BOAS PRÁTICAS PARA DADOS NA WEB:
análise do portal Dados Abertos Capes

DATA ON THE WEB BEST PRACTICES: portal Dados Abertos Capes analysis

Emanuelle Torino[1]
Silvana Aparecida Borsetti Gregorio Vidotti[2]

RESUMO
O presente estudo objetivou discutir o atendimento de conjuntos de dados abertos governamentais acerca da Avaliação da Pós-Graduação Stricto Sensu disponíveis no portal Dados Abertos Capes às boas práticas para de dados na web. A análise apresentou a adequação dos 29 conjuntos de dados às 35 boas práticas para a disponibilização de dados na web, recomendadas pelo *World Wide Web Consortium*, bem como os benefícios alcançados pelo atendimento, de forma a fornecer aportes teóricos da Ciência da Informação para subsidiar os fornecedores de dados, no tocante à melhoria dos dados disponibilizados na web. A partir da análise foi possível verificar que, das 35 boas práticas, 7 não se aplicam aos conjuntos de dados analisados; 20 foram consideradas não atendidas ou parcialmente atendidas e 8 consideradas atendidas. Os dados disponíveis no portal Dados Abertos Capes necessitam de ajustes para que possam atender à terceira estrela dos dados abertos.
Palavras-chave: Dados abertos governamentais. Boas práticas para dados na web. Lei de Acesso à Informação.

ABSTRACT
The present study aimed to discuss the attendance of government open datasets related to the *Stricto Sensu* Graduate Evaluation available in the portal Dados Abertos Capes to data on the web best practices. The analysis presented the adequacy of the 29 data sets to the 35 data on the web best practices, recommended by the World Wide Web Consortium, as well as the benefits achieved by the attendance, in order to provide theoretical contributions from Information Science to support the data providers, with regard to improving the data made available on the web. From the analysis it was possible to verify, of the 35 best practices, 7 do not apply to the analyzed datasets; 20 were considered did not met or partially met and 8 considered met. The data available on the portal Dados Abertos Capes need adjustments to be able to meet the third star of open data.
Keywords: Government open data. Data on the Web Best Practices. Access to Information Law.

Artigo submetido em 22/02/2020 e aceito para publicação em 18/02/2021

1 Doutoranda no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil. ORCID https://orcid.org/0000-0002-3791-9884. E-mail: etorino@gmail.com
2 Docente permanente no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil. ORCID https://orcid.org/0000-0002-4216-0374. E-mail: silvana.vidotti@unesp.br

1 INTRODUÇÃO

A transparência nos atos governamentais tem sido uma discussão social necessária e uma das formas da sua materialização se dá por meio da abertura de dados, quer seja por iniciativa do órgão público que o gera ou pela possibilidade de solicitação de acesso à informação disponível a qualquer cidadão.

Nessa discussão, dois instrumentos jurídicos brasileiros devem ser destacados, a Lei de Acesso à Informação (LAI) que regula o acesso a informações, conforme previsto no art. 5, inciso XXXIII da Constituição Federal "todos têm direito a receber dos órgãos públicos informações de seu interesse particular, ou de interesse coletivo ou geral, que serão prestadas no prazo da lei, sob pena de responsabilidade, ressalvadas aquelas cujo sigilo seja imprescindível à segurança da sociedade e do Estado" (BRASIL, 1988), e, o Decreto nº 8.777, que institui a Política de Dados Abertos do Poder Executivo Federal (BRASIL, 2016).

Os referidos dispositivos legais regulamentam a abertura dos dados governamentais brasileiros, tornando-a obrigatória para instituições subordinadas à LAI, conforme preconizado na Constituição Federal e definem que tais dados devem ser disponibilizados na web para que sejam acessíveis e processáveis por humanos e agentes computacionais.

A World Wide Web (web), por sua vez, é espaço profícuo para a disponibilização de dados. Contudo, para que eles sejam recuperados, acessíveis, compreensíveis e processáveis por humanos e agentes computacionais é necessário que haja uma estrutura mínima, o que se caracteriza em um desafio.

Para auxiliar fornecedores e consumidores de dados de forma a superar ao desafio apontado, o *World Wide Web Consortium* (W3C), por meio de um grupo de especialistas, disponibiliza um documento no qual estabelece um conjunto de boas práticas a serem seguidas no estruturação da oferta e do consumo de dados.

Nesse sentido, o presente estudo objetiva discutir o atendimento de conjuntos de dados abertos governamentais acerca da Avaliação da Pós-Graduação *Stricto Sensu* disponíveis no portal Dados Abertos Capes às boas práticas para dados na web.

O portal Dados Abertos (2019), disponibiliza "[...] dados e informações sobre a pós-graduação brasileira, sobre a formação de professores para educação básica e outros temas relacionados à educação".

2 DADOS ABERTOS GOVERNAMENTAIS

No Brasil foi sancionada a Lei nº 12.527 (BRASIL, 2011), conhecida como Lei de Acesso à Informação (LAI), com base no disposto no art. 37, parágrafo 3, inciso II da Constituição Federal, "A lei disciplinará as formas de participação do usuário na administração pública direta e indireta, regulando especialmente: [...] o acesso dos usuários a registros administrativos e a informações sobre atos de governo, observado o disposto no art. 5º, [...] XXXIII", e no art. 216, parágrafo 2 "Cabem à administração pública, na forma da lei, a gestão da documentação governamental e as providências para franquear sua consulta a quantos dela necessitem." (BRASIL, 1988).

A 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), tendo o acesso à informação como direito e o sigilo como exceção. Estão subordinadas à LAI: instituições públicas dos poderes executivo, legislativo e judiciário; autarquias, fundações e empresas públicas, sociedades de economia mista, e demais entidades controladas pela união, estados, municípios ou distrito federal; instituições privadas e sem fins lucrativos que recebem recursos públicos para a realização de ações de interesse público.

Fica estabelecido no art. 8 da LAI, "É 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." Para tanto, tais instituições devem utilizar-se de todos os meios e instrumentos disponíveis, sendo obrigatória a disponibilização de informações em sites oficiais da web.

Vale destacar que, no art. 4º da LAI considera-se "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).

Adicionalmente, a Política de Dados Abertos do Poder Executivo Federal (BRASIL, 2016) define no art. 2, inciso III: "dados abertos - dados acessíveis ao público, representados em meio digital, estruturados em formato aberto, processáveis por máquina, referenciados na internet e disponibilizados sob licença aberta que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte". E, ainda, estabelece seus objetivos no art. 1, dentre os quais, destacam-se o inciso I "promover a publicação de dados contidos em bases de dados de órgãos e entidades da administração pública federal direta, autárquica e fundacional sob a forma de dados abertos" e IV "facilitar o

intercâmbio de dados entre órgãos e entidades da administração pública federal e as diferentes esferas da federação", que são focos desse estudo.

No que tange à abertura de dados, é importante esclarecer que "Dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras." (OPEN KNOWLEDGE FOUNDATION, 2020), consonante com a Open Definition (THE OPEN DEFINITION, 2020).

Em 2007, foram estabelecidos oito princípios que os dados governamentais devem atender para serem considerados abertos: completos, primários, oportunos, acessíveis processáveis por máquina, não discriminatórios, não proprietários e livres de licença, sendo a eles adicionados posteriormente outros sete: online e gratuitos, permanentes, verdadeiros, com presunção de abertura, documentados, seguros, projetados para a necessidade pública (THE ANNOTATED 8 PRINCIPLES OF OPEN GOVERNMENT DATA, 2020). Dentre os princípios há a indicação de que os dados sejam livres de licença, o que contradiz a definição apresentada pela *Open Knowledge Foundation*.

A esse respeito, cumpre lembrar que, considerando a legislação de direitos autorais brasileira, os dados não são protegidos, uma vez que não se tratam de criações de um autor, contudo, as bases de dados estão protegidas, conforme art. 7º, inciso XIII e suas formas legais de uso são determinadas no capítulo VII (BRASIL, 1998), tornando necessária a autorização para uso, o que pode ser feito pelo licenciamento, o que reforça a definição da *Open Knowledge Foundation*.

Assim, todas as definições e dispositivos legais apresentados elucidam que os dados abertos governamentais devem estar publicamente acessíveis na web, ainda que não haja solicitação expressa, atendendo à necessidade de transparência e, que sejam passíveis de uso e processamento por humanos e agentes computacionais. Para tanto, devem atender às boas práticas para a disponibilização de dados na web.

3 BOAS PRÁTICAS PARA A DISPONIBILIZAÇÃO DE DADOS ABERTOS NA WEB

A web se apresenta como espaço adequado para a disponibilização de dados, favorecida pela grande oferta de softwares/plataformas e até mesmo pelo fenômeno do *big data*. Adicionalmente, no tocante aos dados governamentais, a disponibilização é requerida por um conjunto de dispositivos legais e deve ser realizada conforme estabelecido pela LAI.

Contudo a disponibilização de dados na web necessita de estrutura e consistência para que possibilite a descoberta de recursos, as conexões e os relacionamentos entre eles. Nesse sentido, é

imprescindível que fornecedores de dados os adequem para que, após a disponibilização, possam ser recuperados, compreendidos e utilizados pelos consumidores, gerando ganhos múltiplos. Nesse sentido, o W3C fornece informações adequadas para que fornecedores e consumidores de dados possam estruturar a disponibilização e a coleta de dados na web, e superar os desafios dessa ação.

Desse modo, o World Wide Web Consortium (W3C) por meio de um Grupo de Trabalho definiu em um documento um conjunto de práticas para a disponibilização de dados na web, visando estruturar um ecossistema de dados abertos que possibilite a recuperação e a compreensão por humanos e agentes computacionais, possibilitando a interoperabilidade entre consumidores e fornecedores de dados, o que beneficia a confiabilidade, o uso e reuso. Por ter sido submetido à análise de especialistas e considerado um documento com recomendações estáveis, pode ser utilizado como referência (LÓSCIO; BURLE; CALEGARI, 2017).

O referido documento considerou os 13 principais desafios relacionados aos dados na web: 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, que precisam ser considerados por fornecedores e consumidores de dados visando disponibilizá-los de forma compreensível e recuperável, e, por isso, estabelecem as 35 boas práticas para a disponibilização de dados na web que nortearam a análise.

Segundo Lóscio, Burle e Calegari (2017) a adoção das boas práticas gera alguns benefícios:

a. compreensão: a estrutura de dados, seu significado, os metadados e a natureza do conjunto de dados serão compreensíveis a humanos;
b. processabilidade: dados contidos em um conjunto de dados serão processáveis e manipuláveis por agentes computacionais;
c. facilidade de descoberta: agentes computacionais serão capazes de encontrar automaticamente um conjunto de dados ou dados nele contidos;
d. reutilização: diferentes consumidores poderão reutilizar o conjunto de dados;
e. confiança: ampliação da confiabilidade dos consumidores de dados;
f. conexão: possibilidade de criar conexões entre conjuntos de dados;
g. acesso: dados atualizados poderão ser acessados de diferentes maneiras por humanos e agentes computacionais; e
h. interoperabilidade: fornecedores e consumidores de dados terão consenso na oferta e no consumo.

Há crescente interesse e discussão acerca da oferta e do consumo de dados na web, 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."

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Para a consecução do objetivo de discutir o atendimento de conjuntos de dados abertos governamentais disponíveis no portal Dados Abertos Capes às boas práticas para a disponibilização de dados na web, e considerando a representatividade foram selecionados os dados do tema Avaliação da Pós-Graduação *Stricto Sensu*[3], 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; Autor da Produção Intelectual, totalizando 29 conjuntos de dados.

A coleta de dados para o presente estudo aconteceu no segundo semestre de 2019, quando estavam disponíveis no portal Dados Abertos Capes[4] 35 conjuntos de dados, organizados por tema: 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 2 conjuntos de dados; e, Orçamento e Finanças CAPES, com 1 conjunto de dados.

A análise apresenta a adequação dos 29 conjuntos de dados às 35 boas práticas para a disponibilização de dados na web recomendadas pelo W3C, bem como os benefícios alcançados pelo atendimento, de forma a fornecer aportes teóricos da Ciência da Informação para subsidiar os fornecedores de dados, no tocante à melhoria dos dados disponibilizados na web.

O Quadro 1 apresenta uma análise da adequação desses conjuntos de dados às boas práticas para a disponibilização de dados na web preconizadas pelo W3C. Destaca-se que, a adequação aos 13 desafios e às 35 boas práticas assegura que o conjunto de dados possa ser compreensível e facilmente processável por humanos e agentes computacionais.

3 Disponível em: https://dadosabertos.capes.gov.br/organization/diretoria-de-avaliacao. Acesso em: 1 out. 2019.
4 Disponível em: https://dadosabertos.capes.gov.br. Acesso em: 1 out. 2019.

Quadro 1 – Adequação do conjunto de dados Avaliação da Pós-Graduação *Stricto Sensu* às boas práticas para a publicação de dados na web

Desafio	Boas Práticas		Atendimento ao desafio
Metadados	1	Fornecer metadados	Atende parcialmente
	2	Fornecer metadados descritivos	Não atende
	3	Fornecer metadados estruturais	Atende parcialmente
Licenças de Dados	4	Fornecer informações de licença de dados	Atende parcialmente
Proveniência de Dados	5	Fornecer informações de proveniência de dados	Não atende
Qualidade de Dados	6	Fornecer informações de qualidade de dados	Não atende
Versionamento de Dados	7	Fornecer indicador de versão	Atende
	8	Fornecer histórico de versões	Atende
Identificadores de Dados	9	Usar URIs persistentes como identificadores de conjunto de dados	Atende
	10	Usar URIs persistentes como identificadores dentro de conjuntos de dados	Não atende
	11	Atribuir URIs a versões dos conjuntos de dados e séries	Atende parcialmente
Formatos de Dados	12	Usar formatos de dados padronizados legíveis por máquina	Atende parcialmente
	13	Usar representações de dados com localidade neutra	Atende parcialmente
	14	Fornecer dados em formatos múltiplos	Atende parcialmente
Vocabulários de Dados	15	Reutilizar vocabulários, preferencialmente padronizados	Não atende
	16	Escolher o nível de formalização correto	Não atende
Acesso a dados	17	Fornecer o *download* em massa	Atende parcialmente
	18	Fornecer subconjuntos para conjuntos de dados extensos	Atende
	19	Usar a negociação de conteúdo para disponibilizar dados em formatos múltiplos	Não atende
	20	Fornecer acesso em tempo real	Não atende
	21	Fornecer dados atualizados	Não atende
	22	Fornecer uma explicação para os dados que não estão disponíveis	Não se aplica
	23	Disponibilizar dados por meio de uma *Application Programming Interface* (API)	Atende
	24	Usar padrões web como base para a construção de APIs	Atende
	25	Fornecer documentação completa para a API	Atende
	26	Evitar alterações que afetem o funcionamento da API	Não se aplica
Preservação de Dados	27	Preservar identificadores	Não se aplica
	28	Avaliar a cobertura do conjunto de dados	Não se aplica
Feedback	29	Coletar *feedback* dos consumidores de dados	Não atende
	30	Disponibilizar *feedback*	Não atende
Enriquecimento de Dados	31	Enriquecer dados gerando novos dados	Não atende
	32	Fornecer visualizações complementares	Atende
Republicação de Dados	33	Fornecer *feedback* ao fornecedor original dos dados	Não se aplica
	34	Seguir os termos de licença	Não se aplica
	35	Citar a publicação original do conjunto de dados	Não se aplica

Fonte: Autoria própria (2019).

O primeiro desafio, relativo aos **metadados**, é destacado por Lóscio, Burle e Calegari (2017) como requisito fundamental para especificar o contexto dos dados em um sistema de informação, sem eles, as possibilidades de descoberta e reuso são minimizadas e muitas vezes restringidas ao fornecedor dos dados. Os metadados possuem como característica a estrutura e a representação, capazes de fornecer informações pertinentes aos consumidores (humanos e agentes computacionais) para que possam compreender os dados, considerando os tipos de metadados: administrativos, descritivos, de preservação, técnicos, de proveniência e de uso (GILLILAND, 2008); para este estudo, deve-se considerar ainda os estruturais (RILEY, 2017).

Esse desafio está dividido em três boas práticas, a primeira consiste em fornecer metadados compreensíveis a usuários humanos e agentes computacionais, para que possam ser compreensíveis e processáveis, quer seja apresentando-os como parte de uma página HTML ou em um arquivo adicional. Alcançando os benefícios de reuso, compreensão, descoberta e processabilidade. A segunda boa prática recomenda que sejam fornecidos metadados descritivos acerca do conjunto de dados possibilitando que consumidores de dados, quando agentes computacionais, tenham acesso à uma estrutura padronizada e semanticamente formal, que privilegia a descoberta, a interoperabilidade e o desenvolvimento de aplicações para o consumo automático dos dados; e, quando humanos, acessem informações que representem os dados, sua natureza e seu contexto. Os benefícios da adoção de metadados descritivos são reuso, compreensão e descoberta. E, a terceira boa prática recomenda o fornecimento de metadados estruturais para que as propriedades do conjunto de dados sejam legíveis e processáveis por humanos e agentes computacionais. Quando legíveis por agentes computacionais, os metadados estruturais podem ser fornecidos incorporados ao documento ou em documentos separados. Os benefícios desta boa prática são reuso, compreensão e processabilidade.

No que tange à área Avaliação da Pós-Graduação Stricto Sensu disponível no portal Dados Abertos Capes, avaliada no presente estudo, verificamos que, ao abrir, individualmente os 29 conjuntos de dados, estão disponíveis, em cada um deles, dados organizados por tipo e subtipo, disponíveis sempre nos formatos *Comma-separated values* (CSV) e planilha XLS. Tais dados estão acompanhados de um arquivo em formato Portable Document Format (PDF) cuja nomenclatura está iniciada pela palavra 'Metadados' seguida do nome do conjunto de dados. Consiste de metadados estruturais dos dados disponíveis, que explicita as variáveis codificadas, a descrição e o tipo de dados contidos no conjunto de dados, utilizando um código e não um elemento padronizado, com semântica formal,

apesar disso, considera-se que tal documento pode ser útil, embora necessite de maior dispêndio de tempo para a interpretação e o processamento do conjunto de dados.

Destaca-se, ainda, que cada dado disponível está acompanhado de metadados administrativos que representam informações como a data de criação, formato e licença, sem, contudo, apresentar os metadados descritivos básicos disponíveis no ambiente digital - como: título, palavras-chave, descrição, instituição responsável pelos dados - de forma que há prejuízos à compreensão por humanos e agentes computacionais.

Nesse desafio, os conjuntos de dados atendem parcialmente ao fornecimento de metadados, não atende ao fornecimento de metadados descritivos e atende parcialmente ao fornecimento de metadados estruturais.

No desafio **licença de dados**, a prática recomendada é fornecer informações acerca da licença de dados, por meio de um *link* ou cópia dos termos da licença (LÓSCIO; BURLE; CALEGARI, 2017). A adoção de uma licença protege o titular dos dados, bem como os consumidores, uma vez que, a partir do estabelecimento, o titular determina quais os limites de uso dos dados disponibilizados sem que haja qualquer tipo de infração aos direitos de autor e aos que lhes são conexos. Para tanto é extremamente relevante que a licença esteja disponível em um campo de metadado para que seja legível por humanos e por agentes computacionais.

Nos 29 conjuntos de dados analisados, disponíveis no tema Avaliação da Pós-Graduação *Stricto Sensu*, há a presença de uma licença do tipo Creative Commons que permite, a partir de um ícone, incorporar ao sistema a licença em três camadas: um texto jurídico embasado na lei de direitos autorais e direitos conexos; um texto legível por humanos leigos e expresso por um ícone que representa a licença adotada; e um texto legível por máquinas. Mais especificamente, a Licença Creative Commons Atribuição (CC BY), uma licença de cultura livre, utilizada para maximizar a disseminação dos conteúdos licenciados, que permite qualquer uso, desde que seja atribuído crédito pela criação original (CREATIVE COMMONS BRASIL, 2019).

A licença Creative Commons está disponível no ambiente do portal Dados Abertos Capes e nos metadados, contudo não está contida no conjunto de dados, dificultando sua identificação e uso após realizado o *download* do(s) arquivo(s), quer seja por humanos ou agentes computacionais; além disso, não está clara a versão da licença adotada, imprescindível para a compreensão dos termos de uso especificados, o que caracteriza o atendimento parcial da boa prática. A adequação, indicando a versão da licença nos metadados, assegura os benefícios de reuso e veracidade.

O desafio **proveniência dos dados** consiste em uma boa prática utilizada para manter informações acerca da origem dos dados e de alterações que tenham sido realizadas (LÓSCIO; BURLE; CALEGARI, 2017), e é utilizada para que os consumidores, quer sejam humanos ou agentes computacionais, possam, por meio do contexto histórico, ter asseguradas a qualidade, a integridade e a credibilidade dos dados.

Destaca-se que a proveniência deve ser apresentada como um elemento de metadados e para tanto, podem ser utilizados metadados administrativos baseados, por exemplo, na família PROV (ARAKAKI, 2019) ou, ainda, a ontologia PROV (LEBO; SAHOO; MCGUINNESS, 2013). PROV é conjunto de padrões recomendados pelo W3C com objetivo de suportar o intercâmbio de informações de proveniência na web.

Na área Avaliação da Pós-Graduação *Stricto Sensu* no portal Dados Abertos Capes não há metadados descritivos ou administrativos que representem a proveniência dos dados. Contudo, em cada um dos 29 conjuntos de dados analisados está disponível um arquivo em formato PDF cujo nome inicia por 'Metadados' no qual algumas informações importantes estão disponíveis. Dentre elas é possível identificar que os dados disponíveis na área analisada são coletados da Plataforma Sucupira[5] e estabelece ainda que serão definidas diretrizes para cada quadriênio de avaliação, a metodologia utilizada para a carga de dados, a versão dos dados e demais informações pertinentes. Há, ainda, um arquivo com mesmo nome, em formato *Hypertext Markup Language* (HTML), mas com uma quantidade inferior de informações. Os dois arquivos fornecem informações de proveniência de dados, legíveis apenas por humanos e, além disso, não as fornecem nos metadados para que sejam facilmente localizáveis de modo a gerar benefícios de reuso e veracidade de compreensão, assim, a boa prática relativa à proveniência de dados foi considerada não atendida.

No que tange ao desafio **qualidade de dados**, Lóscio, Burle e Calegari (2017) consideram como boa prática fornecer informações acerca da qualidade dos dados e sua adequação para aplicações específicas, considerando esse um fator preponderante na seleção dos dados e que, por isso, deve ser devidamente documentada nos metadados para que o uso possa ser avaliado pelo consumidor.

Neste sentido, Albertoni e Isaac (2016) disponibilizam um vocabulário de qualidade de dados e esclarecem que o referido material não objetiva definir qualidade de modo objetivo e ideal, do contrário, pretende auxiliar editores de dados na representação de informações acerca da qualidade dos conjuntos

5 Disponível em: https://sucupira.capes.gov.br/sucupira/. Acesso em: 01 out. 2019.

de dados abertos, informando-as nos metadados, para que humanos e agentes computacionais possam acessá-lo(s) e tomar a decisão de seleção/uso, considerando a adequação à sua necessidade; enfatizam ainda a relevância da adoção de políticas e de *feedbacks* de usuários. Os autores elucidam que a ISO/IEC 25012[6] apresenta 15 dimensões da qualidade dos conjuntos dados agrupadas em três categorias: qualidade inerente de dados, qualidade de dados inerente e dependente do sistema, e qualidade de dados dependente do sistema. Tais dimensões podem auxiliar o editor de dados na sua produção e disponibilização, e ainda ser útil aos interessados na ligação de dados, considerando, é claro, as condições de reuso estabelecidas na licença.

Neste sentido, verifica-se a ausência de informações acerca da qualidade de dados na área analisada no portal Dados Abertos Capes o que, além de não atender à boa prática impede de atingir os objetivos de reuso de veracidade.

O desafio **versionamento dos dados** é relevante à medida que contribui como um dos elementos de qualidade. Os conjuntos de dados disponíveis na web podem ser alterados ou atualizados facilmente, tornando necessário informar ao consumidor quando os dados forem alterados, utilizando as boas práticas de indicador de versão e histórico de versões.

Um indicador de versão consiste em um número ou data que individualize cada conjunto de dados tornando-o exclusivo e identificável. Cabe ao fornecedor dos dados estabelecer o indicador de versão a ser utilizado, recomenda-se, contudo, que se utilize uma padronização, para auxiliar os consumidores na coleta dos dados, que esteja disponível em um campo de metadados (LÓSCIO; BURLE; CALEGARI, 2017). Vale mencionar que, no caso de consumidores com processamento automático é imprescindível que o indicador de versão seja único.

Já o histórico de versão, consiste em uma descrição completa e detalhada das alterações realizadas em cada versão disponibilizada do conjunto de dados, visando que sua compreensão seja aprimorada pela compreensão da dinâmica (LÓSCIO; BURLE; CALEGARI, 2017). Os benefícios alcançados com o atendimento a esse desafio são reuso e veracidade.

A este respeito, na área analisada no portal Dados Abertos Capes os conjuntos de dados estão organizados por nomes e, ao final, por período de cobertura, e os dados individuais seguem a mesma estrutura, seguida da data de geração dos dados, no padrão internacional (AAAA-MM-DD), tipo e subtipo. Além disso, cada arquivo disponibiliza nos metadados a data de criação e de última

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

atualização. Em princípio, essa forma de indicador de versão atende ao indicado nas boas práticas W3C, contudo, a longo prazo, haverá muitas páginas para a navegação e localização dos conjuntos de dados, o que possivelmente irá gerar a necessidade de rever a organização da informação. Quanto ao histórico de versões, o campo de última atualização, nos dados analisados apresenta data idêntica à de disponibilização, o que nos leva a crer que não houve alterações de versões, considerando assim que a disponibilidade atende à boa prática estabelecida.

O desafio **Identificadores de dados** agrega três boas práticas, a primeira consiste no uso de *Uniform Resource Identifier* (URI) persistente como identificador individual para cada conjunto de dados (LÓSCIO; BURLE; CALEGARI, 2017). Este identificador persistente (*Persistent Identifier* - PID) pode ser mantido pela instituição fornecedora dos dados ou gerenciada por um serviço de redirecionamento, como o *Digital Object Identifier* (DOI)[7] ou o *Handle System*[8]; o que se deve ter em mente é assegurar aos consumidores que os dados estarão disponíveis e acessíveis via URI ao longo do tempo, independente de *status* ou formato, o que assegura os benefícios de reuso, conectividade, descoberta e interoperabilidade.

Como segunda boa prática, nesse desafio, Lóscio, Burle e Calegari (2017), estabelecem, reutilizar URIs persistentes de outras bases como identificadores em conjuntos de dados, para tanto, retomam Berners-Lee (2006) para elucidar que na web de dados é possível criar uma rede de hipertextos para conectar dados relacionados. Isso permite que, a partir de dados ligados haja um espaço de informação global acessível por humanos e agentes computacionais, sem a necessidade que todos os dados sejam gerados por uma única organização, processo designado como *linked data*. Ao conectar dados é imprescindível certificar-se da qualidade dos dados aos quais se conecta, bem como da adoção de URIs persistentes, para manter a credibilidade e confiabilidade dos seus dados. Essa boa prática assegura os benefícios de reuso, conectividade, descoberta e interoperabilidade.

E, finalmente, a terceira, consiste em atribuir URIs a versões individuais de conjuntos de dados, bem como à série geral (LÓSCIO; BURLE; CALEGARI, 2017), que consiste em determinar uma URI genérica para o conjunto de dados e, utilizá-la como prefixo das suas versões, de modo a estabelecer, para humanos e agentes computacionais, informações consistentes acerca dos dados específicos que acessam nos conjuntos disponíveis, o que trará benefícios de reuso, descoberta e veracidade.

7 Disponível em: https://www.doi.org/. Acesso em: 4 abr. 2019.
8 Disponível em: https://www.dona.net/handle-system. Acesso em: 4 abr. 2019.

Na área analisada neste estudo, a URI utilizada é própria, mantida pela Capes, e adota um padrão de prefixo para todos os conjuntos de dados. Consideramos que a primeira boa prática é atendida, embora não seja possível afirmar que a instituição adota uma política para que a URI seja acessível ao longo do tempo. Os dados disponíveis no portal Dados Abertos Capes não são ligados a dados disponíveis em bases externas e não atendem à boa prática de reuso de URIs. No que tange a atribuir URIs individuais para versões individuais, que atende parcialmente, visto que cada formato de arquivo referente a um mesmo dado possui uma URI própria e, ao observar que, após o prefixo do conjunto de dados, as versões individuais são identificadas por códigos alfanuméricos, recomenda-se o uso de uma forma mais amigável de identificação.

No que se refere ao desafio **formatos de dados**, Lóscio, Burle e Calegari (2017) elucidam que o formato em que os dados são disponibilizados influenciam na sua utilização, por isso incentivam o uso de formatos que possam ser utilizáveis pelo maior público, considerando ainda a facilidade de leitura e processamento por ferramentas computacionais que auxiliarão os humanos na análise e interpretação dos dados. Nesse sentido, esse desafio está dividido em três boas práticas, a primeira consiste na disponibilização dos dados em formatos padronizados, legíveis por máquinas e adequados ao uso pretendido ou potencial. Recomendam ainda algumas sintaxes, como CSV, XML, HDF5, JSON, RDF, RDF/XML, JSON-LD e Turtle. O atendimento a esta boa prática gera benefícios de reuso e processabilidade.

Os dados analisados no presente estudo estão disponibilizados nos formatos CSV, HTML, PDF, XLSX, XLS, atendendo parcialmente às boas práticas do W3C, uma vez que atende apenas a uma recomendação, o CSV.

Ainda sob esse desafio, a boa prática usar representações de dados com localidade neutra ou, na impossibilidade, fornecer metadados acerca da localidade utilizada nos valores dos dados. Ao disponibilizar dados é importante que eles possam ser claros para que sejam interpretados e utilizados de maneira unívoca e sem duplas interpretação. Assim, dados como datas, horas, moedas, números, que, dependendo do formato utilizado na representação pode ser interpretado de forma indevida por humanos ou agentes computacionais devem ser representados com localização neutra (LÓSCIO; BURLE; CALEGARI, 2017). Seu uso adequado gera benefícios e reuso e compreensão.

Nos conjuntos de dados analisados, os valores como data são apresentados no padrão internacional, o que torna a representação padronizada, ou seja, neutra. Contudo, nos metadados administrativos, inseridos junto aos conjuntos de dados de dados em uma tabela, sem uso adequado de um esquema de metadados, os mesmos valores são apresentados no padrão brasileiro, o que

demonstra falta de padronização e uso inadequado dos metadados. Desta forma, atende parcialmente à boa prática, visto que pode haver prejuízo na interpretação desses dados por humanos e agentes computacionais.

Ainda nesse desafio, há uma boa prática que indica a necessidade de fornecimento de dados em formatos múltiplos, considerando o uso pretendido ou potencial, visando reduzir custos ou erros decorrentes da transformação de dados em outros formatos. Esta análise já foi iniciada anteriormente e, embora os dados analisados estejam disponíveis nos formatos CSV, XLSX, XLS, apenas o CVS é um formato adequado para a boa prática referente aos dados legíveis por máquinas. O benefício alcançado é reuso e processabilidade. Destaca-se que os arquivos HTML e PDF constantes da plataforma, referem-se a documentos textuais referente aos metadados, não se tratando de formatos de disponibilização dos dados propriamente ditos, considera-se com isso que a boa prática é parcialmente atendida.

O desafio **vocabulários de dados** refere-se aos conceitos e relacionamentos (atributos) utilizados para descrição e representação em uma área de interesse (LÓSCIO; BURLE; CALEGARI, 2017). Na web semântica são utilizados para estabelecer estruturas padronizadas que permitam a ligação e a interoperabilidade entre os dados; tais estruturas podem ser simples ou complexas, considerando os requisitos e objetivos da aplicação (W3C, 2019).

Esse desafio está dividido em duas boas práticas, a primeira delas recomenda o uso de termos de vocabulários compartilhados, preferencialmente padronizados, para codificar dados e metadados (LÓSCIO; BURLE; CALEGARI, 2017), visando a ampliação da interoperabilidade e reutilização dos dados, dado o consenso entre fornecedores e consumidores de dados. A adoção de metadados padronização facilita a compreensão e auxilia o processamento automático de dados e metadados. Os autores citam como exemplo de vocabulários preexistentes: VOCAB-DCAT, Dublin Core, FOAF, SKOS e vCard. Os benefícios do atendimento dessa boa prática são reuso, processabilidade, compreensão, veracidade e interoperabilidade.

Nos 29 conjuntos de dados analisados no Portal Dados Abertos Capes, embora haja um vocabulário, com metadados administrativos, descritivos e de versão, não são utilizados vocabulários padronizados de dados e metadados, de forma que não atende à boa prática recomendada pelo W3C.

E, a segunda boa prática recomenda a adoção de um nível de semântica formal que se adapte aos dados e aos aplicativos de uso potencial. A semântica formal permite estabelecer especificações claras em significado e a adoção de um vocabulário pode sustentar o processamento automático dos dados. É relevante encontrar equilíbrio entre a adoção de um vocabulário simples ou complexo,

visto que o primeiro pode omitir informações relevantes e o segundo pode exigir esforço demasiado para o reuso dos dados, por isso é importante ressaltar que o objetivo consiste no reuso dos dados e não apenas na disponibilização. Os benefícios alcançados com a adoção dessa prática são reuso, compreensão e interoperabilidade.

Considerando a análise da boa prática anterior, na qual identifica-se que os conjuntos de dados selecionados no Portal Dados Abertos Capes, não utilizam vocabulários padronizados de dados e metadados, verifica-se ainda que não há semântica formal, com isso todo o desafio referente ao vocabulário de dados não é atendido.

Disponibilizar dados na web permite que consumidores humanos ou agentes computacionais possam utilizar-se deles, seja por meio de um simples *download* de um arquivo, ou em massa utilizando uma *Application Programming Interface* (API). A disponibilização dos dados deve considerar se são estáticos ou dinâmicos, disponíveis em tempo real, o tamanho do conjunto e sua granularidade, informações estas definidas por políticas estabelecidas pelos fornecedores de dados. Compete ainda aos fornecedores de dados a definição da forma de acesso, para a qual podem ser requeridos dados do consumidor. Dessa forma, o **acesso a dados** também consiste em um desafio, dividido em 10 boas práticas que definem os comportamentos esperados dos atores envolvidos nesse contexto. Considerando a extensão desse desafio, seis boas práticas serão discutidas individualmente.

Os dados podem ser disponibilizados como conjuntos ou distribuídos, resultando em um conjunto de URIs, dessa forma, a primeira boa prática consiste em permitir que o *download* de conjuntos de dados completos possa ser realizado em massa, a partir de uma única requisição, permite ao consumidor maior liberdade na definição da forma de acesso aos dados, individualmente ou em conjunto. Os benefícios desta prática são reuso e acesso. Os dados analisados, embora apresentem abaixo do conjunto de dados um *link* para os diferentes formatos de arquivos disponíveis, ao clicar direciona para uma página que disponibiliza os *links* para os dados individuais, de modo que, para humanos, não atende à boa prática preconizada pelo W3C. Por outro lado, a ferramenta utilizada no portal Dados Abertos Capes, o software *Comprehensive Knowledge Archive Network* (CKAN), disponibiliza uma API que permite que agentes computacionais façam o *download* em massa, o que atende à boa prática. Considerando que o consumo de dados em massa é possível apenas a aplicações computacionais, essa boa prática é considerada parcialmente atendida.

Ainda considerando grandes conjuntos de dados e a possibilidade de consumo de um subconjunto, a segunda boa prática consiste em fornecer subconjuntos para grandes conjuntos de

dados, de modo a auxiliar os consumidores no armazenamento, processamento e análise. Afirmam que Lóscio, Burle e Calegari (2017, tradução nossa) que "Os dados que levam mais de dez segundos para serem entregues provavelmente farão com que os usuários suspeitem de falha.". Os benefícios dessa boa prática são reuso, conectividade, acesso e processabilidade. Na área analisada, os conjuntos de dados estão organizados por nomes e, ao acioná-los via *hiperlink* os dados estão disponíveis individualmente, atendendo à boa prática estabelecida.

A disponibilização de dados em uma interface pode resultar em vários formatos de dados legíveis por humanos no mesmo espaço que dados legíveis por aplicações computacionais. Assim a terceira boa prática para acesso a dados é utilizar negociação de conteúdo para disponibilizar os dados em diferentes formatos (LÓSCIO; BURLE; CALEGARI, 2017), elucidada por BERNERS-LEE (2009) como uma flexibilidade na arquitetura web[9], em que "De forma simples, o cliente envia um cabeçalho *Accept*: com uma lista de tipos de conteúdo que ele entende, e o servidor retorna as informações usando um deles.". Vale ressaltar que os dados e suas representações devem estar disponíveis a partir da mesma URI, em diferentes formatos legíveis por máquinas e adequado ao uso, para que tenham os benefícios de reuso e acesso. Os conjuntos de dados analisados não atendem a essa boa prática.

Sempre que os dados forem produzidos em tempo real, eles devem ser disponibilizados na web em tempo real ou considerando apenas o atraso gerado pelo processamento e transmissão necessários à disponibilização, e o acesso deve ser possibilitado por pesquisa, API ou *streaming* (LÓSCIO; BURLE; CALEGARI, 2017); tal forma de disponibilização é relevante para dados que necessitam de monitoramento. Os benefícios gerados são reuso e acesso. A área analisada disponibiliza dados estáticos, por faixa temporal, e, por isso, não atende à boa prática indicada.

A quinta boa prática para acesso a dados é disponibilizar dados atualizados, visando estimular o interesse e o uso pelos consumidores; e informar a frequência de atualização, legível por humanos e agentes computacionais, atentando-se à prática de utilização de indicador de versão. Os benefícios gerados são reuso e acesso. A área analisada não disponibiliza qualquer informação acerca da frequência de atualização e, os dados do período anterior, referente ao ano civil de 2018, cuja coleta já foi encerrada ainda não estão completamente disponíveis, dessa forma, não atende à boa prática indicada pelo W3C.

Um desafio frequente no acesso aos dados é a indisponibilidade definitiva ou momentânea do conjunto de dados, ainda que utilizando a URI específica. Para tanto, a sexta boa prática de

9 Disponível em: https://www.w3.org/TR/webarch/. Acesso em: 6 abr. 2019.

acesso a dados é fornecer uma explicação para os dados que não estiverem disponíveis para que os consumidores possam ser informados da razão da indisponibilidade; ainda, o acesso aos dados pode ter sido redirecionado a outro ambiente ou arquivado e acessível mediante solicitação. Qualquer que seja a causa, é imprescindível que o consumidor tenha como retorno à sua solicitação um código de resposta HTTP, lembrando que cada código de *status* se refere a uma resposta à requisição do consumidor (FIELDING, 1999). Os benefícios trazidos por esta boa prática são reuso e veracidade. Nesse estudo, considerando que o portal analisado é muito recente, ainda não há dados retirados, assim, consideramos que essa boa prática não se aplica à análise.

E, para encerrar o desafio referente ao acesso a dados, há quatro boas práticas diretamente relacionadas a APIs, que serão analisadas em conjunto. A primeira delas consiste em disponibilizar dados por meio de API, gerando maior flexibilidade e capacidade de processamento aos consumidores (LÓSCIO; BURLE; CALEGARI, 2017). Cientes da complexidade do desenvolvimento de APIs específicas, os autores mencionam que alguns softwares utilizados para a disponibilização de dados, a exemplo do CKAN já dispõe de API e manual[10] específico para uso. Os benefícios do atendimento a esta prática são reuso, processabilidade, interoperabilidade e acesso.

A segunda, recomenda que as APIs utilizem padrões da web, facilitando a manutenção e o entendimento pelos desenvolvedores e consumidores de dados (LÓSCIO; BURLE; CALEGARI, 2017). Nesse sentido, citam como exemplo a *Representational State Transfer* (REST), uma representação padronizada, que consiste em princípios, regras e *constraints* que permitem que aplicações se comuniquem (FIELDING, 2000). *Web services* que utilizam a arquitetura REST ou RESTful, possibilitam a interoperabilidade entre sistemas na internet. Os benefícios do atendimento a esta prática são reuso, conectividade, interoperabilidade, descoberta, acesso e processabilidade.

A terceira boa prática relacionada à API refere-se ao fornecimento de documentação completa para que os consumidores de dados possam compreendê-la e utilizá-la, de igual maneira, é recomendado que as alterações sejam destacadas para facilitar a identificação (LÓSCIO; BURLE; CALEGARI, 2017). Os benefícios do atendimento a esta prática são reuso e veracidade.

E, por fim, no que tange ao acesso a dados via API, Lóscio, Burle e Calegari (2017), recomendam que alterações que afetem o uso das APIs sejam evitadas ou ao menos indicadas aos consumidores. Os consumidores dos dados utilizam-se da documentação da API para desenvolver códigos e implementar

10 Disponível em: https://docs.ckan.org/en/latest/maintaining/datastore.html. Acesso em: 6 abr. 2019.

clientes para a API e as alterações que afetam o funcionamento podem gerar uma quebra no código do cliente. Desenvolver uma comunicação prévia permite que os consumidores criem mecanismos para que o consumo dos dados não seja prejudicado pela alteração e isso aumentará a confiança. Os benefícios desta prática são veracidade e interoperabilidade.

Quando tratamos de acesso aos dados, estamos abordando o acesso em duas camadas: a interface, utilizada por humanos e, interoperabilidade, integração, arquitetura cliente-servidor ou API, por agentes computacionais. Nesse estudo analisamos a área Avaliação da Pós-Graduação *Stricto Sensu* do portal Dados Abertos Capes, que utiliza como ferramenta para a disponibilização dos dados o *software* CKAN. O CKAN disponibiliza API, baseada em JSON, cuja utilização está adequadamente documentada para administradores[11] e usuários[12] do sistema, de modo que as boas práticas preconizadas pelo W3C são atendidas. Vale destacar que a API está atrelada ao *software*, de forma que, se o portal Dados Abertos Capes, por alguma razão passar a adotar outra ferramenta para a disponibilização dos dados, este conjunto de boas práticas poderá estar em risco.

Embora não seja possível assegurar a presença definitiva de um dado na web, o desafio designado **preservação de dados** demonstra uma preocupação no retorno aos consumidores de dados, para que não tenham como retorno uma mensagem 'recurso não encontrado', por isso estabelecem duas boas práticas.

A primeira delas estabelece que, ao remover dados da web, o identificador deve ser preservado e fornecidas informações acerca do recurso (LÓSCIO; BURLE; CALEGARI, 2017). O que se pretende é oferecer informações ao consumidor, para que não haja dúvidas sobre a indisponibilidade temporária ou permanente do recurso, deve-se, então, manter a URI e, ao remover os dados, incluir uma informação acerca dessa remoção. Nesse sentido, há a possibilidade da exclusão permanente do dado, adotando o retorno do código de resposta http 410, que notifica que o recurso não está disponível intencionalmente e que os *links* remotos para ele devem ser removidos (FIELDING, 1999); ou do acesso por outra via, a exemplo de uma solicitação. De toda forma, é imprescindível que a URI seja mantida permanentemente, em atendimento à boa prática de uso de identificador persistente. Essa boa prática tem como benefícios reuso e veracidade.

A segunda boa prática considera a característica de dependência entre os dados disponíveis na web a um conjunto ou contexto, o que gera a necessidade de preservação conjunta do conjunto

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

de dados, seu contexto e dos vocabulários utilizados, por isso, a boa prática consiste em avaliar a cobertura do conjunto de dados antes da sua preservação (LÓSCIO; BURLE; CALEGARI, 2017). Ao planejar a preservação deve-se considerar tudo o que o conjunto de dados necessitará para ser acessível e compreensível ao longo do tempo, é relevante ainda verificar se há dados conectados e se esses estão sendo preservados. Essa boa prática gera benefícios de reuso e veracidade.

Nesse estudo, há limitação de acesso às políticas do portal analisado, de forma que o planejamento da preservação dos dados não foi consultado. De igual maneira, considerando o tempo de implantação, ainda não há dados retirados do portal. Desta forma, considera-se que as boas práticas relacionadas à preservação de dados não se aplicam à análise.

A disponibilização de dados na web permite reuso e compartilhamento por diferentes consumidores. O desafio ***feedback*** consiste na possibilidade de coletar comentários desses consumidores, quer seja por formulários ou caixas, nas quais possam descrever suas experiências de uso, que tragam subsídios capazes de gerar melhorias na disponibilização de dados; ou ainda pela coleta de métricas ou informações de aplicações que coletam dados automaticamente (LÓSCIO; BURLE; CALEGARI, 2017). Para tanto, os autores recomendam duas boas práticas, a primeira consiste em fornecer um mecanismo facilmente detectável de *feedback* disponível para que os consumidores possam fornecer suas avaliações, tendo como benefícios reuso, compreensão e veracidade. Enquanto a segunda, consiste na disponibilização pública dos *feedbacks* fornecidos pelos consumidores visando demonstrar que o fornecedor dos dados está atento aos problemas apresentados e auxiliar os consumidores em questões que possam afetar o consumo dos dados. Os benefícios gerados por esta boa prática são reuso e veracidade.

O portal Dados Abertos Capes não disponibiliza qualquer mecanismo de *feedback* e, como limitação, não temos acesso às políticas do portal o que impede a avaliação da coleta de métricas de uso por agentes computacionais. Dessa forma, considera-se que a área analisada não atende às boas práticas estabelecidas para esse desafio.

O **enriquecimento de dados** consiste em processos que podem ser utilizados para aprimorar os dados brutos, capazes de torná-los um recurso extremamente valioso. Contudo, é extremamente relevante que haja cautela no enriquecimento, visando manter a integridade, o sigilo e a privacidade de dados sensíveis. Sob esse desafio estão duas boas práticas.

Enriquecer dados gerando novos dados quando isso aumentar seu valor (LÓSCIO; BURLE; CALEGARI, 2017), prática que amplia significativamente a processabilidade, a partir do preenchimento

de valores omissos ou da adição de atributos não disponíveis inicialmente. Para tanto podem ser aplicados aprendizagem de máquina, inferências e outros métodos de enriquecimento. Desta ainda que a disponibilização do código utilizado para o enriquecimento é recomendável sempre que permitido pela licença adotada. Essa prática assegura os benefícios de reuso, compreensão, veracidade e processabilidade.

E, fornecer visualizações complementares dos dados os dados, utilizando, além da opção de *download* e API, visualizações, tabelas, aplicativos da web ou resumos, prática essa que pretende beneficiar consumidores humanos ao fornecer formas alternativas de visualização para a interpretação imediata dos dados sem a necessidade de baixar o arquivo, gerando benefícios de reuso, compreensão, acesso e veracidade.

Ao analisar a área Avaliação da Pós-Graduação *Stricto Sensu* no portal Dados Abertos Capes, verifica-se que não há enriquecimento de dados, contudo, há formas complementares de visualização, por meio de uma pré-visualização da tabela e gráficos configurados pelo usuário, sendo essas funcionalidades disponíveis por meio do CKAN[13], *software* utilizado pelo portal.

O último desafio consiste na **republIcação de dados**, considerando que o reuso é outra forma de disponibilização, Lóscio, Burle e Calegari (2017) estabelecem três boas práticas a serem seguidas para os interessados em republicar dados que foram disponibilizados inicialmente em outras plataformas.

Cumpre esclarecer que a terminologia republicação de dados é utilizada neste estudo por ser a adotada pela W3C, contudo, considerando Kratz e Strasser (2014) a publicação utiliza-se da disponibilização dos dados para que sejam acessíveis, contudo, para que sejam considerados publicados, os dados necessitam de maior rigor e formalização, como: estruturação para que os dados possam ser reutilizados, adoção de licença de direitos autorais, documentação que explicite minunciosamente todos os elementos necessários à compreensão dos dados (*data paper*), utilização de um repositório digital confiável que realize preservação digital dos dados, representação exaustiva dos dados utilizando metadados estruturados; adoção de identificador persistente e validação dos dados.

Isso posto, verifica-se que alguns desses elementos são mencionados nas boas práticas para dados na web (LÓSCIO; BURLE; CALEGARI, 2017), contudo, as boas práticas relacionadas a esse desafio não o caracterizam como publicação e sim como disponibilização de dados.

13 Disponível em: https://ckan.org/. Acesso em: 15 out. 2019.

A primeira boa prática desse desafio é fornecer *feedback* ao fornecedor de dados, que consiste em informá-lo sobre a reutilização dos dados e, caso haja algum erro no consumo. Para os fornecedores de dados é necessário avaliar se os dados disponibilizados são úteis e podem ser consumidos, assim, haverá estatísticas de uso e conhecimento da utilidade dos dados disponibilizados, subsídios importantes para o constante aprimoramento dos conjuntos de dados. Como benefícios, podem ser alcançados reuso, interoperabilidade e veracidade.

A segunda é identificar e seguir os termos estabelecidos na licença adotada nos dados reutilizados. Ao licenciar o recurso, o editor estabelece a forma autorizada de uso dos dados sem que haja infração aos seus direitos, assim, atender à licença assegura ao fornecedor e ao consumidor garantias legais que os mantém seguros na disponibilização, uso e reuso de dados. Recomenda-se assim, que seja realizada uma leitura dos termos de licenciamento para que não haja qualquer infração ao consumir os dados e sejam assegurados os benefícios de reuso e veracidade.

E, finalmente, citar a fonte original dos dados nos metadados e, caso utilize uma interface para usuários, incluir citação visível. A confiabilidade dos dados pode ser atrelada à sua fonte original de publicação ou à sua proveniência, além disso, é de extrema relevância que os fornecedores de dados tenham seu trabalho reconhecido nos créditos, assim como frequentemente se faz com as publicações textuais, a publicação de dados ganha espaço no meio governamental e científico e a autoria deve ser reconhecida. Os benefícios dessa prática são reuso, descoberta e veracidade.

Esse desafio também não pode ser avaliado na área Avaliação da Pós-Graduação *Stricto Sensu* no portal Dados Abertos Capes, uma vez que os dados disponibilizados referem-se à Plataforma Sucupira, mas não podem ser considerados republicação, por não estarem nela disponíveis e sim direcionados ao dela ao portal Dados Abertos Capes, por isso, tais boas práticas não se aplicam à análise.

5 CONSIDERAÇÕES FINAIS

O presente estudo objetivou discutir o atendimento de 29 conjuntos de dados abertos governamentais acerca da Avaliação da Pós-Graduação *Stricto Sensu* disponíveis no portal Dados Abertos Capes às 35 boas práticas para a disponibilização de dados na web, recomendadas pelo W3C, bem como os benefícios alcançados pelo atendimento, de forma a fornecer aportes teóricos da Ciência da Informação para subsidiar os fornecedores de dados, no tocante à melhoria dos dados disponibilizados na web.

A partir da análise, apresentada no quadro 1, foi possível verificar que, das 35 boas práticas, 7 não se aplicam aos conjuntos de dados analisados, sendo consideradas apenas 28. Destas, 20 foram consideradas não atendidas ou parcialmente atendidas e apenas 8 consideradas atendidas, estando 3 delas atreladas ao *software* CKAN utilizado para a disponibilização dos dados e não aos dados propriamente ditos.

Destaca-se que dados abertos requerem tratamento adequado para que possam ser recuperados, compreensíveis e processáveis por humanos e agentes computacionais, de forma que a simples disponibilização na web não atende adequadamente o que preconiza a LAI.

Assim, primeira fragilidade a se destacar no portal Dados Abertos Capes refere-se aos metadados. As boas práticas para dados na web (LÓSCIO; BURLE; CALEGARI, 2017) explicitam a necessidade da adoção e do fornecimento de metadados adequadamente estruturados em duas dimensões: metadados e vocabulários de dados, contudo outras 8 dimensões dependem da estrutura de metadados para que as boas práticas possam ser atingidas: licenças de dados, proveniência de dados, qualidade de dados, versão de dados, identificadores de dados, formatos de dados, acesso a dados, preservação de dados. Evidencia-se com isso que a compreensão dos dados e seu contexto devem ser estruturados e adequadamente representados por meio dos metadados administrativos, estruturais, técnicos, de preservação, de uso, estruturais, de proveniência, que devem ser utilizados em consonância com a necessidade do conjunto de dados.

A esse respeito, considerando que a análise comparativa entre os metadados estruturais dos conjuntos de dados disponíveis no portal Dados Abertos Capes não é foco principal desse estudo, verifica-se que há identificadores (ID), campos (código) e descrição (nome) idênticos, mas utilizando variáveis distintas, o que dificulta o desenvolvimento de processos automatizados, por isso, recomenda-se a adoção de um padrão, preferencialmente amplamente utilizado, para que não haja inconsistência ou falha na implementação de aplicações para consumo ou interoperabilidade dos dados.

Além disso, verifica-se que a Infraestrutura Nacional de Dados Abertos (INDA), estabeleceu um padrão de metadados[14] obrigatórios e opcionais para descrever os conjuntos de dados que compõem o catálogo do Portal Brasileiro de Dados Abertos[15], visando a padronização para fornecedores e consumidores de dados; e, ainda que sendo uma organização partícipe do referido portal a Capes não atende sequer aos metadados descritivos obrigatórios requeridos pela INDA.

14 Disponível em: http://wiki.dados.gov.br/Padroes-de-metadados.ashx. Acesso em: 02 fev. 2019.
15 Disponível em: http://dados.gov.br/. Acesso em: 2 fev. 2019.

Adequar a semântica formal dos dados, adotar um padrão padronizado e fornecer metadados, que consiste em atender às boas práticas (1-3; 15-16) trará melhorias significativas à estrutura dos dados e com isso benefícios de descoberta, compreensão, veracidade, processabilidade, interoperabilidade e reuso. O que certamente auxiliará aos fornecedores e consumidores de dados, que poderão fazê-lo de forma manual ou automatizada.

Considerando ainda os metadados, é imprescindível que a proveniência dos dados seja adequadamente registrada. A confiabilidade, a veracidade, a compreensão e o reuso estão dependentes dessa prática, capaz de assegurar a integridade e a autenticidade do dado, garantindo que ele não tenha sido adulterado e evidenciando a confiabilidade, o que demonstra direta relação com a sua qualidade e preservação.

Destaca-se ainda a relevância de estabelecer adequadamente a licença de dados, especificando o tipo e a versão, por meio de metadados legíveis por humanos e agentes computacionais, além de fornecer acesso ao texto jurídico, para que o reuso possa ser realizado atentando-se às questões legais para que fornecedores e consumidores tenham assegurada a proteção.

Outra fragilidade é a inexpressiva variedade de formatos de arquivos legíveis por máquinas, com a presença apenas de arquivo CSV, em detrimento de formatos como XML, HDF5, JSON, RDF ou Turtle, o que beneficiaria o reuso e a processabilidade.

Ainda como aspecto a ser melhorado, recomenda-se que *feedbacks* sejam coletados e disponibilizados, visando ampliar a transparência no processo de abertura dos dados, ampliando reuso e veracidade.

Há uma questão que merece alerta, trata-se da utilização de URIs amigáveis, persistentes e estáveis, evitando quebrar os *links* existentes, pois os dados, quando conectados, necessitam de maior garantia de estabilidade da URI. Torna-se assim necessária a adoção de um serviço de PID, a exemplo do DOI ou do *Handle System*.

Como aspectos positivos, verificam-se a manutenção dos indicadores e históricos de versões, que permitem que os consumidores de dados possam definir os dados a serem utilizados, e possibilitam com isso o reuso e veracidade. Além disso, os dados são disponibilizados por meio de API do *software* CKAN, cujos benefícios atingidos são: reuso, acesso, descoberta, confiabilidade, veracidade, interoperabilidade e conexão. No que tange às APIs, são totalmente dependentes do *software*, fator que deve ser considerado no momento da escolha da ferramenta a ser utilizada em um ambiente de disponibilização de dados.

Destaca-se ainda o relatório de implementação das boas práticas para dados na web (LÓSCIO; BURLE; CALEGARI, 2019) corroboram com os resultados encontrados no presente estudo ao apresentarem como desafios metadados e vocabulários de dados, licenças e formatos de dados, proveniência, versionamento, *feedback* e preservação de dados.

Por fim, destaca-se a necessidade de adequação da disponibilização dos conjuntos de dados disponíveis no portal Dados Abertos Capes para que possam atender às boas práticas para dados na web e com isso assegurar que o objetivo da disponibilização dos dados seja realmente atingido e com isso atender às 5 estrelas dos dados abertos propostas por Berners-Lee (2012).

REFERÊNCIAS

ALBERTONI, R.; ISAAC, A. (ed.). **Dados sobre as melhores práticas da web**: vocabulário de qualidade de dados. 2016. Disponível em: https://www.w3.org/TR/vocab-dqv/. Acesso em: 1 abr. 2019.

ARAKAKI, F. A. **Metadados administrativos e a proveniência dos dados**: modelo baseado na família PROV. 2019. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2019. Disponível em: http://hdl.handle.net/11449/180490. Acesso em: 1 abr. 2019.

BERNERS-LEE, T. **Content negotiation of content-type**. 2009. Disponível em: https://www.w3.org/DesignIssues/Conneg. Acesso em: 4 jun. 2018.

BERNERS-LEE, T. **Linked data**. 2006. Disponível http://www.w3.org/DesignIssues/LinkedData.html. Acesso em: em: 4 jun. 2018.

BERNERS-LEE, T. **5 [stars] open data**. 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: 23 jan. 2020.

BRASIL. Lei nº 9.610, de 19 de fevereiro de 1998. **Diário Oficial da União, Brasília**, DF, 20 fev. 1998. Seção 1. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L9610.htm. Acesso em: 15 nov. 2018.

BRASIL. Decreto nº 8.777, de 11 de maio de 2016. **Diário Oficial da União**, Brasília, DF, 12 maio 2016. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/D8777.htm. Acesso em: 23 jan. 2020.

BRASIL. [Constituição (1988)]. Constituição da República Federativa do Brasil de 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 23 jan. 2020.

CREATIVE COMMONS BRASIL. **Sobre as licenças**. Disponível em: https://br.creativecommons.org/licencas/. Acesso em: 01 abr. 2019.

DADOS abertos da Capes. Disponível em: https://dadosabertos.capes.gov.br. Acesso em: 31 mar. 2019.

FIELDING, R.T. **Hypertext transfer protocol**: HTTP/1.1. 1999. Disponível em: https://www.w3.org/Protocols/rfc2616/rfc2616.html. Acesso em: 5 abr. 2019.

FIELDING, R T. **Representational State Transfer (REST)**. 2000. Disponível em: https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm. Acesso em: 5 abr. 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.

KRATZ, J; STRASSER, C. Data publication consensus and controversies [versão3]. **F1000Research**, n. 94, 2014. Disponível em: https://doi.org/10.12688/f1000research.3979.1. Acesso em: 14 jan. 2020.

LEBO, T.; SAHOO, S.; MCGUINNESS, D. (ed.). **PROV-O**: a ontologia PROV. 2013. Disponível em: https://www.w3.org/TR/prov-o/. Acesso em: 01 abr. 2019.

LÓSCIO, B.F.; GUIMARÃES, C.B. dos S.; CALEGARI, N. Boas práticas para dados na web: desafios e benefícios. **Revista Principia - Divulgação Científica e Tecnológica do IFPB**, [S.l.], n. 32, p. 9-18, dez. 2016. Disponível em: http://periodicos.ifpb.edu.br/index.php/principia/article/view/1023/578. Acesso em: 13 abr. 2019.

LÓSCIO, B.F.; BURLE, C.; CALEGARI, N. (ed.). **Data on the web best practices**. 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 26 abr. 2018.

LÓSCIO, B.F.; BURLE, C.; CALEGARI, N. (ed.). **DWBP Implementation Report**. 2019. Disponível em: http://w3c.github.io/dwbp/dwbp-implementation-report.html. Acesso em: 13 abr. 2019.

OPEN KNOWLEDGE FOUNDATION. **The open data handbook**. Disponível em: http://opendatahandbook.org/. Acesso em: 14 jan. 2020.

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/Understanding%20Metadata.pdf. Acesso em: 15 jul. 2018.

THE ANNOTATED 8 PRINCIPLES OF OPEN GOVERNMENT DATA. Disponível em: https://opengovdata.org/. Acesso em: 23 jan. 2020.

THE OPEN DEFINITION. Disponível em: http://opendefinition.org/. Acesso em: 23 jan. 2020.

W3C. **Vocabulary**. Disponível em: https://www.w3.org/standards/semanticweb/ontology. Acesso em: 12 abr. 2019.