2ª COLETÂNEA DE ARTIGOS WEB

ceweb.br nic.br cgi.br

2ª COLETÂNEA DE ARTIGOS WEB

2024

FICHA TÉCNICA

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR — NIC.BR

Diretoria:

Demi Getschko (Diretor Presidente)

Hartmut Richard Glaser (Diretor de Assessoria às atividades do CGI.br)

Ricardo Narchi (Diretor Administrativo)

Frederico Neves (Diretor de Serviços e Tecnologia)

Milton Kaoru Kashiwakura (Diretor de Projetos Especiais e de Desenvolvimento)

PRODUÇÃO DESTA PUBLICAÇÃO

Centro de Estudos sobre Tecnologias Web - Ceweb.br

COORDENAÇÃO EXECUTIVA

Vagner Diniz

Selma de Morais

Beatriz Rocha

PREPARAÇÃO E REVISÃO TEXTUAL

Érica Santos Soares de Freitas

PROJETO GRÁFICO

Maricy Rabelo (Comunicação NIC.br)

DIAGRAMAÇÃO

Giuliano Galves (Comunicação NIC.br)

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Núcleo de Informação e Coordenação do Ponto br
2ª Coletânea de artigos Web / Núcleo de Informação e Coordenação do Ponto br; [organização Selma Morais ; coordenação Vagner Diniz ; tradução Érika dos Santos Soares de Freitas]. -- São Paulo: Núcleo de Informação e Coordenação do Ponto br, 2024

Título original: Web.
Bibliografia.
ISBN 978-65-85417-40-2

1. Acessibilidade 2. Artigos - Coletâneas 3. Dados abertos 4. Inteligência artificial 5. Tecnologias digitais 6. WEB (Linguagem de programação) I. Morais, Selma. II. Diniz, Vagner. III. Título

24-196228 CDD-004

Índices para catálogo sistemático:
1. Artigos : Coletâneas : Tecnologias Web 004
Tábata Alves da Silva - Bibliotecária - CRB-8/9253

FICHA TÉCNICA DOS AUTORES DOS ARTIGOS

ADRIANO C. M. PEREIRA - Professor associado do departamento de Ciência da Computação (DCC) da Universidade Federal de Minas Gerais (UFMG).

ANA ELIZA DUARTE - Analista de projetos Web do Centro de Estudos sobre Tecnologias Web (Ceweb.br) e do Núcleo de Informação e Coordenação do Ponto BR (NIC.br).

ANA LUÍSA FREITAS - Pesquisadora do Laboratório de Neurociência Cognitiva e Social (SCN Lab) da Universidade Presbiteriana Mackenzie (UPM) e do Instituto Nacional de Ciência e Tecnologia em Neurociência Social e Afetiva (INCT SANI).

CAROLINE BURLE - Co-Fundadora da IiBertha.org e da Coalizão Licença Paternidade (CoPai).

DIOGO CORTIZ - Especialista em projetos Web do Centro de Estudos sobre Tecnologias Web (Ceweb.br) e do Núcleo de Informação e Coordenação do Ponto BR (NIC.br), e professor da Pontifícia Universidade Católica de São Paulo (PUC-SP).

FERNANDA N. PANTALEÃO - Mestranda do Programa de Distúrbios do Desenvolvimento da Universidade Presbiteriana Mackenzie (UPM) e membro do Laboratório de Neurociência Cognitiva e Social (SCN Lab) da Universidade Presbiteriana Mackenzie (UPM) e do Instituto Nacional de Ciência e Tecnologia em Neurociência Social e Afetiva (INCT SANI).

HENRIQUE S. XAVIER - Especialista em projetos Web do do do Centro de Estudos sobre Tecnologias Web (Ceweb.br) e do Núcleo de Informação e Coordenação do Ponto BR (NIC.br).

JOÃO BÁRBARA - Analista de Dados do departamento de Ciência da Computação (DCC) da Universidade Federal de Minas Gerais (UFMG).

LETÍCIA DA SILVA MACEDO ALVES - Bolsista do Projeto TIC Web/Universidade Federal de Minas Gerais (UFMG)

LETÍCIA Y. N. MORELLO - Mestre em Distúrbios do Desenvolvimento pela Universidade Presbiteriana Mackenzie (UPM) e membro do Laboratório de Neurociência Cognitiva e Social (SCN Lab) e do Instituto Nacional de Ciência e Tecnologia em Neurociência Social e Afetiva (INCT SANI).

PAULO S. BOGGIO - Diretor do do Laboratório de Neurociência Cognitiva e Social (SCN Lab) da Universidade Presbiteriana Mackenzie (UPM) e do Instituto Nacional de Ciência e Tecnologia em Neurociência Social e Afetiva (INCT SANI).

REINALDO FERRAZ - Gerente adjunto do Centro de Estudos sobre Tecnologias Web (Ceweb.br) e do Núcleo de Informação e Coordenação do Ponto BR (NIC.br).

WAGNER MEIRA JÚNIOR - Professor titular do departamento de Ciência da Computação (DCC) da Universidade Federal de Minas Gerais (UFMG).

Sumário

APRESENTAÇÃO

Na primeira edição da Coletânea de Artigos Web (NIC.BR, 2022)[1] , apresentamos os principais artigos escritos, publicados e apresentados pelos profissionais do Centro de Estudos sobre Tecnologias Web (Ceweb) e seus parceiros no ano de 2022, cujo foco abordou a importância do conceito de tecnologias abertas da Web, seja no compartilhamento de informações, na eliminação de barreiras de acesso a essas tecnologias ou mesmo em seu uso no campo da Inteligência Artificial (IA).

Quem ainda não leu a primeira edição, vale a pena ver, por que os artigos são atemporais.

Nessa segunda edição, reunimos artigos do 2023 e estão disponíveis para acesso não só como forma de compartilhar conhecimento, mas também para celebrar os nove anos de existência do Ceweb. São anos de estudos, experimentos e muita escrita para mostrar o potencial transformador da Web.

Dessa vez, os leitores encontrarão artigos com proposições inovadoras de uso de tecnologias web bem interessantes. São experimentos no campo de dados abertos, livro digital e IA. Ademais, antenados com o tema do momento, concluímos com uma reflexão sobre as limitações "morais" da IA.

O artigo **Modelo de Avaliação de Dados Abertos nos Portais Governamentais Brasileiros** foca na falta de dados bem estruturados na Web, sem considerar os padrões existentes. Nesse sentido, aborda e propõe modelo e mecanismos capazes de orientar os publicadores a utilizarem princípios e boas práticas sobre publicação de dados abertos na Web.

O que muitos ainda não sabem é que o livro digital ainda não consegue se igualar ao livro impresso quando se trata de livro didático. O artigo **Um livro didático digital acessível e conectado** revela o potencial da interatividade do formato ePub para uso como material didático, a fim de que alunos respondam exercícios nos livros digitais.

O artigo **Uma solução para verificação de conformidade ao padrão de acessibilidade entre *sites* governamentais brasileiros** apresenta uma solução inovadora para verificação de acessibilidade de sites e suas páginas. É uma plataforma aberta que compreende todo o processo de verificação: a coleta, o processamento e a apresentação de resultados do nível de conformidade dos sites aos padrões internacionais de acessibilidade na Web.

---

1 NÚCLEO DE INFORMAÇÃO E COMUNICAÇÃO DO PONTO BR (NIC.BR). Coletânea de Artigos Web. São Paulo: NIC.br, 2022. Disponível em: https://acervo.ceweb.br/acervos/conteudo/1a5817c-5-e06c-4305-9b08-e88388b10551

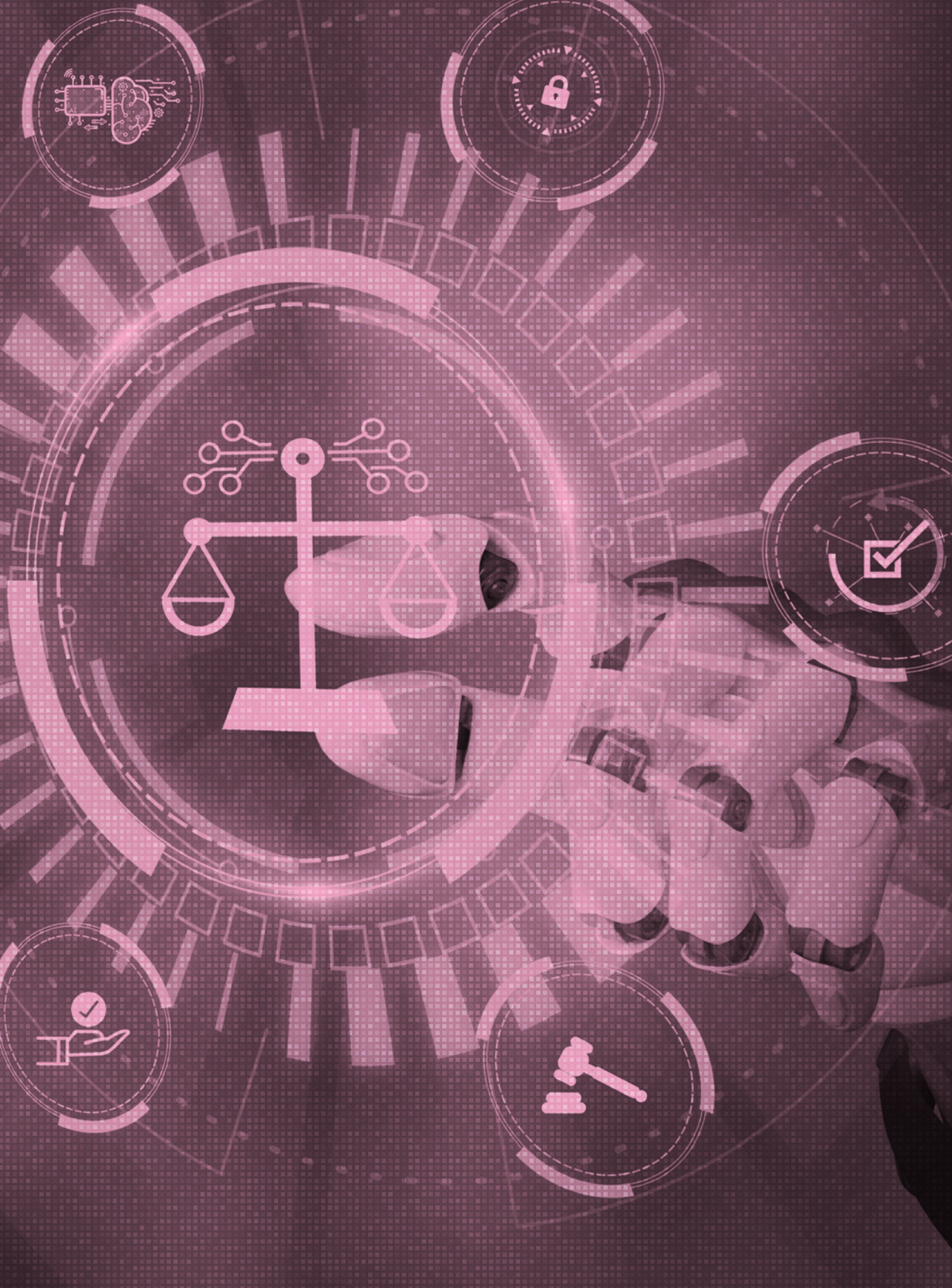
No campo do uso das tecnologias de IA, em especial do Processamento de Linguagem de Máquina e Aprendizagem de Máquina, o artigo **Fiscalização do governo com IA** apresenta uma solução simples para controle social de atos governamentais que monitora e seleciona as matérias mais relevantes, segundo o interesse do usuário, tendo o Diário Oficial da União como parte de um trabalho de fiscalização do Poder Executivo.

Para concluir a nossa coletânea de artigos, a partir da possibilidade de transferir para aplicações baseadas em IA a capacidade de tomar decisões, o artigo **Estamos preparados para os agentes morais artificiais?** propõe uma reflexão sobre ética e IA e como usar uma "inteligência" 100% baseada em regras, em um contexto no qual as variáveis em larga medida são abstratas, por exemplo as emoções e os sentimentos. Conforme o texto, "atribuir às IA a tomada de decisões em contextos morais é uma tarefa difícil devido à falta de consenso social sobre o que constitui uma decisão justa."

Desejamos uma boa leitura! Não deixem de compartilhar nossa revista. Queremos que ela circule. E, como qualquer recurso na Web, ela tem um endereço único para ser encontrada e compartilhada.

Boa leitura!

Vagner Diniz,
Gerente do Centro de Estudos sobre Tecnologias Web (Ceweb.br)



ESTAMOS PREPARADOS PARA OS AGENTES MORAIS ARTIFICIAIS?[2]

Por **Letícia Y. N. Morello**, **Fernanda N. Pantaleão**, **Ana Luísa Freitas**, **Paulo S. Boggio** e **Diogo Cortiz**

2 Tradução do artigo *"Are we ready for artificial moral agency?"* apresentado originalmente no *workshop Moral Agents* da Conferência CHI'23 (MORAL AGENTS, 2023).


AI

Ao discutirmos a moralidade em relação à Inteligência Artificial (IA), idealizamos uma máquina capaz de incorporar e aplicar princípios morais para que possa tomar decisões por nós. Os princípios morais referem-se a uma construção complexa e abstrata que descreve a base das regras morais; são o núcleo da moralidade e ajudam a proteger o bem-estar, os direitos, a equidade e a justiça para todos e em vários contextos (KILLEN; DAHL, 2018). Além disso, são diretrizes para o comportamento ético e respeitam a autonomia, a não maleficência, a beneficência, a justiça, a honestidade e a equidade, tornando, assim, os julgamentos éticos menos arbitrários e mais racionais (UHLMANN *et al.*, 2009).

Os humanos são seres morais porque têm consciência de seu livre-arbítrio e a capacidade de escolher e agir de uma forma ou de outra (BAUMEISTER, 2018), com base em inferências sobre os estados mentais dos demais (WAYTZ; YOUNG, 2018), em cenários hipotéticos. A moralidade também requer a capacidade de tomada de decisão autônoma, o que significa que as regras arbitrárias nem sempre se aplicam a situações morais e devem ser reavaliadas (KILLEN; DAHL, 2018). Os seres morais devem, ainda, ter uma percepção das necessidades, como a autopreservação ou a sobrevivência, relacionadas com experiências básicas e integradas, como a capacidade de sentir dor e prazer (KNOBE, 2018). Desse modo, surgem os agentes morais: eles enfrentam uma interação social pela qual têm a oportunidade de escolher como agir, de forma correta ou incorreta, com relação a outra pessoa (GRAY; WEGNER, 2009).

A espécie humana, quando confrontada com contextos morais, é suscetível de sentir orgulho, culpa, medo de retaliação e compaixão por vítimas, pois sabe que i) pode sentir dor e que outras pessoas podem nos causar danos, ii) depende de outras pessoas para viver bem, e iii) eventualmente morrerá. Nada disso, todavia, é aplicável às IA. Os valores e as emoções pessoais guiam e influenciam as decisões humanas. O cérebro, por exemplo, percebe a injustiça mais rapidamente do que é possível se pensar conscientemente

sobre ela, uma atribuição de valor subjetivo muito precoce que pode evocar indignação e vontade de corrigir as coisas. Há um conjunto de emoções ao qual se pode denominar "emoções morais": aquelas ligadas ao bem-estar e aos interesses da sociedade ou de outros indivíduos para além da pessoa que está julgando ou agindo (HAIDT, 2003), as quais servem como força motivacional que capacita os indivíduos a se envolverem em atos virtuosos e a se absterem de causar danos (KROLL; EGAN, 2004).

Um bom exemplo de uma emoção moral é a culpa. Quando se faz algo considerado errado, surgem alguns sinais inatos que podem provocar arrependimento quanto à ação e até insônia, consequências que podem levar a um pedido de desculpa ou à tentativa de desfazer ou diminuir as consequências negativas do erro. As IA, por sua vez, limitam-se a tomar decisões "sem olhar para trás", o que leva a inferir que, embora possam ser grandes decisores, ou "agentes de decisão", não são agentes morais em sua essência (pelo menos ainda não): não têm as capacidades mentais que os seres humanos tanto apreciam, como a autoconsciência, a experiência subjetiva e capacidades de aprendizagem sofisticadas (SULLINS, 2009).

A forma de raciocínio das IA baseia-se nos dados introduzidos para seguir uma função matemática que gera resultados. No entanto, considerando que elas podem aprender a partir de dados existentes (aprendizagem supervisionada) ou de interações diretas com o ambiente (aprendizagem por reforço), é realmente difícil, se não impossível, determinar a partir de quais regras aprendem para basear suas decisões (CORTIZ, 2019). Seu comportamento é determinado por programas que interagem com dados e ambientes, tornando impossível assumi-los como agentes com qualquer grau de livre-arbítrio (SULLINS, 2009), uma caraterística central para o desenvolvimento da moralidade. Além disso, as IA não têm acesso a todo o contexto quando tomam uma decisão nem têm emoções para guiá-las. Ainda que sejam capazes de tomar decisões generalizadas e basea-

das em algoritmos, não têm a capacidade de fazer concessões, reconsiderar, relevar e perdoar, o que também faz parte da moralidade. Por isso, até hoje, defende-se que os humanos são os únicos seres que podem ser considerados agentes morais *stricto sensu*.

Atribuir às IA a tomada de decisões em contextos morais é uma tarefa difícil devido à falta de consenso social sobre o que constitui uma decisão justa. Consequentemente, alguns investigadores de IA sugerem que uma forma de abordar essa questão é, ao menos, evitar consequências injustas – decisões que prejudiquem ou beneficiem indivíduos ou grupos com base em características irrelevantes (CORTIZ, 2022). Embora este seja um excelente ponto de partida, o que devemos considerar exatamente como relevante em diferentes circunstâncias?

Quando os seres humanos precisam tomar uma decisão, baseiam-se, mesmo que inconscientemente, em diversos fatores, tais como cultura, estado emocional no momento da decisão, crenças, história pessoal, presença de outras pessoas, opinião de líderes, racionalidade sobre julgamentos, entre outros (SAATY, 2008; LE TEXIER, 2019). Mesmo que duas pessoas considerem exatamente os mesmos fatores para tomar uma decisão, podem escolher caminhos diferentes devido ao peso que atribuem a esses fatores.

Ao criar ou treinar uma IA, é preciso considerar que a racionalidade das pessoas é limitada, portanto é altamente improvável, para não dizer impossível, que as máquinas construídas por seres humanos estejam livres dessas limitações. Mesmo partindo do princípio de que não se pretende que uma IA atinja a complexidade da mente humana (DIGNUM, 2019), não se está tentando criar máquinas que ajudem a resolver problemas; em vez disso, tudo indica que se almejam sistemas que possam tomar o lugar das pessoas como agentes.

Assim, tentamos delegar nossos deveres morais para evitar a responsabilidade e a eventual culpa, um fenômeno conhecido

como "difusão de responsabilidade" (BLEHER; BRAUN, 2022). Mais grave ainda: estamos transferindo nossa responsabilidade humana para um sistema, que, em sua essência, não pode ser diretamente culpabilizado e julgado caso tome uma decisão tendenciosa. Como as IA são criadas por seres humanos, são suscetíveis, por conseguinte, a vieses (ANGWIN *et al*., 2016; LEAVY, 2018; OBERMEYER *et al*., 2019).

Para além das diferenças individuais, o construto "moralmente bom" varia culturalmente (GREENE, 2013), o que impõe o dever e o grande esforço de se construírem modelos que contemplem especificidades culturais. Atualmente, a maioria dos modelos de IA utilizados em todo o mundo é desenvolvida por empresas do Norte Global. No entanto, é incerto se esses modelos foram treinados com conjuntos de dados que representem as culturas locais. Essa disposição, portanto, pode resultar em um cenário no qual determinada perspectiva moral seja enraizada nas culturas locais mediante a agência algorítmica. Mesmo que se construam modelos com base em várias culturas, quem decidirá se o resultado é o melhor? Antes disso: quem definirá os valores e seus níveis de prioridade ao escrever as regras?

Em verdade, o mais sensato é supervisionar constantemente as decisões morais tomadas pelas IA. Em geral, automatizamos as funções humanas para otimizar o tempo e o esforço; contudo, no em que cenários sociais devemos utilizar a IA? O mais importante: mesmo que, no futuro, as IA aprendam a tomar decisões morais reais por si próprias, as pessoas estarão realmente dispostas a enfrentar as consequências de deixarem que as máquinas tomem esse tipo de decisões em seu lugar?

REFERÊNCIAS

ANGWIN, J. *et al*. Machine bias. *ProPublica*, 23 maio 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em 9 out. 2023.

BAUMEISTER, R. F. Free will and moral psychology. *In*: GRAY, K.; GRAHAM, J. (ed.). *Atlas of Moral Psychology*. New York: The Guilford Press, 2018. p. 332-337. Disponível em: https://philpapers.org/archive/BISNAA-4.pdf. Acesso em 9 out. 2023.

BLEHER, H.; BRAUN, M. Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, v. 2, p. 747-761, 24 jan. 2022. Disponível em: https://doi.org/10.1007/s43681-022-00135-x. Acesso em 8 out. 2023.

CORTIZ, D. O Design pode ajudar na construção de Inteligência Artificial humanística? *In*: Congresso Internacional de Ergonomia e Usabilidade de Interfaces Humano-Tecnologia, 17; Congresso Internacional de Ergonomia e Usabilidade de Interfaces e Interação Humano-Computador, 17, Rio de Janeiro, 11-13 dez. 2019. *Anais* [...]. São Paulo: Blucher, 2019. p. 14-22. Disponível em: https://doi.org//10.5151/ergodesign2019-1.02. Acesso em 8 out. 2023.

CORTIZ, D. A narrative review of fairness and morality in neuroscience: Insights to artificial intelligence. *AI and Ethics*, v. 3, p. 769-780, 10 ago. 2022. Disponível em: https://doi.org/10.1007/s43681-022-00203-2. Acesso em 9 out. 2023.

DIGNUM, V. *Responsible artificial intelligence*: how to develop and use AI in a responsible way. Cham: Springer, 2019.

GRAY, K.; WEGNER, D. M. Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of personality and social psychology*, v. 96, n. 3, p. 505-520, 2009. Disponível em: https://psycnet.apa.org/doi/10.1037/a0013748. Acesso em 9 out. 2023.

GREENE, J. *Moral tribes*: Emotion, reason, and the gap between us and them. New York: Penguin, 2013.

HAIDT, J. Elevation and the positive psychology of morality. *In*: KEYES C. L. M.; HAIDT, J. (ed.). *Flourishing*: Positive psychology and the life well-lived. Worcester: American Psychological Association, 2003. p. 275-289. Disponível em: https://doi.org/10.1037/10594-012. Acesso em 9 out. 2023.

KILLEN, M.; DAHL, A. Moral judgment: Reflective, interactive, spontaneous, challenging, and always evolving. *In*: GRAY, K.; GRAHAM, J. (ed.). *Atlas of moral psychology*. New York: The Guilford Press, 2018. p. 20-30. Disponível em: https://philpapers.org/archive/BISNAA-4.pdf. Acesso em 9 out. 2023.

KNOBE, J. There is no important distinction between moral and nonmoral cognition. *In*: GRAY, K.; GRAHAM, J. (ed.). *Atlas of moral psychology*. New York: The Guilford Press, 2018. p. 556-564. Disponível em: https://philpapers.org/archive/BISNAA-4.pdf. Acesso em 9 out. 2023.

KROLL, J.; EGAN, E. Psychiatry, moral worry, and the moral emotions. *Journal of Psychiatric Practice*, v. 10, n. 6, p. 352-360, 2004. Disponível em: https://pubmed.ncbi.nlm.nih.gov/15583516/. Acesso em 9 out. 2023.

LEAVY, S. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. *In*: International workshop on gender equality in software engineering, 1, maio 2018. *Anais* […], New York: Association for Computing Machinery, 2018. p. 14-16. Disponível em: https://doi.org/10.1145/3195570.3195580. Acesso em 9 out. 2023.

LE TEXIER, T. Debunking the Stanford Prison Experiment. *American Psychologist*, v. *74*, n. 7, 823-839, 2019. Disponível em: http://dx.doi.org/10.1037/amp0000401. Acesso em 9 out. 2023.

MORAL Agents for Sustainable Transitions: a CHI 2023 Workshop. *Moral Agents*, 2023. Disponível em: https://www.moralagents.org/home. Acesso em 20 nov. 2023.

OBERMEYER, Z. *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, v. 366, n. 6464, p. 447-453, 2019. Disponível em: https://doi.org/10.1126/science.aax2342. Acesso em 9 out. 2023.

SAATY, T. L. Decision making with the analytic hierarchy process. *International journal of services sciences*, v. 1, n. 1, p. 83-98, 2008. Disponível em: https://www.rafikulislam.com/uploads/resourses/197245512559a37aadea6d.pdf. Acesso em 9 out. 2023.

SULLINS, J. P. Artificial moral agency in technoethics. *In*: LUPPICINI, R; ADEL, R. *Handbook of research on technoethics*. Hershey: Information Science Reference, 2009. p. 205-221. Disponível em: https://www.academia.edu/63302916/Artificial_moral_agency_in_technoethics. Acesso em 9 out. 2023.

UHLMANN, E. L. *et al*. The motivated use of moral principles. *Judgment and Decision making*, v. 4, n. 6, p. 479-491, out. 2009. Disponível em: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/F6BEE6E1359B4FDFC9220863CEF09AC4/S1930297500004022a.pdf/the-motivated-use-of-moral-principles.pdf. Acesso em 9 out. 2023.

WAYTZ, A.; YOUNG, L. Morality for us versus them. *In*: GRAY, K.; GRAHAM, J. (ed.). *Atlas of moral psychology*. New York: The Guilford Press, 2018. p. 186-192. Disponível em: https://philpapers.org/archive/BISNAA-4.pdf. Acesso em 9 out. 2023.


11:40
IMPRENSA NACIONAL
1808
Diário Oficial da União
A informação oficial
ao alcance de todos
PRESIDÊNCIA DA REPÚBLICA

FISCALIZAÇÃO DO GOVERNO COM IA?

Lições de uma experiência brasileira [3]

Por **Henrique S. Xavier**

3 Artigo originalmente publicado em Xavier (2023a).

RESUMO

Este artigo analisa um projeto desenvolvido no Congresso brasileiro que, desde 2020, monitora e seleciona matérias relevantes do Diário Oficial da União usando Processamento de Linguagem Natural (*Natural Language Processing* – NLP) e Aprendizado de Máquina (*Machine Learning* – ML) como parte de um trabalho de fiscalização do Poder Executivo. As matérias selecionadas também são usadas para produzir um boletim diário aberto ao público em geral. São apresentados os pormenores técnicos e destacados os êxitos e as possíveis melhorias do projeto. Em termos gerais, o projeto é avaliado como um uso positivo da inteligência artificial (IA) na administração pública, com efeitos secundários adversos mínimos.

Palavras-chave: Ordenação de documentos. Supervisão governamental. Aprendizado de máquina aplicado. Processamento de linguagem natural. Estudo de caso.

I - INTRODUÇÃO

Como em muitos países (PELIZZO; STAPENHURST, 2011), uma das funções do Poder Legislativo no Brasil é fiscalizar o Poder Executivo (BRASIL, 1988). Os deputados federais e senadores brasileiros, por exemplo, devem monitorar – e, eventualmente, reagir a – todas as ações tomadas pelo gabinete presidencial, ministérios, agências governamentais federais e conselhos. Essas ações incluem: os ajustes orçamentários; a aprovação de leis, vetos, decretos, regulamentos, normas administrativas e ordens executivas; e a nomeação e demissão de vários cargos. Uma das principais formas de acompanhamento do Executivo é a leitura diária do Diário Oficial da União (DOU), em que todos esses atos devem ser publicados para, então, entrarem em vigor. Outros inquéritos e reações (denominados "instrumentos de fiscalização") incluem o requerimento de informações, as comissões parlamentares temporárias de inquérito, o depoimento obrigatório de funcionários públicos, os decretos legislativos e o *impeachment* (LEMOS; POWER, 2011).

Em 2022, o DOU publicou, em média, 280 atos normativos e 630 atos de pessoal por dia útil, totalizando 910 matérias diárias a serem examinadas[4]. Na prática, porém, muitas matérias são simplesmente burocráticas ou tratam de questões menores, enquanto apenas uma pequena parcela delas é de interesse dos representantes e do público em geral. Nesse cenário, um tipo de triagem de matérias pode ser bastante útil para reduzir o tempo de leitura do periódico, melhorar a eficiência e a agilidade, e tornar esse trabalho menos maçante.

Em 2020, os assessores de um grupo de congressistas brasileiros criaram uma ferramenta de filtragem desse tipo e lançaram-na na forma de um *software* de código aberto, o qual procurava novas matérias no *website* do DOU a cada 30 minutos e enviava-as para um filtro construído com técnicas de Processamento de Linguagem Natural (*Natural Language*

4 Valores calculados pelo autor a partir de Brasil (s.d.)

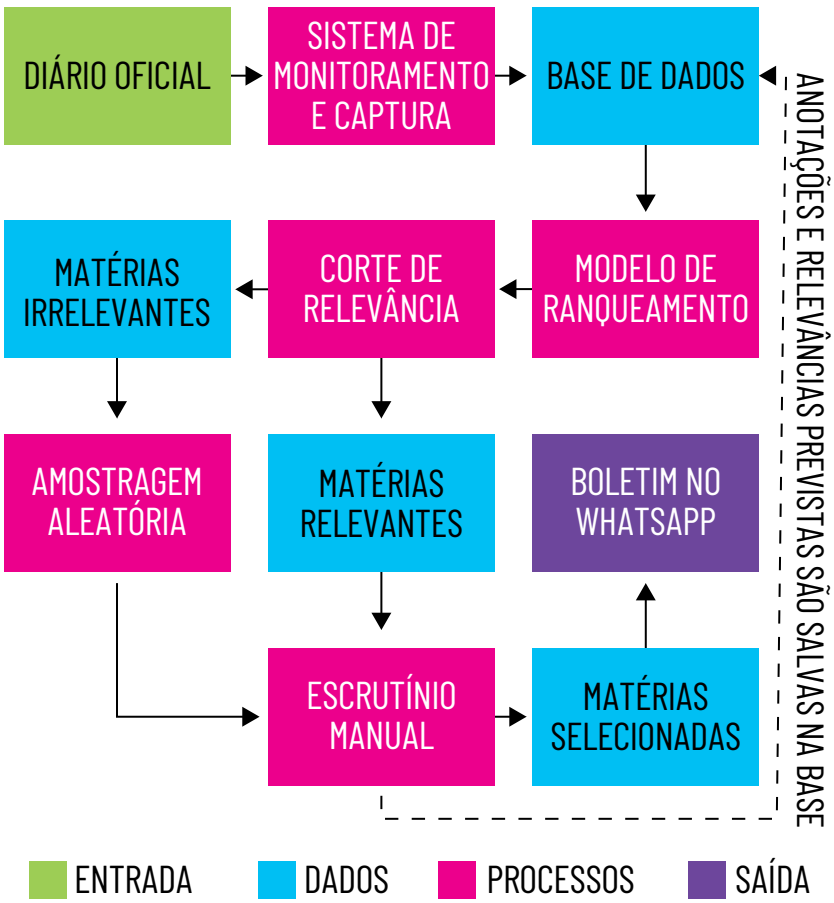
Processing – NLP) e Aprendizado de Máquina. Das 910 matérias diárias, apenas cerca de 40 eram selecionadas para análise humana. Além de ser utilizada pelo grupo de congressistas, a seleção final de matérias foi enviada gratuitamente a aproximadamente 1.300 subscritores – incluindo acadêmicos, consultores privados e funcionários do governo –, como um boletim em um aplicativo de mensagens instantâneas. O fluxograma da Figura 1 apresenta uma visão global do sistema. A ferramenta funcionou de março de 2020 até a data de finalização desta pesquisa (janeiro de 2023).

Na literatura sobre o governo eletrônico (e-GOV), a aplicação da tecnologia à supervisão "Governo a Governo" (G2G) é um pouco negligenciada. Por um lado, a supervisão governamental em um marco de "Governo para Cidadãos" (G2C) é geralmente promovida a partir de medidas de transparência (MARGETTS, 2006). Por outro lado, os estudos G2G preocupam-se, normalmente, com a integração interna, como o compartilhamento de bases de dados entre agências ou a melhoria da eficiência burocrática (LEE; TAN; TRIMI, 2005). Por conseguinte, a ferramenta mencionada oferece uma nova perspectiva para o âmbito do e-Gov, não apenas como um exemplo de supervisão G2G, mas, também, por meio da publicação do boletim para um público mais amplo, como uma combinação de G2G e G2C.

Esse documento avalia e descreve o funcionamento dessa ferramenta de filtragem, destacando-lhe requisitos, limitações, vantagens, oportunidades de melhoria e impactos. Considera-se que se trata de uma aplicação bem-sucedida da tecnologia e do aprendizado de máquina na administração pública que pode ser implementada em diferentes países e esferas de governo. Conquanto a classificação de textos com o aprendizado de máquina seja um campo de estudo ativo e muito utilizado em outros contextos, não se tem conhecimento de nenhuma aplicação publicada para a supervisão governamental. Estudos semelhantes incluem a extração de texto da seção de licitações

e contratos do DOU para detectar irregularidades (ROCHA, 2011), a procura de matérias no Diário que correspondam a consultas do usuário utilizando técnicas de índice invertido (SANTOS NETO, 2013) e a recuperação de documentos jurídicos, a partir de uma *query*, por meio de *embeddings* de documentos (SUGATHADASA , 2018). A principal contribuição deste trabalho é a demonstração dos benefícios e das práticas da aplicação da classificação de textos à supervisão do DOU.

FIGURA 1 - FLUXOGRAMA DO SISTEMA DE MONITORAMENTO E FILTRAGEM IMPLEMENTADO

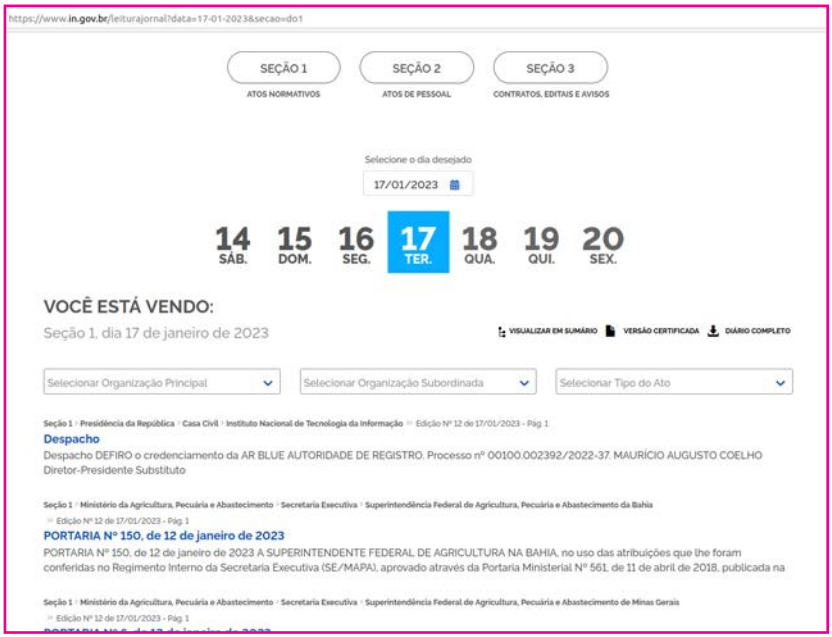


Fonte: Elaboração própria.

II - MONITORAMENTO E CAPTURA DE MATÉRIAS

Um requisito óbvio para a ferramenta de monitoramento é a presença *online* do Diário Oficial. Na melhor das hipóteses, o diário pode fornecer uma *Application Programming Interface* (API), como o DOU começou a fazer em janeiro de 2020 (INLABS, s.d.). Entretanto, no momento do desenvolvimento da ferramenta, o DOU tinha apenas um *website* que precisava ser raspado, como talvez seja para outros diários. Felizmente, complementando a URL do *website* com a *query* ?data=17-01-2023&secao=do1, por exemplo, acessa-se a lista de matérias (e suas *Uniform Resource Locator* – URL) publicadas em determinada data e seção (as quais são: 1 – atos normativos; 2 – atos de pessoal; 3 – licitações e contratos), como ilustra a Figura 2.

FIGURA 2 - PÁGINA DO DIÁRIO OFICIAL DO GOVERNO FEDERAL DO BRASIL



Fonte: BRASIL (2023b).

A Figura 3 apresenta o exemplo de uma matéria. Informações importantes, como o título (que inclui o tipo de matéria – por exemplo, decreto, lei, portaria etc.), a data de publicação, a seção e o órgão governamental associado são identificadas por *tags* de *HyperText Markup Language* (HTML), o que facilita a sua análise. Como as URL das matérias são únicas, podem ser utilizadas como seus identificadores.

As matérias do DOU são, geralmente, disponibilizadas *online*, todas juntas, aproximadamente, às 3 horas da madrugada, mas podem aparecer em lotes ao longo do dia. Assim, uma ferramenta de captura sem perdas pode exigir o monitoramento em tempo real do *website* do diário.

O sistema de monitoramento e captura de matérias foi escrito em Python e executado na Amazon Web Services (AWS) usando funções Lambda (AWS, 2023b): a cada 30 minutos, o sistema verificava a lista de matérias publicadas na data atual e baixava-as para o serviço de armazenamento S3 da Amazon (AWS, 2023a). Para evitar o *download* duplicado de matérias, o sistema mantinha uma lista das URL baixadas com sucesso naquele dia.

FIGURA 3 - EXEMPLO DE MATÉRIA DO DOU


Voltar

Compartilhe

VERSÃO CERTIFICADA | DIÁRIO COMPLETO | IMPRESSÃO

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 17/01/2023 | Edição: 12 | Seção: 1 | Página: 1

Órgão: Ministério da Agricultura, Pecuária e Abastecimento/Secretaria Executiva/Superintendência Federal de Agricultura, Pecuária e Abastecimento da Bahia

PORTARIA Nº 150, DE 12 DE JANEIRO DE 2023

A SUPERINTENDENTE FEDERAL DE AGRICULTURA NA BAHIA, no uso das atribuições que lhe foram conferidas no Regimento Interno da Secretaria Executiva (SE/MAPA), aprovado através da Portaria Ministerial Nº 561, de 11 de abril de 2018, publicada na Seção I do DOU de 13 de abril de 2018 e nos termos da Instrução Normativa nº 22 de 20 de junho de 2013,publicada no DOU de 21 de junho de 2013, resolve:

Art. 1º - Habilitar o Médico Veterinário RÔMULO MENEZES AMORIM inscrito(a) no CRMV-BA nº 07583-VP (BA), para emitir Guia de Trânsito Animal -GTA no Estado da Bahia, em conformidade com os autos do processo nº 21012.000263/2023-54;

Art. 2º - Esta Portaria entra em vigor na data da sua publicação.

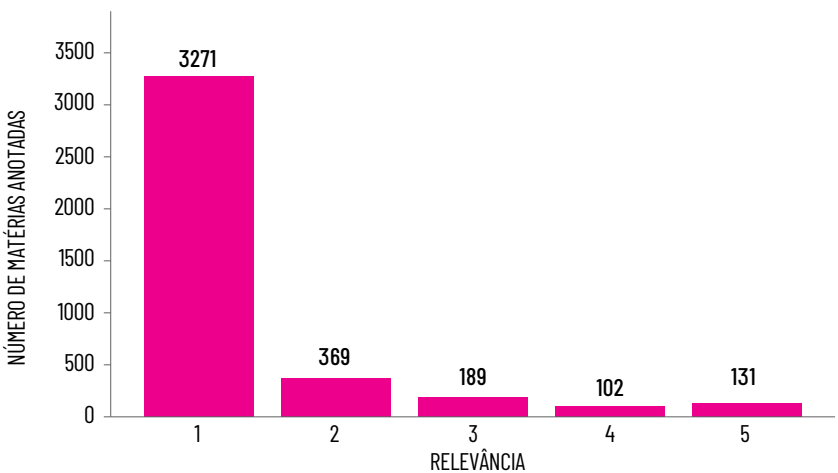
Fonte: BRASIL (2023a).

III - SELEÇÃO INICIAL E CLASSIFICAÇÃO

O processo de treinamento de modelos de aprendizado de máquina requer instâncias rotuladas. No que diz respeito à relevância, uma medida altamente subjetiva (LIN; NOGUEIRA; YATES, 2021), a classificação deve ser feita de acordo com o interesse e o conceito dos usuários finais do sistema. Logo, esse processo deve ser realizado pelos próprios usuários finais (no caso em análise, o grupo de congressistas) ou por pessoas que compreendam e apliquem seus pontos de vista, como os assessores.

Ao longo do primeiro mês do projeto, aproximadamente cinco assessores de congressistas se revezaram para ler e classificar as matérias conforme a relevância, de 1 (menos relevante) a 5 (mais relevante). Como se demonstra no Gráfico 1, a maioria dessas matérias iniciais foi classificada com relevância 1. Durante tal processo, eles também identificaram recortes sobre tipos de matérias e órgãos governamentais para remover aquelas consideradas irrelevantes e evitar que se perdesse demasiado tempo com elas. Esses recortes foram incorporados como um primeiro passo na ferramenta de filtragem.

GRÁFICO 1 - NÚMERO DE MATÉRIAS DE TODAS AS SEÇÕES COLETADAS E ANOTADAS DURANTE O PRIMEIRO MÊS DO PROJETO, POR RELEVÂNCIA INDICADA



Fonte: Elaboração própria.

O projeto de filtragem foi inicialmente elaborado como uma tarefa de classificação (uma escolha comum em trabalhos semelhantes, por exemplo (CAÇÃO, 2022); assim, as classificações de relevância foram limitadas a números inteiros. Percebeu-se, mais tarde, que o filtro funcionava melhor como uma tarefa de regressão seguida de um corte de relevância: as regressões fornecem naturalmente um esquema de ranqueamento para as instâncias, permitindo uma ordenação minuciosa; além disso, calculam a média das avaliações subjetivas feitas por vários analistas, levando a um tipo de acordo sobre a relevância da matéria (pelo menos entre os analistas). Por conseguinte, em futuras implementações, recomenda-se a utilização de uma pontuação contínua para a classificação de relevância em vez das discretas utilizadas[5].

Esse acordo mencionado ainda reflete um determinado ponto de vista, pois todos os analistas eram assessores de um determinado grupo de congressistas. Apesar disso, o filtro continua a ser útil para outras pessoas fora do grupo, visto que muitas delas têm interesses semelhantes e compartilham pontos de vista similares sobre a importância de certas ações governamentais, o que é especialmente verdadeiro para matérias localizadas nos extremos da escala de relevância.

IV - CONSTRUÇÃO E AVALIAÇÃO DO MODELO

A natureza e o conteúdo das matérias de seções distintas do DOU são consideravelmente diferentes entre si, o que motivou a criação de modelos diferentes para as seções 1 e 2 (a seção 3 não fez parte do projeto); entretanto, seus processos de construção foram os mesmos.

As matérias classificadas foram divididas em conjuntos de treinamento, validação e teste. Ao longo do projeto, foram adotadas duas diferentes estratégias de divisão. Em uma pri-

5 Para uma discussão aprofundada sobre o tema da análise de relevância, consultar Lin, Nogueira e Yates (2021).

meira fase, o conjunto de treinamento era constituído por todas as matérias publicadas antes de uma determinada data, enquanto o restante era dividido entre os conjuntos de validação e de teste utilizando uma função de *hashing* (GÉRON, 2019). A estratégia de divisão por datas teve como objetivo simular o ambiente de produção em que o modelo deveria avaliar matérias não provenientes da mesma população; desse modo, ele deveria focar em características mais gerais. Já a utilização de uma função de *hashing* para dividir o conjunto de dados garante que os dois conjuntos não se misturem quando são adicionadas novas instâncias rotuladas ao conjunto de dados, para fins de retreinamento do modelo. Posteriormente, no desenvolvimento do projeto, a estratégia mudou para que todos os conjuntos fossem divididos utilizando uma função de *hashing*. Não se sabe ao certo qual a estratégia que fornece melhores resultados ou se existe de fato alguma diferença.

Os modelos de aprendizado de máquina utilizados no projeto pertencem à família do *bag-of-words*, em que o texto da matéria é representado por contagens de *n*-gramas. Vários algoritmos de regressão foram testados e ajustados utilizando o conjunto de validação: os melhores foram os algoritmos *Ridge* para a seção 1 e uma combinação dos algoritmos de regressão *Ridge*, *Support Vector Machine* e *Random Forest* para a seção 2. Em ambos os casos, o *bag-of-words* binário (ou seja, a representação do texto pela presença ou ausência de *n*-gramas) funcionou melhor. Ao ajustar o modelo, atribuir pesos de amostra mais elevados às matérias mais relevantes (e, em geral, mais raros) também melhorou os resultados[6].

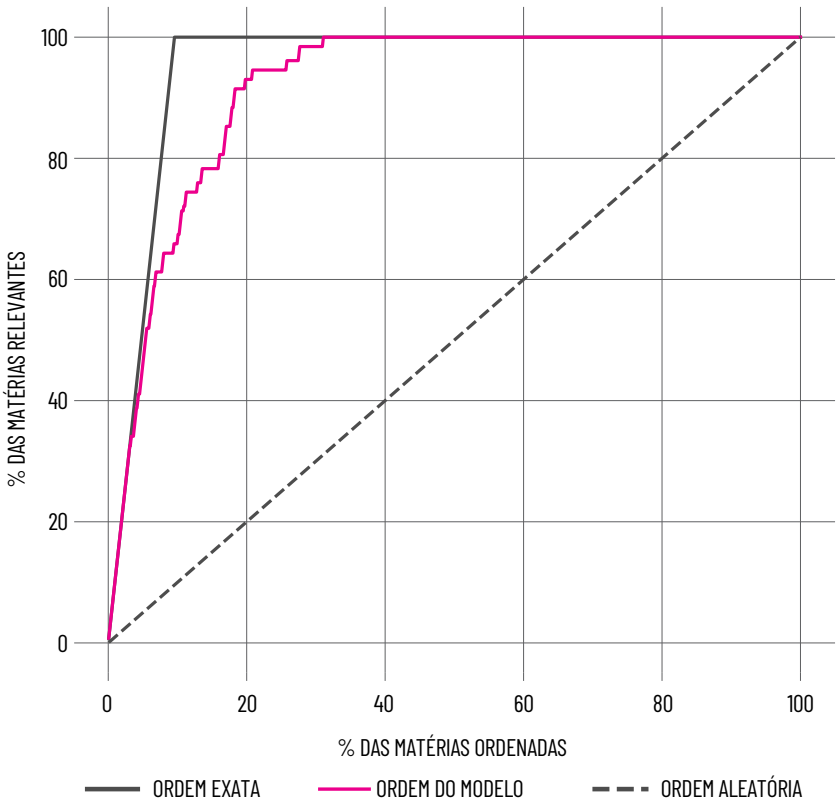
Uma vez que o objetivo do filtro era reduzir a quantidade de leitura necessária para identificar todas as matérias relevantes publicadas em determinado dia, um método de ava-

6 O processo completo de construção e avaliação do modelo para a seção 1 está disponível em Xavier (2023b).

liação útil foi estimar, usando o conjunto de teste, a fração de matérias classificadas como 4 e 5 (ou seja, consideradas "relevantes") que seriam lidas à medida que a tarefa de leitura avançasse pelas matérias ordenadas por relevância prevista decrescente: uma métrica conhecida como Recall@*k*, em que *k* é o número de matérias lidas. Essa avaliação é apresentada nos Gráficos 2 e 3 para as seções 1 e 2, respectivamente, as quais demonstram o funcionamento da modelagem para ambas as seções, dado que agrupa as matérias mais relevantes no topo da lista ordenada (ou seja, as matérias relevantes não estão distribuídas uniformemente, como aconteceria em uma ordenação aleatória). No entanto, a ordenação não é perfeita: as curvas coloridas não estão exatamente em cima das curvas pretas sólidas. Na métrica *Normalized Discounted Cumulative Gain* (nDCG), os modelos atingiram os valores 0,973 e 0,974 para as seções 1 e 2, respectivamente.

Verificou-se, também, que o modelo para a seção 2 funciona melhor do que para a seção 1, ou seja: depois de ordenar as matérias, é preciso ler uma quantidade menor delas para percorrer uma dada fração de matérias relevantes (por exemplo, para encontrar 80% das matérias relevantes, é preciso ler cerca de 4% da seção 2, contra 15% da seção 1). Essa diferença resulta de duas características: em primeiro lugar, as matérias consideradas relevantes são menos frequentes na seção 2 do que na seção 1 (representam 4% da seção 2 e 9% da seção 1); em segundo lugar, as razões pelas quais as matérias foram consideradas relevantes na seção 2 estão associadas à terminologia e à estrutura estáveis (os assessores estavam principalmente interessados em seguir as nomeações e demissões de funcionários de alto nível – ministros, presidentes, secretários, diretores – e essas ações seguem e utilizam normas e termos padrão).

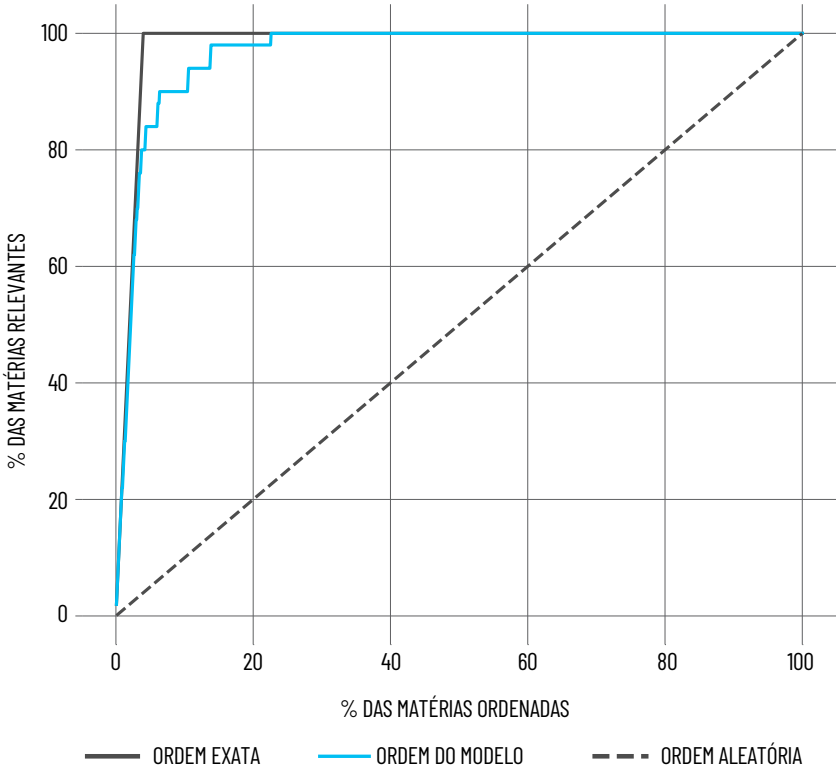
GRÁFICO 2 - FRAÇÃO DE MATÉRIAS DA SEÇÃO 1 COM RELEVÂNCIA 4 OU SUPERIOR QUE CONSTA NA FAIXA SUPERIOR DE MATÉRIAS DA SEÇÃO 1 CLASSIFICADAS SOB DIFERENTES ESQUEMAS DE ORDENAÇÃO DECRESCENTE, EM FUNÇÃO DO TAMANHO DA FAIXA. OS ESQUEMAS DE ORDENAÇÃO SÃO: ALEATÓRIO (LINHA PRETA TRACEJADA), SEGUINDO A RELEVÂNCIA PREVISTA PELO MODELO (LINHA VERMELHA SÓLIDA) E SEGUINDO A RELEVÂNCIA REAL (LINHA PRETA SÓLIDA)



Fonte: Elaboração própria.

Na seção 1, pelo contrário, a relevância derivou de um conjunto diversificado de razões, por exemplo: grandes alterações orçamentárias, revogação de regulamentos sobre armas, alterações fiscais, criação de reservas naturais, fechamento de hospitais, sanções de novas leis, restrições de vivagens devido à pandemia Covid-19 etc.

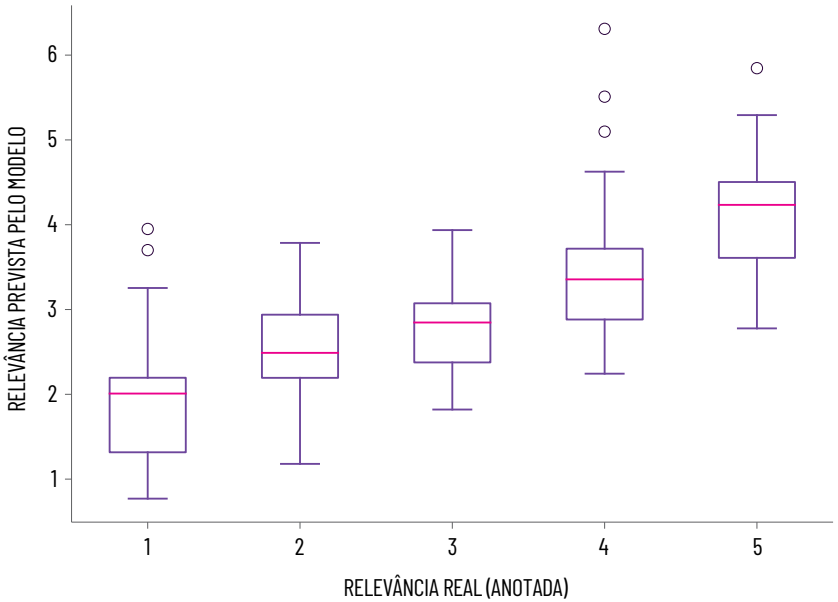
GRÁFICO 3 - SEMELHANTE AO GRÁFICO 2, MAS PARA A SEÇÃO 2. A CURVA ASSOCIADA À ORDENAÇÃO QUE SEGUE AS PREVISÕES DO MODELO É APRESENTADA EM AZUL



Fonte: Elaboração própria.

No Gráfico 4, as matérias da seção 1 com diferentes níveis de relevância estão distribuídas de formas diferentes em termos de relevância prevista. Aproximadamente 75% das matérias classificadas como nível 1 têm uma relevância prevista inferior a 2,2, enquanto todas as matérias do conjunto de teste nos níveis 4 e 5 situam-se acima desse limiar. Portanto, o valor de 2,2 foi definido como um ponto de corte para selecionar o subconjunto de matérias mais relevantes. Um resultado e uma estratégia semelhantes foram observados para a seção 2 do DOU.

GRÁFICO 4 - *BOXPLOT*[7] MOSTRANDO A DISTRIBUIÇÃO DA RELEVÂNCIA PREVISTA PARA AS MATÉRIAS DA SEÇÃO 1 NO CONJUNTO DE TESTE, CONSIDERANDO SUA RELEVÂNCIA REAL (OBSERVADA)



Fonte: Elaboração própria.

Apesar do sucesso do modelo de aprendizado de máquina, ainda há espaço para melhorias, especialmente na seção 1. Uma alternativa pode ser o uso de modelos de aprendizagem profunda pré-treinados, como o BERT (DEVLIN, 2022) e sua versão em português brasileiro, o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020a; 2020b). Esses modelos são facilmente implementáveis usando o Hugging Face (s.d.), podem ser pré-treinados ao contexto do DOU por meio do treinamento em tarefas autossupervisionadas de *Masked Language Modeling* (MLM) e *Next Sentence Prediction* (NSP), e ajustados

7 De acordo com a documentação do pacote Pandas (s.d.), de Python, "a caixa se estende do quartil Q1 ao Q3 dos dados, com uma linha interna na mediana (Q2). As linhas verticais (*whiskers*) se estendem além das bordas da caixa para mostrar o intervalo dos dados. Por padrão, eles não se estendem mais do que 1,5 * IQR (onde IQR = Q3 - Q1) das bordas da caixa, terminando no dado mais distante dentro desse intervalo. Outliers são representados como pontos separados".

para estimar a relevância usando as 5.577 e 9.247 matérias das seções 1 e 2 anotadas pelos assessores dos congressistas ao longo de dois anos. Os modelos BERT são recomendados para a classificação de textos (LIN; NOGUEIRA; YATES, 2021) e foram testados em tarefas semelhantes, superando os modelos mais simples de aprendizado de máquina, como o *Naive Bayes* (CAÇÃO, 2022).

V - ROTINA DE MONITORAMENTO

Os modelos foram armazenados e executados na mesma infraestrutura AWS mencionada na Seção II. Todos os dias úteis, por volta das 9 horas, quando todas as novas matérias estavam habitualmente publicadas no *website* do DOU e coletadas no S3, um assessor acessava um aplicativo *web* de fácil utilização e acionava os modelos para estimar a relevância das matérias coletadas. Cada matéria avaliada era transformada em uma versão estruturada (sendo a relevância prevista um campo e com outros: como o título, o órgão governamental, a seção, o resumo, o trecho, o texto na íntegra, a edição do Diário e a URL) e salva no S3. O sistema construía, então, uma tabela para cada seção do DOU contendo todas as matérias (um por linha), ordenadas por relevância prevista decrescente até o ponto de corte especificado.

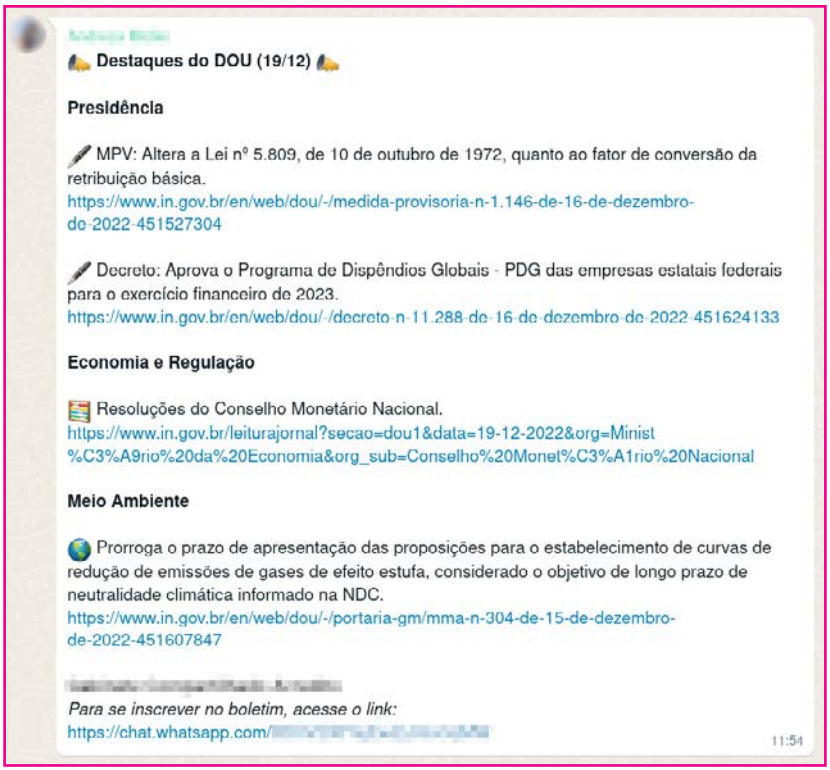
Ao final de cada tabela, o sistema anexava cinco matérias selecionadas aleatoriamente com relevância prevista inferior ao valor de corte, com o objetivo de monitorar a proporção em que matérias relevantes eram erroneamente eliminadas. Para diversificar essas matérias aleatórias e evitar a seleção repetida daquelas com conteúdo semelhante (e, muitas vezes, burocrático), foi selecionada na amostragem, no máximo, uma matéria por órgão governamental. Uma estratégia provavelmente mais eficaz para evitar os mesmos conteúdos seria agrupar as matérias com o algoritmo *k-Means* e amostrar uniformemente cada agrupamento.

Após esse processo, o sistema mostraria ambas as tabelas ao assessor, que leria as matérias. A tabela da seção 1 conteria tipicamente 12 ± 7 matérias filtradas, mais 5 aleatórias, enquanto a tabela da seção 2 conteria 17 ± 9 matérias filtradas, mais 5 aleatórias. Um assessor experiente poderia ler todas as matérias em cerca de uma hora ou menos. À medida que a leitura era efetuada, o assessor preencheria uma coluna com a relevância que atribuía a cada matéria. De vez em quando, essas novas anotações seriam utilizadas para reavaliar o desempenho do modelo e para voltar a treiná-lo.

Algumas vezes, uma edição extra do DOU era publicada depois de a edição normal ter sido lida pelo assessor. Nessas situações, o sistema enviava um aviso ao assessor responsável pelo acompanhamento do DOU, permitindo, assim, uma resposta rápida, se necessário, e adicionava essas matérias extras ao lote do dia seguinte. Essas edições extra são normalmente curtas, com um máximo de 5 matérias, de forma que a sua filtragem não é normalmente necessária.

Após a curadoria das matérias selecionadas pelo modelo, o assessor enviava, pelo aplicativo WhatsApp, um boletim – que era gratuito e aberto a todos – para cerca de 1.300 pessoas com os destaques do DOU do dia (a Figura 4 apresenta um exemplo da seção 1). A quantidade de subscritores do boletim mostra que o interesse na seleção de matérias feita pelo projeto não ficou restrito aos parlamentares que o propuseram, além de as informações produzidas pelo processo de filtragem terem o potencial de auxiliar milhares de profissionais sem custo adicional, tornando-o um serviço altamente escalável. Uma pequena pesquisa realizada com os subscritores indicou que eles incluem outros assessores de parlamentares, funcionários do governo, jornalistas, consultores privados, acadêmicos e profissionais de organizações sem fins lucrativos.

FIGURA 4 - EXEMPLO DE UMA MENSAGEM DO BOLETIM DO DOU ENVIADA POR WHATSAPP COM AS MATÉRIAS MAIS RELEVANTES DA SEÇÃO 1 DESSE DIA



Fonte: Elaboração própria.

VI - CONCLUSÃO

O projeto descrito neste artigo tem algumas características interessantes. Como mencionado na seção I, trata-se de uma aplicação de supervisão governamental G2G e G2C que surgiu a partir das necessidades dos congressistas brasileiros e de seus assessores, e não de uma decisão institucional vinda de órgãos superiores. Requer pouca infraestrutura, está em funcionamento há quase 3 anos e cerca de 1.300 pessoas beneficiam-se diretamente dele diariamente. Em seu período de funcionamento, foram registrados poucos casos de erros importantes. A diversidade dos assinantes do boletim aponta

que o projeto é útil para uma variedade de profissionais, apesar de a relevância ser um conceito subjetivo. Além disso, sua aplicabilidade e utilidade parecem ser mais universais do que apenas o contexto do Governo Federal brasileiro.

A partir da implementação prática do projeto, foram identificados vários acertos e erros técnicos que poderão ser úteis em projetos futuros. Destaca-se a utilização de uma escala contínua para medir a relevância e o enquadramento como uma tarefa de regressão como melhores caminhos para selecionar matérias relevantes. A geração de modelos independentes para diferentes tipos de matérias (atos normativos *versus* atos de pessoal, por exemplo) e a utilização de pesos amostrais proporcionais à relevância da matéria também se apresentam como boas práticas. Além disso, é possível que a utilização de modelos BERT, em vez de modelos *bag-of-words*, possa melhorar a precisão da classificação das matérias (LIN; NOGUEIRA; YATES, 2021; CAÇÃO, 2022), especialmente para um conjunto de matérias mais diversificado e menos padronizado. Todas essas recomendações visam apenas melhorar o desempenho do filtro, são independentes umas das outras e não devem ter impacto em outros aspectos do projeto.

Do ponto de vista humano, a experiência demonstrou que a anotação de algumas matérias pode ser melhor conduzida caso especialistas efetuem diretamente as avaliações ou aconselhem os anotadores. A área de conhecimento abrangida pela seção 1 do DOU é vasta e inclui decisões técnicas econômicas, ações de proteção do ambiente, funcionamento interno de vários órgãos governamentais etc. Julgar, com precisão, a relevância desses atos pode ser difícil e impreciso sem um conhecimento prévio do assunto. Ao rotular determinada categoria de matérias, o modelo também pode se beneficiar da diversidade de anotadores, pois é capaz de levar a uma relevância média, mais representativa da interpretação de um grupo maior de pessoas.

Em tarefas subsequentes, como a publicação das matérias selecionadas, a utilização de redes sociais públicas, tais como o Twitter, pode ser um caminho para aumentar o acesso dos cidadãos à informação produzida. As seleções mais padronizadas – como nomeações e demissões de funcionários públicos importantes – podem ser totalmente automatizadas, não só diminuindo tempo de leitura, mas também tornando o fluxo de informação extremamente rápido.

Em sua fase atual, o projeto de filtragem do DOU não parece ser afetado por problemas comumente relacionados com a utilização da IA na administração pública: não há direitos básicos diretamente afetados por ele nem políticas públicas diretamente definidas com base nele. Ademais, ele não é uma ferramenta especificamente desenvolvida para ajudar na gestão e nas decisões políticas. Por outro lado, caso certo modelo de filtragem ganhe escala ao ponto de se tornarem escassas as avaliações independentes do Diário Oficial, mesmo que também automatizadas, seus inevitáveis enviesamentos podem desviar a atenção da sociedade de atos relevantes, porém não detectados pela ferramenta. Assim, ainda que o método de filtragem do DOU por meio de IA possa se tornar predominante, uma implementação específica não deveria ser amplamente adotada.

REFERÊNCIAS

AMAZON WEB SERVICES (AWS). *Amazon S3*. Seattle: AWS, 2023a. Disponível em: https://aws.amazon.com/s3. Acesso em 5 nov. 2023.

AMAZON WEB SERVICES (AWS). *AWS Lambda*. Seattle: AWS, 2023b. Disponível em: https://aws.amazon.com/lambda. Acesso em 5 nov. 2023.

BERTimbau Base (aka "bert-base-portuguese-cased"). *Hugging Face*, s.d. Disponível em: https://huggingface.co/neuralmind/bert-base-portuguese-cased. Acesso em 5 nov. 2023.

BRASIL. *Leitura do Jornal*. Brasília: IN, s.d. Disponível em: https://www.in.gov.br/leiturajornal. Acesso em 20 nov. 2023.

BRASIL. *Constituição da República Federativa do Brasil de 1988*. Brasília: Presidência da República, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em 8 out. 2023.

BRASIL. *Portaria n. 150, de 12 de janeiro de 2023*. Brasília: DOU, 17 jan. 2023a. Disponível em: https://www.in.gov.br/web/dou/-/portaria-n-150-de-12-de-janeiro-de-2023-458160624. Acesso em 5 nov. 2023.

BRASIL. *Seção 1, dia 17 de janeiro de 2023*. Brasília: IN, 17 jan. 2023b. Disponível em: https://www.in.gov.br/leiturajornal?data=17-01-2023&secao=do1. Acesso em 5 nov. 2023.

CAÇÃO, F. N. *et al*. Tracking environmental policy changes in the Brazilian federal official gazette, *In*: Computational Processing of the Portuguese Language, 15, 2022, Fortaleza. *Anais* [...]. 21 mar. 2022. p. 256-266. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-98305-5_24. Acesso em 8 out. 2023.

DEVLIN, J. *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Conference of the North American Chapter of the Association Computational Linguistics: Human Language Technologies, 17, jun. 2019. *Anais* [...]. 26 set. 2022. v. 1, p. 4171-4186. Disponível em: https://browse.arxiv.org/pdf/1810.04805.pdf. Acesso em 8 out. 2023.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow*: Concepts, Tools, and Techniques to Build Intelligent Systems. 2. ed. Sebastopol: O'Reilly, 2019.

IMPRENSA NACIONAL (INLABS). O objetivo do INLABS é permitir o acesso aos arquivos das edições completas do Diário Oficial da União, em formato PDF e XML, que é livre e gratuito desde o dia 1 de janeiro de 2020. *Github*, 15 set. 2021. Disponível em: https://github.com/Imprensa-Nacional/inlabs. Acesso em 5 nov. 2023.

LEE, S. M.; TAN, X.; TRIMI, S. Current practices of leading e-government countries. *Communications of the ACM*, v. 48, n. 10, p. 99-104, out. 2005. Disponível em: https://dl.acm.org/doi/abs/10.1145/1089107.1089112. Acesso em 8 out. 2023.

LEMOS, L. B.; POWER, T. *Determinants of oversight in a reactive legislature*: The case of Brazil (1988-2005). GEG Working Paper No. 2011/62. Oxford: University of Oxford; GEG, 2011. Disponível em: https://www.econstor.eu/handle/10419/196322. Acesso em 8 out. 2023.

LIN, J.; NOGUEIRA, R.; YATES, A. Pretrained transformers for text ranking: BERT and beyond. *In*: Annual Conference of the North American Chapter of the Association Computational Linguistics: Human Language Technologies: Tutorials, 2021, 6-11 jun. 2021. *Anais* [...], Stroudsburg: ACL, 2021. p. 1-4. Disponível em: https://aclanthology.org/2021.naacl-tutorials.1.pdf. Acesso em 8 out. 2023.

MAKE a box plot from DataFrame columns. *Pandas*, s.d. Disponível em: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html. Acesso em 5 nov. 2023.

MARGETTS, H. Transparency and Digital Government. *In*: Hood, C.; Heald, D. (ed.). *Transparency*: The Key to Better Governance? Londres: British Academy, set. 2006. p. 196-207. Disponível em: https://academic.oup.com/british-academy-scholarship-online/book/13315. Acesso em 8 out. 2023.

PELIZZO, R.; STAPENHURST, F. *Parliamentary Oversight Tools*: a Comparative Analysis, Londres: Routledge, 2011.

PIERRI, G.; LAFUENTE, M. *Digital Government and Corruption: The Impact of Citizen Oversight and Infobras on the Efficiency of the Execution of Public Works in Peru*. Washington: Inter-American Development Bank, nov. 2020. Disponível em: https://publications.iadb.org/en/digital-government-and-corruption-impact-citizen-oversight-and-infobras-efficiency-execution-public. Acesso em 8 out. 2023.

ROCHA, J. P. L. *Inteligência de fontes abertas*: um estudo de caso sobre descoberta de conhecimento no Diário Oficial da União. 2011. Dissertação (Mestrado em Informática) – Universidade Católica de Brasília, Brasília, 2011. Disponível em: https://bdtd.ucb.br:8443/jspui/handle/123456789/1336. Acesso em 8 out. 2023.

SANTOS NETO, F. C. *Monitoramento de publicações no Diário Oficial da União*. 2013. Trabalho de conclusão de curso (Bacharelado em sistemas de Informação) – Universidade Federal da Paraíba, João Pessoa, 2013. Disponível em: https://repositorio.ufpb.br/jspui/bitstream/123456789/17028/1/FCSN04042013.pdf. Acesso em 8 out. 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pre-trained BERT models for Brazilian Portuguese. *In*: Cerri, R.; Prati, R. C. (ed.). *Intelligent Systems*. New York: Springer International Publishing, 2020a. p. 403-417. Disponível em: https://www.researchgate.net/publication/345395208_BERTimbau_Pretrained_BERT_Models_for_Brazilian_Portuguese. Acesso em 8 out. 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Portuguese named entity recognition using BERT-CRF. *arXiv preprint*, arXiv:1909.10649, v. 2, 27 fev. 2020b. Disponível em: https://browse.arxiv.org/pdf/1909.10649.pdf . Acesso em 8 out. 2023.

SUGATHADASA K. *et al*. Legal document retrieval using document vector embeddings and deep learning. *In*: Computing Conference 2018, Londres, 10-12 jul. 2018. *Anais* […]. Londres: SAI, 2018. Disponível em: https://browse.arxiv.org/pdf/1805.10685.pdf. Acesso em 8 out. 2023.

XAVIER, H. S. Overseeing Government with AI : Lessons learned from a Brazilian experience. *In*: Iberian Conference on Information Systems and Technologies (CISTI), 18, Aveiro, jun. 2023. *Anais* […], jun. 2023a. p. 1-6. Disponível em: https://ieeexplore.ieee.org/document/10211905. Acesso em 5 nov. 2023.

XAVIER, H. S. Tutorial on how to use machine learning to rank notices from the government gazette according to their relevance. *Github*, 7 abr. 2023b. Disponível em: https://github.com/cewebbr/text_ranking_in_gov. Acesso em 5 nov. 2023.

UM LIVRO DIDÁTICO DIGITAL ACESSÍVEL E CONECTADO

Por **Reinaldo Ferraz** e **Ana Eliza Duarte**

RESUMO

O formato eletrônico de publicação (Electronic Publication – ePub) possibilita o desenvolvimento de publicações digitais utilizando recursos da plataforma aberta da Web. São recursos desenvolvidos por entidades como o International Digital Publishing Forum (IDPF) e o World Wide Web Consortium (W3C) para a evolução de documentos em formatos abertos, além de eliminar barreiras de acesso a esse tipo de formato para pessoas com deficiência. Embora seu uso esteja aumentando, principalmente em livros de ficção, ainda existem barreiras para a implementação de livros didáticos em formato digital que contemplem os mesmos exercícios da versão impressa. O objetivo deste estudo é explorar o potencial da interatividade do formato ePub para uso como material didático, a fim de que alunos respondam exercícios nos livros digitais. Foi desenvolvido um arquivo em formato ePub que permite que os leitores respondam as perguntas por meio de formulários e obtenham as respostas às suas perguntas de forma acessível. Também foi implementada uma forma de enviar os dados dos exercícios respondidos para uma plataforma que permite o acesso dos professores às respostas, bem como a possibilidade de exportação de dados para outras plataformas.

Palavras-chave: ePub. Livro digital. Acessibilidade.

I - INTRODUÇÃO

No ano de 2007, o International Digital Publishing Forum (IDPF) lançou o padrão aberto para livros eletrônicos: *Electronic Publication* (ePub) (KASDORF, 2013). Produzido inicialmente em formato *eXtensible HyperText Markup Language* (XHTML), um padrão de linguagem de marcação robusta que oferece diversos benefícios em relação ao controle de design e diagramação, por meio de recursos como as Folhas de Estilo em Cascata (*Cascading Style Sheets* – CSS).

A utilização de um formato como o ePub contempla a acessibilidade na publicação. Por ser desenvolvido com tecnologia Web (LEPORINI; MINARDI; PELLEGRINO, 2019), existem diretrizes internacionais para esse tipo de conteúdo, como o "*Web Content Accessibility Guidelines*" (WCAG) (GARRISH, 2012), e orientações específicas para acessibilidade nesse formato (KIRKPATRICK, 2023). O suporte a imagens vetoriais, como *Scalable Vector Graphics* (SVG), também foi um grande avanço, pois imagens vetoriais são mais leves e permitem a ampliação de figuras pelo usuário sem a perda de qualidade. Isso deixa o arquivo final com tamanho menor (em *megabytes*) e permite customizações do usuário.

Considerando usuários com deficiência visual, como pessoas cegas ou de baixa visão, o formato ePub traz o benefício do aumento de fontes e a possibilidade de uso de tecnologia assistiva, como leitores de tela, disponíveis em *smartphones* e computadores. Fazer uso desse formato também beneficia pessoas com limitações motoras, já que os campos de formulários estão disponíveis por toques simples e/ou atalhos de movimentos do *smartphone*.

Atualmente, os dispositivos de leitura são os mais diversos, desde *hardware* específico para leitura (como os *eReaders* da Kobo) até dispositivos multifuncionais (como computadores e *smartphones*). A partir do acesso à Internet no dispositivo, o leitor tem uma série de possibilidades de acesso ao conteúdo.

Estima-se que, no Brasil, em 2021, o número de domicílios com acesso à Internet era cerca de 59 milhões (atingindo aproximadamente 82% dos domicílios brasileiros). Desse total, 99% dos acessos foram realizados por meio de *smartphones* (NIC.BR, 2022). Essas porcentagens evidenciam que um número expressivo de cidadãos brasileiros tem acesso a um telefone com tecnologia suficiente para a execução de aplicativos leitores de livros eletrônicos.

Esse cenário chama atenção para as possibilidades ainda não exploradas que o formato ePub viabiliza, como utilizar recursos educacionais de forma acessível, visto que livros de papel são uma barreira, por exemplo, para pessoas com deficiência que dependem de digitalização ou alteração do conteúdo para seu uso (HARPUR, 2016).

A pandemia Covid-19 evidenciou uma lacuna entre a educação padrão e a tecnologia, já que muitas tarefas presenciais passaram a depender de recursos digitais para mediar o processo de aprendizagem remota (SANTOS JUNIOR; MONTEIRO, 2020). Nesse sentido, por ser um padrão aberto e executável em diversos dispositivos, o formato ePub se apresenta como um potencial recurso para a modernização dos instrumentos pedagógicos utilizados na atualidade.

II - PROCEDIMENTOS METODOLÓGICOS

Para o desenvolvimento e a exploração do padrão ePub, foram estabelecidos alguns critérios para a escolha da obra que serviria de experimento.

O primeiro critério foi que a licença do título deveria ser aberta, para permitir a investigação das possibilidades disponíveis pelo padrão ePub, sem nenhum tipo de impedimento relacionado a direitos autorais.

O outro critério estabelecido para o desenvolvimento da pesquisa foi a utilização do formato ePub, que segue padrões abertos desenvolvidos pelo World Wide Web Consortium (W3C),

cujo uso tem uma série de benefícios, como a estabilidade e a interoperabilidade (W3C, 2012). Esses padrões também são exigidos pelo Plano Nacional do Livro Didático (PNLD) brasileiro para livros digitais (BRASIL, 2021).

A obra *Frações no Ensino Fundamental – Volume 1* (RIPOLL *et al.*, 2021), desenvolvida pelo projeto Um Livro Aberto (2023), foi selecionada para esse experimento. O livro foi criado por professores de Matemática dos ensinos Básico e Superior, cujo objetivo foi a produção de materiais didáticos com licença aberta. A proposta era transformar o livro, que estava em formato *Portable Document Format* (PDF) para um livro em formato ePub, com exercícios que pudessem ser respondidos pelos alunos e o professor pudesse ter acesso às respostas. Também foi considerado o desenvolvimento de um leitor *eReader* que contemplasse os recursos previstos no documento ePub.

III - DESENVOLVIMENTO DO LIVRO EM EPUB

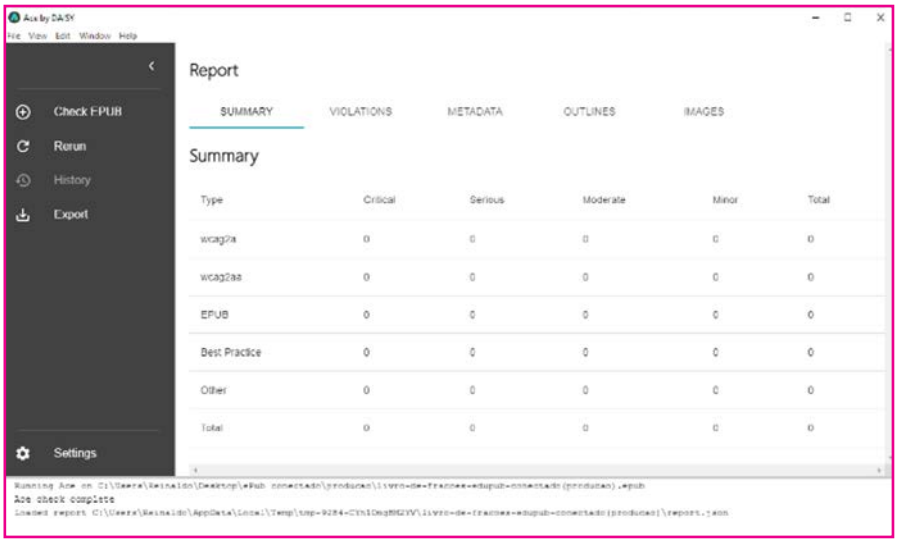
O projeto começou com o desenvolvimento do arquivo ePub. A base do livro selecionado está originalmente em formato PDF e LaTex: esse último permitiu transformar o conteúdo do livro em ePub com mais facilidade. O acesso às imagens em *Scalable Vector Graphics* (SVG) a partir do repositório do livro no Github (2021) permitiu a importação das imagens para o livro.

Foi mantida a estrutura de capítulos do livro para a construção da semântica do *HyperText Markup Language* (HTML), a fim de organizar o conteúdo no ePub, além de terem sido seguidas as boas práticas da especificação técnica do ePub 3 (GARRISH; HERMAN, 2023).

O desenvolvimento do arquivo ePub também seguiu as boas práticas de acessibilidade baseadas nas especificações das Diretrizes de Acessibilidade para o Conteúdo da Web (Web Content Accessibility Guidelines – WCAG) 2.1 (KIRKPATRICK, 2023) e ePub Accessibility 1.1 (GARRISH, 2023). Para a verificação de conformidade com os padrões técnicos, foi utilizado

o verificador da ferramenta Sigil (GITHUB, 2023) e o Ace by DAISY (s.d.). A Figura 1 apresenta o resultado da verificação, sem nenhum erro de conformidade com padrões de acessibilidade. Esse resultado ocorreu devido à conformidade com os padrões técnicos durante o desenvolvimento. Erros como imagens sem descrição ou campos de formulário sem rótulo são detectados por esse tipo de ferramenta.

FIGURA 1 - CAPTURA DE TELA DO RESULTADO DE VERIFICAÇÃO DE CONFORMIDADE DO EPUB COM A FERRAMENTA ACE BY DAISY.

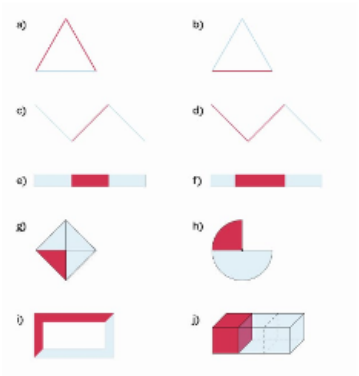


Fonte: Elaboração própria.

Além da verificação automática, foi feita a verificação manual de cada um dos capítulos do livro, principalmente com relação a estrutura de cabeçalhos, rótulos em elementos interativos e descrição de imagens, o que possibilita a alunos com deficiência visual compreenderem a imagem por meio do texto alternativo[8]. A Figura 2 mostra como foram descritas as imagens dos exercícios.

8 Descrições de imagens devem ser feitas com o acompanhamento de profissionais de ensino, para garantir que sigam adequadamente o material didático.

FIGURA 2 - REPRESENTAÇÃO DO TEXTO ALTERNATIVO DE UMA IMAGEM EM SVG CONTENDO GRÁFICOS DE EXERCÍCIOS DO LIVRO.



```
<img class="tasks"
src="../Images/licao01_
ativ6_vetor01.svg"
alt="
Letra A - triângulo com
dois lados pintados.
Letra B - triângulo com
um lado pintado.
Letra C - linha com três
segmentos de reta sendo
um deles pintado. ..."/>
```

Fonte: Elaboração própria.

Para possibilitar a interatividade da plataforma, foram adicionados campos de formulário em algumas questões. Todos estão com seus rótulos relacionados a seus devidos campos e utilizam um simples script que envia os dados para um servidor de sistema Learning Management System (LMS), através da tecnologia Experience API (xAPI (XAPI.COM, 2023),

O arquivo ePub foi testado com os recursos de interatividade em diversos leitores de *eReader*. Os testes consistiram na abertura do livro produzido em ePub para verificar se o aplicativo permite o acesso aos campos de resposta e o envio de dados das respostas. Em alguns casos, certos *eReaders* não permitem o acesso a campos de entrada de dados; em outros, não permitem a conexão de dados na plataforma. Identificamos a compatibilidade com Lithium Reader para Android, ePub Reader para iOS e Thorium Reader para computadores.

Os três *eReaders* mencionados são suportados por grande número de dispositivos. O Lithium Reader, por exemplo, é suportado a partir do sistema operacional Android 4.1 (até a conclusão deste artigo, a versão mais atual do Android era 13).

Não foi desenvolvido um leitor de *eReader* específico para o projeto, pois o objetivo foi permitir que o usuário use os já disponíveis nas lojas de aplicativos, com suporte aos recursos do projeto. Os três aplicativos que tem suporte aos recursos do projeto são gratuitos para uso do usuário.

Uma das limitações encontradas foi a impossibilidade de armazenamento de dados no dispositivo do usuário. Por exemplo, os *eReaders* testados não permitem essa funcionalidade; assim, as respostas preenchidas no livro dependem do acesso à Internet para o envio de dados na plataforma.

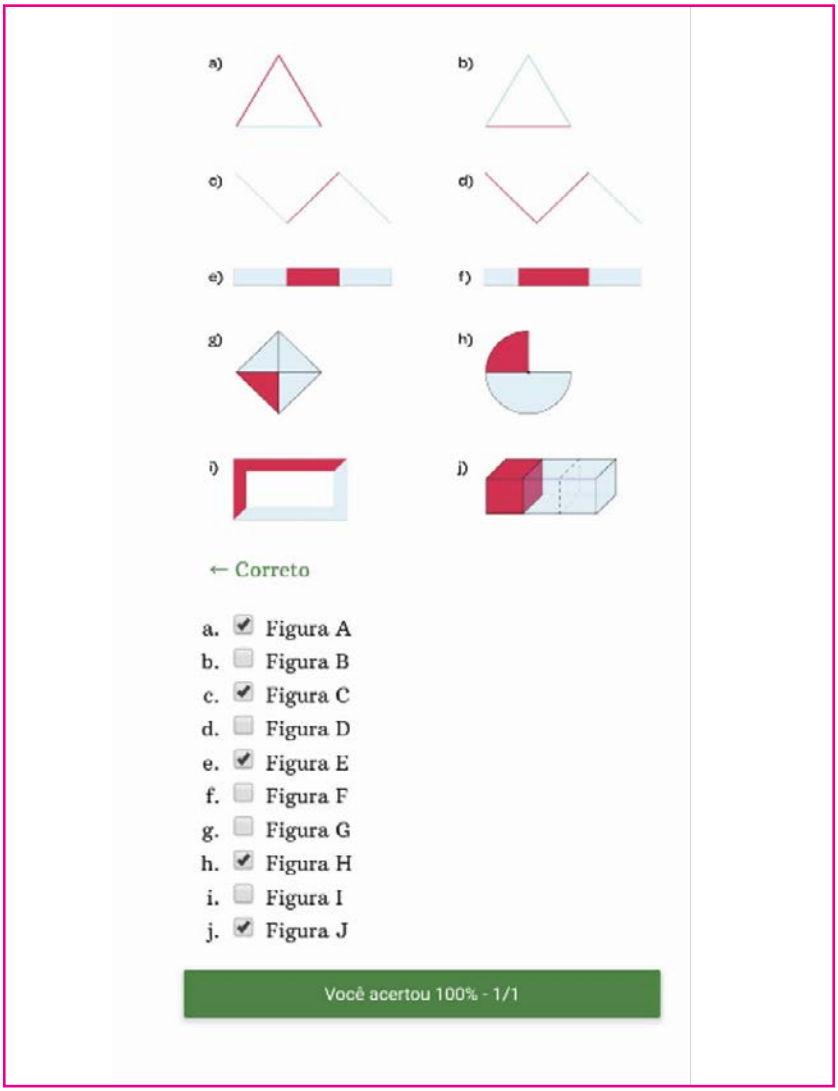
A partir dessa etapa de desenvolvimento, o ePub já permite a leitura e a interação com os exercícios, como mostrado na Figura 3. Um aluno pode utilizar o livro para fazer os exercícios e, em retorno, recebe na hora a informação se sua resposta está certa ou errada. Para que os dados da resposta do aluno sejam armazenados, é necessário conectá-lo à plataforma por meio da xAPI.

Dentre as diversas plataformas pesquisadas, foi selecionado o *Learning Locker* (LRS, 2023), que permite uma série de recursos compativeis com esse projeto. Sua instalação é simples e está muito bem documentada em seu *site*.

Depois da instalação, é necessário configurar a plataforma, criando as credenciais do livro que vai se comunicar com o *Learning Locker*, de modo muito simples. Após adicionar um novo *client*, os dados já estão disponíveis para uso, como mostrado na Figura 4. Essas informações são adicionadas ao arquivo "credenciais.js", no arquivo ePub, o qual permite configurar diversas informações, como dados do aluno e do professor.

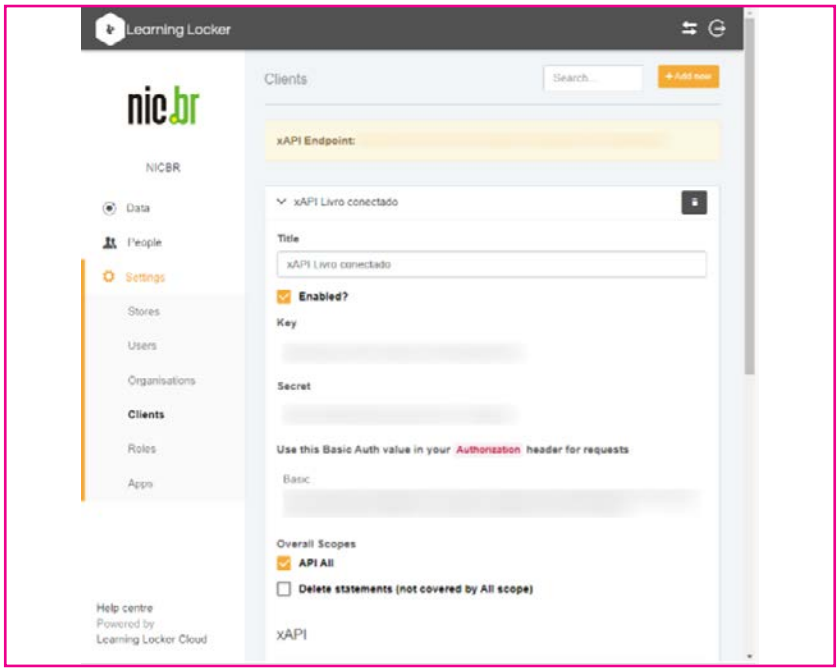
Ao término dessa configuração e da geração de um arquivo em formato ePub, é possível fazer uso do livro e enviar dados respondidos no arquivo ePub para a plataforma *Learning Locker* (LRS, 2023).

FIGURA 3 - EXEMPLO DE EXERCÍCIO RESPONDIDO COM TODAS AS RESPOSTAS CORRETAS.



Fonte: Elaboração própria.

FIGURA 4 - CAPTURA DE TELA DA INTERFACE DO *LEARNING LOCKER* PARA A CRIAÇÃO DE CREDENCIAIS.



Fonte: Elaboração própria.

IV - RESULTADOS

Nesta sessão, apresentamos os resultados da implementação dos recursos de conectividade em um livro digital.

O livro digital em formato ePub teve sua acessibilidade analisada por ferramentas automáticas, verificação manual de código com base nas orientações internacionais de acessibilidade (como demonstrado no Capítulo III) e verificação manual com tecnologia assistiva (leitores de tela). A navegação por leitores de tela, ferramenta utilizada por pessoas cegas ou com baixa

visão em computadores e dispositivos móveis, não encontrou barreiras de acesso na navegação.

A partir da conexão entre o livro e a plataforma, os exercícios respondidos pelo aluno têm seus dados exibidos, também diretamente na plataforma.

Todas as atividades do aluno geram um registro no banco de dados da aplicação que, por conseguinte, gera dados estruturados em formato *JavaScript Object Notation* (*JSON*) (ECMA INTERNATIONAL, 2017), como demonstrado na Figura 5.

FIGURA 5 - EXEMPLO DE CABEÇALHO DE DADOS EXIBIDO PELA PLATAFORMA.

Fonte: Elaboração própria.

A plataforma armazena uma série de dados, inclusive cada uma das respostas, mesmo se o aluno tenha respondido uma vez e depois alterado sua resposta (Figura 6).

FIGURA 6 - EXEMPLO DE CABEÇALHO DE DADOS EXIBIDO PELA PLATAFORMA

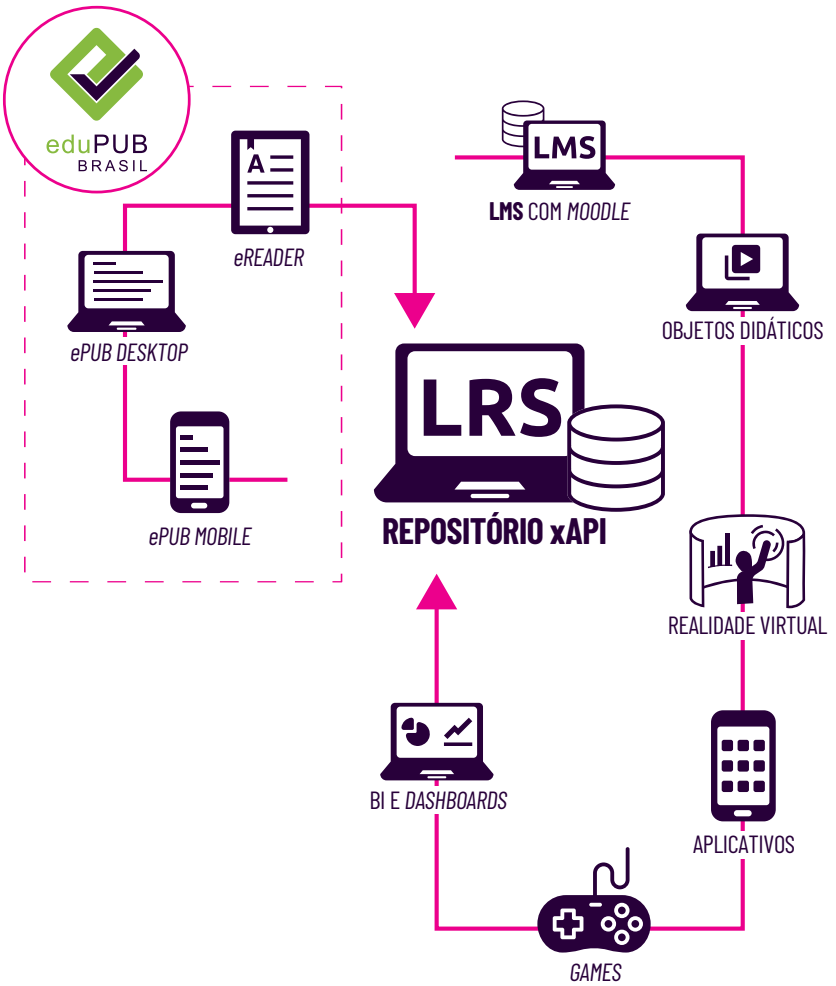

```
"actor": {
  "name": "Reinaldo Teste",
  "mbox": "mailto:aluno01@email.com",
  "objectType": "Agent"
},
"timestamp": "2022-03-15T19:33:06.115Z",
"version": "1.0.0",
"id": "8bad23c1-b7f8-47de-bc93-ea971efad0e4",
"result": {
  "success": true,
  "response": "1/3"
},
```


Fonte: Elaboração própria.

É possível utilizar os dados exibidos em *JSON* das mais diversas formas. A plataforma gera gráficos com base nas respostas dos exercícios do livro, os quais podem ser customizados. Os dados podem ser exportados para uma plataforma de ensino, jogos,ou qualquer aplicação com suporte ao formato *JSON*. No final do livro, há uma sessão que exibe um *dashboard* de respostas do aluno do próprio livro em formato ePub.

Algumas ferramentas, como o Moodle, por exemplo, permitem gerar conteúdos para aprendizados de sua plataforma. O objetivo do Livro Digital Conectado é permitir que a publicação em formato ePub seja interoperável e possa exportar dados para outras aplicações, como para o *Moodle* ou outras plataformas que suportam formatos de dados abertos.

FIGURA 7 - IMAGEM DO ECOSISTEMA DA APLICAÇÃO QUE PERMITE A EXPORTAÇÃO DE DADOS PARA DIVERSAS APLICAÇÕES.



Fonte: Elaboração própria.

O livro em formato ePub e as instruções para instalação e configuração do *Learning Locker* estão disponíveis no *website* do projeto (CEWEB.BR, s.d.).

V - CONCLUSÃO

A partir do experimento descrito, foi possível observar que o formato ePub para livros didáticos pode ser robusto e permite, de forma acessível, a interatividade entre o aluno e as obras. Essa interatividade pode ser compartilhada com o professor, por meio de uma plataforma que acessa os dados das respostas e dá ao docente mais flexibilidade para a correção de exercícios.

O formato ePub interativo também permite que recursos de acessibilidade sejam implementados (como a descrição de imagens), para garantir que uma pessoa com deficiência consiga responder as perguntas com os mesmos privilégios de uma pessoa sem deficiência.

A possibilidade de exportação de dados em formato aberto permite o uso de tecnologias acessíveis para visualização e interação entre o aluno e professor.

TRABALHOS FUTUROS

A aplicação ainda tem algumas limitações. Somente alguns leitores de ePub suportam a interatividade e a conexão com a plataforma (Lithium Reader para Android, ePub Reader para IOs e Thorium Reader para computadores). Desse modo, a pesquisa com outros leitores será ampliada para se estudar a necessidade do desenvolvimento de um leitor de ePub mais robusto, que permita explorações mais ousadas, como novos formatos de exercícios, interatividade e multimídia.

Devido às limitações dos leitores, a aplicação ainda depende de conexão com Internet para o envio das respostas. O objetivo é que a próxima versão do livro digital conectado faça uso de *IndexedDB API* e *Web Storage API*, para que o envio das respostas não dependa desse formato de conexão, já que essas tecnologias permitem o armazenamento das respostas no dispositivo do usuário e o envio para o servidor quando o dispositivo se conectar à Internet.

Esse foi um simples experimento com *scripts* que verificam se as respostas estão certas. No futuro, pretendemos relacionar as respostas diretamente com a plataforma e não no próprio livro, a fim de que professores possam modificar perguntas na plataforma que serão exibidas nos livros dos usuários.

Também será ampliado o conjunto de metadados no livro didático, com vocabulário que explore seu uso no ambiente acadêmico. Ademais, será desenvolvida uma nova obra conectada explorando outros aspectos do formato ePub com outros tipos de livro, como os de ficção.

REFERÊNCIAS

ACE by DAISY. *The DAISY Consortium*, s.d. Disponível em: https://daisy.org/activities/software/ace. Acesso em 5 nov. 2023.

BRASIL. *Edital PNLD 2023*. Brasília: Fundo Nacional de Desenvolvimento da Educação, 16 março 2021. Disponível em: https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/programas/programas-do-livro/consultas-editais/editais/edital-pnld-2023-1. Acesso em 5 nov. 2023.

CENTRO DE ESTUDOS SOBRE TECNOLOGIAS WEB (CEWEB.BR). *Livro Digital Conectado*. Uma proposta de publicação digital para a educação. São Paulo: Ceweb.br, s.d. Disponível em: https://edupub.ceweb.br/. Acesso em 5 nov. 2023.

ECMA INTERNATIONAL. *ECMA-404*. The JSON data interchange syntax. 2. ed. Genebra: ECMA, dez. 2017. Disponível em: https://www.ecma-international.org/publications-and-standards/standards/ecma-404/. Acesso em 3 out. 2023.

GARRISH, M. *Accessible epub 3*. Sebastopol, O'Reilly Media, Inc., fev. 2012. Disponível em: https://www.oreilly.com/library/view/accessible-epub-3/9781449329297/. Acesso em 3 out. 2023.

GARRISH, M. EPUB Accessibility 1.1. *W3C*, 25 maio 2023. Disponível em: https://www.w3.org/TR/epub-a11y-11. Acesso em 5 nov. 2023.

GARRISH, M.; HERMAN, I. EPUB 3 Overview. *W3C*, 19 maio 2023. Disponível em: https://www.w3.org/TR/epub-overview-33. Acesso em 5 nov. 2023.

HARPUR, P. Disability, Access, and Libraries in the Digital Age. *SSRN*, 16 fev. 2016. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2744358. Acesso em 3 out. 2023.

KASDORF, B. Key Issue-EPUB 3's coming of age. *Insights*, v. 26, n. 2, 2013. Disponível em: https://insights.uksg.org/articles/10.1629/2048-7754.06. Acesso em 3 out. 2023.

KIRKPATRICK, A. *et al*. Web Content Accessibility Guidelines (WCAG) 2.1. 2023. *W3C*, 21 set. 2023. Disponível em: https://www.w3.org/TR/WCAG21/. Acesso em: 5 nov. 2023.

LEARNING RECORD STORE (LRS). *Learning Locker* – Open Source Documentation. *Confluence*, 2013. Disponível em: https://learninglocker.atlassian.net/wiki/spaces/DOCS/overview.

LEPORINI, B.; MINARDI, L.; PELLEGRINO, G. Interactive EPUB3 vs. Web Publication for Screen Reading Users: the Case of 'Pinocchio' Book. *In*: GoodTechs '19: EAI International Conference on Smart Objects and Technologies for Social Good, 5, 2019, Valencia. *Anais* […]. New York, Association for Computing Machinery, set. 2019. p. 235-238. Disponível em: https://dl.acm.org/doi/abs/10.1145/3342428.3342701. Acesso em 3 out. 2023.

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR (NIC.BR). *Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nos domicílios brasileiros*: TIC Domicílios 2021. São Paulo: NIC.br, 2022. Disponível em: https://cetic.br/media/docs/publicacoes/2/20221121125504/tic_domicilios_2021_livro_eletronico.pdf. Acesso em: 3 out. 2023.

OVERVIEW. What is the Experience API? *xAPI.com*, s.d. Disponível em: https://xapi.com/overview. Acesso em 5 nov. 2023.

RIPOLL, C. C. *et al*. *Frações no Ensino Fundamental*. Rio de Janeiro: IMPA, fev. 2021. v. 1. Disponível em: https://educacao.amop.org.br/abrir_arquivo.aspx/Fracoes_no_Ensino_Fundamental_Volume_1?cdLocal=2&arquivo=%7B218E61BB-AED6-D3B6-16CC-5D74B3D1AAE4%7D.pdf. Acesso em 5 nov. 2023.

UMA SOLUÇÃO PARA VERIFICAÇÃO DE CONFORMIDADE AO PADRÃO DE ACESSIBILIDADE ENTRE *SITES* GOVERNAMENTAIS

Por **Reinaldo Ferraz**, **Ana Eliza Duarte**, **João Bárbara**, **Adriano C. M. Pereira** e **Wagner Meira Júnior**

RESUMO

A acessibilidade em *sites web* ainda é um desafio em muitos países, inclusive no Brasil, a despeito, por exemplo, da existência do Modelo de Acessibilidade do Governo Eletrônico (eMAG) brasileiro que orienta como *sites* governamentais devem garantir a acessibilidade e eliminar as barreiras de acesso. Em paralelo, é também fundamental verificar a acessibilidade dos *sites* de forma a subsidiar o aperfeiçoamento das políticas e identificar falhas a serem sanadas. Este artigo apresenta uma plataforma para verificação de acessibilidade de *sites* e suas páginas. A plataforma compreende todo o processo de verificação, desde a coleta, o processamento e a apresentação de resultados para vários níveis de abstração, permitindo análises no nível de página, *site* ou mesmo domínio. A plataforma habilita gestores, desenvolvedores e cidadãos a verificarem a acessibilidade dos *sites* governamentais de forma rápida, flexível e objetiva. Finalmente, apresenta-se um estudo de caso da plataforma proposta sendo aplicada à *web* governamental brasileira (gov.br), demonstrando sua aplicabilidade e efetividade, além de identificar pontos de atenção em relação à acessibilidade.

Palavras-chave: Acessibilidade. ASES. WCAG.

I - INTRODUÇÃO

Seguir padrões de acessibilidade é fundamental para a eliminação de barreiras de acesso para pessoas com deficiência (BAPTISTA *et al*., 2016; TAKAGI *et al*., 2009; AIZPURUA *et al*., 2009); porém, esses padrões são pouco usados durante o desenvolvimento de *sites* no Brasil (FREIRE; RUSSO; FORTES, 2008).

Materiais e recursos sobre acessibilidade em língua portuguesa sempre foram muito requisitados no Brasil. A ferramenta de verificação de acessibilidade Web AccessMonitor, desenvolvida pelo governo de Portugal, tem sido utilizada em chamadas públicas e documentos que atestam a conformidade com padrões internacionais de acessibilidade (FREIRE; RUSSO; FORTES, 2008).

Apesar da disponibilidade de ferramentas para a verificação de acessibilidade, uma grande barreira é a falta de preocupação com a acessibilidade das aplicações pelos desenvolvedores no Brasil. A falta de conhecimento dos padrões de acessibilidade é uma das principais causas do baixo número de *sites* acessíveis no país (ANTONELLI, 2018).

Ademais, ainda há a limitação do idioma. Por mais que a língua inglesa seja amplamente conhecida, ela é uma barreira para boa parte da população, inclusive desenvolvedores. Logo, a falta de documentação técnica em português dificulta ainda mais a compreensão sobre a implementação da acessibilidade digital (PICHILIANI; PIZZOLATO, 2021).

No âmbito governamental brasileiro, existe o Modelo de Acessibilidade do Governo Eletrônico (eMAG) (BRASIL, 2021), que orienta como *sites* do governo devem garantir a acessibilidade e eliminar barreiras de acesso a eles. Essa documentação é baseada nas Diretrizes de Acessibilidade para Conteúdo Web (WCAG) 2.0 (W3C, 2021).

Conforme as diretrizes do eMAG, o Governo Federal brasileiro desenvolveu o Avaliador e Simulador de Acessibilidade em Sítios (ASES) (BRASIL, 2017), uma ferramenta de verificação

de acessibilidade de *sites*, a qual ajuda desenvolvedores a identificar barreiras de acesso em *sites* para torná-los acessíveis para as pessoas com deficiência.

Uma limitação dessa ferramenta é a dificuldade de verificar um *site* inteiro, já que as interfaces disponíveis na Web não permitem a verificação de mais de uma página ao mesmo tempo. A partir desse cenário, foi criado o projeto TIC Web Acessibilidade, com o objetivo de obter um panorama da acessibilidade em *sites* governamentais brasileiros. O projeto existe desde 2010 e fazia coletas uma vez ao ano para a verificação de diversos indicadores, como conformidade com padrões de marcação do World Wide Web Consortium (W3C) (W3C, 1997), suporte ao Internet Protocol version 6 (IPv6), idioma da página e conformidade com os padrões de acessibilidade do governo eletrônico brasileiro. Os dados até 2017 estão publicados na plataforma do projeto (NIC.BR, s.d.).

Com a evolução das ferramentas e um maior interesse por indicadores de acessibilidade, em 2020 o projeto foi direcionado para indicadores mais precisos de acessibilidade, com o objetivo de atuar de forma semelhante ao Observatório de Acessibilidade do governo português (PORTUGAL, 2020). A ideia foi desenvolver uma plataforma que colete os *sites* governamentais e verifique a acessibilidade de diversas páginas de cada um. Assim, será possível haver uma publicação com um diretório de *sites* e sua conformidade com o modelo brasileiro de acessibilidade.

Na primeira versão da TIC Web, só foi possível verificar a conformidade com os padrões de acessibilidade. Isso significa que só foram atestadas páginas com 100% de conformidade ou não. Essa situação gerava um desafio, pois páginas com 1 erro tinham o mesmo peso de páginas com 1.000 erros: isso foi aprimorado e será apresentado no decorrer deste artigo.

Em dezembro de 2021, foi lançada a plataforma TIC Web Acessibilidade (CEWEB, 2023), com o objetivo de apresentar o nível de conformidade das páginas e *sites* coletados e os erros mais comuns encontrados durante a verificação.

Nas próximas seções, serão descritos o desenvolvimento da aplicação e os resultados obtidos nas primeiras coletas efetuadas.

II - PLATAFORMA DE ACESSIBILILDADE

Esta seção apresenta e descreve a plataforma de acessibilidade. Na Seção II-A, será explicada a metodologia e a ferramenta AsesWeb, que faz parte da plataforma de acessibilidade desenvolvida. Em seguida, será apresentada a arquitetura da ferramenta na Seção II-B. Na Seção II-C, as funcionalidades da ferramenta são descritas e, por fim, na Seção II-D, são explicados alguns detalhes de implementação.

II-A. METODOLOGIA

Os sítios a serem avaliados partem de uma lista contendo candidatos pertencentes ao domínio-alvo. A primeira etapa do processo consiste na coleta de todas as páginas (ou de um conjunto significativo) contidas em cada um dos sítios candidatos. O objetivo é ampliar ao máximo a cobertura do *site*, desconsiderando páginas muito profundas e pouco acessadas e otimizando os recursos computacionais. A lista de sítios e páginas candidatos é atualizada de forma contínua e ininterrupta, considerando aspectos como taxa de atualização e cobertura. A coleta contínua também permite uma maior confiabilidade e atualidade dos resultados. Ao final dessa etapa, também são realizados testes de disponibilidade, por meio da avaliação do *status* de resposta para as requisições *Hypertext Transfer Protocol* (HTTP) e filtro de compatibilidade de tipo de arquivo, eliminando aqueles que

não estejam no formato *HyperText Markup Language* (HTML). A segunda etapa consiste na avaliação das páginas coletadas pela ferramenta de análise de acessibilidade ASES, que sumariza os resultados da avaliação em uma "nota de acessibilidade", além de apresentar o número de erros encontrados e sua localização no código fonte da página. Na terceira e última etapa, esses resultados são novamente filtrados e armazenados para serem agregados e visualizados.

II-B. ARQUITETURA

A plataforma tem os seguintes requisitos básicos:

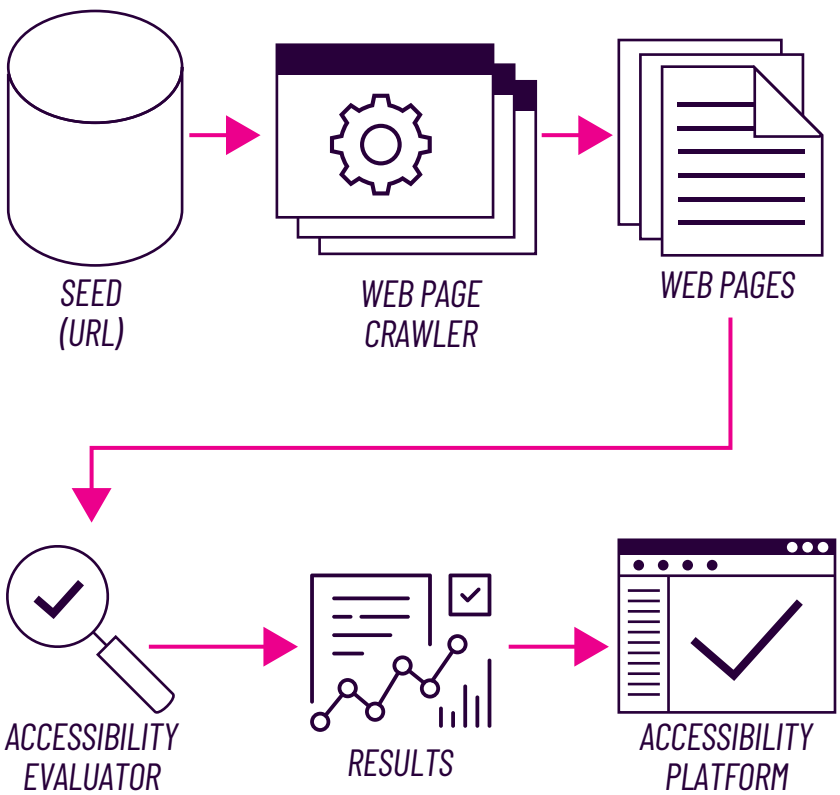
- as páginas devem ser coletadas continuamente, buscando manter a atualidade dos indicadores gerados;
- páginas coletadas devem estar disponíveis e funcionais, requisito restrito a páginas HTML com código de retorno 200;
- as diretrizes constantes de arquivos, como robots.txt, devem ser seguidas, sendo necessário eventualmente contatar o administrador do sítio;
- o conteúdo da página deve estar inteiramente em HTML, pois os *scripts Javascript* atualmente não são executados pelo Coletor.

A arquitetura da plataforma, ilustrada na Figura 1, tem três módulos principais, descritos a seguir.

- Coletor Web: um módulo da ferramenta que coleta e armazena um conjunto das páginas de um determinado *site* a partir de uma lista semente de *Uniform Resource Locator* (URL), extrai todos os endereços das páginas coletadas e os armazena para fins de coletas futuras. É baseado nos conceitos de Crawler e Scraper, estratégias que percorrem recursivamente em profundidade todas as páginas de um *site*.

- Avaliador Assíncrono: um módulo que gerencia o processamento de páginas coletadas. O avaliador instancia módulos ASES e os define, levando em conta o número de páginas coletadas, filas de avaliação. Os arquivos *JavaScript Object Notation* (JSON), resultantes do avaliador, são armazenados.
- Portal de Acessibilidade: uma aplicação *web* que armazena os resultados de avaliação, exibe as páginas analisadas pelo ASES e constrói indicadores globais.

FIGURA 1 - ARQUITETURA DA PLATAFORMA DE ACESSIBILIDADE



Fonte: Elaboração própria.

II-C. FUNCIONALIDADES

O portal de acessibilidade apresenta as seguintes funcionalidades:

- Faixas de Conformidade: percentual de sítios e páginas em conformidade com as métricas de avaliação da ferramenta AsesWeb.
- *Sites* Verificados: lista com todos os sítios verificados pelo portal TIC Web Acessibilidade. Para cada um, há informações detalhadas sobre as páginas coletadas, bem como avisos e erros de análise de acessibilidade de acordo com as recomendações do eMAG.
- Erros Mais Comuns: sumário de erros e avisos encontrados em todas as páginas verificadas, organizadas por seções. Em cada uma, há uma descrição detalhada das recomendações que serviram de base de avaliação.
- 10 melhores *sites*: lista dos 10 *sites* com melhor média na avaliação de acessibilidade, para servir de referência na avaliação das recomendações adotadas.
- Portal www.gov.br: filtro que apresenta uma lista das páginas analisadas que fazem parte do portal único do Governo Federal brasileiro.
- Histórico de Avaliações: histórico temporal do número de sítios e páginas cujo resultado da avaliação é adicionado ao portal de acessibilidade.

II-D. DETALHES DE IMPLEMENTAÇÃO

Na versão atual, o processo de coleta utiliza a biblioteca Scrapy (ZYTE, s.d.); futuramente, pretende-se utilizar o arcabouço DynWebStats (GUERRA, 2016), que escalona páginas a serem coletadas, considerando várias métricas e a taxa de mudança das páginas. O Portal de Acessibilidade utiliza o *framework web* Ruby On Rails (2023) e o SGBD não-relacio-

nal MongoDB (s.d.). Tanto o Avaliador Assíncrono quanto o Portal de Acessibilidade são executados em contêineres Docker (s.d.) para maior facilidade e flexibilidade.

Os resultados da avaliação de acessibilidade foram mapeados em objetos utilizando o *Object Relational Mapper* (ORM), a fim de facilitar a modelagem da interface de inserção dos dados. Para o *frontend*, foram utilizados o *framework* Bootstrap (GETBOOTSTRAP, s.d.) e o *Javascript*.

A ferramenta AsesWeb segue os parâmetros do documento "*Métricas para avaliação de acessibilidade virtual*" (BRASIL, 2015), que dá peso a cada tipo de erro encontrado nas páginas. Erros de peso 1 são menos impeditivos que os de peso 3. Por exemplo, erros de marcação de HTML tem peso 3, enquanto erros de marcação de Cascading Style Sheets (CSS) tem peso 1. A partir da identificação e da marcação de peso de erros, a ferramenta gera uma nota da página, a qual vai de 0 a 100, a partir dos seguintes critérios:

- acima de 95: *site*/página com poucos erros;
- entre 85 e 94,99: o resultado requer atenção;
- entre 70 e 84,99; há um número considerável de erros;
- abaixo de 70, o *site*/página com muitos erros.

II-E. ESTUDO DE CASO: GOV.BR

A primeira coleta de páginas aconteceu em janeiro de 2021 e seguiu até dezembro do mesmo ano, quando a plataforma foi lançada publicamente. Na data de lançamento, o portal de acessibilidade contava com 418 *sites* e 267.090 páginas *web*.

Naquele momento, havia apenas 2.407 páginas com pontuação de conformidade acima de 95 e somente uma página com pontuação 100 (sem nenhum erro). Também não havia nenhum *site* na faixa de 95 e 100 pontos.

No decorrer de 2022, mais *sites* foram coletados para a plataforma. No momento da redação deste artigo, havia 1.445 *sites* e 415.117 páginas avaliadas.

Na análise para esse artigo, 7 *sites* foram identificados na faixa de conformidade acima de 95. A quantidade de páginas de cada *site* varia entre 38 e 139 páginas avaliadas, o que dá uma média de aproximadamente 77 páginas de cada *site* nessa faixa de conformidade. O sítio com a melhor nota média, 98.83, tem 93 páginas avaliadas, sendo 48 delas sem nenhum erro: o restante das páginas tem em sua maioria poucos erros.

Na avaliação por Unidade Federativa brasileira, ainda há muito trabalho a ser feito. O estado que tem a melhor média de *sites* tem a nota 82.41. Apesar de ser a melhor pontuação entre os 27 estados brasileiros, esse número está na terceira faixa de conformidade, sendo um resultado não muito bom.

Também foram avaliados os erros mais comuns encontrados nas páginas da plataforma. O principal deles é descrever os *links* clara e sucintamente, que significa que a descrição do *link* não deve ter textos como "leia mais" ou "saiba mais" (referente à Recomendação 3.5 – Descrever *links* clara e sucintamente do eMAG e Success Criterion 2.4.9 Link Purpose WCAG). A Tabela 1 apresenta os cinco erros mais comuns encontrados, detalhando o total de erros e o número de páginas em que esses erros foram encontrados.

TABELA 1 - ERROS MAIS COMUNS ENCONTRADOS NAS PÁGINAS

DESCRIÇÃO	TOTAL DE ERROS	NÚMERO DE PÁGINAS
1 - Descrever *links* clara e sucintamente	6.689.759	398.463
2 - Disponibilizar todas as funções da página via teclado	5.589.379	180.685
3 - Organizar o código HTML de forma lógica e semântica	3.053.675	413.001
4 - Fornecer alternativa em texto para as imagens do sítio	2.323.020	405.687
5 - Utilizar corretamente os níveis de cabeçalho	747.952	403.602

Fonte: Elaboração própria.

O eMAG considera que uma página com barreira de acesso por teclado não tem as funções JavaScript definidas para teclado. Possivelmente são páginas com funções com "*onmouseover*" sem "*onfocus*", que possibilita o foco por teclado. Isso pode explicar a segunda linha da Tabela 1, em que um número de páginas menor tem uma incidência grande de erros.

Nesse sentido, verifica-se que a TIC Web Acessibilidade é uma poderosa ferramenta para promoção da transparência em *sites* e aplicações Web; por exemplo: *sites* governamentais com baixa pontuação podem ser acionados pela justiça. Também é possível servir de apoio para gestores cobrarem dos desenvolvedores a conformidade com a acessibilidade em suas aplicações. Além disso, o diagnóstico da baixa pontuação na plataforma pode ser um vetor de transformação que vai desde a capacitação de desenvolvedores até a criação de políticas públicas efetivas sobre acessibilidade na Web.

III - CONSIDERAÇÕES E IMPLEMENTAÇÕES FUTURAS

Este artigo apresentou uma plataforma[9] para avaliação da acessibilidade da Web, em especial de *sites* governamentais brasileiros, contendo várias funcionalidades.

Os resultados obtidos foram positivos e, considerando o caso de uso da Web brasileira, serão úteis para cidadãos interessados em verificar o nível de acessibilidade nos sítios públicos e para desenvolvedores e administradores tomarem conhecimento das boas práticas e de soluções de possíveis erros.

Futuramente, pretende-se adicionar à semente, que atualmente compreende apenas o gov.br, *sites* dos domínios jus.br, mp.br e mil.br. Além do ASES, novos analisadores com modelos diferentes podem ser adicionados, como as Diretrizes de Acessibilidade para Conteúdo Web (Web Content Accessibility Guidelines – WCAG) (W3C, 2021). Ademais, existe um plano de evoluir a coleta estática das páginas para um mecanismo dinâmico usando um arcabouço de coleta dinâmica DynWebStats (W3C, 1997).

9 Essa ferramenta está publicada sob a Licença Creative Commons 4.0 (CC BY-SA 4.0), disponível em Ceweb (2023).

REFERÊNCIAS

AIZPURUA, A. *et al*. Transition of accessibility evaluation tools to new standards. *In*: 2009 International Cross-Disciplinary Conference on Web Accessibililty (W4A) (W4A '09), abr. 2009. *Anais* […]. New York: Association for Computing Machinery, 20 abr. 2009. p. 36-44. Disponível em: https://doi.org/10.1145/1535654.1535662. Acesso em 3 out. 2023.

ANTONELLI, H. L. *et al*. A survey on accessibility awareness of Brazilian web developers. *In*: International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2018), 8, jun. 2018. *Anais* […]. New York: Association for Computing Machinery, 20 jun. 2018. p. 71-79. https://doi.org/10.1145/3218585.3218598. Acesso em 3 out. 2023.

BAPTISTA, A. *et al*. Web accessibility challenges and perspectives: a systematic literature review. *In*: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), 11, jun. 2016, Gran Canaria. *Anais* […], 2016. p. 1-6. Disponível em: https://ieeexplore.ieee.org/document/7521619/similar#similar. Acesso em 3 out. 2023.

BRASIL. *Métricas para avaliação de acessibilidade virtual*. Brasília; Bento Gonçalves: MPOG; UFRGS, mar. 2015. Disponível em: http://ticwebacessibilidade.ceweb.br/ases/metricas_ases_eselo_junho_2016.pdf. Acesso em 4 out. 2023.

BRASIL. *Avaliador e Simulador de Acessibilidade em Sítios (ASES)*. Brasília: Departamento de Governo Eletrônico, 2017. Disponível em: https://asesweb.governoeletronico.gov.br/. Acesso em 4 out. 2023.

BRASIL. *eMAG* – Modelo de Acessibilidade em Governo Eletrônico. Brasília: Departamento de Governo Eletrônico, 2021. Disponível em: http://emag.governoeletronico.gov.br/. Acesso em 4 out. 2023.

BUILD fast, responsive sites with Bootstrap. *Getbootstrap*, s.d. Disponível em: https://getbootstrap.com. Acesso em 4 nov. 2023.

CENTRO DE ESTUDOS SOBRE TECNOLOGIAS WEB (CEWEB). *TIC Web Acessibilidade*. São Paulo: Ceweb, nov. 2023. Disponível em: https://ticwebacessibilidade.ceweb.br/. Acesso em 29 jan. 2024.

COMPRESS the complexity of modern web apps. *Ruby on Rails*, 11 out. 2023. Disponível em: https://rubyonrails.org. Acesso em 4 nov. 2023.

CRIE mais rapidamente. Crie com mais inteligência. *MongoDB*, s.d. Disponível em https://www.mongodb.com. Acesso em 4 nov. 2023.

FREIRE, A. P.; RUSSO, C. M.; FORTES, R. P. M.. A survey on the accessibility awareness of people involved in web development projects in Brazil. *In*: 2008 International cross-disciplinary conference on Web accessibility (W4A) (W4A '08), abr. 2008. *Anais* […]. New York: Association for Computing Machinery, 21 abr. 2008. p. 87-96. Disponível em: https://doi.org/10.1145/1368044.1368064. Acesso em 3 out. 2023.

GUERRA, I. *et al*. DynWebStats: A Framework for Determining Dynamic and Up-to-date Web Indicators. *In*: Brazilian Symposium on Multimedia and the Web (Webmedia '16), 22 nov. 2016. *Anais* […]. New York: Association for Computing Machinery, nov. 2016. p. 247-254. Disponível em: https://doi.org/10.1145/2976796.2976857. Acesso em 4 out. 2023.

MAKE better, secure software from the start. *Docker*, s.d. Disponível em https://www.docker.com. Acesso em 4 nov. 2023.

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR (NIC.BR). *TIC Web Acessibilidade*. São Paulo: NIC.br; UFMG, s.d. Disponível em: http://mapaweb.speed.dcc.ufmg.br/relatorios/index. Acesso em 4 out. 2023.

PICHILIANI, T. C. P. B.; PIZZOLATO, E. B. Cognitive disabilities and web accessibility: a survey into the Brazilian web development community. *Journal on Interactive Systems*, Porto Alegre, v. 12, n. 1, p. 308-327, 2021. Disponível em: https://sol.sbc.org.br/journals/index.php/jis/article/view/982. Acesso em: 6 jan. 2023.

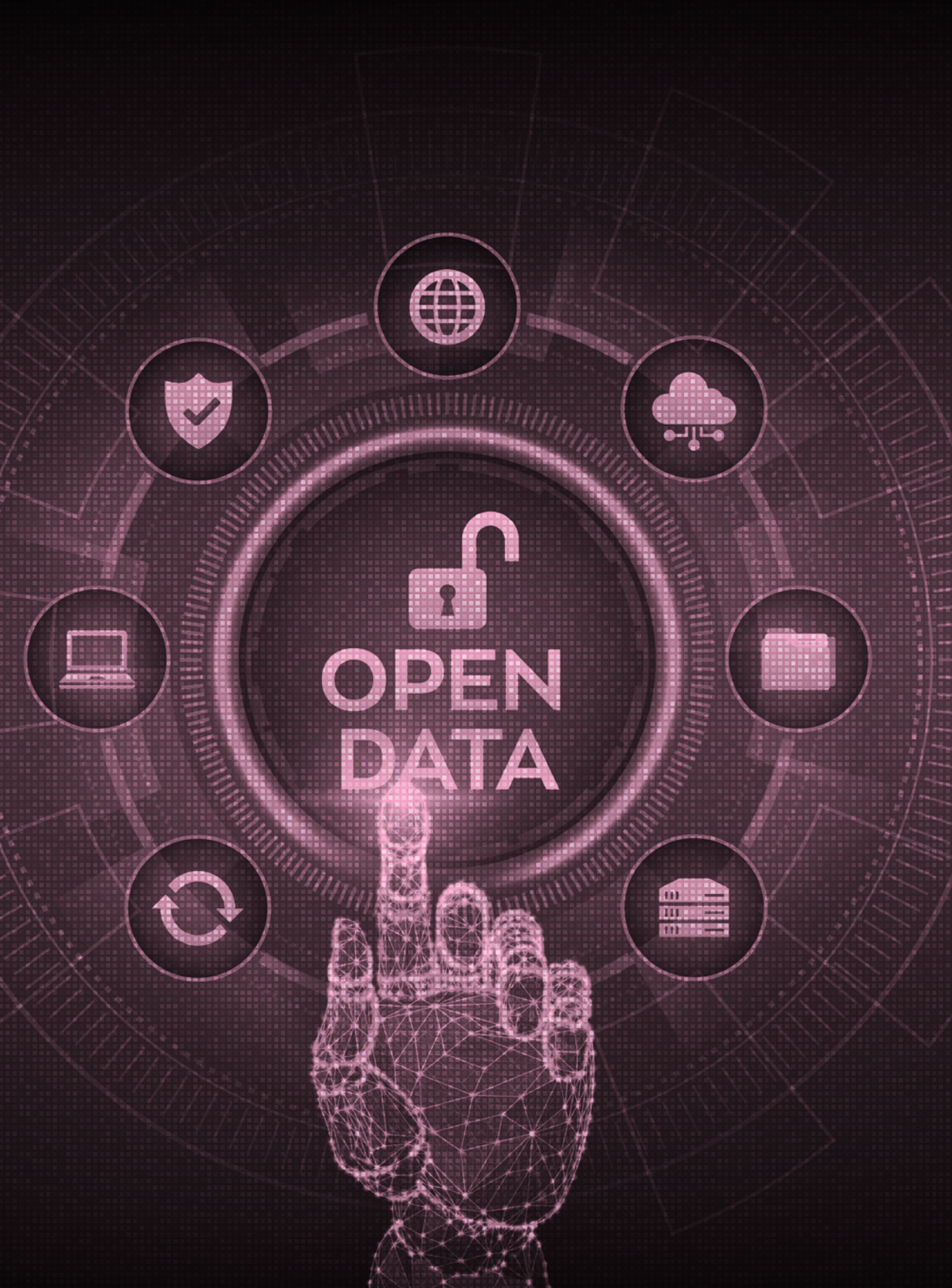
PORTUGAL. *Observatório Português da Acessibilidade Web*. Lisboa: AMA, 2020. Disponível em: https://observatorio.acessibilidade.gov.pt/. Acesso em 4 out. 2023.

SCRAPY. An open source and collaborative framework for extracting the data you need from websites. *Zyte*, s.d. Disponível em: https://scrapy.org. Acesso em 4 nov. 2023.

TAKAGI, H. *et al*. Collaborative web accessibility improvement: challenges and possibilities. *In*: International ACM SIGACCESS conference on Computers and accessibility (Assets '09), 11, out. 2009. *Anais* […]. New York: Association for Computing Machinery, out. 2009. p. 195-202. Disponível em: https://doi.org/10.1145/1639642.1639677. Acesso em 3 out. 2023.

WORLD WIDE WEB CONSORTIUM (W3C). Launches International Web Accessibility Initiative. *W3C*, 7 abr. 1997. Disponível em: https://www.w3.org/Press/WAI-Launch.html. Acesso em 4 out. 2023.

WORLD WIDE WEB CONSORTIUM (W3C). WCAG 2 Overview. *W3C*, 2021. Disponível em: https://www.w3.org/WAI/standards-guidelines/wcag/. Acesso em 4 out. 2023.


OPEN
DATA

MODELO DE AVALIAÇÃO DE DADOS ABERTOS NOS PORTAIS GOVERNAMENTAIS BRASILEIROS[10]

Por **João Bárbara**, **Caroline Burle**, **Adriano César Machado Pereira**, **Ana Eliza Duarte**, **Wagner Meira Júnior** e **Letícia da Silva Macedo Alves**

10 Documento apresentado no XI Congreso Internacional en Gobierno, Administración y Políticas Públicas GIGAPP (2022).

RESUMO

Um dos grandes desafios relacionados à publicação e ao consumo de dados na Web é a padronização das estruturas informacionais. Muitas organizações publicam seus conjuntos de dados de uma forma particular, sem considerar os padrões e os princípios de publicação desenvolvidos estabelecidos pela comunidade, com o objetivo de facilitar o compartilhamento e o aproveitamento da informação. Tendo em vista esse cenário, é necessária a criação de mecanismos capazes de orientar os publicadores a utilizarem princípios e boas práticas sobre publicação de dados abertos na Web e, futuramente, permitir uma melhor governança de dados em seus portais de informação. Isso resultaria em uma série de benefícios, como uma melhor compreensão dos dados, maior facilidade de descoberta e processamento, reuso e compartilhamento dos dados pelos usuários consumidores. Ao se analisar um conjunto de portais publicadores de dados abertos, percebeu-se que não há uma padronização dos dados publicados, o que impede ou dificulta o acesso a dados abertos por parte dos usuários. Nesse contexto, este trabalho propõe uma metodologia para avaliar esse ambiente heterogêneo e apresentar uma estimativa capaz de mapear o uso e as recomendações de boas práticas para fins de publicação e disponibilização de dados abertos em portais governamentais brasileiros. Criou-se um Modelo de Avaliação de Dados Abertos nos Portais Governamentais Brasileiros e um processo de análise que apresentam uma abordagem ágil e simplificada, capazes de identificar e avaliar os dados comumente apresentados nos diversos portais de publicação de dados abertos, tendo em vista Os 10 Princípios de Governança de Dados Abertos, documento construído pela Universidade Federal de Minas Gerais (UFMG, no prelo), e a recomendação do W3C Data on the Web Best Practices (DWBP, 2017).

Palavras-chave: Dados Abertos. Governo Aberto. Boas Práticas para Dados na Web. Governança de Dados. Governo Brasileiro.

I - INTRODUÇÃO

Diante da falta de padronização nas estruturas de informações, apresenta-se um dos grandes desafios relacionados à divulgação e ao acesso de dados na Web. Muitas organizações publicam seus conjuntos de dados de maneira única, sem aderir aos padrões e princípios de publicação estabelecidos pela comunidade web, cujo propósito é facilitar o compartilhamento e a utilização da informação.

Nesse contexto, nosso trabalho propõe uma metodologia para avaliar esse ambiente heterogêneo e apresentar uma estimativa que identifica o uso e as diretrizes de boas práticas para a publicação e disponibilização de dados abertos em portais governamentais brasileiros. Desenvolvemos um novo Modelo de Avaliação de Dados Abertos nos Portais Governamentais Brasileiros com um processo de análise ágil e simplificado. Esse modelo tem a capacidade de identificar e avaliar os dados normalmente apresentados nos diversos portais de divulgação de dados abertos, levando em consideração os *10 Princípios de Governança de Dados Abertos*, um documento elaborado pela Universidade Federal de Minas Gerais (UFMG, no prelo), e as *Recomendações das Melhores Práticas de Dados na Web do W3C* (DWBP, 2017).

Para resolver o desafio da falta de padronização, é fundamental criar mecanismos que orientem os responsáveis pela publicação a adotarem princípios e boas práticas para a divulgação de dados abertos na internet, com o objetivo de promover uma governança mais eficaz dos dados em seus portais de informações. Isso traria diversos benefícios, como uma compreensão mais clara dos dados, maior facilidade na sua descoberta e utilização, além de incentivar sua reutilização e compartilhamento por parte dos usuários. Ao se analisar uma variedade de portais que disponibilizam dados abertos, observamos a falta de padronização nos dados publicados, o que dificulta ou impede o acesso a esses dados por parte dos usuários.

Ao analisarmos uma variedade de portais que disponibilizam dados abertos, observamos a falta de padronização nos dados publicados, o que dificulta ou impede o acesso a eles por parte dos usuários.

Este artigo está organizado da seguinte forma: além da introdução, apresentada nesta seção, na Seção II é descrita a metodologia do trabalho, enfatizando seus principais pontos e conexão entre as etapas. A Seção III apresenta o desenvolvimento do trabalho, com a aplicação da metodologia apresentada e o uso de dados reais de sites governamentais brasileiros. Por fim, a Seção IV apresenta as conclusões e as principais contribuições do projeto, e discute os benefícios, os desdobramentos do trabalho, com foco nos gestores públicos, e o potencial de uso desses resultados para criação e fomento de políticas públicas e governança metropolitana.

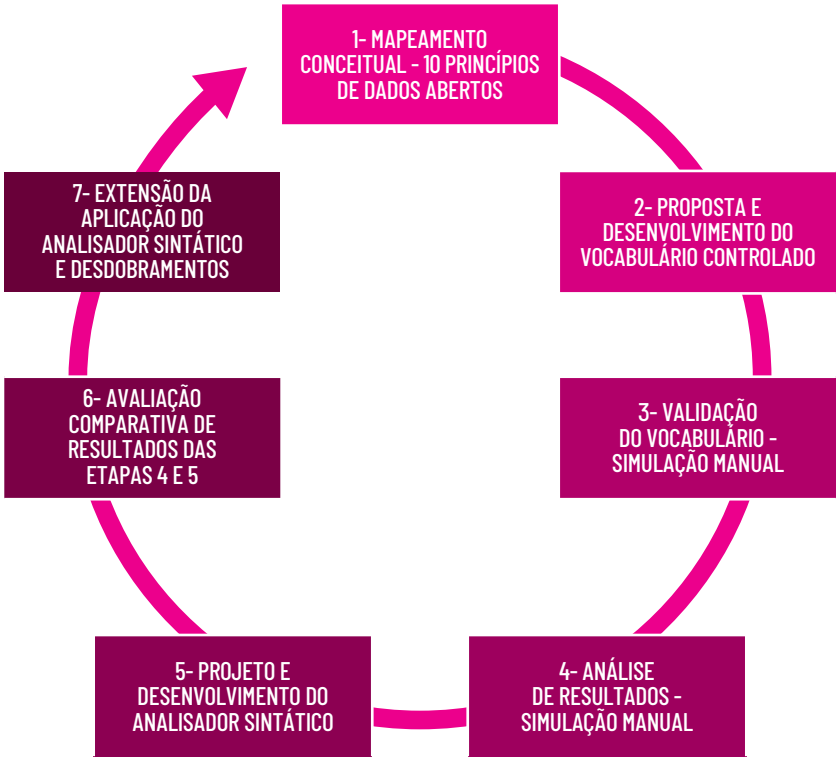
II - METODOLOGIA

O Modelo de Avaliação de Dados Abertos nos Portais Governamentais Brasileiros tem por objetivo pesquisar, projetar e implementar um arcabouço de ferramentas que permita avaliar as páginas da Web brasileira a partir da extração de dados e metadados disponíveis nos próprios portais. Esse arcabouço possui indicadores utilizados para quantificar critérios de avaliação da dinâmica e da evolução de um ou mais domínios da Web brasileira. A definição de indicadores considera não apenas seu significado, sua semântica, mas também sua viabilidade em termos de coleta de dados. Finalmente, verifica-se a qualidade da estimativa de cada indicador. Todos esses passos são acompanhados pela equipe à frente do projeto e, quando for o caso, são executados conjuntamente com parceiros.

Essa frente de trabalho tem por objetivo evoluir o processo de construção de indicadores de dados abertos para páginas e *site* da Web ".gov.br". Originalmente, no plano de trabalho do projeto (LÓSCIO; BURLE; CALEGARI, 2019), está prevista a ela-

boração de uma metodologia inicial proposta e adaptada para contemplar um novo Modelo de Avaliação de Dados Abertos nos Portais Governamentais Brasileiros.

FIGURA 1 - ARQUITETURA DA PLATAFORMA DE ACESSIBILIDADE



Fonte: Elaboração própria.

A Figura 1 ilustra a metodologia do Modelo de Avaliação de Dados Abertos, formada por 7 Etapas:

i) Mapeamento conceitual dos 10 princípios de governança em dados abertos, tendo como base teórica o documento Os 10 Princípios de Governança de Dados Abertos (UFMG, no prelo).

ii) Proposta e desenvolvimento de um vocabulário controlado como um modelo que sintetize a descrição dos princípios de governança indicados, por meio de um conjunto de metada-

dos capaz de identificar os dados explicitamente legíveis dos portais de publicação de dados abertos e classificá-los.

iii) Validação do vocabulário controlado por meio da simulação manual de analisador sintático em um subconjunto de portais de publicação de dados abertos, verificando sua capacidade de reconhecimento, processamento e classificação dos dados disponibilizados a seus consumidores. Essa validação manual será executada com uma amostra contendo 33 portais com uma média de cinco Uniform Resource Identifier (URI) por portal de publicação.

iv) Análise dos resultados obtidos pela simulação manual e proposição de uma implementação computacional do analisador sintático e, consequentemente, do vocabulário de termos. Esse último visa a automatização do processo de análise e classificação, com objetivo de ampliar o conjunto amostra de portais de publicação de dados abertos governamentais.

v) Projeto e desenvolvimento do analisador sintático proposto na Etapa 4. Para validação, utilizando a mesma amostra de portais de publicação de dados abertos utilizados na Etapa 3, com a validação das mesmas URI. Os resultados obtidos serão disponibilizados para posterior avaliação comparativa.

vi) Avaliação comparativa dos resultados apresentados nas Etapas 4 e 5, cujo objetivo é verificar a eficiência da implementação do vocabulário de termos e regras ao analisar a equivalência dos resultados.

vii) Avaliação da escalabilidade do analisador sintático em uma nova amostra de portais. Para isso, um novo conjunto de páginas web será escolhido a partir da coleta extensiva de dados executada.

A próxima seção descreve o desenvolvimento da metodologia apresentada, apresentando os detalhes de cada Etapa e resulta-

dos obtidos com sua aplicação a um conjunto de dados reais de sites governamentais brasileiros.

III - DESENVOLVIMENTO E ANÁLISE

Esta seção descreve o desenvolvimento de cada etapa da metodologia, apresentada na seção anterior, com uso de uma amostra de dados reais da Web governamental brasileira.

ETAPA 1 – MAPEAMENTO CONCEITUAL

A primeira etapa realizada foi o mapeamento dos 10 princípios de governança de dados abertos, considerando a DWBP. Seu objetivo foi apresentar uma proposta capaz de caracterizar cada princípio, tendo como base um subconjunto de recomendações, a partir das recomendações principais.

Essa modelagem visa reduzir ao máximo a subjetividade que aparece na tentativa de classificar os portais de publicação: é feita apenas por meio dos conceitos apresentados pelos 10 Princípios de Governança de Dados Abertos (UFMG, no prelo). Como as boas práticas foram baseadas em análises implementáveis[11], o mapeamento dos princípios em subconjuntos de recomendações reduz a subjetividade e torna viável a classificação dos portais a partir de uma perspectiva quantitativa.

O vocabulário de termos baseia-se em um subconjunto de indicadores, extraídos dos dados contidos no conjunto das boas práticas.

ETAPA 2 – VOCABULÁRIO CONTROLADO

O objetivo das DWBP é estimular e possibilitar a expansão da Web como um ambiente para a troca de dados. Dessa forma, as boas práticas fornecem a publicadores e consumidores de dados

11 Mais informações disponíveis em Lóscio, Burle e Calegari (2019).

um conjunto de padrões que permitem a melhoria da qualidade da informação nos portais de dados abertos, de forma a facilitar o compartilhamento e o aproveitamento da informação.

A principal premissa é que a descrição da informação disponibilizada seja o mais legível possível, tanto para pessoas quanto para máquinas, a fim de facilitar a compreensão dos dados e proporcionar uma maior qualidade.

O vocabulário controlado é uma lista de palavras-chave (nomeadas neste trabalho como tokens e regras) derivada de cada princípio de governança e das boas práticas para dados na Web, que, ao ser aplicada em analisador sintático, permite avaliar a descrição dos dados explicitamente disponibilizados em um determinado portal de publicação e classificá-lo.

Se no final do processo de avaliação do analisador (parsing) forem encontrados tokens vinculados a um determinado princípio, infere-se que a probabilidade do portal de publicação se enquadrar a ele é grande.

Ao final, teremos uma estimativa básica do nível da padronização, baseada nos princípios de governança e, consequentemente, nas DWBP para um conjunto de portais de publicação.

Conforme mencionado, ter uma referência sobre o grau de padronização dos dados dos portais de publicação permitirá mapear quais diretrizes os publicadores têm utilizado para apresentação dos dados de pesquisa, além de visualizar como o uso, ou o não uso, das DWBP (LÓSCIO; BURLE; CALEGARI, 2017) impacta na qualidade dos dados disponibilizados aos consumidores. Além disso, esses resultados possibilitarão novas e mais eficientes iniciativas no estímulo do uso de recomendações e princípios de governança pelos publicadores.

A seguir, descrevem-se alguns exemplos dessa série de vocabulários de termos vinculados a cada um dos princípios de governança.

TABELA 1 - PRINCÍPIOS DE GOVERNANÇA MAPEADOS DE ACORDO COM AS DWBP

Completos
Informações sobre Metadados (DWBP 8.2): 1) Palavras-chave (*tokens*): metadados, informações adicionais, dicionário(s), dicionário de dados, taxonomia, critério(s), descrição do conjunto de dados, título, URI, palavra(s) chave, data de publicação, data de criação, criação, frequência, atualização, data de atualização, contato, granularidade, referência(s), responsável(is), idioma, fonte(s), versão, mantenedor(es), tema, formato data, metadado(s) estrutural(is), campo, tipo de dados, métrica, última modificação, última atualização, descrição, cobertura geográfica, cobertura temporal, escopo geopolítico, autor(es), criado, entidade responsável, ponto de contato, período temporal, data da última modificação, temas, categorias, formato, formato de mídia, licença, identificador, relação, tipo de conteúdo, recursos; 2) Verificação da presença de URI para arquivos de metadados anexos.
Informações sobre Licença (DWBP 8.3): 1) Palavras-chave (*tokens*): licença, tipo de licença, termos da(e) licença, restrições/ restrição.
Informações sobre Procedência dos dados (DWBP 8.4): 1) Palavras-chave (*tokens*): fonte(s), criador(es), responsável(is), área responsável, mantenedor(es), autor(es), data de publicação, editor(es/as).
Informações sobre Qualidade da publicação dos dados (DWBP 8.5): 1) Palavras-chave (*tokens*): qualidade dos dados, qualidade, integridade, integridade dos dados, métrica, disponibilidade, disponibilidade dos dados.
Informações sobre Versionamento dos dados (DWBP 8.6): 1) Palavras-chave (*tokens*): versão, versão atual, atualização, última atualização, data, data criação, última versão, histórico, obsoleto, frequência, frequência de atualização, histórico de mudanças, histórico de modificações, histórico de modificação, última modificação(ões), periodicidade.
Informações sobre *Feedback* (DWBP - 8.12): 1) Palavras-chave (*tokens*): contato, *feedback*, formulário, *rank*, ranqueamento, esperado, avaliação, avaliações, botão, botões, qualidade dos dados, qualidade, comentário, questionamento, classifica, classificação, correção, revisão, compartilhar, compartilhe, informe, fale com, entre em, sugestões.

Primários

Informações sobre Procedência dos dados (DWBP 8.4):

1) Palavras-chave (*tokens)*: fonte(s), criador(es), responsável(is), área responsável, mantenedor(es), autor(es), data de publicação, editor(es/as).

Informações sobre Qualidade dos dados (DWBP 8.5):

1) Palavras-chave (*tokens*): qualidade dos dados, qualidade, integridade, integridade dos dados, métrica, disponibilidade, disponibilidade dos dados (sugestão de ter uma forma de testar a disponibilidade dos dados – *link* disponível e com código HTTP que garanta essa disponibilidade).

Informações sobre *Feedback* (DWBP - 8.12):

1) Palavras-chave (*tokens*): Contato, *feedback*, formulário, *rank*, ranqueamento, esperado, avaliação, avaliações, botão, botões, qualidade dos dados, qualidade, comentário, questionamento, classifica, classificação, correção, revisão, compartilhar, compartilhe, informe, fale com, entre em, sugestões.

Atuais

Informações sobre Versionamento dos dados (DWBP 8.6):

1) Palavras-chave (*tokens*): versão, versão atual, versões, atualização, última atualização, data, data criação, última versão, histórico, frequência, frequência de atualização, histórico de mudanças, histórico de modificações, histórico de modificação, última modificação(ões);

Informações sobre Garantir acesso aos dados (DWBP 8.10):

1) *Download*:

a) Presença de URL direcionada ao dado, acessado com apenas uma solicitação.

i) *Download* de grandes volumes de dados em uma única requisição: Presença de URI que aponta para conjunto de dados direcionados a arquivos com extensão: zip, rar, 7z, tar, gz.

ii) Dados Atualizados:

(1) Palavras-chave (*tokens*): data, atualização, última, frequência, criado, criação, atualizado, cobertura, temporal, validade.

Atuais

2) API:

a) Palavras-chave (*tokens*): API, documentação, documentação da API, manual, descrição, parâmetros, especificação, *webservice*, REST, RESTful, consumo, retorno, resultados, método, GET, URL, URI, cabeçalhos, *headers*.

b) Dados indisponíveis:

i) URI com código de resposta HTTP na faixa 400 ou 500 ao verificar o acesso, que identifica o conjunto de dados. (É necessário fazer teste de disponibilidade para os URI de conjunto de dados encontrados).

c) URI para documentação das API:

i) Palavras-chave (*tokens*): API, documentação, *swagger*, docs, io-docs, *openApis*.

Facilidade de Acesso Físico ou Eletrônico

Informações sobre Garantir acesso aos dados (DWBP 8.10):

1) *Download:*

a) Presença de URL direcionada ao dado, acessado com apenas uma solicitação.

i) *Download* de grandes volumes de dados em uma única requisição: Presença de URI que aponta para conjunto de dados direcionados a arquivos com extensão: zip, rar, 7z, tar, gz.

i) Dados Atualizados:

(1) Palavras-chave (*tokens*): data, atualização, última, frequência, criado, criação, atualizado, cobertura, temporal, validade.

2) API*:*

a) Palavras-chave (*tokens*): API, documentação, documentação da API, manual, descrição, parâmetros, especificação, *webservice*, REST, RESTful, consumo, retorno, resultados, método, GET, URL, URI, cabeçalhos, *headers*.

b) Dados indisponíveis:

i) URI com código de resposta HTTP na faixa 400 ou 500 ao verificar o acesso ao URI que identifica o conjunto de dados. (É necessário fazer teste de disponibilidade para os URI de conjunto de dados encontrados).

c) URI para documentação das API:

i) Palavras-chave (*tokens*): API, documentação, *swagger*, docs, io-docs, *openApis*;

Fonte: Elaboração própria.

Cada índice do vocabulário de termos representa um dos *10 Princípios de Governança de Dados Abertos* (UFMG, no prelo) e abrange um subconjunto das DWBP.

Cada item do subconjunto apresenta uma ou mais recomendações das boas práticas organizadas em uma lista de *tokens* e/ou regras. Ao analisar as páginas com a publicação de dados abertos dos portais, o analisador busca por dados explícitos que coincidam com os termos ou regras contidas no vocabulário. Quanto maior o pareamento de termos, maior a confiança de que aquele portal de publicação segue a orientação dos padrões de boas práticas vinculadas ao respectivo princípio de governança.

ETAPA 3 – VALIDAÇÃO MANUAL DO VOCABULÁRIO DE TERMOS E REGRAS

A primeira etapa de validação do vocabulário de termos consiste na **aplicação manual do analisador sintático** em uma amostra contendo um subconjunto de portais de publicação de dados abertos, conforme a Tabela 2.

TABELA 2 - AMOSTRA DE PORTAIS DE DADOS ABERTOS GOVERNAMENTAIS

TÍTULO	URL
Alagoas em dados e informações	http://dados.al.gov.br/
Fortaleza Dados Abertos	http://dados.fortaleza.ce.gov.br/
Dados abertos – TCM-CE	https://api-dados-abertos.tce.ce.gov.br/docs/
Dados abertos Distrito Federal	http://dados.df.gov.br/
Dados abertos – Governo do ES	https://transparencia.es.gov.br/DadosAbertos/BaseDeDados#
Dados abertos – Goiás Transparente	http://www.transparencia.go.gov.br/portaldatransparencia/institucional/dados-abertos

TÍTULO	URL
Dados abertos – Assembleia Legislativa de Minas Gerais	http://dadosabertos.almg.gov.br/ws/ajuda/sobre
Dados abertos – Estado de MG	https://dados.mg.gov.br/
Dados abertos do SAGRES – TCE/PB	https://tce.pb.gov.br/servicos/dados-abertos-do-sagres-tce-pb
Dados abertos – Governo de Pernambuco	https://dados.pe.gov.br/
Dados da Prefeitura de Recife	http://dados.recife.pe.gov.br
Dados Abertos Curitiba	http://www.curitiba.pr.gov.br/dadosabertos/
data.rio	http://data.rio/
Dados Geográficos Abertos da Cidade do Rio de Janeiro	https://datariov2-pcrj.hub.arcgis.com/
Dados RS	http://dados.rs.gov.br/
Dados Abertos SC	https://dados.sc.gov.br/
Dados Abertos POA	https://dados.portoalegre.rs.gov.br/
Governo Aberto SP	http://www.governoaberto.sp.gov.br/
Programa de Dados Abertos do Parlamento	http://www.camara.sp.gov.br/transparencia/dados-abertos/
Dados Abertos Legislativos – Senado Federal	http://dadosabertos.senado.gov.br/
Dados abertos da Câmara dos Deputados	https://dadosabertos.camara.leg.br/
Dados Abertos TCE-RS	http://dados.tce.rs.gov.br/
Dados Abertos MPRS	http://dados.mprs.mp.br/

TÍTULO	URL
Portal da Transparência Municipal	http://transparencia.tce.sp.gov.br/
Portal dos Dados Abertos TCE/RN	http://apidadosabertos.tce.rn.gov.br/
Portal de Dados Abertos da Cidade de São Paulo	http://dados.prefeitura.sp.gov.br/
Câmara de Itabira – Dados Abertos	http://www.itabira.cam.mg.gov.br/dados-abertos
Portal dos Dados Abertos da Alesp	https://www.al.sp.gov.br/dados-abertos/
Transparência Anápolis – Dados Abertos	https://transparencia.anapolis.go.gov.br/transparencia/dadosAbertos.jsf
Download de Bases – TRANSPARÊNCIA	http://www.transparencia.mt.gov.br/downloads-de-bases
Dados PB	http://dados.pb.gov.br
Portal da Transparência – Dados Abertos	http://transparencia.campinas.sp.gov.br/
Dados Abertos – TCEMG	https://dadosabertos.tce.mg.gov.br/
Dados Abertos – Ceará Transparente	https://cearatransparente.ce.gov.br/portal-da-transparencia/dados-abertos/conjuntos-de-dados?locale=pt-BR
Portal da Transparência e Mobilidade Urbana de Natal	http://dados.natal.br/
Dados Abertos – Portal da Transparência de Pernambuco	http://web.transparencia.pe.gov.br/dados-abertos/
Dados Abertos – Tribunal de Contas do Estado de Pernambuco	https://www.tce.pe.gov.br/internet/index.php/dados-abertos
Portal de Dados Abertos – Prefeitura de Belo Horizonte	https://dados.pbh.gov.br/

TÍTULO	URL
Portal de Dados Abertos da Câmara Legislativa do Distrito Federal	http://dadosabertos.cl.df.gov.br/
Portal Opendatasus	https://opendatasus.saude.gov.br/
Portal Brasileiro de Dados Abertos	https://dados.gov.br/

Fonte: Elaboração própria.

Sabemos que, como premissa, as DWBP definem que os dados disponibilizados precisam ser legíveis tanto por pessoas e, na maioria das situações, por máquinas. A partir disso, cabe ao analista simular o analisador sintático buscando nos dados descrições legíveis e explícitas ao consumidor e associação de termos que coincidam com os *tokens* ou regras definidas em cada índice do princípio de governança.

A classificação do portal de publicação ocorrerá pela análise do número de associações sintáticas encontradas para cada subconjunto de termos vinculados aos princípios de governança, sem qualquer subjetividade.

A ideia é avaliar se, ao final do processo analítico das páginas, associando exclusivamente os dados legíveis e explicitados aos usuários, é possível classificar os portais de publicação de dados abertos de acordo com Os 10 Princípios de Governança de Dados Abertos (UFMG, no prelo) e com as DWBP.

Cada página de publicação terá seu próprio formulário de resposta, simulando o resultado obtido pelo analisador sintático. Devem constar informações relativas a cada um dos princípios de governança, descrevendo claramente quais os tokens e/ou as regras encontrados e suas respectivas recomendações de boas práticas. O atributo Resultado confirmará se o princípio foi validado ou não.

O campo Observações deve atender a qualquer aspecto que gerou dúvida ou situações que podem ser levadas à discussão no instante da apresentação dos resultados.

TABELA 3 – FORMULÁRIO DE ANÁLISE PARA AVALIAÇÃO MANUAL

PORTAL: **DD/MM/AAAA** **URI:** **DATA:** **RESPONSÁVEL:** ***SCREENSHOT* DA PÁGINA DE PUBLICAÇÃO:**
Princípio: Completos *Tokens* Encontrados: Boas Práticas Encontradas: Resultado: Observações:
Princípio: Primários *Tokens* Encontrados: Boas Práticas Encontradas: Resultado: Observações:
Princípio: Atuais *Tokens* Encontrados: Boas Práticas Encontradas: Resultado: Observações:
Princípio: Facilidade de acesso Físico ou Eletrônico *Tokens* Encontrados: Boas Práticas Encontradas: Resultado: Observações:
Princípio: Processáveis por Máquina *Tokens* Encontrados: Boas Práticas Encontradas: Resultado: Observações:

PORTAL:	
DD/MM/AAAA	
URI:	**DATA:**
RESPONSÁVEL:	
***SCREENSHOT* DA PÁGINA DE PUBLICAÇÃO:**	

Princípio: Não Discriminatório
Tokens Encontrados:
Boas Práticas Encontradas:
Resultado:
Observações:

Princípio: Formatos de propriedade comum ou abertos
Tokens Encontrados:
Boas Práticas Encontradas:
Resultado:
Observações:

Princípio: Licenças livres
Tokens Encontrados:
Boas Práticas Encontradas:
Resultado:
Observações:

Princípio: Permanência
Tokens Encontrados:
Boas Práticas Encontradas:
Resultado:
Observações:

Princípio: Custos de Utilização
Tokens Encontrados:
Boas Práticas Encontradas:
Resultado:
Observações:

Fonte: Elaboração própria.

Conforme explicitado na Metodologia, foram avaliadas em média cinco URI por portal de publicação, sendo o resultado final a intersecção de todos os dados contidos em cada formulário. Os dados que compõem o resultado foram normalizados em uma tabela, da qual serão extraídas

informações a serem apresentadas nas discussões que se seguirão, avaliando a eficiência do processo.

Todos esses aspectos servem de aprendizado e base para o aperfeiçoamento do vocabulário e, consequentemente, para a validação da implementação computacional.

ETAPA 4 – ANÁLISE DOS RESULTADOS DA VALIDAÇÃO MANUAL DO VOCABULÁRIO DE TERMOS E REGRAS

Esta etapa compreende a análise dos resultados encontrados na validação manual. É importante concluir o grau de eficiência do analisador sintático e, consequentemente, do vocabulário, quanto à capacidade de filtrar tokens e regras em todo o conjunto de dados explícitos apresentados em cada URI.

Embora seja uma amostra pequena, o resultado encontrado serve como crítica ao modelo proposto, sendo uma excelente oportunidade para correção de erros e adaptação a ambientes heterogêneos.

Toda a expertise alcançada com o experimento tornou-se base para a etapa seguinte e expõe com clareza os principais desafios que a análise automatizada precisa solucionar.

ETAPA 5 – VALIDAÇÃO AUTOMATIZADA DO VOCABULÁRIO DE TERMOS E REGRAS

Nesta etapa, será proposta uma solução capaz de automatizar a análise sintática dos portais utilizando o vocabulário de termos. A solução pode compreender ferramentas de análise sintática existentes ou a implementação de uma aplicação específica.

Para validação da solução escolhida, será utilizada a mesma amostra de portais da etapa manual e o mesmo formato de re-

sultados. Dessa forma, as validações apresentadas poderão ser comparadas, a fim de ser possível propor melhorias tanto à metodologia quanto à solução automatizada.

O processo de análise de requisitos para escolha, adaptação e ou desenvolvimento da solução foge ao escopo dessa documentação e será discutido em uma outra oportunidade.

ETAPA 6 – AVALIAÇÃO COMPARATIVA DOS RESULTADOS

Cabe a esta etapa comparar os resultados encontrados na validação manual e automatizada e aperfeiçoar a metodologia proposta.

Essa avaliação responde, mesmo que num contexto minimizado, se o uso de um modelo baseado num vocabulário de termos baseado nas DWBP é capaz representar o cenário no qual os dados são disponibilizados nos diferentes portais governamentais brasileiros.

ETAPA 7 – AVALIAÇÃO DA ESCALABILIDADE DA SOLUÇÃO AUTOMATIZADA

Nesta etapa, a escalabilidade da solução automatizada será avaliada em um contexto no qual o número de URI processadas aumenta consideravelmente a cada ciclo de execução.

O objetivo é avaliar a performance do analisador e ampliar o conjunto de amostras para o universo ".gov.br". Os resultados obtidos até o momento desta publicação permitiram validar o modelo e a metodologia propostos. As principais conclusões e os desdobramentos dessa validação são apresentadas na próxima seção.

CONCLUSÃO E DESDOBRAMENTOS

Este artigo traz como contribuição principal um novo Modelo de Avaliação de Dados Abertos nos Portais Governamentais Brasileiros, consolidando um processo de análise que apresenta uma abordagem ágil e simplificada, capaz de identificar e avaliar os dados comumente apresentados nos diversos portais de publicação de dados abertos, tendo em vista os Ten Principles For Opening Up Government Information (SUNLIGHT FOUNDATION, 2010) e considerando as DWBP.

Uma das contribuições da análise dos portais de dados abertos é o mapeamento das diretrizes utilizadas pelos publicadores quando se trata do acesso à informação. Nesse sentido, ao mesmo tempo que a publicação de dados abertos facilita o acesso à informação e a tomada de decisões, a heterogeneidade de portais de acesso aos dados pode levar à diminuição da qualidade da informação. Geralmente, a forma como os dados abertos institucionais são disponibilizados levam em conta a estrutura de cada órgão público. Ministérios, governos estaduais e prefeituras apresentam os dados de forma diferente e com critérios de atualizações particulares.

Além disso, é complexo verificar critérios subjetivos tais como: a compreensão, relacionada ao entendimento sobre a estrutura e o significado dos dados; a confiança, que busca a melhoria da qualidade dos dados apresentados; e a conexão, que se destina ligação de dados de diferentes fontes, a partir apenas dos critérios de Classificação em 5 Estrelas (BERNERS-LEE, 2006) e dos Ten Principles For Opening Up Government Information (SUNLIGHT FOUNDATION, 2010).

Logo, a utilização de um modelo de avaliação capaz de sinalizar critérios que evidenciam o uso de boas práticas de publicação, vinculados aos 10 princípios de dados abertos (UFMG, no prelo) contempla 3 importantes benefícios: facilitação do acesso aos dados, transparência e o reuso.

O grande diferencial da contribuição das DWBP para a administração pública é a presença de um elemento de caráter normativo que complementa uma abordagem para implementação, isto é, a descrição de uma possível estratégia ou modelo computacional, como descrição do resultado pretendido. Este resultado serve como guia para a avaliação de que uma determinada boa prática foi seguida.

A presença de diretrizes objetivas aos publicadores promove o gerenciamento de dados mais consistente; portanto, tais orientações ajudam a promover a reutilização dos dados e a fortalecer a confiança nos dados entre os desenvolvedores, independentemente da tecnologia escolhida, além de possibilitar um ambiente de publicação e troca de dados mais homogêneo e previsível.

Assim, diferentes organizações podem especificar uma forma comum de comunicar informações entre comunidades, permitindo que dados de diferentes fontes possam ser relacionados e facilitando seu pareamento. Esse processo incentiva a implementação de ferramentas inovadoras e mais abrangentes, pois facilita a tomada de decisões e, consequentemente, a administração pública.

Dada a falta de padronização de estruturas informacionais no fornecimento de dados na Web, é bastante relevante o estudo da publicação de dados abertos governamentais para facilitar o acesso e o consumo de informação. Ao se observarem diferentes formas de estruturação de dados, é possível inferir quais são os portais que mais se aproximam do cumprimento de padrões e boas práticas de publicação de dados estabelecidas, a fim de nortear o compartilhamento dessas informações.

O estudo de publicação de dados exposto neste artigo, guiado pelo Modelo de Avaliação de Dados Abertos nos portais governamentais brasileiros, permitiu a criação de uma metodologia para identificar a presença das DWBP. Por meio dessa filtragem, avalia-se em que medida um portal segue os princípios

de governança esperados como índices de qualidade de publicação de dados abertos.

A avaliação dos portais, por sua vez, abre espaço para, por exemplo, a tomada de decisão por parte das organizações governamentais a fim de rever seus formatos de publicação de dados e se aproximar da padronização estabelecida. A padronização de informação facilitará o acesso a dados abertos para os usuários, o que é positivo para aprimorar a transparência e fomentar o consumo dos dados.

Dessa forma, partindo de um ambiente heterogêneo de compartilhamento de dados na Web, seguir uma metodologia permitiu estimar a qualidade de publicação de dados em cada portal avaliado e, então, mapear o uso de boas práticas em portais governamentais brasileiros. O repasse dessa avaliação para os portais pode ser útil para auxiliá-los na tomada de decisão acerca de futuros aprimoramentos na estruturação de dados a serem compartilhados com a população.

Este trabalho traz diferentes contribuições e possibilidades para criar e fomentar políticas públicas e governança metropolitana, cerne do Grupo de Trabalho (GT) do XI Congreso Internacional en Gobierno, Administración y Políticas Públicas GIGAPP 2022. Destacamos seus principais desdobramentos:

i) A diplomacia com dados passa pelo fato de ter transparência com dados de sites governamentais, permitindo que os gestores públicos e a sociedade tenham dados abertos da gestão pública disponíveis para conhecer, explorar e usá-los como insumos para processos de tomada de decisão.

ii) A governança pública demanda o acesso a dados e ferramentas de gestão pública (por exemplo, sistemas de informação e algoritmos computacionais) para estudar, avaliar, entender e propor políticas públicas baseadas em dados reais e conhecimentos adquiridos.

iii) A importância de gestores públicos conhecerem as boas práticas de outros gestores públicos, aprender com eles e adaptá-las para prover melhor governança metropolitana é primordial.

REFERÊNCIAS

BERNERS-LEE, T. *Design Issues*. Architectural and philosophical points. Wakefield: W3C, 2006. Disponível em: https://www.w3.org/DesignIssues/. Acesso em 5 out. 2023.

CONGRESO GIGAPP 2022, XI, Madrid, 21-23 set. 2022. *Anais* [...], Madrid: Universidad Compluense; GIGAPP, 2022. Disponível em: https://www.gigapp.org. Acesso em 13 nov. 2023.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. *Data on the Web Best Practices*. Wakefield: W3C, 31 jan. 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em 5 out. 2023.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. *DWBP Implementation Report*. Wakefield: W3C, 5 mar. 2019. Disponível em: https://w3c.github.io/dwbp/dwbp-implementation-report.html. Acesso em 5 out. 2023.

SUNLIGHT FOUNDATION. *Ten Principles For Opening Up Government Information*. Washington: SunLight Foundation, 11 ago. 2010. Disponível em: https://sunlightfoundation.com/wp-content/uploads/sites/2/2016/11/Ten-Principles-for-Opening-Up-Government-Data.pdf. Acesso em 5 out. 2023.

UNIVERSIDADE FEDERAL DE MINAS GERAIS (UFMG). *Os 10 princípios de Governança de Dados Abertos*. (no prelo).

ISBN: 978-65-85417-40-2