Process Model with Quality Control for the Production of High Quality Linked Open Government Data

B. E. Penteado, J. C. Maldonado and S. Isotani, *Senior Member, IEEE*

***Abstract* — Governments worldwide have invested resources in publishing open data to promote new business and services and to promote transparency and accountability of public policies. However, due to different factors, such as different file formats and different information granularity, these data end up in informational silos, without having additional value besides what is contained in the data file itself. The linked data technology supports addressing this sort of problem, providing principles - based on Web standards - to connect distributed and heterogeneous data sources. Nevertheless, studies in the literature have shown that the perception of quality around open linked datasets is low, what impacts the consumption and reuse of these information by the society. In this work, we propose a model of process which embeds quality control activities (verification and validation) during the process. We adopted the Design Science Research (DSR) methodology to conceive and assess the method, using an illustrative case study. A quality assessment framework was applied and the case study's results were compared to others in the literature, showing an improvement in overall quality, through the selected metrics.**

***Index Terms*—linked data, open government data, process**

I. Introdução

Na última década (2010-2020), governos de vários países têm dedicado esforços para a produção de dados sobre seu funcionamento, no movimento chamado de dados abertos governamentais, baseado na filosofia de que esses dados são abertos para qualquer cidadão, sem restrições de uso [36]. Espera-se que a liberação desses dados para o público possa oferecer mais transparência e senso de responsabilização dos atores políticos por meio da fiscalização social, com valor social e comercial e participação dos cidadãos no processo democrático [1]. Mais países têm publicado dados em catálogos de dados em nível nacional, que têm sido usados por cidadãos, pesquisadores e empreendedores [2].

O Brasil foi um dos pioneiros nesta empreitada. Em 2004, foi lançado o Portal da Transparência, criado pela Controladoria Geral da União para publicar dados financeiros e de execução orçamentária do poder executivo, com o objetivo de combate a corrupção e o mau uso do dinheiro público. Em 2011, junto a outros 7 países, foi estabelecido um acordo multilateral chamado *Open Government Partnership* (OGP) com compromissos concretos para promover a transparência e criar planos nacionais de dados abertos a partir de diretrizes estabelecidas. Logo depois, foi aprovada a Lei de Acesso à Informação [3], que, dentre outros pontos, estabeleceu a obrigação dos entes públicos de fornecer informações aos cidadãos, de modo ativo ou passivo (ou seja, sem e com provocação por parte do público, respectivamente). Em 2012, foi lançado o Portal Brasileiro de Dados Abertos (www.dados.gov.br) que concentra os conjuntos de dados produzidos por todos os órgãos federais, concretizando um dos compromissos assumidos junto à OGP [37].

Com o aumento da disponibilização dos dados, surgiram outros desafios. Mesmo havendo diretrizes para a produção e formatação de dados abertos, como o e-PING [38] e as práticas estabelecidas pela Infraestrutura Nacional de Dados Abertos (INDA) [39], a descentralização das ações de liberação de dados faz com que os conjuntos de dados sejam produzidos em diferentes formatos, gerando silos de informações, que isolam os dados de outras fontes e impede que se crie conexões para atender a consultas de dados mais complexas [2]. Essa limitação restringe o potencial de reuso das informações, pois dificulta o cruzamento de informações de diferentes fontes, já que o processo de conhecimento das bases, limpeza dos dados, descoberta de conexões e interligação com outras bases de dados fica a cargo de quem os consome.

A abordagem dos dados conectados auxilia na solução deste problema, ao usar a arquitetura e padrões estabelecidos da Web para conectar dados relacionados e presentes nela. Para tornar os dados conectados, foram propostos os seguintes requisitos, reutilizando padrões da Web [5]:

- Usar URIs para identificar recursos;
- Usar URIs HTTP, de modo que as pessoas possam buscar e acessar informações sobre esses recursos;
- Quando alguém buscar por uma URI, deve-se fornecer informações úteis, usando padrões da Web (RDF, SPARQL); e
- Incluir links para outras URIs, permitindo que se possa descobrir mais recursos relacionados.

Esta pesquisa foi parcialmente financiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, CNPq (Processo 307887/2017-0) e FAPESP (Processo 15/24507-2). Bruno Elias Penteado, doutorando no programa de pós-graduação em Ciência da Computação no Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC/USP), São Carlos, 13566–590, Brasil (e-mail: brunopenteado@usp.br). Seiji Isotani é professor titular do Departamento de Computação no Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC/USP), São Carlos, CEP 13566–590, Brasil (e-mail: sisotani@icmc.usp.br).

Em adição às condições citadas, o autor também propôs um esquema com 5 níveis, denominado *5 estrelas*, para denotar o grau de abertura dos dados, variando desde a existência dos dados em formato digital até sua ligação semântica com outros repositórios [5]:

- 1 estrela: dados de qualquer formato, com licença de uso definida (ex.: PDF, imagens, HTML);
- 2 estrelas: dados estruturados legíveis por máquina (ex.: planilhas Excel);
- 3 estrelas: como o anterior, mas em formato não-proprietário (ex.: XML, CSV...);
- 4 estrelas: como o anterior, mas usando padrões da Web para representá-los (ex.: URIs, RDF, SPARQL);
- 5 estrelas: como o anterior, mas que esteja ligado a fontes externas, fornecendo maior contexto aos dados (LOD).

Deste modo, a abordagem de dados abertos conectados estende o conceito de dados abertos. Nos dados abertos, os dados estão publicamente acessíveis via internet, sem barreiras para acessá-los e reutilizá-los. Os dados abertos conectados, por sua vez, permitem que relacionamentos sejam expressos entre diferentes fontes de dados, enriquecendo, assim, o *dataset* com informações complementares vindas de outras fontes [6], tornando a Web uma base de dados global, chamada de "Web de dados". Esta diferenciação traz importantes desafios tecnológicos, como o alinhamento entre os dados das diferentes fontes, limpeza desses dados, armazenamento, diferentes granularidades, dentre outros. Por outro lado, esperam-se, com isso, benefícios como: maior contextualização dos dados com informações complementares e o processamento e interpretação automáticos por agentes de software.

Poucos conjuntos de dados governamentais estão disponibilizados em níveis 4 e 5 do esquema 5 estrelas [7][9], variando entre 2% e 5% em diferentes países pesquisados. No Brasil, ainda são poucos os conjuntos de dados disponibilizados em formato RDF e, em nosso conhecimento, nenhum deles faz ligação com fontes externas de dados (nível 5). Além disso, diversas investigações na literatura apontam que os dados disponibilizados apresentam problemas de qualidade. A percepção geral sobre os dados conectados é que os *datasets* que os compõem são de baixa qualidade para seus consumidores [10], devido a vários fatores – dificuldade em encontrá-los, interpretá-los, acessá-los, confiabilidade, etc. Essa percepção é relativizada em [11], ao argumentarem que os maiores problemas dos *datasets* conectados estão relacionados a um pequeno número de inconformidades a algumas diretrizes e melhores práticas. Na Referência [12], foi criado um arcabouço para a avaliação da qualidade de dados conectados e apontam-se alguns dos principais problemas com bases de dados conectadas em produção. Além disso, os autores argumentam que conjuntos de dados com problemas de qualidade apresentam severas implicações para os consumidores de dados abertos conectados.

O objetivo deste trabalho é abordar a geração de dados abertos governamentais conectados (DAGC) de qualidade, permitindo a realização de consultas mais complexas e extraindo mais valor das bases de dados na Web. Muitos trabalhos na literatura adotam metodologias *ad-hoc*, baseadas nos 4 princípios dos dados conectados e não seguem uma sequência sistemática de passos para atingir esse objetivo.

Neste trabalho, propomos um modelo de processo de produção de dados abertos conectados governamentais que embute atividades de verificação e validação. Com isso, busca-se lidar com os principais problemas levantados em pesquisas empíricas de qualidade de dados conectados.

II. Trabalhos Relacionados

Diversas metodologias e conjuntos de boas práticas para a produção de dados abertos conectados foram propostas na literatura. A W3C (*World Wide Web Consortium*), órgão responsável por definir as diretrizes de funcionamento da Web, publicou recomendações como *Best Practices for Publishing Linked Data* (LD-BP) [13], contendo dez passos macro para a produção e disponibilização desses dados na Web, desde a preparação dos envolvidos, a seleção dos dados a serem abertos, até a publicação e anúncio dos *datasets* junto ao público-alvo. Mais recentemente, foi proposta a recomendação *Data on the Web Best Practices* (DWBP) [14] que apresenta um conjunto de melhores práticas para o compartilhamento de dados online, explorando os padrões e a arquitetura básica da Web. A Referência [15] apresenta um mapeamento de conformidade desta recomendação com dados educacionais brasileiros, apresentando baixo atendimento às práticas. Outras recomendações foram desenvolvidas para tratar de características pontuais desse processo de publicação, como a *Cool URIs for the Semantic Web* [16], que trata do problema do desenho de URIs para representação das informações na Web, e a *Best Practice Recipes for Publishing RDF Vocabularies* [17], com sugestões de como publicar vocabulários e ontologias na Web.

Além dessas recomendações da W3C, outras metodologias foram propostas na literatura. Por exemplo, os trabalhos [18] e [19] serviram de base para a elaboração da recomendação LD-BP da W3C, contendo 6 e 5 fases, respectivamente. A Referência [20] traz uma perspectiva mais ampla, com uma fase para avaliação da qualidade, sem, no entanto, apresentar como ela se encaixa no processo. As metodologias propostas na literatura são compostas por aproximadamente o mesmo número de fases e atividades, variando de acordo com o contexto de cada aplicação. Na Referência [21] é proposto um método para que os consumidores possam fazer a conversão e interligação com diferentes fontes de dados. Porém, essa abordagem é ineficiente, já que faz com que cada consumidor dedique esforços para este fim, com cada consumidor podendo chegar a resultados distintos. Na Referência [22] é feita uma estimativa de esforços e competências necessárias para cada fase do processo (neste caso, 7 fases) baseado em diferentes estudos de caso de aplicação da metodologia. Já as Referências [23] e [24] dão ênfase na atividade de conversão de dados governamentais para o formato RDF ou conectado, reaproveitando os dados legados já publicados. No entanto, as outras fases do processo de produção não são devidamente exploradas.

Tais recomendações e metodologias são muitas vezes consideradas genéricas [25],[26], sem detalhar quais passos e ferramentas devem ser utilizados. Como resultado, a maioria dos estudos não especifica o uso de metodologias sistemáticas para a publicação dos dados [27]. Essas características são importantes para cenários em que a produção de dadosconectados encontra barreiras como falta de recursos e de mão de obra [28]. Algumas iniciativas buscaram mapear e desenvolver ferramentas para o ciclo de vida dos dados abertos conectados, como o *LOD Project* [40] e o *OpenGovIntelligence* [41],porém enfatizando as ferramentas. Além disso, em poucas dessas metodologias são encontradas atividades específicas de controle de qualidade do produto de dados resultante desse processo.

Assim, dada a percepção da baixa qualidade dos dados pelos consumidores, acreditamos que a inserção de atividades de validação e verificação no processo de produção dos dados pode influenciar positivamente na qualidade dos dados resultantes do processo.

III. Metodologia

Para a criação deste processo de produção de dados governamentais conectados de alta qualidade, foi adotada a metodologia de *Design Science Research* (DSR). Trata-se de uma abordagem prescritiva, que busca investigar a criação de artefatos como atividade humana na resolução de problemas de um domínio em particular – ao contrário das *hard sciences* (sociais ou da natureza), que buscam explicar e compreender fenômenos naturais [29]. A DSR foi escolhida pois fornece um arcabouço metodológico para a criação e avaliação de artefatos tecnológicos, como o desejado neste trabalho e está alinhada à questão de pesquisa posta. A DSR é composta por dois eixos [30]: as atividades de pesquisa e os produtos da pesquisa. As atividades de pesquisa compreendem os passos necessários para definir o problema (definido na introdução), projetar uma solução (Seção 4), avaliá-la (Seção 5) e aumentar a base de conhecimento sobre o problema baseado nos resultados (Seção 6). Como produto da pesquisa, o artefato pode assumir alguma das seguintes formas: construtos, modelos, métodos e implementações de sistemas.

Para a criação do método (modelo de processo) proposto, utilizamos uma abordagem integrativa, ao compilar as diferentes atividades das metodologias presentes na literatura. Essa escolha foi feita a partir da observação de que os passos para a produção de dados abertos governamentais conectados não se alteram muito entre as diferentes metodologias propostas. Pelo contrário, elas se complementam conforme são aplicadas em diferentes domínios e com graus diferentes de formalidade. As metodologias mais recentes já incluem algumas das boas práticas recomendadas pela W3C, mas não em sua totalidade. Em seguida, estendemos esse modelo de processo ao incluir passos de validação e verificação durante as atividades, voltados para os principais pontos de qualidade avaliados em dados abertos conectados. Foi escolhido como nível de granularidade as fases do processo, já que cada uma delas dispõe de artefatos que são usados em uma fase posterior durante o ciclo de produção dos dados conectados.

Para a avaliação do artefato de processo proposto, adotamos um estudo com cenário ilustrativo [31]. Este tipo de técnica tem por finalidade aplicar o método em cenários do mundo real, com o objetivo de ilustrar a utilidade do artefato. Para isso, foi selecionado um problema relevante e foram usadas bases de dados presentes na Web para responder a perguntas de pesquisa relacionadas a este problema. O domínio escolhido foi o educacional, um dos domínios em que o governo federal brasileiro disponibiliza há mais tempo suas bases de dados [32].

Para a avaliação de qualidade foi utilizado o referencial proposto em [11], que contém 27 métricas, agrupadas em 13 dimensões e 4 categorias, configurando uma avaliação quantitativa da qualidade dos dados. Esse referencial é uma derivação do trabalho de [12], mais amplo, com 69 métricas, 18 dimensões e as mesmas 4 categorias, que por sua vez, compilou tais métricas a partir de mapeamento sistemático com diferentes trabalhos que exploraram aspectos de qualidade de dados conectados. Ela foi escolhida por ser mais genérica a diferentes contextos, por focar a avaliação em métricas mais detalhadas e mais objetivas de serem avaliadas, facilitando sua replicação. Das 27 métricas desse referencial, 25 são herdadas de [12] e outras 2 concebidas a partir da recomendação DWBP da W3C [14]. A Tabela I mostra as métricas usadas na avaliação e sua classificação. Para detalhes sobre as métricas, consultar [11]. Ao final, computamos o valor agregado de todas as métricas, para estabelecer uma comparação aos *datasets* avaliados em [11].

O cenário para validação é a visualização de dados de movimentação e rotatividade docente nas escolas municipais de São Paulo. Trata-se de um problema complexo, que influencia significativamente no aprendizado dos alunos, ao ocasionar a descontinuidade do trabalho pedagógico de um professor em sua escola [33]. Para este cenário, os *datasets* mínimos são os seguintes: Censo Escolar da Educação Básica, com os arquivos de dados de Docentes e Escolas, publicado pelo INEP; o cadastro de escolas municipais, publicado pela Secretaria Municipal de Educação (SME) de São Paulo e que traz dados georreferenciados das escolas e os resultados municipais IDEB, também publicados pela SME de São Paulo. Este cenário foi escolhido para ilustrar como reutilizar dados mais genéricos (no caso, de nível federal) e estendê-los para um cenário particular (município de São Paulo). Foram adotados os *datasets* de 2016 e 2017 para demonstrar sua visualização.

IV. Modelo do Processo

O modelo de processo proposto, ilustrado na Tabela II, é composto por 6 fases: especificação, modelagem, conversão, publicação, exploração e manutenção. A seguir são detalhadas as fases e processos do modelo, em ordem sugerida de execução. As atividades sublinhadas são obrigatórias e necessárias para obter dados em nível 5 no esquema 5 estrelas.

As outras atividades podem ser consideradas complementares, mas sua execução aumenta muito seu potencial de reuso – um objetivo fundamental para a disponibilização de dados governamentais junto à sociedade,

TABLE I
MÉTRICAS PARA AVALIAÇÃO QUANTITATIVA DE QUALIDADE [11]

Categoria	Dimensão	Métrica
Representacional	Concisão representacional	RC1 - Manter URIs curtas
		RC2 - Uso mínimo de estruturas de dados RDF
	Interoperabilidade	IO1 - Reuso de termos existentes
	Interpretabilidade	IN3 - Uso de classes e propriedades indefinidas
		IN4 - Uso de *blank nodes*
	Versatilidade	V1 - Diferentes formatos de serialização
		V2 - Uso de múltiplas linguagens
	Proveniência	P1 – Informações básicas de proveniência
		P2 – Rastreabilidade dos dados
Contextual	Compreensibilidade	U1- Rótulos e comentários legíveis por humanos
		U3- Definição de URIs por expressão regular
		U5- Indicação dos vocabulários usados
Intrínsecas	Concisão	CN2 - Concisão extensional
	Consistência	CS1- Entidades como membros de classes disjuntas
		CS2- Classes ou propriedades mal colocadas
		CS3- Mau uso de propriedades OWL *Datatype* ou *Object*
		CS4- Uso de classes ou propriedades descontinuadas
		CS5- Uso válido de propriedade funcional inversa
		CS6- Sequestro de ontologia
		CS9- Uso de tipos de dados incorretos em *Domain* ou *Range*
	Validade sintática	SV3- Tipos de dados compatíveis
Acessibilidade	Disponibilidade	A3 – Resolução de URIs
	Licenciamento	L1 - Licença legível por máquina
		L2 – Licença legível por humanos
	Interligação	I1 – Links para fornecedores conectados externos
	Desempenho	PE2 – Alta vazão
		PE3 – Baixa latência

em que não há definições prévias de quem serão seus consumidores.

Na fase de *especificação*, deve-se levantar quais dados são de interesse da comunidade e que serão publicados em formato conectado, os metadados obrigatórios e opcionais que ele deve conter (dependendo do contexto do domínio), os vocabulários (ou ontologias, taxonomias) que deverão descrever esses dados e o padrão de URI a ser adotado pela

TABLE II
MODELO DE PROCESSO PARA A PUBLICAÇÃO DE DAGC.

Fase	Atividades	Saída
Especificação	• Identificar necessidades de dados • Selecionar fonte de dados • Definir bases externas • Levantar diretrizes institucionais • Especificar metadados • Especificar licenças de uso • Planejar padrão das URI • Criar/manter portal de dados	• Arquivos de dados • Diretrizes institucionais • Padrões técnicos
Modelagem	• Pré-processar dados • Normalizar dados • Criar dados • Reusar vocabulários • Criar novos vocabulários • Mapear semanticamente os dados	• Mapeamento dos dados para os vocabulários
Conversão	• Conectar a outras fontes • Enriquecer os dados • Converter para RDF • Limpar RDF	• Dados conectados em formato RDF
Publicação	• Versionar dados • Registrar metadados • Armazenar dados • Divulgar publicação	• Acesso aos arquivos (*endpoint* SPARQL, *dump* ou API)
Exploração	• Criar interface de dados conectados • Criar aplicações	• Aplicações funcionais consumindo dados conectados
Manutenção	• Definir requisitos não funcionais • Definir tarefas de manutenção • Engajar com a comunidade	• Plano de manutenção

organização, as licenças de uso aplicáveis, além de outros padrões institucionais cabíveis ao contexto governamental. Os portais de dados devem ser criados ou mantidos já apartir desta fase. A saída desta fase é o conjunto de especificações a serem adotadas ao longo do processo.

Na segunda fase, dá-se a *modelagem* dos dados. Primeiro, é necessário pré processá-los, de modo a filtrar dados incompletos ou inconsistentes (informações faltando, mal formatados, etc.), normalizá-los ou criar novos dados. Em seguida, são levantadas as necessidades de anotações semânticas que ainda não foram atendidas pelos vocabulários existentes e são criados novos vocabulários. Assim, o próximo passo é anotar os dados originais com sua respectiva anotação semântica. A saída desta fase são todos os mapeamentos entre dados limpos e vocabulários.

Na fase de *conversão*, os dados originais, já mapeados, são transformados para o modelo de dados semântico (RDF). Além disso, nesta fase é possível fazer a ligação com outras bases de dados externas, apontando para URIs desta outra base, usando o mesmo modelo de dados. Uma vez feita a ligação, é possível enriquecer os dados originais, ou seja, importar informações das bases externas, tornando-os mais contextualizados. A saída desta fase é um grafo RDF, com todos os relacionamentos mapeados, conectados externamente e enriquecidos.

Na fase de *publicação*, o conteúdo do grafo gerado é armazenado, seja por meio de upload em um website de catálogo de dados (ex.: CKAN), um arquivo *dump* ou de um *endpoint* SPARQL. Além disso, deve-se divulgar o ponto de acesso a essa fonte de dados, seja a URL do arquivo ou do *endpoint* SPARQL. Por fim, esse grafo deve ser versionado, seus metadados devem ser preenchidos. A divulgação em mecanismos de busca ou catálogos de dados (*data hubs*) também é um passo importante para que os dados sejam reutilizados. O produto desta fase é o acesso a um ou mais desses meios.

Na fase de *exploração*, aplicações são desenvolvidas ou aplicadas aos dados recém-criados. Trata-se de uma fase importante para o publicador, uma vez que por meio dela será feita a validação da qualidade do processo como um todo. O uso de interfaces de dados conectados (ex.: Pubby [42]) ajuda o usuário consumidor a explorar e compreender o que foi publicado, ao mesmo tempo que fornece serviços de negociação de conteúdo e resolução de URIs. O produto desta fase é um ambiente em que agentes de software e humanos possam localizar, explorar e consumir dados conectados.

Por fim, na fase de *manutenção*, deve-se criar um plano de manutenção para manter os dados publicados disponíveis ao público, levando em consideração requisitos não funcionais, como segurança, desempenho e disponibilidade. Os pontos de acesso aos dados (interface de dados conectados, API, arquivos ou *endpoint* SPARQL) devem ser verificados se estão *online*, com baixa latência e com alta vazão e manter um serviço que alerte sobre possíveis oscilações. Publicadores de dados devem engajar com a comunidade consumidora dos dados, para obter feedback e coletar sugestões de melhorias.

A Tabela III traz as extensões de verificação e validação para o modelo de processo, checadas ao final de cada fase. A Fig. 1 traz o modelo de processo, ilustrando cada fase e suas respectivas atividades (em azul), bem como as atividades de controle de qualidade (em laranja).

V. Resultados

A Tabela IV traz uma instanciação do processo a partir do modelo citado na Tabela II, e que foram necessários para realizar o cenário do estudo de caso em questão. Alguns passos do modelo não estão no escopo desta validação, em especial a fase de manutenção, e por isso não foram implementados e estão sinalizados com N/A. Como aplicação sobre os dados conectados, foi criada uma visualização de dados em grafos, uma maneira eficiente de ganhar consciência em dados estatísticos de alta complexidade [33]. Foram construídos grafos de movimentação, tendo como origem a escola em que o docente lecionava em 2016 e como destino a escola em que ele lecionou em 2017. Os docentes que não trocaram de escolas não foram representados. Foi utilizado o software Gephi [43] versão 0.9.2 para gerar as visualizações. A visualização gerada a partir desse mapeamento é ilustrada na Fig. 2, que mapeia todas as movimentações de docentes ocorridas entre 2016 e 2017, considerando apenas as escolas ativas dentro da rede municipal de São Paulo, ou seja, não são considerados os docentes que vieram ou foram para fora do município ou que saíram de escolas que fecharam no ano de 2016.

TABLE III
EXTENSÕES AO MODELO DE PROCESSO PROPOSTO.

Fase	Atividade
Especificação	• Documentações existentes: • Plano de ação para liberação dos dados; • Mapeamento de diretrizes; • Plano de comunicação para coleta de feedbacks; • Conjuntos de metadados obrigatórios e opcionais; • Formato válido de URIs; • Validação da estrutura dos arquivos (formatação de linhas e colunas ou objetos e propriedades); • Validação de URIs por expressão regular; • Bases de dados a serem conectadas e os respectivos pontos de ligação; • Todos os *datasets* contém dados sobre seu criador; • Bases de dados a serem conectadas e os respectivos pontos de ligação;
Modelagem	• Checar tamanho das URI modeladas; • Não usar estruturas RDF; • Evitar *blank nodes*; • Campos normalizados, com dados atômicos; • Especificação de novos dados; • Tratamento de dados faltantes; • Mapear quais vocabulários e termos serão reaproveitados; • Mapear todas as outras classes, termos e propriedades que deverão ser criados em um novo vocabulário; • Modelo de conversão entre dado de entrada e dado anotado, cobrindo todos os dados de interesse;
Conversão	• Metadados de diferentes níveis: do catálogo, dos *datasets* e suas distribuições, das classes e propriedades; • Adicionar expressão regular das URIs nos metadados; • Adicionar vocabulários usados nos metadados de cada *dataset*; • Correspondência entre o modelo de mapeamento e as triplas geradas; • Testar validade dos mapeamentos; • Conectar a outras fontes • Manutenção das estruturas após o enriquecimento; • Verificar resolução de todas classes e propriedades; • Checar se as *strings* estão definidas em todas as linguagens a serem suportadas; • Adicionar os formatos de serialização nos metadados; • Adicionar quem foi o responsável por gerar os dados nos metadados;
Publicação	• Pontos de acesso funcionais, autorizações a usuários configuradas; • Vocabulários, metadados e licenças publicadas; • Dados com versionamento comentado; • Licenças de uso atribuídas aos *datasets*, legíveis por humanos e máquinas; • Conformidade `as diretrizes institucionais; • *Sitemap* registrado;
Exploração	• *Datasets* anunciados ao público; • Interface de dados conectados funcional; • Aplicações criadas usando os dados publicados;
Manutenção	• Plano de manutenção elaborado; • Métricas de nível de serviço estabelecidas; • Serviços de monitoramento configurados; • Registro de feedbacks da comunidade;

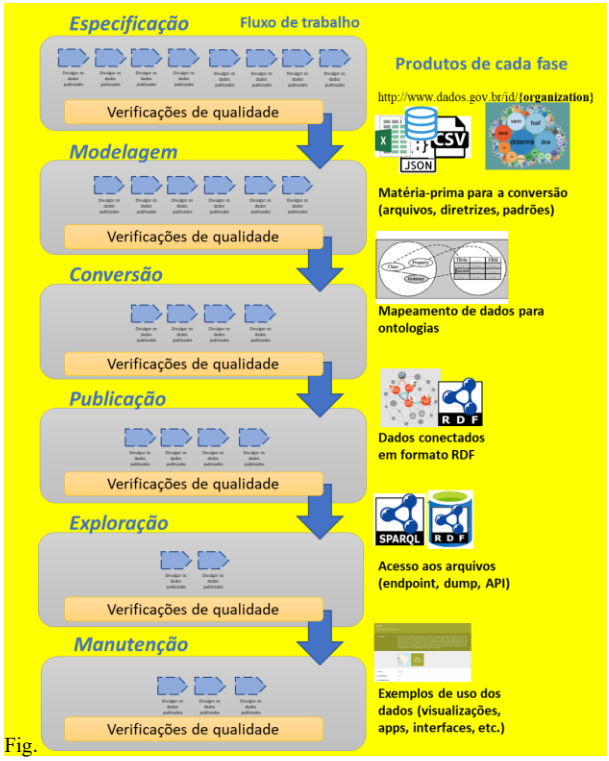

Fig. 1. Visão geral do processo, contendo: as 6 fases especificadas e suas respectivas atividades; as verificações de qualidade de dados abertos conectados, executadas ao final de cada fase e as saídas esperadas.

A Tabela V mostra os resultados (*Sc*) da avaliação da qualidade, levando em conta o referencial adotado e suas métricas *M*. Para efeito de comparação, é fornecido também o valor agregado (média) encontrado nas bases de dados investigadas no trabalho de [11], para cada uma das categorias de qualidade analisadas (M_i, M_r, M_c, M_a). As métricas assumem um valor entre 0 e 100% em relação à proporção de triplas que atendem ao critério dentro da base de dados, com exceção das métricas V1 e V2, que são respectivamente binária (presença de formatos disponíveis nos metadados) e inteira (quantidade de línguas atendidas).

Em estudos anteriores, as avaliações empíricas de qualidade de bases de dados conectadas foram feitas quase todas sobre grandes bases de propósito geral, como a Dbpedia e não em dados governamentais, como neste trabalho. O arcabouço de avaliação de qualidade analisa 27 métricas, relevantes na recente literatura sobre o tema. No entanto, cada aplicação pode adicionar ou remover métricas desta lista, de acordo com o que for valorizado em determinado contexto.

Ressaltamos três limitações deste trabalho. Primeiro, os escores das métricas podem estar superestimados, pois as métricas serviram como base para a construção de determinadas regras de V&V. No entanto, essas mesmas métricas são centrais em outros arcabouços de avaliação de qualidade de dados conectados, fornecendo assim uma generalização entre diferentes avaliações. Segundo, todo o processo, implantação e sua avaliação foram feitos em um único computador, com acesso local, sem estar publicado de fato na internet. Isso influenciou a medição das métricas P2 e P3, de desempenho computacional.

TABLE IV
INSTANCIAÇÃO DO PROCESSO PARA O ESTUDO DE CASO.

Categoria	Dimensão	Métrica
Especificação	Selecionar fonte de dados	Censo Educação Básica (Docentes e Escolas) de 2016 e 2017;
	Levantar diretrizes institucionais	e-PING, e-VOG, Cartilha técnica da INDA (documentos federais) e Plano de Transparência da Prefeitura de São Paulo
	Limpar dados	Os dados georreferenciados apresentaram problema de formação dos decimais de latitude e longitude
	Especificar metadados	Especificados na Cartilha Técnica da INDA
	Planejar padrão das URI	Política de URIs para publicação de dados no governo
	Identificar usuários potenciais	Pesquisadores na área educacional
	Definir bases externas	Cadastro de escolas do município de São Paulo; IDEB; municípios do IBGE. Identificadores compartilhados: ID da escola no INEP e ID do município no IBGE
Modelagem	Reusar vocabulários	Schema.org, FOAF, Dublin Core, Time, VoID, DCAT, RDFS, OWL, Provenance Model, XSD, DQV
	Criar novos vocabulários	Novo vocabulário, contendo as classes e propriedades específicas
	Normalizar dados	N/A
	Mapear semanticamente os dados	Mapeamento D2RQ de todos os campos para os termos dos vocabulários existentes ou ao novo
Conversão	Criar novos dados	N/A
	Converter para RDF	Criação de arquivos RDF, a partir da base de dados e o mapeamento semântico para os vocabulários
	Conectar a outras fontes	Conexão entre os datasets por meio de: ID da escola e ID do município
	Enriquecer os dados	Dados do IDEB e o Nível Socioeconômico das escolas
Publicação	Criar portal de dados	Acesso aos arquivos (*endpoint* SPARQL, *dump* ou API)
	Versionar dados	Datasets versionados a cada upload para o repositório de triplas
	Aplicar licenças	Replicação de licença usada no município de São Paulo
	Armazenar dados	Carregamento dos arquivos RDF para o repositório de triplas Apache *Fuseki*
	Divulgar publicação	Arquivo *sitemap.xml* gerado
Exploração	Criar interface de dados conectados	Configuração do software *LODView* para expor recursos/resolver URIs
	Criar aplicações	Visualizações de dados apresentadas a seguir
Manutenção	Definir requisitos não funcionais	N/A
	Definir tarefas de manutenção	N/A
	Engajar com a comunidade	N/A

Fig. 2. Grafo de transições dos docentes entre as escolas do município de São Paulo entre 2016 e 2017. Cada nó do grafo representa geograficamente uma escola e cada aresta representa a transição de uma escola para outra no período considerado.

TABLE V
RESULTADO DA AVALIAÇÃO QUANTITATIVA DE QUALIDADE, POR CATEGORIA (INTRÍNSECA M_I, REPRESENTACIONAL M_R, CONTEXTUAL M_C, E ACESSIBILIDADE M_A, RESPECTIVAMENTE) E SEUS ESCORES PARA AS MÉTRICAS DA TABELA I E SUA COMPARAÇÃO COM A LINHA DE BASE.

M_i	Sc	M_r	Sc	M_c	Sc	M_a	Sc
RC1	100%	P1	100%	CN2	100%	A3	100%
RC2	100%	P2	100%	CS1	100%	L1	100%
IO1	13%	U1	100%	CS2	100%	L2	100%
IN3	100%	U3	100%	CS3	100%	I1	100%
IN4	100%	U5	100%	CS4	100%	PE2	100%
V1	0	***Agg.***	***100%***	CS5	100%	PE3	100%
V2	1	***Base***	***8.7%***	CS6	100%	***Agg.***	***100%***
Agg.	***82%***			CS9	100%	***Base***	***33%***
Base	***61%***			SV3	100%		
				Agg.	***100%***		
				Base	***81%***		

VI. Discussão

Os resultados da avaliação da qualidade de dados conectados mostram uma melhoria significativa, principalmente nas categorias representacional e acessibilidade em relação às bases públicas investigadas em outros trabalhos empíricos sobre qualidade de *datasets* de dados conectados. O modelo de processo mostra-se flexível, com poucas atividades fixas e as atividades auxiliares podendo ou não serem instanciadas conforme o contexto da implantação dos dados abertos conectados – por ex., certas atividades dependem das características dos dados, conforme também demonstrado em [35]. No caso ilustrativo apresentado, duas atividades não foram utilizadas, pois não eram necessárias para a disponibilização dos dados nem para o cenário com o problema da aplicação. Isso não significa que o processo foi inválido ou que a qualidade do produto de dados foi de menor qualidade; apenas que o contexto não exigiu esse nível de rigor no processo.

Para uma medição mais fidedigna, essa avaliação deverá ser feita em ambiente Web, de modo que a comunidade possa acessar. Terceiro, não foi possível a avaliação de um caso de estudo real, pois nenhum *dataset* foi publicado oficialmente em nível 5 no Brasil. De toda forma, os resultados das atividades de controle de qualidade de dados conectados apresentaram um valor agregado ao processo, fazendo com que os dados resultantes atendessem aos requisitos de qualidade elaborados pela comunidade de conectados.

VII. Conclusão

Demonstramos, com o artefato proposto, a possibilidade de produzir sistematicamente dados abertos governamentais conectados de qualidade. O modelo de processo proposto mostra-se flexível para ser adotado em contextos com diferentes necessidades, com atividades podendo ser adotadas conforme o cenário de publicação. Com isso, estendemos a literatura ao introduzir atividades de verificação e validação durante o processo, sendo possível detectar, o mais cedo possível, desvios de qualidade do produto de dados resultante do processo. Como trabalhos futuros, o processo deverá ser refinado e testado em diferentes conjuntos de métricas, de acordo com a necessidade do cenário em que ele é aplicado.

Agradecimentos

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001, CNPq (Processo 307887/2017-0) e FAPESP (Processo 15/24507-2) pelo apoio financeiro.

Referências

[1] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic survey of open government data initiatives," *Government Information Quarterly*, vol. 32, no. 1, pp. 399–418, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.giq.2015.07.006.

[2] *WWW Foundation. "*Open Data Barometer – 4th edition*".* Available: *https://opendatabarometer.org/leadersedition/report/#introduction*

[3] *Brasil.* Lei de Acesso à Informação (Lei n. 12.527). Available: *http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm*

[4] *B. E. Penteado. "*Correlational Analysis Between School Performance and Municipal Indicators in Brazil Supported by Linked Open Data", in *25th International Conference Companion on WWW,, pp. 507-512, 2016. Available:* https://doi.org/10.1145/2872518.2890459

[5] T. Berners-Lee, "Linked data - design issues," https://www.w3.org/DesignIssues/LinkedData.html, 2008, accessed: 2019-06-06.

[6] N. Konstantinou and D.-E. Spanos, "Deploying linked open data: Methodologies and software tools," in *Materializing the Web of Linked Data*. Springer International Publishing, 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-16074-0_3.

[7] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzazanica, "Arethe methodologies for producing linked open data feasible for public administrations?" in International Conference *on Data Management Technologies and Applications* (KomIS-2014). Springer, 2014, pp. 399–407. [Online]. Available: http://dx.doi.org/10.5220/0005143303990407.

[8] R. Matheus, M. M. Ribeiro, and J. C. Vaz, "Brazil towards government 2.0: Strategies for adopting open government data in national and subnational governments," in *Case Studies in e-Government 2.0*. Springer International Publishing, 2014, ch. 8, pp. 121–138. [Online]. Available: https://doi.org/10.1007/978-3-319-08081-9_8.

[9] C. Alexopoulos, L. Spiliotopoulou, and Y. Charalabidis., "Open data movement in greece: A case study on open government data sources," in *17th Panhellenic Conference on Informatics*. ACM, 2013, pp. 279—286. [Online]. Available: http://dx.doi.org/10.1145/2491845.2491876.

[10] P. Hitzler and K. Janowicz, "Linked data, big data, and the 4th paradigm," *Semantic Web Journal*, vol. 4, no. 3, pp. 233–235, 2013. [Online]. Available: http://dx.doi.org/10.3233/SW-130117.
[11] J. Debattista, C. Lange, S. Auer, and D. Cortis, "Evaluating the quality of the lod cloud: An empirical investigation," *Semantic Web Journal*, vol. 9, no. 6, pp. 859–901, 2018. [Online]. Available: http://dx.doi.org/10.3233/SW-180306.
[12] A. Zaveri, A. Rula, R. Pietrobon, J. Lehmann, and A. Sauer, "Quality assessment for linked data: A survey," *Semantic Web Journal*, vol. 7, no. 1, pp. 63–93, 2016. [Online]. Available: http://dx.doi.org/10.3233/SW-150175.
[13] W3C, "Best practices for publishing linked data" https://www.w3.org/TR/ld-bp/, 2014, accessed: 2019-06-06.
[14] ——, "Data on the web best practices," https://www.w3.org/TR/dwbp/, 2017, accessed: 2019-06-06.
[15] B. E. Penteado, I. I. Bittencourt, and S. Isotani, "Análise exploratória sobre a abertura de dados educacionais no brasil: como torna-los prontos para o ecossistema da web?" *Revista Brasileira de Informática na Educação*, vol. 27, no. 1, 2019. [Online]. Available: http://dx.doi.org/10.5753/rbie.2019.27.01.175.
[16] W3C, "Cool uris for the semantic web," https://www.w3.org/TR/cooluris/, 2008, accessed: 2019-06-06.
[17] ——, "Best practice recipes for publishing rdf vocabularies," https://www.w3.org/TR/swbp-vocab-pub/, 2008, accessed: 2019-06-06.
[18] B. Hyland and D. Wood, "The joy of data: A cookbook fo publishing linked government data on the web," in *Linking Government Data*. Springer New York, 2011, ch. 1, pp. 3–26. [Online]. Available: https://doi.org/10.1007/978-1-4614-1767-5_1.
[19] B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez, "Methodological guidelines for publishing government linked data," in *Linking Government Data*, ser. Lecture Notes in Computer Science. Springer New York, 2011, ch. 2, pp. 27–49. [Online]. Available: https://doi.org/10.1007/978-1-4614-1767-5_2.
[20] A.-C. N. Ngomo, S. Auer, J. Lehmann, and A. Zaveri, "Introduction to linked data and its lifecycle on the web," in *Linking Government Data*, ser. Lecture Notes in Computer Science. Springer Cham, 2014, vol. 8714, ch. 1, pp. 1–99. [Online]. Available: https://doi.org/10.1007/978-3-319-10587-1_1.
[21] F. Maali, R. Cyganiak, and V. Peristeras, "A publishing pipeline for linked government data," in *The Semantic Web: Research and Applications. ESWC 2012*, ser. Lecture Notes in Computer Science. Springer Berlin, 2012, vol. 7295, ch. 59, pp. 1–99. [Online]. Available: https://doi.org/10.1007/978-3-642-30284-8_59.
[22] E. Klein, A. Gschwend, and A. C. Neuroni, "Towards a linked data publishing methodology." in *Conference for E-Democracy and Open Government* (CeDEM),, 2016, pp. 188–196. [Online]. Available: https://doi.org/10.1109/CeDEM.2016.12.
[23] T. Lebo, J. S. Erickson, L. Ding, A. Graves, G. T. Williams, D. DiFranzo, X. Li, J. Michaelis, J. G. Zheng, J. Flores, Z. Shangguan, D. McGuinness, and J. Hendler, "Producing and using linked open government data in the twc logd portal," in *Linking Government Data*. Springer New York, 2011, ch. 3, pp. 51–72. [Online]. Available: https://doi.org/10.1007/978-1-4614-1767-5_3.
[24] P. Salas, J. Viterbo, K. Breitman, and M. A. Casanova, "Stdtrip: Promoting the reuse of standard vocabularies in open government data," in *Linking Government Data*. Springer New York, 2011, ch. 6, pp 113–133. [Online]. Available: https://doi.org/10.1007/978-1-46141767-5_6.
[25] M. Laessig, B. Jacob, and C. AbouZahr, "Opening data for global health," in *The Palgrave Handbook of Global Health Data Methods for Policy and Practice*. Springer Berlin, 2019, ch. 23, pp. 451–468. [Online]. Available: https://doi.org/10.1057/978-1-137-54984-6n 2
[26] M. Jovanovik and D. Trajanov, "Consolidating drug data on global scale using linked data," *Journal of Biomedical Semantics*, vol. 8, no. 3, 2018. [Online]. Available: http://dx.doi.org/10.1186/s13326-016-0111-z.
[27] H. D. A. Santos, M. I. S. Oliveira, L. G. F. A. B., K. M. Silva, R. I. V. C. S. Muniz, and B. F. Lóscio, "Investigations into data published and consumed on the web: a systematic mapping study," *Journal of the Brazilian Computer Society*, vol. 24, no. 14, 2018. [Online]. Available: https://doi.org/10.1186/s13173-018-0077-z.
[28] J. Crusoe and U. Melin, "Investigating open government data barriers: A literature review and conceptualization," in *International Conference on Electronic Government EGOV'18*, ser. Lecture Notes in Computer Science, 2018, vol. 11020, pp. 169–183. [Online]. Available: https://doi.org/10.1007/978-3-319-98690-6_15.
[29] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *Management Information Systems Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.
[30] S. T. March and G. F. Smith, "Design and natural science research on information technology," *Decision Support Systems*, vol. 15, no. 4, pp. 251–266, 1995. [Online]. Available: http://dx.doi.org/10.1016/0167-9236(94)00041-2.
[31] K. Peffers, M. Rothenberger, T. Tuunanen, and R. Reza Vaezi, "Design science research evaluation," in *Design Science Research in Information Systems*, ser. Lecture Notes in Computer Science, 2012, vol. 7286, pp. 398–410. [Online]. Available: https://doi.org/10.1007/978-3-319-98690-6_15.
[32] B. E. Penteado and S. Isotani, "Dados abertos educacionais: que informações temos disponíveis?" in *VI Congresso Brasileiro de Educação*, 2017, pp. 1933–1938.
[33] *E. A. P. Junior, Oliveira, D. A. "Indicadores de retenção e rotatividade dos docentes da educação básica" Cadernos de Pesquisa, vol. 46, n. 6, p. 312-332 http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-15742016000200312&nrm=iso.*
[34] H. Paulheim, "Generating possible interpretations for statistic from linked open data," in *9th Extended Semantic Web Conference* (ESWC), 2012, pp. 560–574. [Online]. Available: https://doi.org/10.1007/978-3-642-30284-8_44.
[35] *B. E. Penteado, I. I. Bittencourt, S. Isotani (2019). "*Metaprocesso para transformação de dados educacionais em dados conectados*" in Simpósio Brasileiro de Informática Educacional*, p. 1601-1610*. Available:* http://dx.doi.org/10.5753/cbie.sbie.2019.1601.
[36] [Online]. Available: http://opendefinition.org
[37] [Online]. Available: https://www.opengovpartnership.org/process/joining-ogp/open-government-declaration/
[38] [Online]. Available: http://eping.governoeletronico.gov.br
[39] [Online]. Available: http://wiki.dados.gov.br/
[40] [Online]. Available: http://lod2.eu/
[41] [Online]. Available: http://www.opengovintelligence.eu/
[42] [Online]. Available: http://wifo5-03.informatik.uni-mannheim.de/pubby
[43] [Online]. Available: https://gephi.org

Bruno E. Penteado é doutorando em Ciências da Computação pelo do Instituto de Ciências Matemáticas e da Computação, da Universidade de São Paulo (ICMC/USP). Possui mestrado em Ciência da Computação pela Universidade Estadual Paulista (UNESP) na área de processamento de imagens e engenharia de software e graduação em Sistemas de Informação pela Universidade Estadual Paulista (UNESP). Trabalhou por 14 anos na área de tecnologias educacionais nas áreas de desenvolvimento de software, controle de qualidade e P&D.

José Carlos Maldonado é professor emérito do Instituto de Ciências Matemáticas e da Computação, da Universidade de São Paulo (ICMC/USP). Possui doutorado em Engenharia Elétrica pela Universidade de Campinas (Unicamp). Foi presidente da Sociedade Brasileira de Computação (SBC), entre 2007-2011. Tem experiência na área de Ciência da Computação, com ênfase em Engenharia de Software, atuando principalmente nos seguintes temas: teste de software, educação em engenharia de software, engenharia de software experimental, sistemas embarcados críticos, e ambientes e métodos de ensino.

Seiji Isotani é professor titular do Instituto de Ciências Matemáticas e da Computação, da Universidade de São Paulo (ICMC/USP). Concluiu seu doutorado em Engenharia da Informação na Osaka University (Japão). Realizou seu Pós-Doutorado em Ciências Cognitivas na Carnegie Mellon University (EUA) onde foi contratado e permaneceu no quadro docente até 2011. É fundador e atual co-coordenador do Laboratório de Computação Aplicada à Educação e Tecnologia Social Avançada (CAEd). É também co-fundador de duas empresas de base tecnológica (startups), sendo uma na área de tecnologias educacionais e outra na área de tecnologias semânticas, ambas premiadas em diferentes oportunidades pela produção e aplicação de inovações tecnológicas no setor.