Boas práticas para dados na WEB: desafios e benefícios

Bernadette Farias Lóscio[1], Caroline Burle S. Guimarães[2], Newton J. Calegari[3]

[1] jbfl@cin.ufpe.br, [2] cburle@nic.br [3] newton@nic.br - [1]UFPE - Center for Informatics at Federal University of Pernambuco - Recife - PE; [2,3] Web Technology Study Center at Brazilian Network Information Center - NIC.br - São Paulo - SP

RESUMO

Este trabalho tem como objetivo compartilhar pesquisas empíricas sobre a publicação e utilização de dados na Web a partir dos casos de uso compilados pelo Grupo de Trabalho do W3C Boas Práticas de Dados Web, do inglês W3C Data on the Web Best Practices (DWBP). O Grupo de Trabalho compilou cenários de como os dados normalmente são publicados na Web e como são usados. Esses casos de uso constituíram a base para os principais desafios enfrentados pelos publicadores e consumidores de dados. A partir dos desafios estabelecidos, um conjunto de requisitos foi definido, e ambos orientaram o desenvolvimento das boas práticas para publicação de dados na Web. O documento com as boas práticas também enfatiza a importância de fornecer informações sobre os conjuntos de dados e distribuições com o intuito de contribuir para o aumento de reutilização dos dados. Nesse contexto, este artigo analisa os benefícios de engajar os publicadores na utilização das boas práticas, bem como o uso das boas práticas para melhorar a forma que os dados são publicados na Web.

Palavras-chave: Dados na Web. Boas Práticas. Desafios. Benefícios. Publicação.

ABSTRACT

This paper aims to share empirical research on publishing data on the Web and its use. Starting from the use cases compiled by the W3C Data on the Web Best Practices (DWBP) working group, which compiled scenarios of how data is commonly published on the Web and how it is used. These use cases were the basis to set up the main challenges faced by data publishers and data consumers. Following the challenges, a set of requirements was defined, and both guided the development of the Data on the Web Best Practices (DWBP). It also discusses the importance of providing information about the datasets and distributions that may also contribute to data reuse. Finally, it analyses the benefits to engage data publishers in using the Best Practices as well as the use of the best practices to improve the way the datasets are published on the Web.

Keywords: ***Data on the Web. Best Practices. Challenges. Benefits. Publication.***

1 Introdução

A abertura e a flexibilidade da Web criam novos desafios para os publicadores e consumidores de dados, tais como a forma de representar, descrever e disponibilizar dados de forma que seja fácil encontrá-los e compreendê-los. Neste contexto, é crucial fornecer orientação para os publicadores de dados na Web. Tal orientação promove a reutilização de dados e fomenta a confiança entre publicadores e consumidores de dados, independentemente da tecnologia utilizada para publicá-los, aumentando o potencial de inovação baseada no uso daqueles dados. Para atender essa necessidade, um conjunto de 35 Boas Práticas (BPs) foi publicado pelo Grupo de Trabalho do W3C - Boas Práticas para Dados Web, do

inglês *W3C Data on the Web Best Practices* (DWBP). As BPs podem ser utilizadas por publicadores e consumidores de dados de forma a ajudá-los a superar os diferentes desafios enfrentados ao publicar e consumir dados na Web. A fim de definir o escopo das boas práticas e obter os recursos necessários para a sua elaboração, o grupo de trabalho DWBP compilou um conjunto de casos de uso (LEE; LÓSCIO; ARCHER, 2015) que representam cenários de como os dados são comumente publicados na Web e como são usados. Com base nesses casos de uso, foram identificados os principais desafios enfrentados pelos publicadores e consumidores de dados e, para cada desafio, um conjunto de requisitos foi definido. Estes desafios e requisitos constituíram a base para o desenvolvimento das Boas Práticas para Dados na Web (DWBP) (LÓSCIO et al., 2016).

Neste artigo, discutiremos os desafios identificados com base nas evidências empíricas coletadas a partir do conjunto de casos de uso compilados pelo Grupo de Trabalho DWBP. Esses desafios demonstram a importância da utilização de boas práticas para publicar dados na Web, bem como contribuem para melhorar a comunicação entre os publicadores e consumidores de dados. Também abordaremos os principais benefícios da utilização do conjunto de boas práticas.

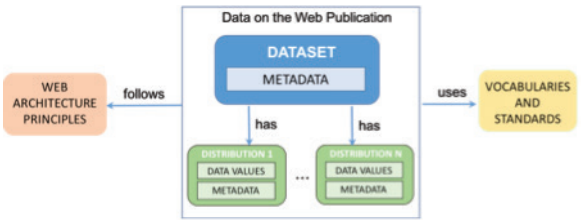
O restante deste artigo está organizado como se segue. Na Seção 2 descrevemos o contexto em que as boas práticas para publicação de dados na Web foram definidas. Na Seção 3 apresentamos os casos de uso de dados na Web. Na Seção 4 introduzimos as boas práticas para dados na Web, enquanto que na Seção 5 discutimos os seus benefícios. Na Seção 6 apresentamos algumas considerações finais.

2 Contexto

As boas práticas propostas para publicação e utilização de dados na Web (DWBP) referem-se a conjuntos de dados, ou seja, "coleções de dados, publicados e gerenciados por um único agente, e disponíveis para serem acessados ou recuperados em um ou mais formatos" (MAALI; ERICKSON, 2014, Tradução Nossa). Por dados, "queremos dizer fatos conhecidos que podem ser registrados e que têm significado implícito" (ELMASRI; NAVATHE, 2010, Tradução Nossa).

Conforme descrito na Figura 1, os dados são publicados em diferentes distribuições, que são uma forma física específica de um conjunto de dados. Essas distribuições facilitam o compartilhamento de dados em larga escala, o que permite que conjuntos de dados possam ser utilizados por vários grupos de consumidores de dados sem levar em conta a finalidade, o público, o interesse ou a licença. Nesse contexto, um consumidor de dados pode ser "uma pessoa ou grupo que acessa, utiliza, e potencialmente executa tarefas de processamento nos dados" (DIANE; LEE; WANG, 1997, Tradução Nossa). Tendo em vista esta heterogeneidade e o fato de que os publicadores e os consumidores de dados podem não se conhecer, é necessário fornecer algumas informações sobre os conjuntos de dados e distribuições que contribuam para aumentar a confiança entre publicadores e consumidores, bem como a reutilização dos dados. Entre essas informações, destacam-se: metadados estruturais, metadados descritivos, acesso à informação, informação sobre a qualidade de dados, informações sobre a proveniência, informações sobre licença e informações sobre o uso.

Figura 1 – Contexto de publicação de dados na Web



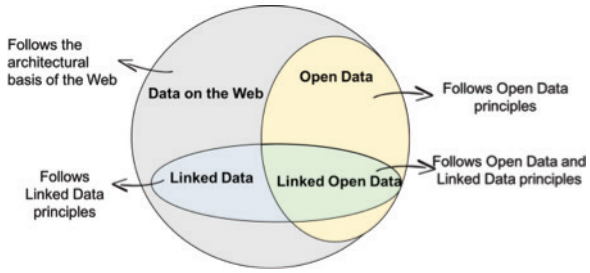
Fonte: DWBP (2016)

Por fim, uma questão importante sobre publicação e compartilhamento de dados na Web diz respeito à base arquitetônica da Web (JACOBS; WALSH, 2004). Um aspecto relevante desta arquitetura é o princípio de identificação, o qual menciona que URIs devem ser utilizados para identificar recursos. Em nosso contexto, um recurso pode ser um conjunto de dados ou um item específico de determinado conjunto de dados. Todos os recursos devem ser publicados com URIs estáveis, de modo que possam ser referenciados e permitam a criação de conexões, via URIs, entre dois ou mais recursos.

Nessa perspectiva, "Dados na Web" é um termo mais geral, que pode ser usado para denotar dados publicados de acordo com a base arquitetônica da Web (JACOBS; WALSH, 2004). Como ilustrado na Figura 2, dados na Web podem ser classificados como Dados Abertos (PIRES, 2015), Dados Conectados e Dados Abertos Conectados (BERNERS-LEE, 2009).

De acordo com o *Open Data Charter*, "dados abertos são dados digitais disponibilizados com as características técnicas e jurídicas necessárias para que possam ser utilizados livremente, reutilizados e redistribuídos por qualquer pessoa, a qualquer hora, em qualquer lugar". Considerando que a Web é o meio mais adequado para disponibilizar dados abertos, logo, dados abertos também são dados na Web. Outra distinção importante diz respeito aos dados publicados na Web de acordo com os Princípios dos Dados Conectados, em inglês Linked Data Principles (BERNERS-LEE, 2009). Uma parte dos dados atualmente disponíveis na Web segue esses princípios, sendo classificada como dados conectados. Finalmente, quando um conjunto de dados é publicado na Web seguindo ambos os princípios dos Dados Abertos e dos Dados Conectados, tais dados podem ser classificados como Dados Abertos Conectados.

Figura 2 – Dados na Web x Dados Abertos X Dados Conectados



Fonte: DWBP (2016)

É importante notar que nem todos os conjuntos de dados publicados na Web são compartilhados abertamente, ou seja, há uma grande parte dos dados publicados na Web que estão "fechados". Segurança, restrições comerciais e, acima de tudo, a privacidade dos indivíduos precisam ser levados em consideração para determinar em quais circunstâncias os dados devem ser publicados, assim como a sua licença de uso.

3 Casos de uso de dados na web

Há um crescente interesse na publicação e consumo de dados na Web. Tanto o governo como as organizações não governamentais disponibilizam uma variedade de dados na Web, alguns abertos, outros com restrições de acesso, abrangendo vários domínios, como Educação, Economia, Segurança, Patrimônio Cultural, eCommerce e Dados Científicos. Desenvolvedores, jornalistas, entre outros, manipulam esses dados para criar visualizações e realizar análises de dados. Porém, 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.

Com o intuito de identificar tais questões, o Grupo de Trabalho DWBP coletou e analisou diversos casos de uso. Cada caso de uso fornece uma descrição sobre a experiência de publicação e uso de dados na Web. Os casos de uso abordam diferentes domínios e ilustram alguns dos principais desafios enfrentados pelos publicadores e consumidores de dados. De acordo com os desafios estabelecidos a partir dos casos de uso, um conjunto de requisitos foi criado, de tal maneira que um requisito motivou a criação de uma ou mais boas práticas recomendadas.

O Quadro 1 mostra os principais desafios identificados a partir do documento de casos de uso [DWBPUCR], juntamente com os principais requisitos associados a cada desafio. É importante notar que um dos desafios – o de dados sensíveis – apresentado no documento Casos de Uso foi considerado fora do escopo do Grupo de Trabalho DWBP e, por isso, não está listado no Quadro 1. Um dos desafios originais, o de Data Usage (Uso de Dados), foi renomeado para Feedback. O desafio de Data Republication não foi identificado como parte da análise de casos de uso. No entanto, durante o desenvolvimento das boas práticas, várias questões relativas à republicação de dados foram discutidas e consideradas no escopo do documento DWBP. Tais problemas também correspondem a alguns dos requisitos previamente estabelecidos. Em algumas situações, o reuso de dados pode ser considerado como um outro modo de publicação de dados, sendo denominado de republicação. Isso acontece quando dados existentes são combinados com outros dados, criando aplicações Web ou visualizações, ou "reempacotando" os dados em um novo formato, como uma tradução.

Quadro 1 – Publicação de dados na Web: desafios e requisitos

Desafio: Acesso aos Dados (Data Access)
Tornar fácil o acesso aos dados na Web a fim de permitir que tanto humanos quanto máquinas aproveitem os benefícios do compartilhamento de dados utilizando a infraestrutura da Web.
Requisitos:
1. Os dados devem estar disponíveis para download em massa.
2. O nível de acesso aos dados deve ser fornecido juntamente com as condições de acesso, por exemplo, aberto, restrito ou fechado.
3. Quando os dados são produzidos em tempo real, devem estar disponíveis na Web em tempo real.
4. Os dados disponíveis devem ser atualizados, bem como seu ciclo de atualização deve ser explicitado.
5. Se os dados estão disponíveis por meio de uma API, a API deve ser documentada.

Desafio: Enriquecimento dos Dados (Data Enrichment)
Aperfeiçoar ou melhorar os dados brutos ou previamente processados para agregar valor aos dados.
Requisitos:
1. Deve ser possível executar algumas tarefas de enriquecimento de dados a fim de agregar valor aos dados, proporcionando maior valor para usuários de aplicações e serviços.

Desafio: Formato de Dados (Data Formats)
Escolher e disponibilizar dados em formatos que permitem o reuso.
Requisitos:
1. Informações sobre parâmetros de localidade (data e número de formatos, linguagem) devem ser disponibilizados.
2. Dados devem estar disponíveis em um formato legível por máquina que seja adequado para o seu uso pretendido ou potencial.
3. Dados devem estar disponíveis em múltiplos formatos.
4. Dados devem estar disponíveis em formato aberto.
5. Dados devem estar disponíveis em formatos padronizados. Com a utilização de dados em formatos padronizados, espera-se também a interoperabilidade.

Desafio: Identificação dos Dados (Data Identification)
Prover identificadores únicos para recursos de dados (conjuntos de dados ou registros que pertencem aos conjuntos de dados) disponíveis na Web.
Requisitos:
1. Cada recurso deve ser associado a um identificador único.

Desafio: Qualidade dos Dados (Data Quality)
Documentar a qualidade dos dados a fim de facilitar o processo de seleção de conjunto de dados e aumentar as chances de reutilização.
Requisitos:
1. Publicadores devem indicar se os dados estão parcialmente comprometidos ou se o conjunto de dados está incompleto.
2. Dados devem ser completos.
3. Dados devem ser associados a um conjunto de métricas de qualidade devidamente documentado, objetivo e, se possível, padronizado. Este conjunto de métricas de qualidade pode incluir definições dos usuários ou métricas específicas do domínio.
4. Opiniões subjetivas sobre a qualidade dos dados devem ser consideradas.
5. Dados disponíveis em diferentes níveis de granularidade devem ser acessíveis e modelados de uma maneira comum.

Desafio: Feedback
Coletar feedback de consumidores de dados e garantir que os dados publicados atendam às necessidades de consumo.
Requisitos:
1. Deve ser possível citar os dados na Web.
2. Deve ser possível rastrear o uso de dados.
3. Consumidores de dados devem ter uma maneira de compartilhar feedback e dados de classificação.

Desafio: Vocabulários de Dados (Data Vocabularies)
Aumentar a interoperabilidade e consenso entre os publicadores e consumidores de dados.
Requisitos:
1. Vocabulários devem ser claramente documentados.
2. Vocabulários devem ser compartilhados de forma aberta.
3. Vocabulários existentes devem ser utilizados quando possível.
4. Vocabulários devem incluir informações sobre versionamento.
5. Dados devem ser passíveis de comparação com outros conjuntos de dados.

Desafio: Licenças (Licenses)
Permitir que humanos compreendam informações sobre as licenças dos dados, descrevendo possíveis restrições de uso em uma determinada distribuição, e permitir que agentes de software possam detectar automaticamente a licença dos dados de uma distribuição.
Requisitos:
1. Dados devem ser associados a uma licença.

Fonte: Os autores.

4 Boas práticas para dados na web

As Boas Práticas para Dados na Web (LÓSCIO et al., 2016) foram desenvolvidas para incentivar e permitir a expansão continuada da Web como um meio para o intercâmbio de dados. O crescimento do compartilhamento on-line de dados abertos pelos governos em todo o mundo, o aumento da publicação de dados científicos na Web, a coleta, análise e publicação de dados de mídias sociais, o aumento crescente da publicação na Web de importantes acervos do patrimônio cultural, como da *Bibliotheque Nationale de France* e o crescimento do *Linked Open Data Cloud* (SCHMACHTENBERG et al., 2014), ilustram o crescimento no uso da Web como plataforma para publicação e compartilhamento de dados.

Em termos gerais, os publicadores de dados visam compartilhar dados abertamente ou com acesso controlado, enquanto os consumidores de dados (que também podem ser eles mesmos publicadores) buscam ser capazes de encontrar, usar e estabelecer conexões entre os dados, especialmente se os dados forem precisos, atualizados e tiverem garantia de alta disponibilidade. Isso cria uma necessidade fundamental para um entendimento comum entre os publicadores e os consumidores de dados. Sem esse acordo, os esforços dos publicadores podem ser incompatíveis com os anseios dos consumidores.

Neste contexto, torna-se crucial fornecer orientações aos publicadores que contribuam para a melhoria na forma como os dados são publicados. Espera-se que essa orientação promova a reutilização de dados e fomente a confiança nos dados por parte dos desenvolvedores, qualquer que seja a tecnologia que eles utilizem, aumentando o potencial de inovação genuína. O conjunto de Boas Práticas para Dados na Web (LÓSCIO et al., 2016) foram desenvolvidas para oferecer orientação técnica para a publicação de dados na Web, contribuindo para melhorar a relação entre publicadores e consumidores de dados.

As boas práticas abrangem diferentes desafios e exigências relacionadas com a publicação e o consumo de dados, como formatos de dados, acesso a dados identificadores de dados, vocabulários e metadados. Por um lado, cada boa prática lida com pelo menos um dos requisitos identificados no documento de casos de uso (LEE; LÓSCIO; ARCHER, 2015), de tal forma que a relevância da boa prática é evidenciada por esses requisitos. Por outro lado, cada requisito é abordado por pelo menos uma boa prática.

Conforme descrito em Lóscio et al. (2016) e ilustrado no Quadro 2, cada boa prática tem um Resultado esperado, que descreve "O que deve ser possível fazer quando um publicador de dados segue a boa prática". Em geral, o Resultado esperado é uma melhoria no modo que um consumidor de dados (humano ou software) pode manipular um conjunto de dados publicados na Web. Em alguns casos, o resultado esperado reflete uma melhoria no próprio conjunto de dados, o que também resultará em um ganho para o consumidor de dados.

Quadro 2 – Publicação de dados na Web: desafios e requisitos

BP1: Fornecer metadados
Os seres humanos serão capazes de compreender os metadados, e os agentes de software serão capazes de processá-los.
BP2: Fornecer metadados descritivos
Os seres humanos serão capazes de interpretar a natureza do conjunto de dados e suas distribuições, e os agentes de software serão capazes de descobrir automaticamente conjuntos de dados e distribuições.
BP3: Fornecer metadados estruturais
Os seres humanos serão capazes de interpretar o esquema de um conjunto de dados, e os agentes de software serão capazes de processar automaticamente os dados das distribuições.
BP4: Fornecer informações sobre a licença de dados
Os seres humanos serão capazes de compreender a licença de dados, descrevendo eventuais restrições impostas à utilização de certos dados, agentes de software serão capazes de detectar automaticamente a licença de dados de uma distribuição.
BP5: Fornecer informações de proveniência dos dados
Os seres humanos serão capazes de identificar a origem dos conjuntos de dados, e agentes de software serão capazes de processar automaticamente informações de proveniência.
BP6: Fornecer informação de qualidade de dados
Os seres humanos e os agentes de software serão capazes de avaliar a qualidade e, portanto, a adequação de um conjunto de dados para a sua aplicação.
BP7: Fornecer indicador de versão
Os seres humanos e os agentes de software poderão facilmente determinar qual versão de um conjunto de dados.

BP8: Fornecer o histórico de versões
Os seres humanos e os agentes de software serão capazes de entender como o conjunto de dados muda de versão para versão e como quaisquer duas versões específicas diferem.
BP9: Usar URIs persistentes como identificadores de conjuntos de dados
Os conjuntos de dados ou informações sobre conjuntos de dados poderão ser descobertos e citados ao longo do tempo, independentemente da sua disponibilidade ou do formato dos dados.
BP10: Usar URIs persistentes como identificadores dentro de conjuntos de dados
Os itens de dados serão relacionados em toda a Web criando um espaço global de informação acessível a humanos e máquinas.
BP11: Atribuir URIs para as versões dos conjuntos de dados e séries
Os seres humanos e os agentes de software serão capazes de referenciar versões específicas de um conjunto de dados, séries de conjunto de dados, bem como a versão mais recente de um conjunto de dados.
BP12: Usar formatos de dados padronizados legíveis por máquina
Máquinas serão capazes de ler e processar dados publicados na Web e os seres humanos serão capazes de usar ferramentas computacionais para manipular os dados.
BP13: Usar representações de dados que sejam independentes de localidade (*locale neutral)*
Os seres humanos e os agentes de software serão capazes de interpretar o significado do conjunto de caracteres (*strings*) que representam datas, horas, moedas, números, entre outros, com precisão.
BP14: Fornecer dados em vários formatos
Tantos usuários quanto possível serão capazes de utilizar os dados sem ter que transformá-los em seu formato preferido.
BP15: Reutilizar vocabulários, dando preferência aos padronizados
Interoperabilidade e consenso entre os publicadores e consumidores de dados serão reforçados.
BP16: Escolher o nível de formalização adequado
Os casos de aplicação mais prováveis serão apoiados com não mais complexidade do que o necessário.
BP17: Fornecer 'bulk download'
Transferências de arquivos grandes, ou seja, que exigem mais tempo do que um usuário típico consideraria razoável, serão possíveis por meio de protocolos de transferência de arquivos dedicados.

BP18: Fornecer subconjuntos para conjuntos de dados grandes
Os seres humanos e as aplicações serão capazes de acessar subconjuntos de um conjunto de dados, em vez de todo o conjunto. Isso proporcionará aos consumidores o acesso aos dados com uma elevada proporção de dados que são realmente necessários em comparação aos dados desnecessários. Conjuntos de dados estáticos considerados muito grandes poderão ser recuperados em porções menores. APIs podem ser usadas para filtrar os dados disponíveis. A granularidade de acesso aos dados poderá ser definida de acordo com as necessidades do domínio e as demandas de desempenho das aplicações.
BP19: Usar '*content negotiation*' para servir os dados disponíveis em vários formatos
Content negotiation permitirá que diferentes recursos ou representações diferentes de um mesmo recurso possam ser servidas de acordo com a requisição feita pelo cliente.
BP20: Fornecer acesso em tempo real
Aplicações serão capazes de acessar os dados em tempo real ou quase em tempo real; em tempo real significa um intervalo de milissegundos até alguns segundos após a criação de dados.
BP21: Fornecer dados atualizados
Os dados na Web serão atualizados em tempo hábil para que os dados disponíveis *on-line* reflitam os dados mais recentes divulgados através de qualquer outro canal. Quando novos dados estiverem disponíveis, logo que possível, serão publicados na Web.
BP22: Fornecer uma explicação para os dados que não estão disponíveis
Os consumidores saberão que os dados que são referenciados a partir do conjunto de dados não estão disponíveis ou se estão disponíveis sob diferentes condições.
BP23: Tornar os dados disponíveis através de uma API
Os desenvolvedores terão acesso aos dados para uso em seus próprios aplicativos, com dados atualizados e sem a necessidade de esforço por parte dos consumidores. As aplicações serão capazes de obter dados específicos por meio de consultas à API.
BP24: Usar padrões Web como base para construção de APIs
Desenvolvedores que tenham alguma experiência com APIs baseadas em padrões Web, tais como o REST, já deverão ter um conhecimento inicial de como usar a API. Além disso, será mais fácil dar manutenção na API.

BP26: Evitar alterações que afetem o funcionamento de sua API
O código do desenvolvedor deve continuar válido após alterações na API. Os desenvolvedores devem ser notificados das melhorias feitas na API e devem ser capazes de fazer uso delas. Quebrar alterações em sua API será raro e, se ocorrer, os desenvolvedores terão tempo e informações suficientes para adaptar o seu código, aumentando a confiança na API. Alterações na API deverão ser anunciadas no site da documentação da API.
BP27: Preservar identificadores
A URI de um conjunto de dados sempre levará para o conjunto propriamente dito ou então redirecionará para um recurso com informações sobre ele.
BP28: Avaliar a cobertura do conjunto de dados
Os usuários serão capazes de fazer uso de dados arquivados no futuro.
BP29: Coletar *feedback* dos consumidores de dados
Os consumidores de dados serão capazes de fornecer *feedback* e avaliações sobre conjuntos de dados e distribuições.
BP30: Compartilhar o *feedback* disponível
Os consumidores serão capazes de avaliar os tipos de erros que afetam o conjunto de dados, avaliar experiências de outros usuários e ter a certeza de que o publicador trata os problemas de forma adequada. Os consumidores também serão capazes de determinar se outros usuários já fizeram comentários semelhantes, poupando-lhes a submissão de relatórios desnecessários e poupando os publicadores de terem que lidar com duplicatas.
BP31: Enriquecer dados por meio da geração de novos dados
Os conjuntos de dados com valores nulos poderão ser "corrigidos" a partir do preenchimento de tais valores. Estrutura poderá ser conferida aos dados e sua utilidade poderá ser melhorada se forem adicionadas medidas ou atributos relevantes. Porém, tal adição só deverá ser feita se não alterar os resultados analíticos, o significado ou o poder estatístico dos dados.
BP32: Fornecer visualizações complementares
Complementar os conjuntos de dados com possíveis visualizações permitirá que os consumidores humanos tenham uma visão imediata sobre os dados, apresentando-os de forma que possam ser facilmente compreendidos.
BP33: Fornecer feedback para o publicador original
Uma melhor comunicação entre publicadores e consumidores fará com que seja mais fácil para os publicadores originais determinar como os dados que eles publicam estão sendo usados. Isso ajudará a justificar o investimento na publicação dos dados. Os publicadores também serão informados de medidas que podem ser tomadas para melhorar a qualidade dos seus dados.
BP34: Obedecer os termos de licença
Os publicadores serão capazes de confiar que seu trabalho está sendo reutilizado de acordo com os seus requisitos de licenciamento, tornando-os mais propensos a continuar com a publicação dos dados. Reutilizadores de dados vão ser capazes de licenciar adequadamente os trabalhos derivados a partir de dados previamente publicados.
BP35: Citar a publicação original do conjunto
Os consumidores finais serão capazes de avaliar a origem dos dados e os esforços dos publicadores originais serão reconhecidos. A cadeia de proveniência para os dados na Web será rastreável de volta ao seu publicador original.

Fonte: Os autores.

5 Benefícios das boas práticas para dados na web

A fim de incentivar os publicadores a adotar as boas práticas para publicação de dados na Web, uma série de benefícios que podem ser alcançados a partir da aplicação das boas práticas foram identificados, são eles: compreensibilidade; facilidade de processamento; facilidade de descoberta; reuso; confiança; capacidade de conexão de dados; facilidade de acesso; e interoperabilidade. Os benefícios são importantes porque ajudam publicadores de dados a ter uma melhor compreensão de "o que será possível" quando as boas práticas são adotadas. Tal como descrito nas subseções a seguir, cada benefício está associado a uma ou mais boas práticas. Por exemplo, a "compreensibilidade" está associada a dez boas práticas, que estão relacionadas a metadados, vocabulários de dados, *feedback* e enriquecimento de dados. Isto significa que se um publicador de dados adotar estas práticas, o nível de compreensibilidade aumentará, isto é, será possível para os seres humanos terem uma melhor compreensão sobre a estrutura e o significado dos dados, bem como a natureza do conjunto de dados. É importante notar que o benefício se torna mais forte na medida em que aumenta a adoção das boas práticas. Considerando que a publicação de

dados na Web é um processo incremental, o nível de cada benefício poderá aumentar após algumas iterações do processo de publicação de dados. A seguir, descrevemos cada um dos benefícios esperados com a adoção das boas práticas, juntamente com as boas práticas que contribuem para esse benefício.

5.1 Compreensibilidade

Os seres humanos terão uma melhor compreensão sobre a estrutura e o significado dos dados, bem como dos metadados e da natureza do conjunto de dados.

Boas práticas:

BP1: Fornecer metadados

BP2: Fornecer metadados descritivos

BP3: Fornecer metadados estruturais

BP5: Fornecer informações de proveniência de dados

BP13: Usar representações de dados que sejam independentes de localidade (locale neutral)

BP15: Reutilizar vocabulários, dando preferência aos padronizados

BP16: Escolher o nível de formalização adequado.

BP29: Coletar feedback dos consumidores de dados

BP31: Enriquecer dados através da geração de novos dados

BP32: Fornecer visualizações complementares

5.2 Facilidade de processamento

Máquinas ou agentes de *software* serão capazes de processar e manipular automaticamente os dados.

Boas práticas:

BP1: Fornecer metadados

BP3: Fornecer metadados estruturais

BP12: Usar formatos de dados padronizados legíveis por máquina

BP14: Fornecer dados em vários formatos

BP15: Reutilizar vocabulários, dando preferência aos padronizados

BP18: Fornecer subconjuntos para grandes conjuntos de dados

BP23: Tornar os dados disponíveis através de uma API

BP24: Usar padrões Web como base para construção de APIs

BP31: Enriquecer dados através da geração de novos dados

5.3 Facilidade de descoberta

Os agentes de *software* serão capazes de descobrir automaticamente um conjunto de dados ou dados dentro de um conjunto de dados.

Boas práticas:

BP1: Fornecer metadados

BP2: Fornecer metadados descritivos

BP9: Usar URIs persistentes como identificadores de conjuntos de dados

BP10: Usar URIs persistentes como identificadores dentro de conjuntos de dados

BP11: Atribuir URIs para as versões do conjunto de dados e séries

BP24: Usar padrões Web como base para construção de APIs

BP35: Citar a publicação original do conjunto de dados

5.4 Reuso

As chances de reutilização do conjunto de dados por diferentes grupos de consumidores de dados tende a aumentar.

Boas práticas:

Todas as 35 Boas Práticas

5.5 Confiança

A confiança que os consumidores têm no conjunto de dados tende a melhorar.

Boas práticas:

BP4: Fornecer informações de licença de dados

BP5: Fornecer informações de proveniência de dados

BP6: Fornecer informação de qualidade dos dados

BP7: Fornecer um indicador de versão

BP8: Fornecer histórico de versões

BP11: Atribuir URIs para as versões do conjunto de dados e séries

BP15: Reutilizar vocabulários, dando preferência aos padronizados

BP22: Fornecer uma explicação para os dados que não estão disponíveis

BP25: Fornecer a documentação completa para a sua API

BP26: Evitar alterações que afetem o funcionamento de sua API

BP27: Preservar identificadores

BP28: Avaliar a cobertura do conjunto de dados

BP29: Coletar feedback dos consumidores de dados

BP30: Compartilhar o feedback disponível

BP31: Enriquecer dados através da geração de novos dados

BP32: Fornecer visualizações complementares

BP33: Fornecer *feedback* para o publicador original

BP34: Obedecer aos termos de licenciamento

BP35: Citar a publicação original do conjunto de dados

5.6 Capacidade de conexão

Será possível criar ligações entre conjuntos de dados e itens de dados.

Boas práticas

BP9: Usar URIs persistentes como identificadores de conjuntos de dados

BP10: Usar URIs persistentes como identificadores dentro de conjuntos de dados

BP24: Usar padrões Web como base para construção de APIs

5.7 Facilidade de acesso

Os seres humanos e as máquinas serão capazes de acessar dados atualizados em uma variedade de formas.

Boas práticas:

BP17: Fornecer *download* em massa

BP18: Fornecer grandes subconjuntos para conjuntos de dados

BP19: Usar '*content negotiation*' para servir os dados disponíveis em vários formatos

BP20: Fornecer acesso em tempo real

BP21: Fornecer dados atualizados

BP23: Tornar os dados disponíveis através de uma API

BP24: Usar padrões Web como base para construção de APIs

BP32: Fornecer visualizações complementares

5.8 Interoperabilidade

Será mais fácil chegar a um consenso entre os publicadores e consumidores de dados.

Boas práticas:

BP9: Usar URIs persistentes como identificadores de conjuntos de dados

BP10: Usar URIs persistentes como identificadores dentro de conjuntos de dados

BP15: Reutilizar vocabulários, dando preferência aos padronizados

BP16: Escolher o nível de formalização adequado

BP23: Tornar os dados disponíveis através de uma API

BP24: Usar padrões Web como base para construção de APIs

BP26: Evitar alterações que afetem o funcionamento de sua API

BP33: Fornecer feedback para o publicador original

6 Considerações finais

Neste artigo, discutimos os principais desafios da publicação e do consumo de dados na Web, bem como os principais benefícios de utilizar boas práticas para publicação de dados na Web. Os desafios apresentados foram extraídos de situações reais de publicação e consumo de dados na Web.

A partir desses desafios foi elencado um conjunto de requisitos, os quais foram tratados na proposta de boas práticas para publicação de dados na Web. A necessidade de cada uma das boas práticas propostas é evidenciada pelos requisitos associados a cada um dos desafios. Finalmente, foram identificados alguns benefícios inerentes ao conjunto de boas práticas, a fim de prover meios que justifiquem o investimento na publicação adequada de dados na Web.

REFERÊNCIAS

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. Data on the Web best practices. W3C Working Draft, **World Wide Web Consortium** (W3C), May 2016. Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 30 ago. 2016.

Bibliothèque nationale de France. Reference information about authors, works, topics. Disponível em: <http://data.bnf.fr/>. Acesso em: 30 ago. 2016. Data on the Web Best Practices Working Group. Main Page. Disponível em: <https://www.w3.org/2013/dwbp/wiki/Main_Page>. Acesso em: 31 ago. 2016.

LEE, D.; LÓSCIO, B. F.; ARCHER, P. **Data on the Web Use Cases and Requirements**. W3C Working Group Note, World Wide Web Consortium (W3C), Feb. 2015. Disponível em: <https://www.w3.org/TR/2015/NOTE-dwbp-ucr-20150224/>. Acesso em: 30 ago. 2016.

MAALI, F.; ERICKSON, J. W3C. **Data Catalog Vocabulary (DCAT)**. 16 January 2014. W3C

Recommendation. Disponível em: <https://www.w3.org/TR/vocab-dcat/>. Acesso em: 31 ago. 2016.

JACOBS, I.; WALSH, N. W3C. **Architecture of the World Wide Web**, Volume One. 15 December 2004. W3C Recommendation. Disponível em: <https://www.w3.org/TR/Webarch/>. Acesso em: 31 ago. 2016.

International Open Data Charter. **Open Data Charter**. Disponível em: <http://opendatacharter.net/principles/>. Acesso em: 31 ago. 2016.

PIRES, M. T. **Open Data Guideline**. 2015. This Guideline is part of the cooperation project between São Paulo State Government and the UK Government. Disponível em: <http://ceWeb.br/guias/dados-abertos/en/>. Acesso em: 31 ago. 2016.

SCHMACHTENBERG, M. et al. **The Linking Open Data Cloud Diagram**. April, 2014. Disponível em: <http://lod-cloud.net/>. Acesso em: 30 ago. 2016.

ELMASRI, R.; NAVATHE, S.; ADDISON WESLEY, B. **Fundamentals of Database Systems**. 2010.

Research Data Alliance. Research Data Sharing Withtout Barriers. Disponível em: <http://rd-alliance.org>. Acesso em: 31 ago. 2016.

STRONG, D. M.; LEE, Y. W.; WANG, R. Y. Data quality in context. **Communications of the ACM**, v. 40, n. 5, p. 103-110, 1997.

BERNERS-LEE, T. **Linked Data**. 2009. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 30 ago. 2016.

World Wide Web Foundation. Open Data Barometer. Disponível em: <http://opendatabarometer.org>. Acesso: 31 ago. 2016.