Um *Survey* sobre Soluções para Publicação de Dados na Web sob a Perspectiva das Boas Práticas do W3C

Lairson Emanuel R. de Alencar Oliveira[1], Marcelo Iury S. Oliveira[1,2], Bernadette Farias Lóscio[1]

[1]Centro de Informática -– Universidade Federal de Pernambuco (UFPE)
Caixa Postal 50.740–560 -– Recife -– PE -– Brasil

[2]Unidade Acadêmica de Serra Talhada – Universidade Federal Rural de Pernambuco

{lerao, bfl, miso}@cin.ufpe.br, marcelo.iury@ufrpe.br

Abstract. *The growing interest in using the Web as platform for data sharing has motivated the development of data publishing solutions. These solutions support various tasks related to the Data on the Web lifecycle. However, since there are plenty of solutions, selecting the most suitable for a specific usage scenario is not a trivial task. This task becomes more complex given that significant subset of the most commonly used solutions are not specified or evaluated in scientific studies. In this work, it is presented a survey and analysis of Data on the Web publishing solutions with regards to the Data on The Web Best Practices recommended by the W3C.*

Resumo. *O crescente interesse no uso da Web para compartilhamento de dados tem motivado o desenvolvimento de soluções para publicação de dados. Essas soluções auxiliam em várias tarefas relacionadas ao ciclo de vida dos dados na Web. Entretanto, como há uma abundância de soluções, selecionar a mais apropriada para um cenário específico não é uma tarefa trivial. Esse trabalho se torna ainda mais complexo pelo fato de que uma parcela significativa das soluções mais usadas não estão especificadas ou avaliadas em trabalhos científicos. Neste artigo, é realizado um survey das soluções de publicação de dados na Web, apresentando um comparativo sob o ponto de vista do conjunto de Boas Práticas para Dados na Web recomendado pelo W3C.*

1. Introdução

Uma série de iniciativas vem sendo desenvolvidas em todo o mundo objetivando o compartilhamento e consumo dos dados na Web, dentre elas destacam-se a publicação de dados abertos (*Open Data*) [Dietrich et al. 2009] e de dados conectados (*Linked Data*) [Isotani and Bittencourt 2015]. Em conjunto com o crescente interesse nessas iniciativas, tem aumentado o interesse no desenvolvimento de soluções para publicação de dados a fim de facilitar tanto a publicação quanto o acesso, o consumo e a manutenção dos dados publicados [Lóscio et al. 2015].

A proposição de soluções para publicação de dados na Web tem sido alvo de vários estudos reportados na literatura [Milić et al. 2015, Goldacre and Gray 2016, Hoxha and Brahaj 2011], entretanto muitos trabalhos ainda estão em fase conceitual com poucas validações práticas. Em contraposição, também existem diversas soluções prontas

para uso e de propósito genérico, tais como CKAN, Socrata, OGDI, Junar e OpenDataSoft, as quais são aplicadas em diversas iniciativas de dados abertos e portais de dados abertos governamentais.

Essas soluções são bastante heterogêneas quanto a diversos aspectos relacionados à publicação e ao consumo de dados, tais como formatos de dados adotados para a publicação (*e.g.*, XML, JSON, CSV), o uso de APIs (do inglês, *Application Programming Interface*) para o acesso aos dados e o uso de vocabulários para descrição dos dados e metadados (*e.g.*, DCAT[1]). Como há uma abundância de soluções, selecionar qual a solução mais apropriada para um cenário especifico não é uma tarefa trivial.

Uma alternativa seria medir a qualidade das soluções. Entretanto, determinar a qualidade e a eficácia de uma solução também não é uma tarefa trivial devido aos diversos aspectos técnicos e humanos envolvidos na sua utilização. Diante dessa problemática, diversas áreas fazem uso de especificações ou recomendações técnicas com o objetivo tanto de guiar processos de avaliação de qualidade quanto reduzir os riscos inerentes em avaliações subjetivas. Exemplos clássicos de especificações de qualidade são os padrões da família ISO 9000[2] que tem sido usado em diversos casos desde avaliação de software até processos de negócio.

Um conjunto de critérios promissor para análise e avaliação das soluções de publicação de dados na Web é a recomendação do W3C intitulada *Data on the Web Best Practices* (DWBP) [Lóscio et al. 2016a]. Essa recomendação propõe um conjunto de boas práticas a serem implementadas pelos produtores de dados com o intuito de produzir conjuntos de dados que sejam úteis e atendam às expectativas dos consumidores de dados [Lóscio et al. 2016a]. No total, são 35 boas práticas que abordam diferentes aspectos relacionados à publicação e ao consumo de dados, tais como formatos de dados, acesso a dados, metadados e uso de identificadores. Ao seguir essas boas práticas, espera-se que uma série de benefícios distintos possam ser alcançados, tais como a compreensão, processabilidade, descoberta, reúso, acesso e interoperabilidade de dados [Lóscio et al. 2016a]. De modo geral, a recomendação DWBP não considera abordagens ou tecnologias específicas para publicação de dados, assim ele pode ser utilizado como modelo de referência para a medição de qualidade de soluções de publicação e consumo de dados na Web.

Diante desse cenário, neste artigo é apresentado um *survey* comparativo de soluções de publicação de dados na Web. Para guiar a realização do *survey*, foi utilizada a seguinte questão: *Quais são as soluções mais utilizadas atualmente para a publicação de dados na Web e como elas estão alinhadas com as Boas Práticas recomendadas pelo W3C?*. Além disso, foram seguidas quatro etapas bem definidas:(i) definição dos critérios de avaliação, (ii) seleção das soluções, (iii) extração dos dados e (iv) análise. Os critérios de avaliação são representados pelas boas práticas e, de maneira geral, busca-se avaliar se a solução de publicação atende, atende parcialmente ou não atende a respectiva boa prática. Para o *survey*, foram selecionadas sete soluções de publicação usadas em uma amostra representativa de portais de publicação de dados na Web.

Os resultados da análise demonstraram que existem boas práticas que são cobertas plenamente ou parcialmente por todas as soluções, como aquelas relacionadas aos metada-

[1]https://www.w3.org/TR/vocab-dcat/

[2]https://www.iso.org/standard/45481.html

dos, informações de proveniência, reúso de vocabulários, uso de representações complementares para os dados e na disponibilização de diferentes mecanismos de acesso como API e *bulk download*. No entanto, algumas boas práticas ainda são pouco ou não são atendidas pelas soluções, como aquelas relacionadas ao versionamento, atualização, qualidade e preservação dos dados.

O restante deste artigo está estruturado como se segue. Na Seção 2, são apresentados os trabalhos relacionados. Na Seção 3, é descrito o método de pesquisa adotado e as soluções encontradas. Na Seção 4, é apresentada a análise das soluções. Por fim, na Seção 5, é apresentada uma pequena discussão sobre o trabalho realizado e, na Seção 6, são apresentadas as conclusões e sugestões para os trabalhos futuros.

2. Trabalhos Relacionados

Dados na Web é um campo de pesquisa que tem sido alvo de muitos estudos reportados na literatura. No que diz respeito à avaliação e *surveys* de soluções para publicação de dados na Web, há trabalhos que fazem um levantamento de algumas soluções, apresentando suas características gerais.

Os estudos mais exaustivos foram realizados por [Stråle and Lindén 2014], [Lisowska 2016] e [Millette and Hosein 2016]. Apesar desses trabalhos serem os mais similares ao nosso em termos de representatividade de amostra de soluções e em relação à interseção de alguns critérios, como os metadados e aspectos técnicos, eles diferem em relação ao foco.

Stråle e Lindén (2014) realizaram uma avaliação comparativa das principais soluções existentes em relação a uma solução proprietária utilizada pelo governo da Suécia. Em outras palavras, levantaram as principais características do CKAN, Socrata e OpenDataSoft com o intuito de verificar quais aspectos a solução usada pelo governo Sueco poderia melhorar ou, alternativamente, substituir em sua solução. Millette e Hosein (2016) seguiram a mesma abordagem ao apresentar um levantamento de características do CKAN, Socrata e Junar e as compararam com uma nova solução conceitual com o foco no consumo de dados. Assim, eles identificaram as características gerais das soluções, verificando a disponibilidade de ferramentas para transformação e enriquecimento de dados, além da disponibilização e criação de APIs.

Por outro lado, o trabalho [Lisowska 2016] analisou as soluções sob a perspectiva da interoperabilidade dos metadados. Nesse trabalho, foram analisados 55 portais de dados abertos gerados a partir de diferentes soluções de publicação de dados com o intuito de testar a interoperabilidade de metadados. Assim, os autores analisaram a interoperabilidade e correspondência entre os vocabulários de representação de metadados (*e.g.*, DCAT) usados por CKAN, Socrata, OpenDataSoft, Junar, DKAN e ArcGis Open Data.

Esses trabalhos destacam a importância e o surgimento de soluções de publicação de dados como campo de pesquisa. No entanto, eles se concentram na análise de características gerais ou analisam uma quantidade pequena de soluções, geralmente mais populares ou citadas em outros artigos. Além disso, nenhum deles cobre a quantidade de soluções avaliadas neste artigo, assim como não consideram aspectos diversos dispostos nas Boas Práticas para Dados na Web do W3C.

3. Método de Pesquisa

O principal objetivo deste trabalho é apresentar um *survey* comparativo das soluções de publicação de dados na Web. Em geral, *surveys* permitem o levantamento de conceitos, modelos, características e tendências emergentes que aparecem de forma desconexa e esparsa na literatura. Visto que a temática deste trabalho envolve o uso de soluções mercadológicas, estas podem não ser cobertas por mapeamentos e revisões sistemáticas da literatura. Então, *surveys* são essenciais, porque possibilitam a descoberta de trabalhos oriundos de diversos veículos, incluindo artigos científicos e literatura cinzenta, tais como sites, manuais, relatórios técnicos e *white papers*.

O presente *survey* foi realizado em quatro etapas bem definidas:(i) definição dos critérios de avaliação, (ii) seleção das soluções, (iii) extração dos dados e (iv) análise. Enquanto a análise das soluções selecionadas é apresentada na Seção 4, as demais etapas são apresentas nas subseções a seguir.

3.1. Critérios de Avaliação

Neste trabalho, os critérios de avaliação das soluções de publicação de dados na Web foram construídos essencialmente a partir da recomendação de Boas Práticas para Dados na Web do W3C. Dessa forma, os critérios de avaliação são representados por cada boa prática e as soluções são avaliadas para verificar se atendem, atendem parcialmente, ou não atendem a respectiva boa prática. Na Tabela 1, é apresentado o conjunto de boas práticas.

Seguindo as boas práticas, alguns benefícios podem ser alcançados mais facilmente e as chances de reutilização dos conjuntos de dados tendem a melhorar. Assim, será possível uma melhor compreensão sobre a estrutura e o significado dos dados, bem como as máquinas ou agentes de software poderão descobrir, processar e manipular automaticamente os dados. Somado a isso, será possível acessar dados atualizados em uma variedade de formas, será mais fácil chegar a um consenso entre os produtores e consumidores de dados e a confiança dos consumidores nos conjunto de dados tende a melhorar [Lóscio et al. 2016a].

Para a realização do *survey*, algumas boas práticas foram desconsideradas, pois não dizem respeito às soluções em si. Este foi o caso das boas práticas 10, 16, 22 e 28, que estão diretamente relacionadas aos dados ao invés das soluções. Também foram desconsideradas as Boas Práticas 31, 33, 34 e 35, que se aplicam aos casos de republicação de dados, dependendo fortemente do consumidor ao invés da solução.

3.2. Seleção das Soluções

A seleção das soluções mostrou-se uma tarefa complexa, pois o presente trabalho teve como desafio cobrir uma amostra representativa de soluções usadas na prática para publicação de dados na Web. Contudo, grande parte das soluções existentes não estão documentadas em artigos científicos, impossibilitando, dessa forma, fazer uma busca utilizando métodos tradicionais de pesquisa como o uso de base de dados científicas (*e.g.*, ACM, IEEE ou Scopus).

Ao invés de procurar diretamente pelas soluções disponíveis na Web, optou-se por identificar as soluções de publicação de dados que estão sendo utilizadas em portais de dados disponíveis na Web. Essa abordagem permite identificar quais são as soluções usa-

Tabela 1. Boas práticas para dados na Web. Fonte: [Lóscio et al. 2016a]

ID	Descrição
BP1	Prover metadados
BP2	Prover metadados que possibilitem interpretar a natureza dos conjuntos de dados e suas distribuições
BP3	Prover metadados que descrevem o esquema e a estrutura interna de uma distribuição
BP4	Prover um link ou cópia do contrato de licença que controle o uso dos dados
BP5	Prover informações de proveniência dos dados (informações completas sobre a origem dos dados e as alterações que realizadas nos mesmos).
BP6	Prover informações sobre qualidade e adequação de dados para fins específicos
BP7	Prover um indicador de versão, que atribua e indique um número ou data de versão para cada conjunto de dados
BP8	Prover um histórico de versões completo que explique as mudanças feitas em cada versão.
BP9	Empregar URI persistentes como identificadores de conjuntos de dados
BP10	Empregar URI persistentes como identificadores dentro de conjuntos de dados
BP11	Atribuir URIs a versões individuais de conjuntos de dados, bem como à série de versões como um todo
BP12	Disponibilizar dados em um formato de dados padronizado e legível por máquina que seja adequado ao seu uso potencial ou pretendido
BP13	Empregar representações de dados neutras que não sejam influenciadas por aspetos culturais ou relacionados a localidade, por exemplo a formatação de datas e números. Quando isso não for possível, forneça metadados sobre as representações e formatações usadas pelos valores de dados.
BP14	Disponibilizar dados em vários formatos quando mais de um formato se adéqua ao seu uso potencial ou pretendido
BP15	Reusar termos de vocabulários, de preferência padronizados, para codificar dados e metadados
BP16	Optar por um nível de semântica formal que possa encaixar os dados e as aplicações mais prováveis
BP17	Prover um serviço de *download* em massa *(bulk download)* para permitir que usuários recuperem múltiplos conjuntos de dados com uma única solicitação
BP18	Prover opções para efetuar operações e realizar *download* de subconjuntos dos dados
BP19	Usar negociação de conteúdo para permitir o *download* dos dados em vários formatos
BP20	Disponibilizar dados em tempo real ou próximo de tempo real, quando os mesmos são produzidos em tempo real
BP21	Prover dados atualizados conforme frequência de atualização do mundo real
BP22	Prover uma explicação do porquê os dados não estão indisponíveis, como os dados podem ser acessados e quem pode acessar
BP23	Disponibilizar dados através de uma API
BP24	Empregar os Padrões da Web como base das APIs, tais como REST and SOAP
BP25	Prover documentação completa para a API
BP26	Evitar alterações na API para impedir quebra de código para quem já as está usando
BP27	Preservar o identificador e fornecer informações sobre o recurso arquivado
BP28	Avaliar a cobertura de um conjunto de dados antes da seu arquivamento
BP29	Prover um meio facilmente descoberto para que usuários ofereçam *feedback* a respeito dos dados
BP30	Tornar *feedback* dos usuários a respeito dos conjuntos de dados e distribuições publicamente disponíveis
BP31	Enriquecer seus dados, gerando novos dados ao fazê-lo, aumentará seu valor
BP32	Prover representações complementares para os dados, como visualizações, tabelas, aplicativos Web ou sínteses estatísticas
BP33	Prover *feedback* para o produtor original dos dados
BP34	Seguir os requisitos da licença original do conjunto de dados quando os mesmos estão sendo re-publicados
BP35	Citar a publicação original em seus metadados

das na prática, bem como aquelas que possuem nível de maturidade suficiente para serem analisadas.

Dessa forma, o levantamento dos portais foi realizado usando como base os sites DataPortals.org[3] e Opendatasoft.com[4]. Ambos os sites apresentam um catálogo de portais de dados de diversos países do mundo. Esses catálogos são reconhecidos como os mais abrangentes em relação a quantidade de portais indexados[5]. No total, identificam mais de 3300 portais de dados.

Contudo, esses catálogos apresentam alguns problemas de inconsistência, tais como portais inativos e redundância entre os portais indexados. Além disso, há portais

[3]http://dataportals.org/ Acesso em 10/05/2017

[4]https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/ Acesso em 10/05/2017

[5]http://okfnlabs.org/projects/dataportals-org/ Acesso em 10/05/2017

que utilizam soluções proprietárias que tornam difícil ou até impossível de serem avaliadas. Por esses motivos, foi realizada uma filtragem com base nos seguintes critérios de exclusão: (i) O portal está inativo; (ii) O portal é redundante (foi identificado previamente); (iii) O portal usa uma solução proprietária para publicação de dados; (iv) O portal é uma página de teste ou representa uma prova de conceito da solução; (v) O portal não faz uso de solução de publicação de dados. Aplicando os critérios de exclusão, restaram ainda 2400 portais para serem analisados.

A identificação da solução de publicação de dados foi realizada de forma manual, analisando individualmente cada portal, ou de forma automática na qual um *crawler* envia requisições HTTP para cada portal e analisa o cabeçalho das respostas recebidas em busca de *tokens* e palavras chaves que identifiquem a solução usada como plataforma de construção dos portais. Essa segunda abordagem, apesar de promissora, não possui uma boa precisão, pois não há uma API padronizada usada por todas as soluções, tornando, assim, difícil detectar soluções que ainda não são tão populares como CKAN ou Socrata.

Para aumentar a precisão da identificação de soluções, optou-se por uma abordagem mista, na qual parte dos portais foi analisada de forma automática e outra parte foi analisada manualmente. Entretanto, como o processo manual de checagem demanda tempo considerável para ser realizado, foi realizada uma amostragem aleatória dos portais para identificação manual das soluções de publicação adotadas pelos portais. Ao todo, foram checados 1091 portais (representando aproximadamente 45% da amostra total de portais).

Tabela 2. Soluções de Publicação de Dados na Web

Nome	**URL**	***Open Source?***	***Cloud computing?***
CKAN	http://ckan.org/	Sim	Não
Socrata	https://socrata.com/	Não	Sim
DKAN	http://getdkan.com/	Sim, mas tem uma versão paga	Não (gratuito) ou Sim (pago)
Junar	http://junar.com/	Não	Sim
ArcGIS Open Data	http://opendata.arcgis.com/	Não	Sim
OpenDataSoft	https://www.opendatasoft.com/	Não	Sim
Knoema	https://knoema.com/	Não	Sim

Após a aplicação das estratégia manual e automática de checagem, foram mapeadas 15 soluções. Entretanto, foram selecionadas 7 soluções que estão sendo usadas em pelo menos 10 portais atualmente. As soluções selecionadas são apresentadas na Tabela 2. É importante ressaltar que o propósito deste artigo não foi realizar uma análise exaustiva de todas as soluções existentes. Mas, identificar as principais soluções usadas na prática em portais de dados disponíveis na Web.

3.3. Extração dos Dados

Nesta fase, buscou-se obter informações das soluções de publicação para responder os critérios de avaliação previamente definidos. Os dados usados para avaliação das soluções foram coletados através de documentação oficial (*e.g.*, manuais, diagramas arquiteturais, *white papers*) e ambientes de demonstração providos pelas instituições responsáveis pelas soluções. A consolidação dos dados extraídos envolveu interpretação subjetiva. Contudo, com o intuito de mitigar o risco de má interpretação, os dados foram triangulados a partir de outras fontes, tais como sites e fóruns de usuários, vídeos, apresentações, páginas de divulgação da solução. Além disso, foram realizadas reuniões para se alcançar o consenso da interpretação dos dados.

4. Análise

Nesta seção, são apresentados os resultados da análise das soluções mais utilizadas para a publicação de dados na Web sob a perspectiva das Boas Práticas para Dados na Web. É importante ressaltar que, ao invés de considerar os conjuntos de dados ou portais de dados específicos como evidências, foram analisadas as soluções adotadas pelos portais para verificar se elas implementam as boas práticas. O resultado da análise é apresentado na Tabela 3 é discutido a seguir.

Tabela 3. Soluções para Publicação de Dados X Boas Práticas. Fonte: os autores

Boa Prática	CKAN	Socrata	DKAN	Junar	ArcGIS O. D.	OpenDataSoft	Knoema
BP1	A	A	A	A	A	A	A
BP2	A	A	A	A	A	A	A
BP3	AP	A	-	AP	A	-	AP
BP4	A	A	A	-	AP	AP	-
BP5	A	A	A	-	AP	AP	A
BP6	-	A	-	-	-	-	-
BP7	A	-	A	-	-	-	AP
BP8	-	AP	A	-	-	-	-
BP9	A	A	A	A	AP	A	AP
BP11	-	A	-	-	-	-	-
BP12	A	A	A	A	A	A	A
BP13	-	-	-	-	-	AP	A
BP14	-	A	-	A	A	A	A
BP15	A	A	A	A	AP	A	AP
BP17	A	A	A	A	A	A	A
BP18	A	A	A	A	A	-	A
BP19	-	A	-	A	A	A	-
BP20	-	-	-	-	A	-	-
BP21	-	AP	-	A	-	-	AP
BP23	AP	A	A	A	A	A	-
BP24	A	A	A	A	A	A	A
BP25	A	A	A	A	A	A	A
BP26	A	A	A	A	A	A	A
BP27	-	-	-	-	-	-	-
BP29	A	A	A	-	-	-	A
BP30	A	A	-	-	-	A	-
BP32	A	A	A	A	A	A	A

Legenda: A = Atende; AP = Atende Parcialmente; - = Não atende.

Os metadados ajudam os usuários a compreenderem o significado dos dados, sua estrutura, licença, instituição que produziu os dados, métodos de acesso e agendamento de futuras atualizações dos conjuntos de dados. Nesse contexto, todas as soluções fornecem metadados de uma forma geral (BP1), assim como fornecem metadados descritivos que ajudam na interpretação da natureza dos conjuntos de dados e suas distribuições (BP2). Com relação ao fornecimento de metadados estruturais (BP3), o Socrata e o ArcGIS Open Data atendem plenamente essa boa prática, sendo destacado o Socrata por fornecer mecanismos para interpretar automaticamente os metadados estruturais. Algumas soluções oferecem apenas uma cobertura parcial ou não disponibilizam os metadados igualmente em formatos legíveis por humanos e por máquinas. Dessa forma, o CKAN, o Knoema e o Junar implementam a BP3 parcialmente, pois não disponibilizam uma versão legível por humanos para os metadados estruturais, ou seja, não disponibilizam metadados que descrevam o esquema e a estrutura interna dos conjuntos de dados.

Dado que podem existir restrições quanto ao compartilhamento e reutilização dos dados, é importante informar a licença dos conjuntos de dados (BP4). Somado a isso, é

importante informar a origem dos dados (BP5), assim como informar metadados de qualidade (BP6). Em relação a disponibilização de informações acerca das licenças do conjunto de dados, apenas o Knoema não fornece esse tipo de informação. Já em relação a informações de proveniência, todas as soluções atendem parcialmente essa boa prática ao fornecer informações a respeito das entidades responsáveis pela produção e manutenção dos dados. No entanto, o ARCGIS e o OpenDataSoft implementam a BP4 e a BP5 de forma parcial por não oferecer uma maneira para representar os metadados de licença e proveniência legíveis por máquina. Finalmente, apenas o Socrata fornece informações a respeito da qualidade dos conjuntos de dados (BP6).

Como os conjuntos de dados também podem mudar ao longo do tempo, são consideradas boas práticas atribuir e indicar o número de versão ou data para cada conjunto de dados (BP7), assim como manter um histórico de todas as versões geradas (BP8). Apenas o CKAN, DKAN e Knoema apresentam metadados que indicam a versão do conjunto de dados. Enquanto o Knoema faz o registro das alterações em nível de dados, o CKAN e DKAN fazem o controle do conjunto de dados como um todo. Em relação a BP8, apenas o Socrata atende parcialmente essa boa prática ao permitir a listagem do histórico de versões, contudo essa funcionalidade não é acessível por API.

Identificadores são amplamente utilizados nos sistemas de informação com o intuito de se identificar sem ambiguidades e independente de localização um determinado recurso. São elencadas como boas práticas o uso de URIs persistentes como identificadores para conjuntos de dados ou para versões individuais (BP9) e para a série de versões como um todo (BP11). Todas as soluções fazem uso de URIs persistentes para identificação dos conjuntos de dados. No entanto, o Knoema atende parcialmente a BP9. Apesar de fazer uso de URIs para identificar os conjuntos de dados, no Knoema, as URIs são geradas através de funções hash[6] não permitindo, assim, que humanos possam reconhecer o conjunto de dados associado a um determinado conjunto de dados. No mais, somente o Socrata realiza a atribuição de URIs para as versões dos conjuntos de dados.

O formato no qual os conjuntos de dados são disponibilizados é fundamental para tornar os dados utilizáveis para os consumidores. Dessa forma, é considerada uma boa prática disponibilizar os dados em formato legível por máquina (BP12), como CSV, XML, RDF e outros. Também é aconselhado usar representações de dados que não sejam influenciadas por localidades ou, quando não for possível, fornecer metadados sobre a localidade ou idioma usado pelos conjuntos de dados (BP13). Todas as soluções permitem o uso de formatos legíveis por máquina para a distribuição dos conjuntos de dados. Por fim, apenas o OpenDataSoft e Knoema atendem a BP13, ainda que parcialmente, pois oferecem uma maneira de representar os metadados de localidade ou idioma para acesso legível por máquina.

Além disso, também constatou-se que as soluções fornecem opções para fornecimento de dados em diferentes formatos (BP14). É importante frisar que os formatos de dados podem depender do formato que o produtor deseja publicar, mas para análise foi considerado se a solução apresentava limitações quanto aos principais formatos usados nesse contexto.

Visando uma maior interoperabilidade e consenso entre os produtores de dados e

[6]https://en.wikipedia.org/wiki/Cryptographic$_{h}ash_{f}unction$

os consumidores, é considerado uma boa prática o reúso de vocabulários (BP15), dando preferência aos padronizados. Os vocabulários definem os conceitos e atributos utilizados para descrever e representar uma área de interesse, tais como o vocabulário DCAT que é utilizado para expressar os metadados relacionados aos conjuntos de dados. Praticamente todas as soluções fazem o uso de vocabulários compartilhados para padronização de metadados. Esses vocabulários tem sido utilizados especialmente para padronização de metadados descritivos dos conjuntos de dados. No mais, todas as soluções analisadas são compatíveis com o vocabulário DCAT, indicando que seus metadados são cobertos por ele.

Por padrão, a Web oferece acesso aos dados através de métodos HTTP, o que permite acesso em um nível de transação atômica. Com o objetivo de oferecer mais opções de acesso, é recomendado permitir o *download* em massa de múltiplos conjuntos de dados com uma única solicitação (BP17) e, em caso de conjuntos de dados grandes, pode-se oferecer opções por meio de APIs para efetuar operações e recuperar subconjuntos dos dados (BP18). Todas as soluções permitem o *download* em massa de múltiplos conjuntos. Já em relação a BP18, todas as soluções permitem acesso a subconjuntos dos dados por meio da API.

Ainda em relação à recuperação de conjuntos de dados, também é considerado uma boa prática o uso da negociação de conteúdo para permitir a recuperação dos conjuntos de dados em vários formatos (BP19). Por exemplo, a partir de uma mesma URI, usando a negociação de conteúdo, podem-se obter os dados em JSON, XML ou CSV. Exceto pelo Knoema, CKAN e DKAN, as demais ferramentas permitem recuperar os dados em diferentes formatos do que foi publicado, ou seja, fazem uso da negociação de conteúdo para pré-processamento interno ou para disponibilizar os dados em formatos distintos.

Nenhuma das soluções apresenta alternativa para tratar dados em tempo real (BP20). Todavia, vale salientar que o ArcGIS Open Data é o único que apresenta referência quanto a essa boa prática, indicando que ao ser integrado ao *ArcGIS for Server*, pode realizar a publicação de dados geográficos em tempo real para consumo.

Quando os dados não são produzidos em tempo real, é considerado uma boa prática fornecer os dados atualizados, deixando a frequência de atualização explícita (BP21). O Socrata e o Knoema apenas apresentam a informação da periodicidade de atualização nos metadados, mas a atualização não é realizada de forma automática. No entanto, a atualização no Socrata pode ser garantida por meio de uma aplicação adicional (DataSync[7]), mas que, ainda sim, não extrai os dados a serem publicados e necessita de integração com outras ferramentas. Contrapondo-se, o Junar permite que a frequência de atualização seja seguida quando os dados são consumidos por meio de uma API. Apesar de não apresentar a informação de quando os dados serão atualizados para os consumidores, por meio dos metadados, é apresentada a data da última consulta que foi realizada a atualização.

O uso de APIs para acesso a dados (BP23) é um consenso entre as soluções. Uma API oferece uma maior flexibilidade e capacidade de processamento para os consumidores de dados, permitindo também o uso de dados em tempo real e realização de filtros. Além disso, todas as soluções fazem uso de padrões da Web (BP24), tais como REST, na construção das APIs, assim como também fornecem uma documentação da API (BP25).

[7]https://github.com/socrata/datasync

Por fim, vale salientar que, até o momento, as soluções tem evitado alterações na API para não quebrar o código de quem está utilizando-a (BP26).

Tendo em vista que é provável que os produtores podem remover os dados da Web, é importante preservar seus identificadores. Assim, espera-se que seja possível dereferenciar o URI de um conjunto de dados mesmo ele não estando disponível e fornecer informações sobre o seu arquivamento (BP27). Foi observado que nenhuma das soluções permite a preservação dos identificadores dos conjuntos de dados. Quando um conjunto de dados não está mais disponível nas soluções, é retornada apenas uma mensagem de erro 404 (*NOT FOUND*). Logo, os identificadores não são mantidos e preservados nas soluções.

Com os dados publicados na Web, os consumidores podem acessá-los e criar suas próprias experiências. No entanto, os produtores de dados muitas vezes não tem o *feedback* dos consumidores sobre como os conjuntos de dados são usados e não oferecem maneiras eficazes para discutir essas experiências [Lóscio et al. 2016b]. São boas práticas o fornecimento de pelo menos um mecanismo para receber *feedback* (BP29) e a disponibilização dos *feedbacks* já coletados ao público (BP30). De forma geral, as soluções CKAN, Socrata, DKAN, Knoema atendem superficialmente a BP29. As mesmas disponibilizam opções muito simples para receber *feedback*, como formulários de contato. Além disso, apenas as soluções CKAN, Socrata e OpenDataSoft permitem a visualização de *feedback* emitidos por outros usuários.

Outro ponto bastante explorado pelas soluções é a possibilidade de gerar visualizações complementares de forma automática (BP32), como gráficos, tabelas interativas e sínteses estatísticas. Nesse ponto, é possível destacar o Knoema como a solução que mais apresenta tipos de visualizações distintas e possibilidades de filtros.

5. Discussão

As boas práticas apresentadas na Seção 3.1 auxiliam o entendimento do cenário de compartilhamento de dados na Web, uma vez que abordam os principais desafios a serem enfrentados na publicação e no consumo de dados na Web. É possível observar que a Web possibilita o compartilhamento de dados, *i.e.*, o acesso e a leitura de dados, de maneira bastante simples, sem a exigência de sistemas para controlar o acesso concorrente aos dados. Porém, a facilidade de compartilhamento em grande escala requer soluções flexíveis que possibilitem o consumo de dados por grupos de usuários previamente desconhecidos [Lóscio et al. 2016a].

De forma geral, foi verificado que as soluções existentes tem um foco maior na publicação dos dados, *i.e.*, no processo de disponibilização dos conjuntos de dados para uso na Web, compreendendo também os metadados, por meio de um catálogo. Contudo, constatou-se que nenhuma das soluções permite a extração e publicação de conjuntos de dados a partir de outras fontes de dados. Um exemplo seria a publicação de um conjunto de dados criado a partir um banco de dados hospedado em um Sistema de Gerenciamento de Banco de Dados. Esta funcionalidade permitiria garantir que os dados publicados na Web estejam sincronizados com a fonte de origem. Assim, à medida que ocorrerem atualizações nas fontes de origem, os dados publicados na Web também devem ser atualizados. Além disso, para possibilitar a publicação dos conjuntos de dados e garantir que estarão atualizados conforme sua fonte de dados de origem, as soluções deveriam prover mecanismos que permitam realizar a extração e transformação dos dados. De modo geral, deve ser possível

extrair os dados automaticamente a partir de diferentes fontes de dados e, posteriormente, realizar a transformação dos dados de origem para publicá-los na Web.

Como mencionado previamente, os conjuntos de dados podem ser alterados ao longo do tempo, seja para realizar uma melhoria, uma correção ou até mesmo a inserção de novos dados. Dessa forma, é importante manter versões dos conjuntos de dados e possibilitar o acesso as mesmas. No entanto, são poucos os meios disponibilizados pelas soluções analisadas para garantir um gerenciamento de versões adequado. Por exemplo, algumas soluções estão de acordo com a BP7 por apresentar um indicador de versão para cada conjunto de dados, mas elas não proporcionam ferramentas que permitam o gerenciamento das versões.

Somado a isso, todas as soluções de publicação de dados analisadas utilizam apenas um subconjunto do DCAT para coletar informações (metadados) sobre os conjuntos de dados e distribuições. Foi também identificado uma carência de maior suporte para coletar informações sobre qualidade e versionamento, além de prover o processo de curadoria de informações relacionadas aos consumidores e aplicações que fazem uso dos dados. Também é importante que as soluções adotem mais mecanismos para coleta de *feedback*, uma vez que boa parte só disponibiliza um simples formulário de contato. Além disso, nenhuma das soluções analisadas garante a preservação dos conjuntos de dados ao longo do tempo, ou seja, permitem a exclusão dos dados, mas não mantém os identificadores ativos para que os consumidores obtenham informações acerca do conjunto de dados.

Portanto, a fim de que as soluções estudadas neste artigo possam prover o compartilhamento de dados na Web de forma mais adequada, é importante que sejam resolvidas questões como as relacionadas à qualidade dos dados, *feedback*, versionamento, atualização e preservação dos dados, que são desafios até então pouco explorados.

6. Conclusão e Trabalhos Futuros

A quantidade de soluções para publicação de dados na Web tem crescido nos últimos anos. Todavia, por apresentarem diversidade de requisitos e de características, encontrar a mais apropriada para um cenário específico não é uma tarefa trivial. Essa tarefa torna-se ainda mais complexa uma vez que boa parte das soluções usadas não são especificadas ou avaliadas em trabalhos científicos. A fim de suprir essa lacuna, foi realizado um levantamento e análise das principais soluções de publicação de dados na Web.

Neste artigo, é apresentado um *survey* comparativo das soluções de publicação de dados sob o ponto de vista do conjunto de Boas Práticas para Dados na Web recomendado pelo W3C. O resultado do levantamento permitiu detectar 15 soluções *Open Source* ou comerciais, das quais 7 foram descritas e analisadas por serem utilizadas em pelo menos 10 portais que estão em uso atualmente. Como resultado da análise, foi identificado que as soluções analisadas atendem aos principais desafios de compartilhamento de dados na Web que foram descritos pelo W3C. De modo geral, a análise mostrou que existem limitações e lacunas nas soluções, como a carência de ferramentas que possibilitem o gerenciamento adequado de versões e a preservação dos conjuntos de dados, por exemplo.

Como trabalho futuro, pretende-se aumentar o número de soluções analisadas. Somado a isso, também pretende-se buscar um entendimento mais amplo dos principais desafios para publicação e consumo dos dados na Web. A partir desse entendimento, será

possível elencar os requisitos esperados de uma solução de publicação de dados na Web a fim de possibilitar o gerenciamento adequado dos conjuntos de dados publicados.

7. Agradecimentos

Este trabalho foi parcialmente apoiado pelas instituições de financiamento brasileiras Fundação de Amparo a Ciência e Tecnologia de Pernambuco (FACEPE) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Marcelo Iury e Lairson Emanuel são bolsistas do CNPq na modalidade doutorado. Marcelo Iury também agradece a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa modalidade doutorado sanduíche. Os autores também gostariam de agradecer aos seus colegas do projeto Aladin por sua contribuição para este artigo.

Referências

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J., and Zijlstra, T. (2009). Open data handbook. http://opendatahandbook.org/guide/en/. Acesso em 05 de dezembro de 2016.

Goldacre, B. and Gray, J. (2016). Opentrials: towards a collaborative open database of all available information on all clinical trials. *Trials*, 17(1):164.

Hoxha, J. and Brahaj, A. (2011). Open government data on the web: A semantic approach. In *Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on*, pages 107–113. IEEE.

Isotani, S. and Bittencourt, I. (2015). *Dados Abertos Conectados: Em busca da Web do Conhecimento*. NOVATEC.

Lisowska, B. (2016). Metadata for the open data portals. Technical report, Joined-up Data Standards. http://juds.joinedupdata.org/wp-content/uploads/2016/12/JUDS-DP6-Metadata-for-the-open-data-portals.pdf. Acesso em 21/05/2017.

Lóscio, B. F., Burle, C., and Calegari, N. (2016a). Data on the Web Best Practices. W3C Recommendation, World Wide Web Consortium (W3C). https://www.w3.org/TR/dwbp/.

Lóscio, B. F., Burle, C., and Calegari, N. (2016b). Data on the web best practices: Challenges and benefits. In *Open Data Reserach Symposium (ODRS 2016)*.

Lóscio, B. F., Oliveira, M. I. S., and Bittencourt, I. I. (2015). Publicação e Consumo de Dados na Web: Conceitos e Desafios. *Tópicos em Gerenciamento de Dados e Informações (Mini Cursos - SBBD 2015)*, d:39–69.

Milić, P., Veljković, N., and Stoimenov, L. (2015). Linked relations architecture for production and consumption of linksets in open government data. In *Conference on e-Business, e-Services and e-Society*, pages 212–222. Springer.

Millette, C. and Hosein, P. (2016). A consumer focused open data platform. In *Big Data and Smart City (ICBDSC), 2016 3rd MEC International Conference on*, pages 1–6. IEEE.

Stråle, J. and Lindén, H. (2014). An evaluation of platforms for open government data. http://kth.diva-portal.org/smash/get/diva2:723341/FULLTEXT01.pdf. Acesso em 20/05/2017.