XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-08 – Informação e Tecnologia

BOAS PRÁTICAS PARA PUBLICAÇÃO DE DADOS NA WEB: APLICAÇÃO NOS DADOS REFERENTES AOS RESULTADOS DE PESQUISA CIENTÍFICA

Caio Saraiva Coneglian (Universidade Estadual Paulista - UNESP)

Larissa Pavarini Luz (Universidade Estadual Paulista - UNESP)

José Eduardo Santarém Segundo (Universidade de São Paulo - USP)

THE BEST PRATICES FOR WEB DATA PUBLISHING: APPLICATION ON RESULTS DATA REFERRING TO SCIENTIFIC RESEARCH

Modalidade da Apresentação: Comunicação Oral

Resumo: Considerando a necessidade da comunidade científica em encontrar meios de localizar e navegar pelos resultados de pesquisa cientifica (no contexto desta pesquisa, artigos de periódicos e trabalhos de eventos), dificultado pela explosão de informações disponíveis na Web, o *Linked Data* se apresenta propício para a divulgação científica, em especial dos dados referentes aos resultados de pesquisa científica, como por exemplo os metadados de um artigo científico. No entanto, a falta de normativas e diretrizes para a publicação dos dados, pode conduzir a serem publicados conjuntos de dados na Web que não tem a qualidade desejada. Desta forma, objetiva-se discutir e analisar as boas práticas de publicação de dados na Web no contexto do *Linked Data* recém-publicado pelo *World Wide Web Consortium*, no contexto dos dados referentes aos resultados de pesquisa científica. Para tanto, procede-se com pesquisa qualitativa exploratória bibliográfica. Desse modo, observa-se que as doze categorias das boas práticas estão vinculadas diretamente ao domínio do trabalho, existindo literaturas capazes de auxiliar na tarefa de conversão dos dados para os formatos de *Linked Data*, o que permite concluir que os projetos que seguirem as boas práticas e as recomendações dadas neste trabalho, conseguirão ter como resultado oito benefícios necessários a estes dados.

Palavras-Chave: Linked Data; Web Semântica; Boas Práticas de Publicação.

Abstract: *Considering the need for the scientific community to find means to locate and navigate the results of scientific research (in the context of this research, papers and paper events), hampered by the explosion of information available on the Web, the Linked Data proved to be conductive to scientific dissemination, related to the data results referring to this scientific research, such as the metadata of a paper. However, the lack of regulations and guidelines for the publication of data may lead to the publication of data sets on the Web that do not have the desired quality. In this way, we aim to discuss and analyze the best practices of data publication in the Web in the context of Linked*

Data publisher to World Wide Web Consortium, recently published, in the context of the data results referring to scientific research. For that, we proceed with qualitative exploratory bibliographical research. In this way, it can be observed that the twelve categories of best practices are directly linked to the work domain, and there are literatures capable of assisting in the task of converting data to Linked Data formats, which allows to conclude that projects that follow the best practices and the recommendations given in this paper, will be able to result in eight necessary benefits to these data.

Keywords: Linked Data; Semantic Web; Best Practices for publishing.

1 INTRODUÇÃO

O XVIII ENANCIB ao apresentar a temática "Informação, Sociedade e Complexidade" nos conduz a debater acerca de como a Ciência da Informação se insere mais ativamente nesta transição que vive a ciência, desafiando os pesquisadores a reverem os métodos oriundos das ciências clássicas, e inserir esta área do conhecimento mais ativamente no âmbito da multi, inter e transdisciplinaridade.

Neste sentido, diversas transformações dentro do campo científico estão diretamente relacionadas ao desenvolvimento das Tecnologias de Informação e Comunicação (TIC), em especial a criação e a popularização da Web. Ainda que passados dezessete anos do início do século XXI com uma evolução extrema das tecnologias, identificamos como atual o que Morin (2005, p. 120) afirma que "[...] estamos no fim de um certo tempo e nós o esperamos, no começo de novos tempos.", necessitando encontrar formas que introduzam com mais frequência o pensamento complexo na ciência.

Ainda que recaiam críticas sobre os monopólios empresariais criados dentro da Web, este ambiente informacional possui em sua essência a liberdade, permitindo com que quaisquer pontos de acesso conectados à rede disponibilizem seus dados abertamente. Neste contexto, a Web pode auxiliar a promover uma revolução científica real, abrindo espaço aos diversos pesquisadores a debaterem e encontrarem informações em novos formatos, expandindo as barreiras tradicionais da comunicação científica.

A Ciência da Informação deve estar ao centro desse debate, pela sua natureza interdisciplinar e com a competência de domínio acerca do tratamento dado aos resultados de pesquisa científica, em especial, para promover e conduzir esta nova visão de como os resultados dessa pesquisa possam ser disponibilizados para a imersão neste novo cenário.

Um caminho que floresce nesta revolução se encontra no *Linked Data*, um cenário vislumbrado por Tim Berners Lee, que a partir dos conceitos e das tecnologias da Web Semântica busca ser uma forma de publicar dados na Web. O *Linked Data* contempla essencialmente diretrizes para a disponibilização de dados na Web, interligando variados conjuntos de dados, seguindo os princípios da Web de Dados A chamada Web de Dados faz com que dentro da Web, as informações estejam contextualizadas e relacionadas a outros recursos, possibilitando que agentes computacionais sejam capazes de compreender o domínio de um dado, aprimorando a recuperação da informação.

Lançado em 2006, o *Linked Data* tem crescido e se popularizado na disponibilização de conteúdo. A partir desta expansão, neste ano, 2017, um grupo do W3C publicou um texto indicando 35 boas práticas para a publicação de dados na Web, a partir dos princípios do *Linked Data* (LÓSCIO; BURLE; CALEGARI, 2017). Essa iniciativa é de suma importância para o aprimoramento os conjuntos de dados publicados no âmbito do *Linked Data*, pois estabelece diretrizes que tornam mais aperfeiçoado o processo de publicação, melhorando a qualidade dos dados publicados na Web.

Diante deste cenário, a publicação de dados referentes a resultados de pesquisa científica, como artigos de periódicos e trabalhos apresentados em eventos científicos, podem usufruir dessas boas práticas. A realização desse processo faz com que de fato o uso das TIC neste contexto científico possa contribuir na instauração desta nova configuração na Ciência, ao encontrar um ambiente livre para disponibilizar e encontrar dados que se referem aos resultados científicos, em novos formatos, mais adequados aos tempos que vivemos.

Os dados referentes a resultados de pesquisas científicas, por exemplo os metadados de um artigo ou de um trabalho publicado em um evento, são elementos chave na divulgação da própria pesquisa, visto que contemplam as informações descritivas de uma publicação, ao mesmo tempo que fornece o *link* para o acesso ao texto original, além de contemplar uma série de outros elementos que não são apresentados nas estruturas tradicionais. Neste contexto, a publicação deste tipo de dado em formatos semânticos pode ser uma importante via para traçar relacionamentos e inseri-los mais eficientemente na Web.

Contudo, apesar dos claros benefícios trazidos pelo emprego de boas práticas em cenários de publicação, a sua aplicação em cenários distintos traz particularidades que devem ser consideradas. Assim, questiona-se: Como as boas práticas de publicação de dados na Web podem ser aplicadas no domínio de dados referentes a resultados de pesquisa científicas?

Desta forma, este trabalho tem como objetivo discutir e analisar as boas práticas de publicação de dados na Web (LÓSCIO; BURLE; CALEGARI, 2017), no contexto dos dados referentes a resultados de pesquisa científica. Para isso, utilizou-se uma metodologia de natureza qualitativa, realizando um estudo exploratório e analítico, que buscou na literatura subsídios teóricos das boas práticas e sua relação com o *Linked Data*, além de realizar uma análise de como as boas práticas podem ser aplicadas ao domínio de resultados de pesquisa científica.

2 WEB SEMÂNTICA E LINKED DATA

Com a expansão da Web, em especial com o aumento exponencial de conteúdos disponibilizados, foi necessário encontrar formas de aprimorar os processos de recuperação da informação neste ambiente. Neste sentido, a Web Semântica foi criada em 2001 com a finalidade de ser "[...] uma extensão da Web atual, em que a informação possui um significado claro e bem definido, possibilitando uma melhor interação entre computadores e pessoas." (BERNERS-LEE; HENDLER; LASSILA, 2001, p. 2, tradução nossa).

O surgimento da Web Semântica se tornou importante uma vez que ela permite que agentes computacionais consigam interpretar, por meio de artefatos, o contexto de uma pessoa ao utilizar um ambiente informacional digital. Para isso, a Web Semântica foi constituída por conceitos e tecnologias, que vêm evoluindo a cada ano de forma consistente, dando alicerce às aplicações, em especial na busca de explorar com precisão bases de dados visando ter uma compreensão do contexto das informações.

O *Linked Data*, criado em 2006 por Berners-Lee, é uma proposta que utiliza os conceitos da Web Semântica e se destacou pela inserção de significado nos dados na Web. Berners-Lee (2006) apresentou algumas diretrizes para a implementação do *Linked Data* com criação de bases de dados seguindo normas que facilitam a inserção de significado nestes dados.

A publicação de dados seguindo os princípios do *Linked Data* vem ocorrendo em diversos cenários, desde dados de ciências da saúde, passando por informações sociais, chegando até dados referentes a publicações científicas. Neste último domínio, uma série de iniciativas surgiu para nortear a criação e o compartilhamento de informações de publicação de dados provenientes de resultados de pesquisa científicas, como artigos de periódicos e trabalhos de eventos.

Neste contexto, a publicação dos dados referentes a resultados de pesquisa pode auxiliar na divulgação do conhecimento científico, que segundo Lynch (2011), permite uma reprodução de resultados, oferecendo comprovações para a qualidade do trabalho científico.

Contudo o sucesso na construção e manutenção desses *datasets* de publicação de dados referentes a resultados de pesquisa científicas, está claramente ligada a adoção das boas práticas, designadas pela W3C, e discutida no âmbito da Ciência da Informação.

A criação e a estruturação dos *datasets* por parte das entidades que promovem a disseminação do conhecimento científico, proporcionam as condições adequadas para que a

gestão eficiente da informação científica ocorra, propiciando ainda que a recuperação da informação seja feita de forma eficiente e organizada, tornando esses conjuntos de dados capazes de auxiliar nas necessidades de pesquisadores e das comunidades científicas.

Esse processo permite que o uso das TIC na publicação científica contribua de forma intensa na Ciência, transformando em um ambiente livre para disponibilizar e encontrar dados sobre resultados de pesquisas científicas em novos formatos, e que possam auxiliar de forma mais adequada a necessidade de cada usuário.

3 BOAS PRÁTICAS DE PUBLICAÇÃO CIENTÍFICA

A publicação de dados na Web, bem como a sua divulgação, contempla processos bastante complexos. Empiricamente falando, realizar uma determinada tarefa para que posteriormente seja publicada, não se resume somente na maneira pela qual será exposta, mas sim se existem métodos que orientam na estruturação que vão além de uma simples veiculação de informações.

Nesse sentido, as boas práticas de publicação de dados são diretrizes que auxiliam na elaboração de subsídios para quem pretende publicar dados na Web, visando tornar os dados publicados com um nível superior de qualidade. A W3C (LÓSCIO; BURLE; CALEGARI, 2017), dispõe de 35 recomendações denominadas como boas práticas (*best practice*s) divididas em 12 categorias que norteiam a organização de dados de forma coerente para obter-se o melhor resultado durante a publicação de dados, apresentadas na Figura 1.

Figura 1: Apresentação das 12 categorias relacionadas as boas práticas.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

Como relatado, as doze categorias apresentadas na Figura 1 classificam as 35 boas práticas quanto aos pontos que elas atendem no processo de publicação dos dados. A seguir está apresentada, as principais características destas doze categorias:

1. Metadados: fornecimento de informações que colaboram para o entendimento do significado dos dados, auxiliando na realização de tarefas;
2. Licença de Dados: auxilia na avaliação e usabilidade do dado, permitindo que seja identificada a existência de restrições de compartilhamento ou reutilização;
3. Procedência de Dados e Qualidade: a origem da informação para poder conjecturar a qualidade do mesmo, se é confiável ou não, além da própria qualidade de dados, implica no modo em como o subsídio será utilizado;
4. Versionamento de Dados: a necessidade de se obter novas versões, pois ao longo do tempo, os dados podem mudar, e por isso é preciso disponibilizar essas novas versões;
5. Identificadores de Dados: visa auxiliar na localização e na identificação das informações;
6. Formatos de Dados: garante que se algum dado for alterado por mais de uma pessoa, esses vários formatos ajudam a evitar erros, além de economizar tempo e dinheiro;
7. Vocabulário de Dados: estabelece termos que podem ser empregados em algum aplicativo específico;
8. Acesso de Dados: concerne na facilidade das pessoas e das máquinas de terem acesso e aproveitarem os benefícios;
9. Preservação de Dados: quando um dado for removido da Web, o identificador deve ser preservado, além de serem fornecidas informações sobre o recurso arquivado;
10. *Feedback*: os comentários dos usuários são fundamentais para que uma publicação consiga atingir uma ampla gama de público de diferentes níveis de experiência;
11. Enriquecimento dos dados: contempla o processo de enriquecimento, para melhorar as relações dos dados, além de ter a possibilidade de tornar-se um subsídio ativo valioso.
12. Republicação: há requisitos que devem ser seguidos para que não haja problemas posteriores na republicação dos dados.

Todas estas categorias divididas em recomendações sobre melhores práticas são extremamente relevantes na estruturação dos dados que serão divulgados.

No que diz respeito às boas práticas de publicação científica é perceptível que essas orientações se aplicam aos diversos cenários existentes, mas vale ressaltar que a publicação científica tem as suas particularidades. Desta forma, além das recomendações da W3C, a divulgação dos dados referentes a resultados de pesquisa científica requer alguns cuidados para que seja feito corretamente, o que será discutido na próxima seção.

4 RESULTADOS E DISCUSSÕES: Aplicação das Boas Práticas no Âmbito dos Dados referentes a Resultados de Pesquisa Científica

As boas práticas de publicação de dados na Web consistem em um marco teórico e prático da aplicação dos princípios do *Linked Data*. A partir do desenvolvimento das boas práticas, a publicação dos dados se torna mais factível e materializável, seguindo diretrizes que definem como os dados devem ser publicados, apontando ainda os benefícios de se seguir cada uma das boas práticas.

Diante deste cenário, um caminho natural é aplicação das boas práticas nos diversos domínios do conhecimento, para que se possa ter, a médio prazo, uma quantidade razoável de dados publicados de acordo com as diretrizes do W3C. Uma consequência disto seria ter dados com uma qualidade superior, ou seja, que permita que mais aplicações utilizem e publiquem dados seguindo tais princípios, e como resultado possuir um ciclo que retroalimenta as próprias iniciativas de *Linked Data*.

Partindo desta perspectiva de aplicação das Boas Práticas nos diversos campos de estudos, identifica-se a necessidade de apontamentos sobre como elas podem ser empregadas em cenários específicos, identificando e pontuando suas especificidades em cada domínio que atua. Assim no que concerne aos dados referentes a resultados de pesquisas científicas, ou seja, informações sobre os artigos de periódicos e os trabalhos de eventos mais especificamente, deve-se verificar particularidades que devem ser consideradas na aplicação das Boas Práticas neste domínio.

Desta forma, na busca de se ter um material científico que explore as boas práticas no contexto da publicação de dados referentes a resultados de pesquisa científica, desenvolveu-se um estudo que investiga como cada categoria de boas práticas pode ser aplicada.

A seguir, uma sequência de quadros é apresentada, demonstrando como as boas práticas são categorizadas, onde a primeira coluna contém a boa prática em si e na segunda apresenta sua definição segundo a W3C (LÓSCIO; BURLE; CALEGARI, 2017). Após a apresentação dos quadros comentários são tecidos sobre o domínio e suas especificidades, de

forma a evidenciar a importância do emprego das boas práticas, além de ferramentas e tecnologias específicas para a aplicação destas diretrizes.

O Quadro 1 mostra as boas práticas na categoria metadados, apresentando três especificações.

Quadro 1: Boas práticas da categoria Metadados

Boas práticas	Definição W3C
Prover metadados	Prover metadados tanto para usuários humanos quanto para aplicações de computadores
Prover metadados descritivos	Prover metadados que descrevam os recursos gerais dos *datasets* e distribuições.
Prover metadados estruturais	Prover metadados que descrevam o esquema e a estrutura interna de uma distribuição.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

As boas práticas expostas no Quadro 1 são fundamentais, pois a aplicação de metadados é essencial para a representação e a recuperação de documentos de publicação científica. No *Linked Data*, a utilização de padrões como o Bibframe, o Dublin Core e outros com características RDF são essenciais, tanto para o fornecimento de metadados descritivos, que estão vinculados às próprias informações bibliográficas, quanto aos estruturais, relacionados às estruturas que promoverão uma melhor recuperação.

Arakaki (2016) aponta algumas propostas que disponibilizaram metadados em formatos de *Linked Data*, e que estão encontrando sucesso neste contexto. O autor ainda aponta alguns padrões que são aderentes ao domínio bibliográfico e que podem ser utilizados para os dados referentes a resultados de publicação científica.

A categoria subsequente trata das licenças dos dados e está exposta no Quadro 2.

Quadro 2: Boas práticas da categoria Licença

Boas prática	Definição W3C
Prover informações de licença de dados	Prover um link ou cópia do contrato de licença que controle o uso dos dados

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

No âmbito do *Linked Data* a apresentação da licença em que os dados estão sujeitos é de fundamental importância para explicitar como tais dados podem ser usados e eventualmente reutilizados. Nos dados de publicações científicas essa informação permanece necessária, visto que as diversas empresas e instituições publicam suas informações com níveis de licenças variados e, que podem ou não serem abertas. O DCTerms possui um termo compatível com o contexto em questão para a explicitação desta informação (dct:license). (DUBLIN CORE METADATA INITIATIVE, 2012).

Relacionado às próprias licenças, é necessário identificar a proveniência e a qualidade dos dados. Este atributo está também contemplado pelas boas práticas conforme e está descrito no Quadro 3.

Quadro 3: Boas práticas da categoria Proveniência e Qualidade

Boas prática	Definição W3C
Prover informações de proveniência de dados	Prover informações completas sobre a origem dos dados e as alterações feitas.
Prover informações de qualidade de dados	Prover informações sobre qualidade e adequação de dados para fins específicos.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

O Quadro 3 contempla duas boas práticas, mas que são fundamentais para o domínio tratado neste trabalho. Isto porque, pode haver a necessidade de alteração nos dados, garantindo que erros cometidos pelos autores ou mesmo publicadores sejam corrigidos, seja no próprio texto ou nos metadados. Desta forma, os dados devem apresentar essas possíveis alterações, para que os usuários e as máquinas sejam sinalizados, e não sejam conduzidos a erros, que façam que se utilizem dados anteriores a sua atualização.

Já a qualidade pode ser medidas por distintas métricas que apontam aos usuários a qualidade dos dados, questão versada na segunda boa prática do Quadro 3. No domínio de publicações, essas métricas são essenciais para apontar aos usuários humanos e não-humanos se aqueles dados têm um nível suficiente para serem utilizados.

Melo (2017) aponta uma metodologia que pode ser utilizada para avaliar a qualidade de dados em *Linked Data*, em especial no contexto dos dados referentes a resultados de publicações científicas, utilizando as métricas: *interlinking*, licenciamento, consistência, precisão sintática, precisão semântica, completude e avaliação temporal. Vale destacar que explicitar essas métricas no próprio *dataset* contribui para que os usuários obtenham com facilidade tal informação.

Outro ponto a ser considerado ao se publicar dados, refere-se ao seu versionamento. É classificada como uma categoria de boas práticas e indica como esse processo deve ocorrer; o quadro 4 apresenta essas informações.

Quadro 4: Boas práticas da categoria Versionamento

Boas práticas	Definição W3C
Prover um indicador de versão	Atribua e indique um número ou data de versão para cada conjunto de dados.
Prover o histórico da versão	Prover um histórico de versões completo que explica as mudanças feitas em cada versão.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

A atualização de versões é tanto inevitável como essencial para que um conjunto de dados se mantenha sempre atual. Neste contexto é de fundamental importância apontar em qual versão os dados estão, para que assim os usuários identifiquem se o conjunto de dados utilizado está atualizado, além de explicitar as alterações realizadas nestas atualizações, como é evidenciado nas boas práticas do Quadro 4.

Para os dados de publicação científica as versões são importantes, pois nas diversas esferas há atualizações de dados lançando novas informações. Como exemplo, um evento que realiza uma nova edição e publica os seus anais, ou uma revista que lança um novo número, sendo importante lançar a versão, apontar que é uma nova versão e afirmar quais foram as atualizações. Meinhart (2015) discute diversos apontamentos acerca do versionamento, demonstrando como criar uma plataforma que pode ser utilizada para o domínio de dados de publicação científica.

Em relação aos identificadores, que são centrais em quaisquer ambientes de *Linked Data*, apresenta um conjunto de três boas práticas que estão expostas no Quadro 5.

Quadro 5: Boas práticas da categoria Identificadores

Boas práticas	Definição W3C
Use URI persistentes como identificadores de conjuntos de dados	Identifique cada dataset por um URI cuidadosamente escolhido e persistente.
Use URI persistentes como identificadores dentro dos *datasets*	Reutilize os URIs de outras pessoas como identificadores dentro dos conjuntos de dados, quando possível.
Atribuir URIs para versões e séries de conjuntos de dados	Atribua URIs a versões individuais de conjuntos de dados, bem como à série geral.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

As boas práticas do Quadro 5 contemplam principalmente as URIs que são centrais no *Linked Data*, pois é por meio desses identificadores que é possível relacionar os dados com outros *datasets* e assim permitir a interoperabilidade. Neste sentido, os dados de publicações científicas devem estar normalizados utilizando identificadores expressivos para os diversos ambientes, isto porque, por vezes os dados de publicações podem estar presentes em *datasets* de revistas como também em *datasets* de repositórios, sendo assim de fundamental importância a utilização de URIs padronizadas e significativas.

Além disso, utilizar URIs contribui neste contexto, para facilitar a obtenção dos dados, o que faz com que o acesso aos publicadores de dados científicos seja realizado mais facilmente. Vale destacar que o uso de identificadores como o DOI e o ORCID auxilia nesta tarefa, uma vez que fornecem identificadores padronizados para a maioria dos centros

acadêmicos. Haak et al. (2012) discute e apresenta o ORCID como o identificador para pesquisadores, inclusive discutindo melhores práticas para o uso e para a interoperabilidade.

Juntamente com os identificadores, os formatos em que os dados são disponibilizados possuem um papel central na adequação dos dados aos princípios do *Linked Data*. Desta forma, o Quadro 6 expõem as boas práticas desta categoria.

Quadro 6: Boas práticas da categoria Formatos

Boas práticas	Definição W3C
Use formatos de dados padronizados legíveis por máquina	Disponibilize dados em um formato de dados padronizado e legível por máquina que seja adequado ao seu uso pretendido ou potencial.
Use representações de dados locais neutras	Use estruturas e valores de dados localmente neutros ou, quando isso não for possível, forneça metadados sobre a localidade usada pelos valores de dados.
Prover dados em múltiplos formatos	Disponibilize dados em múltiplos formatos quando mais de um formato se adequa ao seu uso pretendido ou potencial.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

Uma das bases para o uso efetivo do *Linked Data* é a disponibilização em formatos legíveis por máquinas, como indicado no Quadro 6. Essa característica permite com que agentes computacionais e algoritmos diversos sejam capazes de processar e analisar os dados mais rapidamente, explorando e relacionando as diversas variáveis. Além disso, a utilização de dados neutros, como datas em formatos padrões compreensíveis em todo o mundo, auxilia em tornar os dados utilizáveis em qualquer sistema, permitindo a disponibilização em vários formatos e contribuindo em abarcar um gama de sistemas mais amplo.

Essas três características no domínio de publicações científicas podem ser utilizadas com formatos como o RDF/XML e o JSON-LD, que contemplam atributos principais do *Linked Data* e da Web Semântica e podem ser utilizados pelos principais padrões de metadados compatíveis. Além disso, a disponibilização em XML e JSON contempla a maioria dos sistemas Web, que são facilmente acopláveis em sistemas de publicações científicas como, por exemplo o OJS e DSpace. No *website* do *software* de repositório DSpace (DURASPACE, 2017), há uma página com explicações referentes a relação entre este sistema e o *Linked Data*, apresentando como ele pode ser configurado nos padrões do *Linked Data*, utilizando os formatos de triplas (RDF).

A categoria seguinte, retratada no Quadro 7, trata de vocabulários.

Quadro 7: Boas práticas da categoria Vocabulários

Boas práticas	Definição W3C
Reutilizar vocabulários, de preferência padronizados	Use termos de vocabulários compartilhados, de preferência padronizados, para codificar dados e metadados.

Escolha o nível de formalização correto	Opte por um nível de semântica formal que se adapta aos dados e às aplicações mais prováveis.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

No âmbito de dados referentes a resultados de pesquisa, existe uma série de vocabulários que são reconhecidos e aceitos, contemplando assim as boas práticas apresentadas no Quadro 7. Em especial o DCTerms consegue abranger grande parte das informações que compõem esse domínio, necessitando apenas da complementação de vocabulários que são padrões para todos os dados, o que possibilita um grau de formalização adequado, em especial OWL, SKOS e RDF *Schema* e são capazes de fornecer aos conjuntos de dados a semântica exigida para cada cenário.

Schmachtenberg, Bizer e Paulheim (2014) ao apontar algumas boas práticas para publicação de dados no *Linked Data*, explicitam que na categoria de publicações, o vocabulário DCTerms é utilizado por mais de 83% dos *datasets*, o que demonstra a importância do uso destes vocabulários em domínios que estão vinculados aos dados de publicações.

O Quadro 8 está relacionado às boas práticas de acesso, que contempla também uma especificidade no que tange o acesso aos dados por meio de APIs.

Quadro 8: Boas práticas da categoria Acesso

Boas práticas	Definição W3C
Prover download em massa	Permita que os consumidores recuperem o conjunto de dados completo com um único pedido.
Prover subconjuntos para grandes conjuntos de dados	Se o seu conjunto de dados for grande, habilite usuários e aplicativos a trabalhar facilmente com subconjuntos úteis de seus dados.
Use a negociação de conteúdo para fornecer dados disponíveis em vários formatos	Use a negociação de conteúdo além das extensões de arquivo para fornecer dados disponíveis em vários formatos.
Fornecer acesso em tempo real	Quando os dados são produzidos em tempo real, disponibilize-o na Web em tempo real ou próximo de tempo real.
Forneça dados atualizados	Disponibilize dados de forma atualizada e torne explícita a frequência de atualização.
Forneça uma explicação para os dados que não estão disponíveis	Para dados que não estão disponíveis, forneça uma explicação sobre como os dados podem ser acessados e quem pode acessá-lo.
Disponibilize dados através de uma API	Ofereça uma API para fornecer dados se você tiver recursos para fazê-lo.
Use os Padrões da Web como base das APIs	Ao projetar APIs, use um estilo arquitetônico baseado nas tecnologias da própria Web.
Prover documentação completa para sua API	Forneça informações completas na Web sobre sua API. Atualize a documentação à medida que você adiciona recursos ou faz alterações.
Evite quebrar alterações na sua API	Evite alterações na sua API que quebram o código do cliente e comunique quaisquer alterações na sua API para seus desenvolvedores quando ocorrer a evolução.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

De modo geral, o acesso aos dados referentes a resultados de pesquisa deve ocorrer seguindo o padrão dos demais conjuntos de dados do *Linked Data*, acompanhando as boas práticas apontadas no Quadro 8. Assim, deve ser permitido aos usuários baixarem todo o conjunto de dados, além de possibilitar o acesso aos subconjuntos (que neste caso, poderá ser edições de revistas ou eventos, temáticas publicadas, entre outros) e disponibilizar os dados em diferentes formatos. No caso de resultados de pesquisa, não há dados em tempo real, mas sempre devem ser fornecidos dados atualizados, que conforme ocorre o lançamento de uma nova publicação, estes sejam disponibilizados o mais rápido possível, e caso algum dado não seja publicado, é necessário mostrar com clareza o motivo deste fato e como o usuário pode vir a encontrá-lo.

Neste contexto, o acesso aos dados referentes a resultados de pesquisa científica é essencial para a disseminação da informação e fundamental para que as publicações possam ser acessadas, lidas e citadas. O *Linked Data* se mostra como uma plataforma que pode colaborar com isso, assim, o acesso a esses dados devem seguir as diretrizes citadas, para que a utilização deles possa ocorrer.

As quatro últimas linhas do Quadro 8, apontam boas práticas do acesso aos dados por meio de APIs. Essa disponibilização, bem como a utilização de padrões Web, provém documentação adequada e evitam quebras de sua API, usando recomendações padrões para instituições que desejam publicar os seus dados na Web, em especial no formato do *Linked Data*. Neste contexto, ao publicar dados referentes a resultados de pesquisa científica é importante seguir tais recomendações, pois a utilização de APIs aprimora o acesso, possibilitando mais um meio para que esse processo ocorra, em especial por estar em conformidade com as normas que os principais mecanismos computacionais utilizam.

Lanthaler e Gütl (2012) apresentam e discutem como o JSON-LD, um formato para troca de dados compatível com os princípios da Web Semântica, pode ser base para o desenvolvimento de APIs e serviços Web, especificamente os serviços Restful. O uso do JSON-LD neste cenário ao entregar um formato reconhecido pela comunidade, específico para o *Linked Data*, contribui para a construção de serviços mais eficientes e expressivos.

Outro tópico integrante ao *Linked Data*, diz respeito às boas práticas visando a preservação, que está manifestado no Quadro 9.

Quadro 9: Boas práticas da categoria Preservação

Boas práticas	Definição W3C
Preservar identificadores	Ao remover dados da Web, preserve o identificador e forneça informações

	sobre o recurso arquivado.
Avaliar a cobertura do conjunto de dados	Avalie a cobertura de um conjunto de dados antes da sua preservação.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

A preservação dos identificadores é fundamental para que os dados estejam contextualizados e as suas possíveis relações realizadas não se percam, fruto de um despejo ou arquivamento dos dados, itens apontados no Quadro 9. Neste sentido, os dados referentes a resultados de pesquisa podem estar vinculados aos demais dados, sendo necessário que a preservação dos identificadores e a avaliação da dimensão que um conjunto de dados está relacionado, ocorra de modo com que os usuários consigam sempre navegar pelos conjuntos de dados e encontrar as informações desejadas. Neste contexto, o uso de identificadores permanentes como o DOI e o ORCID podem auxiliar, pois são mantidos por instituições que buscam essa preservação.

Obter o retorno dos usuários e dos demais editores é essencial na publicação de dados. Assim, as boas práticas que contemplem essa questão estão classificadas como *feedback*, apresentadas no Quadro 10.

Quadro 10: Boas práticas da categoria *Feedback*

Boas práticas	Definição W3C
Reúna comentários de consumidores de dados	Fornecer um meio facilmente descoberto para que os consumidores forneçam *feedbacks*.
Torne os comentários disponíveis	Torne os comentários dos consumidores sobre conjuntos de dados e distribuições publicamente disponíveis.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

Permitir com que os usuários sejam capazes de dar *feedbacks* sobre o conjunto de dados, pode aprimorar sua qualidade, uma vez que o editor dos dados é capaz de ver algum problema, ou retorno do impacto dos dados nos consumidores. Os dados referentes a resultados de publicação científica estão sujeitos e podem ser aprimorados a partir desses feedbacks, sendo que a aplicação das boas práticas pode contribuir na aproximação dos próprios resultados de publicação dos usuários, visto que eles poderão de alguma forma colaborar na manutenção desses dados.

Outro ponto importante para a publicação de dados no domínio do trabalho, tange ao enriquecimento dos dados visando explicitar relações e dar mais semântica aos dados. No entanto, este processo também deve seguir boas práticas, que estão expostas no Quadro 11.

Quadro 11: Boas práticas da categoria Enriquecimento

Boas práticas	Definição W3C

Enriqueça dados gerando novos dados	Enriquecer seus dados, gerando novos dados ao fazê-lo, aumentará seu valor.
Fornecer apresentações complementares	Enriquecer dados, apresentando-o de forma complementar, imediatamente informativa, como visualizações, tabelas, aplicativos da Web ou resumos.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

O enriquecimento dos dados referentes a resultados de publicação científica é capaz de relacionar as informações com diversas outras bases, possibilitando que a semântica do conjunto seja mais rica e explícita, pontos apresentados no Quadro 11. Neste contexto, este domínio específico pode se beneficiar pela definição de axiomas que permitem a futura geração de inferências dos dados, permitindo com que os agentes computacionais explorem e identifiquem relações ainda não explicitadas, bem como auxiliem a outras disciplinas, como a bibliometria, a encontrarem novas informações (como métricas e índices de citações) a partir da disponibilização e do enriquecimento dos dados. Como relata Lóscio, Burle e Calegari (2017) as técnicas de enriquecimento são complexas, mas há uma série de pesquisas tratando dessa temática, em especial a aprendizagem de máquinas que podem contribuir neste cenário.

Além disso, o enriquecimento dos dados no domínio tratado por esta pesquisa pode promover níveis de relacionamentos com outras bases de dados, inclusive com outros domínios, como *datasets* de domínios específicos como da saúde e das ciências da natureza. Tal possibilidade permite com que os dados dos resultados de publicações científicas, ou seja, os metadados dos artigos, estejam vinculados aos próprios dados produzidos nas pesquisas científicas.

Por fim, a última categoria de boas práticas trata da republicação, demonstrada no Quadro 12.

Quadro 12: Boas práticas da categoria Republicação

Boas práticas	Definição W3C
Fornecer comentários ao editor original	Deixe o editor original saber quando você está reutilizando seus dados. Se você encontrar um erro ou ter sugestões ou elogios, informe-os.
Seguir os Termos de Licenciamento	Encontre e siga os requisitos de licenciamento do editor original do conjunto de dados.
Cite a publicação original	Reconheça a origem de seus dados em metadados. Se você fornecer uma interface de usuário, inclua a citação visivelmente na interface.

Fonte: Adaptado (LÓSCIO; BURLE; CALEGARI, 2017).

A reutilização e republicação de dados originais podem ocorrer no contexto de resultados de publicação científica, pois informações complementares aos dados dos resultados em si podem ser utilizadas para enriquecer e traçar relações nestes dados. Desta

forma, a sinalização que usa tais dados, bem como seguir os termos de licenciamento e a citação a publicação original é essencial, para que assim os dados estejam seguindo as normas científicas e legais, e os editores dos outros conjuntos de dados saibam que os seus dados estão sendo utilizados em outros conjuntos.

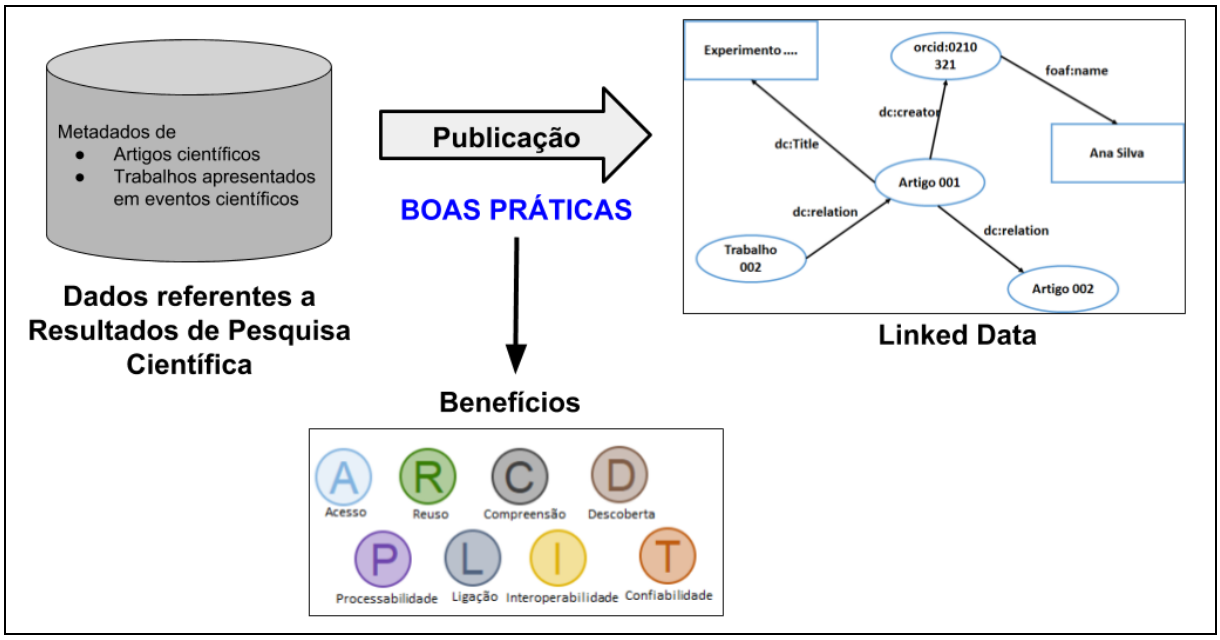
Por meio dos quadros e das explanações realizadas é possível visualizar uma série de citações que apontam casos ou considerações da aplicação dessas boas práticas no âmbito do domínio de dados referentes a resultados de pesquisas científicas. As referências citadas podem ser utilizadas para auxiliar na aplicação das boas práticas neste domínio, por apresentarem estudos detalhados dos pontos abordados. Neste contexto, o próprio documento das Boas Práticas (LÓSCIO; BURLE; CALEGARI, 2017) traz mais detalhes da aplicação destas diretrizes, porém de uma visão mais generalista.

As discussões apresentadas em cada categoria de boas práticas evidenciam o domínio dos dados referentes a resultados da pesquisa, apresentando argumentos que devem ser considerados na aplicação destas práticas neste cenário. Justifica-se o olhar nestes pontos, uma vez que as boas práticas recomendam ou citam inclusive o uso de tecnologias, de formatos e de sistemas que podem ser utilizados, e nestas questões cada domínio vai apresentar padrões, softwares e ferramentas distintos, que são reconhecidos pelas comunidades utilizadoras.

Neste âmbito, em diversos momentos das discussões são citados padrões de metadados, vocabulários e sistemas, como o BIbframe, o Dublin Core, o DCTerms, o DSpace e o OJS, que estão envoltos nas questões relativas à publicação de resultados de pesquisa científicas.

Diante dos pontos apresentados, é possível identificar diretrizes quanto à aplicação das boas práticas no âmbito dos dados referentes a resultados de pesquisa científica, especificando e detalhando peculiaridades da aplicação das boas práticas neste domínio. A Figura 2 esquematiza a aplicação do processo de publicação dos dados em *Linked Data*, a partir da aplicação das 35 boas práticas seguindo as particularidades apontadas anteriormente.

Figura 2: Esquema de aplicação das boas práticas na transformação dos dados referentes a resultados de pesquisa científicas em *Linked Data*



Fonte: Elaborado pelos autores.

No esquema apresentado na Figura 2 é destacado que o domínio da pesquisa abrange essencialmente metadados de artigos científicos e trabalhos apresentados em eventos, que estão a princípio dispostos em esquemas tradicionais, de banco de dados relacionais, que devem ser transpostos para o *Linked Data*. Esse processo ocorre por meio de técnicas que transformam os dados nos seus formatos para o RDF, seguindo os princípios destacados de cada vocabulário e padrão de metadados. Neste processo, a aplicação das boas práticas é fundamental para que a conversão ocorra sem perder os principais ganhos que o uso do *Linked Data* fornece aos dados.

Além disso, a Figura 2 demonstra os benefícios que a aplicação das boas práticas, considerando as particularidades do domínio demonstradas nos diversos quadros pode trazer. Tais benefícios são apresentados por Lóscio, Burle e Calegari (2017), como resultantes da aplicação das 35 boas práticas que são elas: compreensão, reuso, descoberta, interoperabilidade, acesso, ligação, confiabilidade e processabilidade.

Estes oito benefícios são essenciais na aplicação de todos os domínios, pois permite com que os dados em formato de *Linked Data* sejam contemplados por estas características, e consequentemente tenham utilidade nos domínios em que forem aplicados. Neste contexto, o foco dado neste trabalho ao domínio de dados referentes a resultados de pesquisa científica, destaca que cada domínio tem particularidades e a aplicação das boas práticas exige um estudo aprofundado do cenário de aplicação.

O presente trabalho buscou trazer especificações mais claras da aplicação destas boas práticas no domínio dos dados referentes a resultados de pesquisa científica, que consequentemente trará aos dados os oito benefícios que são introduzidos e discutidos por Lóscio, Burle e Calegari (2017).

5 CONSIDERAÇÕES FINAIS

A Web se tornou uma plataforma que modificou como os cientistas localizam, recuperam e acessam as produções acadêmicas, sendo necessário encontrar meios para que estes processos ocorram de forma mais eficiente, ou seja, que os usuários na Web sejam capazes de encontrar aquilo que necessitam e navegar entre os recursos que estão interligados.

Neste contexto, as bases de dados e iniciativas como o Google *Scholar* buscam facilitar o acesso a estas informações, mas partindo de um princípio sintático, em que a busca ocorre principalmente pelos termos descritos pelos usuários. Desta forma, é natural que se procure por meios que tornem o acesso mais fácil e inteligente, em especial em tempos que as tecnologias da Web Semântica estão cada vez mais maduras e eficientes.

Disponibilizar os dados em formatos aderentes ao *Linked Data* tem se mostrado como uma solução para o cenário apresentado, em que o uso das tecnologias da Web Semântica pode promover buscas com significado semântico, considerando o contexto dos dados tanto quanto o contexto dos usuários. Assim, o debate sobre como deve ocorrer o processo de publicação de dados seguindo os princípios do *Linked Data* é fundamental, pois traz diretrizes de como deve ocorrer e o que deve ser feito ao disponibilizar esses os dados.

Partindo das Boas Práticas publicadas por Lóscio, Burle e Calegari (2017), o presente trabalho buscou identificar como elas podem ser aplicadas ao cenário dos dados referentes a resultados de publicações científicas, identificando e tratando as particularidades que contornam este domínio. A partir disto, no presente trabalho pode-se verificar que todas as boas práticas são aplicadas e essenciais para o domínio em questão, apontando vocabulários, sistemas e literaturas que podem contribuir na publicação destes dados para o *Linked Data*.

Nas doze categorias de boas práticas discutiu-se pontualmente a relação entre os pontos apresentados pelos autores destas boas práticas, frente às especificidades do domínio. Ao final, foi apresentado um esquema que indica a relação entre a publicação dos dados para

formatos compatíveis com o *Linked Data* seguindo as boas práticas, com os oito benefícios que esse processo traz ao final.

Portanto, este trabalho discute e demonstra a aplicação de uma recomendação oficial da W3C recente, as Boas Práticas, apontando como isso pode ser efetivamente aplicado em casos reais. Partindo deste trabalho, tem-se como trabalhos futuros a publicação de conjuntos de dados de revistas científicas e anais de eventos vinculados aos autores seguindo as recomendações dadas durante esta pesquisa e do documento oficial das Boas Práticas.

REFERÊNCIAS

ARAKAKI, F. **Linked Data**: Ligação de Dados Bibliográficos. 2016. 146 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2016. Disponível em: <https://repositorio.unesp.br/handle/11449/147979>. Acesso em: 02 ago. 2017.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific american**, v. 284, n. 5, p. 28-37, 2001.

BERNERS-LEE, T. **Linked data principles**. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 02 ago. 2017.

DUBLIN CORE METADATA INITIATIVE. **DCMI Metadata Terms**. 2012. Disponível em: <http://dublincore.org/documents/dcmi-terms/>. Acesso em: 02 ago. 2017.

DURASPACE. **Linked (Open) Data**. 2017. Disponível em: <https://wiki.duraspace.org/display/DSDOC5x/Linked+%28Open%29+Data>. Acesso em: 02 ago. 2017.

HAAK, L. L. et al. ORCID: a system to uniquely identify researchers. **Learned Publishing**, v. 25, n. 4, p. 259-264, 2012. Disponível em: <http://www.ingentaconnect.com/contentone/alpsp/lp/2012/00000025/00000004/art00004?crawler=true&mimetype=application/pdf>. Acesso em: 02 ago. 2017.

LANTHALER, M.; GÜTL, C. On using JSON-LD to create evolvable RESTful services. In: Proceedings of the Third International Workshop on RESTful Design. **ACM**, 2012. p. 25-32. Disponível em: <https://pdfs.semanticscholar.org/ba69/b6c33792344fb189903792ec955af4aa0a98.pdf>. Acesso em: 02 ago. 2017.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. **Data on the Web Best Practices**. W3C, 2017. Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 02 ago. 2017.

LYNCH, C. **O quarto paradigma de Jim Gray e a construção do registro científico.** In: HEY, T.; STEWARD, T.; TOLLE, K. (Org.). O quarto paradigma: descobertas científicas na era da eSience. Tradução Leda Beck. São Paulo: Oficina de textos, 2011. p. 187-193.

MEINHART, P. **Versioning Linked Datasets**: Towards Preserving History on the Semantic Web. 2015. Dissertação (Mestrado em Engenharia de Sistema) – Hasso Plattner Institut, Universität Potsdam, Potsdam. 2015. Disponível em: <https://hpi.de/fileadmin/user_upload/fachgebiete/meinel/Semantic-Technologies/theses/Masterthesis-Meinhardt-2015.pdf>. Acesso em: 02 ago. 2017.

MELO, J. O. S. F. **Metodologia de avaliação de qualidade de dados no contexto do linked data**. 2017. 111 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2017. Disponível em: <https://repositorio.unesp.br/handle/11449/150870>. Acesso em: 02 ago. 2017.

MORIN, E. **Introdução ao pensamento complexo**. Porto Alegre: Sulina, 2005. 120p.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. Adoption of the Linked Data Best Practices in Different Topical Domains. In: MIKA, P. et al. (Eds.). . The Semantic Web – ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. **Proceedings**, Part I. Cham: Springer International Publishing, 2014. p. 245–260. Disponível em: <https://link.springer.com/chapter/10.1007%2F978-3-319-11964-9_16>. Acesso em: 02 ago. 2017.