

Dados abertos CAPES: um olhar à luz dos desafios para publicação de dados na web

Emanuelle Torino

Doutoranda em Ciência da Informação, Universidade Estadual Paulista (Unesp), Marília, São Paulo, Brasil.
Bibliotecária, Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Paraná, Brasil.
<https://orcid.org/0000-0002-3791-9884>
E-mail: emanuelle@utfpr.edu.br

Gustavo Lunardelli Trevisan

Doutorando em Ciência da Informação, Universidade Estadual Paulista (Unesp), Marília, São Paulo, Brasil.
<https://orcid.org/0000-0002-4175-7910>
E-mail: g.trevisan@unesp.br

Silvana Aparecida Borsetti Gregorio Vidotti

Doutora em Educação, Universidade Estadual Paulista (Unesp), Marília, São Paulo, Brasil.
Docente do Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil.
<https://orcid.org/0000-0002-4216-0374>
E-mail: silvana.vidotti@unesp.br

Data de submissão: 12/09/2019. Data de aprovação no ConfOA: 12/06/2019. Data de publicação:

RESUMO

A Lei de Acesso à Informação instaurou, no Brasil, a necessidade da abertura dos dados governamentais, em benefício do acesso às informações públicas. Esses dados devem ser disponibilizados visando à descoberta, compreensão e reuso. Logo, o estudo objetivou discutir o atendimento de conjuntos de dados abertos governamentais aos desafios para a publicação de dados na web, apresentados pelo World Wide Web Consortium. A análise verificou a adequação de 29 conjuntos de dados acerca da avaliação da pós-graduação *stricto sensu* disponíveis no portal Dados Abertos CAPES aos 13 desafios para a disponibilização de dados na web, bem como os benefícios alcançados pelo atendimento, a fim de auxiliar os produtores na melhoria dos dados. A partir da análise foi possível verificar que, dos 13 desafios, apenas um foi considerado atendido, seis são parcialmente atendidos, quatro não atendidos e dois não se aplicam. Os dados disponíveis no portal Dados Abertos CAPES atualmente não atendem à primeira estrela dos dados abertos, contudo, possuem condições de atender à terceira estrela, indicando adequadamente a versão da licença Creative Commons adotada, visto que disponibilizam os dados em formato CSV. Além disso, há a necessidade de reestruturar os metadados para favorecer a compreensão e a processabilidade dos dados.

Palavras-chave: Dados abertos governamentais. Desafios para a disponibilização de dados na web. Lei de Acesso à Informação.

Dados Abertos CAPES: a look at the challenges of web data publishing

ABSTRACT

The Law on Access to Information established, in Brazil, the need to open government data, in order to access public information, such data should be made available for discovery, understanding and reuse. Thus, the present study aimed to discuss the attendance of government open datasets to the challenges for data publishing on the web, presented by the World Wide Web Consortium. An analysis has demonstrated the suitability of 29 datasets about the Stricto Sensu Graduate Evaluation available in the portal Dados Abertos CAPES to 13 challenges for data web availability, as well as the benefits achieved by the attendance, to assist producers in improving the data. From the analysis, it was possible to check that of the 13 challenges, only 1 was considered met, 6 partially met, 4 did not meet the parameters and 2 not applicable. The data available on the site Dados Abertos CAPES don't meet the requirements of the first open data star, however, they are able to meet the third star. They use records of the version of Creative Commons adopted, visualized that offers the data in the CSV format, in addition, there is a need to restructure metadata to facilitate understanding and processability of data.

Keywords: *Government open data. Challenges for providing web data. Access to Information Law.*

Dados Abertos CAPES: una mirada a los desafíos de la publicación de datos web

RESUMEN

La Ley de Acceso a la Información estableció, en Brasil, la necesidad de abrir los datos del gobierno, para acceder a la información pública, dichos datos deben estar disponibles para su descubrimiento, comprensión y reutilización. Por lo tanto, el presente estudio tuvo como objetivo discutir la asistencia de los conjuntos de datos abiertos del gobierno a los 13 desafíos para la publicación de datos en la web, presentados por el World Wide Web Consortium. El análisis presentó la adecuación de 29 conjuntos de datos sobre la Evaluación de Pos Grados de Stricto Sensu disponibles en el sitio Dados Abertos CAPES a 13 desafíos para la disponibilidad de datos web, así como los beneficios logrados, para ayudar a los productores a mejorar los datos disponibles en la web. Del análisis se encontró que de 13 desafíos, solo 1 se consideró cumplido, 6 parcialmente cumplidos, 4 insatisfechos y 2 no aplicables. Los datos disponibles en el sitio Dados Abertos CAPES actualmente no cumplen con la primera estrella de datos abiertos, sin embargo, pueden cumplir con la tercera estrella, lo que indica la versión de la licencia Creative Commons adoptada adecuadamente, ya que hacen que los datos estén disponibles en formato CSV, además, es necesario reestructurar los metadatos para facilitar la comprensión y la capacidad de procesamiento de los datos.

Palabras clave: *Datos abiertos del gobierno. Desafíos para disponibilización de datos web. Ley de Acceso a la Información.*

INTRODUÇÃO

A Lei de Acesso à Informação (LAI) instaurou, no Brasil, a necessidade da abertura dos dados governamentais, em benefício do acesso às informações públicas. Dada a obrigatoriedade, instituições subordinadas à LAI disponibilizam seus dados em sites da World Wide Web (web) para diferentes finalidades.

É válido destacar que a simples disponibilização de dados na web não os torna abertos, tampouco passíveis de uso por diferentes consumidores. É necessário identificar se esses dados podem ser facilmente detectáveis na web, se são compreensíveis e processáveis por humanos e agentes computacionais e, ainda, se estão atualizados.

A disponibilização de dados na web cumpre o que foi estabelecido pela LAI. Contudo, para que os dados sejam considerados dados abertos, eles precisam atender à características jurídicas que assegurem o uso, reuso e distribuição livres, mediante adoção de uma licença e técnicas que permitam que os dados possam ser reutilizados. Por isso, eles precisam atender às boas práticas para a produção e disponibilização de dados na web, preconizadas pelo World Wide Web Consortium (W3C) (LÓSCIO; BURLE; CALEGARI, 2017).

Essas boas práticas devem ser utilizadas para que os dados possam ser disponibilizados e consumidos na web a fim de superar os 13 desafios enfrentados no processo e que o atendimento exponha os benefícios apresentados por Lóscio, Burle e Calegari (2017): 1) compreensão: humanos compreenderão a estrutura de dados, seu significado, os metadados e a natureza do conjunto de dados; 2) processabilidade: aplicações computacionais poderão processar e manipular automaticamente dados contidos em um conjunto de dados; 3) máquinas de descoberta: serão capazes de encontrar automaticamente um conjunto de dados ou dados nele contidos; 4) reutilização: ampliação da possibilidade de reuso do conjunto de dados por diferentes usuários; 5) confiança: melhoria na confiabilidade dos consumidores de dados; 6) conexão: possibilidade

de criar links entre conjuntos de dados; 7) acesso: humanos e agentes computacionais serão capazes de acessar dados atualizados em uma variedade de formas; e 8) interoperabilidade: haverá consenso entre editores e consumidores de dados.

Para Lóscio, Guimarães e Calegari (2016), “[...] apesar de ser um assunto bastante discutido, várias questões importantes precisam ser abordadas a fim de satisfazer os requisitos de ambos publicadores e consumidores de dados na web.”

Assim, o presente estudo objetivou discutir o atendimento de conjuntos de dados abertos governamentais aos desafios para a publicação de dados na web, apresentados pelo W3C.

DADOS ABERTOS GOVERNAMENTAIS

No Brasil, a Lei de Acesso à Informação (LAI) “[...] dispõe sobre os procedimentos a serem observados pela União, Estados, Distrito Federal e Municípios, com o fim de garantir o acesso a informações” (BRASIL, 2011), estando a ela subordinada a instituições públicas dos poderes executivo, legislativo e judiciário, além de autarquias, fundações e empresas públicas, sociedades de economia mista e demais entidades controladas pela união, estados, municípios ou Distrito Federal; aplica-se ainda às instituições privadas e sem fins lucrativos que recebem recursos públicos para a realização de ações de interesse público. Além disso, objetiva assegurar o acesso à informação, tendo a publicidade como preceito e o sigilo como exceção.

A partir da LAI, tornou-se “[...] dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas.” (BRASIL, 2011) Para tanto, devem utilizar-se de todos os meios que dispuserem, sendo obrigatória a disponibilização em sites da web. O Art. 4º da LAI estabelece “I - informação: dados, processados ou não, que podem ser utilizados para produção e transmissão de conhecimento, contidos em qualquer meio, suporte ou formato” (BRASIL, 2011).

DESAFIOS PARA A DISPONIBILIZAÇÃO DE DADOS ABERTOS

A web se apresenta como um lugar aberto e flexível para disponibilização e compartilhamento de dados, de maneira que utilização deles está diretamente atrelada a variáveis, como a descoberta, a compreensão, a veracidade e a autenticidade, entre outros.

Assim, é necessário estabelecer um conjunto de requisitos padronizados para que fornecedores e consumidores de dados, que podem alternar seus papéis em determinados momentos, possam localizar conjuntos de dados, acessá-los, compreendê-los, considerá-los confiáveis, processá-los, torná-los interoperáveis e ainda conectá-los com outros dados ampliando o contexto, exercendo todas as possibilidades de reuso.

Evidentemente, nem todos os dados devem ser abertos. Antes disso, devem ser atendidos critérios legais, de anonimização, sensibilidade comercial, segurança e demais aspectos concernentes à privacidade dos dados pessoais ou sigilosos.

Portanto, considerando que fornecedores e consumidores de dados podem estar inclusive geograficamente dispersos, um dos objetivos das boas práticas preconizadas pelo W3C é dispor de informações que contribuam para a confiabilidade, a compreensão e o reuso dos dados, como: “[...] metadados estruturais, metadados descritivos, acesso informação, informação de qualidade de dados, informação de proveniência, informação de licença e informação de uso” (LÓSCIO; BURLE; CALEGARI, 2017, tradução nossa).

Nesse sentido, o W3C criou um grupo de trabalho para estabelecer boas práticas de dados na web e estruturar um ecossistema de dados abertos que beneficie a interoperabilidade entre fornecedores e consumidores de dados visando melhoria no processo de reuso, confiabilidade e ampliando o potencial de uso, independente da tecnologia adotada. O documento consiste em um conjunto de recomendações estáveis, que aprimoram a funcionalidade e a interoperabilidade na web (LÓSCIO; BURLE; CALEGARI, 2017).

Foram estabelecidas 35 boas práticas para a publicação de dados na web, estruturadas em 13 desafios, denominados: metadados, licenças de dados, proveniência de dados, qualidade de dados, versão de dados, identificadores de dados, formatos de dados, vocabulários de dados, acesso a dados, preservação de dados, feedback, enriquecimento de dados e republicação.

METODOLOGIA

O portal Dados Abertos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) é uma instituição governamental que, segundo a LAI, deve disponibilizar dados à sociedade. Foi definido como fonte para análise dos conjuntos de dados.

De acordo com o portal Dados Abertos CAPES (2019), no ambiente estão disponíveis “[...] dados e informações sobre a pós-graduação brasileira, sobre a formação de professores para educação básica e outros temas relacionados à educação”.

A coleta aconteceu no primeiro trimestre de 2019, quando estavam disponíveis 35 conjuntos de dados, organizados em: Avaliação da Pós-Graduação Stricto Sensu, com 29 conjuntos de dados; Acesso ao Portal de Periódicos, com 3 conjuntos de dados; Bolsas e Auxílios, com dois conjuntos de dados; e Orçamento e Finanças CAPES, com um conjunto de dados.

Considerando a expressividade, optou-se por selecionar os 29 conjuntos de dados do tema Avaliação da Pós-Graduação Stricto Sensu, que estão divididos em grupos: Projetos da Pós-Graduação; Detalhes da Produção Intelectual; Catálogo de Teses e Dissertações; Discentes da Pós-Graduação; Docentes da Pós-Graduação; Produção Intelectual da Pós-Graduação; Cursos da Pós-Graduação; Programas da Pós-Graduação; e Autor da Produção Intelectual.

Os dados coletados foram analisados considerando os 13 desafios para a disponibilização de dados na web preconizados pelo W3C (LÓSCIO; BURLE; CALEGARI, 2017) aos quais estão atreladas 35 boas práticas, que consistem em orientações direcionadas a fornecedores e consumidores de dados na web. Essas práticas visam auxiliá-los ao representar, descrever e disponibilizar dados que têm como objetivo a descoberta e a compreensão para que possam ser processáveis e reutilizados.

ANÁLISE E DISCUSSÃO DOS

RESULTADOS

A análise dos 29 conjuntos de dados do tema Avaliação da Pós-Graduação Stricto Sensu disponíveis no portal Dados Abertos CAPES permitiu identificar que, dos 13 desafios, apenas um foi totalmente atendido, seis são parcialmente atendidos, quatro não atendidos e dois não se aplicam. A discussão pretende sensibilizar os fornecedores de dados para que possam compreender os desafios e às boas práticas estabelecidas pelo W3C, de modo a auxiliá-los na melhoria dos dados disponibilizados na web.

O único desafio completamente atendido pelos conjuntos de dados analisados é a versão de dados. Lóscio, Burle e Calegari (2017) esclarecem que os conjuntos de dados podem ser alterados e ou atualizados por diferentes razões, fator preponderante para a manutenção de indicadores de versão que individualizam o conjunto de dados, tornando-os únicos. Para tanto, deve ser utilizado um padrão para a representação, disponível nos metadados, visando auxiliar o consumidor na coleta e contribuir para a qualidade do conjunto de dados. Além disso, os dados devem apresentar o histórico de versão, que consiste na descrição detalhada das alterações realizadas em cada versão. Os benefícios alcançados com o atendimento a esse desafio são reuso e veracidade. Os 29 conjuntos de dados analisados atendem completamente a esse desafio, uma vez que apresentam indicadores e fornecem históricos de versões.

O desafio metadados é considerado por Lóscio, Burle e Calegari (2017) como requisito fundamental, visto que a web é um espaço aberto de disponibilização de dados e informações e os metadados fornecem um contexto específico que auxilia a descoberta e utilização. Além disso, fornecem informações adicionais que auxiliam o consumidor, seja humano ou aplicação computacional, na compreensão do significado e da estrutura dos dados. Ao desafio estão atreladas três boas práticas, o fornecimento de metadados e a presença de metadados descritivos e estruturais. Vale destacar que as características fundamentais dos metadados são a estrutura e a representação, que fornecem informações relevantes para que os consumidores possam compreender os dados, considerando os tipos de metadados: administrativos, descritivos, de preservação, técnicos, de proveniência e de uso (GILLILAND, 2008), além dos estruturais (RILEY, 2017). Os benefícios da adoção são reuso, compreensão, descoberta e processabilidade.

Na área analisada, os 29 conjuntos de dados apresentam um arquivo em formato Portable Document Format (PDF), cuja nomenclatura está iniciada pela palavra “metadados” seguida do nome do conjunto de dados. O documento descreve os metadados estruturais utilizados no conjunto de dados, as variáveis codificadas, a descrição e os tipos de dados contidos no conjunto. Embora o documento norteie o consumidor de dados, não é adotado nenhum padrão de metadados com semântica formal, o que exige interpretação e tratamento para processar os dados, dificultando o reuso.

Além disso, na interface do portal estão disponíveis alguns metadados administrativos, como data de criação, formato e licença. Contudo, não há metadados descritivos imprescindíveis à compreensão por humanos e agentes computacionais, como: título, palavras-chave, descrição e instituição responsável pelos dados. Assim os conjuntos de dados atendem parcialmente ao desafio metadados.

A licença de dados é que determina quais usos podem ser realizados sem infração dos direitos. Por isso, o desafio requer a especificação da licença como metadados e/ou em documento vinculado, visando assegurar reuso e veracidade (LÓSCIO; BURLE; CALEGARI, 2017). Nos conjuntos de dados analisados está disponível uma licença Creative Commons do tipo Atribuição (CC BY), considerada a licença mais aberta, indicada para ampliar a disseminação e o uso do objeto licenciado. Contudo, não há indicação da versão adotada, o que impede a interpretação do texto legal que respalda a permissão de acesso e uso. Nesse caso, o desafio foi considerado parcialmente atendido.

O desafio dos identificadores de dados possui três boas práticas relacionadas, que asseguram os benefícios de reuso, conectividade, descoberta e interoperabilidade. Delas, apenas o uso de Uniform Resource Identifier (URIs), persistentes como identificadores de conjunto de dados, foi atendida, o que consiste no uso de identificador individual para cada conjunto de dados (LÓSCIO; BURLE; CALEGARI, 2017). Vale destacar que o identificador deve assegurar que os dados permanecerão acessíveis aos consumidores utilizando a mesma URI de modo permanente, o que pode ser realizado pela instituição fornecedora dos dados ou gerenciado por um serviço externo, como o Handle System ou Digital Object Identifier (DOI). Assim, o desafio foi considerado parcialmente atendido, visto que não disponibilizam URIs como identificadores nos conjuntos de dados e não são atribuídas URIs a versões e séries de conjuntos de dados.

Os formatos de dados são essenciais para que os dados consumidos possam ser processados e reutilizados, sobretudo por aplicações computacionais, visando reuso, compreensão e processabilidade (LÓSCIO; BURLE; CALEGARI, 2017). As três boas práticas relacionadas ao desafio reforçam a necessidade de uso de formatos de dados padronizados e legíveis por máquinas utilizando sintaxes, como CSV, XML, HDF5, JSON, RDF, RDF/XML, JSON-LD e Turtle; fornecimento de dados em mais de um formato, considerando o uso potencial;

e representação de dados neutra, independente de idiomas ou contendo especificações para a interpretação como metadados. Os dados analisados no portal Dados Abertos da CAPES são disponibilizados em diferentes formatos: CSV, HTML, PDF, XLSX, XLS, e atendem parcialmente às boas práticas do W3C, pois disponibilizam apenas o formato um formato recomendado, o CSV.

Já no desafio de acesso a dados, que contém 10 boas práticas — relevantes à medida que representam a maneira que os consumidores utilizarão para acessar os dados — apresenta os seguintes benefícios alcançados pelo atendimento: reuso, conectividade, acesso, veracidade, interoperabilidade, descoberta e processabilidade. Os 29 conjuntos de dados analisados atendem parcialmente ao desafio, por meio de quatro boas práticas. São elas: fornecer subconjuntos para conjuntos de dados grandes; disponibilizar dados por meio de uma Application Programming Interface (API); usar padrões da web como base de APIs; e fornecer documentação completa para sua API. Nesse sentido, o fornecimento de subconjuntos é imprescindível, segundo Lóscio, Burle e Calegari (2017, tradução nossa), porque “os dados que levam mais de dez segundos para serem entregues provavelmente farão com que os usuários suspeitem de falha”. Os conjuntos de dados analisados estão organizados por nomes e, em cada hiperlink, há conjuntos de dados disponíveis individualmente, atendendo à boa prática estabelecida. No que tange ao uso de APIs, o Portal Dados Abertos CAPES utiliza o software CKAN, que disponibiliza API baseada em JSON, o qual tem sua utilização documentada para administradores¹ e usuários², atendendo assim às boas práticas preconizadas pelo W3C. Destaca-se que a API está atrelada ao CKAN e, se o portal Dados Abertos CAPES alterar a ferramenta para a disponibilização dos dados, esse desafio necessitará de nova análise.

¹ Disponível em: <https://docs.ckan.org/en/latest/maintaining/datastore.html#the-datastore-api>. Acesso em: 11 abr. 2019.

² Disponível em: <https://docs.ckan.org/en/latest/api/index.html>. Acesso em: 11 abr. 2019.

O último desafio parcialmente atendido foi o enriquecimento de dados, que consiste na adoção de processos e técnicas para melhora de dados. Nele, uma das boas práticas foi atendida, a que consiste em oferecer apresentações complementares, que pretendem beneficiar consumidores humanos ao fornecer alternativas de visualização imediata dos dados, gerando benefícios de reuso, compreensão, acesso e veracidade. Na área analisada é possível visualizar os dados em tabelas antes de baixá-los.

O desafio de proveniência dos dados consiste em fornecer informações, nos metadados, acerca da origem dos dados, a fim de que após o reuso seja possível identificar o fornecedor e consumidor dos dados, auxiliando ainda na avaliação da qualidade e confiabilidade. Os benefícios da adoção são o reuso, a compreensão e a veracidade (LÓSCIO; BURLE; CALEGARI, 2017). Na área avaliada do portal Dados Abertos CAPES, no documento PDF relativo a metadados, há informações textuais na seção “Metodologia de carga”, que permitem identificar, por humanos, que os dados foram coletados a partir da Plataforma Sucupira³. Contudo, não há metadados de proveniência, o que impede a compreensão dessa boa prática por aplicações computacionais.

O desafio relacionado à qualidade de dados requer que sejam indicados, nos metadados, informações de qualidade dos dados e sua adequação para finalidades específicas, visando assegurar reuso e veracidade. Nesse sentido, há vocabulários de qualidade de dados, como o de Albertoni e Isaac (2016) e as dimensões de qualidade de dados ISO/IEC 25012⁴, capazes de auxiliar os editores de dados nesta representação. Os dados disponíveis no portal Dados Abertos CAPES não fornecem informações acerca da qualidade dos dados.

Para Lóscio, Burle e Calegari (2017), os vocabulários de dados definem os conceitos e relacionamentos (atributos) utilizados para a descrição e representação de determinada área de interesse, utilizados para definir relacionamentos e restrições ao uso desses termos em determinada aplicação, no intuito de auxiliar na integração de dados e organização do conhecimento. Para tanto, o W3C recomenda a adoção de vocabulários de dados padronizados, como: VOCAB-DCAT, Dublin Core, FOAF, SKOS e vCard. Esse desafio conta com duas boas práticas — o reuso de vocabulários padronizados e a adoção do nível de semântica formal correta — e seu atendimento gera como benefícios reuso, processabilidade, compreensão, veracidade e interoperabilidade. Nos 29 conjuntos de dados analisados neste estudo, apesar da presença de um vocabulário de dados que contempla metadados administrativos, descritivos e de versão, não são utilizados vocabulários padronizados de dados e metadados. Além disso, não há semântica formal, o que faz com que o desafio preconizado pelo W3C não seja atendido.

A disponibilização de dados na web permite que eles sejam acessados e consumidos em grande escala. Assim, Lóscio, Burle e Calegari (2017) indicam, no desafio feedback, que o retorno dos consumidores quanto ao atendimento às suas necessidades é imprescindível para atingir aos benefícios de reuso, compreensão e veracidade. Nesse desafio recomendam-se duas práticas: coletar feedback e disponibilizá-los publicamente. O portal Dados Abertos CAPES, analisado neste estudo, não disponibiliza formas de feedback pela interface e não é possível avaliar se existe coleta automática de feedback por aplicações computacionais. Logo, não atende a este desafio.

³ Disponível em: <https://sucupira.capes.gov.br/sucupira/>. Acesso em: 01 abr. 2019.

⁴ Disponível em: <http://iso25000.com/index.php/en/iso-25000-standards/iso-25012>. Acesso em: 04 abr. 2019.

A disponibilização e exclusão de dados na web podem ser facilmente realizadas, por isso Lóscio, Burle e Calegari (2017) recomendam duas boas práticas relacionadas ao desafio da preservação de dados: a utilização de identificadores e o fornecimento de informações sobre os dados, caso sejam excluídos ou disponibilizados mediante solicitação; avaliar a cobertura dos conjuntos de dados antes da preservação, considerando a dependência do conjunto de dados a um contexto, o que leva à necessidade da preservação dos dados, contexto e vocabulário de dados utilizados. O atendimento a essa dimensão gera benefícios de reuso e veracidade. O estudo teve como limitação a indisponibilidade de acesso às políticas do portal analisado, sobretudo à política de preservação dos dados. Destaca-se que, por se tratar de um portal implantado recentemente, ainda não há dados retirados. Assim, considera-se que o desafio não se aplica à análise.

Por sua vez, o último desafio apresentado por Lóscio, Burle e Calegari (2017) é a republicação de dados, na qual o reuso se faz pela disponibilização dos dados fornecidos em outra fonte. Para tanto, estabelecem-se três boas práticas: informar ao fornecedor dos dados da reutilização e enviar feedbacks; republicar os dados atentando-se aos termos estabelecidos na licença; e citar a origem dos dados (proveniência). Essas práticas asseguram os benefícios de reuso, descoberta, interoperabilidade e veracidade, de modo que tal dimensão não pode ser avaliada nos 29 conjuntos de dados analisados, pois os dados disponíveis no portal Dados Abertos CAPES são coletados por meio da Plataforma Sucupira, contudo, são disponibilizados apenas no portal Dados Abertos CAPES, o que não consiste em republicação. Portanto, considera-se que o desafio não se aplica aos dados analisados.

CONCLUSÕES

O presente estudo objetivou discutir o atendimento de 29 conjuntos de dados abertos governamentais acerca da Avaliação da Pós-Graduação Stricto Sensu disponível no portal Dados Abertos CAPES aos 13 desafios para a publicação de dados na web recomendados pelo W3C, bem como os benefícios alcançados pelo atendimento, de modo a auxiliar os produtores na melhoria dos dados disponibilizados na web.

Uma fragilidade é a ausência ou tratamento inadequado dos metadados. Embora apenas duas dimensões (metadados e vocabulário de dados) explicitem a necessidade da adoção e fornecimento adequado de estrutura de dados, semântica formal e fornecimento de metadados, outras dimensões, de igual maneira, necessitam de estruturas de metadados para que as boas práticas possam ser atingidas, tais como: licenças de dados, proveniência de dados, qualidade de dados, versão de dados, identificadores de dados, formatos de dados, acesso a dados, preservação de dados. Enfim, no que tange ao próprio dado e ao seu contexto, todos os dados estarão estruturados e representados para que possam ser facilmente compreendidos pelos consumidores. Para tanto, devem ser utilizados, sempre que necessário, metadados administrativos, estruturais, técnicos, de preservação, de uso, estruturais e de proveniência.

De igual maneira, é necessário estabelecer adequadamente a licença dos dados, não apenas indicando o tipo de licença adotada, mas especificando a versão da licença e utilizando metadados para expressá-la de modo a ser adequadamente compreendida por humanos, por aplicações computacionais e pelo texto jurídico, o que assegura a veracidade e reuso sem qualquer infração à questões legais, mantendo protegidos os fornecedores e consumidores dos dados.

Como aspectos positivos, destacam-se a relevância da manutenção dos indicadores e históricos de versões, que possibilitam aos consumidores a adequada definição do dado a ser utilizado, fator atendido pelos conjuntos de dados analisados, que atingem reuso e veracidade.

Cumpra-se ressaltar que a simples disponibilização de dados na web não assegura que eles possam ser consumidos, em atendimento ao que preconiza a LAI, sobretudo pela ausência de metadados descritivos, estruturais e de proveniência, capazes de assegurar ao consumidor confiabilidade no acesso e processamento dessas informações.

Ressalta-se que os resultados apresentados estão em conformidade com o relatório de implementação das boas práticas para a publicação de dados na web (LÓSCIO; BURLE; CALEGARI, 2019), segundo o qual os desafios consistem em metadados e vocabulários de dados, licenças e formatos de dados, proveniência, versionamento, feedback e preservação de dados.

Por fim, recorreremos ao plano de desenvolvimento para as cinco estrelas dos dados abertos, de Berners-Lee (2012), que consistem em: 1) disponibilizar os recursos na web, em qualquer formato, sob uma licença aberta; 2) disponibilizar os recursos na web em um formato estruturado, ainda que proprietário (ex.: XLS), sob uma licença aberta; 3) disponibilizar os recursos na web, em um formato estruturado não proprietário (ex. CSV), sob uma licença aberta; 4) disponibilizar os recursos na web, em um formato estruturado não proprietário, utilizando Resource Description Framework (RDF) sob uma licença aberta e utilize URIs para identificá-los; e 5) cumprir os dados abertos que têm quatro estrelas e conectar seus dados com outros dados (linked data).

Com isso, podemos afirmar que os dados disponíveis no portal Dados Abertos CAPES atualmente não atendem sequer à primeira estrela dos dados abertos de Berners-Lee. Entretanto, possuem condições de atender à terceira estrela, realizando uma alteração simples na disponibilização dos seus dados: estabelecer a versão da licença Creative Commons Atribuição (CC BY) adotada para os conjuntos de dados, com possibilidade de expandir para a quarta estrela caso passem a disponibilizar dos dados em RDF, utilizando URIs persistentes. Vale destacar que a compreensão e a processabilidade da informação estão diretamente dependentes da reestruturação dos metadados, conforme mencionado anteriormente.

REFERÊNCIAS

- ALBERTONI, R.; ISAAC, A. (ed.). *Dados sobre as melhores práticas da web: vocabulário de qualidade de dados*. [S.l.], 2016. Disponível em: <https://www.w3.org/TR/vocab-dqv/>. Acesso em: 1 abr. 2019.
- BERNERS-LEE, T. *5 [stars] open data*. [S.l.], 2012. Disponível em: <http://5stardata.info/en/>. Acesso em: 13 abr. 2019.
- BRASIL. Lei nº 12.527, de 18 de novembro de 2011. *Diário Oficial da União*, Brasília, DF, 18 nov. 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 1 maio 2018.
- CAPES. *Dados Abertos Capes*. [S.l.], 2017. Disponível em: <https://dadosabertos.capes.gov.br>. Acesso em: 31 mar. 2019.
- GILLILAND, A. J. Setting the stage. In: BACA, M. (ed.). *Introduction to metadata*. Los Angeles: Getty, c2008. p. 1-19. Disponível em: <http://d2aohiyo3d3idm.cloudfront.net/publications/virtuallibrary/0892368969.pdf>. Acesso em: 15 jul. 2018.
- LÓSCIO, B. F.; GUIMARÃES, C. B. dos S.; CALEGARI, N. Boas práticas para dados na web: desafios e benefícios. *Revista Principia*, João Pessoa, n. 32, p. 9-18, dez. 2016. Disponível em: <http://periodicos.ifpb.edu.br/index.php/principia/article/view/1023/>. Acesso em: 13 abr. 2019.
- LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. (ed.). *Data on the web best practices*. [S.l.], 2017. Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 26 abr. 2018.
- RILEY, J. *Understanding metadata: what is metadata, and what is it for?*. Baltimore, MD: National Information Standards Organization (NISO), c2017. Disponível em: https://groups.niso.org/apps/group_public/download.php/17446. Acesso em: 15 jul. 2018.