Inf. Pauta Fortaleza, CE v. 3 Número especial nov. 2018 ISSN 2525-3468

DOI: https://doi.org/10.32810/2525-3468.ip.v3iEspecial.2018.39721.117-140

ARTIGO

WEB SEMÂNTICA: fluxo para publicação de dados abertos e ligados[i]

SEMANTIC WEB: flow for publishing open and linked data

José Eduardo Santarém Segundo [1]

[1] Doutor em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)-Marília/SP.

E-mail: santarem@usp.br

Conflito de interesses: O autor declara que não há conflito de interesses.

Financiamento: Não há

Declaração de Disponibilidade dos dados: Todos os dados relevantes estão disponíveis neste artigo.

Recebido em: 20/09/2018.
Revisado em: 01/10/2018.
Aceito em: 10/10/2018.

Como citar este artigo:
SANTAREM SEGUNDO, José Eduardo. Web semântica: fluxo para publicação de dados abertos e ligados. **Informação em Pauta**, Fortaleza, v. 3, número especial, p. 117-140, nov. 2018. DOI: https://doi.org/10.32810/2525-3468.ip.v3iEspecial.2018.39721.117-140.

RESUMO

Publicar dados em formato aberto e semântico tem se tornado um grande desafio as organizações ao redor do mundo. Há uma grande variedade de ações que devem ser executadas para que um projeto de publicação de dados possa ser concluído. Esta pesquisa tem como objetivo principal apresentar uma proposta de fluxo organizacional, segmentado em fases, que descreva as atividades que devem ser desenvolvidas no processo de publicação de dados em formato aberto e semântico seguindo as melhores práticas de dados ligados. A metodologia utilizada é baseada em pesquisa descritiva e analítica, baseada em análise documental. Como resultado espera-se que o fluxo apresentado possa contribuir com o desenvolvimento de novos projetos de publicação de dados em formato aberto e semântico.

Palavras-chave: Linked data. Web semântica. Publicação de dados. Dados ligados. Fluxo organizacional.

ABSTRACT

Publishing open and semantic data has become a major challenge for organizations around the world. There is a wide variety of actions that must be performed before a data publishing project can be completed. The main objective of this research is to present a phase-oriented organizational flow proposal that describes the activities that should be developed in the process of publication of data in an open and semantic format following the best practices of linked data. The methodology used is based on descriptive and analytical research, based on documentary analysis. As a result, it is expected that the presented flow can contribute to the development of new projects of publication of data in open and semantic format.

Keywords: Linked data. Semantic web. Data publishing. Connect data. Organizational Flow.

1 INTRODUÇÃO

A Ciência da Informação se transformou após a chegada da Internet, não há dúvidas que há uma revolução nos objetos de estudo e em grande parte dos processos em relação ao que era discutido algumas décadas atrás. Isso não significa impacto que altere as suas teorias, que continuam e continuarão totalmente aderentes as pesquisas realizadas atualmente, mesmo com as mudanças significativas que a Internet nos proporcionou, como pode ser visto nas citações a seguir.

Borko (1968, p. 3) afirma que "a Ciência da Informação é uma disciplina que investiga as propriedades e o comportamento da informação, as forças que governam seu fluxo, e os meio de processá-la para otimizar sua acessibilidade e uso".

Saracevic (1996, p. 43) diz que:

> [...] uma vez que a ciência e a tecnologia são críticas para a sociedade (por exemplo, para a economia, saúde, comércio, defesa) é também crítico prover os meios para o fornecimento de informações relevantes para indivíduos, grupos e organizações envolvidas com a ciência e a tecnologia, já que a informação é um dos mais importantes insumos para se atingir e sustentar o desenvolvimento em tais áreas. Posteriormente, essa justificativa, baseada na importância estratégica da informação, foi estendida a todos os campos, a todas as tarefas humanas e a todos os tipos de empreendimentos. Esta justificativa foi e é aplicada globalmente.

Fica evidente assim, que a Internet fortalece ainda mais o que já fora dito algumas décadas atrás. Nota-se ainda mais recentemente, como com Guimarães (2000), o valor estratégico da informação, independente do meio:

> Em tempos de informação com valor estratégico, cabe criar instrumentos que se adequem a uma concepção de disponibilização de conhecimento registrado para geração de novo conhecimento, em que a vertente temática assume papel preponderante, visto resgatar a essência do conteúdo informacional. (GUIMARÃES, 2000, p. 9)

Uma rápida análise nas bases de dados mais significativas da área nos mostra que muitos dos termos que nem existiam alguns anos atrás, agora são utilizados com frequência nas mais variadas subáreas da Ciência da Informação. Apesar deste texto se propor a tratar de um tema relacionado à tecnologia no contexto da Ciência da Informação, alguns termos têm tido significativo destaque de uma forma geral dentro da área, são eles: Web Semântica, Ontologias, Dados Ligados e Ciência dos Dados (*Semantic Web, Ontologies, Linked Data and Data Science*).

Esse relativo crescimento de interesse por esses termos, e consequentemente, pelos conceitos e tecnologias que envolvem Web Semântica, Ontologias, Dados Ligados, além claro, de uma nova era baseada em dados, tem pautado novos caminhos para os estudos da Ciência da Informação. A área entra definitivamente na rota de interesse de muitas outras áreas, que tem entendido que grande parte desses estudos competem e dependem de pesquisas realizadas estritamente na Ciência da Informação, com significativo apoio da Ciência da Computação.

Dentro deste contexto de estudo há uma relação de proximidade e às vezes de possível conflito entre estudos que vem sendo realizados na Ciência da Informação e na Ciência da Computação, entretanto, há uma clara diferença em como essas duas áreas podem contribuir diretamente nas pesquisas. Com importância, registre-se que muitos pesquisadores que atualmente trabalham com estes temas, tem suas raízes na Ciência da Computação e atualmente atuam na Ciência da Informação; enquanto outros tem sua formação básica em Biblioteconomia, Arquivologia e/ou Museologia e posteriormente migraram para a Ciência da Computação, em geral para uma ala mais aplicada e menos pura de estudos da Computação. Dessa forma, sim, estudos relacionados a Dados, Web Semântica, Ontologias e Dados Ligados precisam ser compartilhados entre pesquisadores das duas áreas.

Os últimos anos têm sido bastante significativos em como as tecnologias da Web Semântica e as possibilidades propostas pelas práticas de Dados Ligados tem evoluído e refletido diretamente numa crescente necessidade de se publicar dados. Os dados governamentais de alguns países, disponibilizados em formato aberto e semântico, tem tido impacto perante a sociedade e despertado um conjunto de iniciativas pelo desenvolvimento de aplicações que possam efetivamente levar o cidadão a consumir esses dados para os mais variados propósitos no seu dia a dia.

Há uma variada gama de aplicações que podem ser acessadas via browser, ou então por meio de aplicativos para dispositivos móveis, que se utilizam de dados publicados de forma aberta e semântica. Há também iniciativas de esforços para que mais aplicações sejam desenvolvidas no intuito de consumir dados que passam a ser publicados pelas mais variadas fontes.

Além dos dados de governo, há uma clara movimentação de interesse de algumas comunidades em publicações de informações de uso geral, que inclusive envolvem dados variados e de vários segmentos, algumas bases de dados, como por exemplo:

DBPedia, Wikidata, Bio2RDF, Europeana, Unesco e bases relativas a dados de mídias sociais tem crescido constantemente tanto em tamanho quanto em uso.

Todo esse novo contexto de publicação de dados não pode ser tratado exatamente como uma novidade, há muitas teorias e pesquisas, algumas nem tão mais recentes, que tem pautado os estudos baseados em dados como um novo paradigma de pesquisa (quarto paradigma da ciência) e desenvolvimento, e também por isso é notável um grande interesse por vários segmentos de comunidades distintas em também publicar seus dados.

O chamado quarto paradigma da ciência, e-Science ou ainda Data-Driven Science, que entende os dados como grande aliado e impulsionador para o avanço da ciência moderna fora previsto por Jim Gray em 2007 (HEY; TANSLEY; TOLLE, 2009), e tem estado cada dia mais presente nas ações do mundo atual.

O processo de publicação de dados, que em diversas situações parece uma tarefa trivial, tem se tornado o grande problema das equipes ou pessoas que se propõe a realiza-lo, e é esse o ponto que tem justificado fortemente o desenvolvimento desta pesquisa. Esse problema motivou também este pesquisador a oferecer uma disciplina chamada "Conceitos e Tecnologias para Publicação de Dados Abertos e Semânticos seguindo as melhores práticas do Linked Data" no Programa de Pós-Graduação em Ciência da Informação da Unesp de Marília, tento tido uma grande audiência nas oportunidades em que a disciplina foi oferecida.

Desta maneira, o objetivo principal desta pesquisa e apresentar uma proposta de fluxo organizacional, segmentado em fases, que descreva as atividades que devem ser desenvolvidas no processo de publicação de dados em formato aberto e semântico, seguindo as melhores práticas de Dados Ligados.

Os objetivos específicos são:

- Descrever um processo completo para publicação de dados em formato aberto e semântico;
- Identificar os atores que farão parte do processo.
- Segmentar o processo em fases, identificando cada uma delas de forma que possa ficar mais claro aos publicadores de dados todo o processo;
- Apresentar, em forma de diagrama, o processo de publicação de dados em formato aberto e semântico.

Importante ressaltar que não é objetivo deste trabalho selecionar ou indicar ferramentas nem tampouco orientar em como usá-las em cada processo, visto que este é um procedimento que depende muito do andamento do processo e dos objetivos da equipe de publicação de dados.

A metodologia utilizada para construir essa pesquisa foi baseada principalmente em análise de literatura nacional e internacional, principalmente as dedicadas a apresentar estudos de casos, além da experimentação de ferramentas e técnicas de publicação de dados. Assim, consideramos como uma pesquisa descritiva e analítica, com base em análise documental.

Espera-se que o fluxo organizacional, apresentado como resultado desta pesquisa, possa contribuir para que mais dados sejam publicados seguindo as melhores práticas de Dados Ligados. Espera-se ainda que os respectivos donos, gestores, responsáveis pela custódia ou pessoas que tenha qualquer outro tipo de relação com dados passíveis de publicação, possam encontrar nos resultados dessa pesquisa, os caminhos necessários para facilitar o processo de publicação de dados.

2 WEB SEMANTICA E DADOS LIGADOS

Desde 2001, quando Berners-Lee, Hendler e Lassila publicaram o primeiro texto sobre Web Semântica, onde diziam que "A Web Semântica é uma extensão da Web atual em que cada informação é dada por um significado bem definido, fazendo com que computadores e pessoas trabalhem melhor em cooperação", houve uma evolução constante dos processos e tecnologias que permitem que a Web Semântica atualmente faça parte da nossa vida cotidiana.

Para disponibilizar dados numa estrutura semântica é necessário pensar em partes do modelo descrito por Berners-Lee em 2001, no chamado bolo de noiva, estrutura de camadas que apresenta a Web Semântica. Destaca-se neste quesito a linguagem RDF, também indicada para representação de dados abertos, o uso de metadados e principalmente a construção e aplicação de ontologias de domínio. (SANTARÉM SEGUNDO, 2015).

Em 2006 Berners-Lee publicou um conjunto de princípios para publicação de dados usando as tecnologias da Web Semântica, que chamou de Linked Data (Dados

Ligados). Esses princípios, que representam a materialização da Web Semântica, são regras para publicação de dados, de forma que estes possam ser mais facilmente recuperáveis e possam estar ligados entre si:

- Usar URIs como nomes para os itens.
- Usar URIs HTTP para que as pessoas possam consultar esses nomes.
- Quando alguém consulta uma URI, prover informação RDF útil.
- Incluir sentenças RDF com links para outras URIs, a fim de permitir que itens relacionados possam ser descobertos.

Grande parte das pesquisas e projetos nos últimos anos se dedicaram principalmente a infraestrutura de organização e recuperação de dados em formato semântico, entretanto é sempre importante lembrar que a Web Semântica tem um papel social muito importante, é por meio dela que agentes computacionais (softwares, bots, aplicativos) podem realizar tarefas para facilitar a vida diária dos seres humanos.

Quando falamos de Web Semântica, falamos de uma mistura de interoperabilidade; padronização, organização e reuso da informação; inferências e de serendipidade.

A serendipidade se refere a descobertas feitas ao acaso, capacidade que as tecnologias da Web Semântica e principalmente do Dados Ligados trazem à tona e possibilitam através da ligação semântica entre dados de fontes diversas espalhadas pelo mundo. Enquanto a inferência diz respeito a capacidade de se deduzir ou tomar decisões, baseadas na consolidação de uma verdade de uma proposição que não é conhecida, mas é tida a partir de sua relação direta com outras verdades existentes, podendo ser considerada uma das cerejas do "bolo de noiva da Web Semântica". (SANTARÉM SEGUNDO; CONEGLIAN, 2016).

3 DADOS ABERTOS, DESAFIOS E REQUISITOS PARA PUBLICAÇÃO DE DADOS

É difícil desassociar dados ligados e dados abertos, são temas que apresentam muitos pontos em comum, principalmente quando se pensa que a primeira proposta de melhores práticas para ligar dados seja de Tim Berners-Lee. Entretanto sabe-se que os processos de ligação de dados podem ser aplicados a dados privados.

Apesar de ainda ser difícil pensar em dados ligados de forma privada, há muitas bases de dados que usam os princípios de ligação de dados para gestão de dados de forma restrita.

Esta pesquisa aborda a publicação de dados em formato aberto e semântico, tendo como ideal o uso de dados abertos e passíveis de consumo pela comunidade, portanto todo o contexto aqui apresentado é pensando na publicação de dados abertos.

O acesso à informação tem sido pautado como grande propulsor do desenvolvimento no século XXI. As instituições, sejam elas públicas ou privadas, tem investido na organização e no acesso à informação como o grande diferencial na tomada de decisão em várias de suas instâncias.

Há também no mundo uma tendência de publicação de dados governamentais, com o objetivo de criar a cultura de participação do cidadão na gestão do Estado, construindo um modelo conhecido como transparência.

Atuando desde 2004 a *Open Knowledge Foundation* tem se dedicado a trabalhar com projetos que envolvem o conceito de conhecimento aberto. Segundo eles "Conhecimento Aberto é qualquer informação, conteúdo ou dados que as pessoas são livres para utilização, reutilização e redistribuição - sem qualquer restrição legal, tecnológica ou social".

O movimento de abertura de dados governamentais está embasado em 3 leis propostas pelo especialista em políticas públicas David Eaves (2009):

- Se o dado não pode ser encontrado e indexado na web, ele não existe.
- Se não estiver aberto e em formato compreensível por máquina, ele não pode ser reaproveitado.
- Se algum dispositivo legal não permitir sua reaplicação, ele não é útil.

Apesar da clara necessidade de uso, dados abertos, especialmente os governamentais, constituem-se como um ótimo recurso, ainda timidamente explorado. Muitos indivíduos e organizações coletam uma ampla gama de diferentes tipos de dados para executar suas tarefas. O governo é particularmente importante nesse contexto, tanto por causa da quantidade e da centralidade dos dados que coleta quanto pelo fato de que tais dados são públicos, um direito garantido no artigo 5º da Constituição Federal brasileira (MANUAL..., 2011).

Não estamos tratando aqui apenas de dados governamentais, mas entende-se que eles são uma grande parte dos dados que se deseja publicar. Há também muitos outros

dados, geridos pelos mais derivados entes, que precisam e poderiam ser publicados, entretanto há uma série de fatores que implicam em desafios e requisitos para que possam ser publicados.

Quais são as questões que envolvem diretamente um projeto de publicação de dados? Quais são exatamente os passos para se publicar dados em formato aberto e semântico? Alguns desafios importantes, para que um processo de publicação de dados em formato aberto e semântico possa acontecer, podem ser facilmente listados por equipes ou pessoas responsáveis por tais atividades, como as questões que seguem.

Será que instituição ou organização que mantém os dados tem intenção ou um plano para publicar dados? Existe um modelo padronizado que uma equipe possa usar para publicar dados? De que setor da instituição/organização é a equipe que ficará responsável pelo processo de publicação dos dados? Como sincronizar os interesses de quem publica e do público que vai consumir os dados, ou seja, será que os dados que tenho disponíveis atendem efetivamente quem gostaria de consumi-los? Quais dados serão publicados? Onde são gerados (fonte) os dados que quero publicar, com que frequência eles são gerados? Os dados podem ser disponibilizados? Que licença devo usar para publicar meus dados? Como tornar os dados interoperáveis? Quais formatos de dados utilizar? Como restringir ou permitir acesso quando os dados forem sensíveis? Como transmitir confiança a quem vai consumir os dados (qualidade e proveniência)? Como garantir a preservação dos dados? Como enriquecer os dados, com quais outras bases se conectar? Como garantir e propor que os dados possam ser usados e reutilizados? Como obter feedback a partir do uso dos dados publicados? Essa são apenas algumas das perguntas que fazem parte de um projeto para publicar dados em formato aberto e semântico.

Esta pesquisa não tem a intenção de responder a todas essas perguntas, pelo contrário, essa é uma tarefa que precisa ser tratada em partes, entretanto é necessário que esse tipo de atividade tenha um mínimo de organização, de procedimentos, e principalmente de fluxo organizacional que possa dar uma linha de condução ao processo de publicação de dados.

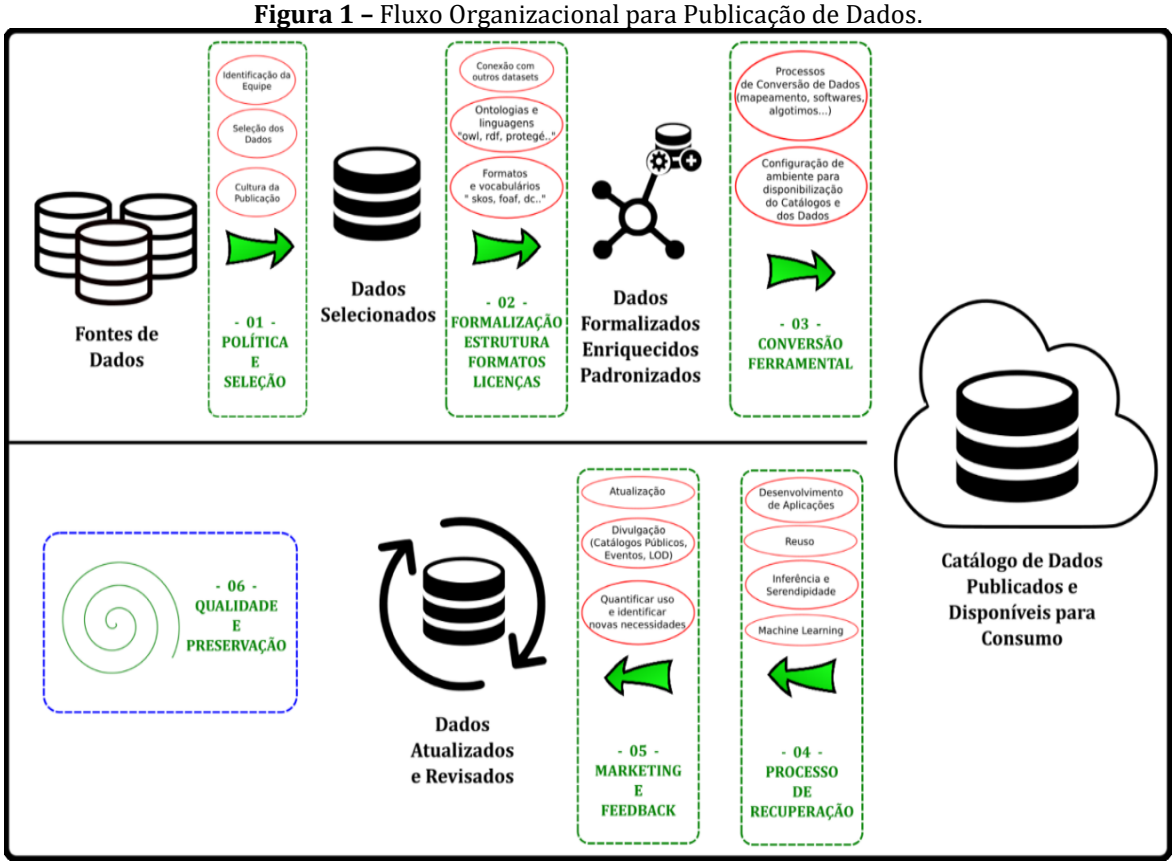
4 FLUXO ORGANIZACIONAL PARA PUBLICAÇÃO DE DADOS

Há uma variedade grande de *papers* que apresentam procedimentos para publicação de dados, dos mais variados tipos, das mais variadas fontes e usando uma grande gama de ferramentas diferentes.

Grande parte dessas pesquisas tem forte apelo no uso de uma ou outra ferramenta, e grande parte das vezes gira em torno do funcionamento da mesma, não havendo uma preocupação com todo o processo de publicação de dados.

Por meio da figura 1 apresenta-se a sugestão de modelo de um fluxo organizacional, segmentado em fases, que organiza o caminho por qual um projeto de publicação de dados deve passar.

Figura 1 – Fluxo Organizacional para Publicação de Dados.



Fonte: Dados da pesquisa.

Entende-se que um fluxo organizacional de publicação de dados, que pode ser chamado de projeto de publicação de dados, não é simples, e pode envolver várias pessoas ou divisões de uma organização.

Por meio da figura 1 apresenta-se a proposta de divisão do fluxo organizacional com várias tarefas, que se organizam divididas em 6 fases:

1 – Política e seleção;

2- Formalização, Estrutura, Formatos e Licenças;

3 – Conversão Ferramental;

4 – Processo de Recuperação;

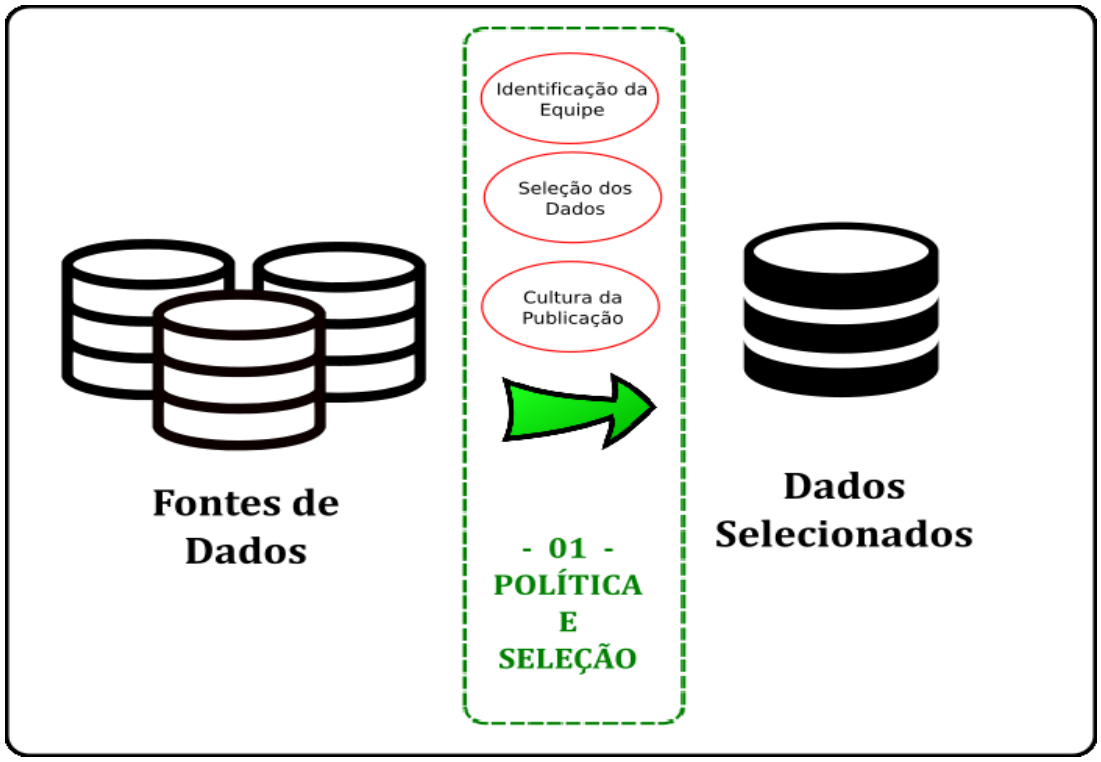
5 – Marketing e Feedback;

6 – Qualidade e Preservação.

4.1 Política e Seleção

A primeira fase, vista por meio da figura 2, de qualquer projeto de publicação de dados vem muito antes da parte técnica, o que as vezes dificulta o processo quando esse nasce dentro da área de TI de uma organização.

Figura 2 – Fase 1 do fluxo organizacional de publicação de dados



Fonte: Dados da pesquisa.

Mesmo que exista uma grande massa de dados que poderia ser publicada de forma aberta e semântica, ou ainda que seja apenas uma pequena fatia de uma base de dados, é importante inicialmente identificar quais os dados que realmente serão

publicados. Ressalta-se que o procedimento de publicar dados, depois de iniciado, deve ser contínuo, ou seja, deverá fazer parte da rotina da organização.

Portanto uma das primeiras tarefas é organizar um grupo de pessoas, uma equipe, de preferência multidisciplinar, com capacidade e responsabilidade técnica e administrativa para executar e tomar decisões sobre o tema, e que possa conduzir todo o processo, além de dar conta de constituir uma cultura de publicação de dados na organização.

Constituir uma cultura de publicação de dados, é levar ao conhecimento de todos os colaboradores de uma instituição, o ideal de divulgar dados de forma aberta na internet, e conscientizar que esse tema deve ser discutido frequentemente, e que pode ser rediscutido a qualquer momento. Importante que fique claro que para algumas situações será necessário algum tipo de esforço de pessoas específicas, para que os dados possam ficar disponíveis.

É importante que o ideal de disponibilizar dados seja uma intenção da organização, independente se a necessidade é por desejo de publicação de dados ou por força de lei.

A partir do momento que se tem uma equipe, é necessário identificar o público que pode ter interesse nos dados que serão disponibilizados, quais dados da organização serão disponibilizados, qual é a granularidade do dado que será entregue e, principalmente, quais são os colaboradores da empresa que tecnicamente darão acesso ou entregarão frequentemente os dados a serem publicados.

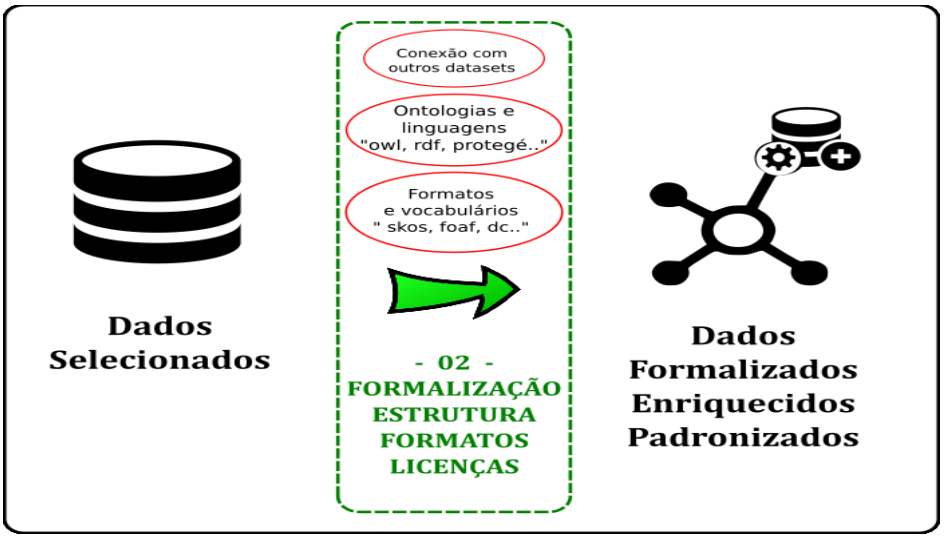
Essa fase estará completa quando for possível ter uma equipe responsável pelo projeto e um pacote de dados que seja um exemplo ou que represente pelo menos parte do que será publicado. Esse pacote de dados pode ser um arquivo, um canal de consulta, uma abertura por API (*Application Programming Interface*) ou qualquer outro tipo de tecnologia que permita com que a equipe responsável tenha uma porção dos dados para trabalhar.

4.2 Formalização, Estrutura, Formatos e Licenças

A segunda fase é o momento em que o dado já selecionado receberá todo o tratamento técnico necessário para que possa efetivamente ser publicado de forma aberta e semântica.

Conforme pode ser visto por meio da figura 3, é nessa fase que acontecerá uma transformação no dado, incluindo, quando necessário, uma mudança de formato, a ligação com outros dados para que possa ser enriquecido com a ligação com outra base de dados aberta e onde serão atribuídas licenças de uso, para que a comunidade que vai consumir tenha certeza de que pode utilizar o dado.

Figura 3 – Fase 2 do fluxo organizacional de publicação de dados



Fonte: Dados da pesquisa.

É nessa fase que se atribuem aos dados as características técnicas que o transformam em semânticos. Também é nesta fase, que ao finalizada, teremos o dado no formato que será disponibilizado para a comunidade.

Destaca-se aqui nesta fase uma grande necessidade de trabalho técnico, que dependerá muito da equipe de TI da organização. O interessante da fase 2 é que ela normalmente é uma fase de destaque e que exige muita dedicação da equipe de trabalho, porque ela tem muita responsabilidade em mudanças que poderão impactar diretamente no interesse da comunidade pelos dados a serem consumidos.

Há, em muitos casos, uma falsa ilusão quando se inicia um projeto de publicação de dados, que os procedimentos compreendidos na fase 2, sejam os únicos dentro de um projeto desta natureza. Ou seja, uma falsa ideia de que as atividades técnicas desta fase

são as únicas que um projeto precisa ter para transformar qualquer base de dados em um conjunto de dados abertos e semânticos.

Há um documento produzido por um grupo de trabalho do W3C (LÓSCIO; BURLE; CALEGARI, 2017), publicado como recomendação W3C a partir de 31 de janeiro de 2017, chamado "Melhores práticas para dados na Web" (Data on the Web Best Practices[ii]), que teve como objetivo constituir procedimentos para ajudar a suportar um ecossistema autossustentável de publicação de dados. Os dados devem ser descobertos e compreensíveis por seres humanos e máquinas. Este documento oferece uma grande quantidade de informações e sugere técnicas (práticas) para lidar com a fase 2 apresentada aqui.

O documento em si aborda de forma bem didática 8 possíveis benefícios que podem ser atingidos utilizando-se das 35 práticas propostas. Os 8 benefícios apresentados são: reuso, acesso, conexão, descoberta, processamento, confiança, interoperabilidade e compreensão.

As práticas são apresentadas uma a uma, inicialmente com um *template* muito claro e definido, onde se indica o porque aquele item é especificamente relevante para a publicação ou reutilização de dados na Web e porque pode incentivar a publicação ou reutilização de dados na Web. Posteriormente indica-se o resultado esperado e descreve sobre uma possível estratégia de implementação.

Cada uma das práticas ainda indica como ela pode ser testada, apresenta informações sobre a relevância da aplicação daquela prática específica e por fim lista os benefícios (entre os 8) que aquela prática agrega aos dados a serem publicados.

Nesta fase (2) deve-se abordar as questões relativas a estrutura do dado e seu formato, esse é um item altamente técnico e que pode ter grande impacto posteriormente no momento de consumo dos dados publicados.

A estruturação, formalização e formatação dos dados é um processo importante para que se possa atribuir semântica aos mesmos, e normalmente acontece de forma sequencial. O dado que foi selecionado na fase 1 deve ser destrinchado, atribuindo-se a ele uma nova estrutura, incluindo a normalização das informações, que é um processo muito importante.

Nem sempre os dados a serem publicados são oriundos da mesma fonte, em geral esses dados podem ter como fontes as planilhas, as bases de dados (ou tabelas originadas por elas), arquivos dos mais variados formatos, incluindo alguns oriundos de

mineração e, portanto, é necessário que eles sejam reorganizados. Uma das tarefas da organização é justamente a padronização dos dados, ou seja, usar os mesmos tipos de informações, vocabulários controlados, associar termos que sejam similares (ou iguais) mas estejam explanados em formatos diferentes, usar as mesmas unidades de medidas para dados numéricos (ou financeiros). Esse processo de normalização é de fundamental importância.

A formalização dos dados envolve a construção de um modelo conceitual para o conjunto de dados. É na formalização que os dados devem passar a fazer parte de uma estrutura lógica como as ontologias. No começo de um projeto de publicação de dados, pode ainda não haver uma definição sobre a ontologia a ser utilizada, portanto é necessário que a equipe responsável pelo projeto tenha em mente que deverá desenvolver uma ontologia, utilizar-se de uma que já esteja em uso ou ainda adaptar uma já existente para o conjunto de dados a ser publicado.

Não é objetivo deste texto apresentar metodologias para desenvolvimento e uso de ontologias, entretanto é importante ressaltar que ter uma formalização por meio de uma ontologia e com uso de vocabulários internacionalmente reconhecidos é muito importante para o sucesso do projeto de publicação de dados. O projeto Linked Open Vocabularies[iii] (LOV) é um ótimo recurso para identificar vocabulários conhecidos e usá-los para dar significado (semântica) na formalização dos dados e construção (ou adaptação) de ontologias.

Após os dados estarem normalizados e formalizados é importante que sejam formatados dentro de uma estrutura técnica que seja possível recuperá-los. Dentro desse contexto é necessário coloca-los dentro de um formato de serialização computacional, usando uma linguagem computacional. Os dados podem ser disponibilizados em OWL, XML, JSON, JSON-LD, vai depender muito de como será a disponibilização desses dados ao público que irá consumi-los.

A formatação dos dados também depende de quais ferramentas serão utilizadas para prover acesso aos dados. Entretanto essa parte do projeto será discutida na fase 3.

Ainda na nesta fase é importante que se definam as licenças que serão atribuídas aos dados. Note que atribuir licença é garantir ao consumidor que o dado possa ser utilizado e indicar como ele pode ser utilizado.

Apesar de muitos projetos de dados não deixarem claro qual é a licença atribuída, considera-se esse um ponto de extrema relevância para quem vai consumir os dados.

Entende-se que a organização que está publicando dados tenha realmente interesse que esses dados sejam consumidos, sendo assim, é de fundamental importância garantir segurança e a liberdade de uso aos consumidores.

Há uma infinidade de licenças que permitem acesso e uso dos dados, entretanto cada uma delas tem características diferentes e em muitos casos precisam ser estudadas e entendidas para que sejam atribuídas sem risco nem a quem publica nem a quem consome os dados.

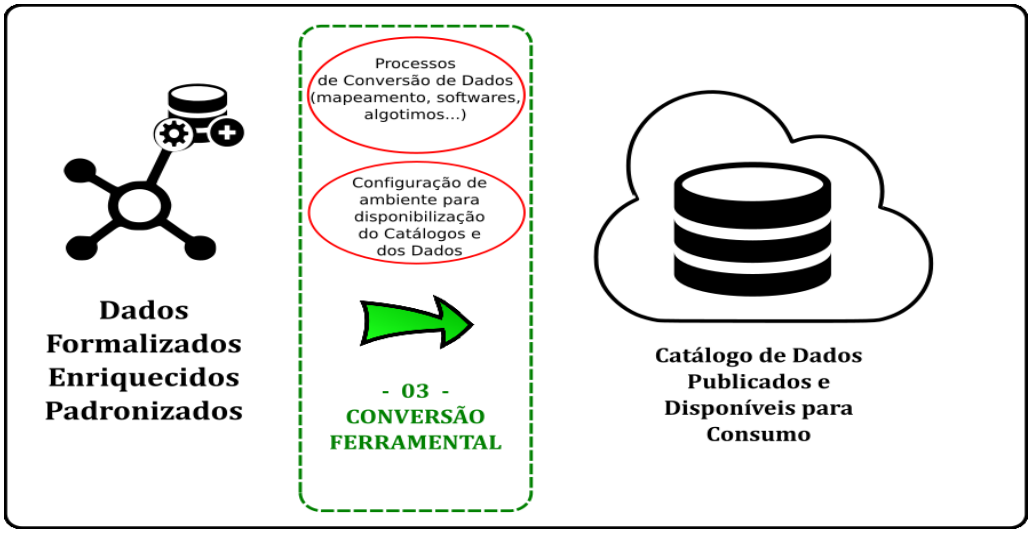
Com o advento da publicação de dados algumas licenças têm sido criadas especificamente para essa nova realidade, é o caso das Licenças Open Data Commons[iv]. Algumas outras licenças já conhecidas como as Creative Commons continuam sendo uma ótima opção também.

A partir do momento que os dados já foram normalizados, estruturados, formalizados, formatados e já tem uma licença passa-se a fase 3 do projeto de publicação de dados.

4.3 Conversão Ferramental

A fase 3 é uma etapa muito importante do projeto, pois o dado deixa de ser restrito e será efetivamente publicado na Web, ou seja, ele ultrapassa o muro da organização para passar a integrar uma grande nuvem de dados e ser consumido por quem tenha interesse, como pode ser visto na figura 4.

Figura 4 – Fase 3 do fluxo organizacional de publicação de dados



Fonte: Dados da pesquisa.

Essa fase do projeto é marcada pela seleção de ferramentas que deverão permitir o acesso ao dado pelos interessados em consumi-los. Há várias maneiras de disponibilizar os dados para que possam se consumidos.

As maneiras mais simples e básicas remetem a simples disponibilização de arquivos de dados, ou pacotes compactados que contemplam os arquivos, que mesmo nesse formato podem ser semânticos, via arquivos em serializações adequadas.

O que se espera é que tenhamos um conjunto de ferramentas que permitam acesso ao dado das mais variadas formas, tanto para acesso por humanos quanto por máquinas.

O acesso pra humanos, nem tanto trivial nessa fase do projeto, poderá ser fornecido por uma interface Web que garanta acesso diretamente a URI dos recursos e através delas as suas propriedades.

O acesso para máquinas pode ser feito diretamente através de ferramentas que disponibilizem um Sparql EndPoint, ou seja, uma interface para consultas via linguagem de consulta semântica (Sparql), ou ainda por meio de Webservices e APIs.

Nessa fase é importante também que se escolha uma ferramenta que possa servir como catálogo de dados para os consumidores. Uma ferramenta do tipo catálogo (exemplo mais utilizado é o CKAN, mas há outras) permite que haja uma visualização completa de toda a informação a respeito dos dados que estão sendo publicados.

O processo nomeado conversão ferramental tem seu ponto crítico a partir do momento que já foram escolhidas as ferramentas para disponibilizar e permitir acesso aos consumidores de dados, e também já se tem a disponibilidade dos dados prontos para serem carregados nas ferramentas.

Carregar os dados é uma tarefa que pode parecer simples, e se ela for realizada apenas uma vez ela realmente será. Em geral, as ferramentas que são interfaces de acesso dos usuários aos dados, disponibilizam interfaces para que os dados possam ser carregados diretamente via arquivo. O que torna o processo mais trabalhoso é justamente pensar em um procedimento recorrente e cíclico, de forma que possa haver alimentação frequente de dados nas ferramentas.

A criação de uma rotina que possa varrer dados disponíveis e carrega-los nas ferramentas normalmente não é uma tarefa disponível na maioria das ferramentas, o que implica que para essa tarefa seja executada possa ser necessário o desenvolvimento de algum tipo de script de programação que possa operar o processo.

Escrever um script de programação é uma tarefa para os componentes de TI da sua equipe, e muitas vezes será necessário construir um diagrama de como o processo funcionará, envolvendo um cronograma de atividades que inicie na coleta dos dados já formalizados e em seguida com a inserção do mesmo na ferramenta que torna o dado disponível para a comunidade.

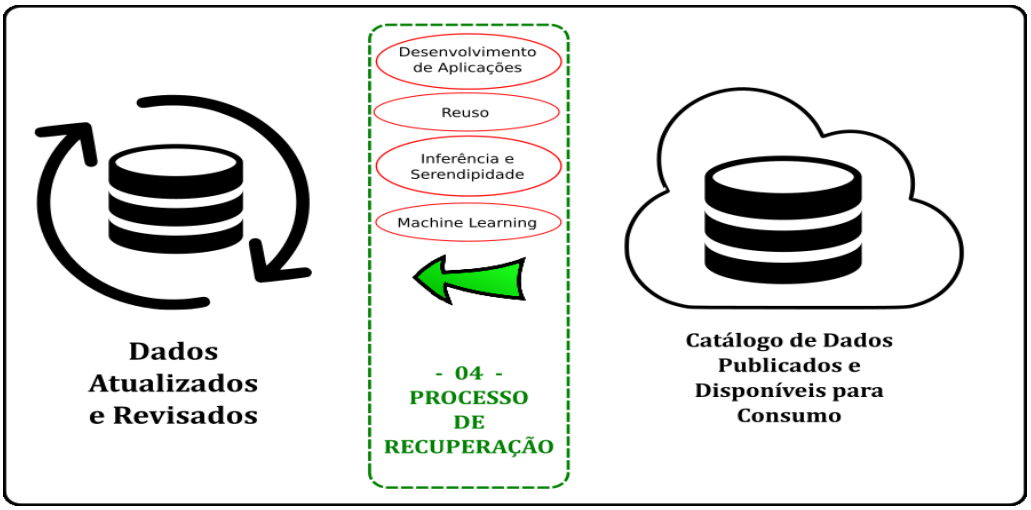
O processo de conversão ferramental dependerá muito das escolhas feitas pela equipe do projeto, e ele com certeza dará ritmo a frequência e volume de dados que ficarão disponíveis a comunidade.

Criar uma rotina que dependa menos do trabalho humano é um dos fatores primordiais na fase 3, e portanto, quanto maior for a dedicação nesta fase do projeto maior será a automatização da sua linha de produção e publicação de dados.

4.4 Processo de Recuperação

As três primeiras fases do processo de publicação de dados tinham foco excessivamente na estruturação e publicação de dados. As duas próximas fases estarão concentradas no momento posterior a publicação de dados, exatamente quando o usuário que vai consumi-los começa a ter acesso a esses dados, conforme pode ser visto na figura 5.

Figura 5 – Fase 4 do fluxo organizacional de publicação de dados



Fonte: Dados da pesquisa.

As tarefas da fase 4 não devem ser atribuídas a equipe de publicação de dados, entretanto é importante que se tenha a noção exata do que pode ser feito com os dados e como a comunidade vai consumir esses dados.

Além do consumo por humanos, o processo natural de consumo de dados ocorre via software, ou seja, aplicações que emitirão comandos e receberão como retorno os dados disponibilizados. Aqui no processo de recuperação, que pode inclusive ser parte dos interesses da própria organização que está publicando os dados, há uma séria de elementos que podem ajudar no uso desses dados para uma diversidade de aplicações.

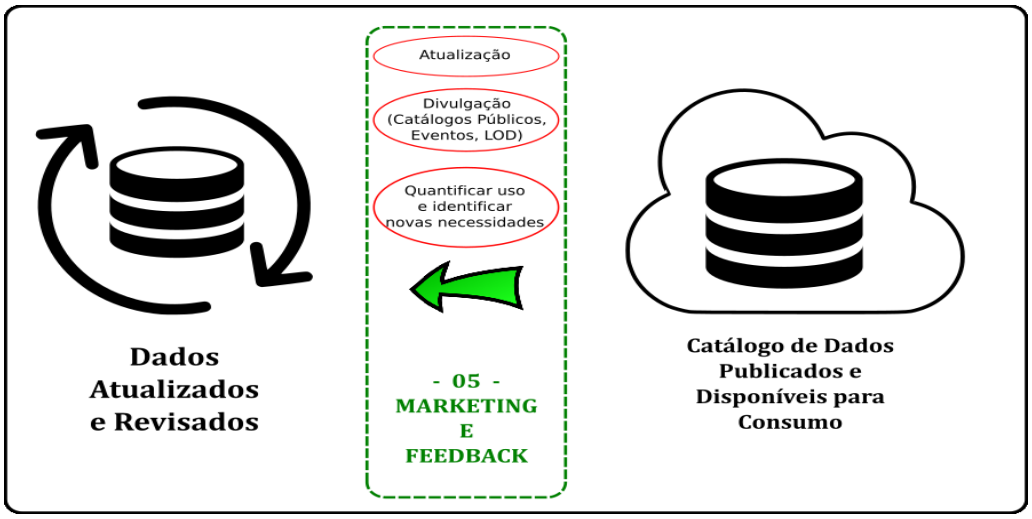
Cabe nesse caso o desenvolvimento de aplicações (web, dispositivos móveis, etc) que possam levar o dado a usuários que nunca teriam acesso senão por aplicações para usuários leigos. Cabe também a criação de aplicações para que os dados possam ser reusados pela própria organização, entretanto já com uma carga semântica e principalmente com o enriquecimento de dados oriundos de outras bases.

Nessa fase é importante destacar o uso de técnicas como aprendizado de máquina (*Machine Learning)* e também da criação de axiomas que possam permitir constituir inferências nos dados. O uso de aprendizado de maquinas juntamente com as possibilidades de inferências nos dados pode gerar uma gama de padrões e resultados, inclusive preditivos, que permitem encontrar padrões informacionais até então não percebidos.

4.5 Marketing e Feedback

É importante acompanhar se o conjunto de dados publicados está atendendo a demanda da comunidade e se realmente eles estão sendo úteis ou se são de conhecimento das comunidades que poderiam ter interesse. Como pode visto na figura 6, a fase 5 permite que haja considerações acerca do projeto inicial, baseado nas necessidades dos usuários que consomem os dados.

Figura 6 – Fase 5 do fluxo organizacional de publicação de dados



Fonte: Dados da pesquisa.

Muitas vezes há uma grande demanda de trabalho para que se possa publicar dados porém a maneira como eles ficam disponíveis para a comunidade não são interessantes o suficiente para que seja feito o uso, ou consumo.

Divulgar nas principais mídias, informar os possíveis interessados, publicar os dados em catálogos de grande acesso, oferecer os dados para serem trabalhados em eventos como Hackatons são algumas das técnicas de marketing que podem ser utilizadas para que a comunidade tenha conhecimento sobre os dados que estão sendo publicados. É de fundamental importância que quem precise do dado saiba exatamente onde encontrá-lo.

A partir do momento que os dados passam a ser consumidos há um outro fator que pode ser muito importante para que seus dados possam ser cada vez mais utilizados pela comunidade que os consome, é o processo de *feedback*. Criar uma estrutura que seja possível receber informações da comunidade que está consumindo os dados é muito importante.

Criar rotinas que analisem o consumo também é bastante importante, visto que esse tipo de atividade permite entender o que realmente tem despertado interesse da comunidade, quais são os dados de maior interesse, qual a granularidade do dado que mais interessa, qual a forma de acesso mais utilizada.

Criar canais de feedback também é muito importante. O simples fato de disponibilizar um e-mail de contato (que seja respondido) ou ainda um formulário em

uma página Web, já permite que consumidores de dados possam se relacionar com a equipe responsável pelo projeto de publicação de dados.

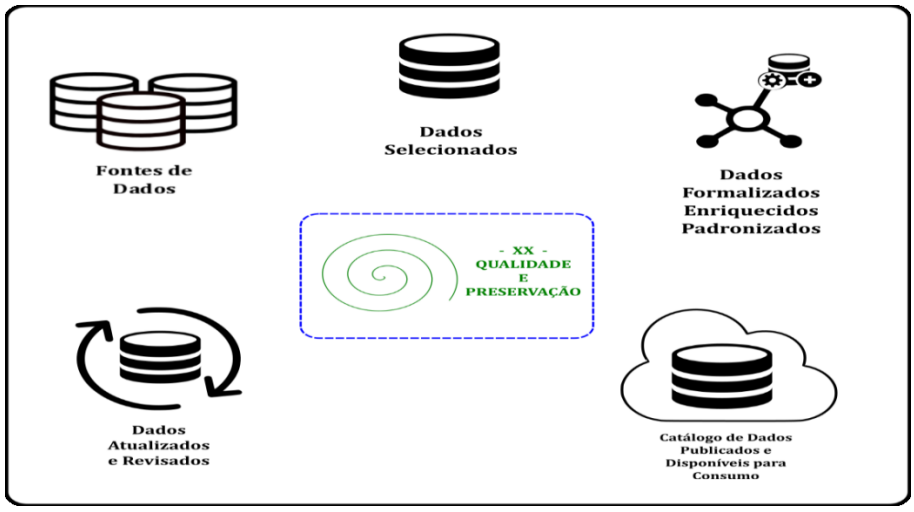
Todo e qualquer feedback que possa ser recebido deve ser discutido em grupo e quando necessário pode gerar alterações nas três primeiras fases do projeto de publicação de dados. Ressalta-se que as solicitações da comunidade podem ser desde parte dos dados que não ficaram disponíveis, passando pela granularidade dos mesmos, o formato de disponibilização, sugestões de alterações na ontologia ou ainda nas ferramentas utilizadas para que o dado fique disponível.

É muito importante que se feche o ciclo de publicação de dados com o máximo de atenção as necessidades de quem consome os dados.

4.6 Qualidade e Preservação

A fase 6, conforme pode ser visto na figura 7, tem um contexto completamente diferente das anteriores. Ela está numerada como fase 6, mas poderia ser também a fase 0, ou ainda qualquer outro tipo de identificação não numerada. Acontece que a fase 6 transcende a todas as outras fases anteriores, e está diretamente relacionada a todas elas. Entende-se que dar qualidade e preservar os dados publicados é um item da maior relevância que pode existir em um projeto.

Figura 7 – Fase 6 do fluxo organizacional de publicação de dados



Fonte: Dados da pesquisa.

Entende-se que pensar nas questões que envolvem qualidade e preservação estão diretamente relacionadas a confiança que se pode ter na base. Confiança é um dos elementos mais significativos no processo de uso e reuso de uma base de dados, e tem sido pauta constante de discussão dada a grande variedade de problema encontrados em base de dados publicadas em formato aberto e semântico ao redor do mundo, mais especificamente quando tratamos de databases publicados no Linked Open Data.

O texto "Metodologia de avaliação de qualidade para dados conectados" de Souza, Botega e Santarém Segundo (2017) faz uma abordagem específica sobre a questão da qualidade na publicação de dados.

Souza, Botega e Santarém Segundo (2017) indicam que

> [...] a qualidade pode ser definida como medidas para que o produto oferecido esteja de acordo com o que se espera dele, podendo este ser uma informação, um dado, um serviço ou um processo. Estando ele livre de problemas, possibilita que as atividades dependentes sejam executadas com sucesso. É notado que a forma como os dados, informações, produtos, etc., são manuseados influenciará na qualidade das atividades desempenhadas nos sistemas de diferentes domínios.

Sobre a questão da qualidade na disponibilização de dados Souza, Botega e Santarém Segundo (2017) ainda afirmam que:

> A literatura aponta problemas de qualidade não somente nos dados, mas também na estrutura provida para sua publicação, fator que pode dificultar seu acesso e até mesmo inviabilizar sua utilização, evidenciando o fato de que a qualidade consiste em um fator de extrema importância.

A preservação dos dados, que acaba sendo intrínseca a qualidade dos dados, e em alguns casos compõe parte dos requisitos de qualidade, é um fator determinante para garantir a longevidade dos dados publicados. Sayão e Sales (2012) confirmam que "o principal desafio recai na necessidade de se preservar não somente o conjunto de dados, mas de preservar, sobretudo, a capacidade que ele possui de transmitir conhecimento para uso futuro das comunidades interessadas".

Os dados, portanto, devem estar disponíveis no momento do uso e também devem permitir que futuros usuários reanalisem os dados dentro de novos contextos. Porém, para que ocorra um processo de preservação em que os significados dos dados possam atravessar a barreira do tempo, é necessário assegurar que os usuários no futuro estejam instrumentados com as informações essenciais para o efetivo reuso dos dados (CONWAY, 2011 *apud* SAYÃO; SALES, 2012).

A qualidade e a preservação garantem a integridade do dado, e além da confiança estimulam o reuso da base publicada. Em geral uma base integra, que garante preservação e tem qualidade, receberá conexões advindas de outras bases, servindo também como referência para enriquecer dados de bases de outrem.

5 CONSIDERAÇÕES FINAIS

Há uma constante evolução na necessidade de se publicar dados, as demandas por publicação de bases abertas com dados ligados são cada vez mais latentes, entretanto é bem interessante notar que ainda encontremos muitas dificuldades para constituir um projeto robusto de publicação de dados, que possa fazer parte da cultura organizacional das organizações.

Constituir um projeto robusto é garantir uma rotina de publicação de dados, dentro de um contexto que atenda ás necessidades da comunidade que tem interesse nos dados e que seja revisto constantemente de forma a melhorar ainda mais o atendimento a essa comunidade, além disso é garantir que os dados atendam a padrões de formalização, tenham licenças apropriadas, dotados de requisitos mínimos de qualidade e que sejam preservados para uso perene.

Constituir uma equipe responsável pelo projeto de publicação de dados é essencial quando a organização pretende efetivar a publicação de dados. Ressalta-se aqui que muitas vezes não haverá pessoal suficiente para uma equipe multidisciplinar, ou ainda, a equipe será formada por apenas uma única pessoa, mas é de fundamental importância que o projeto seja conhecido na organização e que outros colaboradores, se houverem, saibam que a organização tem o intuito de publicar dados na Web para serem consumidos livremente.

O fluxo organizacional apresentado é resultado de um conjunto de pesquisas que apresentam projetos de publicações de dados, entretanto grande parte dessas pesquisas apresentam algumas dessas fases ou então parte das tarefas que se misturam nesse fluxo apresentado aqui. A ideia da apresentação desse fluxo nasceu justamente da junção de partes de múltiplas pesquisas de forma que pudesse constituir um ponto de partida e um entendimento de um projeto de publicação de dados por completo.

Esse fluxo organizacional, que não se apega as técnicas e tecnologias, não indicando ou sugerindo ferramentas na maior parte das vezes, tem como objetivo principal dar a compreensão exata de que um projeto dessa natureza envolve muitas tarefas e pode ser menos trivial do que o imaginado pelas organizações ou pessoas que pretendem publicar dados, entretanto ele representa e sugere o que se considera um conjunto de fases ideias para projetos de publicação de dados.

Não há dúvidas que pode haver projetos que funcionem sem passar por todas as fases, entretanto entende-se que as fases aqui apresentadas no fluxo organizacional de publicação de dados é o ponto de partida mínimo para constituição de um projeto robusto.

Destaca-se ainda a fase 5, que pouco aparece em grande parte dos projetos de publicação de dados, entretanto é de fundamental importância para que todo o trabalho estrutural realizado nas três primeiras fases possa ter valido a pena. A fase 5, representada principalmente pelos processos de marketing e feedback, garante ao projeto a reponsabilidade de dar visibilidade a todo o trabalho feito e também o compromisso de atender as necessidades da comunidade. Um projeto caracterizado com os princípios de Dados Ligados e que tenha realmente o intuito de atender a comunidade que quer consumir os dados tem como principal requisito o fator de atender as demandas do que a comunidade realmente quer e como ela precisa para que os dados possam ser bem utilizados.

Como última consideração, entende-se que esse fluxo possa contribuir como norteador em projetos de publicação de dados e que possa ser ponto de partida para outras pesquisas que possam evoluir com o fluxo proposto.

REFERÊNCIAS

BERNERS-LEE T.; LASSILA, O.; HENDLER, J. The semantic web. **Scientific American**, New York, v. 5, 2001.

BERNERS-LEE, T. **Linked data principles**. 2006. Disponível em: <https://bit.ly/1x6N7XI>. Acesso em: 09 jun. 2018.

BORKO, Harold. Information science: what is it? **American Documentation**, [*S.l.*], v. 19, n. 1, p. 3-5, 1968. Disponível em: <https://bit.ly/2DLQfkL>. Acesso em: 10 jul. 2018. DOI: http://doi.org/d9zjg3.

EAVES, D. **The Three laws of open government data**. 2009. Disponível em: <https://bit.ly/2ftyZUW>. Acesso em: 10 jul. 2018.

GUIMARÃES, J. A. C. Perspectivas de ensino e pesquisa em organização do conhecimento

em cursos de Biblioteconomia do Mercosul: uma reflexão. *In*: ENCUENTRO DE INVESTIGADORES DE BIBLIOTECOLOGIA Y CIENCIA DE LA INFORMACIÓN DE IBEROAMERICA Y EL CARIBE, 5., 2000, Granada. **Anais**...Granada: EDIBCIC, 2000.

HEY, T. *et al*. (Org.). **The Fourth Paradigm:** Data-Intensive Scientific Discovery. Redmond, Washington: Microsoft Research, 2009. Disponível em: <https://bit.ly/1iD63DJ>. Acesso em: 10 ago. 2018.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. **Data on the Web Best Practices**: challenges and benefits. W3C Recommendation, 2017. Disponível em: <https://bit.ly/2FG1EoK>. Acesso em: 10 jun. 2018.

MANUAL dos dados abertos: desenvolvedores. Cooperação técnica científica entre Laboratório Brasileiro de Cultura Digital e o Núcleo de Informação e Coordenação do Ponto BR (NIC.br). São Paulo: Comitê Gestor da Internet no Brasil, 2011. Disponível em: <https://bit.ly/2Ai8oTB>. Acesso em: 10 abr. 2018.

MELO, J. O. S.; BOTEGA, L. C.; SANTAREM SEGUNDO, J. E. Metodologia de avaliação de qualidade para dados conectados. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais...** Marília Unesp: ANCIB, 2017.

OPEN KNOWLEDGE FOUNDATION. **About OKF**. 2004. Disponível em: <http://okfn.org/about/>. Acesso em: 25 ago. 2018.

SANTARÉM SEGUNDO, J. E. Web Semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente as iniciativas internacionais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, João Pessoa, v. 8, p. 219–239, 2015.

SANTARÉM SEGUNDO, J. E; CONEGLIAN, C. S. Web Semântica e Ontologias: um estudo sobre construção de axiomas e uso de inferências. **Informação & Informação**, Londrina, v. 21, n. 2, p. 217–244, dez. 2016. Disponível em: <https://bit.ly/2uLpbgL>. Acesso em: 09 jun. 2018.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**. Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SAYÃO, L. S. F.; SALES, L. F. Digital curation: a new platform for digital preservation of research data. **Informação & Sociedade**: Estudos, Paraíba, v. 22, n. 3, 2012. Disponível em: <https://bit.ly/2KLvOIq>. Acesso em: 09 set. 2018.

NOTAS

[i] A revisão ortográfica, gramatical e em Língua Portuguesa é de responsabilidade do autor.

[ii] https://www.w3.org/TR/2017/REC-dwbp-20170131/

[iii] https://lov.linkeddata.es/

[iv] https://opendatacommons.org