE – Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC

alex.marinho.natal@gmail.com, carina.dorneles@ufsc.br

Abstract. *A considerable volume of data remains outside the knowledge of the governmental organizations, without curation, turning into dark data. In the public control area, where there are silos from several sources, with a growing volume, including citizens, dark data has been a topic not explored in the literature. This article brings the main concepts in dark data topic, listing its characteristics and risks, thus elaborating a conceptual map for the public control area. There is also the proposal of an approach that offers high abstraction for identification, classification, and monitoring of dark data, especially for the public control.*

Resumo. *Um volume considerável de dados permanece à margem do conhecimento de organizações governamentais, sem recurso de curadoria, transformando-se em dados escuros (dark data). Na área de controle público, onde há silos de diversas fontes, com um volume crescente, inclusive de cidadãos, dados escuros têm sido um tema não explorado pela literatura. Este artigo traz os principais conceitos na área de dados escuros, listando suas características e riscos, elaborando um mapa conceitual para a área de controle público. No decorrer do artigo, é apresentada uma abordagem de um pipeline para manipulação de dados escuros, que oferece alta abstração para identificação, classificação e monitoramento de dados escuros, especialmente para área de controle público.*

1. Introdução

O crescimento exponencial de informações eletrônicas nos últimos anos por meio de uma sociedade cada vez mais digital, só tende a aumentar, com o uso massivo de redes sociais, internet das coisas, entre outros aplicativos. Esse cenário se repete também para o Governo, especialmente, decorrente de políticas e tecnologias de Governo Eletrônico. No Brasil, o governo vem adotando o uso de tecnologias de informação para realização dos seus atos. À medida que sistemas entram em operação, cada vez mais dados são gerados, e a necessidade de adoção de plataformas de Big Data e recurso humano especializado para poder realizar a curadoria de tais dados se torna mais crítica [Henriques 2021].

Contudo, dada a rigidez do governo no processo de contratação, tanto de recurso tecnológico, como de pessoal por meio de concurso público, a realização de um processo de curadoria de dados se torna mais difícil. Aliado a esse cenário, a falta de padrão de desenvolvimento de software e a rotatividade de terceirizados e servidores tem contribuído para a perda de conhecimento dos dados. Tal fator é apenas um dos tantos exemplos que resultam no desconhecimento por parte do governo de dados que estão sob sua posse, dada a ausência de um processo de curadoria. As razões para a não

curadoria são várias, podendo estar relacionadas à falta de recursos financeiro/pessoal ou ausência, indisponibilidade de documentação [Heidorn 2008], ou um processo de falha ou abandono de determinado projeto [Goetz 2007].

Acredita-se que estes dados desconhecidos estejam concentrados na chamada "Longa cauda" do governo [Stahlman, Heidorn e Steffen 2018], se caracterizando assim como dados escuros. A "Longa cauda" consiste em um número relativamente pequeno de dados que têm como uma distribuição assimétrica à direita, conforme demonstra a Figura 1 [Heidorn 2008]. O início da "Longa cauda" é caracterizado pela separação do "limite conhecido", área que possui como característica a separação entre os dados conhecidos daqueles desconhecidos pela organização [Heidorn 2008].

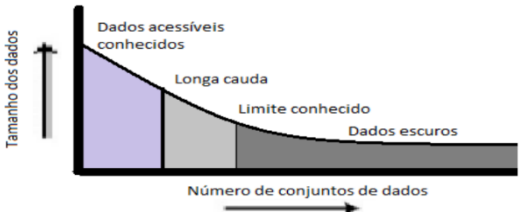

Figura 1 - "Longa cauda" de dados escuros. [HAWKINS et al 2020].

Tal área representa em sua grande parte um alto número de conjuntos de dados, mas de tamanhos não tão significativos quanto da área conhecida. Em grande parte da "Longa cauda" predominam os dados escuros, o que difere uma cauda da outra é relativo ao domínio de informação. Estima-se que, por exemplo, para dados da área biomédica, cerca de 85% são dados escuros [Macleod et al. 2014].; já no âmbito de dados oriundos de dispositivos de Internet das Coisas, cerca de 99% são escuros [Manyika et al. 2015].

Dados escuros representam um problema significativo, pois não são facilmente localizáveis, acessíveis e interoperáveis [Schembera 2021]. Os dados escuros podem ser tanto estruturados como não estruturados, mas sua parcela de maior significância tende à forma não estruturada. De acordo com Stewart[1] (2013), dados escuros geralmente são gerados por humanos e orientados para pessoas e não se encaixam perfeitamente em modelos relacionais, ou outro modelo mais estruturado. Como exemplo de tais dados, estão as conversas de bate-papo em linguagem natural, e-mail, imagens e vídeos. Para o governo, documentos digitalizados oriundos de processos administrativos, por exemplo. Tais informações não foram especificamente desenvolvidas para serem processáveis por máquinas, mas sim por humanos.

O governo precisa saber o que são os dados escuros e onde estão localizados. *Entendê-los* é o primeiro passo para assumir seu controle e posteriormente extrair seu valor para o exercício de política pública e especialmente de seu controle público. A função de controle de dados, exercida pelo governo, pode ser classificada como interno e externo. O interno é aquele exercido pela própria organização sobre seus atos, enquanto o externo é exercido por outra organização que goza de autonomia e independência. Como exemplo de controle interno e externo, estão respectivamente a

[1] Disponível em: https://blogs.gartner.com/darin-stewart/2013/05/01/big-content-the-unstructured-side-of-bigdata/ . Acessado em: 20/05/2022

Controladoria Geral da União (CGU) e o Tribunal de Contas da União (TCU). Segundo O'Donnell (1998), o mau funcionamento desses controles tem várias consequências (corrupção, ineficiência etc.).

Considerando que a avaliação da adesão/conformidade de atos do governo com as leis e normas é uma das principais funções do controle público. Em um cenário de dados escuros, essa avaliação é prejudicada pela dificuldade (ou impossibilidade) de obtenção da informação. Existem alguns trabalhos na literatura acerca de dados escuros [Heidorn 2008, Munot et al. 2019, Schembera 2021], contudo, nenhum possui como foco a sua intersecção com a área de controle público.

Este artigo discute essa lacuna, propondo a criação de uma abordagem adaptativa, de um pipeline para manipulação destes dados, que pode ser aplicada à área de controle público para identificar a presença de dados escuros e explorá-los. O artigo explora os principais conceitos de dados escuros, a fim de demonstrar evidências da literatura que possam caracterizar dados escuros no governo, detalhando suas características e tipos de dados com os respectivos exemplos e formatos.

Este artigo está organizado como segue. A Seção 2 traz uma visão geral acerca de dados escuros, mostrando características discutidas na literatura. A Seção 3 traz um primeiro resultado do levantamento bibliográfico através de um mapa conceitual de dados escuros para o controle público. A Seção 4 lança luz sobre a abordagem proposta. Por fim, a Seção 5 apresenta a conclusão e os trabalhos futuros.

2. Visão Geral

A seguir, são apresentadas algumas características importantes sobre o tópico de dados escuros, bem como uma breve discussão sobre os riscos e valores de tais dados para as organizações.

2.1. Características

Uma das primeiras caracterizações de dados escuros foi discutida em 2008, a partir de um trabalho de Bryan Heidorn [Heidorn 2008]. Nesse trabalho, dados escuros são definidos como dados que não são previamente indexados e armazenados de modo que são de difícil acesso para cientistas e outros usuários em potencial, e, portanto, é mais provável que permaneça subutilizado e eventualmente perdido. Outra definição vem do trabalho de Thomas Goetz [Goetz 2007], que associa dados escuros a dados que foram "cortados". Tais dados são escuros por conta de que seu conhecimento não alcança o âmbito científico e a comunidade. Em uma de suas colocações, o autor destaca: "o dado escuro pode ser um elo perdido de um pesquisador/engenheiro, o pedaço indescritível de dado que ele precisava".

Para a indústria, dados escuros podem ser definidos como dados que são coletados e armazenados a fim de utilizá-los no futuro para diversos fins, mas que acabam não sendo utilizados devido à falta de tempo e de recurso de máquina. Tais dados precisam ser processados, pesquisados e interpretados [Munot et al. 2019]. De acordo com Leonelli (2013), as organizações têm aplicado seus recursos para o gerenciamento a curto prazo dos dados. Para o armazenamento e gestão a longo prazo, muito menos recursos são empreendidos. Tal processo facilita o esquecimento, ou perda, desses dados, apesar dos custos envolvidos na produção, coleta, curadoria e no próprio armazenamento [Leonelli 2013].

Na literatura, percebe-se uma ampla gama de conceitualização de dados escuros, que são demonstradas na Tabela 1. Apesar das características serem distintas e até mesmo, em alguns cenários, equivalentes, nada impede que um determinado dado escuro tenha ao mesmo tempo várias delas.

Tabela 1 - Características mapeadas na literatura de dados escuros

Autor	Conceitos/características
[Hernández et al. 2018, Heidorn et al. 2018, Liu et al. 2021, Shukla et al. 2015, Trajanov et al. 2018, Macleod et al. 2014]	Dados nunca analisados ou inexplorados
[Liu et al 2021, De Sa et al. 2016]	Dados tipicamente não estruturados
[Zhang et al. 2016]	Dados que não são processáveis por meio de bancos de dados tradicionais
[Heidorn et al. 2018, Cafarella 2016]	Dados de difícil acesso
[Schembera 2021]	Dados abandonados ou oriundos de projetos fracassados
[Schembera 2021]	Dados não documentados
[Hawkins et al. 2020]	Dados nunca publicados
[Gimpel e Alter 2021]	Dados perdidos
Gartner, 2022[2]	Dados coletados e utilizados para uma finalidade, mas não para outras.
[Munot et al. 2019]	Dados que serão utilizados no futuro, sem curadoria atualmente

Diante das várias características apresentadas, os artigos buscam evidenciar uma pergunta em comum: como diminuir a "Longa cauda" de modo que dados escuros passem a ser conhecidos, e bem utilizados? As possíveis soluções que visam lançar luz sobre tais dados começaram com a proposta de criação de centros de ciência segregados por disciplinas, bibliotecas e museus [Heidorn 2008]. Em seguida, as primeiras propostas de ferramentas, tais como as demonstradas por Shukla (2015), De Sa (2016) e Schembera (2021). Há também proposição da criação de um novo papel de trabalho nas organizações [Schembera 2020]. Outras iniciativas envolvem uso de algoritmos de aprendizado de máquina e processamento de linguagem natural [Munot et al. 2019]. Todas essas soluções abordam tratamentos que visam dar suporte à descoberta de dados escuros. Para que haja a diminuição da "Longa cauda", é necessário não somente tratar

[2] Definição de Dark Data pela Gartner. Disponível em: https://www.gartner.com/en/information-technology/glossary/dark-data. Acesso em: 02/05/2022

a descoberta do dado escuro, mas abordar o problema holisticamente, desde o fato gerador da sua criação, bem como, o de sua "escuridão".

2.2. Riscos e valor

Alguns trabalhos exploram o valor singular que dados escuros podem possuir para contribuições científicas [Goetz 2007, Hawkins et al. 2020]. No cenário de organizações, Moumeni et al. (2021) cita a importância de extrair novos conhecimentos de dados escuros que podem agregar valor aos negócios das organizações, e contribuir para uma melhor tomada de decisão.

A tomada de decisão é um fator em que dados escuros podem desempenhar um papel crucial. Há de fato um potencial ganho de valor em explorá-los, bem como, há um potencial risco em não explorar. Munot (2019) associa dados escuros com risco de falhas no suporte à decisão, considerando que tais dados podem possuir informações relevantes para a tomada de decisão, e que uma vez no estado de dado escuro, permanece oculto para a organização.

Há também o risco associado por conta da sensibilidade que a informação pode ter, podendo assim comprometer a segurança da informação, bem como impedir a conformidade com normas ou leis [Moumeni et al. 2021]. Correlato à conformidade, também existe o risco de vazamento, considerando que o desconhecimento de tais dados implica também na ausência de monitoramento dos mesmos, podendo assim impactar na imagem da organização, bem como, em sanções legais considerando a vigência das normas da *General Data Protection Regulation* (GDPR) e Lei Geral de Proteção de Dados (LGPD) [Moumeni et al. 2021].

3. Dados escuros no controle público

A partir das características levantadas na literatura, foi possível estabelecer a criação de um mapa conceitual de dados escuros aplicados ao controle público. Vale destacar que nesse cenário, o que normalmente é escuro para o governo, também o é para o controle. Nesse mapa, é possível destacar as principais características associadas a dados escuros, conforme demonstrado na Figura 2.

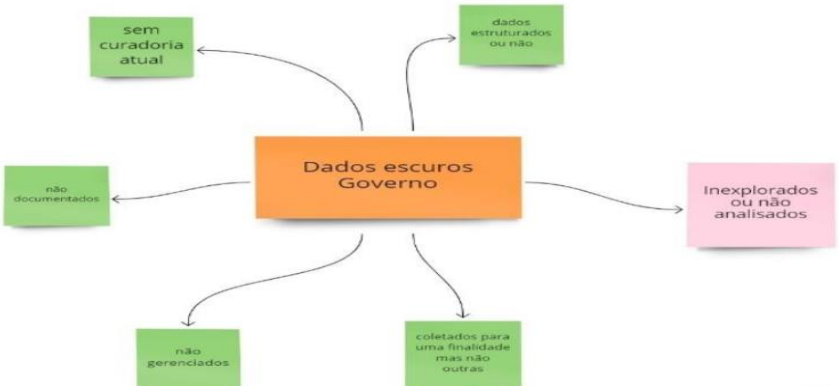

Figura 2 - Mapa conceitual de dados escuros para o controle público governamental

A Figura 2 demonstra a variedade de conceitos que pode ser aplicado para dados escuros no governo, em sua área de controladoria. No que tange a situação dos dados inexplorados ou não analisados, o governo tem uma enorme capacidade de criar

sistemas, coletar informações, contudo, boa parte desses dados permanece inexplorados. Esse cenário decorre tanto pela falta de curadoria apropriada, quanto pela ausência de mão de obra especializada, bem como, decorrente da ausência de tecnologias analíticas ou plataformas de big data.

Outra característica fundamental que explora o possível valor para dados escuros para o controle, está no fato do governo ter sob custódia vários dados coletados para uma finalidade, mas não usada para outras. Um exemplo disso, está nos dados de folha de pagamento dos servidores onde o CPF pode ser usado para cruzar com dados dos registros de óbitos, gerando assim uma nova informação "Falecidos na folha de pagamento", o que indicaria uma possível irregularidade cuja materialidade pode chegar a milhões de reais. Esse tipo de dado escuro é de grande representatividade para o exercício do controle público.

Apesar da literatura fazer algumas menções a dados não estruturados [Liu et al. 2021, De Sa et al. 2016], tal conceito não pode ser aplicado de forma restritiva ao governo. No âmbito governamental, é possível a existência de dados escuros de forma estruturada, em tabelas de bancos de dados, considerando as outras características apontadas no mapa conceitual, em especial, na amplitude de sistemas legados.

Os dados legados oriundos desses sistemas representam um grande desafio no setor público, demonstrando a dificuldade com o que o governo tem de compartilhar e organizar tais dados [Kim et al. 2014]. Os fatores como ausência de documentação, tecnologias legadas e rotatividade de pessoal contribuem para que a curadoria desses dados se perca, transformando-os assim em dados escuros. O valor e risco de tais dados ficam à margem do conhecimento da organização. Esse cenário se agrava com a rigidez de contratação de mão de obra e tecnologia.

O fato de uma parcela significativa dos dados escuros serem não estruturados reflete basicamente a grande capacidade do governo realizar seus atos a partir da geração de documentos eletrônicos (contratos, atos de admissão, empenhamento, emissão de nota fiscal, memorandos, ofícios etc.), os inúmeros dados decorrentes de interações sociais por meio de aplicativos de mensagens corporativos. Especialmente no contexto na continuidade do trabalho remoto, iniciado em decorrência da pandemia de COVID-19 em meados de março de 2020. Tais documentos não estruturados tem um potencial enorme para o exercício de controle público. Sua natureza não estruturada, volume e variabilidade, dificulta o alcance do controle. Nesse cenário, artefatos probatórios para fins de investigação que poderiam ser o elo para responsabilização ou expansão de atividades de fiscalização/investigação permanecem escuros ao controle.

Demonstra-se na Tabela 2, a partir de tais características, a seguinte distribuição de dados escuros no âmbito do controle público. Os dados refletidos na Tabela 2 são comuns de grande parte das organizações governamentais. No entanto, tal situação pode ser acrescida dada a natureza do órgão em questão, por exemplo, órgãos de governo ligados a exploração espacial [Gallaher et al. 2015, Heidorn 2018] ou de meio ambiente contam com dados de satélites e/ou de sensores, podendo ser considerados em parte, dados escuros.

Tabela 2 - Tipos de dados escuros para o controle público

Tipo de dado	Formato
Documentos de comunicação oficial	PDFs, arquivos texto
Documentos de admissão de pessoal, despesa e comprovação, processos licitatórios	PDFs, imagens, arquivos e textos
E-mails e aplicativos de comunicação oficial	Arquivos texto, imagens e vídeos
Sistemas legados	Tabelas relacionais
Documentos de arrecadação	Arquivos PDF e texto
Logs de sistemas	Arquivos texto
Informações obtidas de cruzamentos de dados	Arquivos, PDFs, tabelas relacionais
Tabelas relacionais legadas	Tabelas relacionais

4. Abordagem proposta

A seguir, é apresentada uma abordagem de um pipeline de um processo para manipulação de dados escuros. A abordagem proposta tem como aplicação o controle público governamental, contudo, possui adaptabilidade para outras organizações, considerando as características transversais de dados escuros em praticamente todos os domínios de informações. Possui uma iniciativa distinta das soluções da literatura, ao focar tanto na diminuição da "Longa cauda", quanto na sua expansão. A abordagem institui um processo de forma integrada para que se tenha maior governança sobre os dados, dificultando sua transformação para dados escuros e facilitando sua descoberta. Uma das principais contribuições para a literatura é trazer uma nova abordagem integrativa e holística para os dados escuros.

A visão holística dessa abordagem é uma estratégia crítica para o negócio. A partir do mapa conceitual elaborado anteriormente e dos fatores geradores discutidos como causas do surgimento de dados escuros, deve-se ter uma visão de toda arquitetura de geração e transformação desses dados, e não mais visualizá-los como componentes individuais e independentes de dados.

Para se alcançar a cobertura holística de ponta-a-ponta da organização, torna-se necessário que a abordagem tenha três eixos: estratégia, pessoal e processo. Em seu eixo "Estratégia", há a instituição de dois artefatos: a Política de Governança de Dados, Informações e Conhecimento (PGDIC). A PGDIC estabelece em um nível tático/estratégico os princípios, diretrizes, atribuições e responsabilidades para a gestão de curto e longo prazo de todo e qualquer ativo que contém dados, informação ou conhecimento. Outro artefato produzido nesse eixo é o Plano Diretor de Dados (PDD), documento revisto com periodicidade mínima anual. O PDD possui um viés operacional e deve operacionalizar o PGDIC, estando alinhado ao planejamento estratégico da organização. No PDD, conterão os projetos e operações, além das contratações necessárias para sua manutenção. A criação desses dois artefatos é importante pois marca um compromisso e patrocínio da alta administração com os dados, informações ou conhecimento que estão na organização. Em uma estrutura rígida ainda para muitas

organizações públicas, a instituição dessas duas políticas garante maior segurança e celeridade para as contratações tanto de pessoal como ferramental e serviço.

O eixo "Pessoal" traz a instituição de um escritório de *DataOps*, um termo recentemente cunhado por cientistas e engenheiros de dados, que se refere a um papel destinado a encurtar o tempo de vida do ciclo analítico de dados ponta-a-ponta, introduzindo a automação no processo de coleta, validação e verificação dos dados [Munappy et al. 2020]. *DataOps* promove a implementação de fato do PDD, e de forma colaborativa suporta todo trabalho com dados relacionados ao desenvolvimento de software e ativos de armazenamento. Para o governo, a criação desses papeis permitirá com que tenha recurso pessoal dedicado a atividade analítica, quer seja de forma terceirizada como de quadro próprio.

Por fim, o eixo "Processo" permite a integração da atuação dos *DataOps* e de artefatos tecnológicos. Nele é possível detalhar seis áreas de ações: identificar, coletar, centralizar, processar, estruturar e avaliar, conforme demonstra a Figura 3.

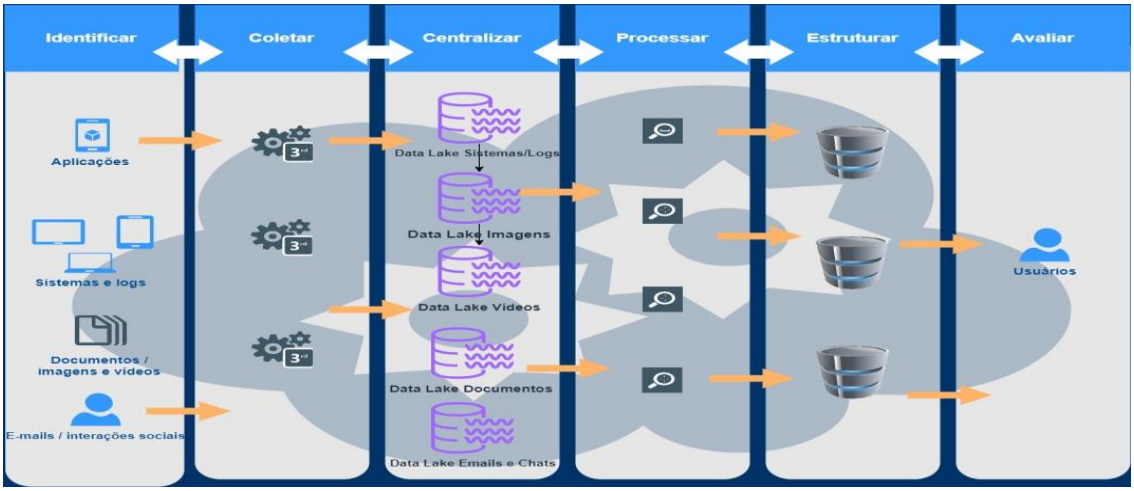

Figura 3 – Áreas de ações do eixo Processo, na abordagem proposta

A área de Identificar envolve a descoberta e catalogação de todo recurso computacional que gere informações, tais como: sistemas, e-mails, bate-papos, documentos vídeos, imagens e logs, por exemplo. Essa fase pode ser feita aliada ao mapeamento de catálogo de dados da Lei Geral de Proteção de Dados (LGPD) [Menegazzi 2021].

Após a área de Identificar, entra em operação a área de Coletar, que envolve a realização de mecanismos para a coleta contínua ou em lote de tais dados, a considerar a criticidade da informação. Nessa área, ferramentas de *Extraction Transform Load* (ETL) e demais ferramentas que constam em plataformas de big data podem contribuir para automatizar a integração das fontes identificadas na área de Identificar.

Após a saída da área de Coletar, existe a atuação da área Centralizar. Essa área visa agrupar a cópia dos dados coletados da área anterior em estruturas de lagos de dados. Possui também uma separação em lagos de dados distintos, ativos de informação estruturados e não estruturados. Bem como, separação de lagos em que há um domínio de seu conteúdo com processos de curadoria já estabelecidos e amadurecido dos que não possuem.

A área de Processar vem em seguida à área de Centralizar. Tal área permite que haja de fato a implementação da extração de valor nessas fontes. Importante destacar que tal processo é fundamental que seja executado em clusters de *big data*, para usufruir do poder de processamento paralelo e distribuído, dado o cenário de alta volumetria de variabilidade, especialmente considerando o volume de dados escuros para governo.

Essa área pode se beneficiar de soluções já existentes ou adaptadas da literatura [De SA et al. 2016 Shukla et al. 2015]. Nesse processo também, pode fazer uso de técnicas de aprendizado de máquina, mineração de texto, inferência estatística, e outras técnicas computacionais, a depender do tipo de extração de valor que a organização governamental necessite. O processamento de dados escuros contribui para agregar valor ao exercício do controle. Em alguns cenários típicos de exercício de atividades ligadas ao controle público, a busca de documentos cujo conteúdo contém determinados valores como CPF/CNPJ, nomes de pessoas ou empresas pode definir o rumo de uma ação de investigação/auditoria, especialmente sua relação semântica com outras palavras. Técnicas computacionais ligadas a descoberta de relações semânticas similares entre estruturas de dados distintas também são fundamentais. Destaca-se também que tal área não visa apenas extrair valor dos dados escuros, mas também possíveis riscos que estão em dados escuros, que podem dar suporte à área de segurança de informação e proteção de dados.

A área Estruturar tem como objetivo principal transformar a informação processada pelo estágio anterior, de uma fonte tipicamente não estruturada em uma informação estruturada mapeada em um *schema*, permitindo que o resultado da busca e o monitoramento das informações extraídas seja efetivo.

Por fim, há a área Avaliar, que possui dois desdobramentos: o primeiro, a partir da execução das áreas anteriores, institui-se variáveis para estabelecimento de métrica de avaliação do grau de exposição daquela organização em dados escuros. O segundo desdobramento consiste em avaliar de forma contínua as métricas a partir dos resultados obtidos das técnicas/ferramentas da área de Processar. Dessa forma, estabelece-se um monitoramento da expansão/contração da "Longa Cauda" e permite a descoberta de dados escuros.

5. Conclusão e Trabalhos Futuros

Dados escuros na seara do controle público governamental é um tópico inexplorado na literatura. A partir da investigação das principais características que envolvem dados escuros, foi estabelecido quais delas e que tipos de dados escuros estão associados no âmbito do governo. O resultado foi a geração de um mapa conceitual. O mapa conceitual demonstra os possíveis fatores geradores de dados escuros, para que se possa adotar ações com vias de atender um dos problemas principais dos dados escuros: como diminuir a "Longa Cauda" de dados escuros?

Essa problemática atinge o governo, em especial, sua função de controle, o qual é fundamental para o correto funcionamento do Estado e da conformidade com as normas/leis. Os dados escuros dificultam o exercício do controle pois dada sua natureza de difícil acesso e/ou processamento, estão à margem do conhecimento das organizações públicas. Nesse cenário, seu conteúdo pode conter valor significativo suficiente para responsabilização de irregularidades, elementos probatórios e qualquer suporte informacional à atividade de controle.

Diante dessa problemática e das características levantadas na elaboração do mapa conceitual, foi necessário trazer uma abordagem distinta das soluções encontradas na literatura: tratar o problema de dados escuros de forma holística. Fundamental também, dada a natureza *sui-generis* de dados escuros que essa abordagem seja adaptável não só para o governo, mas também para as demais organizações.

A natureza holística da abordagem visa atender fatores críticos de geração de dados escuros. Tais fatores foram mapeados e traduzidos em três eixos: estratégia, pessoal e processo. A estratégia é fundamental pois em entes governamentais, todas as ações, inclusive aquelas no nível operacional, devem estar amparadas em normas e leis. Para isso, a criação de políticas e planos chancelados pela alta administração promove respaldo legal e permite melhor planejamento de compras e contratações para a execução.

O eixo pessoal permite preencher um "vácuo" na administração pública, de forma geral, no quesito de mão de obra. Diante da dificuldade de contratação de pessoal de quadro próprio decorrente do limite de despesa de pessoal oriundo da Lei de Responsabilidade Fiscal (LRF). A instituição de *DataOps* por esse eixo visa criar uma cultura analítica no âmbito do governo, contribuindo para a gestão tanto curto quanto longa dos dados da organização. A atuação de *DataOps* em dados escuros ainda não foi relatada, permanecendo uma oportunidade de contribuição científica inédita.

Por fim, o eixo de processo, contribui para que de fato os dois eixos anteriores consigam materializar suas ações. Esse eixo institui áreas que se desdobram em ações, métodos/técnicas/ferramentas de forma que conjuntamente consigam não somente controlar a expansão da "Longa Cauda" de dados escuros, mas também, diminuí-la. Nesse sentido, permite maior conhecimento do órgão governamental sobre seus dados, maximizando o valor e diminuindo riscos. A instituição de uma métrica de dados escuros por esse eixo abre uma oportunidade inédita de contribuição para a literatura. Outra possível contribuição de destaque da abordagem reside no aspecto de tratamento de dados legados, que são considerados escuros de acordo com levantamento feito na Tabela 2. A abordagem permitirá criar um "*pipeline*" de dados contribuindo para seu acesso e conhecimento, podendo ser avaliada como um modelo de referência de migração de dados legados.

Torna-se fundamental, como trabalho futuro, a implementação dessa abordagem para colher os resultados, avaliá-los e discuti-los para possíveis melhorias e obtenção de feedbacks dos gestores da organização. No aspecto tecnológico, cabe ter como premissa o uso de software *open-source*, permitindo assim com que a área de governo se beneficie em termos de custos. Estruturas que permitam armazenar uma grande variabilidade de dados como o Hadoop, pode cumprir um papel transversal nas áreas identificadas, especialmente na área Centralizar. Sua estrutura HDFS permite armazenar dados tanto estruturados com diferentes motores de armazenamento/busca de acordo com a natureza do dado (Apache Impala, Kudu e ORC), como não-estruturados a partir de motores de indexação como o Apache Solr. Considerando a possibilidade de se ter conjuntos de dados muito grandes, a abordagem deve ter como premissa a execução sob uma arquitetura Lambda, com uma camada em lote e outra de *streaming*. Ferramentas como Apache Kafka e Sqoop podem cumprir o papel que permeia a área de Coletar. Uma das ferramentas a serem estudadas da área de Identificar é o Apache Atlas. Ele permite a catalogação automática e classificação de qualquer dado dentro do hadoop,

atuando como uma base de gestão de conhecimento para dados escuros. Para a camada de Processar e Estruturar, o Apache Spark e Solr podem contribuir para a busca de informações e novas descobertas. Nessa área, técnicas ligadas a área de *Data Science*, citadas na literatura [Munot et al. 2019] podem ser aplicadas e validadas, com foco em atender os principais tipos de dados escuros elencados na Tabela 2. Para além do aspecto tecnológico, torna-se necessário aprofundar o eixo de pessoal e de estratégia e como eles podem juntamente com a eixo de processo criar métricas para o diagnóstico de dados escuros na organização.

Referências

Cafarella, M., et al. (2016). "Dark Data: Are we solving the right problems?". *In 2016 IEEE 32nd International Conference on Data Engineering (ICDE)* (pp. 1444-1445). IEEE.

De Sa, C., et al, C. (2016). "Deepdive: Declarative knowledge base construction." ACM SIGMOD Record, 45(1), 60-67.

Gallaher, D., et al. (2015). "The process of bringing dark data to light: The rescue of the early Nimbus satellite data." GeoResJ, v. 6, p. 124-134, 2015.

Gimpel, G.; Alter, A. (2021). "Benefit From the Internet of Things Right Now by Accessing Dark Data". IT Professional, v. 23, n. 2, p. 45-49, 2021.

Goetz, T. (2007). "Freeing the dark data of failed scientific experiments." *Wired Magazine*, 15(10), 15-10.

Hawkins, B.; et al. (2020). "Data dissemination: shortening the long tail of traumatic brain injury dark data*". Journal of neurotrauma*, 37(22), 2414-2423.

Heidorn, P. (2008). "Shedding light on the dark data in the long tail of science." Library trends, 57(2), 280-299

Heidorn, P., et al. (2018). "Astrolabe: curating, linking, and computing astronomy's dark data". *The Astrophysical Journal Supplement Series*, 236(1), 3.

Henriques, A. (2021). *Big data analytics para o desenvolvimento humano: um estudo no Governo Federal Brasileiro*. Tese de Doutorado. Fundação Getúlio Vargas – FGV.

Hernández, D., et al. (2018). "Bauspace: a scalable infrastructure for soft sensors development". *In Proceedings of the 47th International Conference on Parallel Processing Companion* (pp. 1-4).

Leonelli, S. (2013). "Why the current insistence on open access to scientific data?". Big data, knowledge production, and the political economy of contemporary biology. *Bulletin of Science, Technology & Society*, 33(1-2), 6-11.

Liu, Y., et al. (2021). "Deep Hash-based Relevance-aware Data Quality Assessment for Image Dark Data". *ACM/IMS Transactions on Data Science*, 2(2), 1-26.

Kim, G., et al. (2014). "Big-data applications in the government sector". Communications of the ACM, v. 57, n. 3, p. 78-85.

Macleod, M., et al (2014). "Biomedical research: increasing value, reducing waste". *The Lancet*, 383(9912), 101-104.

Manyika, J., et al (2015). "The Internet of Things: Mapping the value beyond the hype". (Vol. 24). New York, NY, USA: McKinsey Global Institute.

Menegazzi, D. (2021). "Um guia para alcançar a conformidade com a LGPD por meio de requisitos de negócio e requisitos de solução." Dissertação de Mestrado. Universidade Federal de Pernambuco.

Moumeni, L., et al. (2021). "Dark data as a new challenge to improve business performances: review and perspectives". *In 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)* (pp. 216-220). IEEE.

Munappy, A., et al. (2020). "From ad-hoc data analytics to dataops". In *Proceedings of the International Conference on Software and System Processes* (pp. 165-174).

Munot, K., et al. (2019). "Importance of Dark Data and its Applications". *In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1-6). IEEE.

O'Donnell, G. (1998). "Accountability horizontal e novas poliarquias". In Revista Lua Nova, Nº 44; São Paulo. CEDEC

Schembera, B. (2021). "Like a rainbow in the dark: metadata annotation for HPC applications in the age of dark data". *The Journal of Supercomputing*, 77(8), 8946-8966.

Schembera, B., e Durán, J. (2020). "Dark data as the new challenge for big data science and the introduction of the scientific data officer". *Philosophy & Technology*, 33(1), 93-115.

Shukla, M., et al. (2015). "POSTER: WinOver enterprise dark data". *In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (pp. 1674-1676).

Stahlman, G., Heidorn, P., e Steffen, J. (2018). "The astrolabe project: identifying and curating astronomical 'dark data' through development of cyberinfrastructure resources". *In EPJ Web of Conferences* (Vol. 186, p. 03003). EDP Sciences.

Trajanov, D., et al (2018). "Dark data in internet of things (IOT): challenges and opportunities". *In 7th Small Systems Simulation Symposiu*m (pp. 1-8).

Zhang, C., et al. (2016). "Extracting databases from dark data with deepdive". *In Proceedings of the 2016 International Conference on Management of Data* (pp. 847-859).