Pós-Graduação em Ciência da Computação

Karina Moura da Silva

Um Modelo de Ciclo de Vida de Dados na Web

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2019

Karina Moura da Silva

Um Modelo de Ciclo de Vida de Dados na Web

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Banco de Dados
Orientador: Profa. Dra. Bernadette Farias Lóscio

Recife
2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S586m Silva, Karina Moura da
Um modelo de ciclo de vida de dados na web / Karina Moura da Silva. – 2019.
107 f.: il., fig., tab.

Orientadora: Bernadette Farias Lóscio.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.
Inclui referências e apêndices.

1. Banco de dados. 2. Dados na web. 3. Publicação de dados. I. Lóscio, Bernadette Farias (orientadora). II. Título.

025.04 CDD (23. ed.) UFPE- MEI 2019-085

Karina Moura da Silva

“Um Modelo de Ciclo de Vida de Dados na Web”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 14/03/2019.

BANCA EXAMINADORA

_____

Prof. Dr. Kiev Santos da Gama
Centro de Informática/UFPE

_____

Profa. Dra. Damires Yluska de Souza Fernandes
Instituto Federal da Paraíba/Campus João Pessoa

_____

Profa. Dra. Bernadette Farias Lóscio
Centro de Informática/UFPE
(**Orientadora**)

Dedico esta dissertação a todas as pessoas que, de alguma forma, me ajudaram a chegar até aqui. Em especial, a minha orientadora, meus pais e minha irmã.

AGRADECIMENTOS

O caminho percorrido até chegar a versão final desta dissertação não foi fácil. Mas, graças a Deus chegou o fim! Foram dias e noites de preocupação, ansiedade e angústia. Esperava ansiosamente pelo momento em que começaria a escrever esses agradecimentos. Pois, para mim, era o sinal de que este trabalho estaria chegando ao fim e poderia agradecer a todos que colaboraram para a sua conclusão.

Em primeiro lugar, gostaria de agradecer a minha orientadora, Profª Bernadette Farias Lóscio (Berna), por ter me orientado, me apoiado e confiado em mim durante toda essa jornada. Agradeço principalmente por ter acreditado em mim até nos momentos em que eu mesma não acreditava. Foram muitas reuniões, principalmente nesses últimos seis meses, e muitas vezes, eu me perguntava se seria o momento de desistir. Mas, a "senhora", sempre com seu jeitinho, conversava comigo e me fazia vê que ainda era possível. E foi!! Berna, muito obrigada por tudo! Sem a "senhora", com certeza eu não teria chegado ao fim deste trabalho.

Agradeço também aos meus pais por sempre me apoiarem e permitirem que hoje eu estivesse aqui. A minha irmã, que me ajudou durante toda a construção desta dissertação lendo todos os capítulos concluídos. E a minha avó, que apesar de ter nos deixado, sempre se fez presente em meus pensamentos.

Agradeço aos meus amigos, Lairson e Marcelo, por terem me ajudado com as revisões, discussões e orientações durante toda a construção deste trabalho. Além disso, agradeço pela paciência e disponibilidade todas as vezes em que fui a procura de vocês, seja para tirar dúvidas ou, até mesmo, para desabafar sobre a ansiedade de chegar ao fim desse ciclo. Vocês são demais!!!

Agradeço aos meus amigos, Glória, Wilker, Rayelle e Helton que, além de contribuírem diretamente para a realização desta dissertação, sempre se mostraram presentes e dispostos a ajudar. Muito obrigada pessoal!

Agradeço também a toda a equipe do grupo de pesquisa aLADIN, pelas reuniões, discussões e, não poderia esquecer, as confraternizações. Vocês são incríveis!

Agradeço as minhas amigas, Tássia e Magda, por compreenderem a minha ausência durante esses últimos meses ao qual estive totalmente dedicada a este trabalho. Sei que para conclusão de algumas etapas, muitas vezes, precisamos abdicar de outras. E, durante o mês de Janeiro, deixar de ir visitar minha afilhada (Larinha), que estava em Limoeiro-PE para passar as férias, foi uma abdicação muito difícil de ser realizada.

Agradeço aos meus amigos, Gabriel e Matheus, por, durante esses últimos dois meses, segurarem as pontas dos projetos que desenvolvemos juntos para que eu pudesse estar exclusivamente dedicada ao mestrado. Não tenho palavras para agradecer meninos, vocês são demais!!!

Agradeço aos meus colegas de trabalho do Núcleo Estadual de Telessaúde da SES-PE pela compreensão durante os últimos meses. Pois, sempre que precisei me ausentar para a escrita desta dissertação, tive o apoio e a compreensão de todos. Muito obrigada pessoal! Em especial, agradeço aos meus chefes, Carlos e Dulce, pela flexibilidade que me foi dada durante esses dias mais complicados.

Por último, e não menos importante, gostaria de agradecer a todos que de alguma forma contribuíram para que esta dissertação fosse realizada. Muito obrigada!!

"Não importa o que aconteça, continue a nadar." (WALTERS, 2013)

RESUMO

A Web é um ambiente capaz de oferecer um grande volume de informações dos mais diversos domínios e, devido ao seu crescimento exponencial, transformou-se na principal plataforma para compartilhamento e troca de dados. Nesse cenário, dois papéis merecem destaque: os provedores e os consumidores de dados. Em termos gerais, os provedores são os usuários que visam a publicação e o compartilhamento de dados, enquanto os consumidores desejam fazer uso desses dados agregando ainda mais valor aos mesmos. Com o crescimento do compartilhamento de dados na Web, várias iniciativas foram propostas com o intuito de facilitar tanto a publicação quanto o consumo de dados, como metodologias, *guidelines* e boas práticas. Entretanto, não foram encontradas iniciativas que definam de forma mais precisa o que são os dados na Web e as fases percorridas pelos dados ao longo de sua trajetória na Web. Desse modo, uma alternativa promissora para preencher essa lacuna, consiste em adotar uma abordagem de ciclo de vida. Neste contexto, esta dissertação de mestrado propõe um Modelo de Ciclo de Vida de Dados na Web (Data on the Web Lifecycle Model - DWLM). O modelo proposto tem como principais objetivos prover um entendimento comum do processo de publicação e consumo de dados na Web e especificar claramente os papéis, fases e atividades envolvidas nesse processo.

Palavras-chaves: Dados na Web. Publicação de Dados. Consumo de dados. Ciclo de Vida.

ABSTRACT

The Web is an environment capable of offering a large amount of information about diverse domains and, due to its exponential growth, became the main platform for sharing and exchanging data. In this scenario, two roles deserve to be highlighted: the providers and consumers of data. In general terms, providers are users who aim at publication and data sharing, whereas the consumers want to make use and aggregate more value to the data. With the growth of data sharing on the Web, initiatives have been launched to benefit both publication and data consumption, such as methodologies, guidelines and best practices. However, it was not found initiatives that define more precisely what are data on the Web and the phases visited by these data. Thus, a promising alternative to discard this gap is to adopt a lifecycle approach. In this context, this master's dissertation propound a Data on the Web Lifecycle Model - DWLM. The main objective of the model is to provide a common understanding of the process of publication and consumption of data on the Web and specify the roles, phases and activities involved.

Keywords: Data on the Web. Data Publication. Data Consumption. Lifecycle.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

CKAN	Comprehensive Knowledge Archive Network
CSV	Comma Separated Value
DCAT	Data Catalog Vocabulary
ETL	Extract Transform Load
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
ODS	Open Document
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
URIs	Uniform Resource Identifiers
VPN	Virtual Private Network
WWW	World Wide Web
XLS	Microsoft Excel file format
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

Este Capítulo fornece uma visão geral desta pesquisa e apresenta o contexto no qual este trabalho está inserido. A Seção 1.1 apresenta uma breve motivação para o desenvolvimento deste trabalho. A caracterização do problema é descrita na Seção 1.2 e os objetivos e contribuições são apresentados na Seção 1.3. A Seção 1.4 discorre sobre a método de pesquisa usado e, por fim, a Seção 1.5 apresenta a estrutura desta dissertação.

1.1 MOTIVAÇÃO

A World Wide Web (WWW), ou apenas Web, proporciona um ambiente aberto e compartilhado para a publicação de dados dos mais variados domínios. Para Bouquet e Stoermer (2012), ela se tornou uma ferramenta muito importante para colaboração e uma fonte de inovação sem precedentes. Nesse cenário, muitos movimentos e técnicas surgiram para incentivar a publicação e o compartilhamento dos dados na Web, como: os movimentos de Dados Abertos, Web das Coisas, Dados Conectados, e-business, dentre outros. Na literatura, podemos encontrar uma série de trabalhos voltados para esses contextos, como o trabalho de Zhao (2010) que aborda a publicação de informações da área médica chinesa utilizando a metodologia de dados conectados, o de Fernández, Martínez-Prieto e Gutiérrez (2011) apresenta uma proposta para publicação de dados do censo da Espanha, e também o de Andersen et al. (2014) que descreve a publicação de dados relacionados à agricultura dinamarquesa considerando os princípios dos dados conectados. Com base nesses trabalho, podemos perceber a crescente necessidade de publicação de dados na Web em todo o mundo, bem como a diversidade de domínios existentes.

Nesse contexto de compartilhamento de dados, é importante ressaltar que as atividades de publicação e consumo podem ser executadas por um conjunto de atores. Esses atores, por sua vez, poderão assumir dois papéis fundamentais: publicador e/ou consumidor de dados. O publicador será responsável por compartilhar/disponibilizar os dados na Web, enquanto o consumidor fará algum uso desses dados. Porém, é importante deixar claro que um mesmo ator poderá exercer os dois papeis, tanto o de consumidor, quanto o de publicador (LÓSCIO; OLIVEIRA; BITTENCOURT, 2015). Por exemplo, vamos imaginar que Beatriz é uma bióloga e publicou seus dados referentes a experimentos realizados em algumas plantas. Rodrigo, estudante de Ciências Biológicas, deseja usar os dados publicados por Beatriz para uma análise que será apresentada em uma disciplina do curso. Logo, nesse cenário identificamos os dois papéis, Beatriz como publicadora e Rodrigo como consumidor. Entretanto, após finalizado o curso de Ciências Biológicas, Rodrigo decide publicar os resultados dos experimentos do seu trabalho de conclusão do curso. E, nessa mesma época, Beatriz estava realizando uma pesquisa com o mesmo tema de Rodrigo e,

enquanto realizava buscas na Internet se depara com o conjunto de dados publicado por Rodrigo. Assim, ela resolve usar o conjunto de Rodrigo para obter alguns resultados e complementar sua pesquisa. Nesse segundo cenário, podemos identificar que os papéis se inverteram, Beatriz assumiu o papel de consumidor enquanto Rodrigo assumiu o papel de publicador. Dessa forma, fica claro que uma mesma pessoa poderá sim desempenhar os dois papéis.

O crescimento no interesse da publicação de dados na Web tem tornado seus benefícios mais evidentes. Em uma pesquisa feita por Santos et al. (2018) foram identificados cerca de 28 benefícios relacionados a publicação e consumo de dados na Web. Dentre eles, os mais citados foram: melhor reutilização de dados, descoberta fácil, benefícios econômicos, interoperabilidade e benefícios sociais.

Contudo, para que esses benefícios possam ser alcançados ainda há vários problemas que precisam ser solucionados. Apesar dos dados serem um bem valioso, muitas vezes, eles acabam sendo publicados de forma equivocada e em formatos que não são aptos a serem processados por máquina, o que pode ocasionar problemas de interoperabilidade e reúso. Além disso, a falta de metadados e a incompletude nos dados afetam diretamente na sua qualidade (SANTOS et al., 2018). Em paralelo, alguns consumidores não qualificados podem apresentar dificuldades na hora de manipular os dados por desconhecimento de características importantes do conjunto de dados.

Com a finalidade de minimizar alguns desses desafios da publicação e consumo de dados na Web, várias iniciativas foram propostas, desde metodologias (*e.g.* (NECASKÝ et al., 2013)), *guidelines* (*e.g.* (RADULOVIC et al., 2015)) e boas práticas (*e.g.* (LÓSCIO; BURLE; CALEGARI, 2017)) para publicação de dados a Sistemas para Gerenciamento de Dados na Web (*e.g.* (OLIVEIRA et al., 2018)). Entretanto, não há nenhuma iniciativa que mostre o que constitui dados na Web, as fases que ele percorre ao longo de sua trajetória e suas características. Nessa linha de pensamento do que são dados e o que eles constituem, Möller (2013) os define como "*uma coisa viva que se move ao longo de vários estágios, como criação, publicação, uso ou término*".

Uma alternativa promissora para descrever esse comportamento dos dados, bem como os papéis envolvidos e suas fases seria uma abordagem de ciclo de vida. De acordo com Higgins (2008) uma abordagem de ciclo de vida garante que todas as etapas necessárias sejam identificadas e planejadas, e que as ações pertinentes sejam implementadas na sequência correta. Além disso, Möller (2013) disse não ter conhecimento de nenhuma definição de ciclo de vida específica para dados, entretanto, uma definição genérica do termo no *Oxford English Dictionary* seria "*ciclo de vida, n. [...] Uso prolongado: um curso ou evolução desde o começo, através do desenvolvimento e produtividade, até a decadência ou o fim.*"

Dessa forma, alguns modelos de ciclo de vida foram propostos para domínios como: pesquisa (HUMPHREY, 2006), software (ISO/IEC/IEEE..., 2017), curadoria de dados (HIG-

GINS, 2008), dentre outros. No domínio de dados, Möller (2013) propôs um Modelo Abstrato de Ciclo de Vida de Dados (*Abstract Data Lifecycle Model* - ADLM). O ADLM foi desenvolvido para servir como um modelo de ciclo de vida genérico que pode ser usado como um meio de classificar, comparar e relacionar outros modelos de ciclo de vida de dados, bem como fornecer a base para elaborar novos modelos (MöLLER, 2013). Porém, no ADLM identificamos fases que, ao nosso ver, para o contexto de dados na Web, não seriam fases, mas sim uma atividade que iria compor uma fase (*e.g.* o feedback, para nós não é uma fase do ciclo de vida mas sim, uma atividade da fase de consumo.) Ademais, foi sentida a necessidade de estipular as atividades que devem ser executadas em cada fase do ciclo de vida e o ADLM também não nos oferece isso. Em consequência disso, buscamos desenvolver um modelo de ciclo de vida para dados na Web, o qual fosse genérico ao ponto que os diferentes domínios de dados na Web pudessem utilizá-lo.

Diante desse contexto, esta dissertação propõe um Modelo de Ciclo de Vida de Dados na Web (*Data on the Web Lifecycle Model* - DWLM). A partir do uso do modelo proposto, esperamos especificar claramente os papéis e responsabilidades que estarão envolvidos em todas as fases do modelo de ciclo de vida, bem como auxiliar na aplicação das Melhores Práticas de Publicação de Dados (*Data on the Web Best Practices* - DWBP). Além disso, o DWLM almeja estabelecer um canal de comunicação entre publicadores e consumidores, a fim de prover um entendimento comum do processo de publicação e consumo. Convém pontuar, ainda, que em cada fase definida no modelo serão descritas as atividades que devem ser executadas, as boas práticas que estarão envolvidas e suas entradas e saídas (*outputs*).

1.2 CARACTERIZAÇÃO DO PROBLEMA

Sayao (1999) descreve que *"um modelo é uma criação cultural, um "mentefato", destinado a representar uma realidade, ou alguns dos seus aspectos, a fim de torná-los descritíveis qualitativa e quantitativamente e, algumas vezes, observáveis"*. O Dicionário da Língua Portuguesa da Porto Editora, diz que um modelo *"é um protótipo ou exemplo que se pretende reproduzir ou imitar."* Para Gouveia (1999), a utilidade de um modelo é dita de acordo com algumas variáveis, ele relata que quanto maior o seu uso prático (valor de uso), possibilidade de previsão (valor preditivo) e similaridade com o fenômeno proposto (valor de face) melhor ele será considerado. Em resumo, podemos dizer que um modelo é bom quando funciona para os fins propostos (GOUVEIA, 1999).

Existe um consenso geral quanto ao papel crucial que os modelos podem desempenhar em diversas áreas. Apostel (1960 apud SAYAO, 1999) diz que os modelos são necessários por constituírem uma ponte entre os níveis da observação e o teórico, e tratam da simplificação, redução, concretização, experimentação, ação, extensão, globalização, explicação e formação da teoria. Hoje, diversas áreas utilizam modelos para abstraírem alguma reali-

dade, alguns exemplos de modelos utilizados são: Modelo atômico[1], Modelo matemático[2], Modelo econômico[3].

Em particular, os Modelos de Ciclo de Vida são utilizados em várias áreas com objetivo de acompanhar e descrever as etapas identificadas na evolução de "coisas" em cada domínio (ZHONGJIE; XIAOFEI, 2009). Uma área muito conhecida pelo uso de modelos de ciclo de vida para apoiar o desenvolvimento dos seus produtos é a Engenharia de Software. Nela, existem vários modelos que são amplamente utilizados, como o Modelo em Cascata[4], Modelo em Espiral[5] e o Modelo em V[6]. Esses modelos podem variar de acordo com a complexidade do cenário, sistema ou contexto no qual eles estão inseridos. Além disso, de acordo com Meschankina (2018) os modelos de ciclo de vida de desenvolvimento de software ajudam a equipe a ter uma ideia clara do que precisa ser feito, destacam os problemas e definem cada fase do processo de desenvolvimento antes que uma única linha de código seja escrita.

No cenário dos dados na Web, existem algumas metodologias voltadas para publicação de dados (e.g. (NECASKÝ et al., 2013), (ZENGENENE; CASAROSA; MEGHINI, 2013), (RAO; NAYAK, 2017)). Contudo, essas soluções ainda não estão consolidadas, vários trabalhos permanecem apenas no contexto da publicação e não fornecem detalhes do que acontece com os dados a partir disso. Assim, a escassez de metodologias voltadas ao acompanhamento do processo evolutivo dos dados na Web apresenta-se como uma oportunidade para pesquisa.

Dessa forma, um modelo de ciclo de vida de dados na Web é importante para garantir esse acompanhamento do dado na Web. Além disso, ele também servirá como um guia para auxiliar a publicação e consumo de conjuntos de dados, bem como o uso das melhores práticas.

Por fim, este trabalho teve seu desenvolvimento guiado pela seguinte questão de pesquisa: *"Como a publicação e o consumo de dados na Web podem ser estruturados e sistematizados por meio de um ciclo de vida?"*

1.3 OBJETIVOS

O principal objetivo dessa dissertação é propor um Modelo de Ciclo de Vida de Dados na Web. Esse modelo, leva em consideração alguns aspectos do ADLM proposto por Möller (2013) e as Melhores Práticas de Publicação de Dados proposta pela W3C, a fim de estabelecer um entendimento comum do processo de publicação e consumo de dados na Web.

[1] https://pt.wikipedia.org/wiki/Modelo_atômico
[2] https://pt.wikipedia.org/wiki/Modelo_matemático
[3] https://pt.wikipedia.org/wiki/Modelo_econômico
[4] https://en.wikipedia.org/wiki/Waterfall_model
[5] https://en.wikipedia.org/wiki/Spiral_model
[6] https://en.wikipedia.org/wiki/V-Model_(software_development)

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Realização de um mapeamento sistemático da literatura na área de Dados na Web;
- Levantamento dos principais Modelos de Ciclo de Vida de Dados existentes;
- Definição de um Modelo de Ciclo de Vida de Dados na Web;
- Aplicação do modelo proposto em um cenário da UFPE;
- Avaliação do modelo proposto por meio de um grupo focal.

1.4 MÉTODO DE PESQUISA

Normalmente, vários métodos podem ser aplicados aos problemas de pesquisa, e em muitos casos, é necessário uma combinação de métodos para um entendimento completo do problema. Para Easterbrook et al. (2008), um método de pesquisa é um conjunto de princípios organizacionais em torno dos quais dados empíricos são coletados e analisados. Marconi e Lakatos (2010) pontua que o método de pesquisa deve estar diretamente relacionado ao problema a ser estudado. Ou seja, uma pesquisa deve ser rigorosamente analisada mesmo antes de sua execução real.

A realização desta dissertação, foi baseada em um paradigma filosófico pragmático. De acordo com Easterbrook et al. (2008), o pragmatismo valoriza o conhecimento prático sobre o conhecimento abstrato e usa todos os métodos apropriados para obtê-lo. Além disso, Easterbrook et al. (2008) e Creswell (2010) descrevem que essa postura filosófica é caracterizada pela aceitação de diferentes conceitos para apoiar a pesquisa. Desse modo, os pesquisadores que optam por esse paradigma filosófico acreditam que o problema é mais importante e usam todos os meios disponíveis para conseguir entendê-lo. Easterbrook et al. (2008) complementa afirmando que os pragmatistas preferem fortemente a pesquisa de métodos mistos, onde vários métodos são usados para esclarecer a questão em estudo.

Em resumo, essa pesquisa foi realizada em quatro etapas, como mostra a Figura 1. Inicialmente, foi realizado um mapeamento sistemático da literatura com o intuito de obter uma visão geral do cenário de publicação e consumo de dados. Segundo Keele et al. (2007), um mapeamento sistemático é uma metodologia orientada por protocolos que tem o objetivo de revisar e sintetizar trabalhos de uma área de pesquisa. Essa mapeamento teve como objetivo oferecer um *snapshot* das pesquisas sobre a publicação e consumo de dados na Web (i) identificando e analisando como os dados têm sido publicados e consumidos na Web, (ii) descobrindo os benefícios e limitações da publicação e do consumo de dados na Web, (iii) analisando a evolução da pesquisa sobre publicação e consumo de dados na Web e (iv) classificando os estudos em categorias relacionadas à sua contribuição. Esta etapa foi de suma importância, pois formou um embasamento teórico inicial para a continuidade da pesquisa e, a partir das publicações analisadas, foi possível identificar que, até então,

não há uma especificação clara de um modelo de ciclo de vida para dados na Web. Os resultados desse mapeamento foram publicados em (SANTOS et al., 2018).

Figura 1 – Etapas da Metodologia de Pesquisa. Fonte: Autor

A próxima fase foi caracterizada por uma revisão *ad-hoc* da literatura para encontrar trabalhos relacionados ao nosso que pudessem nos fornecer um embasamento teórico. Uma revisão *ad-hoc* é uma busca informal para encontrar insumos de um determinado assunto ou área temática. Essa revisão resultou no Capítulo 3.2.

Após realização das pesquisas, a terceira fase teve como objetivo construir e melhorar continuamente o modelo proposto nesta dissertação. Essa etapa passou por vários ciclos de verificação e refinamento até ter como resultado uma versão inicial do Modelo de Ciclo de Vida de Dados na Web. Por último, uma avaliação utilizando o método de grupo focal foi realizada. Um grupo focal pode ser definido como um grupo de indivíduos reunidos para avaliar produtos, sistemas, conceitos ou para evidenciar problemas (KONTIO; LEHTOLA; BRAGGE, 2004). Com base nos resultados obtidos nesta avaliação, o modelo foi refinado a fim de atender as melhorias propostas pelos participantes e sua versão final é apresentada no Capítulo 4.

1.5 ESTRUTURA DA DISSERTAÇÃO

Os próximos capítulos estão organizados da seguinte forma. No Capítulo 2, apresentamos o que constituí os Dados na Web, enquanto que no Capítulo 3 descrevemos os modelos de ciclo de vida no domínio de desenvolvimento de software e dados. Já no Capítulo 4, apresentamos o modelo de ciclo de vida de dados na Web proposto nesta dissertação. No Capítulo 5 aplicamos o modelo proposto em um cenário da UFPE e no Capítulo 6 apresentamos a técnica de grupo focal utilizada para avaliação do modelo. Por fim, no Capítulo 7 é feito uma pequena discussão sobre o trabalho realizado e sugestões para os trabalhos futuros.

2 DADOS NA WEB

Neste Capítulo, iremos apresentar os conceitos essenciais para o entendimento desta dissertação. Inicialmente, na Seção 2.1, apresentamos uma visão geral do contexto de Dados na Web. Na Seção 2.2 descrevemos os Dados Abertos, enquanto que na Seção 2.3 apresentamos os Dados Conectados. Em seguida, na Seção 2.4 detalhamos as Boas Práticas para Publicação de Dados na Web. Na Seção 2.5 apresentamos os conceitos acerca de Ecossistemas de Dados na Web. E, por fim, na 2.6 apresentamos as Considerações Finais do Capítulo.

2.1 VISÃO GERAL

A publicação e compartilhamento de dados na Web fornecem inúmeros benefícios, um exemplo disso são os mais variados tipos de aplicações que estão surgindo proveniente desses dados. Por exemplo, o observatório de oncologia[1] que faz uso dos dados compartilhados na Web para realizar estudos como o aumento das mortes por câncer, detecção precoce do Papanicolau e da Colonoscopia em Salvador, infecções virais, dentre outros.

Nesse contexto, nos últimos anos, a Web tem se consolidado cada vez mais uma grande plataforma de compartilhamento e consumo de dados. De acordo com Lóscio, Guimarães e Calegari (2016), dados na Web pode ser entendido como o termo mais geral que pode ser usado para denotar dados publicados de acordo com a base arquitetônica da Web[2]. Além disso, os autores complementam definindo que os dados na Web podem ser classificados como dados abertos[3], dados conectados e dados conectados abertos[4], conforme mostra a Figura 2. Assim, os Dados na Web são todos os dados que estão disponíveis, sejam eles publicados em formato aberto ou fechado (LÓSCIO; GUIMARÃES; CALEGARI, 2016).

Segundo Lóscio, Burle e Calegari (2017), os dados devem ser publicados em diferentes distribuições, que são a forma física específica de um conjunto de dados, para facilitar o compartilhamento de dados e contribuir com atividades de pré-processamento dos dados. Além disso, como os consumidores e produtores podem ser desconhecidos entre si, é necessário fornecer informações sobre os conjuntos de dados e distribuições. Tais informações fomentarão a confiabilidade e reutilização dos dados, ou seja, é necessário fornecer a maior quantidade de informações possíveis por meio de metadados, como metadados descritivos, de acesso, qualidade, proveniência e licença (LÓSCIO; BURLE; CALEGARI, 2017). Somado a isso, é importante seguir os princípios da arquitetura Web e usar vocabulários e padrões de dados. Lóscio, Burle e Calegari (2017), enfatiza a importância de cada recurso, que

[1] https://observatoriodeoncologia.com.br/
[2] https://www.w3.org/TR/webarch/
[3] http://ceweb.br/guias/dados-abertos/en/
[4] https://www.w3.org/DesignIssues/LinkedData.html

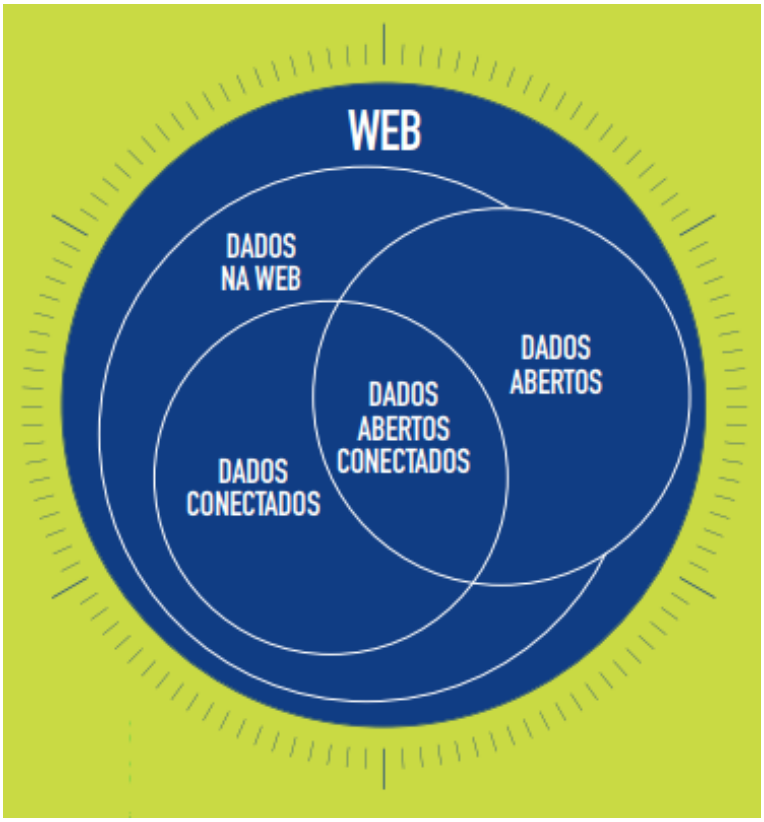

Figura 2 – Dados na Web x Dados Abertos x Dados Conectados. Fonte: (LÓSCIO et al., 2018)

pode ser um conjunto de dados inteiro ou um item específico de um determinado conjunto de dados, ser publicado com uma Uniform Resource Identifiers (URIs) estável, para que possam ser referenciados e conectados a outros recursos. Além disso, é importante usar vocabulários e padrões de dados, como o Data Catalog Vocabulary (DCAT) [5], para promover a interoperabilidade entre conjuntos. A Figura 3 ilustra o contexto de publicação dos dados na Web.

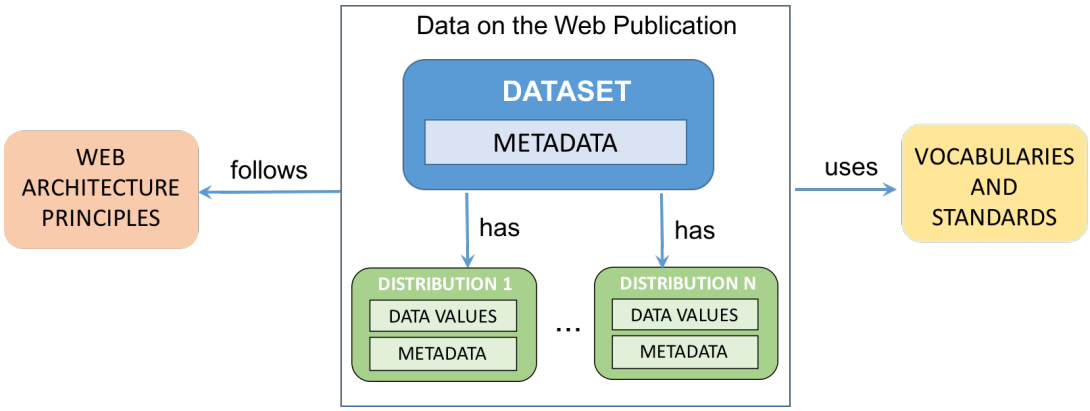

Figura 3 – Contexto de Publicação dos Dados na Web. Fonte: (LÓSCIO; BURLE; CALEGARI, 2017)

[5] https://www.w3.org/TR/vocab-dcat/

Desse modo, a Web oferece funcionalidades eficientes e menos complexas para a publicação de dados (LÓSCIO; OLIVEIRA; BITTENCOURT, 2015) e a quantidade de dados disponível está em crescente evolução. No entanto, existem particularidades que são inerentes ao contexto da Web e serão discutidas nas próximas seções.

2.2 DADOS ABERTOS

O conceito de Dados Abertos corresponde à ideia de que certos dados devem estar disponíveis para que todos usem e publiquem, sem restrições de direitos autorais e patentes. Segundo Dietrich et al. (2009), dados abertos são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa, estando sujeitos a, no máximo, exigências de atribuição a autoria e compartilhamento pelas mesmas licenças em que as informações foram apresentadas. Os dados, quando abertos, promovem transparência, inovação, democracia e podem ajudar no aperfeiçoamento de várias questões sociais. Além disso, quando usados em conjunto com outros dados, é possível ocorrer descobertas de novas informações que estavam implícitas nos dados isolados. Essas novas descobertas auxiliam entidades públicas e privadas na criação de novos produtos e serviços ou no melhoramento dos existentes.

Dietrich et al. (2009) concluiu que um dado é considerado aberto quando apresenta as seguintes características:

- *Disponibilidade e Acesso:* os dados devem estar disponíveis de uma forma conveniente e modificável. Ademais, deve haver a opção de download.
- *Reutilização e redistribuição:* os dados devem ser fornecidos sob termos que permitam a reutilização e a redistribuição, assim como, a junção com outros conjuntos de dados.
- *Participação Universal:* todos devem poder usar, reutilizar e redistribuir, sem discriminação contra campos de atuação, pessoas ou grupos.

Berners-Lee (2006) propôs cinco níveis de classificação para a publicação de dados na Web na qual, claramente, é observada uma preocupação não só com a distribuição dos dados na Web, mas também com a conexão entre eles. Esses níveis ficaram conhecidos como “Esquema de 5 estrelas”, que classifica por meio de estrelas o grau de abertura dos dados e um nível vai complementando o outro até chegar aos dados abertos conectados. Os níveis de classificação são representados na Figura 4.

De acordo com o esquema proposto, um dado disponível na Web, independente de formato, sob uma licença aberta, é avaliado com 1 estrela. Se for disponibilizado de forma estruturada, como um Excel em vez de uma imagem escaneada, recebe como avaliação 2 estrelas. Quando são utilizados formatos não-proprietários como o CSV, é avaliado como

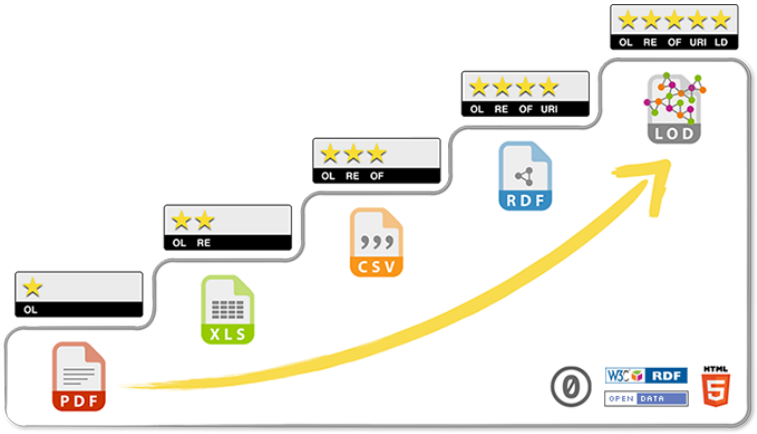

Figura 4 – Esquema de 5 estrelas para Dados Abertos. Fonte: http://5stardata.info/en/

3 estrelas. A partir do momento que os dados passam a utilizar identificações únicas - URIs, são avaliados como 4 estrela. E, por último, os dados que estão conectados a dados de outras fontes fornecendo um contexto, recebe a avaliação de 5 estrelas.

Seguindo esse movimento de dados abertos, vários governos de diversos países estão utilizando a Web como meio para a publicação de Dados Abertos Governamentais. Esses dados, geralmente são compartilhados em Portais de Dados Abertos que oferecem uma interface amigável para catalogação e acesso (LÓSCIO et al., 2018). Esses portais cresceram muito nos últimos anos, hoje diversos países tem portais de dados abertos já consolidados, como EUA[6], França[7], Brasil[8], Chile[9], Holanda[10], dentre outros. No Brasil, o Portal de Dados Abertos foi lançado em 2012 e hoje conta com mais de 6 mil conjuntos publicados.

Com o objetivo de estabelecer alguns requisitos necessários para dados abertos governamentais o grupo de trabalho, *Open Government Working Group*, estipulou 8 princípios para os dados governamentais (TAUBERER; LARRY, 2007), são eles:

1. *Completo*: todos os dados públicos devem ser disponibilizados. Um dado público é o dado sem estar sujeitos a limitações de privacidade, segurança ou privilégio.

2. *Primário*: os dados devem ser coletados diretamente na fonte, sem agregações ou modificações.

3. *Atuais*: os dados devem ser prontamente disponibilizados para preservar o seu valor.

4. *Acessível*: os dados devem estar disponíveis para o maior número de usuários e propósitos possíveis.

[6] http://data.gov
[7] http://data.gouv.fr
[8] http://dados.gov.br
[9] http://datos.gob.cl
[10] http://dataoverheid.nl

5. *Processáveis por máquina*: os dados devem ser um minimamente estruturados para permitir o processamento por máquinas.

6. *Não-discriminatórios*: os dados devem estar disponíveis para todas as pessoas, sem exigência de registro.

7. *Não-proprietários*: os dados devem estar em formatos abertos, ou seja formatos onde nenhuma entidade possua controle exclusivo.

8. *Livres de Licença*: os dados não devem estar sujeitos a nenhum regulamento de direitos autorais, patente, marca comercial ou segredo industrial. Apenas são permitidas restrições de segurança e privilégio.

Para Lóscio et al. (2018) os dados abertos governamentais abordam diversos assuntos desde despesas orçamentarias a dados sobre censo escolar, reclamações de consumidores, pontos turísticos, dentre outros. Além disso, esses dados são fortemente usados para desenvolvimento de aplicativos que, posteriormente, são utilizados pela população e/ou governo.

2.3 DADOS CONECTADOS

O conceito de Dados Conectados (*Linked Data*) origina-se da ideia de utilizar a Web para conectar os dados publicados. Dessa forma, assegurando que os conjuntos de dados não sejam apenas ilhas de dados isolados, mas que eles se comuniquem e possibilitem uma integração. Bizer, Heath e Berners-Lee (2009) diz que, tecnicamente, Dados Conectados referem-se a dados publicados na Web em formatos legíveis por máquina, com significados explicitamente definidos, e que permitam vínculos a conjuntos de dados externos.

Além disso, Bizer, Heath e Berners-Lee (2009) acrescenta que o conceito de Dados Conectados também pode ser definido como um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na Web. Para realizar a conexão desses dados é necessário que eles sejam publicados em formatos de dados estruturados, como o *Resource Description Framework* (RDF). O RDF é uma estrutura de representação de informações baseado em triplas (KLYNE; CARROLL, 2004), onde para cada tripla é informado o sujeito, predicado e objeto de um recurso a ser descrito. Para a identificação de cada tripla são utilizados *Universal Resource Identifier* (URIs)[11]. Os URIs, além de serem utilizados para a identificação, também são usados para realizar relacionamentos entre recursos de um RDF.

Para Bizer, Heath e Berners-Lee (2009) os Dados Conectados baseiam-se em duas tecnologias que são fundamentais para a Web: os URIs e o *HyperText Transfer Protocol* (HTTP)[12]. Os URIs, como falado anteriormente, são utilizados para identificação e

[11] https://tools.ietf.org/html/rfc3986
[12] https://www.w3.org/Protocols/rfc2616/rfc2616.html

vínculos de recursos RDF, mas ele também se destaca por ser semelhante aos endereços URLs (*Uniform Resource Locators*) que conhecemos. No entanto, eles fornecem um meio mais genérico para identificar qualquer entidade existente no mundo de forma que, essas entidades possam ser consultadas simplesmente desreferenciando o URI sobre o protocolo HTTP. Desta forma, o protocolo HTTP fornece um mecanismo simples, mas universal, para recuperar recursos que podem ser serializados como um fluxo de *byte* (como uma fotografia de um cachorro) ou recuperar descrições de entidades que não podem ser enviadas pela rede desta maneira (como o próprio cão).

Berners-Lee (2006) propôs um conjunto de regras para publicar dados conectados na Web, que ficaram conhecidas como *Linked Data principles*, essa regras são:

i) Fornecer URIs como nome de cada recurso a ser representado;

ii) Usar HTTP URIs para que os usuários possam procurar esses recursos;

iii) Quando alguém acessar uma URI, fornecer informações úteis, usando os padrões (RDF, SPARQL Protocol and RDF Query Language (SPARQL));

iv) Incluir links para outros URIs, para que os usuários possam descobrir novos recursos.

Outro conceito muito utilizado é o de Dados Abertos Conectados (*Linked Open Data* - LOD), esse termo se refere a Dados Conectados que possuem conteúdo aberto. Essa definição foi proposta por Tim Berners-Lee como um expansão aos Dados Conectados, onde nesse cenário por se tratar de dados abertos eles possuem licença livre. Atualmente existe um grande volume de dados abertos conectados disponíveis na Web, um exemplo disso é o projeto LOD[13]. Ele foi iniciado em 2007 por Chris Bizer e Richard Cyganiak com o apoio da W3C. O projeto tem como objetivo publicar conjuntos de dados realizando ligações entre diversas bases de dados que temos disponíveis na Web. Ele conta com uma nuvem de dados que contém fontes de variadas áreas de conhecimento como: publicações científicas, dados governamentais, dados geográficos, dentre outros. Hoje, ele é tido como o maior exemplo de utilização de Dados Conectados, sua última atualização foi realizada em junho de 2018 e naquele momento ele continha 1.234 conjuntos de dados com 16.136 links.

2.4 BOAS PRÁTICAS PARA PUBLICAÇÃO DE DADOS NA WEB

Com o crescente interesse de publicar e consumir dados na Web muitas discussões surgiram a respeito da abordagem mais apropriada para a publicação de dados. Além disso, segundo Lóscio, Oliveira e Bittencourt (2015), a heterogeneidade dos dados e a falta de padrões para descrição e acesso aos conjuntos de dados tornam o processo de publicação,

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

compartilhamento e consumo uma tarefa complexa. Desse modo, em busca de alternativas que possibilitem um entendimento comum entre os atores desse contexto, a W3C criou um grupo de trabalho denominado *Data on the Web Best Practices*. Esse grupo teve como objetivo propor uma recomendação que servisse como um guia para a publicação e o consumo dos dados na Web.

Quadro 1 – Desafios da publicação de dados na Web.

Desafio	**Descrição**
Metadados	Permitir que os seres humanos entendam os metadados, interpretando a natureza e a estrutura dos dados, e que as máquinas também possam processá-los
Licença	Permitir que os seres humanos compreendam as informações da licença e que as máquinas possam detectar automaticamente
Proveniência	Permitir que os seres humanos conheçam a origem ou o histórico do conjunto de dados e que as máquinas possam processar automaticamente tais informações
Qualidade	Documentar a qualidade dos dados, para facilitar o processo de seleção dos conjuntos de dados e chances de reutilização
Versionamento	Permitir que versões dos dados sejam geradas e seja possível o acesso a cada versão
Identificação	Fornecer identificadores únicos para os conjuntos de dados e distribuições
Formato	Escolher formatos que permitam o uso e o reuso
Vocabulários	A fim de melhorar a interoperabilidade e manter terminologia comum entre os produtores e consumidores
Acesso	Permitir o fácil acesso aos dados usando a infraestrutura da Web tanto para seres humanos quanto para máquinas
Preservação	A fim de indicar corretamente se os dados foram removidos ou arquivados
Feedback	Receber feedback dos consumidores e assegurar que os dados atendam as suas necessidades
Enriquecimento	Enriquecer, melhorar ou refinar os dados brutos agregando valor
Republicação	Permitir que os dados utilizados possam ser republicados

Fonte: Derilinx, Lóscio e Archer (2015)

Para alcançar esse objetivo, o grupo de trabalho do W3C, selecionou um conjunto de casos de uso[14] que representam cenários de como os dados são publicados e consumidos na Web. Com esses casos de uso, foi possível identificar os principais desafios enfrentados por produtores e consumidores de dados (ver Quadro 1), assim como um conjunto de requisitos necessários para a publicação. A partir dos desafios e requisitos encontrados nesses casos de usos, foi desenvolvido o documento de Boas Práticas para Publicação de Dados na Web (*Data on the Web Best Practices* - DWBP).

[14] https://www.w3.org/TR/dwbp-ucr

De acordo com Lóscio, Guimarães e Calegari (2016), as Boas Práticas para Dados na Web foram desenvolvidas para oferecer orientação técnica para a publicação de dados na Web, contribuindo para melhorar a relação entre publicadores e consumidores de dados. Além disso, elas são independentes de domínio e aplicação, ou seja, é aplicável a todos os domínios, podendo ainda ser estendidas ou complementadas com outros documentos ou normas mais especializadas. Para cada desafio apresentado no Quadro 1 foram propostas uma ou mais práticas. No total, são 35 boas práticas que discursam sobre diferentes aspectos relacionados à publicação e consumo de dados, como acesso aos dados, identificadores, metadados, formatos, dentre outros. O Quadro 2 apresenta as boas práticas estipuladas para cada desafio encontrado.

Conforme apresentado por Lóscio, Burle e Calegari (2017), cada boa prática (BP) tem um resultado esperado com sua aplicação e possíveis formas de implantação da prática. Além disso, são descritas a motivação para o seu uso e quais testes podem ser realizados para verificar se a prática foi implementada de forma adequada. Ao final, ainda apresenta as evidências que comprovam a relevância da prática e os benefícios que serão alcançados com o seu uso.

Quadro 2 – Boas Práticas para publicação de dados na Web

Desafio	**Boas Práticas**
Metadados	BP1 - Fornecer metadados
	BP2 - Fornecer metadados descritivos
	BP3 - Fornecer metadados estruturais
Licença	BP4 - Fornecer informações de licenciamento de dados
Proveniência e Qualidade	BP5 - Fornecer informações sobre a proveniência dos dados
	BP6 - Fornecer informações sobre a qualidade dos dados
Versionamento	BP7 - Fornecer um indicador de versão
	BP8 - Fornecer histórico de versão
Identificação	BP9 - Utilizar URIs constantes como identificadores de conjuntos de dados
	BP10 - Utilizar URIs constantes como identificadores dentro dos conjuntos de dados
	BP11 - Designar URIs para versões e séries de conjuntos de dados
Formato	BP12 - Utilizar formatos de dados padronizados inteligíveis por máquinas
	BP13 - Utilizar representações de dados de localidade neutra
	BP14 - Fornecer dados em formatos múltiplos
Vocabulários	BP15 - Reutilizar vocabulários preferencialmente padronizados
	BP16 - Escolher o nível correto de formalização
Acesso	BP17 - Fornecer download em massa
	BP18 - Fornecer subconjuntos para conjuntos de dados extensos
	BP19 - Utilizar a negociação de conteúdo para disponibilizar dados em formatos múltiplos
	BP20 - Fornecer acesso em tempo real
	BP21 - Fornecer dados atualizados
	BP22 - Fornecer uma justificativa para dados não disponíveis
	BP23 - Disponibilizar dados por meio de uma API
	BP24 - Utilizar padrões da Web como base para as APIs
	BP25 - Fornecer a documentação completa para sua API
	BP26 - Evitar modificações que quebrem sua API
Preservação	BP27 - Preservar os identificadores
	BP28 - Avaliar a cobertura do conjunto de dados
Feedback	BP29 - Coletar feedback de consumidores de dados
	BP30 - Disponibilizar feedback
Enriquecimento	BP31 - Enriquecer dados por meio da geração de novos dados
	BP32 - Fornecer apresentações complementares
Republicação	BP33 - Fornecer feedback ao editor original
	BP34 - Seguir os termos de licenciamento
	BP35 - Citar a publicação original

Fonte: Lóscio, Burle e Calegari (2017)

2.5 ECOSSISTEMA DE DADOS NA WEB

Desde seu surgimento, a Web vem evoluindo para atender às novas demandas e passando por transformações frequentes, como o suporte a novas formas de serviços e comércio (NATH; DHAR; BASISHTHA, 2014). Em sua origem, a Web possibilitava a publicação, propagação e visibilidade de conteúdos como sites corporativos e institucionais. Mas, às transformações ao longo dos anos, tanto acompanhada pelo avanço da tecnologia, quanto por novas demandas, possibilitou uma participação efetiva dos usuários e hoje conta com uma grande e crescente quantidade de dados disponíveis. Nesse contexto, Isotani e Bittencourt (2015) aponta que muitos dados não estão estruturados para facilitar a compreensão por

aqueles que podem acessá-los e manipulá-los ou, ainda, não são facilmente descobertos. Segundo Gama e Lóscio (2014), uma das formas para possibilitar a descoberta de dados é a criação de ecossistemas de dados que estimule a interação entre todos os envolvidos na publicação e consumo de dados.

Assim, um Ecossistema de dados na Web pode ser definido como um conjunto de atores e artefatos envolvidos na produção, distribuição e consumo de dados por meio da Web (LÓSCIO; GUIMARÃES; CALEGARI, 2016). Nesse cenário, um ator pode ser tanto um usuário, como um sistema ou ambos. Além disso, pode atuar como um publicador de dados ou como um consumidor de dados. Enquanto o publicador tem o papel de entregar e produzir dados de vários tipos, o consumidor atua no consumo desses dados, seja processando, analisando ou extraindo informações.

Um ecossistema de dados na Web, portanto, deve fomentar a participação e interação desses atores nas suas diversas atividades. Os publicadores de dados são responsáveis por atividades, como a definição de licenças, escolha de formatos e plataformas para publicação dos dados. Além disso, eles podem prover uma ou mais interfaces de acesso para recuperação dos dados, como as APIs. Já os consumidores de dados têm o foco em consumir os dados publicados de acordo com as condições e recursos necessários para resolver um determinado problema ou alcançar um objetivo. Conforme dito anteriormente, a flexibilidade da Web permite que os dados sejam publicados e consumidos de maneira simples, sem a exigência de sistemas para controlar o acesso concorrente aos dados, transações ou manutenção de integridade dos dados. Em contrapartida, são necessárias preocupações adicionais, uma vez que os dados podem ser consumidos por grupos de usuários previamente desconhecidos e com diferentes requisitos (LÓSCIO; OLIVEIRA; BITTENCOURT, 2015).

Um exemplo de Ecossistema de dados Web são as iniciativas de Dados Abertos Governamentais (ATTARD et al., 2015). Nessas iniciativas, os governos atuam como publicadores de dados, disponibilizando seus dados em formatos legíveis por máquina (*machine-readable*) e sob condições abertas de licenciamento que permitem o uso e o compartilhamento dos dados para diversos consumidores. Empreendedores, desenvolvedores de aplicativos ou cidadãos agem como consumidores de dados, criando novos produtos e serviços de informação e visualizações, que permitem monitorar as atividades de seu governo local, como também criar ferramentas para facilitar a vida diária.

2.6 CONSIDERAÇÕES FINAIS

Neste Capítulo, foi apresentado uma visão geral do que constituem Dados na Web, descrevendo o contexto ao qual a publicação de dados está envolvida. Em seguida, apresentamos o conceito de Dados Abertos descrevendo suas características e movimentos derivados, como os dados abertos governamentais. Na Seção 2.3 transcorremos sobre o que constituem Dados Conectados e os seus princípios. Além disso, nela também foi apresentado

o conceito de Dados Abertos Conectados. Na sequencia, apresentamos as Boas Práticas para publicação de dados na Web. Nesta Seção, são descritos os desafios encontrados na publicação de dados na Web e o processo de construção do documento DWBP. Por fim, finalizamos o Capítulo descrevendo o ecossistema de dados na Web e os atores envolvidos na produção, distribuição e consumo dos dados.

3 CICLOS DE VIDA

Neste Capítulo, iremos apresentar os conceitos essenciais sobre Ciclos de Vida. Inicialmente, na Seção 3.1, apresentamos uma visão geral do que são Ciclos de Vida e Modelos de Ciclo de Vida. Na Seção 3.2 abordaremos alguns Modelos de Ciclo de Vida de Desenvolvimento de Software e na Seção 3.3 serão mostrados alguns exemplos de Modelos de Ciclo de Vida de Dados encontrados na literatura. Por fim, na Seção 3.4 é apresentado as Considerações Finais do Capítulo, bem como um comparativo entre os Modelos de Ciclo de Vida de Dados apresentados.

3.1 VISÃO GERAL

O conceito de ciclo de vida é utilizado em diversas áreas (*e.g.* (DOORBAR, 2005), (LESTER; PARNELL; CARRAHER, 2003)). Na Informática, a Engenharia de Software fez uso da ideia de ciclos de vida para determinar as fases que englobam um determinado projeto/produto, desde seu planejamento até sua entrega final ISO/IEC/IEEE... (2017). Para representação desses ciclos, foram propostos vários modelos (*e.g* (ROYCE, 1970), (BOEHM, 1988), (MARTIN, 1991)) que descrevem como o software será desenvolvido, lançado, aprimorado e finalizado. A escolha do modelo dependerá de diversas variáveis, como tempo, modelo de negócio, custo e equipe.

Além disso, para a área de dados também foram propostos alguns modelos (*e.g* (CHEN; CHEN; LIN, 2003), (MöLLER, 2013), (ALSHAMMARI; SIMPSON, 2017)) com o intuito de realizar um acompanhamento da evolução dos dados durante o seu ciclo. Diante disso, nas próximas seções serão abordados alguns modelos de ciclo de vida de desenvolvimento de software e alguns modelos de ciclo de vida de dados.

3.2 MODELOS DE CICLO DE VIDA DE SOFTWARE

A normativa da ISO/IEC/IEEE 12207 determinou que um Modelo de Ciclo de Vida de Software (SLCM) *"é uma estrutura que contém os processos, atividades e tarefas envolvidas no desenvolvimento, operação e manutenção de um produto de software, abrangendo a vida útil do sistema desde a definição de seus requisitos até o término de seu uso"* (ISO/IEC/IEEE..., 2017). Dessa forma, todas as atividades e produtos de trabalho necessários para desenvolver um sistema de software constituem um SLCM.

A ISO/IEC/IEEE... (2017) acrescenta que os ciclos de vida variam conforme a natureza, finalidade, uso e circunstâncias prevalecentes do sistema de software. A ISO/IEC TS 24748-1, complementa que os estágios típicos do ciclo de vida do sistema incluem conceito, desenvolvimento, produção, utilização, suporte e aposentadoria. Dessa forma, usar

estágios simultâneos e em ordens diferentes podem levar a diferentes formas de ciclos de vida com características distintas.

Existem diversos modelos de ciclo de vida de desenvolvimento de software na literatura, cada modelo tem suas vantagens e desvantagens, não existe uma regra que defina qual modelo é melhor para todos os tipos de projetos. Os mesmos tendem a variar de acordo com o contexto do projeto a ser desenvolvido. Modelos como o Cascata, Espiral e Incremental são os mais conhecidos e utilizados no processo de desenvolvimento de software e serão descritos nas próximas seções.

3.2.1 Modelo em Cascata

Esse modelo é o mais antigo dentre os modelos de ciclo de vida de software, ele foi proposto por Royce (1970). O modelo é composto por um conjunto de fases e segue um fluxo que, para iniciar um nova fase você precisa completar a anterior. Nele, a documentação é produzida em cada fase. Isso torna o processo visível para que os gerentes possam monitorar o progresso em relação ao plano de desenvolvimento.

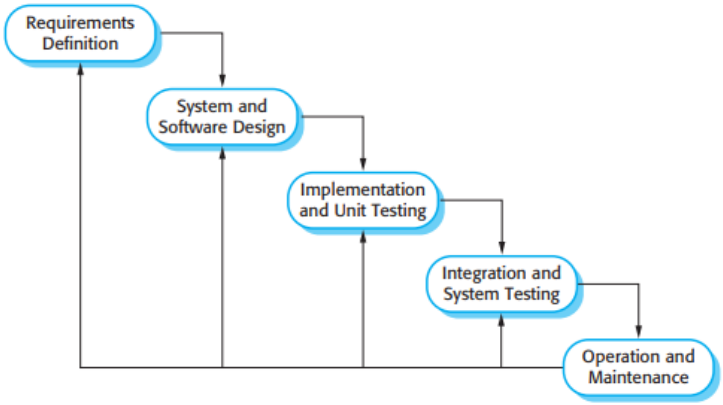

Figura 5 – Modelo em Cascata. Fonte: Sommerville (2011)

Para Sommerville (2011), uma desvantagem é o particionamento inflexível do projeto em etapas distintas. Conforme Figura 5, os requisitos são coletados uma única vez na fase inicial no processo, o que dificulta a resposta às mudanças caso haja alguma alteração no que foi previamente coletado com o cliente (SOMMERVILLE, 2011).

Desse modo, para que o modelo em cascata seja eficaz o usuário deverá conhecer todos os requisitos antes de iniciar o processo de desenvolvimento. Para projetos que não possuem muita complexidade e não requer tanta flexibilidade, o proposta em cascata pode ser apropriada.

3.2.2 Modelo em Espiral

O modelo espiral foi proposto inicialmente por Boehm (1988) (ver Figura 6). Nele o esforço de desenvolvimento é iterativo, ou seja, assim que uma iteração é concluída, outra iteração começa. Boehm (1988) afirmou que a principal característica distintiva do modelo espiral é que *"ele cria uma abordagem orientada ao risco para o processo de software, em vez de um processo orientado principalmente por documento ou por código."*

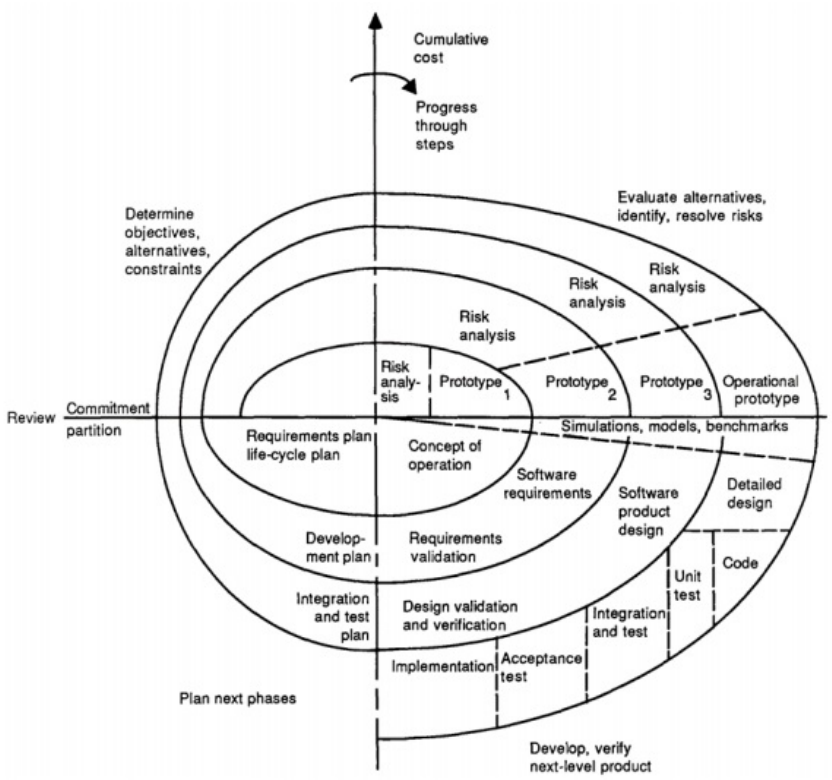

Figura 6 – Modelo em Espiral. Fonte: Boehm (1988)

Dessa forma, além das atividades que Royce (1970) propôs no modelo em cascata, o modelo em espiral possui atividades de gerenciamento de risco, reutilização e prototipagem. Nele o projeto começa em uma escala muito pequena, após isso são explorados os riscos e decidido se é correto passar para a próxima etapa, ou seja, para a próxima iteração da espiral.

3.2.3 Modelo Incremental

Esse modelo, proposto por Mills (1980), descreve uma abordagem incremental que destina-se a criar versões aprimoradas de um sistema. Para Sommerville (2011) o modelo incremental baseia-se na ideia de desenvolver uma implementação inicial, oferecendo ao usuário a possibilidade de comentar e refinar esses comentários através de muitas versões até que o sistema adequado seja desenvolvido. Nele as atividades de especificação, desenvolvi-

mento e validação são intercaladas e não separadas, com um *feedback* rápido em todas as atividades. A Figura 7 ilustra as atividades do modelo.

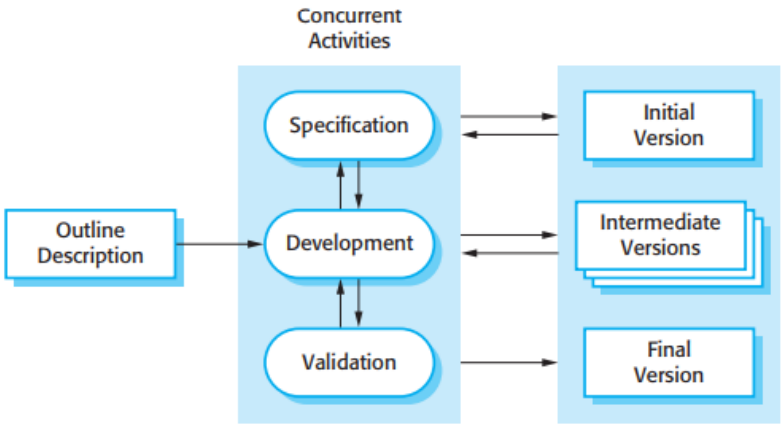

Figura 7 – Modelo Incremental. Fonte: Sommerville (2011)

O modelo incremental é fundamental para as abordagens ágeis, pois por desenvolver versões contínuas ao longo do seu ciclo reflete a maneira como as metodologias ágeis trabalham. Além disso, Sommerville (2011) afirma que, ao desenvolver o software de forma incremental, é mais barato e mais fácil fazer alterações no software conforme ele está sendo desenvolvido.

3.3 MODELOS DE CICLO DE VIDA DE DADOS E METADADOS

Diferentes Modelos de Ciclo de Vida que lidam diretamente com dados e metadados estão sendo discutidos e propostos na literatura por diversos autores. Para Cox e Tam (2018), o favoritismo de usar modelos de ciclo de vida é proveniente do *"apelo à dimensão temporal que a metáfora acrescenta à nossa compreensão das diferentes atividades em vista"*.

Dessa forma, buscando entender e classificar esses modelos alguns autores, como Ma e Wang (2010), tratam o ciclo de vida em duas perspectivas: valor e gerenciamento. Na perspectiva de gerenciamento, ele classifica os ciclos de vida de acordo com sua forma de visualização, que são os modelos de cadeia, matriz, circular, espiral, integrado e onda. Outro autor, Carlson (2014), classifica os ciclos de vida de acordo com os usuários. Ele sugere que os modelos possam ser divididos em três tipos: Individual, Organizacional e em Comunidades.

Contudo, esses modelos de ciclo de vida voltados ao domínio de dados podem variar de acordo com o cenário ao qual determinado projeto está inserido. Assim, cada autor propôs seus respectivos modelos e hoje há vários trabalhos que descrevem modelos de ciclo de vida de dados nos mais variados domínios. Alguns dos trabalhos encontrados durante nossa revisão *ad-hoc* de literatura são apresentados nas próximas seções. Foi

escolhido apresentar esses modelos pois são os que mais se assemelham a essa dissertação de mestrado.

3.3.1 Modelo de Ciclo de Vida de Metadados para Bibliotecas Digitais

Chen, Chen e Lin (2003) descrevem metadados como *"uma abordagem emergente para organizar coleções digitais estruturadas, a fim de apoiar a recuperação precisa, a preservação a longo prazo e a interoperabilidade em uma escala extraordinária da Internet."* Ele complementa que a falta de metodologias que recomende uma prática para o desenvolvimento de metadados em bibliotecas digitais aparece como um problema, pois não há um guia que indique como começar, como adquirir as necessidades de metadados, como escolher um padrão de metadados adequado e adotá-lo, como desenvolver a especificação de metadados, dentre outros questionamentos. Diante desse desafio, Chen, Chen e Lin (2003) propõe um Modelo de Ciclo de Vida de Metadados (Metadata Lifecycle Model - MLM). O modelo proposto (ver Figura 8) é composto por dez etapas, distribuídas em quatro grupos pelo qual os projetos de biblioteca digital podem projetar e implementar a provisão de metadados. Nesse modelo não são estipulados os papéis envolvidos em cada uma das etapas.

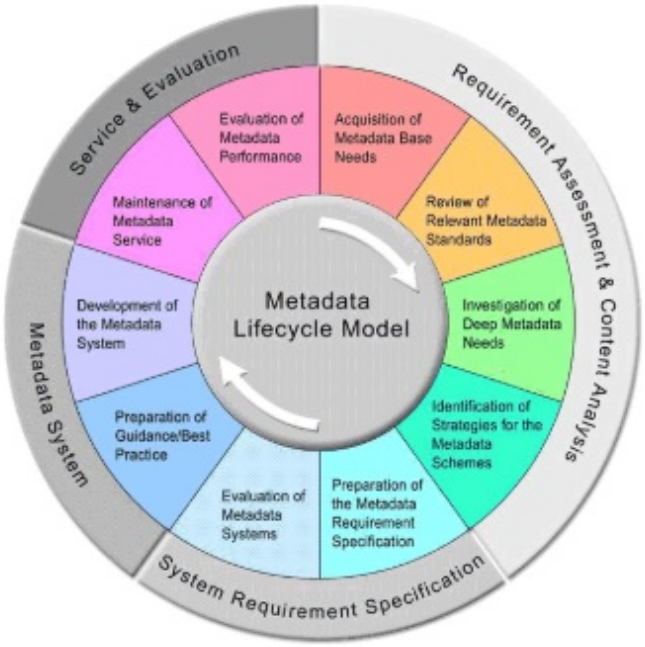

Figura 8 – Metadata Lifecycle Model (MLM). Fonte: Chen, Chen e Lin (2003)

As fases e atividades definidas por Chen, Chen e Lin (2003) foram:

1. Avaliação de Requisitos e Análise de Conteúdo
    - *Atividade 1: Aquisição de Necessidades de Base de Metadados*

É responsável por entrevistar os especialistas de conteúdo ou provedores sobre seus requisitos de metadados para cada projeto de coleta, e analisar os atributos dos projetos de coleta.

- *Atividade 2: Revisão dos Padrões e Projetos de Metadados Relevantes*
  Envolve a identificação de possíveis padrões de metadados, a examinação dos esquemas de metadados existentes e seus casos de uso.
- *Atividade 3: Investigação das Necessidades Profundas de Metadados*
  Essa atividade tem como objetivo identificar uma série de necessidades de metadados do projeto de biblioteca digital com mais precisão.
- *Atividade 4: Identificação de Estratégias para os Esquemas de Metadados e Alcançar a Interoperabilidade com Padrões de Metadados Conhecidos*
  Envolve a formulação da estratégia de metadados para a biblioteca digital, com base nos achados anteriores.

2. Especificação de Requisito do Sistema

   - *Atividade 5: Preparação da Especificação do Requisito de Metadados*
     Essa atividade tem o objetivo de preparar a especificação de requisitos para obter um acordo comum entre os participantes do projeto de coleta, especialistas em metadados e projetistas de sistemas.
   - *Atividade 6: Avaliação de Sistemas de Metadados*
     Esta etapa envolve a avaliação de potenciais sistemas de metadados.

3. Sistema de Metadados

   - *Atividade 7: Preparação da Orientação de Melhores Práticas*
     Essa atividade é responsável pela geração de diretrizes de práticas recomendadas para elementos de metadados individuais fornecidos pela especificação de requisitos de metadados.
   - *Atividade 8: Desenvolvimento do Sistema de Metadados*
     Os desenvolvedores de sistema irão desenvolver as ferramentas e sistemas de metadados de acordo com a especificação de requisitos.

4. Serviço e Avaliação

   - *Atividade 9: Manutenção do Serviço de Metadados*
     A manutenção do serviço de metadados tem como objetivo garantir a qualidade dos mecanismos de metadados.
   - *Atividade 10: Avaliação do desempenho de metadados*
     Esta etapa busca rever os resultados de todo processo de metadados e seu desempenho.

3.3.2 Modelo de Ciclo de Vida de Metadados para Objetos de Aprendizagem

Segundo Catteau, Vidal e Broisin (2006), o Repositório de Objetos de Aprendizagem (*Learning Object Repository* - LOR) permite o armazenamento de conteúdos de aprendizagem e fornece um ambiente para compartilhar e reutilizar Objetos de Aprendizagem (*Learning Objects* - LO). Portanto, com o intuito de alcançar uma representação completa e genérica que possa ser aplicada de forma eficaz a qualquer situação de aprendizagem, Catteau, Vidal e Broisin (2006) propuseram um Modelo de Ciclo de Vida de Metadados para Objetos de Aprendizagem. Esse modelo é apresentado na Figura 9.

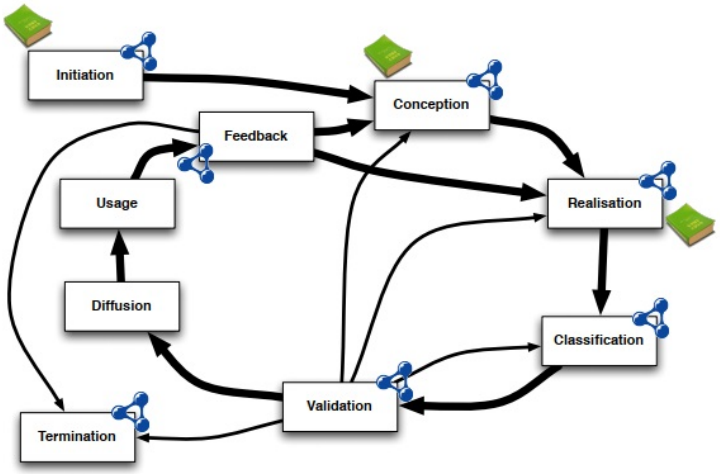

Figura 9 – Modelo de Ciclo de Vida de Metadados para Objetos de Aprendizagem. Fonte: Catteau, Vidal e Broisin (2006)

O modelo de Catteau, Vidal e Broisin (2006) é composto por 9 fases. São elas: iniciação, concepção, realização, classificação, validação, difusão, uso, feedback, término. Contudo, ele não descreve os papéis que serão envolvidos em cada fase.

3.3.3 Modelo Abstrato de Ciclo de Vida de Dados

O Modelo Abstrato de Ciclo de Vida de Dados (Abstract Data Lifecycle Model - ADLM) foi proposto por Möller (2013) e, por sua vez, enfoca os dados no contexto de sistemas técnicos, incluindo diferentes atores sociais e atividades. Nesse trabalho ele observou que nenhuma das literaturas encontradas foi além de seu respectivo domínio e propôs um modelo genérico de ciclo de vida para dados. Com o intuito de preencher essa lacuna, ele sugeriu um modelo chamado ADLM (ver Figura 10). O modelo é descrito por cinco partes: (i) um conjunto de fases que generalizará as etapas e processos definidos nos modelos individuais da pesquisa, (ii) um conjunto de características que podem ser usados para definir modelos de ciclo de vida, (iii) um conjunto de funções que descrevem os diferentes atores no modelo, (iv) um conjunto de características descrevendo os atores no modelo e

(v) um conjunto de características que descrevem os dados e metadados encontrados em um ciclo de vida.

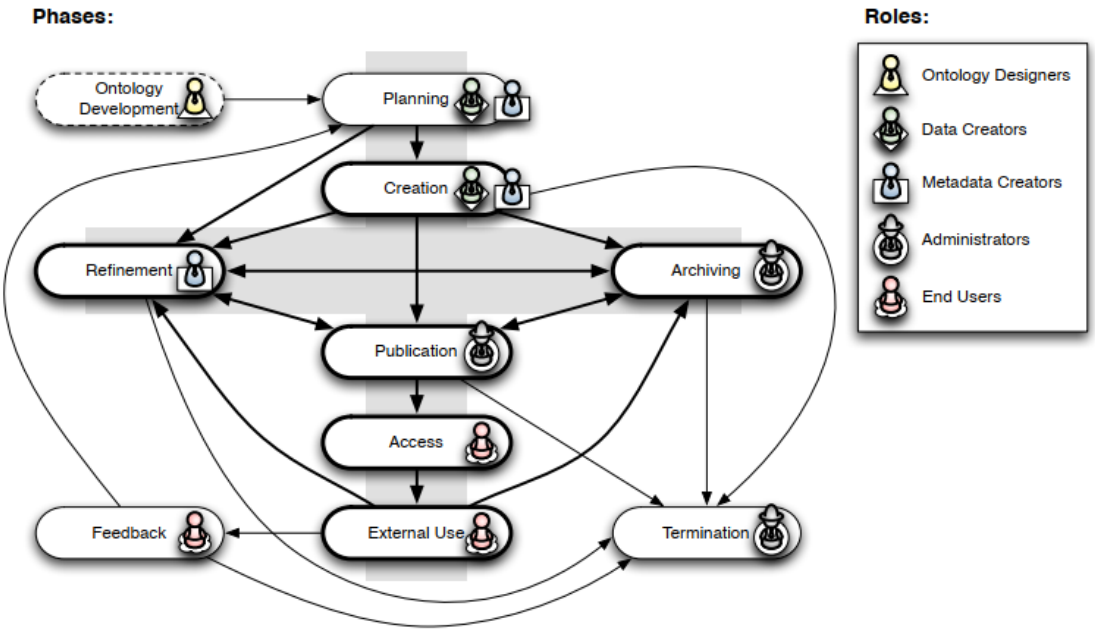

Figura 10 – Abstract Data Lifecycle Model (ADLM) proposto por Möller (2013). Fonte: Möller (2013)

O modelo de Möller (2013) possui 5 papéis e é composto por 10 fases. Os papéis elecandos por Möller (2013) foram: Designer de Ontologia, Criador de Dados, Criador de Metadados, Administradores e Usuários Finais. A seguir apresentamos uma descrição breve do que Möller (2013) estipulou para cada fase do ADLM.

- *Desenvolvimento de Ontologia*

  Esta fase é responsável por definir o modelo formal de representação do domínio.

- *Planejamento*

  O planejamento é responsável por preceder o ciclo de vida real dos dados, ela descreve o momento que a intenção de criar os dados assume uma forma concreta.

- *Criação*

  A fase de criação é o momento em que os dados e metadados são criados.

- *Arquivamento*

  Esta fase é encarregada de realizar o arquivamento de um dado dentro de um sistema por meio de indexação, catalogação ou similar.

- *Publicação*

A publicação é o momento em que os dados são disponibilizados dentro ou fora de algum sistema.

- *Acesso*

  É o momento em que os dados que foram publicados obtêm acesso, por meio de uma consulta ou através de navegação.

- *Uso Externo*

  O uso externo significa que o usuário realiza alguma ação adicionais com os dados, como exportar para outros sistemas.

- *Refinamento*

  O refinamento diz respeito a todo tipo de atividade que fazem acréscimos ou alterações nos dados.

- *Feedback*

  O feedback permite que usuários do sistema comentem os dados ou metadados que foram acessados.

- *Término*

  Como última, a fase de término é o momento em que os dados são removidos do sistema.

3.3.4 Modelo Abstrato de Ciclo de Vida de Dados Pessoais

Como foi objetivado por Möller (2013) ao desenvolver o ADLM, alguns trabalhos o utilizam como base para a modelagem de um novo modelo de ciclo de vida de dados. É o caso do trabalho de Alshammari e Simpson (2017), nele é proposto um Modelo Abstrato de Ciclo de Vida de Dados Pessoais (APDL) (ver Figura 11). O APDL é um modelo abstrato que representa dados pessoais em termos de estados e operações. Ele identifica um conjunto de estágios pelos quais os dados pessoais se movem durante sua vida útil e indica a ordem e a profundidade em que essas atividades podem ocorrer. O APDL tem o potencial de ser usado como um meio para classificar, comparar e relacionar outros modelos de ciclo de vida de dados pessoais, bem como para ser usado como base para definir novos modelos de ciclo de vida de dados pessoais para vários domínios.

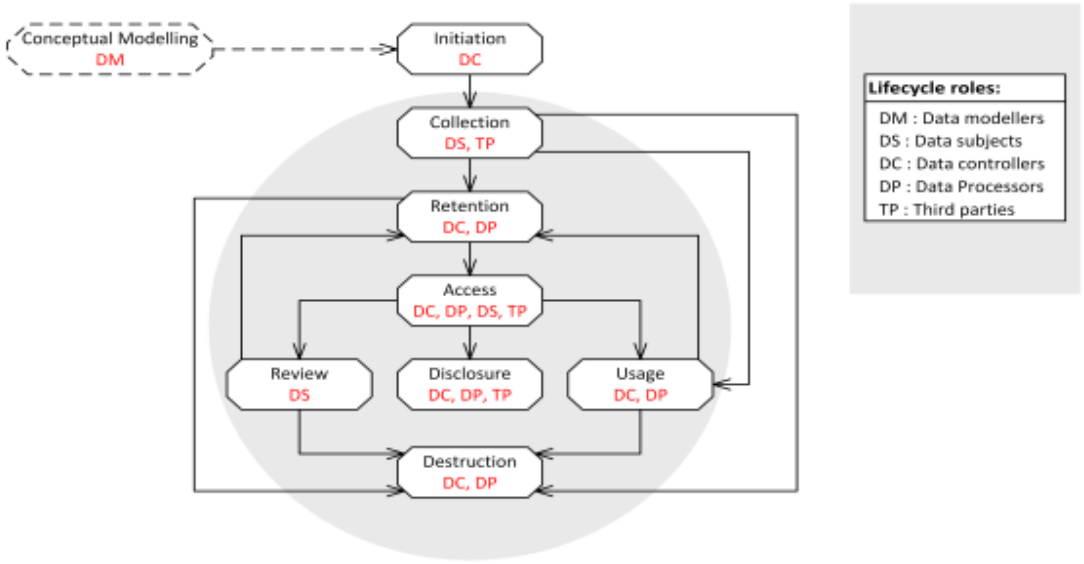

Figura 11 – Abstract Personal Data Lifecycle (APDL). Fonte: Alshammari e Simpson (2017)

Assim como no modelo proposto por Möller (2013), o APDL também descreve papéis e fases. Os papéis definidos por Alshammari e Simpson (2017) para o APDL foram: Modeladores de dados, Sujeitos de dados, Controladores de dados, Processadores de dados e Terceiros. Contudo, o modelo de Alshammari e Simpson (2017) tem diferenciais, visto que para cada fase ele determina suas entradas, saídas e os princípios GPS envolvidos. A sigla GPS significa *Global Privacy Standard*. Alshammari e Simpson (2017) relata que os princípios GPS foram aceitos como um conjunto unificado de princípios que refletem variantes apropriadas dos Princípios do Códigos de Práticas Justas de Informação (Fair Information Practice Principles - FIPPs). Alshammari e Simpson (2017) complementa afirmando que os princípios do GPS foram adotados para colocar limitações nos estágios do ciclo de vida e nas atividades associadas.

3.3.5 Ciclo de Vida dos Dados na Web

Um outro trabalho que merece destaque é o Ciclo de Vida de Dados na Web proposto por Lóscio, Oliveira e Bittencourt (2015). Esse ciclo de vida também foi inspirado no ADLM proposto por Möller (2013) e pode ser definido como "*um conjunto de fases que compõem o processo de publicação e consumo dos dados, contemplando desde o planejamento até o refinamento dos dados*". As fases dessa proposta de ciclo de vida são ilustradas na Figura 12.

Como atores que participam do ciclo de vida de dados na Web, Lóscio, Oliveira e Bittencourt (2015) determinam dois papéis: provedores e consumidores de dados. Os provedores de dados executariam as funções de criar metadados, criar os dados e publicar os

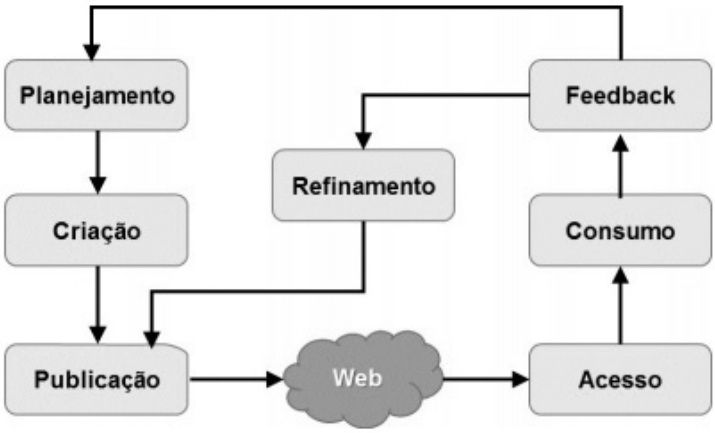

Figura 12 – Ciclo de Vida dos Dados na Web. Fonte: Lóscio, Oliveira e Bittencourt (2015)

dados. Enquanto os consumidores consomem esses dados.

Apesar do modelo de Lóscio, Oliveira e Bittencourt (2015) apresentar uma abordagem interessante para o ciclo de vida dos dados na Web, a proposta não evoluiu, sendo apenas uma breve descrição das fases do ciclo de vida. Dessa forma, a fim de dar continuidade a esse trabalho, nesta dissertação de mestrado o modelo de ciclo de vida proposto por Lóscio, Oliveira e Bittencourt (2015) foi utilizado como um ponto de partida para construção do modelo proposto nesta dissertação.

3.4 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados alguns modelos de ciclo de vida. Esses modelos foram distribuídos entre modelos de ciclo de vida de desenvolvimento de software e modelos de ciclo de vida de dados. Os modelos da área de desenvolvimento de software já existem há vários anos e atualmente é algo mais consolidado. De forma que, os livros de engenharia de software (*e.g* Sommerville (2011)) indicam tais modelos como pontos de referência para o início de desenvolvimento de software.

Entretanto, como observado diante dos modelos de ciclo de vida apresentados para a área de dados, a representação de um processo por meio de modelos de ciclo de vida ainda é algo novo. Muitas iniciativas foram criadas, mas ainda não há algo de referência que os usuários possam utilizar como modelo de partida. Em paralelo, o nível de detalhes que um modelo dispõe para os usuários é muito importante, pois eles serão essenciais para fornecer o máximo de detalhes acerca do conceito por trás do que está representado.

Diante disso, elaboramos um quadro comparativo (ver Quadro 7) entre os modelos de ciclo de vida de dados abordados no Capítulo 3.3, a fim de apresentar as dimensões tratados em cada modelo. Essas dimensões foram classificados em:

- Papéis: Essa dimensão diz respeito aos atores que estarão envolvidos em cada fase do ciclo de vida. Com ela é pretendido identificar os modelos que determinam o uso

de papéis durante o seu modelo de ciclo de vida.

- Fases: Essa dimensão está relacionada ao conjunto de fases de um modelo de ciclo de vida. Ela visa identificar se o modelo estipulou fases para acompanhar a evolução dos dados.
- Atividades: Outra dimensão de detalhamento são as atividades. Ela é responsável por descrever as atividades que serão realizadas em cada fase do modelo de ciclo de vida.
- Entradas: Para cada fase, é importante que o modelo de ciclo de vida determine as entradas. Ou seja, o que seria considerado como um *input* para inicialização de cada fase.
- Saídas: Além das entradas, também é interessante que o modelo determine as saídas de cada fase. Ou seja, seriam os resultados gerados a cada final de iteração de um fase do modelo.
- Características: Por fim, uma dimensão de detalhamento importante seriam as características do modelo, isto é, informações a respeito do modelo que possam ajudar a classificá-lo posteriormente.

Todas essas dimensões foram estipuladas a partir de uma análise do que cada modelo do Capitulo 3.3 dispõem como apoio para a aplicabilidade de um modelo de ciclo de vida de dados.

Quadro 3 – Comparação entre os modelos de ciclo de vida de dados

	Papéis	Fases	Atividades	Entradas	Saídas	Características
Metadata Lifecycle Model Chen, Chen e Lin (2003)	Não	Sim	Sim	Não	Não	Não
Metadata Lifecycle Model for Learning Objects Catteau, Vidal e Broisin (2006)	Não	Sim	Não	Não	Não	Não
Abstract Data Lifecycle Model Möller (2013)	Sim	Sim	Não	Não	Não	Sim
Abstract Personal Data Lifecycle Alshammari e Simpson (2017)	Sim	Sim	Sim	Sim	Sim	Não
Data on the Web Lifecycle Lóscio, Oliveira e Bittencourt (2015)	Sim	Sim	Não	Não	Não	Não

4 MODELO DE CICLO DE VIDA DE DADOS NA WEB

Como mostrado no capítulo anterior, os modelos de ciclo de vida propostos na literatura geralmente estão relacionados a domínios mais específicos. Somado a isso, os poucos modelos existentes, que versam sobre domínios genéricos, não contemplam de forma abrangente as diferentes fases existentes no domínio de Dados na Web ou, ainda, não versam sobre fases que são utilizadas nesse domínio. O modelo que mais se assemelha ao nosso é o proposto por Lóscio, Oliveira e Bittencourt (2015), porém ele é apresentado de forma muito simples e não se aprofunda a nível de determinar atividades, entradas ou saídas em cada fase. Dessa forma, neste trabalho propomos um Modelo de Ciclo de Vida de Dados na Web (*Data on the Web Lifecycle Model* - DWLM), que descreve as fases do ciclo de vida desde a concepção até o remoção de acesso ao conjunto dos dados, bem como os papéis, atividades, entradas e saídas de cada fase. Assim, este Capítulo está estruturado da seguinte forma: na Seção 4.1 será apresentado uma Visão Geral do DWLM, descrevendo a motivação para sua criação e o seu processo de desenvolvimento. Em seguida, na Seção 4.2 são apresentados os papéis que estarão envolvidos em cada fase do DWLM. A Seção 4.3 descreve suas fases, enquanto a Seção 4.4 apresenta um diagrama de atividades mostrando o fluxo principal de atividades do DWLM. Logo após, na Seção 4.5 são descritas as características do modelo e por fim, na Seção 4.6 são apresentadas as considerações finais do Capítulo.

4.1 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. Para isso, o modelo incorpora, ao longo de suas fases, as Boas Práticas para Dados na Web (DWBP) propostas pelo W3C. Como descrito na Seção 2.4, 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 descrito na Seção 2.1, onde 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.

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

des 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.

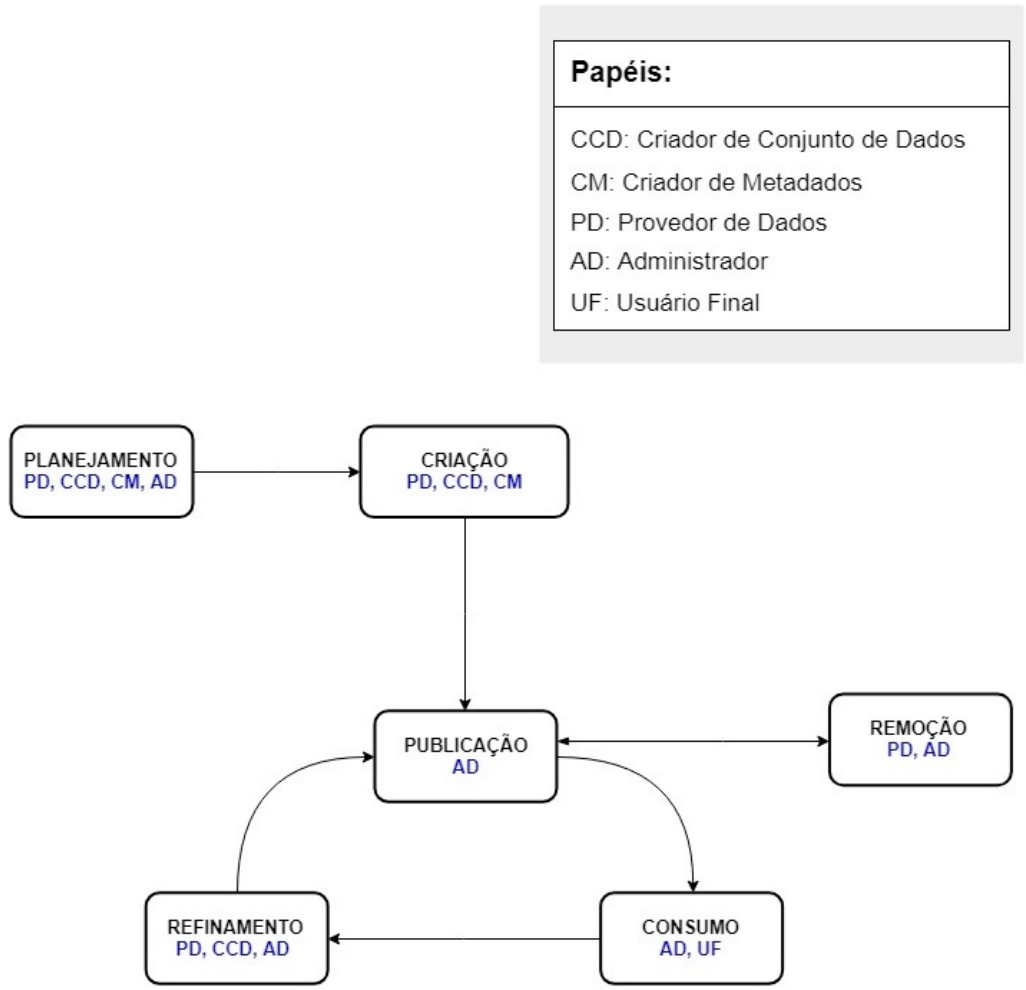

Figura 13 – Modelo de Ciclo de Vida de Dados na Web - DWLM. Fonte: Autor

Seguindo a mesma ideia do ADLM, o DWLM é composto por seis fases, como ilustrado na Figura 13. 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 4 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.

Quadro 4 – As fases, associando as atividades, entradas, saídas e DWBP usadas no DWLM

Fase	Atividades	Entrada	Saída	DWBP
Planejamento	Especificar fontes de dados Descrever conjunto de dados Estabelecer solução para publicação do conjunto de dados	Informações descritivas do conjunto de dados	Documento de Descrição do Conjunto de Dados Solução para publicação do conjunto de dados escolhida	BP1, BP2, BP3, BP4, BP5, BP6, BP12, BP14 BP15, BP16
Criação	Criar o conjunto de dados Avaliar qualidade Validar o conjunto de dados recém criado	Fontes de Dados Documento de Descrição do Conjunto de Dados	Documento de Inconsistência de Dados Termo de consentimento de publicação do conjunto de dados Conjunto de dados criado Documento de Descrição do Conjunto de Dados atualizado Métricas de Qualidade	BP9, BP10, BP11
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 do Conjunto de Dados atualizado	Conjunto de dados publicado	BP1, BP2, BP3, BP4, BP5, BP6, BP7, BP8, 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 Feedback do usuário final	BP29, BP30, BP34, BP35
Refinamento	Corrigir e enriquecer o conjunto de dados Validar o conjunto de dados refinado Versionar o conjunto de dados	Conjunto de Dados Feedback do usuário final	Conjunto de dados Refinado Log de Refinamento Documento de Inconsistência de Dados Termo de Consentimento Nova versão do conjunto de dados	BP7, BP8
Remoção	Remover acesso ao conjunto de dados	Documento de Solicitação de Remoção do Conjunto de Dados	Removido acesso ao conjunto de dados	BP27, BP28

4.2 PAPÉIS DO DWLM

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 13.

- *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 nem todos os modelos de ciclo de vida irão adotar um *Criador de Metadados*, visto 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*.

4.3 FASES DO DWLM

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 4.2. Além disso, cada fase está associada a uma ou mais DWBP 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.

4.3.1 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 14) 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*.

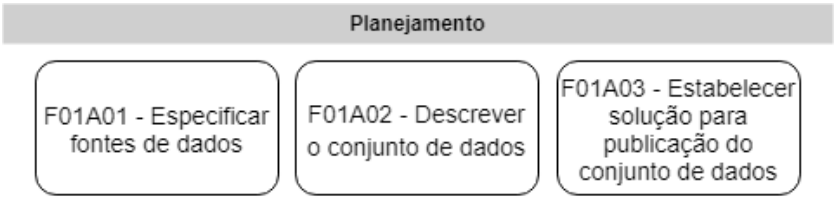

Figura 14 – Atividades da fase de *Planejamento*. Fonte: Autor

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

  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: Comma Separated Value (CSV), TXT, Extensible Markup Language (XML), Resource Description Framework (RDF) e JavaScript Object Notation (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*.
- 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 da Universidade Federal de Pernambuco (UFPE), 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

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

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 Necaskỳ 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 HyperText Markup Language (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 Comprehensive Knowledge Archive Network (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.

4.3.2 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

[2] https://ckan.org/

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-PE. 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 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.

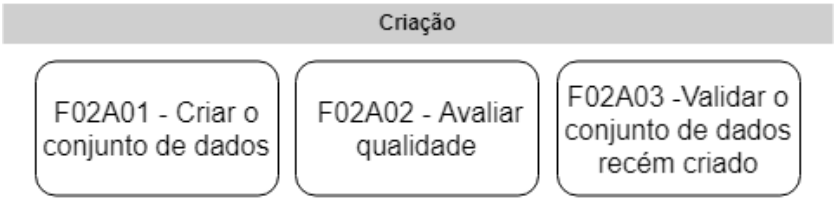

Figura 15 – Atividades da fase de *Criação*. Fonte: Autor

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 15). 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 (Extract Transform Load (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. Necaskỳ 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 Necaskỳ 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:

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

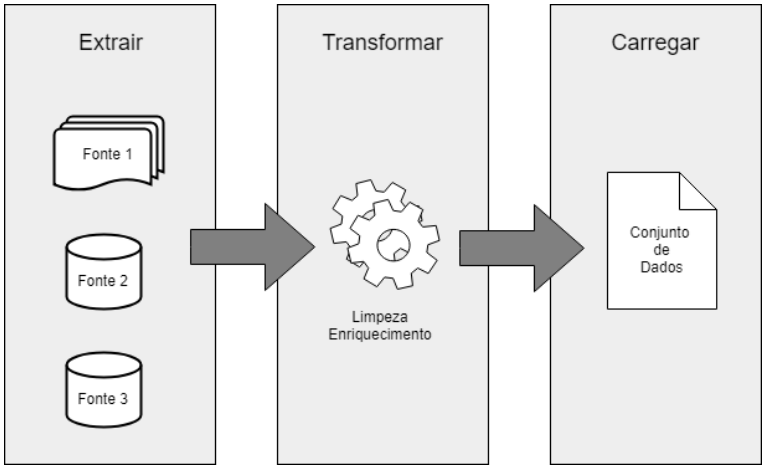

Figura 16 – Processo ETL para Dados na Web. Fonte: Autor

não realizarão nenhuma transformação nos dados. De acordo com Necaskỳ 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, Necaskỳ 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 (Microsoft Excel file format (XLS), Open Document (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*.

4.3.3 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 17), são elas:

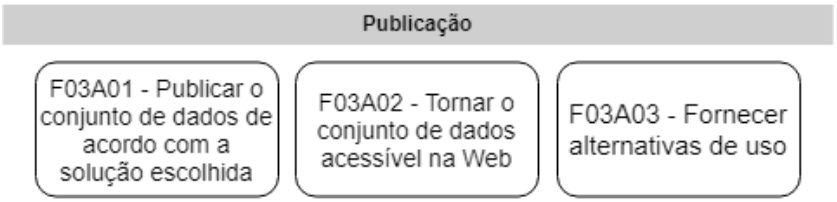

Figura 17 – Atividades da fase de *Publicação*. Fonte: Autor

- ***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 Necaskỳ 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* irá 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 Necaskỳ 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 (NECASKÝ 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. Hypertext Transfer Protocol (HTTP)). Pois, a publicação de conjuntos de dados em ambientes internos como Virtual Private Network (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.

4.3.4 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.

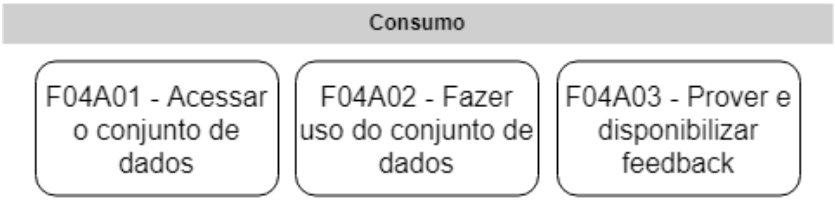

Figura 18 – Atividades da fase de *Consumo*. Fonte: Autor

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).

4.3.5 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 (SANTOS, 2018). Dessa forma, a fase de *Refinamento* do DWLM compreende atividades relacionadas a correções e enri-

quecimento 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 enriquecê-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:

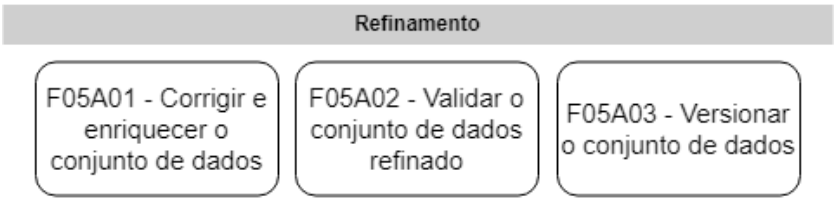

Figura 19 – Atividades da fase de *Refinamento*. Fonte: Autor

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

  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 Santos (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 13), 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.

4.3.6 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.

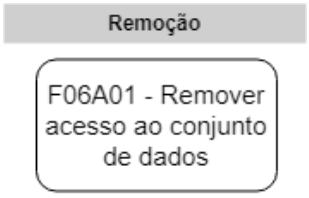

Figura 20 – Atividade da fase de *Remoção*. Fonte: Autor

- ***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.

4.4 DIAGRAMA DE ATIVIDADES DO DWLM

Com o intuito de ilustrar o fluxo principal de atividades do modelo DWLM, elaboramos o diagrama de atividades da Figura 21. Nele, podemos visualizar como as atividades de cada fase se comunicam e também identificar 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*. Outra atividade que também dependerá de uma decisão é a *F06A01 - Remover acesso ao conjunto de dados*, visto que, quando for solicitado a remoção de acesso a um conjunto de dados, seu acesso será removido. Ademais, um dos pontos que também merece destaque no diagrama são as atividades da fase do *Refinamento*. Depois do *Usuário Final* propor 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 dependeram muito do contexto para o qual o DWLM será aplicado.

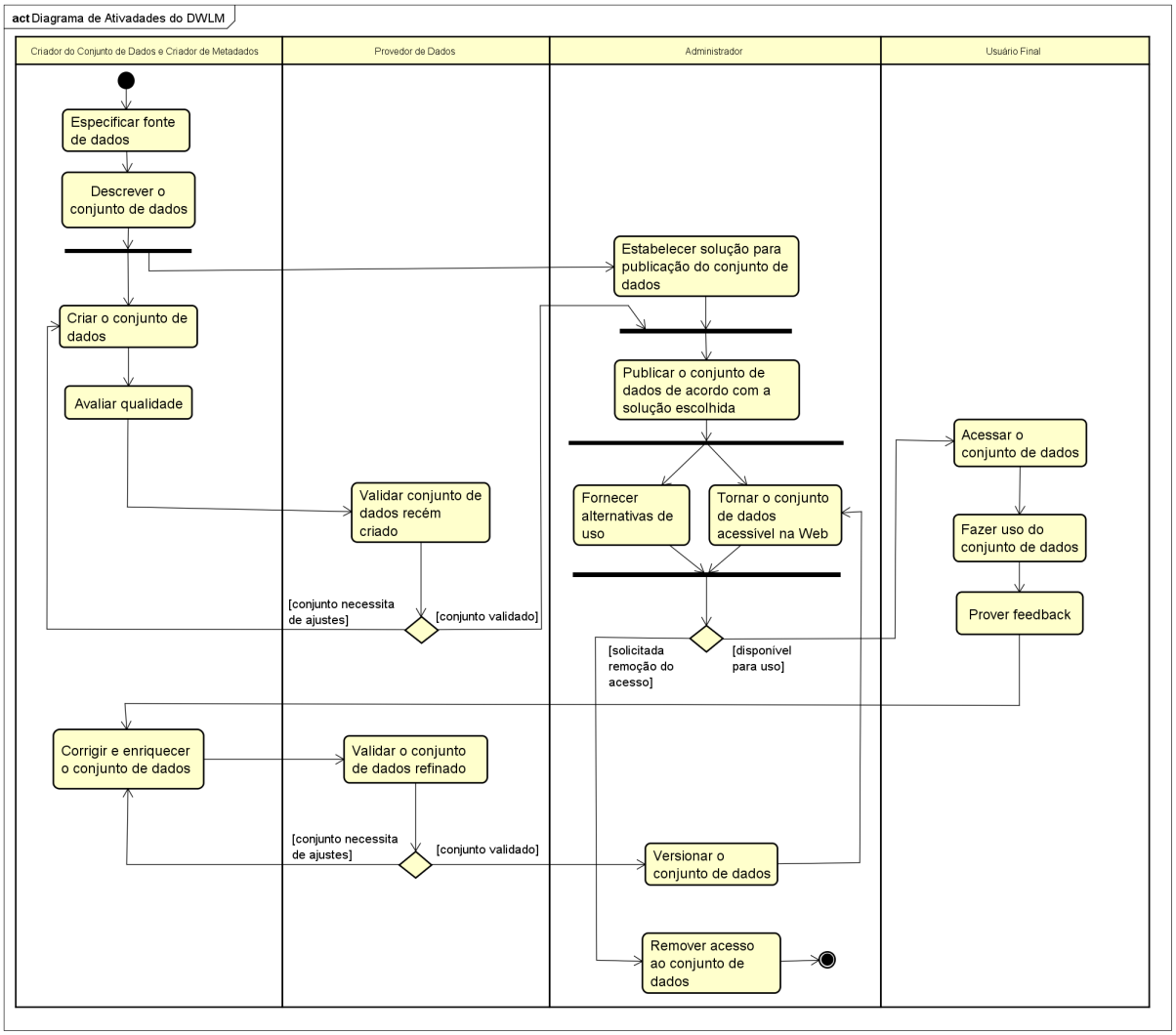

Figura 21 – Diagrama de Atividades do Modelo de Ciclo de Vida de Dados na Web. Fonte: Autor

4.5 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*".

4.5.1 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.

4.5.2 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.

4.5.3 Classificação do DWLM

Com o conjunto de características definidas, classificamos o DWLM em relação a cada característica apresentada na Seção 4.5. No Quadro 5 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.

Quadro 5 – 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*

4.6 CONSIDERAÇÕES FINAIS

Este capítulo apresentou uma proposta de Modelo de Ciclo de Vida de Dados na Web, chamado DWLM. Esse modelo foi construído a partir do estado da arte em publicação e consumo de dados na Web e de modelos de ciclo de vida propostos na literatura. O modelo proposto é o mais genérico possível, para que sua aplicabilidade englobe o máximo de cenários existentes. Além disso, seu desenvolvimento foi constituído por um processo

interativo e incremental, onde as fases e atividades foram identificadas, aprimoradas, refinadas e validadas até chegarmos a versão apresentada no presente trabalho.

O DWLM possui um conjunto de 16 atividades distribuídas dentre suas 6 fases. O modelo também aplica, em todas as suas fases, as boas práticas de publicação de dados na Web e determina as entradas e saídas de cada fase. Ademais, ele apresenta um conjunto de papéis que participam durante seu ciclo e um conjunto de características que, ao final, serviram para classificá-lo.

Em contraste aos modelos apresentados no Capítulo 3, nosso modelo se diferencia por além de estipular as fases, definir atividades, papéis, entradas, saídas e definir características, auxilia na aplicação das melhores práticas. Para realizar uma comparação dentre os modelos apresentados e o nosso, expandimos a tabela apresentada na seção 3.3 adicionando o DWLM a ela (ver Quadro 6).

No Quadro 7 é possível ter uma visão geral da contribuição de alguns dos modelos apresentados no Capítulo 3 para a construção do DWLM.

Quadro 6 – Comparação entre os modelos de ciclo de vida e o DWLM

	Papéis	Fases	Atividades	Entradas	Saídas	Características
Metadata Lifecycle Model Chen, Chen e Lin (2003)	Não	Sim	Sim	Não	Não	Não
Metadata Lifecycle Model for Learning Objects Catteau, Vidal e Broisin (2006)	Não	Sim	Não	Não	Não	Não
Abstract Data Lifecycle Model Möller (2013)	Sim	Sim	Não	Não	Não	Sim
Abstract Personal Data Lifecycle Alshammari e Simpson (2017)	Sim	Sim	Sim	Sim	Sim	Não
Data on the Web Lifecycle Lóscio, Oliveira e Bittencourt (2015)	Sim	Sim	Não	Não	Não	Não
Data on the Web Lifecycle Model DWLM	Sim	Sim	Sim	Sim	Sim	Sim

Quadro 7 – Contribuição de cada modelo para construção do DWLM

Papéis	Abstract Data Lifecycle Model Möller (2013)	Abstract Personal Data Lifecycle Alshammari e Simpson (2017)	Data on the Web Lifecycle Lóscio, Oliveira e Bittencourt (2015)	
Fases	Abstract Data Lifecycle Model Möller (2013)	Abstract Personal Data Lifecycle Alshammari e Simpson (2017)	Data on the Web Lifecycle Lóscio, Oliveira e Bittencourt (2015)	Metadata Lifecycle Model for Learning Objects Catteau, Vidal e Broisin (2006)
Atividades	Abstract Data Lifecycle Model Möller (2013)			
Entradas	Abstract Personal Data Lifecycle Alshammari e Simpson (2017)			
Saídas	Abstract Personal Data Lifecycle Alshammari e Simpson (2017)			
Características	Abstract Data Lifecycle Model Möller (2013)			

5 UM EXEMPLO DE USO DO DWLM

Este capítulo irá apresentar um exemplo de uso do DWLM com dados do Censo 2017 da Universidade Federal de Pernambuco. Na Seção 5.1, será descrito o contexto no qual a UFPE está inserida. Na Seção 5.2, será relatado o processo de aplicação do DWLM aos dados do Censo 2017. E, por fim, na Seção 5.3 serão descritas as considerações finais do Capítulo.

5.1 CONTEXTO DA UFPE

Em 2017, a Universidade Federal de Pernambuco (UFPE) instituiu um Plano de Dados Abertos (PDA). No PDA são estabelecidas as ações para a implementação e incentivo da abertura de dados na instituição. Nesse documento foram estabelecidos alguns objetivos e resultados esperados. Além disso, foram estabelecidos uma lista de critérios para a seleção de dados para abertura. Esses critérios envolvem: relevância para sociedade, relevância para a instituição, normativos legais, dentre outros. A partir dos critérios, algumas estratégias para seleção dos dados foram citadas, assim como uma lista de prioridades dos conjuntos de dados a serem abertos.

No PDA também foi apresentada sua estrutura de governança. Essa estrutura é mostrada na Figura 22. O Comitê Gestor de Tecnologia da Informação da UFPE (CGTI) desempenha a função de acompanhar a execução do PDA em nível estratégico. Enquanto o Laboratório de Dados e Informação da UFPE (aLADIN) ficará responsável pela publicação de novos conjuntos de dados, considerando a lista de prioridades estabelecida. O Núcleo de Tecnologia da Informação (NTI) ficará responsável por hospedar e prestar suporte ao Portal de Dados Abertos, assim como atuar nas decisões tecnológicas. E, por fim, os Responsáveis pelas Unidades Setoriais (RUS) serão encarregados de coordenar os processos de abertura de dados, bem como propor a publicação de novos conjuntos de dados.

Dessa forma, no final do ano de 2018 foi lançada uma versão beta do Portal de Dados Abertos da UFPE[1]. E, como um conjunto de dados *case* para lançamento do portal foi escolhido os dados provenientes do Censo 2017.

[1] https://dados.ufpe.br

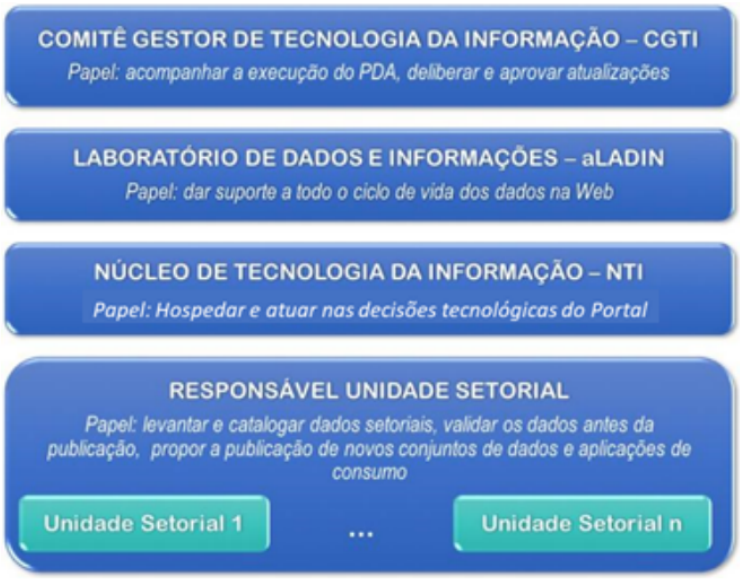

Figura 22 – Estrutura de Governança do PDA - UFPE. Fonte: UFPE (2017)

5.2 APLICAÇÃO DO DWLM NOS DADOS DO CENSO 2017

A Pró-Reitoria de Planejamento, Orçamento e Finanças (PROPLAN) em conjunto com a Pró-Reitoria de Comunicação, Informação e Tecnologia da Informação (PROCIT) e o aLADIN decidiram publicar, na versão beta do Portal de Dados Abertos da UFPE, os dados relacionados ao Censo da Educação Superior do ano de 2017 como um *case* para avaliação e validação do Portal.

O Censo é realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas (INEP) e é o instrumento de pesquisa mais completo do Brasil sobre Instituições de Educação Superior (IES) que ofertam cursos de graduação e sequenciais. O Censo reúne informações sobre as instituições de ensino, seus cursos de graduação presencial ou a distância, cursos sequenciais, vagas oferecidas, inscrições, matrículas, ingressantes e concluintes, dentre outras.

O processo de publicação do conjunto de dados do Censo foi realizado no final de 2018. A ideia dessa Seção é ilustrar o modelo proposto usando um conjunto de dados real. Assim, descreveremos o ciclo de vida do conjunto de dados do Censo a fim de constatar se as fases e atividades propostas no DWLM fazem sentido nesse cenário. É importante ressaltar que o uso do DWLM nos dados do Censo foi realizada de forma descritiva, visto que, o processo de publicação do conjunto se deu em paralelo ao processo de desenvolvimento do DWLM.

Desse modo, nas próximas Seções descreveremos os papeis envolvidos e ilustraremos as fases e atividades do DWLM que foram executadas durante todo o ciclo de vida desse conjunto de dados.

5.2.1 Definição de Papéis

No cenário do Censo 2017 da UFPE os atores participantes durante o ciclo de vida de dados foram: PROPLAN e aLADIN. A PROPLAN, por ser a detentora dos dados assumiu o papel de "*Provedor de Dados*", enquanto o aLADIN, em diferentes fases assumiu papéis distintos. Dessa forma, o aLADIN percorreu entre os papéis de "*Criador de Conjunto de Dados*", "*Criador de Metadados*" e "*Administrador*".

5.2.2 Planejamento

A fase de *Planejamento* tem como principal objetivo coletar todas as informações necessárias do conjunto de dados. E, para o contexto dos dados do Censo, não foi diferente. Essa fase recebeu como uma entrada um arquivo CSV contendo as informações descritivas dos dados e esse documento foi usado como auxilio na atividade de descrição do conjunto de dados. As atividades realizadas na fase de *Planejamento* foram:

- *F01A01 - Especificar fonte de dados*

  Para iniciar o processo de publicação dos dados, foi necessário primeiramente determinar as fontes de dados ao qual os dados eram provenientes. Dessa forma, foi identificado que a PROPLAN enviava anualmente para uma plataforma do INEP um arquivo TXT contendo todos os dados do Censo. Assim, foi determinado que a fonte dos dados seria esse arquivo TXT.

- *F01A02 - Descrever conjunto de dados*

  Na atividade de descrever o conjunto de dados foram coletadas algumas informações a respeito do conjunto e seus metadados. Algumas dessas informações foram coletadas a partir do arquivo CSV contendo as informações descritivas dos dados e colocadas no *Documento de Descrição do Conjunto de Dados*. Um *template* deste documento consta no Apêndice A desse trabalho. Algumas das informações coletadas nessa atividade foram: Título do Conjunto, Descrição, Palavras Chaves, Domínio, Cobertura Temporal, Cobertura Espacial, Linguagem, Formato de Data e Hora, Frequência de Atualização, Licença dos Dados, Provedor dos Dados, Publicador dos Dados e Formatos de Distribuição.

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

  Por ser o primeiro conjunto de dados a ser publicado na UFPE, a instituição ainda não tinha uma ferramenta de publicação de dados na Web. Dessa forma, foi necessário avaliar as ferramentas disponíveis e escolher uma. A ferramenta escolhida foi o IAGO, trata-se de um Sistema de Gerenciamento de Dados na Web proposto por Oliveira (2017). Essa ferramenta se destacou diante as demais por não ser apenas um catálogo de dados, mas também oferecer o gerenciamento e versionamento dos

dados. Assim, por ser uma solução que independe do conjunto de dados para iniciar sua implantação (*i.e* da classificação imposta no DWLM se classifica como *Solução Avançada*), seu processo de implantação já foi iniciado nesta atividade.

Entrada(s): Arquivo CSV com as informações descritivas dos dados.

Saída(s): Arquivo TXT com os dados, Documento de Descrição do Conjunto de Dados e Solução escolhida para a publicação do conjunto de dados (IAGO).

Boa(s) Prática(s) Envolvida(s): BP1, BP2, BP3, BP4, BP5, BP12, BP14.

5.2.3 Criação

Como apresentado no Capítulo 4, a etapa de criação é responsável por criar o conjunto de dados e validá-lo. Para o contexto da UFPE, esta etapa ocorreu de forma semelhante ao que foi descrito no DWLM, porém a atividade de Avaliação de Qualidade não foi realizada.

- *F02A01 - Criar o conjunto de dados*

  Nesta atividade, o aLADIN ficou responsável por criar o conjunto de dados. Para sua criação foi necessário ser executado um processo ETL, visto que haviam informações que não eram importantes para o contexto de publicação e algumas modificações para serem realizadas. Dessa forma, foi necessário realizar a extração desses dados do TXT, executar algumas transformações para ocultar dados pessoais dos discentes e realizar a carga em um arquivo CSV. A Figura 23 apresenta parte do CSV com os dados dos cursos gerados. Após sua criação, o conjunto de dados seguiu para a atividade *F02A03 - Validar o conjunto de dados recém criado.*

A	B	C	D	E	F	G	H	I	J	K	L	M
Tipo_de_r	Semestre_	Codigo_do_Curs	Codigo_	ID_na_IE	Turno_d	Situacao_	Curso_o	Semesti	Aluno_	Semestre_	Tipo_de_	Forma_de
42	2	13573			3	2				22016	0	0
42	1	13581			1	2				12014	1	1
42	2	13581			1	2				12014	1	1
42	2	101129			4	3				22015	0	1
42	1	101129			4	2				22015	0	1
42	2	13573			2	2				22015	0	0
42	1	13573			2	2				22015	0	0
42	1	13576			3	2				22012	0	1
42	2	13576			3	2				22012	0	1
42	2	13599			4	2				12013	0	1
42	1	13599			4	2				12013	0	1
42	2	13625			3	2				22011	0	1
42	1	13625			3	2				22011	0	1
42	2	101129			4	4				12010	2	1
42	1	101129			4	2				12010	2	1
42	1	13609			2	2				12015	0	0
42	2	13609			2	2				12015	0	0
42	2	13576			3	2				22013	0	1
42	1	13576			3	3				22013	0	1
42	1	13573			2	2				12007	2	0
42	2	13573			2	2				12007	2	0
42	1	13597			4	2				12009	2	1

Figura 23 – Parte do CSV com os dados dos cursos da UFPE no Censo 2017. Fonte: Autor

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

  Nesse momento, o conjunto de dados criado foi validado pelo *Provedor de dados*, nesse contexto a PROPLAN, para verificar se todos os dados estavam aptos a serem publicados e se, os dados sensíveis estavam ocultos. A PROPLAN, por sua vez validou o conjunto e assinou o Termo de Consentimento para Publicação do Conjunto de Dados. Assim como o Documento de Descrição, um *template* desse termo está disponível no Apêndice C deste trabalho.

Entrada(s): Arquivo TXT com os dados e Documento de Descrição do Conjunto de Dados.

Saída(s): Termo de Consentimento de Publicação do Conjunto de Dados, Conjunto de Dados Criado em CSV, Documento de Descrição do Conjunto de Dados Atualizado.

Boa(s) Prática(s) Envolvida(s): BP9.

5.2.4 Publicação

Após a criação e validação do conjunto, ele segue para a sua publicação. Nesta fase, ele foi publicado de acordo com a solução escolhida e seu acesso foi disponibilizado.

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

  Nesta fase, o aLADIN ficou responsável por realizar a publicação do conjunto de dados criado na ferramenta de publicação escolhida na atividade *F01A03*. Como dito, a ferramenta escolhida foi o IAGO e enquanto as outras atividades estavam sendo executadas, alguns integrantes do aLADIN, com o papel de "*Administrador*" realizaram a implementação da solução. Dessa forma, após finalizada a implementação da ferramenta, nesta fase é o momento de inserir o conjunto de dados nela. Bem como, realizar o preenchimento dos metadados com o apoio do *Documento de Descrição do Conjunto de Dados* e, se necessário, atualiza-lo.

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

  Após publicado o conjunto de dados e finalizada a implementação do IAGO, o Portal de Dados Abertos da UFPE foi lançado e, consequentemente o conjunto passou a ser acessível na Web. Hoje, o portal encontra-se na seguinte URL: <http://dados.ufpe.br>.

- *F03A03 - Fornecer alternativas de uso*

  No cenário dos dados do Censo 2017, foi criada uma área de indicadores no Portal de Dados da UFPE com o intuito de fornecer visualizações alternativas aos *Usuários Finais* que irão acessá-lo. A Figura 24 mostra alguns desses indicadores.

GRADUAÇÃO - DADOS CENSO 2017

CAMPI	CURSOS	INGRESSANTES	MATRICULADOS	CONCLUINTES
CARUARU	12	1.045	5.237	414
RECIFE	87	5.667	29.274	3.321
VITÓRIA	6	422	2.067	236

Observação: Os dados informados nos indicadores são referentes aos cursos de graduação do ano de 2017.

CAMPI	CENTRO	Nº DE CURSOS (PRESENCIAL, EAD[1], PARFOR[2])
CARUARU	CENTRO ACADÊMICO DO AGRESTE - CAA	12
RECIFE	CENTRO DE ARTES E COMUNICAÇÃO - CAC	27
RECIFE	CENTRO DE BIOCIÊNCIAS - CB	4
RECIFE	CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA - CCEN	9
RECIFE	CENTRO DE CIÊNCIAS JURÍDICAS - CCJ	1
RECIFE	CENTRO DE CIÊNCIAS DA SAÚDE - CCS	11
RECIFE	CENTRO DE CIÊNCIAS SOCIAIS APLICADAS - CCSA	8
RECIFE	CENTRO DE EDUCAÇÃO - CE	8
RECIFE	CENTRO DE FILOSOFIA E CIÊNCIAS HUMANAS - CFCH	14

Figura 24 – Alternativas de uso disponibilizadas no Portal de Dados Abertos da UFPE. Fonte: UFPE (2018)

Entrada(s): Conjunto de Dados Criado em CSV, Documento de Descrição do Conjunto de Dados.

Saída(s): Conjunto de Dados Publicado.

Boa(s) Prática(s) Envolvida(s): BP17, BP18, BP19, BP20, BP21, BP22, BP23, BP24, BP25, BP26, BP32.

5.2.5 Consumo

Esta etapa retrata o momento em que o conjunto de dados começa a ser acessado e, consequentemente, consumido.

- *F04A01 - Acessar o conjunto de dados*

  Caracterizar o acesso a um conjunto de dados é algo bem subjetivo, visto que não temos como definir com precisão os *Usuários Finais* que acessaram o conjunto. Entretanto, foi incluído um mecanismo de contagem a cada visualização da página de descrição do conjunto e, com esse dado da para ter uma ideia de como andam os acessos ao conjunto de dados. Além disso, também há um indicador de quantidade de *download* que um conjunto teve, como podemos observar na Figura 25.

Figura 25 – Página de Detalhamento do Conjunto de Dados do Censo 2017. Fonte: UFPE (2018)

- *F04A02 - Fazer uso do conjunto de dados*

  Após acessar o conjunto de dados e despertar o interesse do *Usuário final*, é natural que, ele venha a criar algum tipo de análise ou processe esse dados de alguma outra forma. Como trata-se de um conjunto de dados publicado recentemente, ainda não se tem nenhuma informação quanto a análises/aplicações criadas utilizando-o.

- *F04A03 - Prover e disponibilizar feedback*

  Para atender essa atividade, foi disponibilizado um ambiente no portal onde os *Usuários Finais* poderão enviar feedback sobre o conjunto. Além disso, os feedback já enviados ficam disponíveis para que outros usuários também possam ver. Dessa forma, O *Usuário Final* poderá verificar se o que ele desejava pedir já foi solicitado por outra pessoa.

Entrada(s): Conjunto de Dados Publicado.
Saída(s): Conjunto de Dados Acessado, *Feedback* do Usuário Final.
Boa(s) Prática(s) Envolvida(s): BP29, BP30.

5.2.6 Refinamento

A fase de refinamento é responsável por realizar alterações no conjunto de dados a fim de corrigi-lo ou enriquece-lo. Entretanto, por se tratar de um conjunto de dados que foi

publicado recentemente, ainda não há históricos de versões e consequentemente, não há refinamentos realizados. Desse modo, nas descrições das atividades será relatado como será realizado esse processo de *Refinamento* no contexto da UFPE.

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

  Após consumir o conjunto de dados, os *Usuários Finais* poderão enviar *feedback* propondo correções e/ou enriquecimento do conjunto por meio da atividade *F04A03 - Prover e disponibilizar feedback*. A partir desse feedback, o *Criador de Conjunto de dados*, neste contexto sendo executado pelo aLADIN, irá realizar as alterações propostas pelos usuários no conjunto e criar um arquivo de Log contendo quais foram as alterações realizadas para que na próxima atividade essas alterações sejam validadas. Além disso, o próprio *Criador do Conjunto de Dados* (aLADIN) poderá, a qualquer momento. após a publicação do conjunto, realizar correções e/ou enriquecimento dos dados.

- *F05A02 - Validar o conjunto de dados refinado*

  Após corrigido e/ou enriquecido, o conjunto de dados precisará passar por uma validação. No contexto da UFPE, esse novo conjunto alterado será enviado a PROPLAN e ela irá validar se as alterações realizadas estão condizentes. Essas alterações serão identificadas a partir do arquivo de Log enviado. Caso as modificações sejam validadas o conjunto segue para o versionamento. Caso não, as alterações voltarão para a atividade de *F05A01 - Corrigir e enriquecer o conjunto de dados* para serem corrigidas.

- F05A03 - Versionar o conjunto de dados

  Com o conjunto validado pelo *Provedor de dados*, ele estará apto a ser versionado. Dessa forma, o aLADIN irá, com o papel do *Administrador*, tornar disponível a nova versão do conjunto de dados por meio da atividade *F03A02 - Tornar o conjunto de dados acessível na Web*.

Entrada(s): Conjunto de Dados, *Feeback* do Usuário Final.

Saída(s): Conjunto de Dados Refinado, Log de Refinamento, Termo de Consentimento de Publicação de Dados, Nova Versão do Conjunto de Dados.

Boa(s) Prática(s) Envolvida(s): BP7, BP8.

5.2.7 Remoção

Esta fase descreve o momento em que a PROPLAN irá solicitar a remoção do conjunto de dados do Portal de Dados Abertos da UFPE via documento de solicitação (Apêndice D). Porém, assim como a fase de *Refinamento*, ainda não houveram solicitações de remoção

de acesso ao conjunto de dados. Dessa forma, na atividade abaixo será relatado a forma como esse processo ocorrerá no contexto da UFPE.

- *F06A01 - Remover acesso ao conjunto de dados*

  Após a PROPLAN enviar o *Documento de Solicitação de Remoção do Conjunto de Dados* a atividade de remover acesso será iniciada. Nesta atividade, o conjunto de dados não será removido, entretanto seu acesso sim. Ao passo que, um *Usuário Final* ao acessar o conjunto de dados que teve seu acesso removido receberá uma página informando que o conjunto não está mais disponível, o porque do seu acesso haver sido removido e que ele poderá solicitar uma cópia do conjunto de dados, se possível. É importante destacar que para envio dessa cópia do conjunto de dados é necessário que a PROPLAN dê consentimento.

Entrada(s): Documento de Solicitação de Remoção do Conjunto de Dados.
Saída(s): Acesso ao Conjunto de Dados Removido.
Boa(s) Prática(s) Envolvida(s): BP27.

5.3 CONSIDERAÇÕES FINAIS

Neste Capítulo apresentamos um exemplo de aplicação do Modelo de Ciclo de Vida de Dados na Web para o conjunto de dados do Censo 2017 da UFPE. Primeiramente, foi apresentado o contexto ao qual a UFPE está inserida em relação ao cenário de publicação e consumo de dados e, em seguida, foi descrito os dados que seriam utilizados para ilustração da aplicação do DWLM.

Para cada fase, foram apresentadas as ações que foram realizadas e mostrado as entradas, saídas e boas práticas envolvidas. Além disso, foram apresentados os papéis que estariam diretamente envolvidos no ciclo de vida. Sendo assim, neste Capítulo foi possível visualizar com mais clareza a aplicação do DWLM e sua importância não apenas no apoio de todo processo de publicação e consumo de dados, mas também na geração de documentos que descreverão toda a evolução do conjunto de dados no decorrer do seu ciclo de vida.

6 AVALIAÇÃO DO DWLM

Neste capítulo é discutida a avaliação do Modelo de Ciclo de Vida de Dados na Web proposto neste trabalho. Para essa avaliação, foi utilizado o método qualitativo de Grupo Focal (*Focus Group*). Dessa forma, a Seção 6.1 apresenta o que constitui um grupo focal, enquanto que a Seção 6.2 descreve sua organização, desde seu planejamento (Seção 6.2.1), condução (Seção 6.2.2) a análise dos dados (Seção 6.2.3). Por fim, na Seção 6.3 são apresentadas as considerações finais do Capítulo.

6.1 GRUPO FOCAL

Neto, Moreira e Sucena (2002) define o Grupo Focal como *"uma técnica de pesquisa na qual o pesquisador reúne, num mesmo local e durante um certo período, uma determinada quantidade de pessoas que fazem parte do público-alvo de suas investigações, tendo como objetivo coletar, a partir do diálogo e do debate com e entre eles, informações acerca de um tema específico"*. Nos últimos anos, o grupo focal tem se tornado popular em diversas áreas como Medicina, Ciências Sociais, Biologia e Ciências da Informação (ZAGANELLI et al., 2015).

Nessa técnica, os participantes irão expôr suas ideias acerca do tema discutido e o moderador da discussão não poderá interferir ou induzir na interpretação deles. Para Chiara (2005) o grupo focal não busca obter consenso, o moderador é que deve criar condições para que diferentes percepções e pontos de vista sejam colocados durante as sessões. Além disso, Chiara (2005) alerta que as discussões inerentes ao processo de adoção dessa técnica devem ocorrer em clima de tranquilidade, sem pressões, de modo que se possa garantir a troca de opiniões em relação ao objeto de estudo. É importante deixar claro que esses participantes geralmente são selecionados devido ao seu nível de conhecimento na área de estudo, ou seja, eles não representam necessariamente uma amostra representativa da população.

De acordo com Kontio, Lehtola e Bragge (2004), o grupo focal é um método empírico rápido e econômico para coletar evidências e realizar avaliações usando os participantes. Esse método também pode fornecer dados qualitativos e informações ricas, além de revelar percepções que são difíceis ou caras de capturar com outros métodos Kontio, Lehtola e Bragge (2004).

6.2 ORGANIZAÇÃO DO GRUPO FOCAL

Para a organização do Grupo Focal foram seguidas as etapas estipuladas por Chiara (2005). Ela definiu três fases: planejamento, condução do grupo focal e análise dos dados. Essas três etapas serão detalhadas nas próximas seções.

6.2.1 Planejamento

Seguindo as recomendações de Chiara (2005), os primeiros passos foram: definir os objetivos do estudo, os critérios para selecionar os participantes, decidir o local de realização das discussões, a duração da sessão, as questões que seriam feitas e elaborar os documentos que seriam entregues aos participantes com os objetivos do estudo.

O objetivo principal do estudo foi realizar uma avaliação de alto nível a respeito da viabilidade, completude e adequação do Modelo de Ciclo de Vida de Dados na Web. Para essa avaliação foram pensados em dez possíveis participantes que poderiam atender aos critérios estipulados. Esses critérios foram:

- conhecimento e expertise em Dados na Web;
- conhecimento e experiência na publicação e consumo de dados;
- disposição em compartilhar suas experiências e opiniões.

Foi enviado um convite para os dez possíveis profissionais que, ao nosso ver, atenderiam aos nossos critérios.

Após realização dos dez convites, oito responderam, mas conseguir conciliar suas agendas foi algo bem complicado, visto que esses profissionais geralmente são muito ocupados. Dessa forma, após muitas discussões e propostas de datas e horários conseguimos fechar em seis profissionais que se dispuseram a participar da pesquisa.

Na tabela 8 são apresentadas as informações gerais dos participantes deste estudo. Todos os participantes são residentes do estado de Pernambuco, Brasil. Quanto aos seus cargos, quatro são estudantes (mestrado e doutorado), onde um dos quatro também desenvolve a função de Analista de Sistemas. Temos também um Professor e um Coordenador de Gestão de Projetos. Dentre os seis, apenas um deles desenvolve atividades de dados na Web apenas no meio profissional. Enquanto três desenvolvem no meio acadêmico e os outros dois em ambos os meios. Em relação a formação, dois possuem mestrado completo, um possui especialização e três graduação. Todos eles tem, no mínimo, 3 anos de experiência com Dados na Web.

Para a realização da sessão, foi escolhido um local de comum acordo para todos, diante disso o lugar escolhido foi o Centro de Informática da Universidade Federal de Pernambuco - CIn/UFPE. A sessão foi realizada em Fevereiro de 2019, com início as 8h30 e término as 11h30, ou seja, foram 3h de duração. Para a realização da sessão, foi escolhido um moderador, que ficou responsável por conduzir toda a sessão e assegurar que o foco de discussão estivesse dentro do tema abordado. Também foram definidos dois observadores, um ficou responsável por realizar todas as anotações pertinentes durante a sessão enquanto o outro pela gravação de áudio da sessão.

Quadro 8 – Informações dos Participantes do Grupo Focal.

Participante	Função/ Cargo	Formação	Experiência com Dados na Web	Meio que desenvolve as atividades
Participante 1	Professor	Mestrado Acadêmico/ Profissional	4 anos	Acadêmico e Profissional
Participante 2	Estudante de Mestrado	Graduação	7 anos	Acadêmico e Profissional
Participante 3	Estudante de Mestrado	Graduação	3 anos	Acadêmico
Participante 4	Estudante de Mestrado/ Analista de Sistemas	Graduação	3 anos	Acadêmico
Participante 5	Coordenador de Gestão de Projetos	Especialização	6 anos	Profissional
Participante 6	Estudante de Doutorado	Mestrado Acadêmico/ Profissional	4 anos	Acadêmico

6.2.2 Condução do Grupo Focal

A sessão do grupo focal começou com uma breve apresentação do moderador, observadores e participantes. Cada um teve a oportunidade de dizer seu nome, o que faz e sua experiência na área. Após isso, foi entregue a cada participante o termo de consentimento para que o participante ficasse a par do que seria discutido e que teriam o seu anonimato garantido. Também foi informado que a sessão seria gravada e perguntado se todos estavam de acordo. Juntamente com o termo, também foi entregue um questionário a cada participante. Esse questionário continha as informações que deveriam ser avaliadas por cada um a respeito dos principais elementos do DWLM.

Na sequência, o moderador abriu a sessão com uma breve explicação da pesquisa descrevendo o assunto que seria discutido e os objetivos do grupo focal. Além disso, também foram apresentadas algumas regras de funcionamento. Essas regras foram:

- Nós queremos ouvir a opinião de todos;
- Todos os comentários são válidos;
- Não existem respostas corretas ou incorretas;
- Não fugir de tópico.

Logo após, foi explicado que o grupo focal seria dividido em três fases. Na primeira, os participantes iriam preencher as suas informações básicas (*i.e* função/cargo, formação, experiência, meio que desenvolve as atividades). A segunda seria um questionário com nove perguntas a respeito dos elementos (em conjunto e separado) do DWLM. E por último, na terceira fase seriam três perguntas gerais, a fim de coletar dados de avaliação de qualidade, bem como comentários gerais. O formulário aplicado pode ser encontrado em: <http://bit.ly/avaliacaoDWLM>

6.2.3 Análise dos Dados

Esta seção apresenta a análise dos dados coletados no grupo focal, discutindo detalhadamente cada pergunta realizada e apontando algumas questões que devem ser levadas em consideração.

6.2.3.1 Avaliação dos Papéis

A primeira pergunta foi relacionada aos papéis definidos no DWLM. Foi perguntado se conjunto de papéis estava descrito corretamente e se eles são viáveis, completos e adequados. Além disso, com esse questionamento almejávamos detectar a existência de lacunas, erros e/ou possibilidades de melhoria. A Figura 26 resume as respostas.

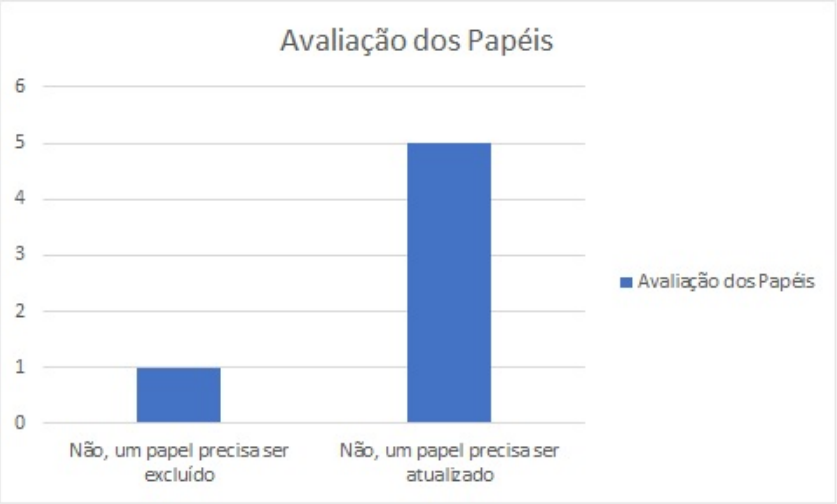

Figura 26 – Avaliação dos Papéis. Fonte: Autor

Participante 1 selecionou "Não, um papel precisa ser atualizado" e informou que "*Tenho dúvidas quanto à noção/nomenclatura de Criador de Dados, sob o ponto de vista da origem: o dado não seria criado "dentro" do provedor, originalmente.*"

Participante 3 selecionou "Não, um papel precisa ser excluído" e afirmou que "*Os papéis de criador de dados e metadados devem ser fundidos.*"

Participante 4 selecionou "Não, um papel precisa ser atualizado" e afirmou que "*Atualizar a descrição do papel do provedor de dados. Não ficou claro a principal diferença entre o criador e o provedor, apesar do documento afirmar que um mesmo papel pode assumir mais de uma função, a forma como foi descrito e discutido, não haveria uma necessidade de se ter um provedor de dados e um criador de dados ao mesmo tempo.*"

Participante 5 selecionou "Não, um papel precisa ser atualizado" e comentou que "*Substituir as palavras "criador" de dados e "criador" de metadados, talvez, por publicador/fornecedor. E, nesse caso talvez rever o perfil administrador. Lembrando que publicador e fornecedor seriam perfis diferentes.*"

Com as melhorias propostas identificamos que houve uma dúvida em relação ao papel do *Criador de Dados*, *Criador de Metadados* e *Provedor de Dados*. Em alguns momentos, os participantes confundiram os papéis e suas atribuições. Notamos que o principal motivo dessa confusão foi devido ao nome do papel de "criador" de dados, visto que no contexto do DWLM é o *Provedor de Dados* que representa o proprietário do dado e o nome "criador" acabou deixando isso um pouco confuso. A partir disso, alteramos o papel de "*Criador de Dados*" para "*Criador de Conjunto de dados*" a fim de sanar essas dúvidas.

Em relação à proposta de fundir os papéis de *Criador de Dados* com *Criador de Metadados*, não concordamos. Os dois representam papéis que trabalham com pontos diferentes e que podem ser executados em paralelo, ou seja, enquanto o *Criador de Dados* está executando um atividade o *Criador de Metadados* poderá executar outra. Entretanto, é importante deixar claro que o DWLM permite que um ator, em um determinado cenário, exerça os dois papéis.

6.2.3.2 Avaliação do Conjunto de Fases

A segunda questão foi relacionada ao Conjunto das Fases e teve como objetivo identificar se ele estava descrito corretamente. Além disso, almejávamos verificar se as fases são viáveis, completas e adequadas para o Ciclo de Vida de Dados na Web, bem como detectar a existência de lacunas, erros e/ou possibilidades de melhoria. A Figura 27 resume as respostas.

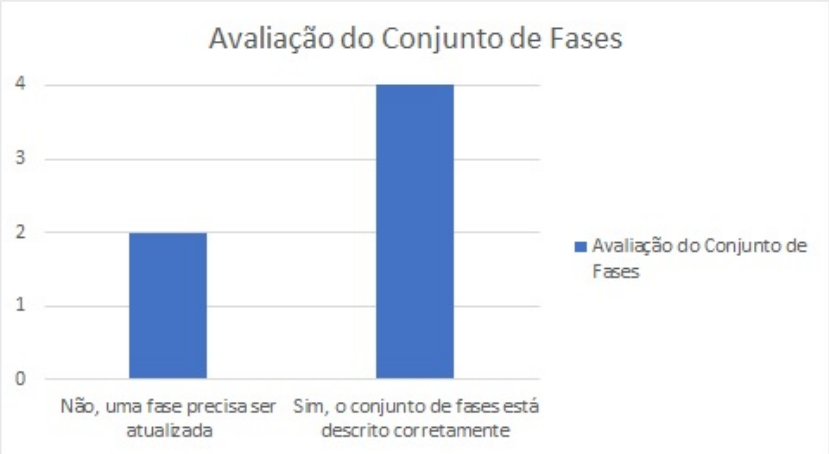

Figura 27 – Avaliação do Conjunto de Fases do DWLM. Fonte: Autor

Participante 2 selecionou "Não, uma fase precisa ser atualizada" e comentou "*...Definir bem o termo "arquivamento" e ver se esse nome é o mais apropriado.*"

Participante 4 selecionou "Não, uma fase precisa ser atualizada" e comentou "*Talvez mudar o nome da fase de arquivamento.*"

Os demais participantes selecionaram a opção "Sim, o conjunto de fases está descrito corretamente" e não pontuaram nada em relação a pergunta.

Na definição do modelo, o nome para a fase de *Arquivamento* também foi uma dúvida. Após realização dessa avaliação, decidimos alterar o nome de *Arquivamento* para *Remoção*. Pois, o principal objetivo dessa fase é remover o acesso ao conjunto e dados. Dessa forma, o modelo foi atualizado para realizar a troca de nome da fase. Ressaltamos que, como essa avaliação foi feita com a fase sendo denominada como "*Arquivamento*", nas próximas seções preservaremos esse nome.

6.2.3.3 Avaliação da Fase de Planejamento

Após a avaliação do Conjunto de Fases, foi iniciada a avaliação de cada uma das fases. Inicialmente, questionamos se a fase de Planejamento estava descrita corretamente. Nesta pergunta, almejávamos verificar se as atividades são viáveis, completas e adequadas, bem como detectar a existência de lacunas, erros e/ou possibilidades de melhoria. A Figura 28 resume as respostas.

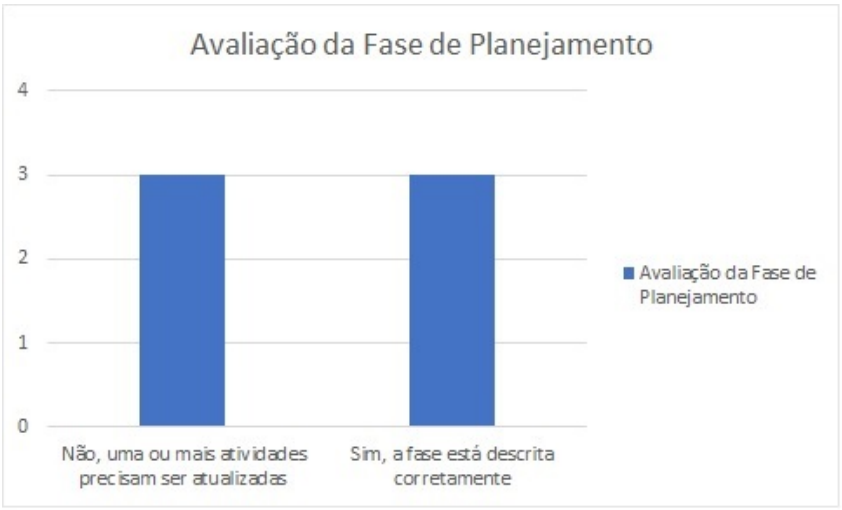

Figura 28 – Avaliação da Fase de Planejamento. Fonte: Autor

Participante 1 respondeu "Não, uma ou mais atividades precisam ser atualizadas" e comentou "*É interessante trazer o provedor de dados para esta fase.*"

Participante 2 respondeu "Não, uma ou mais atividades precisam ser atualizadas" e comentou que "*Lembrar que quando temos um ator criador de dados e outro ator que é o provedor, eles precisam entrar em contato para compartilhar informações descritivas sobre o dataset. Uma sugestão é adicionar a atividade de consulta entre esses dois papéis quando forem dois atores diferentes.*"

Participante 4 respondeu "Não, uma ou mais atividades precisam ser atualizadas" e comentou que "*Quando o provedor e o criador forem atores diferentes, o provedor deve ser incluído, também, na fase de planejamento.*"

Participante 3 e *Participante 4* responderam que "Sim, a fase está descrita corretamente" e não comentaram nada a respeito.

Já o *Participante 6* respondeu "Sim, a fase está descrita corretamente" e acrescentou que "*O provedor de dados deveria ser incluído entre os papéis envolvidos.*"

Nós concordamos com as melhorias propostas pelos participantes e incluímos o papel do *Provedor de Dados* na fase de *Planejamento* do DWLM.

6.2.3.4 Avaliação da Fase de Criação

De forma semelhante à fase de Planejamento, também avaliamos se a fase de Criação estava descrita corretamente e se suas atividades são viáveis, completas e adequadas. Com a pergunta, também pretendíamos detectar a existência de lacunas, erros e/ou possibilidades de melhorias. A Figura 29 resume as respostas.

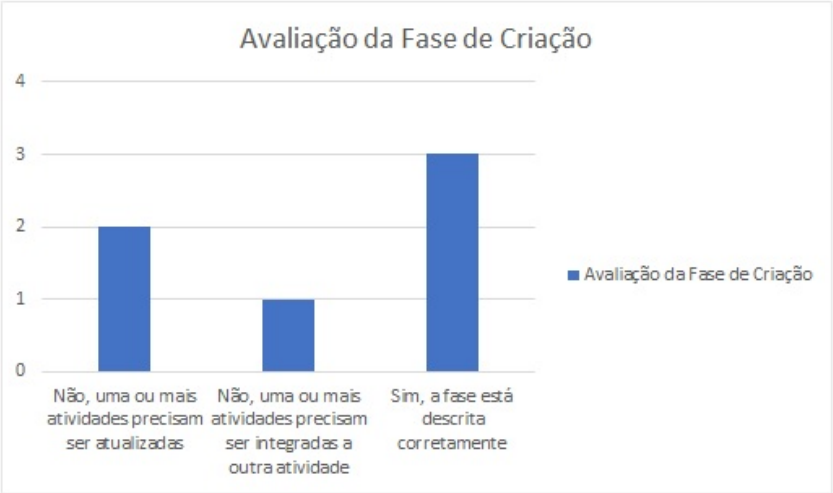

Figura 29 – Avaliação da Fase de Criação. Fonte: Autor

Participante 2 selecionou "Não, uma ou mais atividades precisam ser atualizadas" e comentou "*Na atividade de "Avaliar Qualidade" está faltando definir qual o papel que ficará responsável por isso.*"

Participante 3 selecionou "Não, uma ou mais atividades precisam ser integradas a outra atividade" e comentou "*Fundir a fase de avaliação da qualidade com validação. Para mim, a validação faz parte da avaliação de qualidade.*"

Participante 4 selecionou "Não, uma ou mais atividades precisam ser atualizadas" e comentou "*Na atividade de avaliação da qualidade, não foi descrito o ator responsável pela execução da atividade.*"

Os demais participantes selecionaram "Sim, a fase está descrita corretamente" e não deixaram nenhum comentário no questionário.

Em relação às alterações propostas pelos *Participantes 2 e 4*, nós concordamos que realmente ficou faltando essa definição do papel responsável por executar a Avaliação de Qualidade e o inserimos. Contudo, com relação à possibilidade de fundir a fase de Avaliação de Qualidade com a de Validação não faz sentido ao nosso ver. As duas fases

tem propósitos distintos, de tal forma que uma é responsável por determinar a qualidade do conjunto de dados de acordo com as métricas de qualidade estabelecidas, enquanto a outra tem como objetivo realizar a validação do conjunto por meio do *Provedor de Dados*, que irá verificar se o conjunto não possui inconsistência e se ele está apto a ser publicado.

6.2.3.5 Avaliação da Fase de Publicação

Assim como as fases anteriores, também perguntamos se a fase de *Publicação* estava descrita corretamente e se as suas atividades são viáveis, completas e adequadas. Além disso, pretendíamos identificar a existência de lacunas, erros e/ou possibilidades de melhorias. A Figura 30 resume as respostas.

Figura 30 – Avaliação da Fase de Publicação. Fonte: Autor

Participante 1 selecionou "Não, uma ou mais atividades precisam ser atualizadas" e comentou "*Substituir o termo "COMPLETO" por "AVANÇADO".*"

Os demais participantes selecionaram "Sim, a fase está descrita corretamente" e não informaram nenhum comentário no questionário.

Em relação à alteração proposta pelo *Participante 1*, nós concordamos com ele e realizamos a atualização na fase de *Publicação*.

6.2.3.6 Avaliação da Fase de Consumo

Em sequencia, perguntamos se a fase de *Consumo* estava descrita corretamente e se suas atividades são viáveis, completas e adequadas. Assim como as fases anteriores, também gostaríamos de identificar lacunas, erros e/ou possibilidades de melhorias. A Figura 31 resume as respostas.

Participante 1 selecionou "Não, uma ou mais atividades precisam ser atualizadas" e comentou "*Substituir USO COMPLETO por USO AVANÇADO.*"

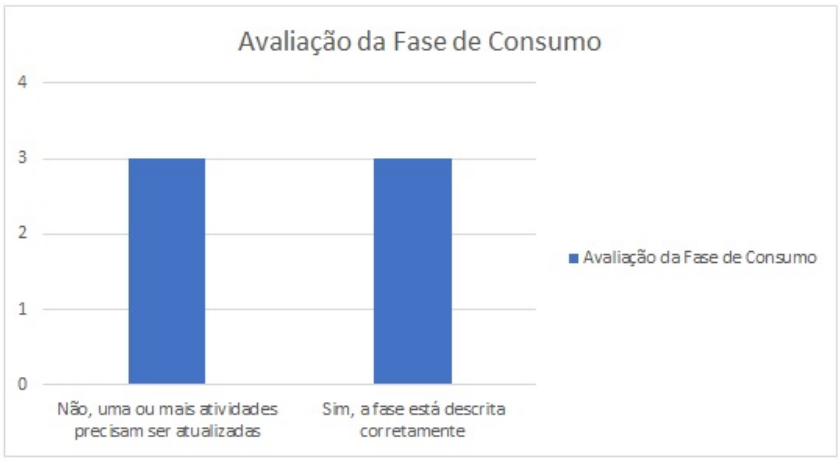

Figura 31 – Avaliação da Fase de Consumo. Fonte: Autor

Participante 2 selecionou "Sim, a fase está descrita corretamente" e acrescentou "*Foi da escolha do autor criar uma classificação de consumo. Talvez embasar mais as justificativas para a classificação e mostrar referências.*"

Participante 3 selecionou "Sim, a fase está descrita corretamente" e não informou nenhum comentário.

Participante 4 selecionou "Sim, a fase está descrita corretamente", mas pontuou que "*Para mim está descrito corretamente, mas vale pensar se realmente há necessidade de categorizar o tipo de uso do consumidor.*

Os *Participantes 5 e 6* selecionaram "Não, uma ou mais atividades precisam ser atualizadas" e também comentaram a respeito dos níveis de uso definidos em uma das atividades da fase de *Consumo*.

Em relação às melhorias propostas pelos participantes, nós concordamos em alguns pontos. Desse modo, analisamos os níveis de uso definidos na fase de *Consumo* e realizamos algumas alterações no sentido que, ao invés de classificarmos os níveis de uso, apenas citamos as possíveis formas de uso. Foi observado que esse tipo de informação vai depender muito de usuário para usuário e, após a avaliação, achamos melhor não haver essa classificação.

6.2.3.7 Avaliação da Fase de Refinamento

Seguindo a mesma ideia, na fase de *Refinamento* questionamos se a fase estava descrita corretamente e se suas atividades são viáveis, completas e adequadas. Além disso, pretendíamos identificar lacunas, erros e/ou possibilidades de melhorias. A Figura 32 resume as respostas.

Participante 2 selecionou "Não, uma ou mais atividades precisam ser atualizadas" e comentou "*Corrigir a descrição da atividade de "correção e enriquecer" para explicar que o usuário final pode alterar o dataset. A descrição de propôr melhorias e limpezas entra*

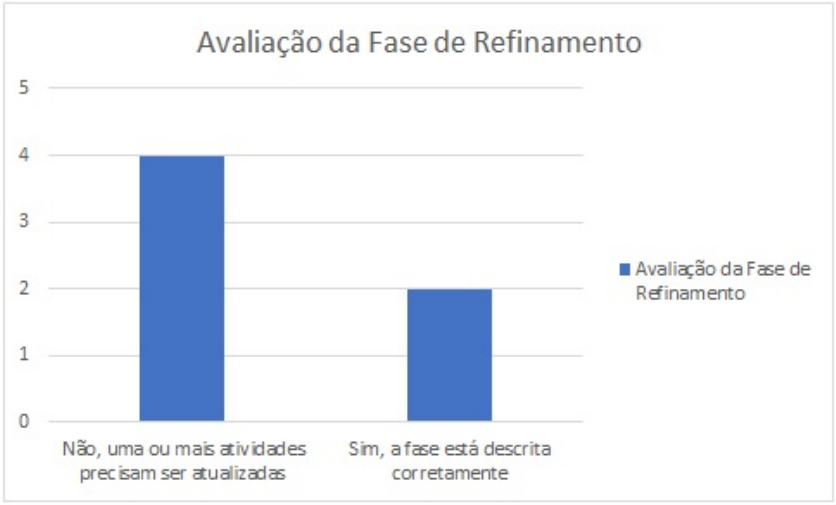

Figura 32 – Avaliação da Fase de Refinamento. Fonte: Autor

em feedback na fase de consumo. ”

Participante 4 selecionou “Não, uma ou mais atividades precisam ser atualizadas” e comentou “*No texto não ficou muito claro que o Usuário Final pode realizar melhorias e correções. É necessário explicar que o usuário final pode ser tanto um usuário interno como externo. E analisar se a atividade de propôr melhorias e limpezas não se encaixaria na atividade de prover e disponibilizar feedback.*”

Participante 5 selecionou “Não, uma ou mais atividades precisam ser atualizadas” e comentou que “*Acredito que o usuário tem um papel de sugerir melhorias, feedback da qualidade, utilidade, etc. Porém, não teria o papel de realizar as melhorias do ponto de vista de republicação.*”

Participante 6 selecionou “Não, uma ou mais atividades precisam ser atualizadas” e comentou “*Nem sempre será possível o usuário final realizar as alterações diretamente no conjunto de dados.*”

Os *Participantes 1 e 3* selecionaram “Sim, a fase está descrita corretamente” e não comentaram nada a respeito.

As dúvidas que surgiram entre os participantes nesta pergunta foi em relação à atividade de *F05A01 - Corrigir e enriquecer o conjunto de dados*, pois nesta atividade descrevemos que o *Usuário Final* poderá realizar correções e enriquecimento dos dados e enviá-lo para uma validação. Os participantes ficaram em dúvida se esses usuários finais poderiam realizar tal alteração ou apenas enviar as sugestões via *feedback*. Dessa forma, com base nas dúvidas levantadas pelos participantes, percebemos que o cenário mais comum é o *Usuário Final* enviar o *feedback* e, a partir dele, ser gerado um *Refinamento*. De modo que, a atividade de *F05A01 - Corrigir e enriquecer o conjunto de dados*, seja realizada pelo *Criador do Conjunto de Dados* embasado pelo *feedback* do *Usuário Final* ou pela detecção de alguma erro e/ou possibilidade de enriquecimento encontrada no conjunto de dados. Assim, melhoramos a descrição da fase de *Refinamento* para incorporar esses

ajustes.

6.2.3.8 Avaliação da Fase de Arquivamento

Para finalizar as perguntas em relação às fases do DWLM, questionamos se a fase de *Arquivamento* estava descrita corretamente e se suas atividades são viáveis, completas e adequadas. Como já mencionado, com esta pergunta pretendíamos identificar lacunas, erros e/ou possibilidades de melhorias. A Figura 33 resume as respostas.

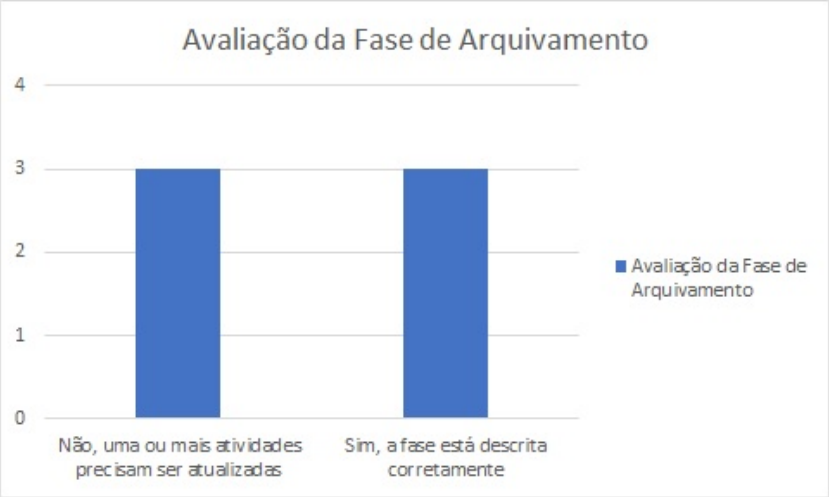

Figura 33 – Avaliação da Fase de Arquivamento. Fonte: Autor

Participante 2 selecionou "Não, uma ou mais atividades precisam ser atualizadas" e comentou "*Está faltando um link entre uma fase anterior e a fase de arquivamento. Uma sugestão é adicionar o papel de provedor na fase de arquivamento para que ele entre com a atividade de geração de documento de solicitação de arquivamento do conjunto.*"

Participante 4 selecionou "Não, uma ou mais atividades precisam ser atualizadas" e comentou que "*Alterar o trecho "... ele poderá solicitar uma cópia" por "... ele poderá solicitar uma cópia, se possível". Além disso, informar ao consumidor o motivo pelo qual foi removido o acesso. Também, acrescentar o papel do provedor de dados na fase de arquivamento.*"

Participante 5 selecionou "Não, uma ou mais atividades precisam ser atualizadas" e comentou "*Apenas ao informar que o dado não está mais disponível, explicar também o motivo da remoção para o usuário. Não será em todos os casos que será possível obter uma cópia do dado removido.*"

O *Participante 1, 3 e 6* selecionaram "Sim, a fase está descrita corretamente" e não acrescentaram nada a respeito da pergunta.

Nós concordamos com a sugestão de adicionar o papel do *Provedor de Dados* na fase de *Arquivamento* e realizamos tais alterações no modelo. Também aprovamos a ideia do *Documento de Solicitação de Arquivamento* do conjunto ser um *link* entre a fase anterior

e a de *Arquivamento*. As demais sugestões em relação à descrição da fase e atividades também foram aceitas e incorporadas ao modelo. Como dito na seção 6.2.3.1, o nome dessa fase no DWLM foi alterado para "*Remoção*", e consequentemente o nome do documento de solicitação foi alterado para *Documento de Solicitação de Remoção do Conjunto de dados*.

6.2.3.9 Avaliação do Conjunto de Características

Por fim, questionamos se o conjunto de características estava descrito corretamente e se elas são adequadas e suficientes para classificar os modelos de ciclo de vida criados a partir do DWLM. Assim como em todas as perguntas anteriores, com essa questão gostaríamos de identificar lacunas, erros e/ou possibilidades de melhorias. A Figura 34 resume as respostas.

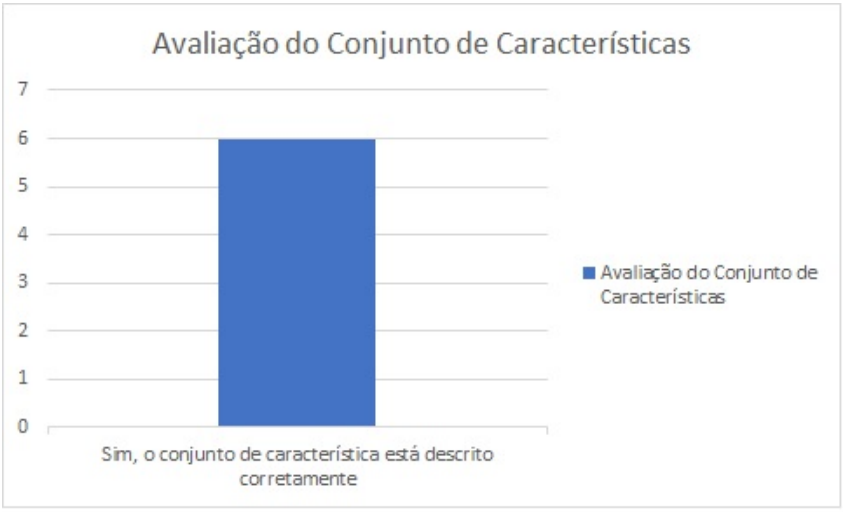

Figura 34 – Avaliação do Conjunto de Características. Fonte: Autor

Para essa pergunta, todos os participantes selecionaram que "Sim, o conjunto de característica está descrito corretamente" e não acrescentaram nenhum comentário a respeito.

6.2.3.10 Perguntas Gerais

A última fase do Grupo Focal foi composta por três perguntas discursivas. Nelas pedíamos numa escala de 1 a 5, na qual 1 significa "Discordo Fortemente" e 5 significa "Concordo Fortemente", como os participantes avaliavam as seguintes afirmações:

1. O DWLM é adequado para orientar e apoiar o ciclo de vida dos dados na Web?
2. A estrutura do DWLM é adaptável?
3. A estrutura do DWLM é flexível?

Quadro 9 – Grupo Focal - Perguntas Gerais sobre o DWLM

	Adequado para orientar e apoiar	**DWLM é adaptável**	**DWLM é flexível**
Participante 1	5	4	4
Participante 2	5	5	5
Participante 3	5	5	3
Participante 4	5	5	5
Participante 5	5	5	5
Participante 6	5	5	5
Média	**5**	**4,83**	**4,5**

A resposta de cada participante a essas perguntas foram apresentadas na Tabela 9.

Participante 1 comentou *"Achei ele bem abrangente, me vejo conseguindo desenvolver algo me enquadrando nas fases, tanto em uma instituição pequena quanto em uma grande porte."*

Participante 2 comentou *"Concordo fortemente com os 3 tópicos acima. Sugestões foram dadas nas discussões anteriores."*

Participante 3 comentou *"Existem soluções disponíveis no mercado que podem influenciar na flexibilidade do modelo. O modelo pode ser definitivo no momento de escolher um software de publicação de dados."*

Participante 4 comentou *"Acredito que o modelo proposto está bem abrangente e as descrições das atividades estão bastante fáceis de serem interpretadas. Desta forma, concordo fortemente com as 3 perguntas acima."*

Os *Participantes 5 e 6* não realizaram comentários em relação as perguntas.

6.3 CONSIDERAÇÕES FINAIS

Esta pesquisa tinha como objetivo avaliar o DWLM a nível de viabilidade, integridade e adequação. No geral, os participantes se mostraram satisfeitos com o modelo proposto e relataram a importância dele para o contexto de publicação e consumo de dados na Web. Mas, também foram sugeridos alguns pontos de melhorias. Assim, como um resultado dessa avaliação, buscamos evoluir nosso modelo para incluir as sugestões mais importantes apontadas pelos participantes. Esta nova versão do DWLM foi apresentada no Capítulo 3.

Embora a maioria das sugestões tenham sido incluídas, algumas foram registradas para serem consideradas posteriormente e outras julgamos como não necessárias.

7 CONCLUSÃO

Neste capítulo descrevemos as considerações finais sobre esta dissertação. Apresentamos as contribuições de pesquisas alcançadas (Seção 7.1) e as diretrizes de trabalhos futuros (Seção 7.3).

7.1 CONSIDERAÇÕES FINAIS

Neste trabalho, propomos um Modelo de Ciclo de Vida de Dados na Web. O modelo proposto se mostra relevante para o contexto de publicação e consumo de dados na Web, uma vez que apresenta uma estrutura para guiar e apoiar o usuário em todo o processo de evolução dos dados na Web. Além disso, ele descreve fases, atividades, entradas, saídas, características e incorpora as boas práticas para Dados na Web em todas as suas fases.

Inicialmente, foi apresentada uma visão geral do que são Dados na Web, transcorrendo por temas como Ecossistema de Dados, Dados Abertos e Dados Conectados. Em seguida, apresentamos outro conceito muito importante para essa dissertação, o Ciclo de Vida. Neste capítulo, dissertamos sobre o que são ciclos de vida e as áreas que mais tem utilizado esse conceito. Além disso, também apresentamos alguns Modelos de Ciclo de Vida de Desenvolvimento de Software e Modelos de Ciclo de Vida de Dados. O resultado dessa análise de modelos de ciclo de vida apontou um défice na área de Dados na Web, visto que não se tinha nada bem fundamento que abordasse tal domínio.

Desse modo, propomos um Modelo de Ciclo de Vida de Dados na Web que tem como intuito estabelecer um canal de comunicação entre os atores participantes desse ciclo. Ademais, ele visa apoiar a aplicação das melhores práticas e oferecer uma abstração de alto nível de todo o processo de publicação e consumo de dados, desde a concepção dos dados a sua remoção. O modelo proposto, intitulado DWLM é composto por 5 papéis, 6 fases e 16 atividades.

Com o objetivo de colocar em prática o modelo proposto, descrevemos sua aplicação para o conjunto de dados do Censo 2017 da UFPE. A publicação desse conjunto de dados foi um trabalho em parceria do Laboratório de Dados e Informação (aLADIN) com a Pró-Reitoria de Planejamento, Orçamento e Finanças (PROPLAN).

Para avaliarmos o DWLM realizamos um grupo focal que contou com 6 participantes da área de Dados na Web e que tinham experiencia na publicação e consumo de dados. Esse grupo focal foi realizado no Centro de Informática da Universidade Federal de Pernambuco e sua condução foi dividade em três etapas. Na primeira etapa foi solicitado que os participantes respondessem ao questionário a respeito dos seus dados acadêmicos e profissionais. A segunda etapa constituiu na aplicação de um questionário contendo 9 perguntas a fim de avaliar os papéis, fases e características do DWLM. Na terceira e

última etapa foram feitas três perguntas gerais com o intuito de avaliar se o modelo é adequado, flexível e adaptável.

Por meio da avaliação, coletamos evidências sobre a importância do Modelo de Ciclo de Vida de Dados na Web, uma vez que foi possível obter bons resultados e pontos de melhorias bem produtivos. Após essa avaliação o modelo foi refinado e algumas das melhorias propostas pelos participantes foram incorporadas.

Por fim, vale salientar que o seguinte artigo foi publicado a partir dos resultados intermediários obtidos durante o desenvolvimento desta dissertação:

- Santos, H.D.; Oliveira, I.S.; Lima, G.F.; Silva, K.M.; Cruz, R.V.; Lóscio, B.F. (2018) **Investigations into data published and consumed on the Web: A Systematic Mapping Study**. Journal of the Brazilian Computer Society, v. 24, n. 1, p. 14.

7.2 LIMITAÇÕES

Durante o estudo algumas limitações puderam ser observadas, mesmo com o esforço de mitigação provida pelos pesquisadores. Primeiramente, não foi fácil encontrar profissionais e pesquisadores para participarem da avaliação do modelo. Infelizmente, todos os participantes tinham alguma familiaridade com Dados Abertos, o que pode haver um certo viés na formação de opinião de cada um deles. Além disso, conseguir conciliar as agendas dos participantes foi algo muito complicado e, consequentemente, não foi possível criar uma grande amostra representativa da população.

Em relação ao método de pesquisa, as limitações são típicas de estudos empíricos, particularmente na generalização dos resultados. A extração e análise de dados pode ser influenciada pelas opiniões pessoais do pesquisador que executa o processo. Para mitigar essa ameaça, dedicamos tempo adequado para executar vários ciclos de avaliação de refinamento para construir o modelo proposto. Outra ação de mitigação consistiu em consultar frequentemente outros pesquisadores do grupo de pesquisa ALADIN para abordar e resolver conflitos que acompanham o progresso da pesquisa.

Com relação ao modelo, mais estudos devem ser realizados em diferentes domínios de dados, a fim de verificar a adequação e completude dele em diferentes contextos. Além disso, por conta do tempo, o uso do modelo no Capítulo 6 se deu de forma descritiva. Desse modo, é interessante realizar a sua aplicação em novos cenários de forma prescritiva para observar o comportamento do DWLM nesses casos.

7.3 TRABALHOS FUTUROS

Como trabalhos futuros, identificamos as seguintes questões:

- **Aplicar o modelo em outros cenários**: Para ter uma melhor visão da importância do modelo proposto é interessante realizar o sua aplicação em outros cenários. Visto que, nesses cenários podem aparecer situações específicas que não foram pensadas durante a sua construção.
- **Instanciar o DWLM para outros domínios**: Como o DWLM foi criado visando ser o mais genérico possível, um trabalho futuro interessante seria realizar uma instanciação do DWLM para outros domínios centrados em dados como, Dados Conectados, Dados de Pesquisa, Dados Privados, detre outros.
- **Desenvolver uma Ferramenta de Acompanhamento do Ciclo de Vida dos Dados**: Um trabalho muito importante é o desenvolvimento de uma ferramenta que possa auxiliar no acompanhamento de toda a evolução do ciclo de vida de dados na Web. A partir dessa ferramenta seria possível haver tanto um repositório contendo todos os artefatos gerados em cada fase do ciclo, quanto um histórico de todos os estágios ao qual o conjunto de dados passou ao longo de sua vida na Web.
- **Replicar a avaliação**: A replicação do grupo focal com outros grupos distintos de participantes permitirá reunir novas evidências. Essa evidências poderão ser usadas para realizar novas melhorias no modelo, quanto para o seu amadurecimento.

REFERÊNCIAS

ALSHAMMARI, M.; SIMPSON, A. J. R. Personal data management for privacy engineering: An abstract personal data lifecycle model. 2017.

ANDERSEN, A. B.; GÜR, N.; HOSE, K.; JAKOBSEN, K. A.; PEDERSEN, T. B. Publishing danish agricultural government data as semantic web data. In: SPRINGER. *Joint International Semantic Technology Conference*. [S.l.], 2014. v. 8943, p. 178–186.

APOSTEL, L. Towards the formal study of models in the non-formal sciences. *International Journal for Epistemology, Methodology and Philosophy of Science*, v. 12, p. 125–161, 1960. ISSN 0039-7857.

ASKHAM, N.; COOK, D.; DOYLE, M.; FEREDAY, H.; GIBSON, M.; LANDBECK, U.; LEE, R.; MAYNARD, C.; PALMER, G.; SCHWARZENBACH, J. The six primary dimensions for data quality assessment. *Semantic Web*, 2013.

ATTARD, J.; ORLANDI, F.; SCERRI, S.; AUER, S. A systematic review of open government data initiatives. *Government Information Quarterly*, Elsevier, v. 32, n. 4, p. 399–418, 2015.

BERNERS-LEE, T. *Linked Data*. 2006. <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em 26 de novembro de 2018.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, p. 205–227, 2009.

BOEHM, B. A spiral model of software development and enhancement. *Computer*, ieee, v. 21, p. 61–72, 1988. ISSN 1558-0814.

BOUQUET, P.; STOERMER, H. Web of data and web of entities: Identity and reference in interlinked data in the semantic web. *Philosophy & Technology*, v. 25, p. 5–26, 2012.

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.

CATTEAU, O.; VIDAL, P.; BROISIN, J. A generic representation allowing for expression of learning object and metadata lifecycle. In: *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*. Kerkrade, Netherlands: [s.n.], 2006. p. 30–32. ISSN 2161-3761.

CHEN, Y.; CHEN, S.; LIN, S. C. A metadata lifecycle model for digital libraries : methodology and application for an evidence-based approach to library research. In: *World Library and Information Congress: 69th IFLA General Conference and Council*. [S.l.: s.n.], 2003.

CHIARA, I. G. D. Grupo de foco. In: *Métodos qualitativos de pesquisa em Ciência da Informação*. São Paulo, Brasil: Polis, 2005. p. 101–117. ISBN 85-7228-021-9.

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.

CRESWELL, J. W. Projeto de pesquisa métodos qualitativo, quantitativo e misto. In: *Projeto de pesquisa métodos qualitativo, quantitativo e misto*. Porto Alegre: Artmed, 2010.

DERILINX, D. L.; LÓSCIO, B. F.; ARCHER, P. W3C Working Group Note, *Data on the Web Best Practices Use Cases Requirements*. 2015. <https://www.w3.org/TR/dwbp-ucr/>. Acessado em 18 de dezembro de 2018.

DIETRICH, D.; GRAY, J.; MCNAMARA, T.; POIKOLA, A.; POLLOCK, R.; TAIT, J.; ZIJLSTRA, T. *Open Data Handbook*. [S.l.]: Open Knowledge Foundation, 2009. <http://opendatahandbook.org/guide/en/>. Acesso em 26 de novembro de 2018.

DOORBAR, J. The papillomavirus life cycle. *Journal of Clinical Virology*, v. 32, p. 7 – 15, 2005. ISSN 1386-6532. Supplement: Human Papillomaviruses. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1386653204003671>.

EASTERBROOK, S.; SINGER, J.; STOREY, M.-A.; DAMIAN, D. Selecting empirical methods for software engineering research. Springer, p. 285–311, 2008.

FERNáNDEZ, J.; MARTíNEZ-PRIETO, M.; GUTIéRREZ, C. Publishing open statistical data: the spanish census. In: *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*. New York, NY, USA: ACM, 2011. p. 20–25. Disponível em: <http://doi.acm.org/10.1145/2037556.2037560>.

FERREIRA, J.; MIRANDA, M.; ABELHA, A.; MACHADO, A. O processo etl em sistemas data warehouse. *INForum 2010 - II Simpósio de Informática*, p. 757–765, 09 2010.

GAMA, K.; LÓSCIO, B. F. Towards ecosystems based on open data as a service. In: *ICEIS*. [S.l.: s.n.], 2014. p. 659–664.

GOUVEIA, A. O conceito de modelo e sua utilização nas ciências do comportamento: breves novas introdutórias. *Rev. Estudos de Psicologia*, v. 6, p. 13–16, 1999.

HIGGINS, S. The dcc curation lifecycle model. *The International Journal of Digital Curation*, v. 3, 2008.

HUMPHREY, C. e-science and the life cycle of research. *The International Journal of Digital Curation*, 2006.

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

ISO/IEC/IEEE International Standard - Systems and software engineering – Software life cycle processes. *ISO/IEC/IEEE 12207:2017(E) First edition 2017-11*, p. 1–157, Nov 2017.

ISOTANI, S.; BITTENCOURT, I. I. *Dados Abertos Conectados: Em busca da Web do Conhecimento*. [S.l.]: Novatec Editora, 2015.

KEELE, S. et al. Guidelines for performing systematic literature reviews in software engineering. In: *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. [S.l.]: sn, 2007.

KLYNE, G.; CARROLL, J. J. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. 2004. <https://www.w3.org/TR/rdf-concepts/>. Acesso em 26 de novembro de 2018.

KONTIO, J.; LEHTOLA, L.; BRAGGE, J. Using the focus group method in software engineering: obtaining practitioner and user experiences. In: IEEE. *Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on*. [S.l.], 2004. p. 271–280.

KOSCH, H.; BOSZORMENYI, L.; DOLLER, M.; LIBSIE, M.; SCHOJER, P.; KOFLER, A. The life cycle of multimedia metadata. *IEEE MultiMedia*, v. 12, p. 80–86, 2005.

LESTER, D. L.; PARNELL, J. A.; CARRAHER, S. Organizational life cycle: A five-stage empirical scale. *The International Journal of Organizational Analysis*, v. 11, p. 339–354, 2003. Disponível em: <https://www.emeraldinsight.com/doi/pdfplus/10.1108/eb028979>.

LÓSCIO, B.; GUIMARÃES, C.; CALEGARI, N. Data on the web best practices: Challenges and benefits. *Open Data Reserach Symposium (ODRS 2016)*, 2016.

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 2018.

LÓSCIO, B. F.; BURLE, C.; IURY, M. S.; CALEGARI, N. C. *Fundamentos para publicação de dados na Web*. [S.l.]: Comitê Gestor da Internet no Brasil CGI.br, 2018. ISBN 978-85-5559-072-6.

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>.

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.

MARCONI, M. d. A.; LAKATOS, E. M. Fundamentos de metodologia cientifica. In: *Fundamentos de metodologia cientifica*. [S.l.]: Atlas, 2010.

MARTIN, J. *Rapid Application Development*. Indianapolis, IN, USA: Macmillan Publishing Co., Inc., 1991. ISBN 0-02-376775-8.

MESCHANKINA, I. *The Software Development Life Cycle: Phases And Methodologies*. 2018. Disponível em: <https://producttribe.com/project-management/sdlc-guide>.

MILLS, H. D. Incremental software development. *IBM Systems Journal*, 1980.

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>.

NATH, K.; DHAR, S.; BASISHTHA, S. Web 1.0 to web 3.0 - evolution of the web and its various challenges. In: *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*. [S.l.: s.n.], 2014. p. 86–89.

NECASKỲ, M.; CHLAPEK, D.; KLiMEK, J.; KUCERA, J.; MAURINO, A.; RULA, A.; KONECNY, M.; VANOVA, L. 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>.

NETO, O. C.; MOREIRA, M. R.; SUCENA, L. F. M. Grupos focais e pesquisa social qualitativa: o debate orientado como técnica de investigação. *XIII Encontro da ABEP*, 2002.

OLIVEIRA, L. E. R.; OLIVEIRA, M. I. S.; SANTOS, W. C. d. R.; LÓSCIO, B. F. Data on the web management system: a reference model. In: ACM. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. [S.l.], 2018. p. 2.

OLIVEIRA, L. E. R. d. A. *Um Modelo de Arquitetura para Sistemas Gerenciadores de Dados na Web*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Centro de Informática, Curso de Pós-Graduação em Ciências da Computação, 2017.

RADULOVIC, F.; POVEDA-VILLALóN, M.; VILA-SUERO, D.; RODRíGUEZ-DONCEL, V.; GARCíA-CASTRO, R.; GóMEZ-PéREZ, A. Guidelines for linked data generation and publication: An example in building energy consumption. *Automation in Construction*, v. 57, p. 178 – 187, 2015. ISSN 0926-5805. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0926580515000801>.

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

RAO, S. S.; NAYAK, A. Linked: A novel methodology for publishing linked enterprise data. *Journal of computing and information technology*, v. 25, n. 3, p. 191–209, 2017.

ROYCE, W. W. Managing the development of large software systems: concepts and techniques. *Proc. IEEE WESTCON, Los Angeles*, Aug 1970.

SANTOS, H. *Uma Estratégia para o Refinamento de Dados na Web Baseada em Social Coding*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Centro de Informática, Curso de Pós-Graduação em Ciências da Computação, Recife, 2018.

SANTOS, H. D. A. d.; OLIVEIRA, M. I. S.; LIMA, G. d. F. A. B.; SILVA, K. M. da; MUNIZ, R. I. V. C. S.; LÓSCIO, B. F. Investigations into data published and consumed on the web: A systematic mapping study. *Journal of the Brazilian Computer Society*, v. 24, n. 1, p. 14, 2018. Disponível em: <https://doi.org/10.1186/s13173-018-0077-z>.

SAYAO, L. F. Modelos teóricos em ciências da informação - abstração e método científico. *Ci. Inf.*, v. 30, p. 82–92, 1999. ISSN 0100-1965.

SOMMERVILLE, I. *Software Engineering*. 9. ed. Harlow, England: [s.n.], 2011. ISBN 978-0-13-703515-1.

SORRENTINO, S.; BERGAMASCHI, S.; FUSARI, E.; BENEVENTANO, D. Semantic annotation and publication of linked open data. In: SPRINGER. *International Conference on Computational Science and Its Applications*. [S.l.], 2013. p. 462–474.

TAUBERER, J.; LARRY, L. *The 8 Principles of Open Government Data*. 2007. <https://opengovdata.org/>. Acesso em 13 de dezembro de 2018.

UFPE, U. F. de P. *Plano de Dados Abertos*. 2017. <https://www.ufpe.br/documents/38982/806616/PDA+-+UFPE.pdf/bcaae838-22dd-42fd-b0da-37a0d4061936>. Acesso em 28 de Janeiro de 2019.

UFPE, U. F. de P. *Portal de Dados Abertos da UFPE*. 2018. <https://dados.ufpe.br>. Acesso em 13 de fevereiro de 2019.

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.

UREN, V.; CIMIANO, P.; IRIA, J.; HANDSCHUH, S.; VARGAS-VERA, M.; MOTTA, E.; CIRAVEGNA, F. 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.

WALTERS, G. Rocurando nem0. 2013.

ZAGANELLI, B. M.; NISENBAUM, M. A.; MARQUES, S. B.; OLINTO, G. O grupo focal na ciência da informação. *Informação & Sociedade: Estudos*, p. 37–47, 2015.

ZAVERI, A.; RULA, A.; MAURINO, A.; PIETROBON, R.; LEHMANN, J.; AUER, S. Quality assessment for linked data: A survey. *Semantic Web*, p. 63–93, 2012.

ZENGENENE, D.; CASAROSA, V.; MEGHINI, C. Towards a methodology for publishing library linked data. In: CATARCI, T.; FERRO, N.; POGGI, A. (Ed.). *Bridging Between Cultural Heritage Institutions*. Berlin: Springer, 2013. p. 81–92.

ZHAO, J. Publishing chinese medicine knowledge as linked data on the web. *Chinese Medicine*, v. 5, n. 1, p. 27, 2010. Disponível em: <https://doi.org/10.1186/1749-8546-5-27>.

ZHONGJIE, W.; XIAOFEI, X. Svlc: Service value life cycle model. In: *2009 IEEE International Conference on Cloud Computing*. [S.l.: s.n.], 2009. p. 159–166. ISSN 2159-6182.

APÊNDICE A – *TEMPLATE* DOCUMENTO DE DESCRIÇÃO DO CONJUNTO DE DADOS

DADOS DE DESCRIÇÃO DO CONJUNTO DE DADOS

1. Título: _____

2. Descrição: _____

_____

3. Palavras Chaves: _____

4. Domínio: _____

5. Cobertura Temporal: _____ 6. Cobertura Espacial: _____

7. Linguagem: _____

8. Formato de Data e Hora: _____

9. Frequência de Atualização: _____

10. Licença dos Dados: _____

11. Provedor dos Dados: _____

12. Publicador dos Dados: _____

13. Dados Abertos: ( ) Sim ( ) Não

14. Dados Conectados: ( ) Sim ( ) Não
Se sim, Vocabulários: _____

15. Formatos de distribuição: ( ) CSV ( ) JSON ( ) RDF ( ) Outro _____

16. Lista de Metadados Estruturais:

| TÍTULO | CAMPO | DESCRIÇÃO | TIPO |

APÊNDICE B – *TEMPLATE* DOCUMENTO DE INCONSISTÊNCIA DE DADOS

DADOS DE IDENTIFICAÇÃO DO CONJUNTO DE DADOS

Título do Conjunto de Dados: _____

Provedor: _____

Publicador: _____

DADOS INCONSISTENTES		
TITULO	VALOR INCORRETO	VALOR CORRETO

Outras Correções: _____

_____

_____

(Assinatura do Provedor)

APÊNDICE C – *TEMPLATE* TERMO DE CONSENTIMENTO DE PUBLICAÇÃO DE DADOS

DADOS DE IDENTIFICAÇÃO DO CONJUNTO DE DADOS

Título do Conjunto de Dados: _____
Provedor: _____
Publicador: _____

Eu _____ declaro estar ciente da publicação do conjunto de dados intitulado como _____ no (informar local de publicação) e concordo com a sua publicação.

_____

(Assinatura do Provedor)

APÊNDICE D – *TEMPLATE* DOCUMENTO DE SOLICITAÇÃO DE REMOÇÃO DO CONJUNTO DE DADOS

DADOS DE IDENTIFICAÇÃO DO CONJUNTO DE DADOS

Título do Conjunto de Dados: _____
Provedor: _____
Publicador: _____
URI: _____

Eu, _____ solicito a remoção do acesso ao conjunto de dados, porque _____ (informar motivo).

_____

(Assinatura do Provedor)