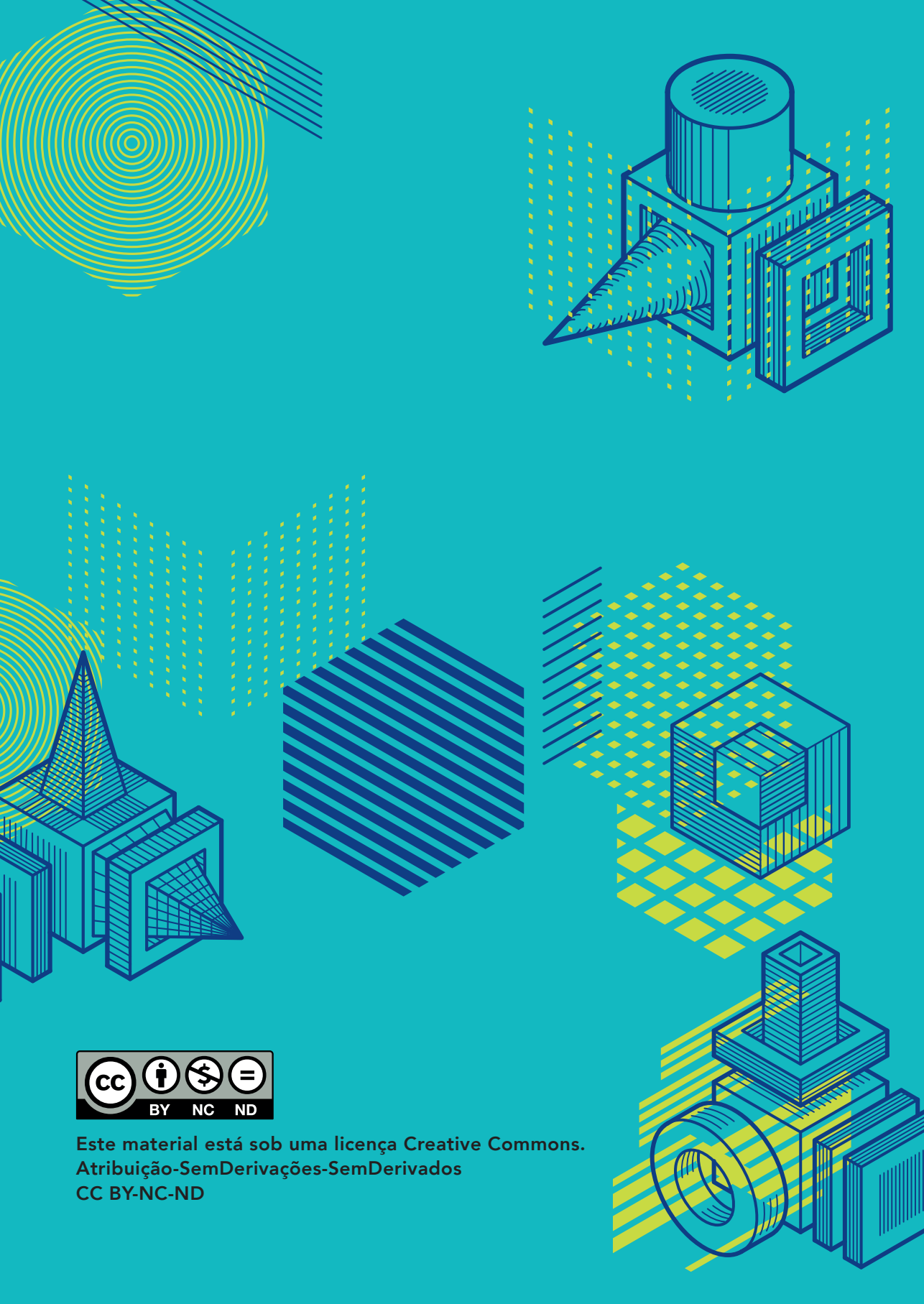
FUNDAMENTOS PARA PUBLICAÇÃO DE DADOS NA WEB

cewebbr nicbr cgibr

TRUST
THE TRUST FOR THE AMERICAS

OEA
Mais direitos para mais pessoas



FUNDAMENTOS PARA PUBLICAÇÃO DE DADOS NA WEB

Bernadette Farias Lóscio (UFPE)
Caroline Burle (Ceweb.br/NIC.br)
Marcelo Iury S. Oliveira (UFRPE)
Newton Calegari (Ceweb/NIC.br)

CGI.br
Comitê Gestor da
Internet no Brasil
2018

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

Fundamentos para publicação de dados na web /Bernadette F. Lóscio ... [et al.] ; organização ; Beatriz Rossi Corrales ; coordenação Vagner Diniz ; Núcleo de Informação e Coordenação do Ponto BR [Autor corporativo]. -- São Paulo : Comitê Gestor da Internet no Brasil, 2018.

Outros autores: Caroline Burle, Marcelo Iury S. Oliveira, Newton Calegari.

Bibliografia

ISBN 978-85-5559-072-6

1. Ciência da computação 2. Dados abertos 3. Web (Publicação) I. Lóscio, Bernadette F. II. Burle, Caroline III. Oliveira, Marcelo Iury S. IV. Calegari, Newton. V. Corrales, Beatriz Rossi. VI. Diniz, Vagner. VII. Núcleo de Informação e Coordenação do Ponto BR.

18-20750 CDD-004.6072081

Índices para catálogo sistemático:
1. Publicação de dados na Web : Tecnologia da informação e comunicação : Comportamento de uso 004.6072081

Esta publicação está disponível também em formato digital em www.ceweb.br.

Este material foi desenvolvido pelo **Centro de Estudos sobre Tecnologias Web** do **Núcleo de Informação e Coordenação do Ponto BR (Ceweb.br/NIC.br)** no marco do projeto "**Do Governo Aberto ao Estado Aberto**", executado por **The Trust for the Americas, a Organização dos Estados Americanos (OEA)** com financiamento da Embaixada dos Estados Unidos em São José, Costa Rica

ceweb.br nic.br cgi.br

AUTORES

Bernadette Farias Lóscio
Centro de Informática – Universidade Federal de Pernambuco (UFPE)
bfl@cin.ufpe.br

Caroline Burle
Centro de Estudos sobre Tecnologias na Web (Ceweb.br)
Núcleo de Informação e Coordenação do Ponto Br (NIC.br)
cburle@nic.br

Marcelo Iury S. Oliveira
Unidade Acadêmica de Serra Talhada – Universidade Federal Rural de Pernambuco (UFRPE)
marcelo.iury@ufrpe.br

Newton Calegari
Centro de Estudos sobre Tecnologias na Web (Ceweb.br)
Núcleo de Informação e Coordenação do Ponto Br (NIC.br)
newton@nic.br

SUMÁRIO

11 INTRODUÇÃO

12 DADOS ABERTOS

19 DADOS CONECTADOS

23 DADOS NA WEB

27 CICLO DE VIDA DOS DADOS NA WEB

31 BOAS PRÁTICAS PARA DADOS NA WEB

47 TÉCNICAS PARA PUBLICAÇÃO DE DADOS NA WEB

51 CONCLUSÃO

53 REFERÊNCIAS

56 ANEXO: ROADMAP DE PUBLICAÇÃO DE DADOS ABERTOS

INTRODUÇÃO

Desde o seu surgimento, a Web tem se destacado como um importante meio para a troca e compartilhamento de informações. Nesse cenário de grande quantidade de dados disponíveis na Web dois papéis merecem destaque: os provedores e os consumidores de dados. Em termos gerais, os provedores de dados visam a publicação e o compartilhamento de dados, com acesso livre ou controlado, enquanto os consumidores de dados (que também podem ser eles mesmos provedores) desejam fazer uso destes dados para a geração de informações úteis e relevantes, bem como para a geração de novos dados.

É importante ressaltar que o interesse na publicação de dados na Web não é algo novo (BERNERS-LEE; CONNOLLY; SWICK, 1999 e ABITEBOUL; BUNEMAN; SUCIU, 2000). Porém, nos últimos anos, este interesse tem se caracterizado pela publicação de dados de maneira a promover o compartilhamento e a reutilização de dados. Dessa forma, apenas disponibilizar o acesso aos dados não é suficiente. De maneira geral, torna-se necessário publicar dados de forma que possam ser prontamente compreendidos e utilizados por consumidores, além da disponibilização dos dados em formatos que possam ser facilmente processados por aplicações. Além disso, fatores como a heterogeneidade dos dados e a falta de padrões para descrição e acesso aos conjuntos de dados, tornam o processo de publicação, compartilhamento e consumo de dados uma tarefa complexa. Neste contexto, esta apostila discute os fundamentos relacionados à publicação de dados na Web, abordando aspectos relevantes, incluindo: os conceitos de Dados Abertos, Dados Conectados (do inglês *Linked Data*), o Ciclo de Vida dos Dados na Web e as Boas Práticas para Dados na Web.

DADOS ABERTOS

Segundo a Open Knowledge International (OPEN KNOWLEDGE, 2012), Dado Aberto é qualquer dado que pode ser livremente utilizado, reutilizado e redistribuído por qualquer um. Assim, dados abertos consistem na publicação e disseminação de informações na Internet, compartilhadas em formatos abertos, legíveis por máquinas, e que possam ser livremente reutilizadas de forma automatizada pela sociedade. Assim, a abertura de dados está interessada em evitar um mecanismo de controle e restrições sobre os dados que forem publicados, permitindo que tanto pessoas físicas quanto jurídicas possam explorar estes dados de forma livre (ISOTANI; BITTENCOURT, 2015). Um dado é considerado aberto quando apresenta as seguintes características (OPEN KNOWLEDGE, 2012):

I. Disponibilidade e acesso: o dado precisa estar disponível por inteiro. Deve estar num formato conveniente e modificável;
II. Reúso e redistribuição: o dado precisa ser fornecido em condições de reúso e redistribuição podendo ser combinado com outros;
III. Participação universal: todos podem usar, reusar e redistribuir o dado sem restrições de áreas, pessoas ou grupos.

Os dados abertos podem ser classificados de acordo com uma escala, baseada em estrelas, proposta por Tim Berners-Lee (BERNERS-LEE, 2006). Segundo essa classificação, apresentada na Figura 1, um dado publicado na Web em qualquer formato (imagem, tabela ou documento) e associado a uma licença que permita o seu uso e reúso sem restrições é avaliado como sendo *1 Estrela*. Apesar de já ser um avanço, os dados com *1 Estrela* precisam ser manipulados manualmente ou por meio de extratores construídos especificamente para o acesso aos dados.

Figura 1:
Essa ilustração foi baseada no esquema proposto por Tim Berners-lee (2006)

A partir do momento em que os dados são publicados em um formato que pode ser processado automaticamente por algum software (por exemplo, planilhas Excel ao invés de uma imagem), os dados passam a ser classificados como *2 Estrelas*. Por um lado, isso pode facilitar o trabalho do consumidor de dados, porém, por outro lado, pode tornar a tarefa de publicação um pouco mais complexa.

Os dados recebem a classificação de *3 Estrelas* quando são publicados em formatos não proprietários (por exemplo, CSV ao invés de Excel). Novamente, a publicação de dados em formatos abertos pode trazer custos adicionais para os provedores. Isso acontece quando o formato de origem é diferente do formato adotado para a publicação, e requer a conversão dos dados, bem como a manutenção da consistência entre a fonte de dados original e os dados publicados em formato aberto.

A medida em que os dados recebem uma identificação única e podem ser conectados com outros dados, eles podem ser classificados como *4 Estrelas*. A criação de links entre os dados permite que eles façam parte de uma rede maior de dados abertos e conectados (BIZER; HEATH; BERNERS-LEE, 2009). Finalmente, os dados abertos recebem a classificação *5 Estrelas* se estiverem conectados com dados já disponíveis na Web. Nesse caso, é necessário identificar dados que representem o mesmo conceito a fim de estabelecer os links entre eles.

Seguindo o movimento dos dados abertos, governos de diversos países estão usando a Web como meio para publicação de dados e informações sobre suas administrações. Esses dados, denominados Dados Abertos Governamentais, podem ser facilmente encontrados nos chamados Portais de Dados Abertos, os quais oferecem uma interface mais amigável para catalogação e acesso aos dados. Como exemplos de portais de dados abertos já consolidados, destacam-se o portal dos EUA [1] e o portal do Reino Unido [2]. Diversos países na Europa, como França [3] e Holanda [4], bem como países na América Latina, como Chile [5] e Uruguai [6], também possuem portais de dados

http://data.gov [1]
http://data.gov.uk [2]
http://data.gouv.fr [3]
http://dataoverheid.nl [4]
http://datos.gob.cl [5]
http://datos.gub.uy [6]

governamentais abertos. No caso do Brasil, o Portal Brasileiro de Dados Abertos[7] foi lançado no início de 2012, e foi liderado pelo Ministério do Planejamento.

A iniciativa de abertura dos dados por parte dos governos foi impulsionada pela procura de transparência, de colaboração e de participação da sociedade/comunidade (GOLDSTEIN; DYSON, 2013). Com o intuito de chegar a um consenso dos requisitos necessários para se caracterizar uma base de dados abertos, o grupo de trabalho, *Open Government Working Group*, elaborou os oito princípios dos dados governamentais abertos (TAUBERER; LESSIG, 2007):

- **Completos:** todos os dados devem estar disponíveis e não limitados. Um dado público é o dado que não está sujeito a limitações válidas de privacidade, segurança ou privilégios de acesso.

- **Primários:** os dados devem estar em formato bruto, sem agregação ou modificação.

- **Atuais:** os dados devem ser publicados tão rapidamente quanto necessário para preservar o seu valor.

- **Acessíveis:** os dados devem ser acessíveis pelo maior número possível de usuários e para o maior número possível de finalidades.

- **Processáveis por máquinas:** os dados devem ser razoavelmente estruturados para permitir processamento automatizado.

- **Não-discriminatórios:** os dados devem ser disponíveis para todos, sem necessidade de cadastro.

- **Não-proprietários:** os dados devem ser publicados em formato aberto sobre o qual nenhuma entidade tem controle exclusivo.

[7] http://dados.gov.br

- **Licenças livres:** os dados não devem estar sujeitos a nenhuma regulamentação de direitos autorais, patentes, propriedade intelectual ou segredo industrial. Restrições sensatas relacionadas à privacidade, segurança e privilégios de acesso podem ser permitidas.

Os dados abertos governamentais dizem respeito a assuntos diversos e podem envolver desde dados sobre despesas e receitas do governo até dados sobre censo escolar, pontos turísticos, reclamações de consumidores, demandas de serviços, entre outros. Em geral, os dados disponibilizados são provenientes de atividades rotineiras realizadas por órgãos governamentais, como ministérios e secretarias.

Uma vez que os dados governamentais sejam disponibilizados em formato aberto, espera-se que sejam usados no desenvolvimento de aplicativos que possam ser facilmente usados e acessados tanto por cidadãos comuns, bem como pelo próprio governo. Os aplicativos oferecem meios para análise dos dados, por meio de filtros, bem como permitem a visualização de dados de forma simples e criativa. Diversos aplicativos e visualizações já estão disponíveis na Web, os quais resultaram, principalmente, de concursos e *hackathons* promovidos para a divulgação e popularização dos portais de dados abertos.

DADOS CONECTADOS

O conceito de Dados Conectados pode ser definido como um conjunto de Boas Práticas para publicar e conectar conjuntos de dados estruturados na Web, com o intuito de criar uma "Web de Dados" (BIZER; HEATH; BERNERS-LEE, 2009). A Web de Dados cria inúmeras oportunidades para a integração semântica dos próprios dados, motivando o desenvolvimento de novos tipos de aplicações e ferramentas, como navegadores e motores de busca (ISOTANI; BITTENCOURT, 2015).

Para um melhor entendimento sobre a Web de Dados, pode-se estabelecer um paralelo entre a Web de Documentos (*i.e.* a Web atual) e a Web de Dados. A primeira faz uso do padrão HTML para publicar dados, enquanto que na segunda os dados são publicados a partir do padrão RDF (ISOTANI; BITTENCOURT, 2015). A Web de Documentos é baseada em um conjunto de padrões, incluindo: um mecanismo de identificação global e único, os URIs (*Uniform Resource Identifier*); um mecanismo de acesso universal, o HTTP; e um formato padrão para representação de conteúdo, o HTML. De modo semelhante, a Web de Dados tem por base alguns padrões, como: o mesmo mecanismo de identificação e acesso universal usado na Web de Documentos (URIs e HTTP, respectivamente); um modelo padrão para representação de dados, o RDF; e uma linguagem de consulta para acesso aos dados, a linguagem SPARQL (ISOTANI; BITTENCOURT, 2015).

Os Princípios de Dados Conectados foram introduzidos por Tim Berners-Lee (2006) e resumem-se em quatro princípios básicos:

I. Usar URIs como nome para recursos;
II. Usar URIs HTTP para que as pessoas possam encontrar esses nomes;
III. Quando uma URI for acessada, garantir que informações úteis possam ser obtidas por meio dessa URI, as quais devem estar representadas no formato RDF;
IV. Incluir links para outras URIs de forma que outros recursos possam ser descobertos.

O primeiro princípio defende o uso de URI para identificar não apenas documentos Web e conteúdos digitais, mas também objetos do mundo real e conceitos abstratos, os quais devem estar representados no formato RDF.

O segundo princípio defende o uso de URIs HTTP para identificar os objetos e os conceitos abstratos definidos pelo Princípio 1, possibilitando essas URIs serem dereferenciáveis sobre um protocolo HTTP. Neste contexto, dereferenciar é o processo de recuperar uma representação de um recurso identificado por uma URI, no qual um recurso pode ter várias representações como documentos HTML, RDF, XML, entre outros.

A fim de permitir que uma ampla gama de aplicações diferentes possa processar dados disponíveis na Web, é importante que exista um acordo sobre um formato padrão para disponibilização dos dados. O terceiro princípio de Dados Conectados defende o uso de RDF como modelo para a publicação de dados estruturados na Web (CYGANIAK; WOOD; LANTHALER, 2014). Com o RDF, é possível descrever significado sobre recursos, habilitando agentes de software a explorar os dados de

forma automática, muitas vezes, agregando, interpretando ou mesclando dados.

O quarto princípio diz respeito ao uso de *links* para conectar não apenas os documentos da Web, mas qualquer tipo de recurso. Por exemplo, um *link* pode ser criado entre uma pessoa e um lugar, ou entre um local e uma empresa. Em contraste com a Web clássica onde os *hyperlinks* são em grande parte não "tipados", *hyperlinks* que conectam os recursos em um contexto de Dados Conectados são capazes de descrever a relação entre eles. *Hyperlinks* no contexto de Dados Conectados são chamados de links RDF, a fim de distingui-los dos *hyperlinks* existentes na Web convencional (HEATH; BIZER, 2011).

É importante destacar que, atualmente, já existe um grande volume de dados abertos conectados disponível na Web. Como exemplo, destacam-se os conjuntos de dados abertos publicados pelo projeto LOD [8]. Como mencionado anteriormente, os Dados Conectados contribuem para a geração de uma Web de Dados, sendo, portanto, a opção mais almejada para a publicação de dados na Web. Nesse contexto, o *W3C Government Linked Data Working Group* propôs um conjunto de Boas Práticas para publicação de Dados Conectados a fim de prover diretrizes para auxiliar o acesso e o reuso de dados governamentais abertos. [9]

[8] http://lod-cloud.net

[9] http://www.w3.org/TR/ld-bp

DADOS NA WEB

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, 2006). 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, em sua maioria, também são dados na Web. Outra distinção importante diz respeito aos dados publicados na Web de acordo com os Princípios dos Dados Conectados. Uma parte dos dados atualmente disponíveis na Web segue esses princípios e é 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, ele pode ser classificado como Dados Abertos Conectados.

DADOS NA WEB

seguem os princípios arquiteturais da Web

DADOS ABERTOS

seguem os princípios de Dados Abertos

DADOS CONECTADOS

seguem os princípios de Dados Conectados

DADOS ABERTOS CONECTADOS

seguem os princípios de Dados Conectados e Dados Abertos

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, sensibilidade comercial e, acima de tudo, a privacidade dos indivíduos precisa ser levada em consideração pelos provedores de dados para determinar a política de publicação de dados e em quais circunstâncias os dados devem ser publicados.

Figura 2:
Intersecção de Dados na Web,
Dados Abertos e Dados Conectados
Fonte: Os autores

CICLO DE VIDA DOS DADOS NA WEB

O processo de publicação e consumo de dados na Web envolve várias fases que vão desde a seleção e publicação dos dados até o uso dos dados e *feedback* sobre os dados utilizados. Esse conjunto de fases que compõem o processo de publicação e consumo dos dados é chamado de Ciclo de Vida dos Dados na Web. A Figura 3 apresenta as fases do ciclo de vida dos Dados na Web, as quais são brevemente descritas a seguir.

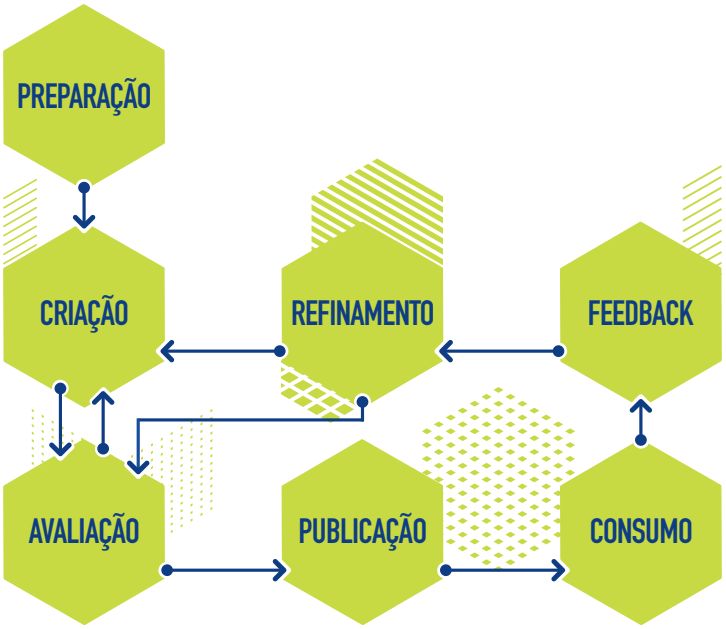


Figura 3:
Ciclo de Vida dos Dados na Web
Fonte:
Os autores

- **Preparação:** Esta fase se estende desde o momento em que surge a intenção de publicar os dados até a seleção dos dados que serão publicados. Vale lembrar que não existem regras que determinem a prioridade dos dados a serem publicados, porém é sempre importante levar em consideração a relevância dos

dados, ou seja, dados que possuem um grande potencial de utilização deveriam ter prioridade no momento da escolha. Dessa forma, sempre que possível, é importante fazer uma consulta prévia junto aos potenciais consumidores de dados para identificar a relevância dos dados.

■ **Criação:** Diz respeito ao momento em que os dados são criados, ou seja, compreende a fase de extração dos dados de fontes de dados já existentes até a sua transformação para o formato adequado para publicação na Web. Durante a fase de criação, além dos dados propriamente ditos, também devem ser criados os metadados que irão descrever os dados. Na fase de criação, também será feita a escolha dos formatos de dados a serem usados para a publicação de dados e metadados. Além disso, é sempre bom considerar a publicação de dados em diferentes formatos, minimizando a necessidade de transformação dos dados por parte dos consumidores.

■ **Avaliação:** Esta fase diz respeito à avaliação dos dados antes da sua publicação. É importante que os especialistas sejam capazes de avaliar os dados a fim de detectar inconsistências ou erros nos dados, bem como apontar dados que sigilosos que não devem ser publicados, por exemplo. Somente após uma avaliação criteriosa, os dados devem ser disponibilizados para publicação. Quando necessário, os dados podem voltar para a fase anterior a fim de resolver os problemas detectados pelos especialistas.

■ **Publicação:** Compreende o momento em que os dados serão disponibilizados de forma pública na Web. Para isso, podem ser usadas ferramentas de catalogação de dados, como CKAN[10] e Socrata.[11] Também podem ser utilizadas APIs (*Application Programming Interface*) que permitam o fácil acesso aos dados publicados, ou páginas Web, por exemplo. Em todos os casos, o provedor de dados deverá oferecer toda a informação necessária para que o con-

[10] http://ckan.org
[11] http://www.socrata.com

sumidor tenha fácil acesso aos dados. Além disso, é importante garantir que os dados serão atualizados de acordo com uma frequência pré-determinada, a qual deverá ser informada juntamente com os dados.

■ **Consumo:** Implica o momento em que os dados são usados para a criação de visualizações, como gráficos e mapas de calor, bem como para aplicações que permitem o cruzamento e a realização de análises sobre os dados. Esta fase do ciclo de vida está diretamente relacionada ao consumidor de dados, que pode ser desde uma grande empresa interessada em usar os dados disponíveis na Web para a melhoria de seus produtos e serviços, até um único desenvolvedor interessado em usar os dados para criar uma aplicação que irá melhorar a qualidade de vida na sua cidade.

■ ***Feedback*:** Esta fase compreende o momento em que os consumidores proveem comentários sobre os dados e metadados previamente utilizados. Esta fase é de fundamental importância, pois a partir do *feedback* dos consumidores será possível identificar melhorias e realizar correções nos dados previamente publicados. Além disso, esse canal de comunicação entre consumidores e provedores de dados também facilita a identificação de novos dados relevantes que devem ter prioridade no momento da escolha de novos dados a serem publicados.

■ **Refinamento:** Esta fase compreende todas as atividades relacionadas a adições ou atualizações nos dados que já foram publicados. É muito importante garantir a manutenção dos dados previamente publicados, a fim de oferecer maior segurança para aqueles que irão consumir os dados. A manutenção pode ser feita de acordo com o *feedback* dos consumidores ou novas versões podem ser geradas a fim de garantir que os dados não fiquem obsoletos. Para isso, é importante fazer o correto gerenciamento das diferentes versões dos dados e garantir que os consumidores tenham acesso à versão correta dos dados.

Com relação aos atores que participam do ciclo de vida dos dados na Web, estes podem desempenhar dois papéis principais: os provedores de dados e os consumidores de dados. O papel de provedor de dados pode ser desempenhado por vários atores, os quais são responsáveis por realizar atividades como criação de metadados, criação e publicação de dados. Os consumidores de dados são atores que recebem e consomem os dados. Ressalta-se que os consumidores de dados também podem ser provedores de dados, uma vez que os consumidores podem realizar melhorias e refinamentos nos dados a fim de oferecê-los novamente para a comunidade. É importante notar que o ciclo de vida proposto não requer que todas as fases sejam seguidas até que uma nova iteração seja iniciada.

BOAS PRÁTICAS PARA DADOS NA WEB

As Boas Práticas para Dados na Web (DWBP, do inglês *Data on the Web Best Practices*), descritas na Recomendação do W3C por Lóscio, Burle e Calegari (2017), foram desenvolvidas para incentivar e permitir a expansão continuada da Web como um meio para o intercâmbio de dados. Em termos gerais, os provedores de dados visam compartilhar dados abertamente ou com acesso controlado. Consumidores de dados 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 provedores e os consumidores de dados. Sem esse acordo, os esforços dos provedores podem ser incompatíveis com os anseios dos consumidores.

Neste contexto, torna-se crucial fornecer orientações aos provedores, de maneira que possam contribuir para a melhoria da coerência na forma como os dados são gerenciados. 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 propostas em Lóscio, Burle e Calegari (2017) foram desenvolvidas para oferecer orientação técnica para a publicação de dados na Web, contribuindo para melhorar a relação entre provedores e consumidores de dados.

As Boas Práticas propostas 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 de dados na Web (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, Burle e Calegari (2017) e ilustrado no Quadro 1, cada boa prática tem um Resultado esperado, que descreve "O que deve ser possível fazer quando um provedor 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.

As Boas Práticas propostas para publicação e utilização de dados na Web referem-se a conjuntos de dados, ou seja, "coleção de dados, publicados gerenciados por um único agente, e disponíveis para serem acessados ou baixados em um ou mais formatos" (MAALI; ERICKSON, 2014, tradução nossa). Por dados, "queremos dizer fatos conhecidos que podem ser gravados e que têm significado implícito" (ELMASRI; NAVATHE, 2010, tradução nossa). Conforme descrito na Figura 4, 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. Ou seja, "uma pessoa ou grupo acessa, utiliza, e potencialmente executa as fases de pós-tratamento dos dados" (STRONG; LEE; WANG, 1997, tradução nossa), sem levar em conta a finalidade, o público, interesse ou licença. Tendo em conta esta heterogeneidade e o fato de que os provedores de dados e os consumidores de dados podem não se conhecer, é necessário fornecer algumas informações sobre

os conjuntos de dados e distribuições que também podem contribuir para a confiabilidade e reutilização, tais como: metadados estruturais, metadados descritivos, acesso à informação, informação sobre a qualidade de dados, informações sobre a procedência, informações sobre licença e informações sobre uso.

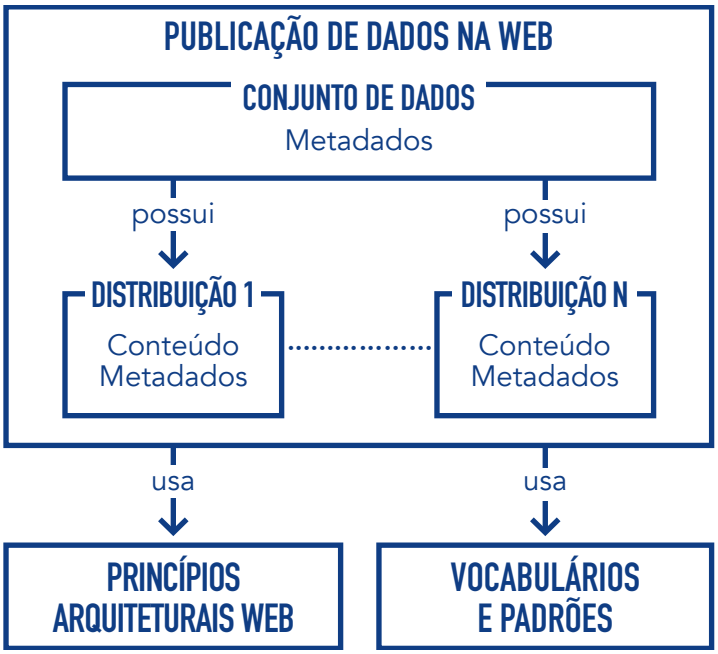


Figura 4: Contexto de publicação de dados na Web. Fonte: Lóscio, Burle e Calegari (2017)

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 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 inteiro 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 fazer conexões, via URIs, entre dois ou mais recursos.

BOAS PRÁTICAS PARA DADOS NA WEB COM SEUS RESPECTIVOS RESULTADOS ESPERADOS

BP1
FORNECER METADADOS

Os seres humanos serão capazes de compreender os metadados e agentes de software serão capaz de processá-los.

BP2
FORNECER METADADOS DESCRITIVOS

Os seres humanos serão capazes de interpretar a natureza do conjunto de dados e suas distribuições, e 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 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 vão saber 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 a 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 descobertas e citadas 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 se referir a 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 de caracteres que representam datas, horas, moedas e números com precisão.

BP14
FORNECER DADOS EM VÁRIOS FORMATOS

Tantos usuários quanto possível serão capazes de utilizar os dados sem primeiro ter que transformá-los em seu formato preferido.

BP15
REUTILIZAR VOCABULÁRIOS, DANDO PREFERÊNCIA AO PADRONIZADOS

Interoperabilidade e consenso entre os provedores 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 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 poderão 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 'NEGOCIAÇÃO DE CONTEÚDO' PARA SERVIR OS DADOS DISPONÍVEIS EM VÁRIOS FORMATOS

Negociação de conteúdo 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, onde 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 em 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 POR MEIO DE 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.

BP25
FORNECER DOCUMENTAÇÃO COMPLETA PARA AS APIS

Os desenvolvedores serão capazes de obter informações detalhadas sobre cada chamada para a API, incluindo os parâmetros que leva e o que é esperado para retornar, isto é, todo o conjunto de informações relacionadas com a API. O conjunto de valores – como usá-lo, avisos de mudanças recentes, informações de contato, e assim por diante – devem ser descritos e facilmente navegável na Web. Também permitirá que as máquinas possam acessar a documentação da API para ajudar os desenvolvedores na criação de softwares clientes da 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. Alterações que afetem o funcionamento da API devem ser raras. Porém, se ocorrerem,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 irá sempre fazer referência ao conjunto de dados ou redirecionar para 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 provedor 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 provedores 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 formas que podem ser facilmente compreendidos.

BP33
FORNECER *FEEDBACK* PARA O PROVEDOR ORIGINAL

Uma melhor comunicação entre provedores e consumidores fará com que seja mais fácil para os provedores originais determinar como os dados que eles publicam estão sendo usados. Isso ajudará a justificar a publicação dos dados. Os provedores também serão informados de medidas que podem ser tomadas para melhorar os seus dados, contribuindo para a melhoria dos dados de uma maneira geral.

BP34
OBEDECER OS TERMOS DE LICENÇA

Os provedores 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 capaz de licenciar adequadamente os trabalhos derivados a partir de dados previamente publicados.

BP35
CITAR A PUBLICAÇÃO ORIGINAL DO CONJUNTO DE DADOS

Os consumidores finais serão capazes de avaliar a confiabilidade dos dados que vêem e os esforços dos provedores originais serão reconhecidos. A cadeia de proveniência para os dados na Web será rastreável de volta ao seu provedor original.

A fim de incentivar os provedores a adotar as Boas Práticas para dados na Web, há uma série de benefícios que podem ser alcançados a partir da aplicação das Boas Práticas, são eles: compreensibilidade; facilidade de processamento; facilidade de descoberta; reúso; confiança; capacidade de conexão de dados; facilidade de acesso; e interoperabilidade. Os benefícios são importantes porque ajudam provedores de dados a ter uma melhor compreensão de "o que será possível" quando as Boas Práticas são adotadas. 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 provedor 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 a 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.

- **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.

- **Facilidade de Processamento:** Máquinas ou agentes de software serão capazes de processar e manipular automaticamente os dados.

- **Facilidade de Descoberta:** Os agentes de software serão capazes de descobrir automaticamente um conjunto de dados ou dados dentro de um conjunto de dados.

- **Reúso:** As chances de reutilização do conjunto de dados por diferentes grupos de consumidores de dados tendem a aumentar.

- **Confiança:** A confiança que os consumidores têm no conjunto de dados tende a melhorar.

- **Capacidade de Conexão:** Será possível criar ligações entre conjuntos de dados e itens de dados.

- **Facilidade de Acesso:** Os seres humanos e máquinas serão capazes de acessar dados atualizados em uma variedade de formas.

- **Interoperabilidade:** Será mais fácil chegar a um consenso entre os provedores e consumidores de dados.

TÉCNICAS PARA PUBLICAÇÃO DE DADOS NA WEB

A medida em que a Web se consolidou como plataforma para publicação e compartilhamento de documentos, organizações passaram a ter interesse no uso da Web como plataforma para publicação de dados. Durante os últimos anos, diversas técnicas emergiram para a publicação de dados na Web que vão desde o uso de formulários para a realização de consultas a um banco de dados até a publicação de Dados Conectados (CERI et al., 2013 e FERRARA et al., 2014). A seguir, algumas dessas técnicas para a publicação de dados são apresentadas (CERI et al., 2013 e FERRARA et al., 2014), incluindo o uso de Web APIs, a inserção de dados diretamente nas páginas HTML e as ferramentas para criação de catálogos de dados.

ACESSO A PARTIR DE WEB APIS

Uma forma de publicação de dados na Web consiste em utilizar Web APIs. Uma das primeiras propostas para padronização de APIs para a Web foram os *Web Services* (ALONSO et al., 2004), inspirados no paradigma de RPC (*Remote Procedure Call*) (NELSON, 1981) e no uso de XML (*eXtensible Markup Language*) para a troca de dados. Posteriormente, surgiu o paradigma REST (*Representational State Transfer*) e o formato JSON (*JavaScript Object Notation*) (MANDEL 2008) passou a ser amplamente adotado. Este novo tipo de API é conhecido como RESTful *service*.

Em geral, dados expostos por meio de APIs não podem ser encontrados pelos mecanismos de busca. Uma das razões para isso é que em muitos casos é necessário realizar uma autenticação antes de ser possível acessar a API. Além disso, existem restrições quanto ao uso da API a fim de evitar acessos exaustivos aos dados. Sendo assim, é possível dizer que os dados disponíveis por meio de APIs são semelhantes aos dados disponíveis na *Deep Web*, ou seja, não podem ser facilmente encontrados e indexados.

Porém, a razão para isso acontecer é bem diferente e consiste na necessidade dos provedores em controlar o acesso aos dados por aplicações externas.

ENRIQUECIMENTO DE PÁGINAS HTML

Uma outra forma de publicar dados na Web consiste em fazer a inclusão dos dados nas páginas HTML. Isso é possível com o uso de microformatos, ou seja, marcadores (*tags*) específicos que tornam explícita a semântica dos dados. O uso de microformatos permite aos mecanismos de busca identificar os dados disponíveis nas páginas HTML e, assim, apresentar melhores resultados aos usuários. Além disso, os provedores de dados podem alcançar maior visibilidade. Diversos microformatos foram desenvolvidos pela comunidade para a publicação de dados de diferentes domínios, incluindo: *hCalendar* para eventos, *hReview* para revisões e ratings, *hRecipe* para receitas culinárias e *hCard* para dados pessoais.[12]

O uso de microformatos é uma solução simples para a publicação de dados na Web, porém também apresenta algumas limitações: I) o uso de diferentes microformatos em uma mesma página pode levar a conflitos de nomes (por exemplo, a class url de CSS e o termo url do microformato *hCalendar*), II) não permite a criação de especializações e generalizações e III) cada microformato requer um *parser* específico.

Esses problemas podem ser solucionados com o uso de RDFa[13], uma solução que permite a especificação de atributos para descrição de dados estruturados em qualquer linguagem de marcação, em particular XHTML[14] e HTML. Enquanto os microformatos combinam a sintaxe para incluir os dados estruturados nas páginas HTML com a própria semântica dos dados, RDFa preocupa-se apenas com a sintaxe para inclusão dos dados estruturados. Para a semântica dos dados, RDFa permite o uso de vocabulários específicos, como o schema.org[15]. RDFa permite que múltiplos vocabulários sejam utilizados em conjunto sem a necessidade de *parsers* específicos para cada um deles.

[12] http://microformats.org
[13] http://w3.org/TR/rdfa-primer
[14] http://w3.org/TR/xhtml1
[15] http://schema.org

Além do uso de RDFa para adicionar metadados estruturados em um documento HTML, a linguagem JSON-LD[16](JSON for *Linked Data*) pode ser utilizada. É um padrão baseado no formato JSON, mas que permite o uso de vocabulários e ontologias para a descrição dos dados. O formato JSON-LD possui grande adoção por parte da comunidade técnica e é recomendado pelo Google[17] para ser adotado como formato padrão de intercâmbio de Dados Conectados nas páginas Web.

FERRAMENTAS PARA CATALOGAÇÃO DE DADOS

Com o crescente interesse na publicação de dados abertos, em especial os dados abertos governamentais, uma nova forma de publicação de dados na Web ganhou destaque: as ferramentas para catálogos de dados, como CKAN[18] e Socrata[19]. A partir dessas plataformas, são criados os portais de dados abertos, os quais oferecem acesso a conjuntos de dados previamente catalogados. Os conjuntos de dados são organizados como uma série de recursos e podem ser classificados de acordo com *tags* que explicitam o domínio dos dados.

Essas plataformas de catalogação são uma ótima ferramenta para indexação de conjuntos de dados, mas deixam a desejar quanto à busca de dados, uma vez que não permitem fazer buscas nos conjuntos de dados propriamente ditos. Em alguns casos, as ferramentas de catalogação oferecem APIs de acesso aos dados, mas isso é feito de forma bastante simplificada. Os conjuntos de dados disponíveis nos catálogos podem ser encontrados pelas ferramentas de busca, porém ainda não é possível encontrar itens de dados específicos armazenados em um conjunto de dados.

Apesar da grande disseminação dos portais de dados abertos, estas soluções apresentam diversas limitações, dentre elas destacam-se: a dificuldade em manter os dados atualizados, a falta de padrões de metadados para descrição dos conjuntos de dados e a impossibilidade de realização de consultas sobre os dados. Além disso, como os conjuntos de dados publicados nos portais geralmente encontram-se disponíveis em diversos formatos, ou seja, múltiplos arquivos para um mesmo conjunto de dados, também pode haver redundância de dados.

[16] https://www.w3.org/TR/json-ld
[17] https://developers.google.com/search/docs/guides/intro-structured-data
[18] http://ckan.org
[19] http://www.socrata.com

CONCLUSÃO

O interesse na publicação de dados na Web não é algo novo. Entretanto, o crescente interesse no uso da Web como plataforma para compartilhamento de dados trouxe novos desafios para a publicação de dados de forma estruturada. Em cenários onde os consumidores de dados não são previamente conhecidos, a publicação de dados deve ser realizada de maneira a atender grupos de consumidores com requisitos e perfis diversos.

Neste contexto, além dos aspectos básicos de disponibilização de dados, devem ser levados em consideração outros aspectos que dizem respeito à compreensão, à confiabilidade e ao processamento dos dados de forma automática. Por um lado, os provedores devem fornecer informações que auxiliem no entendimento dos dados, como metadados estruturais, mas também devem prover informações que permitam aos consumidores conhecer a proveniência e a qualidade dos dados. Por outro lado, os consumidores devem ser capazes de prover *feedback* sobre os dados que foram usados, a fim de contribuir para a melhoria do processo de publicação. Além disso, os consumidores devem prover informações sobre o uso dos dados, ou seja, juntamente com a aplicação ou visualização que foi gerada a partir dos dados publicados, devem ser disponibilizadas informações sobre os dados que foram usados. Para facilitar as tarefas de provedores e consumidores de dados na Web, foram propostas um conjunto de Boas Práticas que abordam aspectos relacionados à todo o ciclo de vida dos na Web. A adoção dessas Boas Práticas leva à criação de um canal de comunicação entre provedores e consumidores, além de contribuir para a melhoria do processo de publicação de dados na Web.

REFERÊNCIAS

ABITEBOUL, Serge; BUNEMAN, Peter; SUCIU, Dan. Data on the Web: from relations to semistructured data and XML. San Francisco: Morgan Kaufmann, 2000.

ALONSO, Gustavo et al. Web Services: Concepts, Architectures and Applications. Heidelberg: Springer, 2004.
BERNERS-LEE, Tim; CONNOLLY, Dan; SWICK, Ralph R.. Web Architecture: Describing and Exchanging Data. 1999. Disponível em: <https://www.w3.org/1999/04/WebData>. Acesso em: 04 set. 2018.

BERNERS-LEE, Tim. Linked Data. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 04 set. 2018.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked Data - The Story So Far. International Journal On Semantic Web And Information Systems, v. 5, n. 3, p.1-22, jul. 2009. IGI Global.

CERI, Stefano et al. Web Information Retrieval. Springer Science & Business Media, 2013.

CYGANIAK, Richard; WOOD, David; LANTHALER, Markus. RDF 1.1 Concepts and Abstract Syntax. 2014. Disponível em: <https://www.w3.org/TR/rdf11-concepts/>. Acesso em: 04 set. 2018.

ELMASRI, Ramez; NAVATHE, Shamkant. Fundamentals of Database Systems. Addison-wesley Publishing Company, 2010.

FERRARA, Emilio et al. Web data extraction, applications and techniques: A survey. Knowledge-based Systems, [s.l.], v. 70, p.301-323, nov. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.knosys.2014.07.007.

GOLDSTEIN, Brett; DYSON, Lauren (Ed.). Beyond Transparency: Open Data and the Future of Civic Innovation. San Francisco: Code For America Press, 2013.

HEATH, Tom; BIZER, Christian. Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers, 2011. 136 p. (Synthesis Lectures on the Semantic Web: Theory and Technology).

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. Dados abertos conectados. São Paulo: Novatec, 2015. 175 p.

JACOBS, Ian; WALSH, Norman. Architecture of the World Wide Web. 2004. Disponível em: <https://www.w3.org/TR/webarch/>. Acesso em: 04 set. 2018.

LEE, Deirdre; LÓSCIO, Bernadette Farias; ARCHER, Phil. Data on the Web Best Practices Use Cases & Requirements. 2015. Disponível em: <https://www.w3.org/TR/dwbp-ucr/>. Acesso em: 04 set. 2018.

LÓSCIO, Bernadette Farias; BURLE, Caroline; CALEGARI, Newton. Data on the Web Best Practices. 2017. Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 04 set. 2018.

MAALI, Fadi; ERICKSON, John. Data catalog vocabulary (DCAT). 2014. Disponível em: <https://www.w3.org/TR/vocab-dcat/>. Acesso em: 04 set. 2018.

NELSON, Bruce Jay. Remote procedure call. 1981. 201 f. Tese (Doutorado) - School Of Computer Science, Carnegie Mellon University, Pa, 1981.

OPEN KNOWLEDGE. Open data handbook. 2012. Disponível em: <http://opendatahandbook.org/>. Acesso em: 04 set. 2018.

PIRES, Marco Túlio. Guia de Dados Abertos. São Paulo: Este Guia é parte integrante do Projeto de Cooperação entre o Governo do Estado de São Paulo e o Reino Unido, 2015. Disponível em: <http://ceweb.br/media/docs/publicacoes/13/Guia_Dados_Abertos.pdf>. Acesso em: 04 set. 2018.

STRONG, Diane M.; LEE, Yang W.; WANG, Richard Y. Data quality in context. Magazine Communications Of The Acm, Nova Iorque, v. 40, n. 5, p.103-110, 05 maio 1997.

TAUBERER, Joshua; LESSIG, Larry. The 8 Principles of Open Government Data. 2007. Disponível em: <https://opengovdata.org/>. Acesso em: 04 set. 2018.

ANEXO

ROADMAP DE PUBLICAÇÃO DE DADOS ABERTOS

1.PREPARAÇÃO

O QUE FAZER?	COMO FAZER?	ARTEFATOS	METADADOS
Identificar demandas de dados	**1.** Interagir com potenciais consumidores por meio de entrevistas ou consultas públicas **2.** Analisar solicitação de acesso à informação **3.** Avaliar portais corporativos ou outras fontes de disseminação de dados	Plano de demandas de dados	Identificação dos dados
Identificar conjuntos de dados em potencial	**1.** Agrupar as demandas que dizem respeito a itens de dados similares em um mesmo conjunto de dados	Lista de conjuntos de dados	Descritivos
Definir a prioridade dos conjuntos de dados a serem abertos	**1.** Definir a prioridade de abertura de cada conjunto de acordo com o número de solicitantes de cada demanda	Lista de prioridades para abertura	

2.CRIAÇÃO

O QUE FAZER?	COMO FAZER?	ARTEFATOS	METADADOS
Modelagem do conjunto de dados	**1.** Avaliar as propriedades de cada demanda associada ao conjunto de dados para definir a estrutura do conjunto como um todo **2.** Agrupar as propriedades semelhantes, eliminar propriedades redundantes	Esquema inicial do conjunto de dados	Estruturais
Identificar fontes de dados de origem	**1.** Avaliar sistemas e documentos existentes a fim de identificar a fonte de origem dos dados	Lista de fontes de dados de origem	Proveniência
Mapeamento entre as fontes de origem e o conjunto de dados	**1.** Estabelecer a correspondência entre as propriedades do esquema do conjunto de dados e as propriedades das fontes de dados de origem	Documento de mapeamento entre fonte de dados de e conjunto de dados	Descritivos
Identificar dados sensíveis	**1.** Consultar especialistas ou legislação correspondente para identificação de dados sensíveis	Lista de dados sensíveis	
Identificar vocabulários	**1.** Avaliar o uso de vocabulários conhecidos (ex: dcterms, foaf, schema.org) na definição das propriedades do conjunto de dados **2.** Fazer busca em respositórios de vocabulários para identificar vocabulários adequados para o domínio	Lista de vocabulários a serem usados no esquema do conjunto	

2.CRIAÇÃO

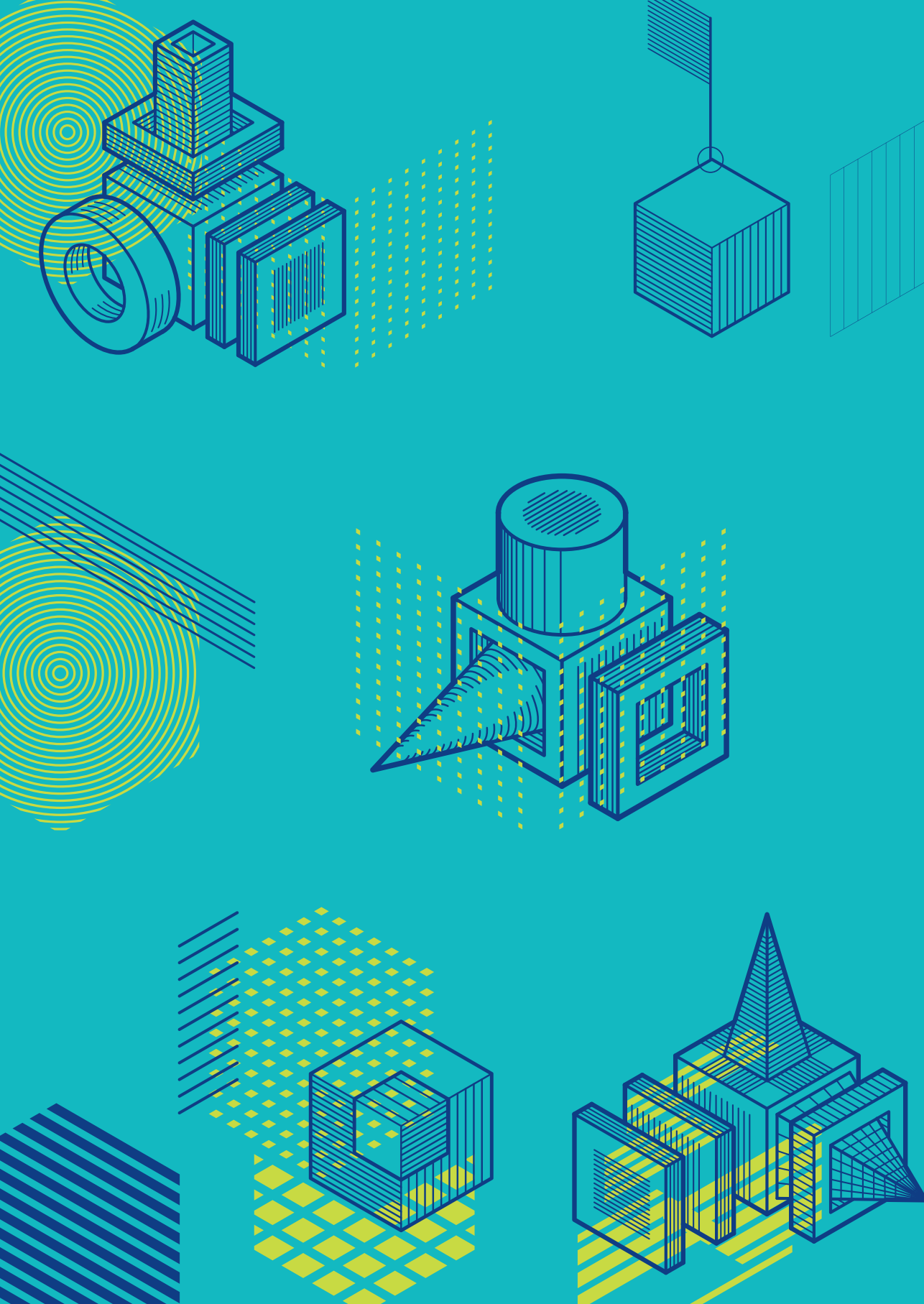
O QUE FAZER?	COMO FAZER?	ARTEFATOS	METADADOS
Mapeamento entre os vocabulários e o esquema do conjunto de dados	**1.** Estabelecer a correspondência entre as propriedades do esquema do conjunto de dados e os termos dos vocabulários previamente escolhidos	Documento de mapeamento entre o esquema e vocabulários	
Definir estratégia de extração dos dados	**1.** De acordo com o tipo de fonte de dados (ex: banco de dados, planilha, documento de texto), especificar a forma de extração dos dados	Plano de extração de dados	Proveniência
Definir subconjuntos de dados	**1.** Caso o volume de dados seja muito grande, definir possíveis conjuntos de dados **2.** A divisão dos subconjuntos pode ser feita com base em algum atributo temporal ou espacial, por exemplo. Outros atributos mais específicos também podem ser usados	Lista de subconjuntos de conjuntos dados	
Gerar distribuições	**1.** Aplicar estratégia de extração previamente definida e gerar as distribuições de dados desejadas	Distribuições do conjunto de dados	Descritivos das distribuições

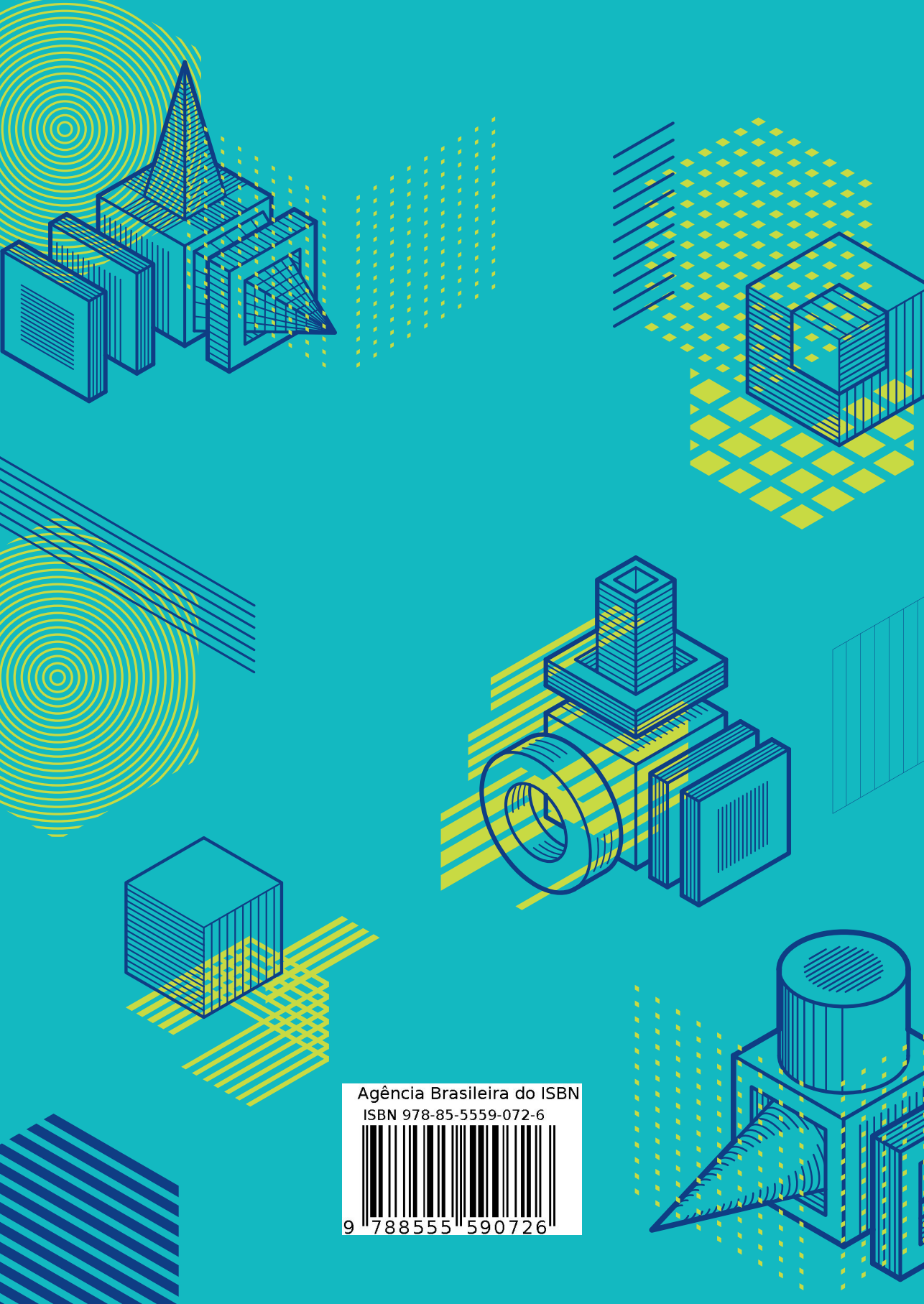
3.AVALIAÇÃO

O QUE FAZER?	COMO FAZER?	ARTEFATOS	METADADOS
Avaliar a qualidade dos dados	**1.** Definir critérios de qualidade a serem avaliados (ex: completude, corretude, atualidade) **2.** Definir métricas para avaliação dos critérios **3.** Definir requisitos mínimos para cada critério de qualidade **4.** Avaliar os critérios de qualidade de forma manual ou automática	Documento de qualidade dos dados	Qualidade dos dados
Liberar dados para publicação	**1.** Preencher documento de liberação do conjunto de dados	Documento de liberação do conjunto de dados	Descritivos
Retornar conjunto de dados para a fase de criação	**1.** Preencher documento de retorno à fase de criação com devida justificativa e descrição de melhorias necessárias	Documento de retorno à fase de criação	

4.PUBLICAÇÃO

O QUE FAZER?	COMO FAZER?	ARTEFATOS	METADADOS
Publicar conjunto de dados em uma ferramenta de catalogação de dados	**1.** O procedimento pode variar de acordo com a ferramenta utilizada. Em geral, é necessário fazer o upload dos arquivos das distribuições e dos metadados do conjunto de dados **2.** Preencher todos os metadados solicitados e, se necessário, acrescentar novos metadados	Conjunto de dados disponível para acesso e download na ferramenta de catalogação	Descritivos, Versionamento
Publicar conjunto de dados em uma página HTML	**1.** Criar a página HTML tanto na versão para o consumo humano quanto para ser processada pela máquina **2.** Inserir tags RDFa no código HTML com as informações semânticas para o processamento pela máquina	Conjunto de dados disponível para acesso e download em uma página HTML	Descritivos, Versionamento
Desenvolver API de acesso aos dados	**1.** Criar API que permita o acesso aos conjuntos de dados **2.** Criar documentação da API	Conjunto de dados disponível para acesso e download por meio de uma API e documentação da API	Descritivos, Versionamento
Estabelecer canal de comunicação com os consumidores de dados	**1.** O canal de comunicação dependerá da forma como o conjunto de dados foi publicado. Caso a ferramenta usada não ofereça um canal de comunicação, crie uma página HTML	Página de contato	Uso dos dados





Agência Brasileira do ISBN
ISBN 978-85-5559-072-6
9 788555 590726