1 Modelo de Ciclo de Vida de Dados na Web

1.1 Visão geral do DWLM

O Modelo de Ciclo de Vida de Dados na Web (*Data on the Web Lifecycle Model* - DWLM) foi proposto com o objetivo de prover um entendimento comum das etapas que um conjunto de dados passa ao longo de sua vida na Web. Além disso, o modelo proposto visa garantir que os conjuntos de dados publicados atendam a alguns requisitos que permitam seu processamento por humanos e máquinas, bem como sua reutilização e confiança entre os consumidores de dados. Para isso, o modelo incorpora, ao longo de suas fases, as Boas Práticas para Dados na Web (DWBP) propostas pelo W3C. As DWBPs referem-se a conjuntos de dados e suas distribuições, onde esses podem ser publicados em diferentes formatos (LÓSCIO; BURLE; CALEGARI, 2017). Portanto, para a elaboração desse trabalho, consideramos o mesmo contexto das DWBPs (Figura 1), no qual a publicação e uso de Dados na Web refere-se a um conjunto de dados que é descrito por metadados e esses dados podem possuir diferentes distribuições.

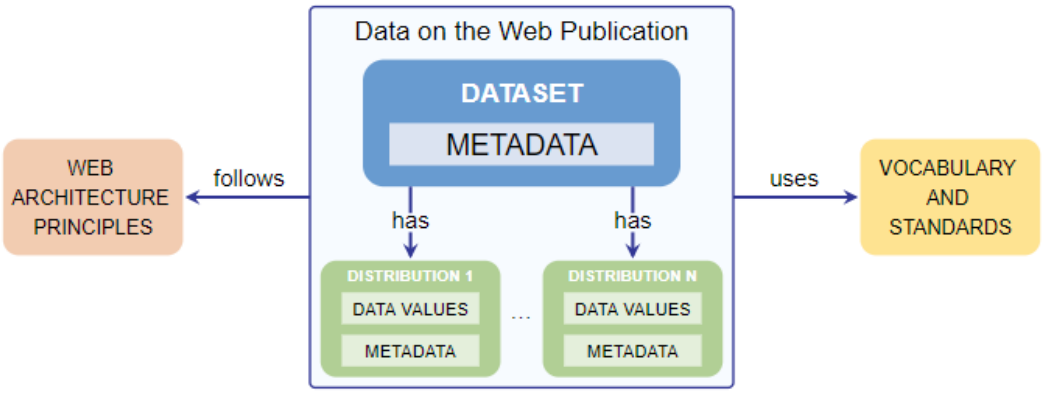

Figura 1 – Contexto de Publicação de Dados na Web. Fonte: (LÓSCIO; BURLE; CALEGARI, 2017)

Para a construção do DWLM, nos baseamos no *Abstract Data Lifecycle Model* (ADLM) proposto por Möller (2013) e utilizamos como ponto de partida o Ciclo de Vida de Dados na Web proposto por Lóscio, Oliveira e Bittencourt (2015). O procedimento de construção do DWLM foi composto por um processo interativo e incremental, onde as fases e atividades foram exaustivamente aprimoradas, refinadas e validadas. Além disso, o DWLM tem o objetivo de ser o mais genérico possível, para que sua aplicabilidade englobe o máximo de cenário. É importante ressaltar que o DWLM foi idealizado para tratar um conjunto de dados a cada iteração, ou seja, se houver três conjuntos de dados haverão três

ciclos de vida, pois cada conjunto terá uma trajetória única na Web.

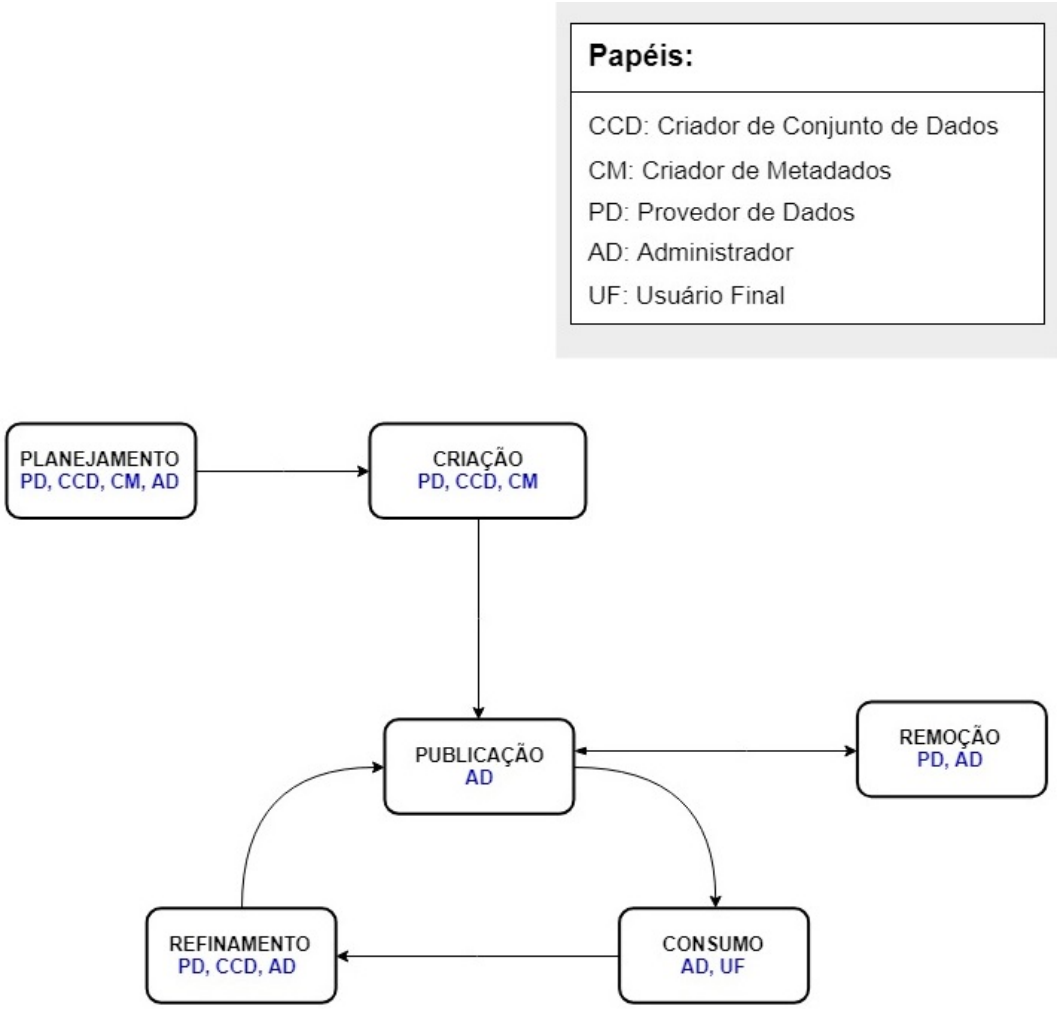

Figura 2 – Modelo de Ciclo de Vida de Dados na Web - DWLM. Fonte: Autor

Seguindo a mesma ideia do ADLM, o DWLM é composto por seis fases, como ilustrado na Figura 2. Cada uma dessas fases tem um papel fundamental no Ciclo de Vida dos Dados na Web. A primeira fase é a de *Planejamento*, nela serão coletadas informações descritivas do conjunto de dados e escolhida a solução para publicação do conjunto. Em seguida, na fase de *Criação*, o conjunto será criado e validado. Após criado, ele segue para a fase de *Publicação*, nesse momento ele será disponibilizado na Web para que possíveis usuários possam consumi-lo e assim chegarmos a fase de *Consumo*. Nela, além do *Usuário Final* ter acesso ao conjunto de dados, serão descritos alguns exemplos de uso, bem como o fornecimento do *feedback*. A próxima fase, intitulada como *Refinamento*, diz respeito às alterações que serão realizadas no conjunto de dados e a geração de uma nova versão. Por último, na fase de *Remoção*, o acesso a esse conjunto será removido da Web encerrando o seu ciclo. Além das seis fases descritas, também destacam-se os seguintes papéis: *Provedor de Dados*, *Criador de Conjunto de Dados*, *Criador de Metadados*, *Administrador* e *Usuário Final*. A Tabela 1 apresenta de forma resumida todas as fases do

modelo DWLM, juntamente com suas atividades, entradas, saídas e melhores práticas a serem aplicadas em cada uma das fases.

Tabela 1 – As fases, associando as atividades, entradas, saídas e DWBP usadas no Modelo ADWLM

Fase	Atividades	Entrada	Saída	DWBP
Planejamento	Especificar fontes de dados Descrever conjunto de dados Estabelecer solução para publicação do conjunto de dados	Dados de fontes distintas	Documento de Descrição dos Dados Solução para publicação do conjunto de dados	BP1, BP2, BP3, BP4, BP5, BP6, BP12, BP14 BP15, BP16
Criação	Criar o conjunto de dados Avaliar qualidade Validar conjunto de dados	Fontes de Dados Documento de Descrição dos Dados	Documento de erros Termo de consentimento do conjunto de dados Conjunto de dados criado Documento de Descrição dos Dados atualizado	BP9, BP10
Publicação	Publicar o conjunto de dados de acordo com a solução escolhida Tornar o conjunto de dados acessível Fornecer alternativas de uso	Conjunto de dados criado Documento de Descrição dos Dados atualizado	Solução escolhida para publicação Conjunto de dados publicado	BP17, BP18, BP19, BP20, BP21, BP22, BP23, BP24, BP25, BP26, BP32
Consumo	Acessar conjunto de dados Fazer uso do conjunto de dados Prover e disponibilizar feedback	Conjunto de dados publicado	Conjunto de dados acessado Áreas para usuários informarem e acessarem os feedbacks	BP29, BP30, BP34, BP35
Refinamento	Corrigir e enriquecer o conjunto de dados Validar o conjunto de dados Versionar o conjunto de dados	Conjunto de Dados	Conjunto de dados Refinado Log de Refinamento Nova versão do conjunto de dados	BP7, BP8
Arquivamento	Remover acesso ao conjunto de dados	Documento de solicitação de arquivamento do conjunto	Acesso ao conjunto de dados Removido	BP27, BP28

1.2 Papéis do ADWLM

No Modelo Abstrato de Ciclo de Vida de Dados (ADLM) proposto por Möller (2013), são definidos cinco papéis para os atores que irão, de alguma forma, interagir com os dados ao longo do ciclo de vida. Com base nisso, para o modelo proposto nesta dissertação foram identificados cinco papéis que terão uma participação direta nas fases do DWLM. Porém, é importante ter em mente que um mesmo ator, dependendo do contexto no qual o modelo for aplicado, pode desempenhar vários papéis. Isto é, um ator, com o papel de *Criador de Conjunto de Dados*, poderia criar o conjunto de dados e, logo após,

com o papel de *Administrador*, publicá-lo e/ou arquivá-lo. Ressaltamos que esses papéis podem ser desempenhados por humanos ou até mesmo por máquinas, quando é usado alguma ferramenta ou sistema para executar as ações.

Cada um desses papéis estará envolvido em fases específicas do modelo de ciclo de vida, como apresentado na Figura 2.

- *Provedor de Dados*

  O *Provedor de dados* é o proprietário e fornecedor dos dados, ou seja, o ator que assumir o papel de *Provedor* irá ceder os dados que, em seguida serão criados em um conjunto de dados e, posteriormente, publicado. Além disso, durante o DWLM esse papel participará de todas as validações necessárias do conjunto de dados. Estas validações são essenciais para verificar se o conjunto de dados que foi criado está de acordo com as suas expectativas e se não existem erros nos dados e metadados. O *Provedor de dados* participará nas fases de *Planejamento*, *Criação*, *Refinamento* e *Remoção*.

- *Criador de Conjunto de Dados*

  O *Criador de Conjunto de Dados* tem como principal objetivo elencar e descrever todas as informações a respeito do conjunto de dados que será criado, assim como, executar todas as atividades relacionadas ao seu processo de criação. Na literatura, esse papel recebeu diferentes nomes como criador de conteúdo, provedores de conteúdo e controladores de dados (MöLLER, 2013). É importante ressaltar que, em algumas situações, o *Provedor de Dados* poderá ser o próprio *Criador de Conjunto de Dados*. No DWLM, esse papel participará da fase de *Planejamento*, *Criação*, *Refinamento*.

- *Criador de Metadados*

  O *Criador de Metadados* irá elencar e descrever os metadados que serão disponibilizados juntamente com o conjunto de dados. Vale ressaltar que, em alguns cenários, suas responsabilidades são comumente executadas pelo ator responsável pelo papel de *Criador de Conjunto de Dados*. No DWLM, o *Criador de Metadados* participará da fase de *Planejamento* e *Criação*.

- *Administrador*

  O *Administrador*, em contraste com os criadores, manipulam o conjunto de dados e seus metadados sem alterar o seu formato e significado (MöLLER, 2013). Ele será responsável por realizar a publicação dos dados e acompanhar o conjunto de dados durante todas as fases seguintes, até chegar ao momento da seu remoção. Assim como o *Usuário Final*, esse papel é essencial para todo o ciclo de vida dos dados na Web, pois estará presente em quase todas as fases. No DWLM, o ator que exercer

esse papel participará das fases de *Planejamento*, *Publicação*, *Consumo*, *Refinamento* e *Remoção*.

- *Usuário Final*

  O último papel identificado é o *Usuário Final*. Esse papel representa os usuários responsáveis por consumi-lo de forma ativa. De acordo com Kosch et al. (2005 apud MöLLER, 2013), o *Usuário Final* está envolvido na "navegação, pesquisa e consumo" de metadados e conteúdo. Além disso, ele poderá enviar *feedback* a respeito do conjunto de dados consumidos e, principalmente, desenvolver aplicações, a fim de oferecer produtos/serviços a outros usuários. Sob essa perspectiva, no DWLM, o *Usuário Final* participará ativamente na fase de *Consumo*.

1.3 Fases do ADWLM

Conforme descrito anteriormente, o DWLM é composto por seis fases e cada fase consiste de uma ou mais atividades que são exercidas por atores desempenhando um dos papéis descritos na Seção 1.2. Além disso, cada fase está associada a uma ou mais DWBPs que fornecem informações adicionais úteis para a realização de cada uma das atividades propostas. Desse modo, as próximas seções descreverão cada uma dessas fases, assim como as atividades envolvidas. Para a identificação de cada atividade em sua respectiva fase, consideramos a nomenclatura "F00A00" onde, "F" representa a fase e "A" a atividade.

1.3.1 Planejamento

A primeira fase proposta para o DWLM é a de *Planejamento*, sendo ela primordial para que o *Criador de Conjunto de Dados* e o *Criador de Metadados* possam se apropriar dos dados que serão posteriormente publicados. Nesse momento, em conjunto com o *Provedor de Dados*, eles farão a descrição do conjunto de dados, bem como a definição dos metadados que serão disponibilizados juntamente com o conjunto. Além disso, serão consideradas informações de proveniência, licença e volume do conjunto de dados.

Para esta fase, foram definidas três atividades (ver Figura 3) que serão realizadas para obter todas as informações necessárias do conjunto de dados. Ao final da etapa de *Planejamento*, teremos como um *output*, o *Documento de Descrição do Conjunto de Dados*. Esse documento é a principal saída da fase de *Planejamento*. Ele é de suma importância, pois é composto por todas as informações coletadas a respeito dos dados, como também os metadados que serão utilizados, licenças e vocabulários. Abaixo estão descritas as três atividades que compõem a fase de *Planejamento*.

- ***F01A01 - Especificar fontes de dados***

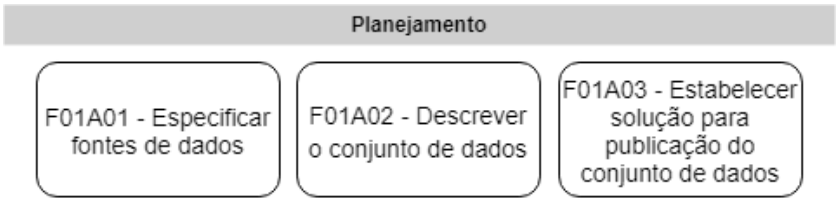

Figura 3 – Atividades da fase de Planejamento Fonte: Autor

Durante esta atividade, o *Provedor de Dados* especificará as fontes de dados que poderão ser usadas para a coleta dos dados que serão publicados. Essas fontes de origem variam desde bancos de dados relacionais, não-relacionais, a até mesmo arquivos em diferentes formatos como: CSV, TXT, XML, RDF e JSON. Além disso, fontes de dados de tempo real, como sensores, também podem ser consideradas.

A realização dessa atividade é importante para que o *Criador de Conjunto de Dados* e o *Criador de Metadados* possam, na fase de *Criação* do conjunto de dados, definir as estratégias de coleta para cada fonte aqui especificada.

- ***F01A02 - Descrever o conjunto de dados***

  Esta atividade consiste na coleta de informações que sejam relevantes para o entendimento e a criação do conjunto de dados. Nela, o *Provedor de Dados*, o *Criador de Conjunto de Dados* e o *Criador de Metadados* deverão descrever todas as propriedades que o conjunto de dados deverá conter. Além disso, as informações aqui descritas poderão fomentar ações que serão realizadas nas fases seguintes do DWLM. Por exemplo, se ao chegar na fase de *Criação* e o conjunto de dados possuir um volume considerado grande pelo seu criador, possivelmente, irá ocasionar na geração de subconjuntos.

  Levando em consideração o documento de Boas Práticas (DWBP)[1], estipulamos alguns pontos que devem ser levados em consideração nesta etapa para a elaboração do *Documento de Descrição do Conjunto de Dados*. São eles:

  – Proveniência dos dados: Atualmente, tem-se tornado natural o usuário se questionar quanto à confiabilidade e integridade de um conjunto de dados na Web. Com isso, é extremamente importante que o *Criador de Conjunto de Dados* informe os processos de derivação dos dados. De acordo com as boas práticas, a proveniência é um meio pelo qual os consumidores de um conjunto de dados julgam sua qualidade. Ademais, o entendimento de seu histórico e origem ajuda a determinar a confiabilidade do consumidor nos dados, além de fornecer um contexto interpretativo importante. Por esse motivo, a BP5, recomenda que os criadores forneçam informações de proveniência de dados para os seus *Usuários Finais*.

[1] https://www.w3.org/TR/dwbp/

– Subconjuntos dos dados: De acordo com as melhores práticas, mais precisamente a BP18, é aconselhável fornecer subconjuntos para conjuntos de dados de grande volume. Ou seja, quando houver um conjunto de dados com volume muito grande, é recomendado que ele seja distribuído em subconjuntos menores. Nesse mesmo sentido, é importante analisar a granularidade do conjunto de dados e verificar se é possível apresentá-lo em diferentes agrupamentos. Por exemplo, um conjunto de dados com todos os professores de uma Universidade, seria interessante que, além de disponibilizar o conjunto completo, fossem disponibilizados subconjuntos agrupando-os pelos centro acadêmicos da universidade.

  Portanto, para auxiliar e servir como apoio ao *Criador de Conjunto de Dados* e ao *Criador de Metadados* nas fases seguintes do modelo, é importante especificar o volume dos dados nesta fase, pois eles necessitarão dessa informação para realizar a publicação do conjunto e/ou subconjuntos.

– Metadados Descritivos: Fornecer metadados descritivos é importante para que os possíveis consumidores dos dados possam compreender com mais facilidade a natureza do conjunto de dados disponibilizado. Além disso, os metadados descritivos permitem que os agentes de busca possam encontrar com mais facilidade o conjunto na Web. O documento de boas práticas, além de recomendar o fornecimento desses metadados, sugere algumas informações que eles devem conter, por exemplo: Título e descrição do conjunto de dados, palavras-chaves que o descrevem, data de publicação, entidade responsável por disponibilizar o conjunto, contato, sua cobertura espacial, período temporal que os dados abrangem, data da última modificação, tema/categoria e frequência de atualização.

– Metadados Estruturais: Os metadados estruturais, como o próprio nome já diz, descrevem a estrutura de uma distribuição do conjunto de dados. Ele é essencial para que as pessoas possam entender os significados dos dados. Por estar ligado diretamente a estrutura do conjunto de dados, as informações diferem conforme os atributos de cada conjunto publicado.

– Licença dos dados: A licença dos dados serve para expressar claramente que o autor abdica de direitos de propriedade originais para dar a outros utilizadores a possibilidade de reutilizar, modificar e partilhar o seu trabalho. Além disso, ela serve para garantir aos consumidores a clareza na utilização das informações disponibilizadas.

– Formatos de distribuição: Os formatos de distribuição são importantes para especificar como os dados serão disponibilizados. Além disso, ter o entendimento de quais formatos serão disponibilizados já nessa fase é essencial para que, na atividade de criar os dados (*F02A01*) os criadores saibam quais distribuições

deverão ser criadas.

- ***F01A03 - Estabelecer solução para publicação do conjunto de dados***

  Para realizar a disponibilização do conjunto de dados na Web, deve-se antes escolher a solução que será utilizada. É importante que essa escolha ocorra na fase de *Planejamento* pois, dependendo da solução, a medida que outras atividades estão sendo realizadas ela pode ser desenvolvida em paralelo. Com o *Documento de Descrição do Conjunto de Dados*, o *Administrador* poderá ter uma visão geral do que almeja com esse conjunto de dados, dessa forma ajudando-o a tomar uma decisão mais coerente sobre qual solução utilizar.

  Com base nas recomendações de abordagens (classificadas como: catálogo de dados primitivo, básico e completo) propostas por Nečaský et al. (2013) em sua metodologia, elaboramos algumas recomendações/práticas que devem ser levadas em consideração na hora de escolher a solução de publicação do conjunto de dados. Para isso, fizemos uma classificação em quatro níveis de soluções de acordo com a abrangência de funcionalidades adotada. Para cada um desses níveis serão descritas alguns recomendações básicas que devem ser atendidas. Essas recomendações são:

  - *Solução primitiva:* Se o *Administrador* optar por utilizar uma solução simples para disponibilizar os dados na Web (*e.g* uma página html), é recomendado que, no mínimo, seja oferecida a opção de *download* do conjunto de dados (BP17). Exportando-o de acordo com as distribuições especificadas no *Documento de Descrição do Conjunto de Dados* (*F01A02*). Além disso, se houver subconjuntos, eles também serão disponibilizados para *download*.
  - *Solução básica:* A solução básica estende a primitiva, de modo que, para cada conjunto de dados seja oferecida uma página HTML onde sejam descritos, em formato legível por máquina, os metadados do conjunto de dados. Nessa abordagem, além de serem oferecidos *links* para o *download* das distribuições, também serão disponibilizados o *download* dos metadados em notação legível por máquina.
  - *Solução intermediária:* Na solução intermediária, além de ser ofertado tudo que está contido na básica, o conjunto de dados e seus metadados também poderão ser acessados por meio de uma API (BP23) que será disponibilizada pela solução. Além disso, a documentação da API (BP25) também será fornecida para que, futuramente, o ator com papel de *Usuário Final*, possa obter informações detalhadas sobre chamadas, parâmetros necessários e retornos esperados.
  - *Solução avançada:* A solução avançada estende a intermediária. Nela serão oferecidas pesquisas (*e.g* filtragens no conjunto dados, buscas por palavras-chaves), pré-visualizações das distribuições do conjunto de dados, ambiente para

coletar *feedback* dos usuários, suporte ao versionamento de conjunto de dados, dentre outras. Ou seja, ela é composta por todo um conjunto de funcionalidades que forneçam um suporte adicional para facilitar o manuseio do conjunto de dados.

Com esses quatro níveis de soluções recomendadas, cabe ao *Administrador* definir o que é primordial para o conjunto de dados que será publicado e, dentre as ferramentas existentes para publicação de dados na Web, escolher qual a mais adequada para as suas necessidades. Ressaltamos que se escolhida a *Solução Avançada*, na maioria das vezes, ela é independente do conjunto de dados, isto é, sua implantação ou desenvolvimento pode iniciar antes mesmo de haver a criação do conjunto de dados. Desse modo, se o *Administrador* julgar necessário, sua implementação já pode ser iniciada nesse momento. Por exemplo, se o *Administrador* escolher utilizar um catálogo de dados como o ckan[2], ele poderá iniciar a implantação do catálogo logo após sua escolha (visto que para iniciar sua implantação não depende do conjunto de dados) e em paralelo dar continuidade as próximas atividades do DWLM.

1.3.2 Criação

Finalizada a etapa de *Planejamento* é iniciada a fase de *Criação*. Nela, o *Criador de Conjunto de Dados* e o *Criador de Metadados*, deverão ter o primeiro contato com os dados propriamente ditos. Por já possuírem detalhes acerca das fontes de dados, nessa fase serão definidas as estratégias de coleta para cada uma delas. Assim como as transformações, atualizações e/ou modificações que poderão ser executadas no momento de manipulação dos dados e, posteriormente sua carga. Além disso, o *Documento de Descrição do Conjunto de Dados* que foi gerado na sessão anterior, servirá como um guia para a criação do conjunto de dados. Ele também poderá passar por atualizações, como o preenchimento e/ou modificação dos metadados descritivos e estruturais.

Para melhor compreensão do que será realizado nesta fase, imaginemos um *Criador de Conjunto de Dados* que precisará publicar um conjunto de dados a respeito dos resultados obtidos por candidatos à Carteira Nacional de Habilitação nas provas práticas realizadas no Detran. Levando em consideração que as fontes de dados sejam uma visão em um banco de dados relacional com as informações dos candidatos e um arquivo CSV que é atualizado com o *status* de aprovação ou reprovação dos candidatos após realização da prova. Para realizar a coleta dos dados a partir destas duas fontes, o *Criador do Conjunto de Dados* precisará estipular suas estratégias de coleta, que possivelmente serão uma consulta SQL para recuperar os dados da visão e o uso de alguma ferramenta para manuseio do CSV. Após isso, ele fará as modificações necessárias nos dados e, dependendo da solução de

[2] https://ckan.org/

publicação, realizará a carga desses dados em um novo banco de dados ou em alguma ferramenta de catalogação de dados, por exemplo.

Mas, antes de ser realizada a publicação de fato desse novo conjunto de dados, o *Criador de Conjunto de Dados* e o *Criador de Metadados* ainda precisarão avaliar a qualidade do conjunto criado, e o *Provedor de Dados*, por sua vez, necessitará validá-lo. Só após essa validação que ele estará apto a ser publicado.

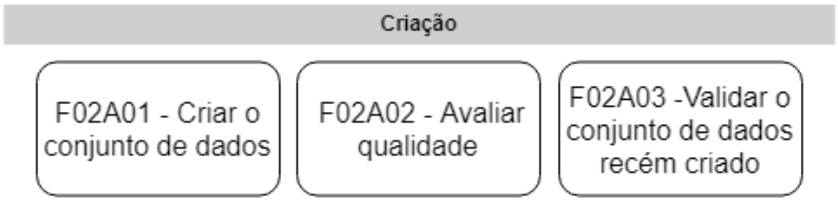

Figura 4 – Atividades da fase de *Criação*. Fonte: Autor

Dessa forma, para guiar o *Criador de Conjunto de Dados* e o *Criador de Metadados* nesse processo de criação, foram definidas três atividades (ver Figura 4). São elas:

- ***F02A01 - Criar o conjunto de dados***

  Esta atividade é responsável por realizar a criação do conjunto de dados. Como já se sabe a(s) fonte(s) de origem dos dados, por meio da atividade *F01A01*, cabe agora extrair os dados conforme a necessidade de cada fonte. Dito isso, para a realização dessa atividade é necessário tomar ciência dos cenários descritos a seguir:

  Cenário 1: Dados que não precisam passar por um processo de transformação, integração ou limpeza

  Nesse cenário, é válido dizer que os dados provenientes da(s) fonte(s) serão utilizados do jeito que estão para a criação do conjunto de dados, ou seja, eles não precisam passar por nenhum processo de transformação, integração e/ou limpeza. Desse modo, nesse cenário só ocorrerá um processo de carga para o armazenamento do conjunto de dados ou para criar o conjunto de dados nas distribuições especificadas no *Documento de Descrição do Conjunto de Dados*.

  Cenário 2: Dados provenientes de APIs em tempo real

  Nesse contexto, não haverá um armazenamento do conjunto de dados, pois os dados devem ser disponibilizados em tempo real. Desse modo, a criação do conjunto de dados consiste em disponibilizar um *link* para que os dados em tempo real possam ser acessados via Web.

  Cenário 3: Dados que necessitam passar por processos de transformação, integração ou limpeza

  Nesse caso, os dados provenientes das fontes não estão prontos para serem publicados e precisam passar por processos de transformação, integração ou limpeza. Identificamos

que nesse cenário, faz-se necessário um processo de Extração, Transformação e Carga (etl). Para uma melhor compreensão do que ocorrerá nesse processo, e por ser um cenário bastante comum no momento de criação, detalhamos todas as etapas e especificamos alguns componentes que podem ser utilizados em cada *pipeline* de ETL.

O processo ETL é uma técnica utilizada em *Data Warehouse* (DW) para realizar a extração de dados de várias fontes, sua limpeza, otimização e carga em um DW (FERREIRA et al., 2010). No contexto de Dados na Web temos um cenário semelhante no qual, na maioria das vezes, têm-se que extrair dados de várias fontes, realizar modificações/transformações e depois consolidado-los em um conjunto de dados. Nečaský et al. (2013) propôs uma metodologia para publicação de conjuntos de dados abertos que envolve um processo de ETL para a criação de conjuntos de dados. Com base nisso, realizamos algumas adaptações do que foi descrito por Nečaský et al. (2013) para o nosso contexto e sugerimos alguns componentes que podem ser usados em cada etapa do processo ETL. Dessa forma, para a criação dos dados será necessário:

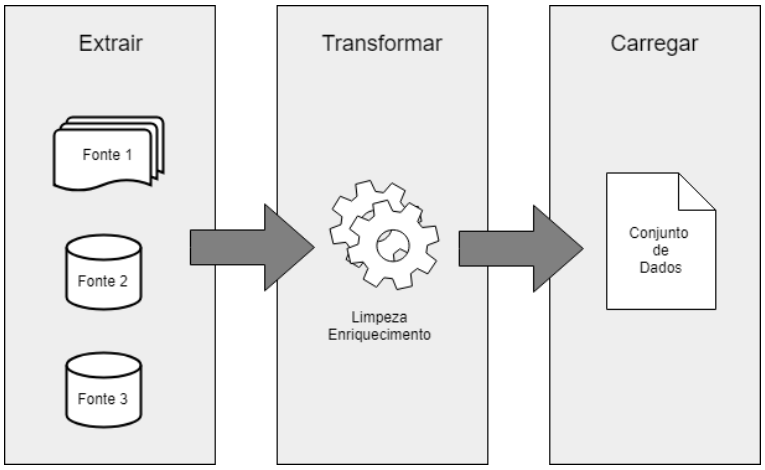

Figura 5 – Processo ETL para Dados na Web.
Fonte: Autor

1. Extrair: Nesse momento serão definidas as rotinas de extração que irão coletar os dados de cada fonte definida na atividade *F01A01*. Para isso, deverão ser projetados alguns extratores (componentes) de procedimentos ETL que acessem essas fontes e realizem a extração dos dados necessários. Esses extratores não realizarão nenhuma transformação nos dados. De acordo com Nečaský et al. (2013), é necessário que a ferramenta ETL escolhida suporte os seguintes componentes que devem ser usados como extratores:
   - Um componente que faça download de um arquivo de dados de uma determinada URL;

- Um componente que leia um arquivo de dados de um sistema de arquivos local;
- Um componente que acessa um banco de dados relacional com consultas SQL (SELECT);
- Um componente que acessa um banco de dados RDF com consultas sparql (SELECT, CONSTRUCT).

Para cada conjunto de dados, é necessário identificar seus extratores e configurá-los. Pois, cada componente necessitará de informações específicas, seja o caminho para um arquivo que está salvo em um sistema local ou uma consulta SQL para extrair os dados de um banco de dados relacional.

2. Transformar: Após coletados, os dados precisarão passar por um processo de limpeza, no qual os valores que estiverem com erros gramáticos ou de formatação serão organizados e ajustados. Além disso, nessa etapa, os dados tidos como sensíveis serão removidos. Nesse momento, os conjuntos serão estruturados e preparados de acordo com os metadados estruturais definidos na atividade *F01A02*. Assim como na etapa de extração, Nečaský et al. (2013) estipulou componentes que podem ser usados como transformadores em *pipelines* de ETL. São eles:
   - Componentes para transformação da estrutura e conversão de formato de dados;
     * Um componente para transformar formatos tabulares proprietários (xls, ods, etc) e resultados de consultas SQL para o formato CSV;
     * Um componente para transformar arquivos json em outros arquivos JSON;
     * Um componente para transformar arquivos JSON em arquivos XML e vice-versa;
     * Um componente para transformar os formatos CSV, XML e JSON em representação de RDF.
   - Componentes para transformar o conteúdo de um conjunto de dados;
     * Componentes para limpeza de dados
     * Componentes para anonimização de dados
   - Componentes para integração de dados;
     * Componentes para vincular conjuntos de dados a outros conjuntos de dados
     * Componentes para enriquecer conjuntos de dados com conteúdo de outros conjuntos de dados na base de links criados

3. Carregar: Com os dados já transformados, a última etapa do processo ETL é a carga. Nela, os dados já prontos, serão carregados em algum repositório, podendo ser um banco de dados e/ou arquivos nos formatos de distribuições especificados na atividade *F01A02*. Algumas recomendações propostas são:
   - Se o conjunto de dados estiver disponível apenas por download em massa de cada distribuição, o procedimento ETL deverá carregar os arquivos de dados em um local que possa ser acessado pelos usuários através do procolo HTTP ou FTP;
   - Se o conjunto estiver disponível por meio de API, o procedimento ETL deverá carregar os dados em um banco de dados relacional ou não-relacional;

Ao final dessa atividade, o *Criador de Conjunto de Dados* e o *Criador de Metadados* poderão atualizar o *Documento de Descrição do Conjunto de Dados* modificando os metadados, que por ventura, foram alterados durante essa atividade. Em seguida, será gerada uma nova versão do documento com os novos dados inseridos e/ou modificados.

- ***F02A02 - Avaliar qualidade***

  A Avaliação de Qualidade dos Dados (*Data Quality Assessment* - QA) é amplamente utilizada em várias áreas de pesquisa, como em bancos de dados relacionais, *data warehouse* e sistemas de gerenciamento de informação (UMBRICH; NEUMAIER; POLLERES, 2015). Ao longo do tempo, muitas áreas estabeleceram diversas métricas e técnicas para avaliar a qualidade de dados e serviços. Hoje, não temos métricas obrigatórias que deverão ser usadas para essa avaliação, o que existe são diversos trabalhos (*e.g* (ZAVERI et al., 2012), (ASKHAM et al., 2013)) que definiram várias métricas e cabe aos detentores dos dados estabelecerem quais serão utilizadas para medir a qualidade do seu conjunto de dados. Além disso, a ISO/IEC... (2014) também fornece um exemplo com 15 dimensões agrupadas em três categorias.

  Dentre as diversas classificações que existem para as dimensões e critérios de qualidade de dados, neste trabalho destacamos a classificação proposta por Zaveri et al. (2012). Esta classificação foi escolhida porque suas dimensões foram propostas a partir de um *survey* realizado com 30 artigos na área de QA. Apesar do autor definir dimensões voltadas para Dados Conectados, muitas delas podem ser aproveitadas para os dados na Web de uma forma geral. Algumas das dimensões citadas por Zaveri et al. (2012) são:

  - Disponibilidade: A medida em que os metadados e o conjunto de dados podem ser obtidos, ou seja, se estão prontos para uso;
  - Licenciamento: Verifica a concessão de permissão para um consumidor reutilizar um conjunto de dados sob condições definida;

- Segurança: É a medida que verifica se os dados são protegidos contra alteração e uso indevido;
- Consistência: Verifica se o conjunto de dados está livre de contradições com relação a mecanismos particulares de representação e inferência de conhecimento;
- Completude: Verifica se todas as informações necessárias estão descritas no conjunto de dados;
- Confiabilidade: É a medida que verifica o grau em que a informação é aceita como verdadeira, correta e confiável;
- Compreensibilidade: Refere-se a clareza de compreensão sem ambiguidades;
- Versatilidade: Verifica disponibilidade dos dados em diferentes representações e de forma internacionalizada.

Desse modo, nesta atividade o *Criador de Conjunto de Dados* ficará responsável por escolher as métricas de qualidade que julgar importantes para o seu contexto e avaliar a qualidade do seu conjunto de dados. É importante salientar que esta atividade não é obrigatória, isto é, ficará a critério do *Criador de Conjunto de Dados* se será necessário uma avaliação de qualidade do seu conjunto de dados antes de ser publicado.

- ***F02A03 - Validar o conjunto de dados recém criado***

  A atividade de validação do conjunto de dados é realizada pelo *Provedor dos Dados*. Ele irá verificar, antes dos dados serem publicados, se o conjunto de dados condiz com o que ele almejava. Além disso, essa fase também é necessária para detectar inconsistências ou erros, bem como apontar possíveis pontos de sensibilidade nos dados (*e.g* dados pessoais, valores). Caso o conjunto não esteja condizente com as expectativas do provedor ou ele ainda identifique algum erro, ele precisará descrever tais erros em um *Documento de Inconsistência de Dados* descrevendo todos os pontos de erros/inconsistências encontrados no conjunto. Se ele julgar o conjunto como correto, o *Provedor de Dados* precisará assinar um *Termo de Consentimento do Conjunto de Dados*, que confirmará sua aceitação para a fase de *Publicação*.

1.3.3 Publicação

Após validado, o conjunto de dados chega à fase de *Publicação*. Nessa fase, o conjunto de dados deverá ser disponibilizado na Web de acordo com a solução escolhida para sua publicação. A publicação não envolve apenas o conjunto de dados em si, mas também a publicação dos metadados relacionados a ele, assim como os possíveis subconjuntos de dados gerados a partir dele. Caso o conjunto de dados possua subconjuntos, é interessante que os administradores, além de fornecerem opções de download para cada subconjunto

separadamente, também forneçam opções de download em massa, de forma que o conjunto de dados possa ser recuperado por completo. Esse tipo de ação pode ocorrer por download a partir de alguma URI ou por solicitação via API.

Muitas das informações que serão solicitadas no momento da publicação estarão contidas no *Documento de Descrição do Conjunto de Dados*. Como, em geral, não se tem nenhum documento que reúna esses dados, ocorre que o *Administrador*, no momento da publicação, necessite ir em busca de todos esses dados de última hora, podendo ocasionar erros e inconsistências.

Para a fase de *Publicação* foram estipuladas três atividades principais (ver Figura 6), são elas:

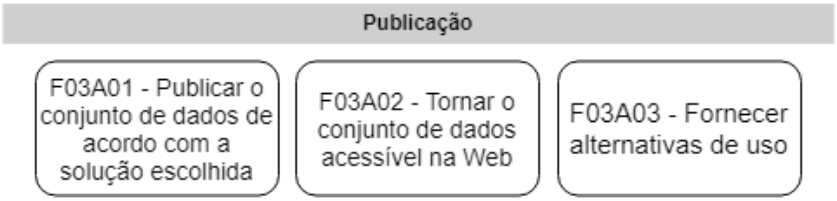

Figura 6 – Atividades da fase de *Publicação*. Fonte: Autor

- ***F03A01 - Publicar o conjunto de dados de acordo com a solução escolhida***

  A primeira atividade a ser realizada na etapa de *Publicação* é a publicação do conjunto de dados na solução escolhida. Essa publicação irá ser realizada de acordo com o nível da solução estabelecida na atividade *F01A03*. Esses níveis foram baseados nas recomendações propostas por Nečaský et al. (2013) em sua metodologia. Dessa forma, detalhamos como a publicação será realizada de acordo com os níveis de soluções propostas.

  - *Solução Primitiva:* A publicação do conjunto de dados na solução primitiva consiste na criação da página HTML e na inclusão das distribuições do conjunto de dados criado na atividade *F02A01*.
  - *Solução Básica:* Nessa solução, além da criação da página HTML e a publicação do conjunto de dados criado, o *Administrador* também deverá publicar os metadados do conjunto.
  - *Solução Intermediária:* Na intermediária, o *Administrador* irá criar a página HTML para os conjuntos de dados e metadados, realizar a publicação dos conjuntos de dados e metadados na solução e desenvolver a API que será disponibilizada, bem como sua documentação.
  - *Solução Avançada:* Essa solução é um pouco diferente das soluções acima, pois, como dito na atividade *F02A01*, a sua implantação já pode ter sido iniciada

na fase de *Planejamento*, visto que ela é independente do conjunto de dados. Desse modo, nesse momento o *Administrador* ir apenas incorporar o conjunto de dados a solução. Seguindo o exemplo da atividade *F02A01*, no contexto do CKAN essa atividade seria a criação do conjunto de dados no catálogo juntamente com o *upload* dos arquivos nas suas respectivas distribuições.

É importante deixar claro que as três primeiras soluções dependem diretamente do conjunto de dados para serem executadas, visto que não faz sentido criar uma página HTML sem ter o conjunto de dados e seus metadados primeiramente criados. Assim como iniciar o desenvolver de uma API sem haver um conhecimento prévio do conjunto. Por esse motivo, para esses três primeiros níveis de solução é recomendado que seu desenvolvimento ocorra nessa atividade. Em contraste, a *Solução Avançada* tem um cenário diferente, como visto na atividade *F02A01*, ela é independe do conjunto de dados. Ou seja, seu processo de implantação/desenvolvimento pode ser iniciado bem antes da criação do conjunto. Dessa forma, temos situações de publicação totalmente diferentes para cada nível de solução escolhida.

- ***F03A02 - Tornar o conjunto de dados acessível na Web***

  Depois de publicar o conjunto de dados e seus metadados na solução escolhida, é recomendado que o *Administrador* estabeleça alguns padrões de URLs para acesso ao conjunto de dados. Alguns exemplos propostos por Nečaský et al. (2013) foram:

  `http://{base-URL}/dataset/{ID}`

  `http://{base-URL}/dataset/{dataset-id}/{distribution-id}`

  Hoje não há um padrão definido para a publicação de dados na Web, o que existem são propostas em algumas subáreas, como Dados Abertos (NEčASKý et al., 2013) e Dados Conectados. Além disso, a depender da solução escolhida, algumas já podem oferecer URLs específicas para o conjunto de dados e seus recursos.

  Ademais, a partir da atividade *F03A01* o conjunto de dados estará publicado. No entanto, para que ele se torne acessível, ele deve ser disponibilizado na Web e fazer uso dos seus protocolos padrões (i.e. http). Pois, a publicação de conjuntos de dados em ambientes internos como vpn ou Intranets, não faz dele acessível na Web, visto que, o *Usuário Final* que não esteja dentro desse ambiente não conseguirão acessá-lo. Desse modo, essa atividade faz-se necessária para assegurar que o conjunto seja disponibilizado de forma que todos os usuários da Web possam acessá-lo.

- ***F03A03 - Fornecer alternativas de uso***

  A partir do momento em que os dados foram publicados, o "*Administrador*" poderá trabalhar no fornecimento de alternativas de uso. Para Lóscio, Burle e Calegari (2017), é interessante fornecer visualizações complementares (BP32) para que os

consumidores possam ter uma visão imediata de uso dos dados, apresentando-os de forma que possam ser facilmente compreendidos. Essas visualizações podem ser desde tabelas com alguns filtros simples, até gráficos estatísticos com análises mais aprofundadas.

1.3.4 Consumo

Após publicados, os conjuntos de dados estarão aptos para serem consumidos. Esta fase representa as diferentes formas de uso e manipulação dos dados, desde consumo, utilizando APIs, acesso a páginas estáticas em HTML ou, até mesmo, por visualizações de dados já definidas. Além disso, os usuários poderão optar por utilizar o conjunto de dados para a criação de novas aplicações e, assim, realizar um uso externo desse conjunto de dados.

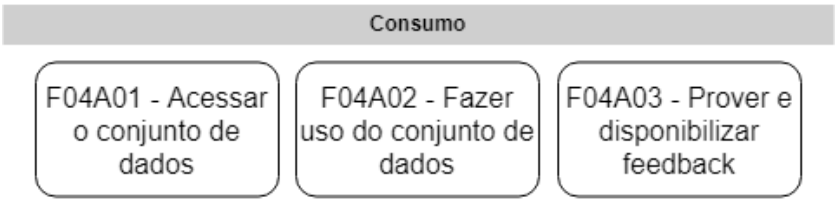

Figura 7 – Atividades da fase de *Consumo*.
Fonte: Autor

Estabelecemos três atividades que compõem essa fase, são elas:

- ***F04A01 - Acessar o conjunto de dados***

  Nessa etapa, os consumidores de dados terão acesso ao conjunto de dados. A partir do momento que os dados são acessados eles estão, automaticamente, sendo consumidos. Esse acesso poderá derivar de diferentes tipos de usuários, desde empresas interessadas em usar os dados para melhoria dos seus serviços e produtos, até um desenvolvedor que vise utilizar os dados para a criação de alguma aplicação. Esses diferentes atores, na fase de *Consumo*, estarão assumindo o papel de *Usuários Finais*, como especificado no modelo DWLM.

- ***F04A02 - Fazer uso do conjunto de dados***

  Após acessar o conjunto de dados, o *Usuário Final* pode optar por usá-lo para a realização de atividades adicionais. Ou seja, o conjunto pode ser reutilizado para a construção de visualizações, criação de gráficos estáticos e dinâmicos, criação de análises ou para o desenvolvimento de aplicações. Nesse cenário, é importante que os usuários sigam os termos de licença impostos no conjunto de dados publicado (BP34). Dessa forma, os provedores de dados poderão presumir que o seu trabalho está sendo reutilizado de acordo com os requisitos de licenciamento (LÓSCIO; BURLE;

CALEGARI, 2017). Ademais, é interessante citar a publicação original (BP35) pois, além de aumentar a confiabilidade dos dados para os usuários que irão consumi-los, ajudará o *Provedor de Dados* a receber o merecido reconhecimento e o incentivará a continuar compartilhando dados na Web.

- ***F04A03 - Prover e disponibilizar feedback***

  Para que os dados estejam em conformidade com as necessidades do consumidor, é importante que os publicadores ofereçam um local onde os usuários possam enviar *feedback* sobre o conjunto de dados consumido (BP29). O *feedback* traz muitos benefícios, pois além de melhorar a integridade dos dados publicados, pode incentivar a publicação de novos dados (LÓSCIO; BURLE; CALEGARI, 2017). Após a coleta desse *feedback*, é recomendado disponibilizá-lo para que outros consumidores de dados possam ter acesso a essas informações (BP30). Torná-lo acessível ao público permite que os usuários tomem conhecimento de outros consumidores de dados, ofereçam suporte para um ambiente colaborativo e permitam experiências entre os usuários da comunidade (LÓSCIO; BURLE; CALEGARI, 2017).

1.3.5 Refinamento

No refinamento de conjuntos de dados publicados na Web são realizadas operações de identificação e correção de erros, adição e atualização de dados, metadados e semântica, visando aumentar da qualidade do conjunto de dados (REVIEW, 2018). Dessa forma, a fase de *Refinamento* do DWLM compreende atividades relacionadas a correções e enriquecimento de um conjunto de dados publicado. Ou seja, quando o usuário tem acesso ao conjunto de dados ele pode sugerir e realizar melhorias, ocasionando um refinamento. Essa etapa irá acontecer após o seu consumo, visto que o usuário para sugerir essas melhorias precisará ter tido algum contato prévio com o conjunto de dados, em outras palavras, ele precisará ter consumido esse conjunto.

Uma das formas de iniciar a fase de *Refinamento* é por meio da atividade de *feedback* (F04A03), contida na fase de *Consumo*. Nessa atividade, o *Usuário Final* poderá enviar sugestões de correção e/ou enriquecimento do conjunto de dados. E, a partir desse *feedback* o *Criador de Conjunto de Dados* irá modificar o conjunto a fim de corrigi-lo ou enriquece-lo. No entanto, essa fase de *Refinamento* também pode ser inciada a partir do *Criador de Conjunto de dados* que, ao verificar alguma irregularidade no conjunto de dados, poderá prontamente refiná-lo.

Para esta fase de *Refinamento*, definimos algumas atividades que a compõem. São elas:

- ***F05A01 - Corrigir e enriquecer o conjunto de dados***

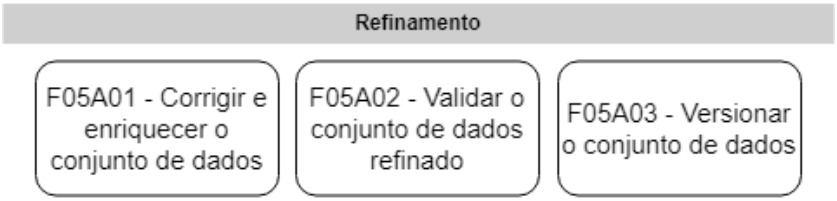

Figura 8 – Atividades da fase de *Refinamento*.
Fonte: Autor

Corrigir o conjunto de dados envolve um processo de limpeza de dados. Essa limpeza consiste na detecção e remoção de erros e inconsistências, objetivando o aumento da qualidade dos dados. De uma maneira geral, os erros podem ser classificados em dois níveis: Nível de Esquema e Nível de Instância. No nível de esquema são problemas relacionados ao esquema ou estrutura do conjunto de dados. Já os problemas relacionados ao conteúdos dos dados são tratados como erros em nível de instância (RAHM; DO, 2000).

Por outro lado, o enriquecimento dos dados tem o objetivo de agregar valor aos conjuntos, seja por meio da adição de novos dados e metadados, ou por anotações semânticas. Existem várias técnicas propostas que podem ser usadas para realizar o enriquecimento, como por exemplo: Anotações Semânticas (UREN et al., 2006), Vinculação e Mapeamento de Recursos (SORRENTINO et al., 2013) e Conversão para modelos de dados semânticos.

No trabalho de Review (2018) foram estipulados alguns procedimentos de limpeza e enriquecimento de dados. No processo de limpeza, são definidas algumas operações de busca e correções de erros, enquanto no processo de enriquecimento foi definido uma operação para enriquecer dados. Além disso, para cada operação foram especificados alguns procedimentos que podem ser realizados. Dito isso, os procedimentos definidos no trabalho foram:

- Procedimentos de Limpeza
    1. Correção de valores falsos;
    2. Correção de ortografia;
    3. Correção de valores ocultos;
    4. Correção de valores abreviados;
    5. Correção de erros referenciais;
    6. Correção de valores agregados;
    7. Correção de valores desviados;
    8. Remoção de registros duplicado.
- Procedimentos de Enriquecimento

1. Adição de dados em atributos com valores vazios;
2. Adição de atributo;
3. Adição de registros;
4. Adição de metadado;
5. Atualização de metadados;
6. Anotação semântica de um valor;
7. Anotação semântica de um metadado.

Após realizadas as correções e/ou enriquecimento do conjunto de dados, é recomendado criar um documento de *log* que especifique o que foi alterado para que o *Provedor de Dados* possa, na atividade seguinte, validar.

- ***F05A02 - Validar o conjunto de dados refinado***

  A fase de validação compreende o momento em que o conjunto de dados é validado pelo *Provedor de Dados*. Ou seja, após o *Criador de Conjunto de Dados* realizar as correções no conjunto, ele será analisado para identificar se não há nenhuma anomalia nos dados que foram alterados. Caso as alterações realizadas sejam convenientes, o *Administrador* irá incorporá-las ao conjunto e um nova versão poderá ser disponibilizada.

- ***F05A03 - Versionar o conjunto de dados***

  Como dito na atividade anterior, após validadas as correções e/ou enriquecimento dos dados é necessário gerar uma nova versão do conjunto de dados alterado. Para realizar esse versionamento, o documento de melhores práticas diz que é necessário fornecer um identificador de versão (BP7) e um histórico das versões (BP8). O identificador de versão é importante para determinar se o conjunto de dados foi alterado ao longo do tempo e para que os consumidores possam identificar qual a versão atual que ele está trabalhando.

Após a finalização dessa fase (como observado na figura 2), o conjunto de dados segue para etapa de *Publicação* (atividade *F03A02*) novamente. Uma vez que, a nova versão, com as atualizações realizadas no refinamento, deve ser tornada acessível.

1.3.6 Remoção

A fase de *Remoção* finaliza o ciclo do DWLM. Levando em consideração que o conjunto de dados não estará disponível sob demanda o tempo todo, essa fase faz-se necessária para realizar a preservação do conjunto de dados que terá o seu acesso removido. Por alguma razão, o *Provedor de Dados* poderá solicitar a remoção de acesso a algum conjunto de dados disponível na Web. Essa solicitação é realizada por meio de

um *Documento de Solicitação de Remoção*, este documento será necessário para que o *Provedor de Dados* informe o motivo do acesso ao conjunto de dados precisar ser removido. Contudo, para realizar essa remoção, é necessário tomar algumas precauções. Para isso, definimos uma atividade que abordará esses cuidados.

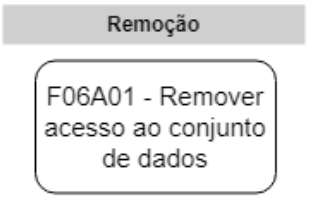

Figura 9 – Atividade da fase de *Remoção*.
Fonte: Autor

- ***F06A01 - Remover acesso ao conjunto de dados***

  O ponto principal dessa atividade é a remoção do acesso ao conjunto de dados. Para isso, é importante realizar a preservação do seu identificador (BP27). Essa preservação é necessária para que ao acessarmos a URI de um conjunto que teve seu acesso removido, não tenhamos como resposta um código 404 (*Not Found*). Com esse tipo de resposta, o usuário não saberá se a falta de disponibilidade é permanente ou temporária, planejada ou acidental (LÓSCIO; BURLE; CALEGARI, 2017). Para resolver esse problema, o documento de melhores práticas propôs que ao remover o acesso de um conjunto de dados deve-se criar uma página de resposta informado-o que o conjunto não está mais disponível, o motivo pelo qual houve essa remoção e que ele poderá solicitar uma cópia, se possível.

  Contudo, quando o conjunto de dados que será removido possuir distribuições em RDF, é recomendado que os administradores realizem uma avaliação de cobertura do conjunto (BP27) antes deles serem preservados. Essa avaliação é primordial para conjuntos que utilizem vocabulários pouco usados. Pois, de acordo com Lóscio, Burle e Calegari (2017) ao preservar o conjunto devemos garantir que todas as informações, que são necessárias para o seu entendimento futuro estejam preservadas junto com ele. Ao apontar para vocabulários ou recursos externos há um risco de, daqui a alguns anos, se for preciso o uso desse conjunto por algum motivo desconhecido, dados tenham se perdido por não estarem mais disponíveis na Web. Portanto, é importante a avaliação da cobertura antes, para que em situações assim, os recursos externos sejam preservados junto com o conjunto de dados.

1.4 Diagrama de Atividades do DWLM

A Figura 10 ilustra o fluxo principal de atividades do modelo DWLM. Neste diagrama de atividades, podemos visualizar como as atividades de cada fase se comunicam

e os papéis responsáveis por executar cada uma delas. Além disso, podemos ter uma visão mais clara do que acontece em alguns pontos de decisão, como é o caso da atividade *F02A03 - Validar conjunto de dados recém criado*, pois, no momento em que o *Provedor de Dados* validar o conjunto de dados, o fluxo segue para a próxima atividade, que é *F03A01 - Publicar o conjunto de dados de acordo com a solução escolhida*. Caso contrário, o conjunto volta para a atividade de *F02A01 - Criar o conjunto de dados*. Ademais, um dos pontos que também merece destaque no diagrama são as atividades da fase do *Refinamento*. Depois do *Usuário Final* fornecer um feedback o *Criador de Conjunto de Dados* irá realizar a atividade de *F05A01 - Corrigir e enriquecer o conjunto de dados*. Logo após, o conjunto segue para uma validação do *Provedor de Dados* e, após essa validação, será criada uma nova versão do conjunto com os dados atualizados. Essa nova versão criada, segue para a atividade *F03A02 - tornar o conjunto de dados acessível na Web*, que por sua vez, deixará a nova versão acessível.

O diagrama também ilustra o caso de atividades paralelas, ou seja, as atividades que podem ser executadas enquanto outras também estão sendo. Além disso, o fluxo representado neste diagrama é, para nós, o fluxo principal de atividades. Por certo, outros modelos de fluxos devem existir, pois eles dependerão muito do contexto para o qual o DWLM será aplicado.

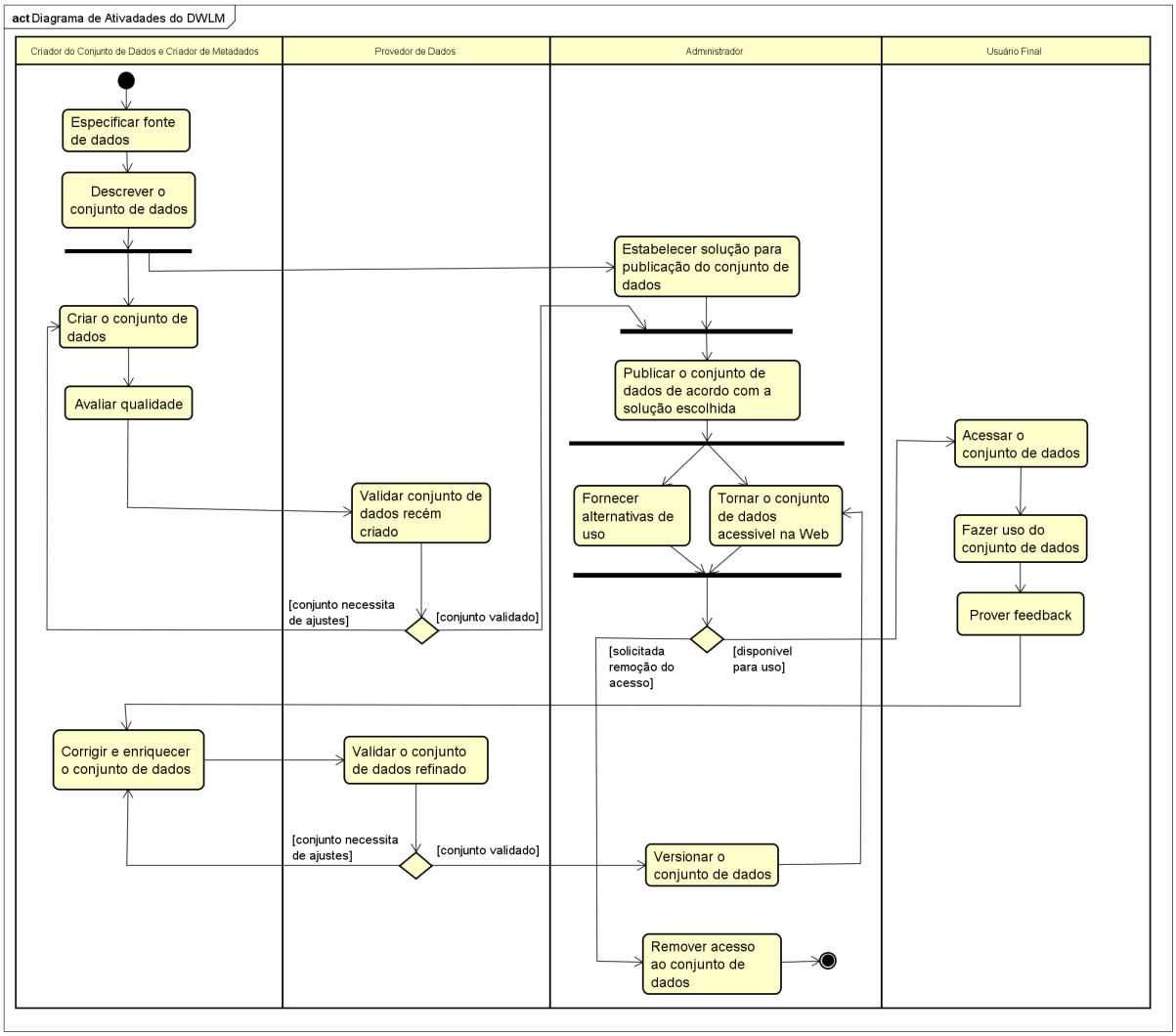

Figura 10 – Diagrama de Atividades do Modelo de Ciclo de Vida de Dados na Web. Fonte: Autor

1.5 Características e Classificação

Há uma série de características que podem ajudar a classificar diferentes tipos de modelos de ciclo de vida. No ADLM, Möller (2013) definiu algumas características do modelo que, posteriormente, são usadas para sua classificação. Além disso, em uma pesquisa realizada por Cox e Tam (2018) acerca de modelos de ciclo de vida, foi realizada uma classificação dos modelos estudados por eles. Essa pesquisa levou em consideração as classificações definidas pelos autores Möller (2013), Ma e Wang (2010) e Carlson (2014). Dito isso, utilizaremos algumas dimensões elencadas na pesquisa de Cox e Tam (2018) para definir as características do DWLM e depois classificá-lo. Seguindo o formato de classificação definido por Cox e Tam (2018) nossas dimensões foram agrupadas em: "*Escopo*" e "*Elementos e Processos*".

1.5.1 Escopo

As características de escopo dizem respeito ao que o modelo aborda, ou seja, o escopo ao qual está voltado o domínio do modelo e ao seu modo de representação. Como características de escopo definimos as seguintes:

- *Indivíduo vs. Organização vs. Comunidade:* De acordo com Carlson (2014) os modelos de ciclo de vida baseados em indivíduos representam as etapas que compreendem um projeto específico. Ou seja, eles servem como uma ferramenta eficaz para projetar e executar um projeto. Nele serão descritas as atividades que precisam ser realizadas, como serão realizadas e quem as executará. Um pouco semelhante ao modelo baseado em indivíduos é o modelo baseado em organização. Contudo, eles servem para um propósito diferente. Os modelos baseados em organização são representações mais gerais dos estágios comuns de ciclos de vida para um determinado campo de prática (CARLSON, 2014). Ele destina-se a especificar passos e etapas em que um usuário poderá se guiar para alcançar seu fim. E por último, temos os ciclos de vida baseados em comunidade. Eles foram desenvolvidos para apoiar ou atender as necessidades de um comunidade específica. Para Carlson (2014) os modelos baseados em comunidade oferecem uma visão geral de alto nível, representando os componentes das melhores práticas recomendadas e suas conexões entre si. Ressaltamos que essas categorias idealizadas por Carlson (2014) não são excludentes, por exemplo, um modelo baseado na comunidade também poderia incluir elementos de apoio organizacional ou individual.

- *Prescritivos vs. descritivos:* Segundo Möller (2013), o termo prescritivo é imposto a modelos de ciclo de vida que estabelecem um conjunto de etapas sugeridas para que outros o utilizem. Em contraste, um modelo descritivo examinará um determinado sistema e localizará nele um ciclo de vida. Ou seja, se um modelo de ciclo de vida está sugerindo uma metodologia de como o processo deve ser executado e descrevendo as melhores práticas que devem serem seguidas, ele deve ser classificado como um modelo de ciclo de vida prescritivo. No entanto, se ele está descrevendo um processo já existente, ele é tido como um modelo descritivo.

1.5.2 Elementos e Processos

Na dimensão de elementos e processos são descritas as características mais voltadas aos dados que compõem o modelo de ciclo de vida. Dessa forma, são definidas algumas características que descrevem desde a granularidade do modelo de ciclo de vida, sua heterogeneidade até a definição se é um modelo de clico de vida com dados centralizado ou distribuído.

- *Granularidade:* Determinar a granularidade de um modelo de ciclo de vida é importante para descobrir se, ao passar por cada etapa individual do modelo, estão sendo manipulados todos os dados dele ou partes. Um modelo em que todos os dados são afetados em cada iteração tem uma granularidade grossa, enquanto um modelo em que somente partes dos dados são afetadas tem uma granularidade fina (MöLLER, 2013).

- *Homogêneo vs. heterogêneo:* Um modelo de ciclo de vida é considerado homogêneo quando os dados que ele descreve são homogêneos, ou seja, quando seu esquema é conhecido de antemão, e nenhum dado de semântica desconhecida entrarão no ciclo. Em contraste, um ciclo heterogêneo é quando os dados descritos nele são heterogêneos, ou seja, quando seu esquema não é conhecido previamente (MöLLER, 2013). Ressaltamos, que essa característica não se restringe ao formato dos dados, mas sim a sua semântica.

- *Aberto vs. fechado:* A ocorrência de um modelo de ciclo de vida ser aberto ou fechado está diretamente relacionado ao fato dele ser homogêneo ou heterogêneo. Essa característica possibilita a inclusão de dados externos que, a priori, não estavam previstos de participar do escopo de domínio do modelo (MöLLER, 2013). Desse modo, caso o modelo de ciclo de vida permita a introdução de dados externos ele será classificado como aberto, do contrário, considere-o fechado.

- *Centralizado vs. distribuído:* Esta característica descreve a natureza física de um modelo de ciclo de vida de dados. Se os conjuntos de dados residirem em uma única infraestrutura controlada centralmente, o ciclo de vida dele será centralizado. No entanto, se for distribuído em uma rede sem ponto único de controle, será classificado como distribuído (MöLLER, 2013).

- *Visualização*: Esta características aborda o tipo de visualização do modelo de ciclo de vida. Segundo Cox e Tam (2018) há três tipos gerais de modelos de ciclos de vida. Que são os: sequenciais, incrementais e evolutivos. Em um modelo do tipo sequencial ou cascata, cada fase só pode ser alcançada se a anterior estiver terminada, dessa forma, uma nova iteração no ciclo só poderá ser iniciada quando todas as etapas forem executadas. Por outro lado, no modelo incremental poderá haver o início de uma nova iteração antes mesmo do ciclo ter terminado completamente. Por fim, o modelo evolutivo indica que os dados podem mudar a qualquer momento, o que indica o início de novas iterações sempre que houver a necessidade.

1.5.3 Classificação do DWLM

Com o conjunto de características definidas, classificamos o DWLM em relação a cada característica apresentada na Seção 1.5. No Quadro 2 apresentamos essa classificação.

Na dimensão de Escopo foram apresentadas duas características: Determinar se o modelo é baseado em Indivíduo, Organização ou Comunidade e se ele é Prescritivo ou Descritivo. Para nós, o DWLM é um exemplo de um modelo baseado em Comunidade e Prescritivo, pois ele foi construído para ser um modelo genérico com o intuito de atender as necessidades da comunidade de Dados na Web. Prescritivo porque ele busca ser um modelo de referencia para que outros possam surgir a partir dele.

A segunda dimensão foi denominada como Elementos e Processos, nela foram apresentadas as seguintes características: Granularidade, Homogêneo ou Heterogêneo, Aberto ou Fechado, Centralizado ou Distribuído. Em relação a Granularidade classificamos o DWLM com uma granularidade Fina, pois a cada etapa do modelo não estaremos manipulando todos os seus dados, mas sim algumas partes. Sobre sua homogeneidade, o classificamos como heterogêneo porque lidamos com Dados na Web e não é possível saber previamente a semântica dos dados que serão trabalhados nesse modelo. Cada domínio empregado poderá utilizar semânticas de dados distintos. Em consequência, o DWLM também é um modelo Aberto e Distribuído, visto que lidamos com um ambiente totalmente aberto que é a Web.

Por último, foi mostrada a característica de Visualização. Essa característica tem o objetivo de identificar se um ciclo de vida é sequencial, incremental ou evolutivo. Para o DWLM, classificamos-o como evolutivo, pois a ideia é que novas iterações possam ser iniciadas quando houver necessidade.

Tabela 2 – Classificação do DWLM

Escopo	***Característica***
Indivíduo vs. Organização vs. Comunidade	*comunidade*
Prescritivos vs. Descritivos	*prescritivo*
Elementos e Processos	
Granularidade	*fina*
Homogêneo vs. Heterogêneo	*heterogêneo*
Aberto vs. Fechado	*aberto*
Centralizado vs. Distribuído	*distribuído*
Visualização	*evolutivo*

Referências

ASKHAM, N. et al. The six primary dimensions for data quality assessment. *Semantic Web*, 2013. Citado na página 14.

CARLSON, J. The use of life cycle models in developing and supporting data services. *Research Data Management: Practical Strategies for Information Professionals*, p. 63–86, 2014. Citado 2 vezes nas páginas 24 e 25.

COX, A. M.; TAM, W. W. T. A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, v. 70, n. 2, p. 142–157, 2018. Citado 2 vezes nas páginas 24 e 26.

FERREIRA, J. et al. O processo etl em sistemas data warehouse. *INForum 2010 - II Simpósio de Informática*, p. 757–765, 09 2010. Citado na página 12.

ISO/IEC 25012. 2014. Http://iso25000.com/index.php/en/iso-25000-standards/iso-25012. Acessado em 13 de janeiro de 2019. Citado na página 14.

KOSCH, H. et al. The life cycle of multimedia metadata. *IEEE MultiMedia*, v. 12, p. 80–86, 2005. Citado na página 6.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. W3C Recommendation, *Data on the Web Best Practices*. 2017. https://www.w3.org/TR/dwbp/. Acessado em 15 de dezembro de 2019. Citado 4 vezes nas páginas 2, 17, 19 e 22.

LÓSCIO, B. F.; OLIVEIRA, M. I. S.; BITTENCOURT, I. I. Publicação e Consumo de Dados na Web: Conceitos e Desafios. *Tópicos em Gerenciamento de Dados e Informações (Mini Cursos - SBBD 2015)*, d, p. 39–69, 2015. Disponível em: <http://dexl.lncc.br/sbbd2015/anais/ShortCourses.pdf>. Citado na página 2.

MA, F.; WANG, J. The review of studies on information lifecycle ii: the perspective of management. *Journal of the China Society for Scientific and Technical Information*, v. 29, p. 1080–1086, 2010. Citado na página 24.

MöLLER, K. Lifecycle models of data-centric systems and domains. *Semantic Web*, v. 4, p. 67–88, 2013. Disponível em: <http://doi.org/10.3233/SW.2012.0060>. Citado 7 vezes nas páginas 2, 4, 5, 6, 24, 25 e 26.

NEčASKý, M. et al. Methodology for publishing datasets as open data. *DELIVERABLE D5.1*, 2013. Disponível em: <https://www.comsode.eu/wp-content/uploads/D5.1-Methodology_for_publishing_datasets_as_open_data.pdf>. Citado 5 vezes nas páginas 9, 12, 13, 16 e 17.

RAHM, E.; DO, H. H. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, v. 23, n. 4, p. 3–13, 2000. Citado na página 20.

REVIEW, B. *Blind Review*. Dissertação (Mestrado) — Blind Review, Blind Review, 2018. Citado 2 vezes nas páginas 19 e 20.

SORRENTINO, S. et al. Semantic annotation and publication of linked open data. In: SPRINGER. *International Conference on Computational Science and Its Applications*. [S.l.], 2013. p. 462–474. Citado na página 20.

UMBRICH, J.; NEUMAIER, S.; POLLERES, A. Quality assessment e evolution of open data portals. *3rd International Conference on Future Internet of Things and Cloud*, 10 2015. Citado na página 14.

UREN, V. et al. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: science, services and agents on the World Wide Web*, Elsevier, v. 4, n. 1, p. 14–28, 2006. Citado na página 20.

ZAVERI, A. et al. Quality assessment for linked data: A survey. *Semantic Web*, p. 63–93, 2012. Citado na página 14.