UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS GRADUAÇÃO EM MODELAGEM COMPUTACIONAL DO CONHECIMENTO

THIAGO JOSÉ TAVARES ÁVILA

Uma proposta de modelo de processo para publicação de Dados Abertos Conectados Governamentais

Maceió
2015

THIAGO JOSÉ TAVARES ÁVILA

Uma proposta de modelo de processo para publicação de Dados Abertos Conectados Governamentais

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Modelagem Computacional do Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Prof. Dr. Ig Ibert Bittencourt Santana Pinto
Coorientador: Profª. Drª. Patrícia Leone Espinheira Ospina

Maceió
2015

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária Responsável: Helena Cristina Pimentel do Vale

A958u Ávila, Thiago José Tavares.
Uma proposta de modelo de processo para publicação de dados abertos conectados governamentais / Thiago José Tavares Ávila. - 2015.
220 f. : il.

Orientador: Ig Ibert Bittencourt Santana Pinto.
Coorientadora: Patrícia Leone Espinheira Ospina.
Dissertação (mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2015.

Bibliografia: f. 209-215.
Apêndices: 216-220.

1. Modelos computacionais. 2. Dados governamentais. 3. Dados abertos. 3. Dados abertos conectados. 4. Dados abertos conectados governamentais. 5. Modelo de processo. I. Título.

CDU: 004.028

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Modelagem Computacional de Conhecimento
Avenida Lourival Melo Mota, Km 14, Bloco 09, Cidade Universitária
CEP 57.072-900 – Maceió – AL – Brasil
Telefone: (082) 3214-1364

Membros da Comissão Julgadora da Dissertação de Mestrado de Thiago José Tavares Ávila, intitulada: "Uma proposta de modelo de processo para publicação de Dados Abertos Conectados Governamentais", apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas, em 25 de novembro de 2015, às 09h00min, no miniauditório do Instituto de Computação.

COMISSÃO JULGADORA

Prof. Dr. Ig Ibert Bittencourt Santana Pinto
UFAL – Instituto de Computação
Orientador

Profa. Dra. Patrícia Espinheira Leone Ospina
UFPE – Departamento de Estatística
Coorientadora

Prof. Dr. Marcus de Melo Braga
UFAL – Instituto de Computação
Examinador

Profa. Dra. Bernadette Farias Lóscio
UFPE – Centro de Informática
Examinadora

Maceió, novembro de 2015.

À Deus e as cinco Marias (mãe, esposa e filhas) que estão sempre ao meu lado, motivando-me a ser uma pessoa melhor.

AGRADECIMENTOS

Agradeço acima de tudo à Deus, por me proporcionar tudo o que tenho e especialmente, ter uma família amada e conviver com pessoas muito especiais em minha vida.

A minha amada mãe, Maria de Fátima, que dedicou a sua vida para me educar e cuidar de mim em todos os momentos, com muito amor e o carinho especial que somente ela me concede. Te amo mãe. Esta conquista é acima de tudo, sua.

As quatro estrelas da minha vida, meus grandes amores Maria Luiza, Laís Maria e Maria Laura (filhas), e Mariana (esposa), por tornarem minha vida mais feliz, gratificante e desafiadora, por me acompanharem nos meus caminhos com muito amor, paciência e dedicação especialmente nesta reta final do Mestrado.

Aos meus quatro irmãos Janvitor e Ernani Ávila (maternos) e Marcílio e Cristiane Ávila (paternos). Amo muito vocês !!!

Ao meu Pai Marcílio *(in memorian)* e ao meu Pai(drasto) Antônio Soares, por contribuirem com minha formação e educação em momentos distintos, mas de forma especial e relevante para a minha vida pessoal.

Ao professor, orientador e amigo, Ig Ibert Bittencourt, pelos ensinamentos de valores e conhecimentos especiais e por desafiar-me à navegar em "mares desconhecidos", em busca do saber e do desenvolvimento profissional e pessoal, bem como pela dedicada orientação que me inspirou no desenvolvimento desta pesquisa e na realização do que era há pouco, um sonho. Para você, minha especial gratidão pela sua liderança e referência e por tudo foi edificado durante este mestrado e as sementes plantadas que certamente serão colhidas após sua conclusão.

Aos companheiros do Núcleo de Excelência em Tecnologias Sociais (NEES) Armando, Judson, Williams, Danila e André, pelas contribuições feitas para esta pesquisa, pela amizade e por todo o aprendizado e vivências compartilhadas sobre nossa área principal de pesquisa (Dados Abertos Conectados), empreendedorismo e muitos outros assuntos.

A professora e co-orientadora Patrícia Ospina, pela dedicação, conhecimento compartilhado e condução destacada em momentos muito relevantes desta pesquisa.

Ao professor Seiji Isotani, pela referência como pesquisador e figura humana que enriquece o nosso NEES sendo, especialmente para mim, um exemplo a ser seguido para os que desejam trilhar os caminhos da vida acadêmica.

Aos colegas pesquisadores Diego Demerval, Ranilson Paiva e Thyago Tenório pela parceria e contribuições fundamentais para esta pesquisa, seja no domínio da metodologia da pesquisa científica, do idioma inglês e das dicas e orientações com o LaTeX

Para colega Inés Mária un agradecimiento especial en español para todo el conocimiento, las discusiones y la amistad compartida a través de nuestras investigaciónes.

Para todos os professores e demais integrantes do NEES por todo o aprendizado con-

junto nesta jornada, em especial ao Denys, Josmário e ao João por contribuições diretas com esta pesquisa.

Para todos os professores, pesquisadores e profissionais que compartilharam seu conhecimento e/ou seu tempo para contribuir com esta pesquisa de alguma maneira.

A professora Bernadette Lóscio e ao professor Marcus Braga pela distinção em participarem das minhas bancas de qualificação e defesa do mestrado compartilhando seu tempo, dedicação e conhecimentos com o aprimoramento desta pesquisa.

Para os professores do Programa de Pós-Graduação em Modelagem Computacional do Conhecimento do Instituto de Computação (IC) da UFAL, especialmente ao Aydano Machado, Alan Pedro, Alejandro Frery, Evandro Costa e Heitor Ramos (e ao meu orientador Ig Ibert) pelos relevantes ensinamentos ao longo desta pós-graduação.

Para todos os colegas da turma 2014.1 do MMCC, em especial aos "Nobres Guerreiros" André, Antônio Carlos, Antônio Marcos, Carlos, Diego, Fábio, Helenilson, Hugo, Paulo Henrique, Pedro, Vilker, Vitor e Wilson por tornar este período de mestrado menos tenso e mais divertido (*'cause the zueira never ends.*)

Aos professores Olival Gusmão e Marcus Braga (novamente) pela orientação e amizade estabelecida desde a minha graduação em Ciência da Computação, bem como, junto com o amigo Victor Heuer e os demais membros do GCI-UFAL, por me abrirem as portas para o início da minha caminhada na pesquisa científica.

Aos demais professores, técnicos administrativos e colaboradores do IC/UFAL.

A professora Silvana Calheiros (IGDEMA) pela amizade, incentivo e orientação em momentos ímpares da minha vida pessoal, profissional e acadêmica.

Ao professor e referência como servidor público e pessoa humana, José Cândido do Nascimento, por toda a amizade, conhecimento compartilhado, e pelo especial incentivo (e cobranças) para que eu voltasse ao mundo acadêmico o quanto antes.

Para todos os companheiros da Superintendência de Produção da Informação do Conhecimento (SINC) da SEPLANDE/AL e SEPLAG/AL pelo apoio nesta empreitada, em especial aos estimados Robson Brandão, Teônia Amorim e Gardênia de Castro pelo incentivo direto durante esta jornada. Meus agradecimentos aos gestores destas instituições pelas condições para que houvesse a conciliação das minhas atuais atividades profissionais com as atividades acadêmicas.

Ao Time de Dados Abertos do Ministério do Planejamento, Orçamento e Gestão, por desenvolverem a iniciativa brasileira de Dados Abertos e por toda a atenção que me foi dispensada ao longo da minha formação nesta temática.

Para toda a comunidade de Dados Abertos (Conectados) (Governamentais) do Brasil, especialmente os integrantes da *Open Knowledge Brasil* e Lista INDA-BR.

À UFAL, pelo seu papel fundamental na formação de pesquisadores e desenvolvimento da ciência, tecnologia e inovação em Alagoas.

In God we trust,
all others must bring data.
(Edward Deming)

RESUMO

Nos últimos anos, governos tem sido estimulados e obrigados a ampliar a disponibilidade de seus dados na *Web* mediante obrigações legais bem como alianças internacionais como a *"Parceria para o Governo Aberto"*. A disseminação do conceito de dados abertos governamentais está contribuindo para uma ampliação desta oferta de dados na *Web*, entretanto sem garantias que tais dados estejam adequados para reuso, processamento automatizado, bem como subsidiar a geração de conhecimento. O conceito de dados conectados, apresentado em 2006 por Tim Berners-Lee, vem ao encontro desta necessidade de aprimorar a qualidade dos dados, mediante a adoção de princípios para estruturar e conectar conjuntos de dados na *Web*. Em 2014, o W3C estabeleceu um conjunto de 10 melhores práticas para a publicação de dados conectados. Na esfera governamental, apesar da existência de alguns processos voltados à publicação de dados abertos conectados, estes priorizam os atores técnicos e não consideram o nível de maturidade da instituição publicadora, sendo aparentemente insuficientes para ampliar a disponibilização de Dados Abertos Conectados Governamentais (DACG). Como solução para estes desafios, esta dissertação propõe um modelo de processo para publicação de DACG que, mediante atividades incrementais, busca guiar instituições governamentais à publicarem seus dados em formatos abertos e conectados. O modelo considera o Esquema 5-Estrelas dos Dados Abertos como escala evolutiva, associando a este esquema, atividades obrigatórias e desejáveis que buscam implementar as melhores práticas para publicação de dados conectados. Como contribuições, a pesquisa apresenta uma revisão de literatura sobre quinze processos de publicação de dados abertos aplicáveis ao setor público que sistematiza atividades que podem ser consideradas na publicação de dados governamentais, bem como uma proposta de modelo de processo que poderá impulsionar a oferta de dados desta natureza, mediante a aprimoração de processos existentes associado às melhores práticas para publicação de dados conectados e incorporação de características iterativas de processos, permitindo a sua utilização em diversas finalidades.

Palavras-chaves: Modelo Computacional, Dados Governamentais, Dados Abertos, Dados Abertos Conectados, Dados Abertos Conectados Governamentais, Modelo de Processo.

ABSTRACT

In recent years, governments have been encouraged and obligated to increase the availability of their data on the *Web* through legal obligations and international alliances as the "Open Government Partnership". The dissemination of the concept of open government data is contributing to an expansion of this data supply on the *Web*. However without guarantees that such data are suitable for reuse, automated processing, and support the production of knowledge. In this way, Linked Data concept, presented in 2006 by Tim Berners-Lee, meets the need to improve data quality, through the adoption of principles for structure and connect datasets on the *Web*. In 2014, the World Wide *Web* Consortium - W3C established a set of 10 best practices for publishing linked data. In the government sector, despite the existence of some processes directed to publishing linked open data, they give priority to the technical people and do not consider the maturity level of the publishing government agency, apparently being insufficient to extend the availability of Government Linked Open Data (GLOD) on the *Web*. As a proposal solution, this research proposes a process model for GLOD publication that, through incremental activities, seeks to guide government agencies to publish their data in open and linked formats. The process model considers the 5-Stars Open Data scheme like as evolutionary scale, aggregating to this scheme, mandatory and desirable activities that seek to implement the Best Practices for Publishing Linked Data. As a contribution, the research presents a literature review on fifteen open data publishing procedures applicable to the public sector which organizes activities that can be considered when publishing government data. Another contribution is a proposal of process model that could boost the supply of this kind of data, by enhance existing processes associated with the best practices for publishing linked data and embedding iterative features, allowing its use for various purposes.

Keywords: Computational Model, Government Data, Open Data, Linked Open Data, Government Linked Open Data, Process Model.

RESUMÉN

En los últimos años, los gobiernos han tenido que disponibilizar sus datos en la *Web* a través de las obligaciones legales y alianzas internacionales como la "Alianza para el Gobierno Abierto". El concepto de datos abiertos de gobierno, presentado por Tim Berners-Lee en 2006 cumple con esta necesidad de mejorar la calidad de los datos a través de la adopción de principios para estructurar y conectar conjuntos de datos en la *Web*. En 2014, el Consorcio World Wide *Web* - W3C estableció un conjunto de 10 mejores prácticas para la publicación de datos conectados. En la esfera gubernamental, a pesar de la existencia de algunos procesos encaminados para la publicación de los datos abiertos conectados, que dan prioridad a los actores técnicos y no tienen en cuenta el nivel de madurez de la institución editora, siendo aparentemente insuficientes para ampliar la disponibilidad de los datos gubernamentales abiertos conectados (DACG). A pesar que la difusión del concepto de datos de gobierno abierto contribuye a la expansión de esta fuente de datos de la *Web*, esto no garantiza que estos datos sean adecuados para su reutilización, tratamiento automatizado, apoyo a la generación de conocimiento y a la toma de decisión. Es por ello, que esta investigación se propone un modelo de proceso para su publicación DACG que, a través de actividades incrementales, busca orientar a las instituciones gubernamentales a publicar sus datos en formatos abiertos y conectados. El modelo de proceso considera el Plan de 5-Estrellas de Open Data como escala evolutiva, la vinculación a este esquema, las actividades obligatorias y deseables que tratan de poner en práctica las 10 mejores prácticas para la publicación de los datos conectados. Como contribuciones, se presenta una revisión de la literatura sobre los procedimientos de publicación de quince de datos abierta aplicable al sector público que organiza las actividades que se pueden considerar al publicar los datos del gobierno, así como una propuesta de modelo de proceso que podría impulsar la oferta de los datos de esta naturaleza, por mejorar los procesos existentes relacionados con las mejores prácticas y la incrustación de características proceso iterativo, lo que permite su uso en diversas realidades.

Palabras clave: Modelo Computacional, Datos Abiertos de Gobierno, Datos Abiertos, Datos Abiertos Conectados, Datos Abiertos Conectados de Gobierno, Modelo de Proceso.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACM Association for Computing Machinery

BPLD Best Practices for Publishing Linked Data

CSV Comma Separated Values/Valores Separados por Vírgula

DACG Dados Abertos Conectados Governamentais

DAC Dados Abertos Conectados

DAG Dados Abertos Governamentais

FOAF Friend-Of-A-Friend

GLOD Government Linked Open Data

GQM Goals, Questions, Metrics

HTML HiperText Markup Language

JSON JavaScript Object Notation

LOD Linked Open Data

ODF Open Document Format

ODS Open Document Sheet

OGD Open Government Data

OGP Open Government Partnership/Parceria para o Governo Aberto

OKF Open Knowledge Foundation/Fundação para o Conhecimento Livre

PDF Portable Document Format

RDF Resource Description Framework/Arcabouço de descrição de recursos

SGBD Sistema Gerenciador de Banco de Dados

SPARQL SPARQL Protocol and RDF Query Language/Protocolo e Linguagem de Consulta para RDF

TICs Tecnologias da Informação e da Comunicação

URI Universal Resource Identifier/Identificador Universal de Recursos

URL Universal Resource Locator/Localizador Padrão de Recursos

W3C World Wide *Web* Consortium/Consórcio da World Wide *Web*

XLS Sheet File Format from Microsoft Excel

XML Extensible Markup Language/Linguagem de marcação extensível

XSD XML Schema Definition/Definição de Esquema XML

SUMÁRIO

1 INTRODUÇÃO

O trabalho apresentado ao longo desta dissertação versa sobre a proposição de um modelo de processo para publicação de dados abertos conectados governamentais, baseado e inspirado nas práticas e conceitos recentes aplicados para o aprimoramento da transparência pública e maior oferta de dados e informações governamentais na *Web*. Este trabalho contribui para a geração de subsídios para a construção de conhecimento relacionado às atividades do setor público. Situa-se na linha de pesquisa de Descoberta do Conhecimento do Mestrado em Modelagem Computacional do Conhecimento, do Instituto de Computação da Universidade Federal de Alagoas.

Esta linha envolve pesquisa em representação e processamento de conhecimento em diferentes áreas, abrangendo o uso de modelos matemático-computacionais e de técnicas de inteligência artificial, numa perspectiva de apoio a processos decisórios. Este trabalho se caracteriza como interdisciplinar, visto que o seu objeto de estudo (dados governamentais) é aplicável às diversas áreas do conhecimento contempladas pela atuação governamental, como a economia, estatística, geografia, planejamento, dentre outras. Ademais, a pesquisa está adequada com os objetivos do mestrado no que tange ao desenvolvimento de *expertise* para a construção de modelos e desenvolvimento de sistemas computacionais que possam contribuir para o avanço do ensino, da pesquisa e da produtividade em alguns setores de nossa economia.

A dissertação visa auxiliar o setor público a melhorar sua oferta de dados na *Web* mediante a utilização de um modelo de processo de publicação de dados abertos conectados governamentais. Este modelo de processo permitirá que as instituições governamentais possam dispor de processos e atividades que facilitem a publicação de novos dados, bem como o aprimoramento de dados existentes.

1.1 Motivação e contextualização do trabalho

Governos têm sido fortemente encorajados a ampliar a sua oferta de dados na *Web*, ora motivados pela demanda da sociedade, ora pela necessidade de ampliar a transparência, a participação e a colaboração dos cidadãos que buscam ter uma atuação mais próxima no acompanhamento de políticas públicas e busca de soluções para problemas de interesse coletivo. Complementarmente, com a economia do conhecimento, a demanda por dados também tem como origem, o desenvolvimento de serviços de valor agregado através do desenvolvimento de aplicações e projetos, realizados pelas comunidades de desenvolvimento, as empresas de tecnologias da informação e da comunicação, comunidade acadêmica e ainda, os cidadãos em geral (COLOMBIA, 2012).

Ademais, a obrigação legal em prover o acesso a informações públicas também tem sido

um importante motivador a uma maior publicação de dados governamentais a exemplo da *Lei de Acesso à Informação do Brasil*[1] e suas regulamentações[2], o *Freedom of Information Act*[3], dos E.U.A., a *Ley Organica de Transparencia Y Acceso A La Información en Equador*[4], do Equador, a *Ley sobre Accesso a La Información Pública*[5], do Chile, dentre outras.

Para isto, melhores práticas para publicação de dados têm sido adotadas, onde se inclui o conceito de dados abertos que são dados acessíveis através da *Web*, com a devida permissão legal e em diversos formatos sendo inclusive processáveis por máquinas.

Especialmente no caso brasileiro, a *Lei de Acesso à Informação* no seu artigo 8º, determina que os sítios que disponibilizam informações do Poder Público ofertem-nas em formatos abertos. O § 3o deste artigo da Lei versa sobre os requisitos de disponibilização dos dados nestes sítios:

> [...]
> *"II - possibilitar a gravação de relatórios em diversos formatos eletrônicos, inclusive abertos e não proprietários, tais como planilhas e texto, de modo a facilitar a análise das informações;*
> *III - possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina;*
> *IV - divulgar em detalhes os formatos utilizados para estruturação da informação;"* [...]

Todavia, segundo Hand (2013) as pessoas não estão interessadas apenas em dados, elas querem respostas, e que os dados possuem valor na medida em que eles podem levar às respostas. Atualmente, com a oferta crescente e descentralizada de dados, impõe-se um grande desafio aos consumidores de dados, pois para se produzir informações que gerem respostas, faz-se necessário a coleta, tratamento e armazenamento local de dados para posteriormente se produzir informações. Somado a este fato, outros requisitos de qualidade de dados precisam ser observados nesta oferta, como o tipo de licença de uso, a disponibilidade destes dados, o seu potencial de reutilização, e ainda, se os dados disponibilizados dispõem de elementos semânticos que permitam ao consumidor (humano ou máquina) entender a sua estruturação e a integração com outros conjuntos de dados, permitindo especialmente a extração de conhecimento a partir de conjuntos de dados de tamanhos variados.

Desta forma, há a necessidade de utilizar um padrão para publicação de dados para que todos possam ler e reutilizar dados na *Web* (WOOD et al., 2013), possibilitando que

[1] Disponível em http://www.planalto.gov.br/ccivil03/ato2011-2014/2011/lei/l12527.htm
[2] Disponível em http://www.governoeletronico.gov.br/biblioteca/arquivos/instrucao-normativa-da-infraestrutura-nacional-de-dados-abertos-2013-inda
[3] Disponível em http://www.foia.gov
[4] Disponível em http://anterior.cdc.gob.cl/wp-content/uploads/documentos/legislacion_internacional/ley_organica_
[5] Disponível em http://www.leychile.cl/Navegar?idNorma=276363

as pessoas não precisem ser especialistas em uma determinada área para poder entender os dados referente às respostas que procuram, assim como permitir que computadores pudessem acessar, consumir e interpretar dados de forma automática.

Ao encontro destes requisitos, os dados abertos conectados têm emergido como um novo conceito que permitem que os dados fiquem disponíveis em formatos legíveis por máquina (BAUER; KALTENBÖCK, 2012; ISOTANI; BITTENCOURT, 2015). Com o uso dos dados abertos conectados é possível conectar dados de fontes diferentes e seguir padrões para representação dos dados que permitem torná-los legíveis por máquina bem como dotá-los de elementos semânticos. Os dados são dispostos mediante uma descrição RDF e apontam para outros dados mediante URIs[6], ou seja, a partir de um dado pode-se conhecer várias outros dados e informações relacionados àquele dado.

Para orientar a publicação de dados abertos governamentais, conectados ou não, diversas nações e comunidades acadêmicas vem propondo e estruturando alguns processos de publicação, que visam organizar o ciclo de produção, publicação e disponibilização de dados, considerando os requisitos de qualidade necessários (BRASIL, 2014c; COLOMBIA, 2012; ECUADOR, 2014; CHILE, 2013b; URUGUAY, 2012). Complementarmente, o *World Wide Web Consortium – W3C* tem desempenhado um papel fundamental para estabelecer melhores práticas para a publicação de dados na *Web* bem como para a publicação de dados conectados (W3C, 2013b; W3C, 2013a; W3C, 2014).

Entretanto, a oferta de dados de natureza conectada ainda é pouco presente nos catálogos de dados abertos governamentais. Dentre outros motivos, isto ocorre devido à baixa disponibilidade de orientações comuns e claras para orientar a publicação de dados conectados (VILLAZÓN-TERRAZAS et al., 2011). Além disso, a publicação de dados conectados é complexa por demandar um maior número de requisitos de qualidade destes dados e conjuntos de dados como o estabelecimento de URIs, vocabulários ou recursos de representação semântica de dados (como as ontologias), garantia que os dados estarão conectados a outros dados relacionados ao seu propósito, bem como a incorporação de novas ferramentas tecnológicas para a disponibilização e consumo de tais dados, como servidores de triplas, *endpoints SPARQL*, dentre outros.

Como consequência desta complexidade, há de se considerar a necessidade de avaliar a maturidade organizacional no que tange a publicação de dados, pois, em sendo uma instituição com pouca experiência na produção e disponibilização de dados na *Web*, devem ser adotadas medidas que permitam guiar a publicação de dados governamentais com atividades essenciais que implementem as melhores práticas para esta natureza de publicação. Por outro lado, instiuições mais experientes, devem dispor de um conjunto mais amplo de atividades que proporcionem uma publicação com maior consistência e qualidade.

Por tais razões, esta pesquisa encontrou neste contexto, motivação para propor um

[6] URI: Identificador Uniforme de Recursos (URI) - *Uniform Resource Identifier (em inglês)* é uma cadeia de caracteres compacta usada para identificar ou denominar um recurso na Internet

modelo de processo, que apresente alternativas de processos de publicação de dados considerando o nível de maturidade da instituição publicadora, contribuindo com as estratégias de produção e disponibilização de dados abertos e dados abertos conectados do setor público.

1.2 Problemática

A publicação de dados governamentais tem sido fortemente estimulada por diversos motivos, especialmente os de natureza legal. Como consequência, inúmeros portais e catálogos de dados governamentais foram desenvolvidos, ofertando milhares de conjuntos de dados online. Em 2012, já existiam 115 catálogos desta natureza disponíveis, ofertando cerca de 710.000 conjuntos de dados (HENDLER et al., 2012). Atualmente, em 2015, segundo o DataPortals.org[7], existem 434 catálogos de dados governamentais abertos disponíveis em todos os continentes, o que comprova a rápida ascensão deste paradigma.

Figura 1 – Mapa Mundi dos Catálogos de Dados Abertos Governamentais

Fonte: Disponível em <http://dataportals.org>. Acesso em: 10 out. 2015

Entretanto, apesar desta oferta disponível de dados abertos, ainda existe uma predominância de dados em formato proprietário e não-processáveis por máquina, dificultando o seu reúso e apropriação para produção de conhecimento. Ademais, outros desafios emergem como decorrência desta vasta oferta de dados como:

- **Restrições ao Consumo:** Em vários casos, são ofertados conjuntos de dados enormes, cujo consumo depende de um *download* demorado. Além disso, ao obter o conjunto de dado, o usuário precisa de um recurso computacional que consiga processá-lo e tenha capacidade de localizar os dados para a obtenção das respostas

[7] Disponível em http://www.dataportals.org

que deseja, conforme sugerido por Hand (2013). Por exemplo, cada conjunto de dados contendo os dados do Orçamento Federal Brasileiro por ano possui em média, 22 MB [8]. Uma consulta para analisar o perfil orçamentário de um órgão público num período de 5 anos resulta no *download* de aproximadamente 110 MB de dados para posterior cruzamento e localização de respostas, algo complexo para a maioria da população. Outros desafios derivados deste são:

  - **Restrições à localização de dados:** A vasta oferta de dados resulta numa sobrecarga de disponibilidade de dados, e como consequência, existe uma dificuldade natural na localização e recuperação de certos dados.
  - **Restrições ao consumo automatizado dos dados:** Apesar de parte desta oferta estar disponível em formatos estruturados, para que haja o consumo destes dados de forma automatizada, é necessário um esforço de extração dos dados das fontes originais, tratamento às necessidades do usuário e carga numa base de dados local. Em que pese haverem APIs voltadas a reduzir este esforço, é necessário o entendimento específico de cada tipo de API disponibilizada, resultando em novos esforços para o consumo de dados.
  - **Restrições ao consumo de vários conjuntos de dados de forma integrada:** Conforme o exemplo inicial, o consumo de dados de várias fontes requer que o usuário visite os catálogos de dados de todas as fontes que deseja, obtenha uma cópia dos conjuntos de dados e posteriormente faça o tratamento e carga numa base local para que possa cruzar os dados e obtenha as respostas que deseja. Neste contexto, seria relevante que o acesso a tais dados fosse realizado mediante consultas a serviços de informação, que integrassem dados de diversas fontes e principalmente, que não obriguem os usuários a fazer o *download* dos conjuntos de dados para cada necessidade de consulta e uso dos dados.

- **Condições legais para o uso dos dados:** A oferta de dados abertos requer o estabelecimento de licenças abertas. Entretanto, o não estabelecimento destas licenças pode resultar numa reutilização indevida dos dados. Por exemplo, um conjunto de dados cuja disponibilidade seja restrita para fins acadêmicos, pode ser utilizada para fins comerciais caso não haja um dispositivo legal que estabeleça claramente esta restrição;

- **Duplicação indiscriminada dos conjuntos de dados:** Por estarem disponíveis para qualquer pessoa, um determinado conjunto de dados pode ser amplamente replicado ao longo da *Web* e de sistemas de informação resultando novos problemas como: impossibilidade de atualização automática das cópias dos conjuntos de dados

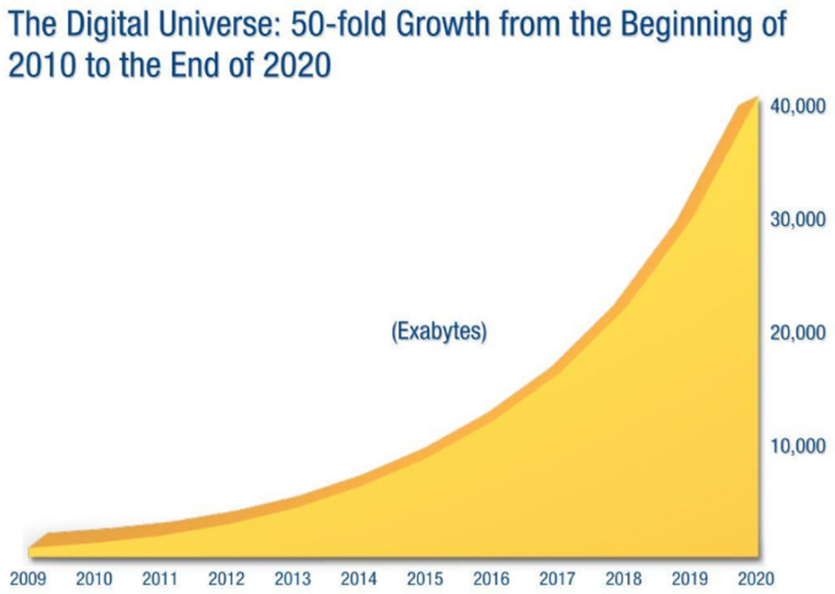
[8] Conjuntos de dados disponíveis em: http://dados.gov.br/dataset/orcamento-federal

quando o original foi atualizado; Risco a utilização de conjuntos de dados defasados, podendo gerar decisões intempestivas, etc.

- **Ausência de elementos semânticos para dar significado aos dados:** Para a obtenção de respostas relevantes, é necessário associar aos dados um contexto de aplicação, para geração de informações e conhecimentos. A oferta tradicional de dados abertos não agrega tais elementos, possibilitando o uso de tais dados em contextos em que o mesmo não se aplica;
- **Proveniência dos dados:** No modelo atual, uma fonte secundária pode se apropriar de conjuntos de dados de uma fonte primária e ofertá-lo em meios mais fáceis para o seu consumo e por tal razão, confundir os usuários quanto a verdadeira origem destes dados.

Num contexto mais amplo, este tema passa a ganhar maior relevância quando são observados os prognósticos referentes ao volume de dados que serão produzidos nas próximas décadas. O estudo *"A Universe of Opportunities and Challenges"*, desenvolvido pela consultoria EMC (EMC, 2012), aponta que de 2006 a 2010, o volume de dados digitais gerados cresceu de 166 Exabytes para 988 Exabytes. Conforme a Figura 2 existe a perspectiva que o volume de dados seja de 40 *Zettabytes* (ou 40 trilhões de *Gigabytes*) em 2020.

Figura 2 – Perspectiva de crescimento da oferta de dados digitais até 2020



Fonte: EMC (2012)

Nesta direção, as perspectivas da economia digital e da oferta de dados são muito promissoras, mas existem problemas relevantes a serem considerados, como:

- Para a tomada de decisão eficaz, as empresas poderão ter que organizar não apenas as suas informações, mas também consumir cada vez mais as informações de terceiros (como fornecedores, governo, etc.), o que vai resultar num esforço ainda maior para a melhoria da oferta de dados e o relacionamento entre eles (estabelecimento de conexões) considerando o caráter cada vez mais descentralizado destes recursos de dados (MCKINSEY, 2011);

- Quanto ao potencial de uso dos dados digitais do mundo, o estudo da EMC aponta um dado preocupante: Em 2012, apenas 23% da informação digital do mundo é útil para gerar novas informações e conhecimento e apoiar a tomada de decisão no âmbito do *Big Data*. Deste total, apenas 3% destas informações são úteis para uso imediato (os demais 20% ainda precisam ser tratados para estar aptas ao uso) (EMC, 2012).

Neste contexto, os modelos vigentes (incluindo os novos catálogos de dados governamentais) aparentemente não tem sido suficientes para melhorar a oferta de dados. Segundo Heath (2011) os consumidores argumentam que a oferta de dados atual vastamente espalhada pela *Web* representa um grande inconveniente, pois existe a necessidade de primeiro obter e armazenar estes dados localmente, antes que possam ser utilizados para a produção de informações relevantes.

Desta forma, se não forem adotadas medidas para a melhoria da qualidade dos dados e especialmente, a adoção de valor semântico que subsidiem a geração de conhecimento, considerando a velocidade da ampliação da oferta de dados prevista pela EMC (EMC, 2012), o impulsionamento à publicação de dados governamentais pode resultar num problema que podemos definir como "um grande **bando** de dados" na *Web* com pouco potencial de reaproveitamento. Assim, torna-se necessário a proposição de conceitos, processos e ferramentas que ampliem o reúso de dados disponível.

Para orientar o aprimoramento de dados abertos para dados abertos conectados, foi estabelecido o esquema *5-Estrelas dos Dados Abertos* (BERNERS-LEE, 2006) que apresenta níveis de maturidade para a oferta de dados abertos na *Web*, sendo necessário o dado alcançar o maior nível (5) para ser considerado conectado.

Segundo Heath e Bizer (2011) a aplicação dos princípios dos dados conectados aos conjuntos de dados governamentais trazem um potencial enorme. Todavia, este potencial é comumente bloqueado devido à falta de recursos para transformar dados brutos em dados conectados de alta qualidade e em larga escala (VILLAZÓN-TERRAZAS et al., 2011).

Um dos principais problemas está relacionado ao fato que a geração e publicação de dados conectados não seguem um conjunto de orientações comuns e claras para dimensionar tais atividades. De resto, ainda existe uma carência de diretrizes pormenorizadas e softwares que apoiem todo o ciclo de vida de publicação de dados conectados no âmbito

governamental. Agrava-se ao fato que, na maioria das vezes, as orientações existentes são destinados a desenvolvedores de software, não para os governos (VILLAZÓN-TERRAZAS et al., 2011).

Ademais, os guias e processos de publicação de dados (abertos) conectados disponíveis costumam apresentar um único fluxo para qualquer situação de publicação de dados, ou seja, independente da tarefa de publicação, ou ainda, da maturidade da instituição publicadora, o procedimento a ser seguido é o mesmo, não dispondo de elementos que facilitem a publicação por parte de iniciantes muito menos que orientem o aprimoramento da oferta de dados existente.

Considerando o exposto nesta seção, esta pesquisa investiga e propõe soluções para as seguintes questões:

Q1- Como guiar instituições governamentais a publicar dados abertos conectados considerando as melhores práticas e procedimentos produzidos e disponíveis na literatura?

Este é o principal problema do trabalho proposto. Em que pese todos os benefícios inerentes aos dados abertos conectados, a oferta destes dados ainda é muito baixa nos catálogos governamentais. Considerando a disponibilidade de dados no formatos RDF e OWL, a Tabela 1 apresenta a participação destes dados em relação à oferta total de dados disponíveis nos catálogos de alguns países ao redor do mundo:

Tabela 1 – Participação de dados conectados nos catálogos de dados abertos governamentais

País	**Recursos de dados abertos conectados**[9]	**Total de recursos de dados**	**Participação (em%)**
Alemanha[10]	4	37.479	0,01%
Austrália[11]	0	2.137	0,00%
Brasil[12]	15	8.582	0,17%
Cingapura[13]	0	11.977	0,00%
E.U.A.[14]	7.051	332.568	2,12%
Espanha.[15]	1.775	9.029	19,66%
Itália[16]	2.055	21.850	9,41%
Japão[17]	0	18.105	0,00%
Reino Unido[18]	362	9.497	3,81%

Fonte: Autor desta dissertação, 2015.

A Tabela 1 apresenta um volume relevante de recursos de dados disponíveis que, de alguma forma, são resultados de processos e atividades de publicação de dados. Entretanto, nota-se uma participação muito baixa de dados nos formatos compatíveis com dados abertos conectados. Desta maneira, justifica-se uma investigação mais detalhada sobre os processos de publicação de dados existentes, bem como as melhores práticas desenvolvidas nos últimos anos, permitindo identificar novos fatores que contribuam para uma maior oferta de dados abertos conectados na esfera governamental.

Preliminarmente, analisando o Esquema 5-Estrelas dos Dados Abertos (BERNERS-LEE, 2006), verifica-se a exigência de requisitos para que um dado seja considerado conectado. Para desenvolver tais requisitos são necessárias diversas tarefas de identificação, modelagem, licenciamento, conversão e publicação de dados que pode ter uma grande complexidade especialmente quando a instituição publicadora não tiver muita experiência em publicação de dados. Em resumo, instituições públicas estão sendo obrigadas a publicar dados abertos que permitam o seu acesso (consumo) de forma automatizada (vide a legislação brasileira, como exemplo).

As instituições que não possuem experiência deparam-se com uma tarefa complexa a cumprir para publicar dados desta forma. A publicação de dados abertos conectados, neste cenário, torna-se ainda mais desafiadora. Ademais, os processos de publicação de dados abertos (e governamentais) analisados nesta pesquisa, aparentemente, propõem um processo único para qualquer tipo de instituição que deseje publicar seus dados, não considerando o porte e a maturidade da organização e de sua equipe técnica neste tipo de atividade.

Além disso, cumpre destacar que, no ano de 2014, o W3C elaborou um conjunto de "Melhores Práticas para Publicação de Dados Conectados" (W3C, 2014). Tais melhores práticas precisam ser incorporadas aos processos de publicação de dados, de tal maneira que o esforço para este tipo de publicação seja mais assertivo e possua melhores resultados.

Neste sentido o modelo proposto nesta pesquisa além de incorporar melhores práticas da literatura disponível, pretende oferecer às instituições publicadoras um conjunto de atividades obrigatórias e desejáveis, voltadas para que tomadores de decisão do setor público possam compreender com maior facilidade os benefícios da publicação de dados em formato conectado, ter orientações adequadas para a sua produção e disponibilização na *Web* e enfim, incentivar a publicação de dados abertos conectados governamentais.

Q2- Que novos procedimentos devem ser propostos para que instituições governamentais publiquem dados abertos conectados considerando o seu nível de maturidade em gestão e publicação de dados?

Com base nos números da Tabela 1 e da constatação que as orientações para publicação de dados abertos conectados são *"destinados a desenvolvedores de software, não para os governos"*(VILLAZÓN-TERRAZAS et al., 2011), justifica-se uma investigação de conceitos complementares que subsidiem o desenvolvimento de procedimentos para publicação de dados voltados preferencialmente para o agente público que atua na camada de negócio.

Além disso, como comentado na questão anterior, a publicação de dados abertos conectados atualmente não é trivial e requer que sejam incorporados, cumulativamente, os requisitos dos níveis do Esquema 5-Estrelas dos Dados Abertos. Logo, há de se considerar que as atividades de publicação devem evoluir ao longo dos cinco níveis até atingir a condição de dado aberto conectado.

Tendo como referência o exposto quanto à necessidade de se considerar à experiência da organização e de seus técnicos em atividades de publicação de dados, serão investigados, diversos processos de publicação de dados, bem alguns modelos de maturidade em software e em dados e modelos de processo de desenvolvimento de software, buscando incorporar conceitos que potencializem o embasamento do modelo de processo de publicação de dados a ser proposto nesta pesquisa.

A incorporação destes conceitos ao modelo proposto visa contribuir para que a instituição publicadora possa alcançar o nível de publicação de dados abertos conectados decorrente de um processo evolutivo, desenvolvendo um conjunto de etapas planejadas que proporcionem o aprimoramento dos seus dados e respectivas atividades de publicação. Espera-se que, com esta proposta, o publicador possa desenvolver sua oferta de dados gradativamente até o nível de dados abertos conectados.

Nesse sentido, a solução proposta visa resolver as duas questões apresentadas, solucionando-as mediante a extração e compilação de atividades de publicação de dados associada às "Melhores Práticas para Publicação de Dados Conectados" do W3C e a um modelo iterativo e incremental que guie atividades de publicação de dados abertos até o nível de dados abertos conectados. Na seção seguinte definiremos os objetivos dessa dissertação.

1.3 Objetivos

A pesquisa tem como objetivo a proposição de um modelo de processo para publicação de dados abertos conectados governamentais que considere o nível de maturidade dos órgãos governamentais, os níveis de maturidade do Esquema 5-Estrelas dos Dados Abertos. Este modelo também apresenta atividades para a implementação de melhores práticas para publicação de dados conectados contribuindo para aprimorar o acesso automatizado por consumidores humanos e sistemas externos, permitindo um melhor aproveitamento destes dados para o desenvolvimento de modelos computacionais de conhecimento a serem aplicáveis em diversas áreas.

Os objetivos específicos deste trabalho são:

1. Realizar uma revisão da literatura sobre processos de publicação de dados abertos, dados abertos governamentais e dados abertos conectados;

2. Analisar comparativamente processos de publicação de dados identificados;

3. Verificar a existência de práticas que recomendem e/ou guiem as instituições a produzirem dados abertos conectados;

4. Extrair recomendações para publicação de dados a partir dos processos analisados;

5. Propor e Classificar um conjunto de recomendações a serem incorporados aos processos de abertura de dados visando estimular a produção de dados abertos conectados;

6. Discutir e propor um modelo de processo de publicação de dados abertos conectados governamentais com características iterativas e incrementais, incorporando as recomendações propostas e classificadas;

7. Validar o modelo proposto.

1.4 Escopo

O trabalho apresentado ao longo dessa dissertação trata da proposição, criação e avaliação de um modelo de processo de publicação de dados abertos conectados governamentais. Situa-se na linha de pesquisa de Descoberta do Conhecimento e Otimização de Decisões do Mestrado em Modelagem Computacional do Conhecimento, do Instituto de Computação da Universidade Federal de Alagoas.

Esta dissertação visa auxiliar agentes e instituições públicas quanto a melhoria da publicação de dados governamentais, mediante a apresentação de uma proposta de modelo evolucionária para guiar a publicação de dados governamentais de forma aberta e conectada, resultando em benefícios para o setor governamental e para os consumidores de tais dados.

1.5 Contribuições do trabalho

O presente trabalho contribui com a comunidade de dados abertos conectados, apresentando uma proposta de modelo que poderá impulsionar a oferta de dados desta natureza, permitindo a sua utilização em diversas finalidades que serão explanadas neste documento.

Além disso, outra contribuição relevante do trabalho é a apresentação de uma revisão de literatura sobre quinze processos de publicação de dados abertos aplicáveis ao setor público. Esta revisão busca sistematizar recomendações a serem consideradas na publicação de dados governamentais, apresentando ainda às instituições publicadoras diversas abordagens sobre publicação de dados abertos.

No campo das políticas públicas, o trabalho ainda apresenta um novo instrumento para impulsionar as políticas de transparência, acesso à informação e governo aberto, potencializando o uso de tais dados e informações em atividades de controle social das ações do poder público, planejamento de novos projetos e desenvolvimento de novos negócios.

1.6 Organização da dissertação

A dissertação aqui apresentada contém sete capítulos organizados e distribuídos da seguinte maneira:

- Capítulo 1: Neste capítulo são apresentados uma motivação, contextualização, problemática, objetivos e as contribuições do trabalho aqui proposto.
- Capítulo 2: São apresentados neste capítulo a fundamentação teórica que aborda os principais conhecimentos utilizados nessa dissertação, enfatizando os temas principais desta pesquisa.
- Capítulo 3: São apresentados trabalhos relacionados à publicação de dados abertos governamentais, bem como alguns modelos de referência que contribuem para o entendimento do contexto de aplicação bem como do modelo a ser proposto.
- Capítulo 4: Este capítulo apresenta um conjunto de recomendações para publicação de dados abertos conectados governamentais extraídos de uma revisão de literatura. Estas recomendações são utilizadas pelo modelo de processo como sugestão de atividades a serem desenvolvidas para a publicação dos dados.
- Capítulo 5: Este capítulo apresenta, de forma detalhada o modelo de processo de publicação de dados abertos conectados governamentais proposto, incorporando as atividades descritas no capítulo anterior.
- Capítulo 6: Este capítulo apresenta a validação do modelo proposto.
- Capítulo 7: Por fim, este capítulo apresenta as conclusões acerca do trabalho apresentado, bem como algumas sugestões de trabalhos futuros.
- No final, serão apresentadas às referências utilizadas na elaboração da pesquisa e os apêndices.

2 FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo são abordados os tópicos que contém os principais fundamentos teóricos dessa dissertação, necessários para a análise e compreensão dos elementos do modelo apresentado, bem como o entendimento de sua implementação e respectivo processo de validação.

Nesse sentido, as seções foram distribuídas da forma como segue. Num primeiro momento são apresentados os conceitos de dados e informações governamentais e sua relevância para todos os demais segmentos produtivos da sociedade na seção 2.1. Em seguida, na seção 2.2, será apresentado o conceito de Governo Aberto, explicitando as iniciativas e perspectivas para a promoção de governos mais transparentes, abertos e voltados à participação cidadã e co-criação social, onde os dados e informações governamentais são elementos-chave deste conceito.

Na seção 2.3 serão apresentados os conceitos de dados abertos e dados abertos conectados como perspectiva de aprimoramento da oferta de dados e informações governamentais que resultam, respectivamente, nos conceitos de Dados Abertos Governamentais e Dados Abertos Conectados Governamentais.

Posteriormente são apresentados diversos modelos de referência para fundamentação da proposta na seção 2.4 e ainda, na seção 2.5, a metodologia GQM, utilizada na validação empírica desta pesquisa.

2.1 Dados e Informações Governamentais

O desenvolvimento socioeconômico, historicamente, é guiado pela disponibilidade e qualidade de dados e informações disponíveis para subsidiar ações de caráter social e econômico. Há de se considerar que desde os primórdios da humanidade que a necessidade de informações para a tomada de decisão consiste numa necessidade inerente ao convívio social (ÁVILA et al., 2012). Neste contexto, a viabilização do desenvolvimento positivo, aquele que visa à promoção da melhoria de uma determinada condição, espaço geográfico, instituição ou comunidade, é pautado por um processo contínuo de tomada de decisões em níveis estratégico, tático e operacional, onde a informação consiste de subsídio fundamental para a realização das melhores escolhas.

Considerando o papel do Estado como indutor do desenvolvimento socioeconômico e o impacto de suas decisões e ações para toda a sociedade, a informação pública possui papel destacado como subsídio ao desenvolvimento, apoiando ações não apenas do segmento governamental, mas também do setor produtivo, acadêmico e da sociedade em geral. Um destaque especial para o papel da imprensa livre, que tem atuação importante na tarefa de comunicar os fatos, bem como denunciar ações inadequadas.

Na Tabela 2 é apresentado o relacionamento entre alguns tipos de informações públicas e as principais demandas e aplicações em diversos segmentos da sociedade.

Tabela 2 – Principais demandas de informações governamentais

Segmento	**Tipo de Informação Pública**	**Finalidade**
Setor Produtivo	Indicadores Sociais, Econômicos, Demográficos, Planos de Governo, Relatórios Fiscais, Informações Geográficas (imagens aéreas, vetores com distâncias entre localidades, mapas e cartogramas sobre dados socioeconômicos), etc.	Projetos de Consultoria; Expansão e/ou Manutenção de Negócios; Desenvolvimento ou aprimoramento de produtos e serviços
Setor Acadêmico	Indicadores Sociais, Econômicos, Demográficos, Planos de Governo, Relatórios Fiscais, Informações Geográficas (imagens aéreas, vetores com distâncias entre localidades, mapas e cartogramas sobre dados socioeconômicos), etc.	Artigos Científicos; Trabalhos Acadêmicos; Projetos de Pesquisa; Monografias; Dissertações; Teses; Projetos de Pesquisa e Extensão; Projetos para captação de recursos em instituições de fomento
Setor Público	Indicadores Sociais, Econômicos, Demográficos, Planos de Governo, Relatórios Fiscais, Informações Geográficas (imagens aéreas, vetores com distâncias entre localidades, mapas e cartogramas sobre dados socioeconômicos), etc.; - Pesquisas acadêmicas, estudos e análises, relatórios de tendência, projeções de cenários.	Diagnósticos governamentais, diagnósticos sobre áreas ou demandas específicas (ex: problemas ambientais); Formulação de planos e programas de governo, execução de ações, monitoramento e avaliação governamental; Publicidade de ações governamentais; Projetos para captação de recursos em instituições de fomento
Imprensa	Dados orçamentários e financeiros; Pesquisas e indicadores socioeconômicos; Dados Populacionais; Relatórios de Monitoramento e Acompanhamento de Ações Governamentais	Matérias e investigações jornalísticas; Publicidade de ações governamentais; Denúncias de não-conformidades em ações governamentais
Sociedade em Geral	Dados orçamentários e financeiros; Pesquisas e indicadores socioeconômicos; Dados Populacionais	Monitoramento e Controle Social do Governo; Elaboração de Projetos para captação de recursos

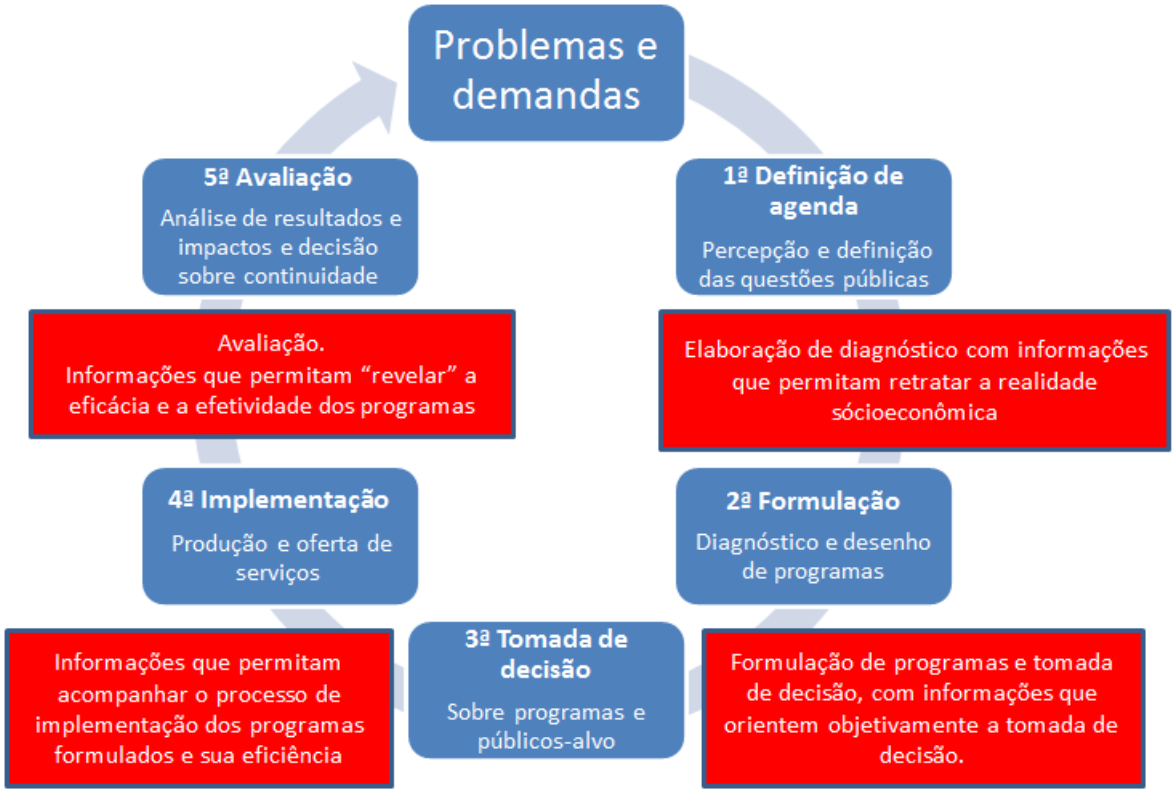
Fonte: Autor desta dissertação, 2015.

Corroborando o que foi apresentado na Tabela 2, Janssen, Charalabidis e Zuiderwijk (2012 apud JANSSEN, 2011) pontuam que as organizações públicas estão entre os maiores

criadores de dados em diversos domínios. Estes domínios variam entre o trânsito, clima, dados geográficos, informações turísticas, estatísticas, negócios, orçamento público, planejamento e avaliação do setor público, bem como diversos tipos de dados sobre políticas públicas especializadas em áreas como alimentação, educação, saúde, segurança, dentre outras (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

Neste contexto, a relevância das informações públicas para a tomada de decisão, especialmente a de natureza estatística, costuma ser bem representada em todas as etapas do ciclo de planejamento e políticas públicas. A Figura 3 explana este ciclo e as necessidades de informações para cada etapa (JANNUZZI, 2012).

Figura 3 – Presença do recurso informação no ciclo de planejamento e políticas públicas



Fonte: Adaptado de Jannuzzi (2012)

Numa visão análoga a apresentada sobre o ciclo de planejamento conforme a Figura acima, a Tabela 2 também destaca o uso das informações públicas para outras finalidades e noutros segmentos. O setor produtivo faz uso de informações em finalidades como a decisão para implantar novos negócios, novas unidades de negócio ou expandir a atuação empresarial, onde as informações são utilizadas nas etapas de diagnóstico, formulação de projetos, execução e avaliação. No segmento acadêmico, este ciclo de uso da informação também se faz presente, em especial na formulação e execução de projetos de pesquisa. Cumpre destacar também o uso da informação pela sociedade, seja no uso de índices econômicos para decisões relacionadas ao planejamento financeiro (aplicações financeiras, índices de inflação), dentre outras finalidades.

Nesta direção, tendo em vista a forte demanda por dados públicos por toda a sociedade, historicamente vem sendo desenvolvidas iniciativas no campo técnico e jurídico visando aprimorar a oferta de informações a sociedade, conforme será explanada na seção a seguir.

2.2 **Governo Aberto**

No histórico recente do desenvolvimento da humanidade há de se ressaltar que o século XX ficou marcado por duas grandes revoluções na história da humanidade: a revolução industrial e a revolução da informação, da comunicação e do conhecimento. A cada etapa de desenvolvimento da história mundial, a produção de dados e a necessidade de informação para tomada de decisão têm crescido e deverá crescer em proporções jamais imagináveis (ÁVILA et al., 2012).

A partir desta segunda revolução, as Tecnologias da Informação e da Comunicação (TICs), em especial a *World Wide Web* (WWW) obtiveram grande protagonismo na sociedade em geral. No que tange ao acesso a informação, tais tecnologias passaram a ser o principal meio de disseminação e compartilhamento de dados e informações e considerando os recursos que proporcionam para a publicação de informações, as TICs tornaram-se uma grande aliada de políticas de disponibilização da informação.

Reilly (2010) complementa ainda que a *Web* criou novos métodos para aproveitar a criatividade das pessoas em grupos, bem como está proporcionando a criação de novos modelos de negócios que estão remodelando nossa economia, ou seja, uma nova geração nasceu na *"era da Web"* e está empenhada em utilizar suas lições de criatividade e colaboração para enfrentar os desafios das nações mundiais.

Na esfera governamental, Reilly (2010) destaca que, com a complexidade e a proliferação das demandas da sociedade e a restrição de recursos para solucioná-los, muitos líderes de governo reconhecem que as oportunidades que as tecnologias *Web* fornecem, em especial com seus recursos atuais conhecidas como *Web 2.0*, não se limitam apenas para ajudá-los a serem eleitos, mas também podem ajudá-los a fazer um governo melhor. Por analogia, muitos estão chamando este movimento de Governo 2.0 (*Government 2.0*).

O Governo 2.0 é definido como o uso das TICs, especialmente tecnologias colaborativas do núcleo da *Web* 2.0 para o desenvolvimento de melhores soluções para os problemas coletivos de cidades, estados, nações e problemas em nível internacional.

Ademais, os cidadãos estão a cada dia, mais dispostos a envolver-se na tomada de decisões políticas de maior complexidade e o surgimento do Governo 2.0 oferece novas oportunidades para que atores sociais atuem na produção de dados, análise e tomada de decisões (MUREDDU et al., 2012).

O conceito de Governo 2.0 é recente o que justifica em partes, sua adesão ainda estar ocorrendo de forma gradual por governos, especialmente nos níveis subnacionais. Um marco histórico no posicionamento político por governos mais efetivos, inovadores e trans-

parentes foi registrada na campanha presidencial americana em 2008, onde o então candidato a presidente, Barack Obama explanou a seguinte proposta em sua campanha:

> *"We must use all available technologies and methods to open up the federal government, creating a new level of transparency to change the way business is conducted in Washington, and giving Americans the chance to participate in government deliberations and decision making in ways that were not possible only a few years ago."*

Posteriormente, após sua eleição, um dos primeiros atos do Governo Obama consistiu na publicação do *"Memorandum of Transparency and Open Government"* em 21 de Janeiro de 2009 (OBAMA, 2009b). Neste memorando, o Governo Obama se comprometeu a promover um nível jamais visto de abertura governamental e estabeleceu os três grandes pilares para o que se entende como um Governo Aberto: transparência, participação e colaboração.

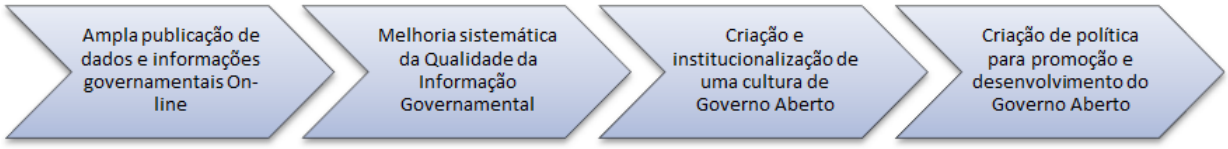
> *[...] "My Administration is committed to creating an unprecedented level of openness in Government. We will work together to ensure the public trust and establish a system of* ***transparency, public participation, and collaboration****." [...]*

Posteriormente, em oito de Dezembro de 2009, foi publicado outro memorando denominado *"Open Government Directive"*, que detalhou como deveriam ser implementados os três princípios para o Governo Aberto supracitados (OBAMA, 2009a). Este documento destaca que a transparência promove e fortalece a responsabilização dos atos governamentais, fornecendo ao público informações sobre o que o governo está fazendo. A participação permite que cidadãos possam contribuir com ideias e conhecimentos para que o governo possa fazer políticas que se utilizem as informações que estão amplamente disponíveis na sociedade. Por fim, a colaboração melhora a eficácia do governo encorajando parcerias e cooperações dentro da Administração Pública, dentre os vários níveis de governo e entre o governo e instituições fora do governo, como a iniciativa privada e outras entidades não-governamentais.

Apesar das nações mundiais desenvolverem diversas ações ora voltados a Transparência Pública, ora voltadas a participação, ora voltadas a colaboração, o conceito apresentado de Governo Aberto passou a influenciar o mundo por causa da conexão estabelecida pelos seus três pilares. Somado a isto, o forte encorajamento ao uso extremo de TICs para o fortalecimento da ação governamental com base neste conceito ensejou um novo ecossistema em torno desta temática em nível global. Bandeira et al. (2014) complementa que um Governo é considerado aberto quando é implementado uma plataforma de políticas públicas voltadas para este fim.

Entretanto, a implementação do Governo Aberto pode requerer investimentos e compromissos não-triviais por parte dos órgãos governamentais que precisarão adquirir novas habilidades, capacitar funcionários, incorporar novas tecnologias e em alguns casos, atualizar a sua infraestrutura de TIC (LEE; KWAK, 2012). Para a implementação de políticas desta natureza são condicionantes a criação e institucionalização de uma cultura de Governo Aberto, onde neste contexto o capital humano precisa ser desenvolvido e encorajado para atuar neste novo paradigma (REILLY, 2010). Isto posto, para que esta cultura seja implantada, são necessários outros condicionantes como o enriquecimento da qualidade da informação governamental e da tomada de decisão, que por ventura, prescinde e estimula que haja uma ampla publicação dos dados e informações governamentais na *Web*. A Figura 4 abaixo demonstra este fluxo de condições estabelecidas por Reilly (2010) para a viabilização do Governo Aberto.

Figura 4 – Princípios para o Governo Aberto



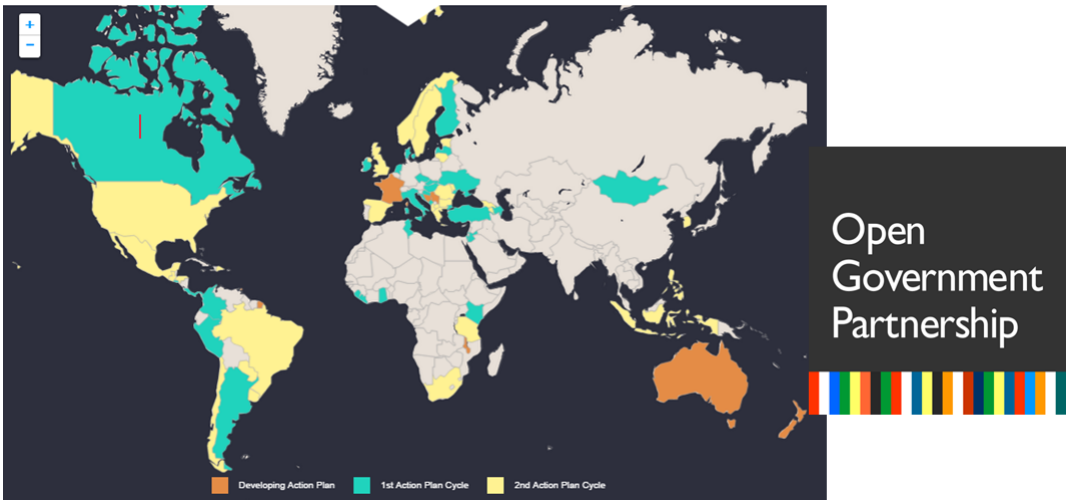
Fonte: Bandeira et al. (2014), a partir de Reilly (2010)

Em escala global, desde 2011 que nações de diversos continentes, liderados pelos E.U.A. e pelo Brasil instituíram a "Parceria para o Governo Aberto" (*The Open Government Partnership* - OGP), como uma iniciativa multilateral que visa assegurar compromissos concretos dos governos para promover a transparência, capacitar cidadãos, lutar contra a corrupção e aproveitar as novas tecnologias para fortalecer a governabilidade no espírito de colaboração entre os vários interessados e que atualmente já possui 65 nações filiadas (BANDEIRA et al., 2014; OGP, 2015a) conforme apresentado na Figura 5. Como principais objetivos, a OGP se propõe a ampliar o acesso a novas tecnologias para fins de transparência e prestação de contas; a aumentar a disponibilidade de informações sobre as atividades governamentais; a apoiar a participação cívica; e a implementar os mais altos padrões de integridade profissional por todas as administrações (GUIMARÃES; DINIZ, 2014). Este conjunto de nações reunidos estabeleceu cerca de 1.000 compromissos para a promoção de governos mais abertos, transparentes e responsáveis (OGP, 2015b).

Neste contexto, considerando os conceitos expostos pode-se deduzir que a efetivação do Governo Aberto passa naturalmente pela transparência e abertura de dados públicos. Todavia, é importante destacar que a dimensão do Governo Aberto está além da camada técnica, envolvendo questões político-institucionais.

A abertura de dados públicos é fundamental para uma política de Governo Aberto, mas não é o único fator responsável pelo desenvolvimento de tais políticas. Dentre os principais desafios associados à relação entre dados abertos e governo aberto temos: (i)

Figura 5 – Mapa mundi ilustrado com os países signatários da OGP



Fonte: OGP (2015a)

Para a tomada de decisão a partir de dados abertos, é necessário possuir habilidades para encontrar os dados desejados, interpretá-los e processá-los corretamente; e ainda (ii) é necessário que hajam canais de diálogo entre os consumidores e provedores de dados, para que haja o devido esclarecimento sobre as características dos dados, bem como para que os governos recebam opiniões dos usuários dos dados que possam melhorar o que estas instituições estão ofertando (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

Ademais, é importante ressaltar que mais informações não necessariamente pode resultar em decisões melhores e/ou mais democráticas (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012). O excesso de informações pode reduzir a capacidade de entendimento a respeito dos dados e informações públicas. Por estas razões, a próxima seção analisará os conceitos de Dados Abertos e Dados Abertos Conectados, buscando o melhor entendimento destes conceitos e sua relação com o Governo Aberto.

2.3 Dados Abertos

Segundo a Open Knowledge Foundation (OKF, 2015c), Dados Abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, a exigência de atribuição da fonte e compartilhamento pelas mesmas regras. GuimarÃes e Diniz (2014) complementam que tais dados devem ser publicados e distribuídos na Internet, compartilhados em formato aberto para que possam ser lidos pessoas e máquinas, permitindo o cruzamento com outros dados de diferentes fontes, para serem livremente reutilizados.

Com dados abertos disponíveis, abrem-se novas possibilidades para a sociedade, que vão desde a análise mais profunda das informações públicas por meio da correlação de diferentes bases de dados, até a criação de aplicativos que fazem uma leitura frequente de

bases de dados públicas para fornecer soluções que beneficiem a sociedade ou que geram oportunidades de negócio (EAVES, 2009; NEVES, 2013).

Entretanto, apesar do crescimento da disponibilidade de dados abertos, a atual oferta de dados na *Web* ainda tem ocorrido em formatos que impõem limitações quanto a sua reutilização, pois em sua maioria são consumidos apenas por humanos, não permitindo que sejam reutilizados de forma automatizada por agentes de software (WOOD et al., 2013). Desta maneira, buscaremos apresentar o conceito de Dados Abertos Conectados como um aprimoramento da oferta de dados abertos.

2.3.1 Dados Abertos Conectados

Considerando o volume de dados e informações conforme destacado na Figura 2, bem como a atual descentralização desta produção como exemplificado na Figura 5, novos desafios emergem no que tange a organização e consumo de dados, pois a tomada de decisão precisa ser subsidiada por informações integradas, comumente decorrentes do cruzamento de várias bases de dados.

Neste contexto, os consumidores de dados visualizam que a oferta de dados atual vastamente espalhada pela *Web* representa um grande inconveniente, pois existe a necessidade de primeiro obter e armazenar estes dados localmente antes que possam ser utilizados para a produção de informações relevantes (HEATH, 2011). O autor ainda ressalta que, mesmo que a informação do setor público esteja disponível em formato aberto, pode estar publicada de forma caótica. Ademais, a mesma informação pode ser encontrada em diferentes locais da *Web* e ainda, sem haver nenhuma conexão entre tais fontes de informações, apresentando, por exemplo, qual é a informação mais atualizada.

Diante desta situação, para que os usuários tenham confiança nos dados disponibilizados buscam analisar a sua procedência, dando preferência àqueles que são originários de fontes confiáveis. Por outro lado, tais dados são ofertados de modo distribuído, não sendo incomum a ausência de hiperlinks para informações relacionadas, ora armazenadas no mesmo repositório de dados ou não (GALIOTOU; FRAGKOU, 2013).

O desafio presente consiste no fornecimento de meios eficazes para acessar dados das origens distribuídas, e ainda, estipular mecanismos através dos quais eles podem ser conectados e integrados (HEATH, 2011). Outro desafio reside na limitação dos seres humanos em processar e conectar a atual oferta de dados e informações disponíveis, considerando que a internet faz com que a riqueza do conhecimento humano esteja disponível para qualquer pessoa, em qualquer lugar. Mais um desafio reside em como classificar e efetivamente utilizar o crescente volume de informação disponível para a obtenção das respostas necessárias.

Neste contexto, Alcantara et al. (2015) posicionam que, a publicação de dados em formato aberto e estruturado não é o bastante para permitir que aplicações enriqueçam suas bases de conhecimento. O ideal é que sejam publicados de acordo com práticas que

permitam a interoperabilidade de dados na *Web* através do uso de vocabulários descritos em RDF, facilitando a sua utilização por serviços automatizados para o consumo dos dados.

Como resposta a estes desafios, o conceito de dados conectados emerge visando orientar as organizações a ofertarem seus dados existentes disponíveis em formatos legíveis por máquina (BAUER; KALTENBÖCK, 2012). Dados Conectados referem-se a um conjunto de boas práticas para publicação e conexão de dados estruturados na *Web* utilizando padrões internacionais do World Wide *Web* Consortium - W3C, permitindo o estabelecimento de uma rede de dados que se conectam e auto enriquecem (HEATH, 2011; WOOD et al., 2013).

Dados conectados para serem considerados como tal, precisam obedecer quatro princípios que são: (1) Devem ser usados Identificadores Universais de Recursos (Universal Resource Identifier - URIs) como nomes para as coisas a serem publicadas; (2) Devem ser usadas URIs HTTP para que os usuários possam localizar estes nomes; (3) Quando a URI for encontrada, ela deve prover informação útil, usando padrões como o RDF (Resource Description Framework) ou o SPARQL; (4) e ainda, as URIs devem incluir hiperlinks para outras URIs, para que os usuários possam descobrir novas coisas que se relacionem a URI que ele esteja buscando (BERNERS-LEE, 2006).

Dados Abertos Conectados respeitam os mesmos princípios aplicáveis aos Dados Conectados (BANDEIRA et al., 2014), incorporando obrigatoriamente requisitos dos Dados Abertos como o uso e reúso livre, podendo ser redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras.

Desta maneira podemos definir os Dados Abertos Conectados como *"um conjunto de práticas para publicação de dados abertos que possuem hiperlinks para outros dados abertos, mediante o uso de URIs que garantem que a partir de um dado, possam ser acessados outros dados relacionados"*.

A evolução dos Dados Abertos para os Dados Abertos Conectados foi estabelecido pela escala de maturidade *5-Stars Linked Open Data* (BERNERS-LEE, 2006) que será explorada mais adiante nesta pesquisa.

A iniciativa ISA - *Interoperability Solutions for European Public Administrations* - conforme apresentado na Tabela 3, estabelece uma importante comparação entre as características dos Dados Abertos Conectados com outros formatos de dados estruturados (ISA, 2014).

Para acompanhar o desenvolvimento dos Dados Abertos Conectados, foi estabelecido o projeto *"The Linking Open Data Cloud Diagram"*, conhecido como *"LOD Cloud"*, mantido pelos pesquisadores Richard Cyganiak (*Insight Centre for Data Analytics at NUI*

Tabela 3 – Comparativo entre características de Dados Conectados e Outros formatos de dados estruturados

Dimensão	**Situação Atual (Dados não-conectados)**	**Situação desejada (Dados Conectados)**
Compartilhamento de dados:	Dados são compartilhados utilizando o XML (eXtensible Markup Language)	Dados são compartilhados utilizando o RDF (Resource Description Framework)
Validação de dados:	Utilização do *XML Schema (XSD)* como meio para validação dos dados	Regras (ex: SWRL) e SPARQL utilizados como meio para validar dados e fazer inferências
Significado de dados:	Estrutura centralizada para prover significado aos dados, mediante a reutilização de arquivos XSD	Estruturas descentralizadas para associação de significado aos dados como vocabulários e outros dados de referência
Provimento de dados:	Serviços *Web* especializados baseados em SOAP permitem o acesso aos dados	Serviços RESTful leves de dados conectados permitem o acesso aos dados
Integração de dados:	Integração de Sistemas	Conexões de dados baseadas em elementos semânticos
Relacionamento com outros conceitos:	Dados e esquemas são considerados completos (mundo fechado)	Dados e esquemas são considerados incompletos (mundo aberto)
Endereçamento de dados:	Cada recurso (ou entidade) de dados possui um único identificador no nível dos sistemas de informação	Cada recurso (ou entidade) de dados possui identificadores comuns e conectados, por diferentes sistemas de informação, no nível da *Web*

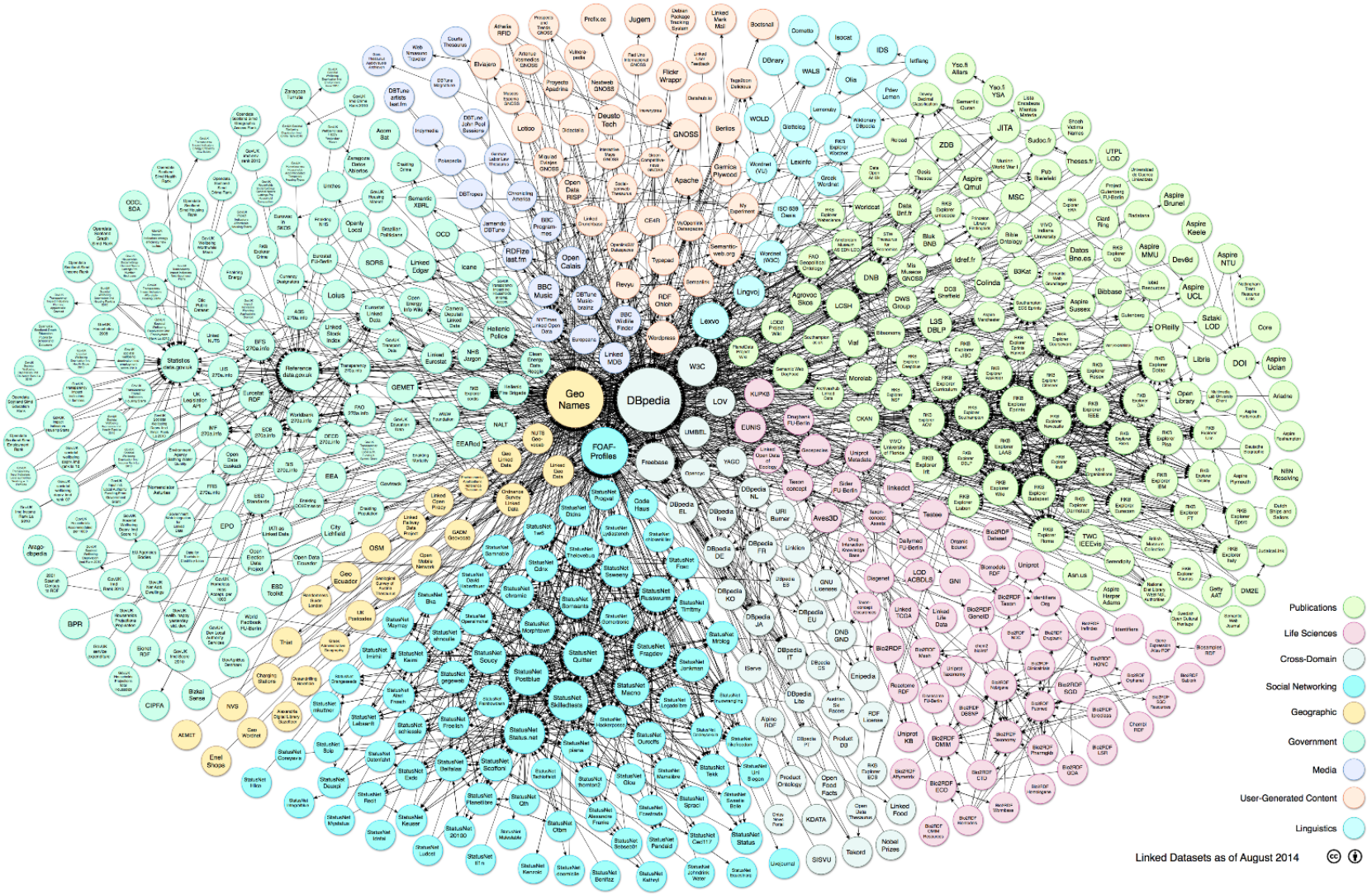
Fonte: ISA (2013)

Galway[1]) e Anja Jentzsch (*Hasso Plattnner Institut*[2]). A *"LOD Cloud"* cataloga e disponibiliza uma imagem que mostra os conjuntos de dados que foram publicados como dados conectados, por contribuintes da comunidade de Dados Abertos Conectados (*Linking Open Data*), bem como outros indivíduos e organizações. A "nuvem" é baseada em metadados coletados mediante a curadoria de contribuintes para o Hub de Dados organizados na *"LOD Cloud"*.

A imagem disponibilizada no projeto é interativa e onde cada conjunto de dados é representado por um circulo contendo um hiperlink para sua página inicial. Atualmente a *"LOD Cloud"*, conforme a Figura 6 conta com 570 datasets, contendo mais de 31 bilhões de triplas e mais de 500 milhões de conexões entre esses conjuntos de dados.

[1] Disponível em http://www.insight-centre.org
[2] Disponível em http://www.hpi.de

Figura 6 – Nuvem de Dados Abertos Conectados estabelecido pelo projeto *LOD Cloud*

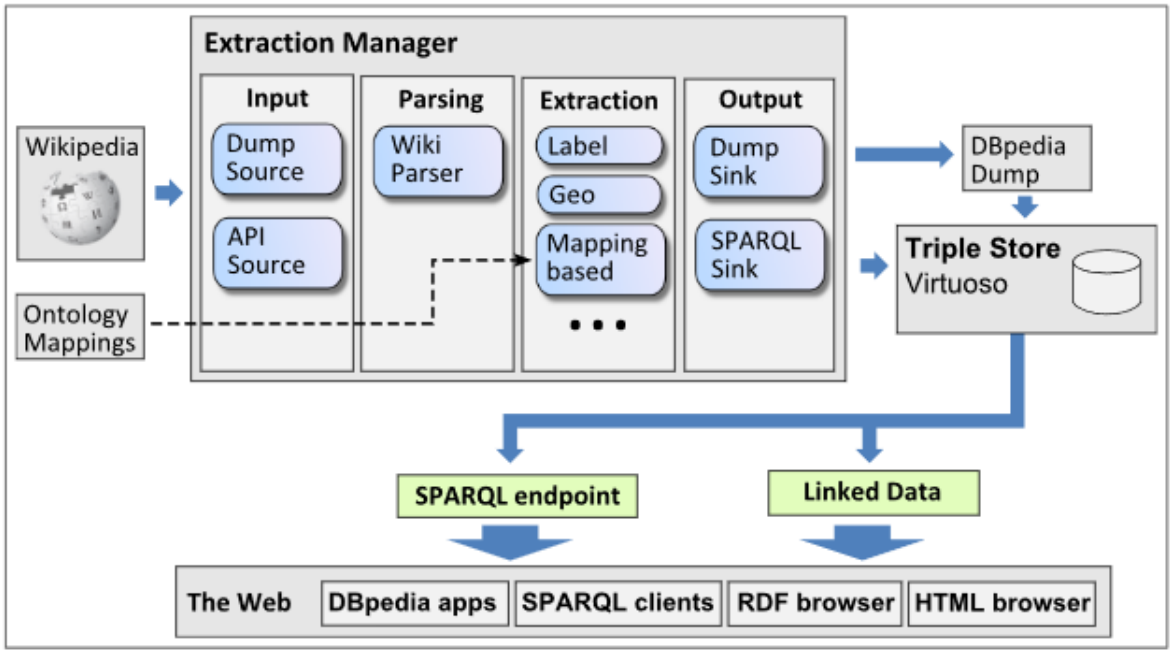


Fonte: *"The Linking Open Data Cloud Diagram"*, disponível em <http://lod-cloud.net/>. Acesso em: 25 out. 2015

A maior iniciativa de Dados Abertos Conectados do mundo é a DBPedia, que é um projeto colaborativo para extração de dados da Wikipédia tornando-os disponíveis e recuperáveis na *Web*, permitindo a realização de consultas sofisticadas conectando diferentes conjuntos de dados existentes na Wikipédia. Dentre outros benefícios, a DBPedia permite novos mecanismos de navegação, conexão de dados e aprimoramento da Wikipédia.

De acordo com Bandeira et al. (2015) a base de conhecimento da DBPedia, versão em inglês, descreve mais de 4 milhões de entidades classificadas em uma ontologia. Essa base de conhecimento contempla mais de 1,4 milhões de pessoas, cerca de 735 mil lugares, descrição de mais de 400 mil conteúdos multimídia (álbuns musicais, filmes, etc.), 241 mil organizações (contemplando 58 mil empresas e 41 mil instituições de ensino), dentre outros conteúdos relevantes. A DBPedia completa, em todos os idiomas atualmente contemplados, já contém 38,3 milhões de entidades armazenadas, contendo 25,2 milhões de links para imagens, 29,8 milhões de links para páginas externas e 80,9 milhões de links para categorias da Wikipédia. Além disso, está conectada com cerca de 50 milhões de outros conjuntos de dados conectados. Em 2014, a DBPedia alcançou um volume de três bilhões de informações estruturadas em triplas RDF (DBPEDIA, 2015).

Figura 7 – Visão geral do framework de geração e disponibilização de dados conectados da DBPedia



Fonte: Bandeira et al. (2015 apud LEHMANN et al., 2014)

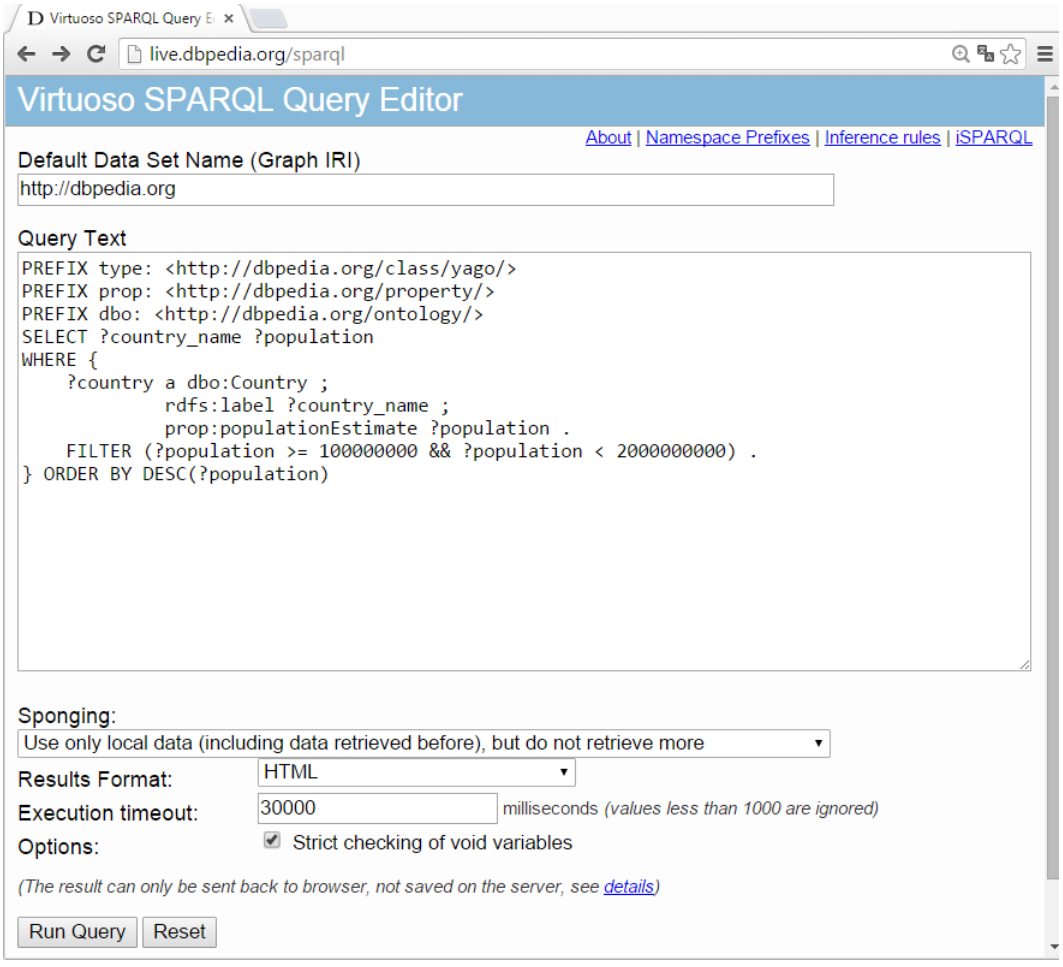
A produção de dados conectados da DBpedia é desenvolvida mediante diversas etapas, que são a leitura, interpretação, extração de dados a partir da Wikipédia, enriquecimento dos dados com o apoio de ontologias e geração de triplas RDF, que são armazenadas em um servidor de triplas. A partir desse servidor é disponibilizado um endpoint SPARQL [3]. Os

[3] Disponível em http://live.dbpedia.org/sparql

dados conectados são consumidos e visualizados por páginas HTML (*HyperText Markup Language*) geradas a partir das triplas que proporcionam uma nova e rica experiência de navegação em torno dos dados da Wikipédia. A Figura 7 apresenta uma visão geral do framework de geração de dados conectados da DBPedia.

Através desse framework, a base da DBPedia possibilita a execução de diversas consultas como: (a) "Qual a população de um determinado país?" ou (b) "Quais os países que possuem população entre 100 milhões e 2 bilhões de habitantes?". As Figuras 8 e 9 apresentam um exemplo de consulta SPARQL para executar a consulta (b) e apresentação da respectiva resposta.

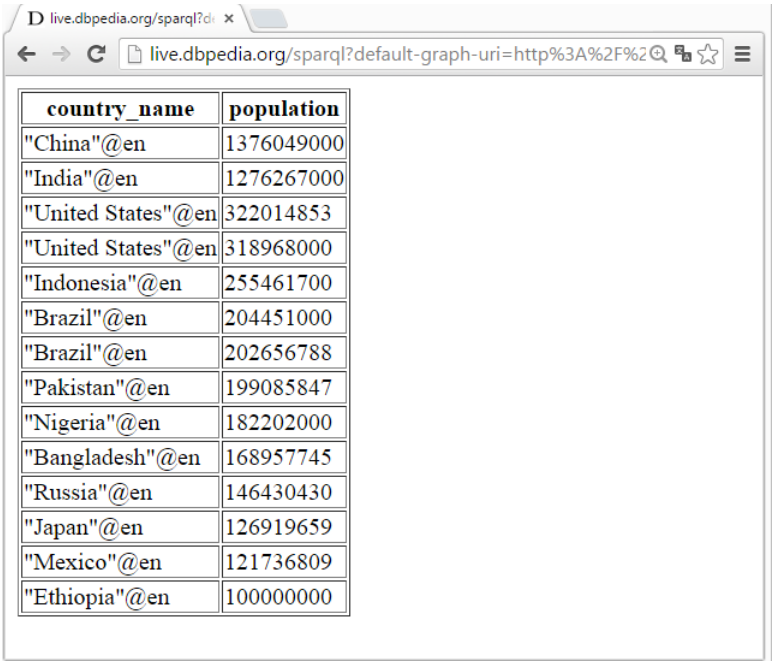
Figura 8 – Consulta SPARQL na DBpedia para obter os países que possuem a população entre 100 milhões e 2 bilhões de pessoas



Fonte: Autor desta dissertação, a partir de http://live.dbpedia.org/sparql, 2015.

A DBpedia, devido a sua relevância quanto à oferta de Dados Conectados no mundo, está sendo utilizada como fonte de conhecimento para cerca de projetos e diversos casos reais de uso, sendo utilizada por grandes empresas ao redor do mundo, como o conglomerado de mídia britânico BBC, dentre muitas outras empresas (KOBILAROV et al., 2009).

Figura 9 – Resultado de consulta SPARQL na DBpedia para obter os países que possuem a população entre 100 milhões e 2 bilhões de pessoas



country_name	population
"China"@en	1376049000
"India"@en	1276267000
"United States"@en	322014853
"United States"@en	318968000
"Indonesia"@en	255461700
"Brazil"@en	204451000
"Brazil"@en	202656788
"Pakistan"@en	199085847
"Nigeria"@en	182202000
"Bangladesh"@en	168957745
"Russia"@en	146430430
"Japan"@en	126919659
"Mexico"@en	121736809
"Ethiopia"@en	100000000

Fonte: Autor desta dissertação, a partir de http://live.dbpedia.org/sparql, 2015.

Ademais, segundo Bandeira et al. (2014), no contexto governamental, os Dados Abertos Conectados contribuem para uma maior exploração de dados governamentais abertos permitindo uma maior garantia de transparência nas transações e permitindo também que os dados sejam legíveis por máquinas.

2.3.2 Dados Abertos Governamentais

Conforme os princípios estabelecidos por Reilly (2010) na Figura 4, a ampla publicação de dados e informações governamentais online e a melhoria sistemática da qualidade da informação governamental são fundamentais para a implantação de um cenário de Governo Aberto. De acordo com GuimarÃes e Diniz (2014), os dados abertos podem contribuir para maior controle social e promover transparência nas ações governamentais, ou seja, os dados abertos podem servir, portanto, para que os cidadãos e as organizações da sociedade possam reutilizá-los com o intuito de verificar, esclarecer, fiscalizar e acompanhar questões de seus interesses.

No cenário atual, o conceito de Dados Abertos vem ao encontro da forte demanda da sociedade da informação e do conhecimento pela promoção da transparência e da abertura governamental. Desta maneira, podemos entender que Dados Abertos Governamentais são dados produzidos pelos governose colocados à disposição das pessoas de forma a tornar possível não apenas sua leitura e acompanhamento, mas também sua reutilização em novos projetos, sítios e aplicativos; seu cruzamento com outros dados de diferentes fontes; e sua disposição em visualizações interessantes e esclarecedoras. (W3C Brasil, 2011). Foram

estabelecidas três leis dos dados abertos governamentais, contendo as condições para que um determinado dado governamental seja considerado como aberto (EAVES, 2009):

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

Complementarmente, a *The Association for Computing Machinery - ACM* publicou uma recomendação sobre dados governamentais onde estabeleceu que

> "Os dados publicados pelo governo devem ser em formatos e abordagens que promovam a análise e reutilização desses dados" OGD (2007 apud ACM, 2009).

A relevância dos dados abertos governamentais vem da habilidade do público de realizar suas próprias análises dos dados brutos, em vez de depender de uma análise do próprio governo. Cumpre destacar esta característica de que os dados governamentais abertos devem ser produzidos e estruturados para serem processados por máquinas, ampliando a sua capacidade de processamento, interpretação e produção de informações e conhecimento a partir deles.

Além das três leis foram estabelecidos os princípios dos dados governamentais abertos, mediante consenso de especialistas no tema. Tais princípios foram reforçados em publicações posteriores, por diversas instituições como a *SunLight Foundation, Association for Computing Machinery, The White House*, dentre outros (OGD, 2007). Segundo este consenso, os dados governamentais abertos precisam ser *Completos, Primários, Atuais, Acessíveis, Processáveis por Máquina, prover Acesso não-discriminatório, possuir formatos não-proprietários e livres de licenças restritivas.*

A não-abertura de dados no setor público pode resultar em diversos problemas para o desenvolvimento socioeconômico. Dentre estes, podemos destacar:

- Gastos duplicados de recursos públicos na produção de dados;
- Alto custo na localização de dados e informações;
- Conflitos de propriedade de dados entre pessoas físicas (funcionários públicos) e pessoas jurídicas (órgãos governamentais);
- Descumprimento da legislação vigente;
- Indisponibilidade de dados suficientes e confiáveis para subsidiar a tomada de decisão;

- Ampliação do custo da produção técnica e científica, decorrente da indisponibilidade de dados;
- Impedimento à criação e aprimoramento de negócios digitais que demandem consumo regular de dados de determinadas fontes;
- Prejuízo à formulação, execução e monitoramento de políticas públicas.

Desta maneira considerando estes problemas listados decorrentes da não-abertura de dados governamentais, associados às suas finalidades e aplicações conforme explanado na Tabela **??**, é necessário que tais dados sejam reutilizáveis e subsidiem a tomada de decisão, ou seja, precisam ser ofertados em formatos e condições que proporcionem a apropriação e o reaproveitamento por parte dos cidadãos e demais segmentos da sociedade. Neste contexto, a próxima subseção apresentará algumas experiências de produção de Dados Abertos Conectados Governamentais como perspectiva à produção de Dados Abertos Governamentais para os próximos anos.

2.3.3 Dados Abertos Conectados Governamentais

Complementando a conceituação sobre Governo Aberto e Dados Abertos Governamentais, Hyland e Wood (2011) comentam que, apesar das diversas e relevantes iniciativas de abertura governamental e maior oferta de dados públicos à sociedade, o atual modelo de publicação de dados estruturados na *Web* é insuficiente, pois para fazer um melhor uso destes dados, as pessoas precisam cruzá-los com outros dados. O cruzamento de dados abertos requer a utilização de diversos softwares além da necessidade de *download*, em alguns casos de grandes arquivos de dados, tornando o consumo dos dados abertos governamentais menos atraente para os cidadãos e consequentemente, reduzindo o seu potencial de uso.

Além deste desafio no cruzamento de grandes arquivos de dados, os dados abertos (sejam governamentais ou não), deixam a desejar quanto aos requisitos semânticos, que permitem seu processamento automatizado. Hyland e Wood (2011) exemplificam este cenário mediante o discurso do pesquisador Robert Schaefer, ao discursar sobre mudanças climáticas na Conferência Internacional para Governo Aberto em 2010, onde mencionou:

> "Dispor de dados abertos governamentais é algo ótimo, mas é muito difícil tomar decisões sensatas sem o contexto do que significa tais dados. Atualmente, é necessário envolver analistas e cientistas para gerar conhecimento a partir de tais dados. O desafio é fazer chegar estas informações aos tomadores de decisões políticas. Então nós temos que simplificar a oferta destas informações aos decisores políticos."

Buscando sintetizar a necessidade da melhoria da oferta de dados abertos governamentais, ISA (2013) complementa que a abertura de dados, como por exemplo, nos portais de Dados Abertos, muitas vezes acontece de uma forma *"ad-hoc"*, resultando, na maioria das vezes, em milhares de conjuntos de dados online que são publicados sem atender aos padrões de dados e metadados mínimos e ainda, sem a reutilização de identificadores comuns. Assim, uma "esfera de dados" fragmentada é criada, onde encontrar, reutilizar, integrar e dar algum sentido a tais dados de diferentes fontes consiste num verdadeiro desafio.

Os Dados Abertos Conectados Governamentais (DACG) são uma resposta a tais desafios e podem ser facilitadores para uma grande transformação no ambiente de Governo Eletrônico (*e-government*). DACG consiste numa forma de identificar e conectar dados abertos governamentais de acordo com os princípios dos dados conectados estabelecidos por Berners-Lee (2006). Ou seja, podemos definir os Dados Abertos Conectados Governamentais como "dados abertos governamentais publicados de acordo com os princípios dos dados conectados" (ISA, 2013).

Considerando a citação de Robert Schaefer, é exatamente em casos como este que os dados de natureza conectada são aplicáveis, permitindo esquemas abertos, extensíveis e com ricas informações explicativas que encaminham o seu usuário para outros dados e informações relacionados e disponíveis na *Web*. A abertura de dados de forma conectada, obedecendo aos padrões internacionais, facilitam a recuperação e compreensão de dados e informações associadas a um contexto que facilitam o seu entendimento (HYLAND; WOOD, 2011).

O ecossistema de dados abertos conectados governamentais é complexo, pois envolve um grande conjunto de partes interessadas, dados, tecnologias de produção, processamento, armazenamento e visualização de dados, recursos semânticos e padrões estabelecidos. Dependendo de sua aplicação, pode beneficiar diversos propósitos como os listados na Tabela 2, atendendo a consumidores individuais e corporativos.

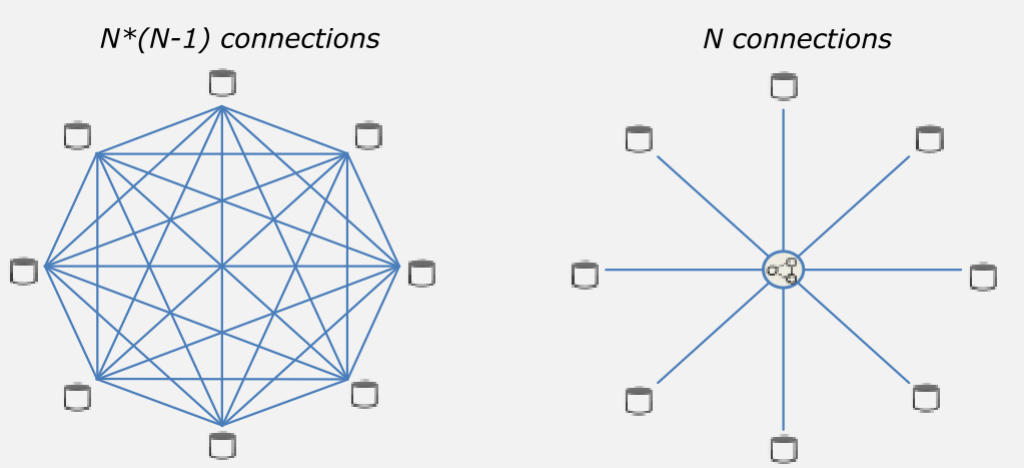
2.3.3.1 Ecossistema de Dados Abertos Conectados Governamentais

De acordo com ISA (2013), no ecossistema DACG as instituições governamentais são fornecedores de dados que disponibilizam dados abertos governamentais como serviços DACG para diversos consumidores de dados - cidadãos, empresas e outros órgãos governamentais. Ao invés de fazer o *download* e processamento de conjuntos de dados inteiros (em alguns casos, em arquivos enormes), DACG permite que um consumidor de dados obtenha informações específicas do seu interesse, de determinada entidade governamental, através da resolução dos Identificadores Universais de Recursos (URIs). Os dados são fornecidos em diferentes formatos legíveis por máquina, pronto para serem conectados e mesclados com outros dados. Este referencial teórico nos permite estabelecer a seguinte proposição de valor para DACG:

- DACG oferece recursos flexíveis de integração de dados governamentais;
- DACG leva a um aumento na qualidade dos dados governamentais;
- O uso de DACG potencializa o desenvolvimento de novos serviços; e
- DACG reduz os custos de integração de dados governamentais.

Entretanto, para tornar possível esta proposição de valor, provedores de DACG precisam ter uma política de Identificadores Universais de Recursos (URIs) para que os consumidores possam contar com serviços DACG confiáveis e que outros fornecedores de dados possam se conectar com estas URIs ou reutilizá-las para identificar conceitos relacionados. O uso de URIs para identificar conceitos semelhantes em conjuntos de dados díspares é um pré-requisito para desbloquear e potencializar o efeito positivo de rede de DACG (ISA, 2013). As Figuras 10 e 11 exemplificam este efeito positivo bem como o potencial de redução de custos de integração de dados.

Figura 10 – Número de conexões de dados não-conectados (à esquerda) e dados conectados (à direita)

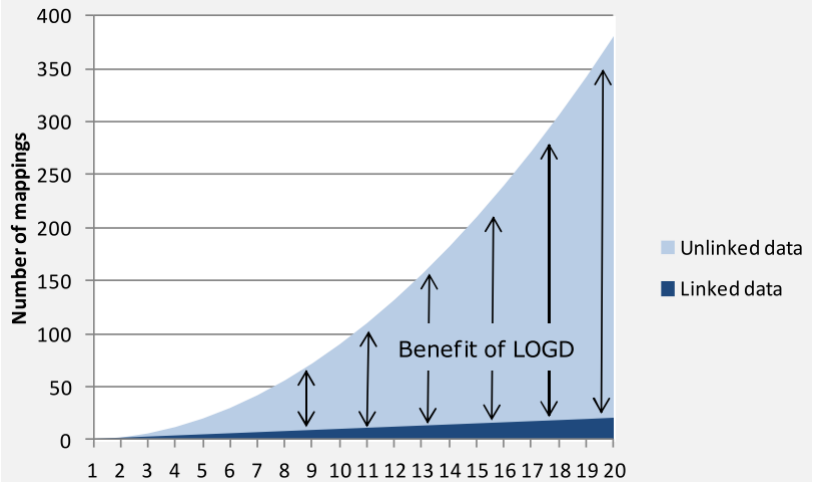


Fonte: ISA (2013)

As consequências do investimento em DACG são relevantes. A relação abaixo apresenta alguns benefícios relevantes para os governos ao considerarem a adoção de DACG (ISA, 2013):

- **Efeitos positivos de rede:** A publicação de DACG pode reduzir os custos da resolução de problemas de interoperabilidade na troca de informações, consequentemente facilitando a integração de dados;
- **Ampliação da transparência governamental e do acesso democrático a informação governamental:** Dados disponibilizados como DACG, como explanados, permitindo que o cidadão consuma exatamente o dado que ele deseja ao

Figura 11 – Gráfico com projeção de custos de uso de dados não-conectados e dados conectados



Fonte: ISA (2013)

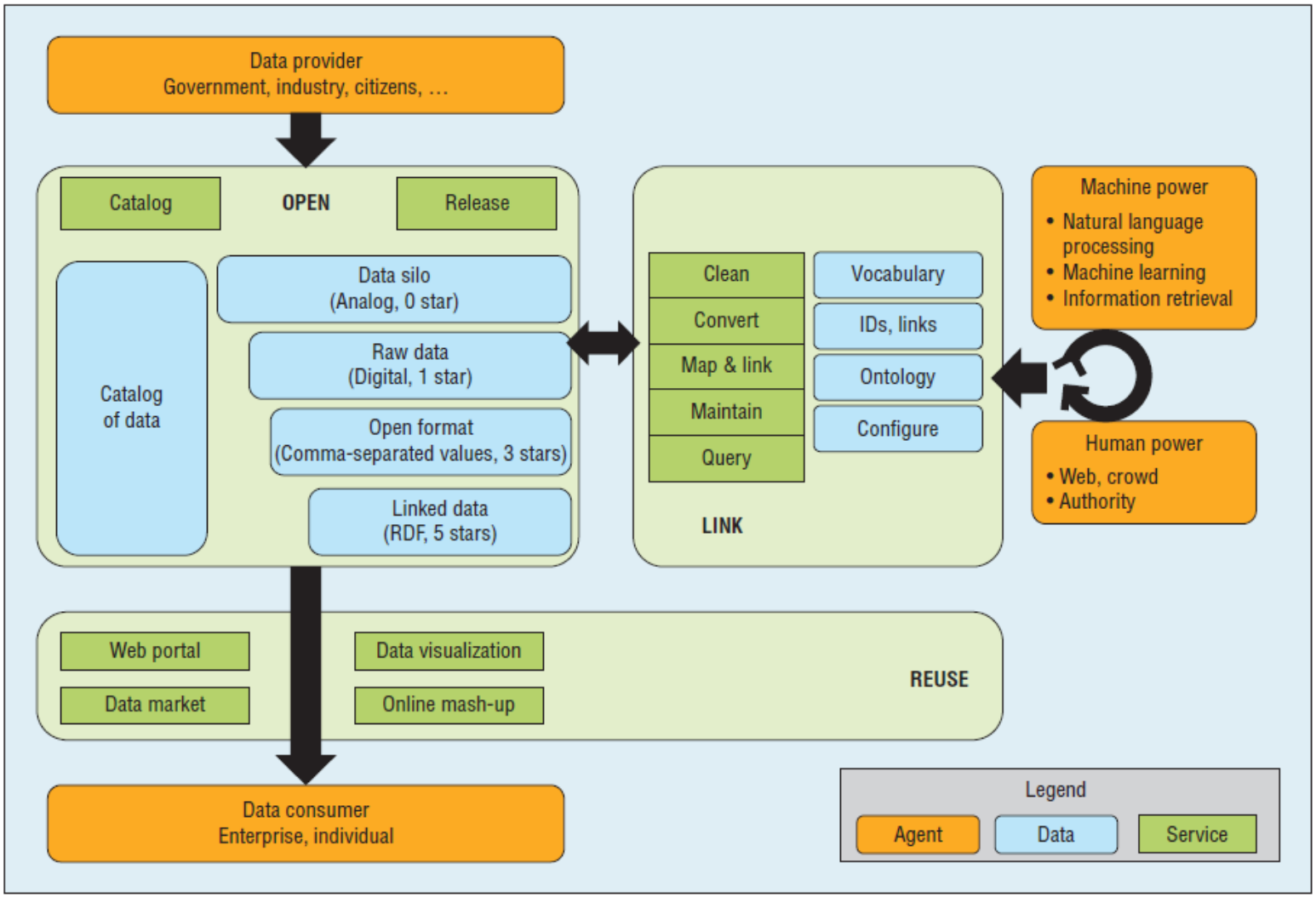
menor custo possível. Além disso, ao garantir serviços de DACG confiáveis, a tendência é que os cidadãos não necessitem fazer mais *downloads* e diversas cópias de dados governamentais, bem como ajudam a combater a corrupção e vantagens econômicas no acesso aos dados públicos, onde agentes públicos mal intencionados podem querer estabelecer vantagens para entregar as informações específicas desejadas pelos cidadãos. Pode ainda contribuir para a redução de atividades econômicas danosas ao governo, como por exemplo, a atuação de empresas que comercializam dados e informações (como imagens de satélite) impondo licenças para que órgãos governamentais sejam obrigados a comprar várias cópias destas informações, não podendo compartilhar tais dados entre si;

- **Economia de escala:** Consequente dos dois benefícios anteriores. Além disso, o custo de produção de dados abertos governamentais de forma conectada irá reutilizar de toda a infraestrutura já existente para a oferta de dados abertos governamentais, sendo necessário poucos investimentos adicionais;

- **Garantia de estabilidade e persistência aos dados:** Proporciona que um determinado serviço de informações estará sempre disponível mediante princípios públicos, tendo o poder público como guardião. Desta maneira, novas atividades e negócios poderão ser desenvolvidos a partir de tais dados sem o risco de que os mesmos possam ficar indisponíveis ou venham a ser cobrados caso sejam fornecidos por alguma empresa privada.

A Figura 12 apresenta uma visão sistêmica do ecossistema DACG destacando quatro atores principais que participam, interagem ou influenciam diretamente, conforme explicado nos tópicos a seguir(DING; PERISTERAS; HAUSENBLAS, 2012; ISA, 2013):

Figura 12 – Ecossistema de dados abertos conectados governamentais



Fonte: Ding, Peristeras e Hausenblas (2012)

- Os **provedores de dados (*data providers*)**, ou seja, as instituições governamentais que abrem seus dados os disponibilizam como DACG (realizando todas as atividades que fazem parte das etapas de abertura (*Open*) e conexão (*Link*) destacadas na Figura 12.

- **Consumidores de dados (*data consumers*)**, ou seja, cidadãos, empresários, empresas e público outras instituições governamentais que reutilizam os dados disponíveis como DACG, realizando atividades como as destacadas na etapa de reúso (*Reuse*) apresentada na Figura 12. Estes consumidores buscam, com DACG, gerar valor agregado para suas aplicações e serviços que venham a desenvolver e disponibilizar para seus clientes. A distinção entre os fornecedores de dados e os consumidores pode não ser nítida, pois uma instituição que fornece dados pode, no mesmo tempo, também consumir os dados de outra instituição. Esse entrelaçamento é comum num ecossistema DACG.

- **Corretores (ou intermediadores) de dados (*data brokers*)**, ou seja, organizações de terceiros, sejam privadas ou públicas, que estabelecem catálogos de dados e espaços de intercâmbio (*marketplaces*) que facilitem o acesso ao DACG disponível. Em alguns casos, estes intermediários também oferecem serviços adicionais, tais como consultas avançadas, visualização de dados, ou exportação/conversão de dados para diferentes formatos,

- **Entidades reguladoras**, nomeadamente os governos nacionais/regionais/locais bem como instituições internacionais, tais como a Comissão Europeia, que regulamenta o provimento de DACG através de políticas, leis e diretrizes.

O estabelecimento de ecossistemas desta natureza permite que haja uma cooperação natural entre publicadores e consumidores de DACG sem haver um grande esforço regulatório ou de conversão de dados para cada necessidade de consumo. Diversos produtores e consumidores (incluindo os protoconsumidores) de dados podem cooperar efetivamente para ofertar dados que podem ser reutilizados e combinados por outras pessoas ou empresas (HYLAND; WOOD, 2011).

2.3.3.2 Iniciativas de Dados Abertos Conectados Governamentais

Apesar de ser um tema recente, DACG está sendo desenvolvendo ao longo de vários países do mundo. Por exemplo, o relatório *"Study on business models for Linked Open Government Data"*, da European Comission (ISA, 2013), identificou diversas iniciativas de uso de DACG em diversos países do continente conforme a Tabela **??**.

Como exemplo, legislation.gov.uk é um serviço que disponibiliza a legislação de todo o Reino Unido. Atualmente estão disponíveis 101 mil registros referentes a leis e outros tipos de atos normativos. Os grandes desafios consistem em, prover acesso rápido às

Tabela 4 – Iniciativas de Dados Abertos Conectados Governamentais em países da União Européia

País:	**Iniciativa:**
Alemanha	German National Library [4]
Áustria	Renewable Energy and Energy Efficiency Partnership (REEEP) [5] Austrian Geological Survey (GBA) - [6]
Bélgica	Vlaams Theater Instituut – Travelogue [7]
Dinamarca	Danish Agency for Digitisation [8]
Espanha	AEMET – Spanish Meteorological Office [9]
Itália	Agenzia per lÍtalia Digitale [10] Regione Emilia-Romagna [11] Trentino government linked open geo-data [12]
Países Baixos	Amsterdam-Amstelland Fire Department [13] Building and address register [14] Stelselcatalogus: linked metadata of Dutch base registers [15]
Reino Unido	BBC [16] National Archives [17] Ordnance Survey [18] Food and Agriculture Organisation of the United Nations (FAO) [19]

Fonte: ISA (2013)

demandas de acesso a legislação do Reino Unido; Prover rastreabilidade sobre as mudanças na legislação; bem como, prover acesso a estes dados por máquinas.

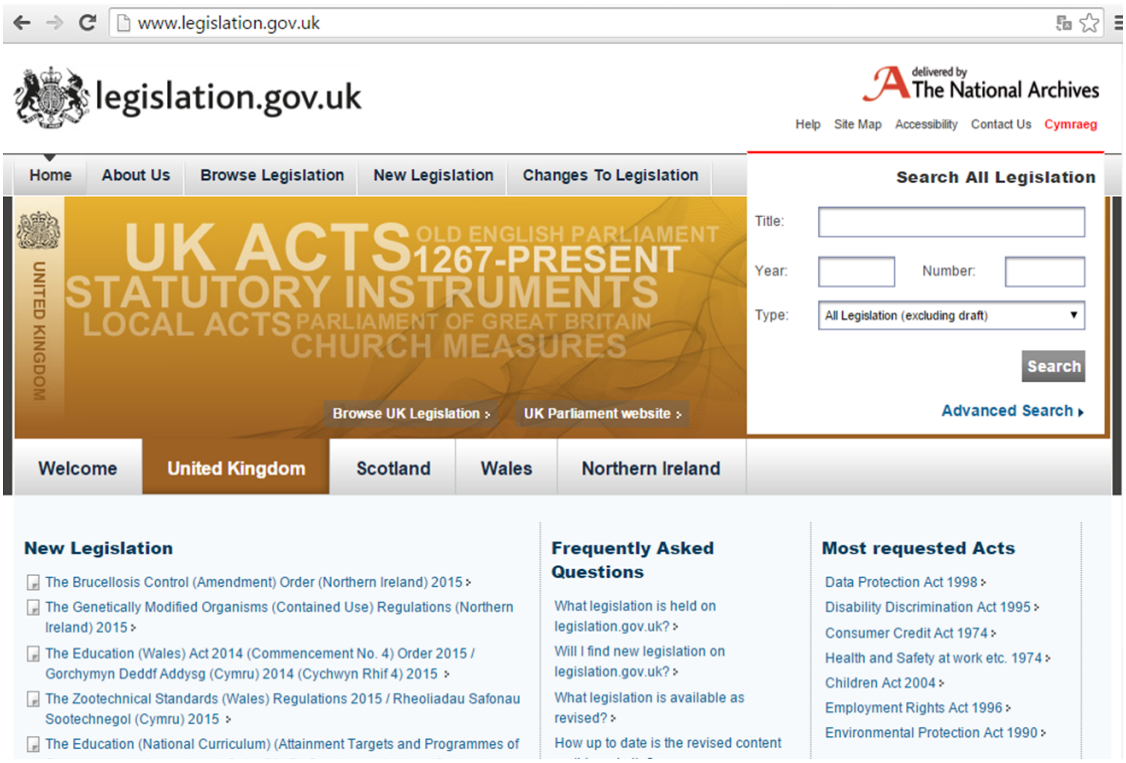
Para isto, o Governo do Reino Unido estruturou uma base de dados conectados contendo toda a legislação disponível neste serviço. Dentre outras medidas, estabeleceu uma política para URIs que estabelece:

- URIs de identificadores: por exemplo, os dados referentes ao ato normativo *"The Transport Act 1985"* estão acessíveis em http://www.legislation.gov.uk/id/ukpga/1985/67
- URIs para documentos: por exemplo, a última versão do ato normativo *"The Transport Act 1985"* está acessível em http://www.legislation.gov.uk/ukpga/1985/67
- URIs para representação: por exemplo, a última versão do ato normativo *"The Transport Act 1985"* no formato XML está disponível em http://www.legislation.gov.uk/ukpga/1985/67/data.xml

Além disto, toda a base de leis e atos normativos está disponível através de um endpoint SPARQL. As Figuras A e B apresentam respectivamente uma consulta SPARQL e o resultado no formato JSON (processável por máquina):

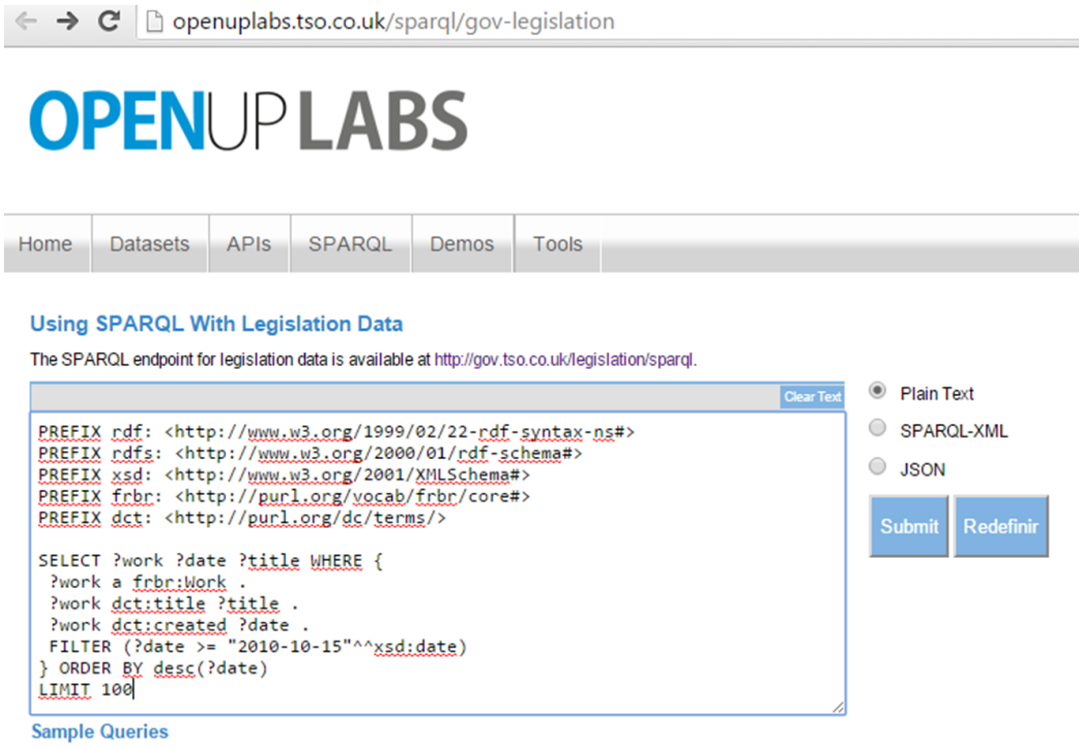
No contexto brasileiro, segundo o CeWeb.BR (2015), o Governo Federal publicou o orçamento federal em formato RDF (no período de 2000 a 2014), a fim de dar maior

Figura 13 – Tela principal do legislation.gov.uk



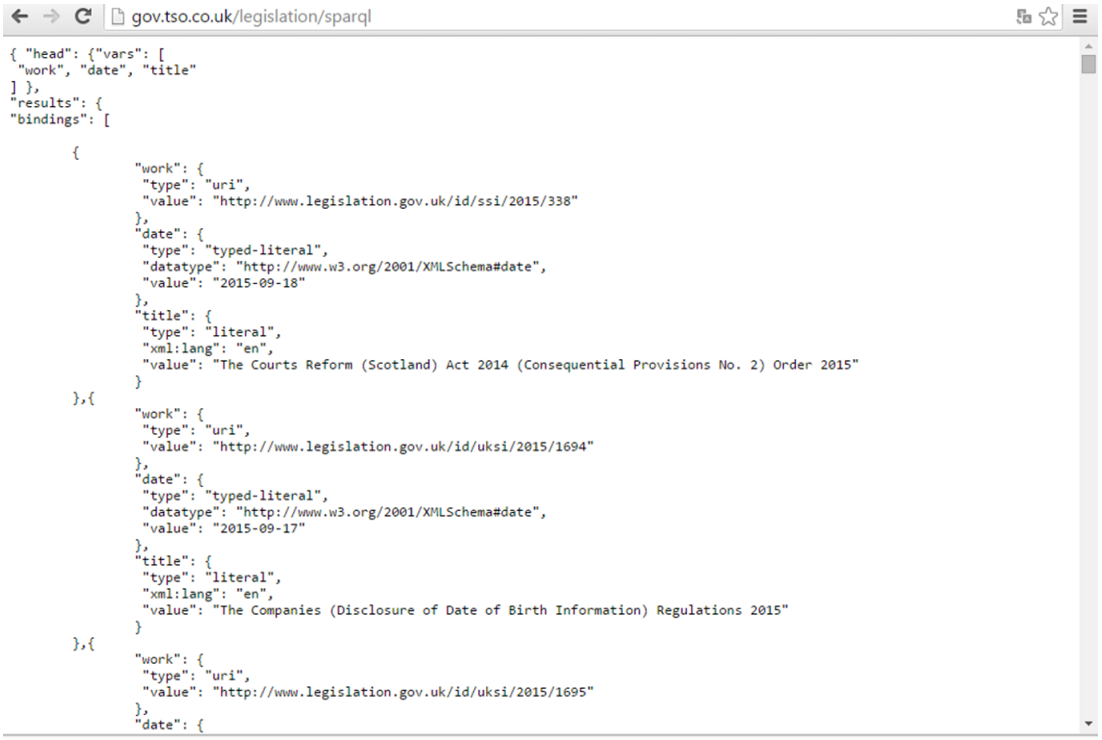
Fonte: http://legislation.gov.uk. Acesso em: 25 out. 2015

Figura 14 – Tela de consulta do endpoint SPARQL do legislation.gov.uk



Fonte: http://legislation.gov.uk. Acesso em: 25 out. 2015

Figura 15 – Resultado de consulta realizada no endpoint SPARQL do legislation.gov.uk



Fonte: http://legislation.gov.uk. Acesso em: 25 out. 2015

transparência e acesso aos dados, de forma que cidadãos e organizações interessados em conhecer melhor os dados do orçamento federal possam realizar pesquisas e análises de forma veloz e eficiente. A estratégia adotada no projeto "Orçamento Federal em Formato Aberto", do Governo Federal, foi desenvolvida mediante a criação de uma ontologia [20] com base na classificação da despesa do orçamento federal, contemplando as categorias e conceitos especificados no Manual Técnico de Orçamento [21]. A Figura 16 apresenta a estruturação do conceito referente ao "ItemDeDespesa" no âmbito da ontologia desenvolvida para experiência brasileira.

De acordo com Araújo et al. (2013), na versão inicial do projeto que considerou o Orçamento Federal do ano de 2012, os conceitos da ontologia foram utilizadas para converter as informações constantes em bancos de dados relacionais em triplas, resultando num arquivo RDF com aproximadamente 825 mil triplas. As informações do orçamento estão disponíveis no do catálogo de dados abertos do governo federal [22], permitindo acesso aos *dumps* dos dados do Orçamento Federal, padrões de URIs e acesso ao endpoint SPARQL do projeto [23].

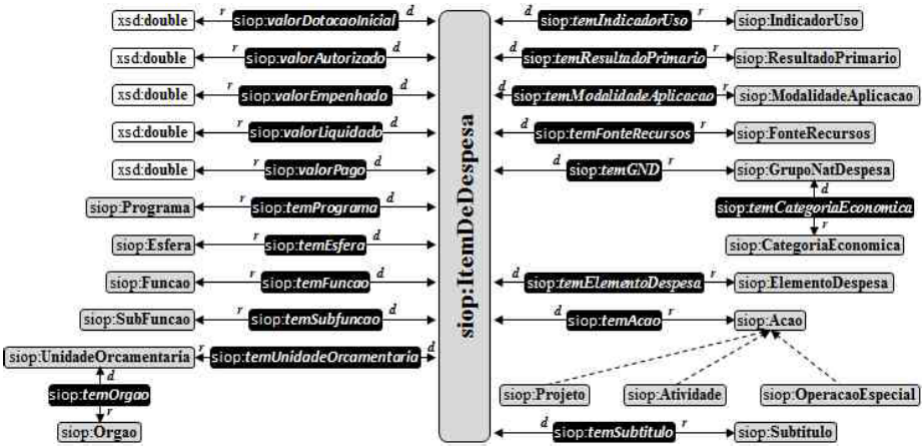
[20] Ontologia do Orçamento do Governo Federal Brasileiro - disponível em http://vocab.e.gov.br/2013/09/loa.owl

[21] Disponível em http://www.orcamentofederal.gov.br/informacoes-orcamentarias/manual-tecnico/mto_2015_1a_edicao-150514.pdf

[22] Disponível em http://dados.gov.br/dataset/orcamento-federal

[23] Orçamento Federal - OpenLink Virtuoso SPARQL Query - disponível em http://orcamento.dados.gov.br/sparql/

Figura 16 – Ontologia das Classificações da Despesa do Orçamento Federal do Brasil



Fonte: Araújo et al. (2013)

Outra iniciativa relevante no Brasil é o projeto SPUK - Melhorando o ambiente de negócios por meio da transparência no Governo de São Paulo [24], decorrente de uma cooperação do Estado de São Paulo com o Governo do Reino Unido. Segundo este projeto, os dados do governo devem servir além do próprio governo, devem ser matéria prima para a transparência, para a construção de novos serviços e para a criação de novos negócios pela sociedade. Ou seja, apresenta uma visão inovadora onde os dados governamentais se apresentam como relevante insumo para o desenvolvimento econômico.

No que trata de Dados Abertos e Dados Abertos Conectados, o SPUK tem como objetivos ampliar em 70% o total de bases de dados governamentais do Governo do Estado de São Paulo em formato aberto, e ainda, ter pelo menos 3% das bases abertas em dados estruturados conforme à *Web* semântica (onde se inclue os Dados Abertos Conectados), com pilotos em pelo menos uma das seguintes áreas: saúde, transporte e educação. Outro objetivo relacionado visa o desenvolvimento de um programa de fomento à utilização de dados abertos governamentais por meio da criação de aplicativos para melhoramento de serviços públicos e promoção de oportunidade de negócios (GOVERNO DO ESTADO DE SÃO PAULO, 2015).

Para concluir esta seção e este capítulo, é relevante destacar o trabalho desenvolvido pelo Grupo de Trabalho para Dados Conectados Governamentais do W3C (*W3C Government Linked Data Working Group*), que atuou entre junho de 2011 e dezembro de 2013 com a missão de *"estabelecer normas e outras informações voltadas a auxiliar governos de todo o mundo a publicar dados conectados governamentais de forma eficaz e utilizável*(W3C, 2011). Este grupo de trabalho é integrante da ação para Governo Eletrônico do W3C (*W3C eGovernment Activity*) e atua na catalogação e disponibilização de informações sobre as atividades governamentais envolvendo dados conectados em todo o mundo. Dentre as contribuições deste grupo para o tema DACG estão as "Melhores Práticas para

[24] Disponível em http://igovsp.net/spuk/

Publicação de Dados Conectados", importante trabalho relacionado para esta pesquisa.

2.4 Modelos de Referência

Com base no objetivo desta pesquisa que visa estabelecer um modelo de processo para publicação de dados abertos governamentais de natureza conectada, é importante destacar trabalhos relacionados enfatizem atividades de melhoria da qualidade de dados. Neste sentido, English (1999 apud HÜNER; OFNER; OTTO, 2009) estabelecem que uma atividade de qualidade de dados é qualquer processo realizado diretamente sobre os dados, a fim de melhorar a qualidade dos dados. Complementarmente sobre dados conectados governamentais, Villazón-Terrazas et al. (2011) estabelece que a o processo de publicação deste tipo de dado deve ter um ciclo de vida, da mesma forma que os ciclos da Engenharia de Software.

Segundo Rocha e Vasconcelos (2004) *"os modelos de maturidade fornecem aos gestores das organizações um poderoso instrumento para determinar em que estágio de maturidade se encontra, de modo a se planejar as ações necessárias para progredir em direção a uma maturidade superior e, por consequência, alcançar os objetivos desejados"*. Tais modelos se baseiam na premissa de que as organizações, pessoas, áreas funcionais, processos e atividades evoluem através de um processo de desenvolvimento ou crescimento em direção a uma maturidade mais avançada, atravessando um determinado número de estágios distintos.

Considerando este preâmbulo, analisaremos alguns modelos relacionados à maturidade em Governo Aberto, maturidade de dados, maturidade de software e processo de software.

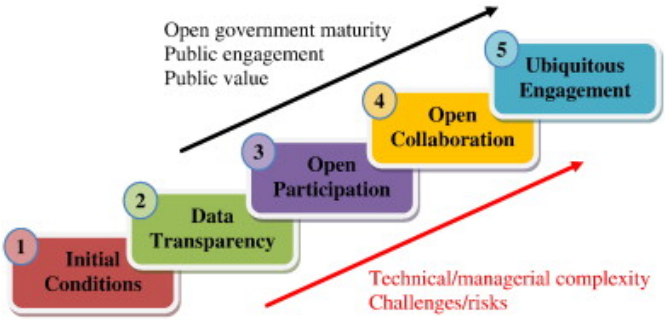
2.4.1 Modelo de Maturidade em Governo Aberto

Considerando o conceito de Governo Aberto explanado no referencial teórico, a sua implantação necessita ser gradual podendo requerer investimentos e compromissos não-triviais por parte dos órgãos governamentais(LEE; KWAK, 2012). Conforme descrito na Figura 4 e complementado pelo *"Memorandum of Transparency"*, a ampliação e o aprimoramento da oferta de dados públicos na *Web* consiste da etapa inicial para a implementação de políticas e programas de Governo Aberto.

A partir deste entendimento, Lee e Kwak (2012) desenvolveram e propuseram um Modelo de Maturidade em Governo Aberto (*Open Government Maturity Model (OGMM)*) que permite guiar as agências governamentais no desenvolvimento de suas ações e políticas voltadas à abertura governamental, como estabelecido na Figura 17. Neste modelo os autores propõem que a transparência dos dados é uma importante condição habilitadora para os demais níveis do Governo Aberto: participação e colaboração.

Os tópicos abaixo descrevem as principais atividades de cada nível deste modelo:

Figura 17 – Modelo de maturidade para o Governo Aberto



Fonte: Lee e Kwak (2012)

1. **Nível 1:** É o estágio inicial onde o órgão governamental foca suas ações na catalogação e disseminação das informações para o público;

2. **Nível 2:** Representa o primeiro passo para uma iniciativa de Governo Aberto, e consiste da promoção da transparência dos dados. Como o volume de dados está crescendo exponencialmente na era da economia digital, neste nível, há de se considerar a relevância e a utilidade dos dados a serem publicados (LEE; KWAK, 2012 apud MEIJER; THAENS; M., 2009), buscando priorizar a publicação do que é mais demandado ou o que representa maior impacto e benefício para a sociedade;

3. **Nível 3:** Busca ampliar a participação aberta do público nas ações e decisões governamentais, através de vários métodos e ferramentas. Enquanto o nível 2 abre os dados governamentais para o público, o nível 3 abre o governo para receber as ideias e o conhecimento oriundo da sociedade, visando melhorar a sua capacidade de respostas.

4. **Nível 4:** Cria mecanismos para que a sociedade atue na co-criação de soluções inovadoras para problemas do setor público ou ainda, na geração de novas oportunidades de negócio baseadas em dados públicos, bem como no incremento da pesquisa científica voltada à melhoria da ação governamental.

5. **Nível 5:** Atinge um ecossistema integrado de melhoramento e inovação do setor público mediante engajamento e articulação entre governo, academia e setor privado, baseado prioritariamente no ciclo virtuoso de transparência e abertura governamental.

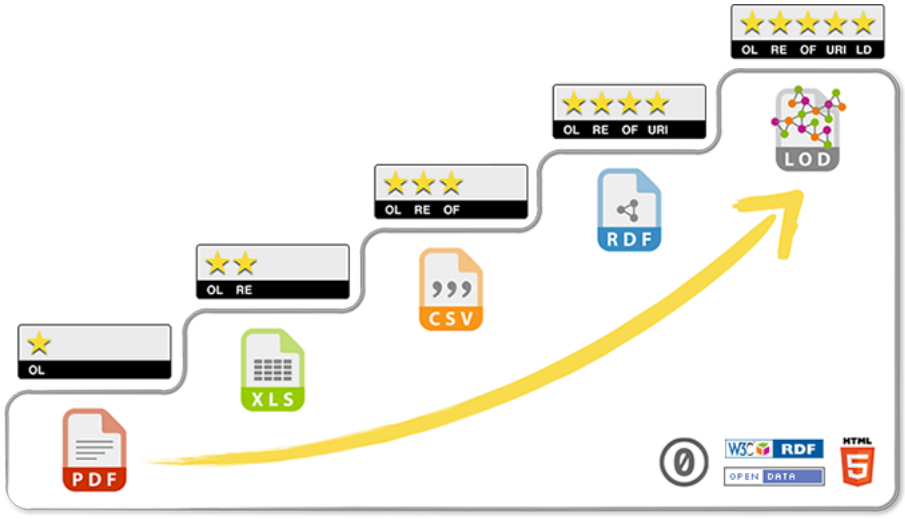
2.4.2 Esquema 5-Estrelas dos Dados Abertos

O *"Esquema 5-Estrelas dos Dados Abertos"*, proposto por Tim Berners-Lee no clássico artigo *"Linked Data"*(BERNERS-LEE, 2006) foi estabelecido como referencial na dimensão de modelo de maturidade de dados abertos. Este trabalho é de suma relevância pois

estabelece um modelo evolucionário para os dados abertos até o nível de dados abertos conectados, além de ter sido proposto pelo criador do conceito de dados conectados.

De acordo com Berners-Lee (2006), sob este esquema, o dado (ou informação) é considerado "uma estrela" se for tornado público de alguma forma, e em qualquer formato de dados (ex: um documento escaneado ou uma fotografia) utilizando uma licença aberta.

Figura 18 – Evolução dos Dados Abertos conforme o Esquema 5-Estrelas



Fonte: Berners-Lee (2006)

Por ser um esquema evolutivo, quão mais aprimorado for o dado e mais fácil de ser consumido pelos usuários, maior o número de estrelas que a ele será associado. Ademais, no esquema 5-Estrelas, para um dado ser considerado conectado ele precisa alcançar o maior nível (5 estrelas) conforme detalhamento abaixo (ISOTANI; BITTENCOURT, 2015):

- 1 Estrela: Disponível na Internet (em qualquer formato; por exemplo, PDF), desde que com licença aberta, para que seja considerado Dado Aberto;
- 2 Estrelas: Disponível na Internet de maneira estruturada (ex: em um arquivo Excel com extensão XLS);
- 3 Estrelas: Disponível na Internet, de maneira estruturada e em formato não proprietário (ex: CSV em vez de Excel);
- 4 Estrelas: Seguindo todas as regras anteriores, mas dentro dos padrões estabelecidos pelo W3C (RDF e SPARQL); usar URIs para identificar coisas e propriedades, de forma que as pessoas possam direcionar para suas publicações;
- 5 Estrelas: Todas as regras anteriores, mais: conectar seus dados a outros dados, de forma a fornecer um contexto.

O Esquema 5-Estrelas busca guiar as instituições publicadoras para que disponibilizem dados visando proporcionar maiores benefícios para os consumidores de dados, mesmo que isto possa proporcionar maior trabalho para os publicadores. A Tabela 5 apresenta um comparativo dos benefícios para consumidores e publicadores de dados para cada um dos cinco níveis do Esquema 5-Estrelas.

Tabela 5 – Benefícios da publicação e consumo de dados no esquema 5-Estrelas dos Dados Abertos

Estrelas:	**Para consumidores:**	**Para publicadores:**
1	Ver os dados Imprimi-los Guardá-los (no disco rígido ou em um pen-drive, por exemplo) Modificar os dados como queira Acessar o dado de qualquer sistema Compartilhar o dado com qualquer pessoa	É simples de publicar Não precisa explicar repetitivamente que as pessoas podem fazer uso dos dados
2	Os mesmos benefícios de quem usa 1 estrela Usar softwares proprietários para processar, agregar, calcular e visualizar os dados Exportá-los em qualquer formato estruturado	É fácil de publicar
3	Os mesmos benefícios de quem usa 2 estrelas Manipular os dados da forma que lhe agrada, sem estar refém de algum software em particular	É ainda fácil de publicar Obs.: Podem ser necessários conversores ou plug-ins para exportar os dados do formato proprietário
4	Os mesmos benefícios de quem usa 3 estrelas Fazer marcações Reutilizar parte dos dados Reutilizar ferramentas e bibliotecas de dados existentes, mesmo que elas entendam apenas parte dos padrões usados por quem publicou Combinar os dados com outros	Há controle dos itens dos dados e pode melhorar seu acesso Outros publicadores podem conectar seus dados, promovendo-os as 5 estrelas
5	Descobrir mais dados vinculados enquanto consome dados Aprender sobre a classificação das 5 estrelas	Torna o dado mais fácil de ser descoberto Aumenta o valor do dado A organização ganha os mesmos benefícios com a vinculação de dados que os consumidores

Fonte: Isotani e Bittencourt (2015 apud W3C Brasil, 2013)

Pela sua característica de guiar os publicadores de dados abertos a produzirem dados abertos conectados, o Esquema 5-Estrelas foi estabelecido como o modelo de maturidade

a ser trabalhado pela proposta de modelo de processo desta pesquisa, cujo detalhamento será apresentada no próximo capítulo.

2.4.3 Modelos de Processo de Software

Corroborando com a afirmação de Villazón-Terrazas et al. (2011) que estabelece que os processos de publicação de dados conectados governamentais devem ter um ciclo de vida como os existentes na Engenharia de Software, e que devem ser de natureza incremental e iterativa, apresentaremos dois ciclos de vida (ou modelos de processo de software) com estas características e que servirão de referência à proposta apresentada nesta pesquisa.

Um modelo de processo de software é uma representação abstrata de um processo de software. Cada modelo de processo representa um processo a partir de uma perspectiva particular, de uma maneira que proporciona apenas informações parciais sobre o processo (SOMMERVILLE, 2007). Desta forma, cada modelo de processo de software define a sequência com que as atividades serão executadas, quais as pessoas estão envolvidas e quais os artefatos são gerados por cada atividade (CARRION; WERNER, 2013).

Cumpre destacar que na literatura de Engenharia de Software os modelos de processo de software também são denominados (i) ciclos de vida de software; (ii) paradigmas de processo de software ou (iii) modelos genéricos de software (SOMMERVILLE, 2007; PRESSMAN, 1995).

Contemplando os requisitos (incremental e iterativo) estabelecido por Villazón-Terrazas et al. (2011), apresentaremos dois modelos de processo projetados explicitamente para apoiar a iteração de processo apresentados em Sommerville (2007):

1. Entrega Incremental: A especificação, o projeto e a implementação de software são divididos em uma série de incrementos desenvolvidos um de cada vez;

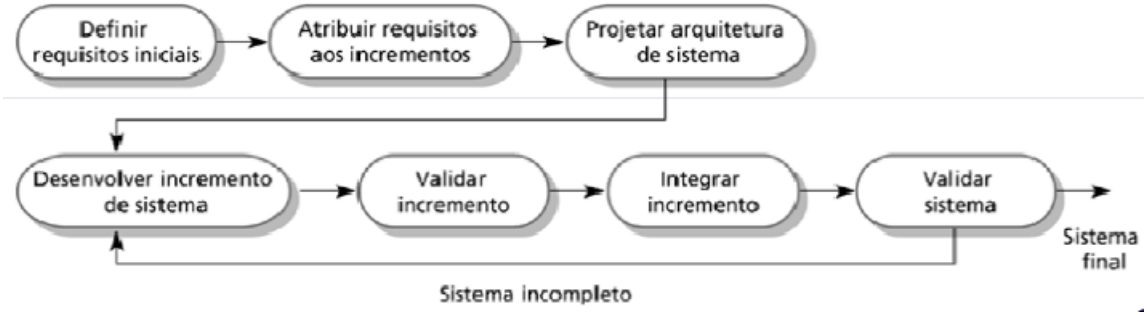
2. Desenvolvimento espiral: O desenvolvimento do sistema evolui em espiral a partir de um esboço inicial até o sistema final.

2.4.3.1 Entrega Incremental

A entrega incremental é uma abordagem intermediária entre outros dois paradigmas de desenvolvimento de software (cascata e evolucionário) que busca combinar as vantagens destes dois paradigmas. Neste modelo de processo, o cliente identifica, em linhas gerais, os serviços a serem fornecidos pelo sistema, conforme exemplificado na Figura 19.

O desenvolvimento mediante este modelo de processo permite que o software seja desenvolvido em incrementos. Os incrementos prioritários são desenvolvidos inicialmente e de forma paralela, os requisitos dos próximos incrementos são estabelecidos. Ao concluir o desenvolvimento de cada incremento, esta parte do software entra em operação, ou seja, fica disponível ao cliente. À medida que novos incrementos são concluídos, estes são

Figura 19 – Modelo de processo de desenvolvimento incremental



Fonte: Sommerville (2007)

integrados aos já existentes, de tal forma que o software é aprimorado a cada incremento entregue (SOMMERVILLE, 2007).

2.4.3.2 Desenvolvimento em Espiral

O modelo de processo em espiral foi apresentado por Boehm (1986). Em vez de representar o processo de software como uma sequência de atividades com alguma saída entre uma atividade e outra, o processo é representado como uma espiral. Cada *loop* (ciclo) na espiral representa uma fase do processo de software (SOMMERVILLE, 2007). Foi desenvolvido para abranger as melhores características tanto do ciclo de vida clássico como da prototipação, acrescentando, ao mesmo tempo, um novo elemento, a análise de riscos que falta a esses paradigmas. De acordo com Sommerville (2007), o modelo define quatro importantes setores representados em quadrantes, conforme tópicos abaixo:

1. Definição de objetivos: Definir os objetivos específicos de cada fase do projeto, com o estabelecimento de restrições sobre o produto e identificação dos riscos. Permite a elaboração de um plano de gerenciamento do projeto;

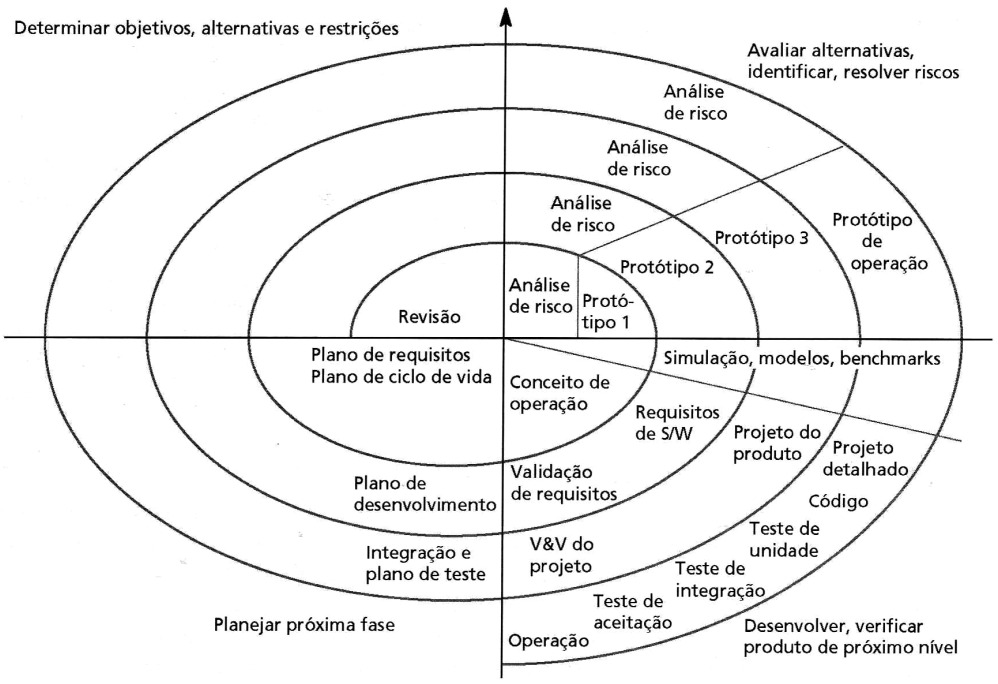
2. Avaliação e redução de riscos: É realizada uma análise detalhada dos riscos do projeto propondo soluções para cada um deles;

3. Desenvolvimento de validação: Desenvolvimento do sistema após a avaliação dos riscos. Dependendo dos tipos de risco, podem ser adotados processos de desenvolvimento distintos. Por exemplo: se os riscos de interface forem dominantes, sugere-se o uso da prototipação;

4. Planejamento: O projeto é revisado gerando decisões para o prosseguimento do próximo loop da espiral. Se o projeto prosseguir, o planejamento é atualizado.

A Figura 20 esclarece o desenvolvimento de software mediante o modelo de processo em espiral conforme (BOEHM, 1986):

Figura 20 – Modelo de processo de desenvolvimento em espiral proposto por Boehm



Fonte: Boehm (1986)

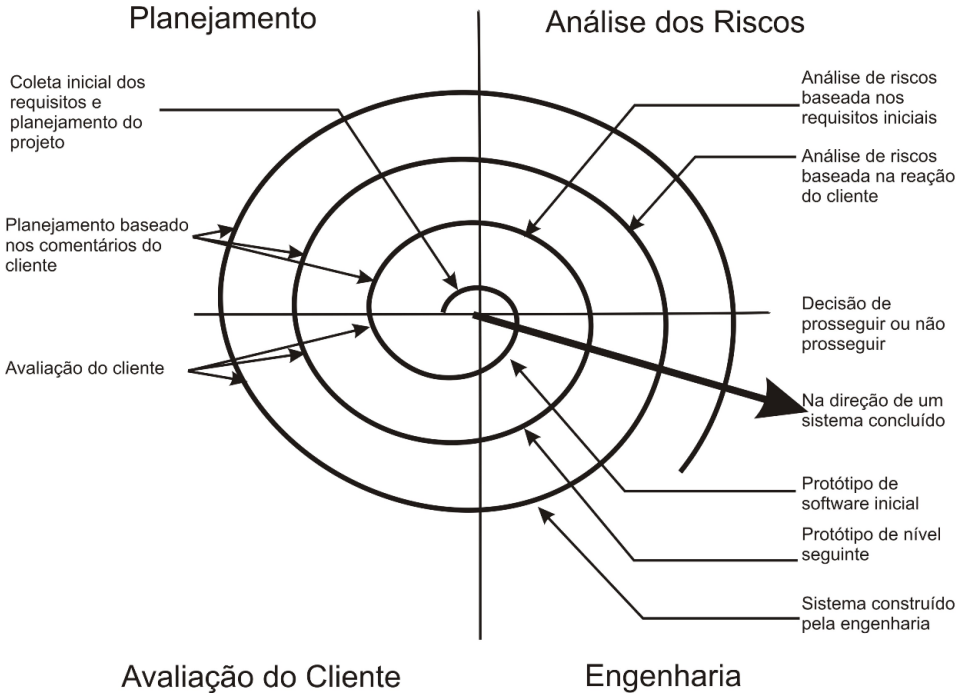
Complementarmente, Pressman (1995) apresenta uma versão simplificada da espiral, composta dos seguintes setores:

1. Planejamento: determinação dos objetivos, alternativas e restrições;
2. Análise de riscos: análise de alternativas e identificação/resolução de riscos;
3. Engenharia: desenvolvimento do produto no "nível seguinte";
4. Atualização feita pelo cliente: avaliação dos resultados da engenharia.

A Figura 21 esclarece o desenvolvimento de software mediante o modelo de processo em espiral conforme Pressman (1995):

Pressman (1995) entende que este modelo de processo também pode ser considerado como uma abordagem "evolucionária" à engenharia de software, capacitando o desenvolvedor e o cliente a entender e reagir aos riscos em cada fase evolutiva. O modelo espiral usa um protótipo (do modelo de processo prototipação) como um mecanismo de redução de riscos, mas, o que é mais importante, possibilita que o desenvolvedor aplique características da abordagem de prototipação em qualquer etapa da evolução do produto. Outro benefício deste modelo é que ele mantém a abordagem de passos sistemáticos sugerida pelo ciclo de vida clássico, incorporando-a numa estrutura iterativa que reflete mais realisticamente o mundo real. O modelo espiral exige uma consideração direta dos riscos

Figura 21 – Modelo de processo de desenvolvimento em espiral proposto por Pressman



Fonte: Pressman (1995) adaptado de Boehm (1986)

técnicos em todas as etapas do projeto e, se adequadamente aplicado, deve reduzir os riscos antes que eles se tornem problemáticos.

2.4.3.3 SCRUM

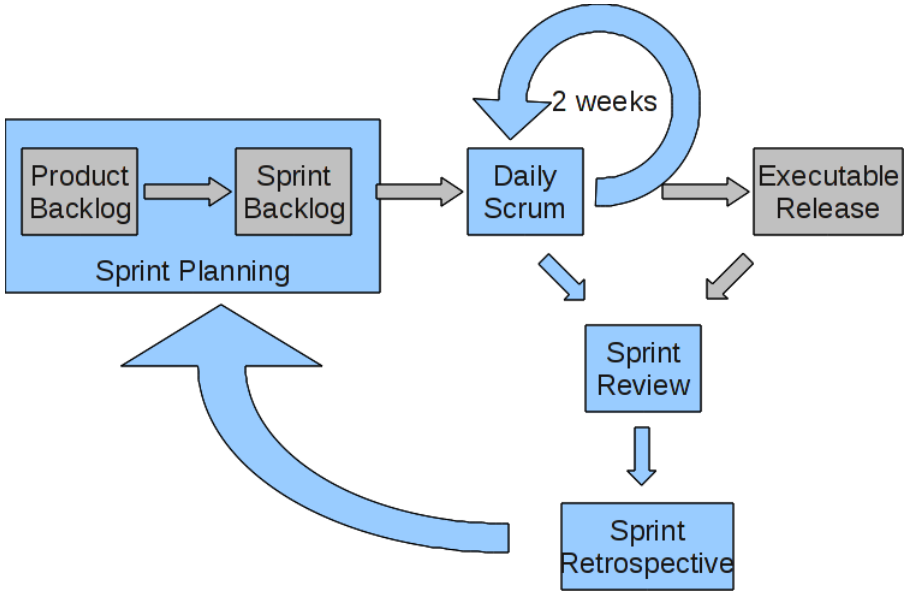
É um modelo de desenvolvimento ágil de software estabelecido por Jeff Sutherland nos anos 90. Tem como principais estratégias a definição de times auto-organizados, progresso do desenvolvimento através de ciclos curtos para atingir metas específicas, conhecidos como *Sprints*, e requisitos de produtos organizados numa lista de itens, conhecido como *"Product Backlog"* (VARASCHIM, 2009).

O *Scrum* tem o progresso de desenvolvimento baseado em iterações de curto prazo. O primeira etapa dentro do *Sprint* consiste de uma reunião de planejamento (*Sprint Planning*), onde o time (*Scrum Team*), em conjunto com o cliente (*Product Owner*) define o que será implementado na iteração, sendo responsabilidade do cliente realizar a priorização do trabalho a ser feito. Posteriormente, é desenvolvida a etapa de execução, com detalhamento das tarefas necessárias para implementar o que foi solicitado. Diariamente são realizadas reuniões para averiguar o andamento do projeto.

Ao final do *Sprint* é realizada uma reunião para a validação da entrega (*Sprint Review*), onde o cliente e quem mais tiver interesse no produto pode verificar se o objetivo do Sprint foi atingido. Logo após, é realizada apenas pelo time uma reunião (*Sprint Retrospective*) onde o *Sprint* é avaliado sob a perspectiva de processo, time ou produto, quais foram os acertos e os erros com o objetivo de melhorar o processo de trabalho.

A Figura 22 apresenta uma visão geral do ciclo de desenvolvimento utilizando o *Scrum*:

Figura 22 – Modelo de processo de desenvolvimento *Scrum*



Fonte: Disponível em <http://paulgestwicki.blogspot.com.br/2011/02/scrum-diagram-rfc.html>. Acesso em: 27 out. 2015

Além dos ciclos iterativos, um dos grandes benefícios do SCRUM baseia-se nas revisões periódicas das atividades, além do estabelecimento de metas de curto prazo. Scrum é recomendado especialmente para projetos de software cujos requisitos não foram ou não podem ser estabelecidos previamente.

2.4.4 Gerenciamento de Projetos

Considerando que uma atividade de publicação de dados é uma atividade temporária, única e resulta num produto final de valor agregado, é relevante para esta pesquisa fundamentar sobre as definições de projeto e gerenciamento de projetos.

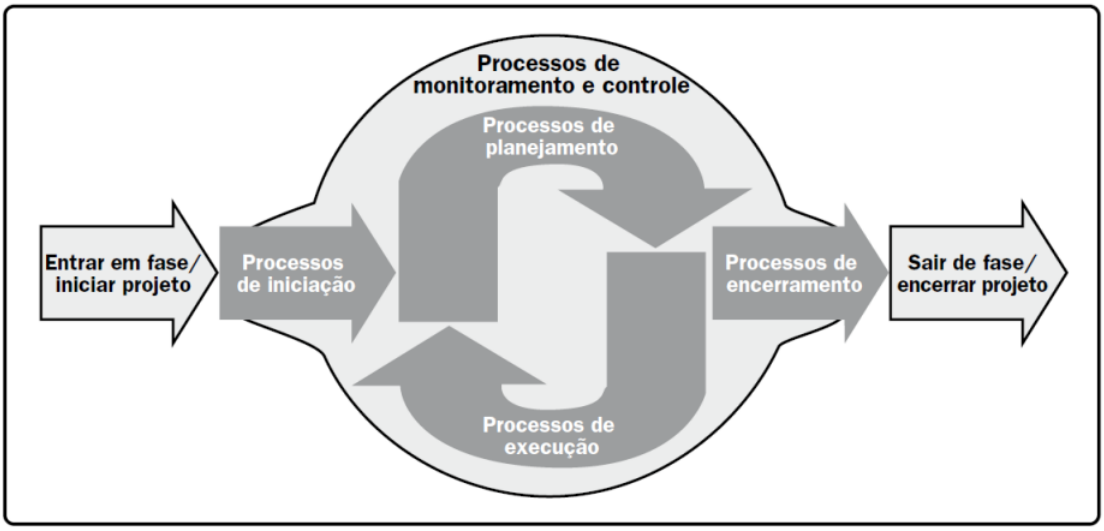
Segundo o PMI (2013), um Projeto é um esforço temporário empreendido para criar um produto, serviço ou resultado exclusivo. A natureza temporária dos projetos indica que eles têm um início e um término definidos. Cada projeto cria um produto, serviço ou resultado único. Embora elementos repetitivos possam estar presentes em algumas entregas e atividades do projeto, esta repetição não muda as características fundamentais e exclusivas do trabalho do projeto.

O Gerenciamento de Projetos pode ser definido como a aplicação do conhecimento, habilidades, ferramentas e técnicas às atividades do projeto para atender aos seus requisitos (PMI, 2013). O gerenciamento de projetos é realizado através da aplicação e integração apropriadas dos 47 processos de gerenciamento de projetos, logicamente agrupados em cinco grupos de processos. Esses cinco grupos de processos são:

- **Processos de iniciação:** Processos executados para definir um novo projeto ou uma nova fase de um projeto existente através da obtenção de autorização para iniciar o projeto ou fase.
- **Processos de planejamento:** Processos necessários para definir o escopo do projeto, refinar os objetivos e definir a linha de ação necessária para alcançar os objetivos para os quais o projeto foi criado.
- **Processos de execução:** Processos realizados para executar o trabalho definido no plano de gerenciamento do projeto para satisfazer as especificações do projeto.
- **Processos de monitoramento e controle:** Os processos exigidos para acompanhar, analisar e controlar o progresso e desempenho do projeto, identificar quaisquer áreas nas quais serão necessárias mudanças no plano, e iniciar as mudanças correspondentes.
- **Processos de encerramento:** Os processos executados para finalizar todas as atividades de todos os grupos de processos, visando encerrar formalmente o projeto ou fase.

A Figura 23 apresenta uma visão integrada dos grupos de processos de gerenciamento de projetos, segundo o *PMBoK© Guide*.

Figura 23 – Grupos de processos de gerenciamento de projetos



Fonte: PMI (2013)

Esta visão integrada do gerenciamento de projetos proporciona maiores condições para que o projeto a ser desenvolvido seja melhor gerenciado com um produto de qualidade ao seu término. A organização nestes grupos de processos proporciona ao gerente de projetos a identificação mais precisa do conjunto de atividades integradas necessárias a ser desenvolvido para o sucesso do projeto.

2.5 Metodologia GQM

A metodologia GQM é assim denominada por representar uma abordagem sistemática, orientada para metas. As metas (ou objetivos) são denominados *"Goals"*. A partir de cada *"Goal"*, são definidas questões relevantes de investigação, denominadas *"Questions"*, que possuem unidades mensuráveis que são as métricas (*"Metrics"*). Em resumo, através de cada meta definida, são refinadas questões, sendo que as métricas fornecem a informação para responder a essas questões.

Originalmente, a abordagem GQM foi definida para avaliar defeitos em um conjunto de projetos no Centro Espacial da NASA, nos Estados Unidos. Apesar de ter sido usada para definir e avaliar metas para um projeto particular, a abordagem teve seu uso expandido para um contexto maior, como em programas de avaliação de qualidade de software (BASILI, 1993).

A GQM (BASILI; CALDIERA; ROMBACH, 1994) é orientada às metas de avaliação de produtos e processos de software, que segue a definição top-down de um programa de avaliação. A análise e a interpretação dos dados, por sua vez, seguem uma visão bottom-up, isto é, interpretam-se os dados separadamente. Os resultados são formulados ao finalizar a avaliação, após o relacionamento e comparação dos dados.

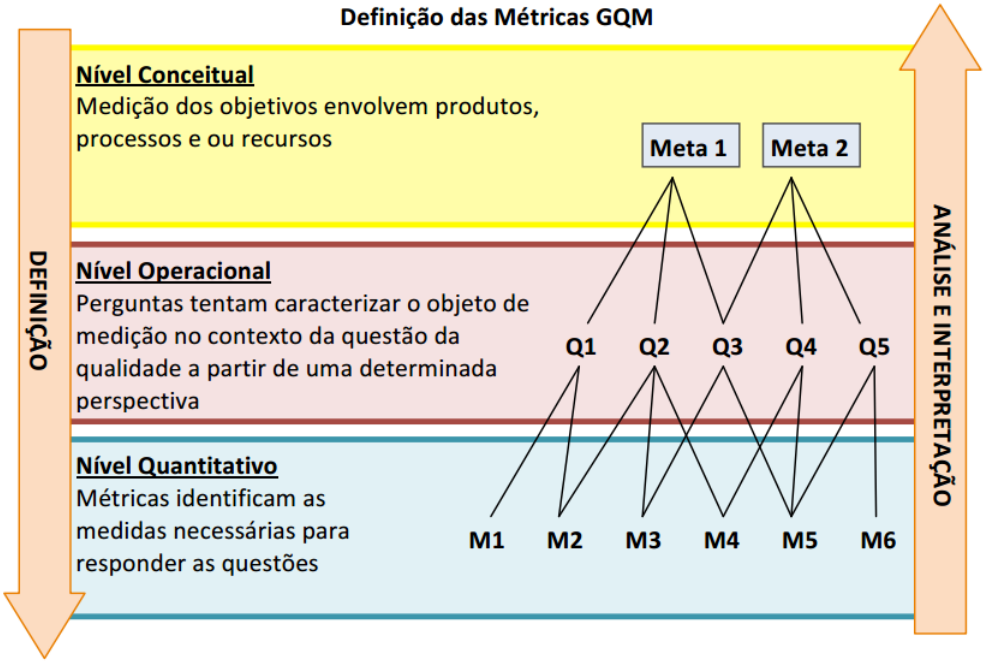
Esta abordagem abrange informações necessárias para a realização das tarefas de análise segundo o paradigma da avaliação orientada por metas, tendo como componentes elementares os objetivos, as questões e as métricas, conforme destacado a seguir (ABIB; KIRNER, 1998):

- Objetivo: Sua definição envolve o objeto, o propósito, o foco de qualidade, o ponto de vista e o ambiente.
- Questão: A questão anuncia a necessidade de se obter informações em uma linguagem natural, podendo-se formular uma ou mais questões para cada categoria de questões; quanto à resposta, deve estar de acordo com o objetivo.
- Métrica: Sua função é especificar os dados ou as informações que se deseja obter durante as avaliações, em termos quantitativos e avaliáveis, podendo-se, utilizar uma ou mais métricas para cada questão.

É muito importante para o sucesso da aplicação do GQM que os objetivos estejam bem traçados, pois somente assim a escolha das métricas e posterior avaliação dos dados será bem sucedida. Sendo assim o GQM considera um modelo com três níveis de realização:

- **Conceitual:** Definição do escopo da avaliação, ou seja, do objeto a ser medido. (Processos, Produtos ou Recursos)
- **Operacional:** Definição de um conjunto de questões que auxilie na caracterização do objeto de estudo e como ele deve ser enxergada dentro do contexto da qualidade.

Figura 24 – Desenvolvimento do modelo GQM



Fonte: Silva et al. (2009)

- **Quantitativo:** Definição de um conjunto de dados a serem obtidos, relacionado a cada uma das questões definidas anteriormente, a fim de respondê-las de forma quantitativa, ou seja, as métricas.

O próximo capítulo abordará os trabalhos relacionados mais relevantes para esta investigação, contribuindo também para a consolidação da base teórica utilizada para a formulação da proposta apresentada nesta dissertação.

3 PRINCIPAIS TRABALHOS RELACIONADOS

Este capítulo abordará os principais trabalhos relacionados com esta dissertação, os quais foram utilizados como base para a construção da proposta de modelo desenvolvida nesta pesquisa de mestrado. Foram destacados dois trabalhos de natureza acadêmica voltados à publicação de dados abertos conectados governamentais e outros trabalhos de origem governamental voltados a publicação de dados abertos. Além disto, a revisão de literatura, explanada no capítulo 4 apresentará outros trabalhos relacionados com a temática desta pesquisa.

Inicialmente, buscaremos delimitar que o escopo considerou prioritariamente atividades relacionadas à publicação de dados. Num espectro mais amplo, um processo de abertura de dados envolve várias fases que vão desde a seleção e publicação dos dados até o uso dos dados e *feedback* sobre os dados utilizados. Este conjunto de fases que compõem o processo de publicação e consumo dos dados abertos é chamado de Ciclo de Vida dos Dados Abertos. Em geral, tanto os responsáveis pela publicação quanto os consumidores de dados ou não têm conhecimento sobre o ciclo de vida dos Dados Abertos ou não existe um consenso entre eles sobre tais fases e seu significado (LÓSCIO, 2014). Veremos abaixo alguns trabalhos relacionados à publicação de dados abertos e dados abertos conectados na esfera governamental.

3.1 Melhores Práticas para Dados Conectados - W3C

A pesquisa estabeleceu este trabalho como referencial comparativo na dimensão de publicação de dados, ou seja, é um dos dois principais trabalhos relacionados. As *"Melhores Práticas para Dados Conectados"* possuem grande relevância e relação com esta pesquisa por terem sido desenvolvidas por um grupo de especialistas convidados pelo W3C no âmbito do grupo de trabalho *"W3C Government Linked Data Working Group"*.

Este trabalho foi elaborado para compilar as práticas mais relevantes para a publicação e utilização de dados de alta qualidade publicados por governos ao redor do mundo, como os Dados Abertos Conectados, com o propósito de orientar a publicação, uso e reutilização de dados abertos governamentais. Tais práticas tem como público-alvo administradores e desenvolvedores de sítios *Web* bem como gestores e técnicos que atuam com gestão de informações governamentais(W3C, 2014).

É composto de 10 melhores práticas que consideram o nível de envolvimento de partes interessadas na publicação de dados, cuidados com a legalidade, modelagem, processamento e organização dos dados, semântica, tempestividade, alta disponibilidade dos dados e ainda, como deve ser feita a sua divulgação para o público e sua manutenção. São descritas resumidamente a seguir (W3C, 2014):

1. **Preparar partes interessadas (*stakeholders*):** Etapa que visa preparar as partes interessadas, em especial os decisores, sobre o processo de criação de manutenção de dados abertos conectados;

2. **Selecionar conjuntos de dados:** Etapa que visa à seleção dos dados que serão publicados bem como irão prover benefícios para os usuários, que poderão utilizá-lo e reusá-lo para diversas finalidades;

3. **Modelar os dados:** Esta etapa visa estabelecer uma melhor representação dos objetos de dados e como eles serão utilizados por aplicações de forma independente a sua origem;

4. **Especificar uma licença apropriada:** Visa o estabelecimento das condições de uso (e não uso) dos dados que serão publicados. O reúso dos dados ocorre com maior frequência quando existe clareza sobre a sua origem, propriedade (autoria) e demais condições relacionadas ao uso destes dados;

5. **Estabelecer bons identificadores universais (URIs) para dados conectados:** O núcleo dos dados conectados se baseia num planejamento bem feito de identificação e referenciamento dos dados na *Web*, baseados em URIs HTTP. Devem ser estabelecidos os requisitos para objetos de dados, suporte a diversos idiomas, alteração de dados ao longo do tempo e estratégia de persistência

6. **Utilizar vocabulários padrão:** Descrever objetos com vocabulários previamente definidos, sempre que possível. Quando necessário, ampliar tais vocabulários de acordo com a necessidade. Podem ser criados novos vocabulários (somente quando necessário) seguindo as melhores práticas para este fim.

7. **Converter e enriquecer dados:** Esta etapa visa estabelecer as condições para a conversão e o enriquecimento de dados para dados conectados. Estão contempladas atividades técnicas de melhoria dos dados, como conversão pra formatos mais apropriados, desenvolver serializações, dentre outras. Para esta pesquisa, serão consideradas tarefas técnicas que preparem os dados para que sejam conectados ou que sejam melhorados para serem conectados futuramente. Pode ser entendida como uma etapa de enriquecimento de dados. É uma atividade comumente realizada por scripts ou processos automatizados;

8. **Prover acesso automatizado aos dados:** Devem ser providos diversos meios e formatos para que ferramentas de busca e outros recursos de processamento e consumo automatizado possam utilizar os dados mediante recursos padrões da *Web*;

9. **Anunciar os conjuntos de dados:** Devem ser divulgados os novos dados publicados para que a comunidade possa fazer o devido uso;

10. **Estabelecer um contrato social para os dados publicados:** Esta etapa visa destacar a necessidade de se reconhecer a responsabilidade em garantir a disponibilidade, manutenção e atualização dos dados publicados. Por serem conectados, tais dados precisam ter a garantia que vão ficar disponíveis onde a organização publicadora estabelece e ainda, será mantido ao longo do tempo.

O trabalho tem como pontos fortes: ser a compilação de práticas para publicação de DACG desenvolvidas por um grupo de trabalho bastante capacitado que trouxe conceitos e técnicas de diversas experiências de publicação de DAC e DACG em todo o mundo.

Poderiamos entender como limitação não ser apresentado um ciclo de vida, mas apenas um conjunto de melhores práticas que podem ser utilizadas adequando-as ao contexto de aplicação. Entretanto, isto se deve pela natureza do trabalho. Outra limitação está voltada a inexistência de diretrizes que considerem o nível de maturidade dos dados que serão publicados, ou seja, é considerado que toda a publicação deva ser concluída com dados abertos conectados, não apresentando estágios intermediários que poderiam ser utilizados especialmente por instituições com menor experiência na publicação de dados abertos e dados abertos conectados

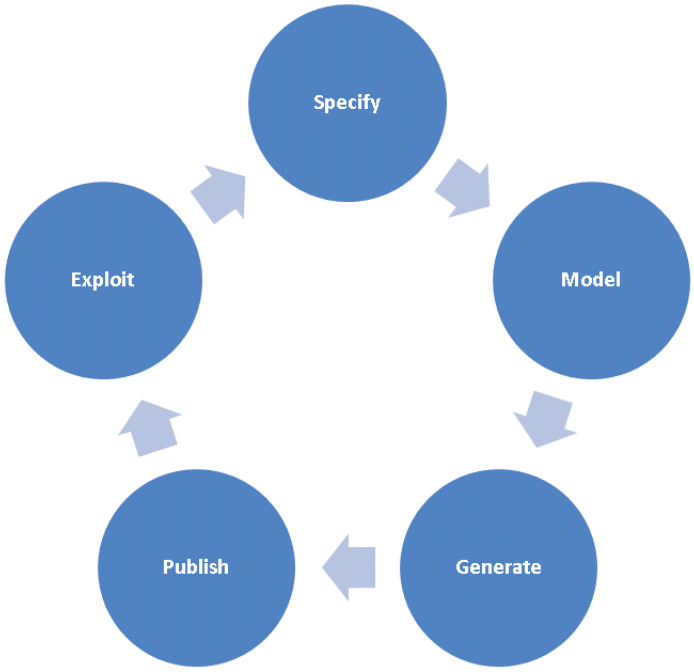
Ademais, desta parte em diante deste trabalho, faremos referência a este trabalho utilizando a sigla BPLD, que é o acrônimo da definição oficial (em inglês) destas melhores práticas (*Best Practices for Publishing Linked Data*).

3.2 Methodological Guidelines for Publishing Government Linked Data

Em *"Methodological Guidelines for Publishing Government Linked Data"* Villazón-Terrazas et al. (2011) apresentam uma proposta inicial para formalizar sua experiência no desenvolvimento de dados conectados governamentais. Seu ciclo de vida é composto pelas seguintes atividades: (1)Especificar, (2)Modelar, (3)Gerar, (4)Publicar, e (5) Explorar. Cada atividade é decomposta em uma ou mais tarefas, e algumas técnicas e ferramentas são apresentadas para sua execução. O autor ressalta que a ordem das atividades e tarefas podem ser alteradas conforme as necessidades específicas dos órgãos governamentais (VILLAZÓN-TERRAZAS et al., 2011). A Figura 25 as principais atividades deste processo.

O ponto forte do trabalho referente ao ciclo de vida proposto por Villazón-Terrazas et al. (2011) é a estruturação de um conjunto resumido de etapas que podem ser desenvolvidas ciclicamente possibilitando a melhoria da qualidade dos dados conectados governamentais a serem produzidos. Outro ponto muito relevante consiste na fundamentação referente a indisponibilidade de oferta e cultura de publicação de DACG. Destacamos ainda o detalhamento quanto a apresentação de diversas técnicas e ferramentas que foram utilizadas nos estudos de caso referentes a validação do artigo científico.

Figura 25 – Ciclo de Vida de dados estabelecido em *"Methodological Guidelines for Publishing Government Linked Data"*



Fonte: Villazón-Terrazas et al. (2011)

No entanto, a limitação deste trabalho para o contexto atual da pesquisa em DACG está relacionado a ter sido desenvolvido anteriormente à publicação das BPLDs. Por consequência, não incorpora conceitos relevantes propostos nas BPLDs como a "Preparação das Partes Interessadas" e o estabelecimento do "Contrato Social entre publicadores e consumidores de dados".

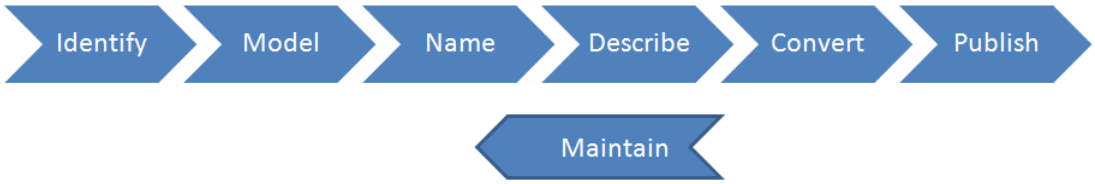
Além disso, aparentemente este ciclo de vida considera que toda a publicação deva ser concluída com dados abertos conectados, não apresentando estágios intermediários que poderiam ser utilizados especialmente por instituições com menor experiência na publicação de dados abertos e dados abertos conectados. Ademais, a pesquisa foi validada num contexto específico na Espanha, onde não foram apresentados novos elementos que subsidiem a aplicação deste ciclo de vida num contexto diferente ao da validação.

3.3 The Joy of Data - A Cookbook for Publishing Linked Government Data on the Web

Em *"The Joy of Data - A Cookbook for Publishing Linked Government Data on the Web"* Hyland e Wood (2011) é proposto um "livro de receitas" de seis passos para modelar, criar, publicar, e anunciar os dados conectados governamentais, que consiste nas atividades de (1)Identificar,(2)Modelar (3)Denominar com URIs, (4)Descrever, (5)Converter, (6)Publicar, e (7)Manter os dados, conforme apresentado na Figura 26:

O ponto forte do trabalho referente ao ciclo de vida proposto por Hyland e Wood (2011)

Figura 26 – Ciclo de Vida de dados estabelecido em *"The Joy of Data - Cookbook for Publishing Linked Government Data on the Web"*



Fonte: Hyland e Wood (2011)

está relacionado a contemplar as etapas mais relevantes para a publicação de DACG e ter sido o trabalho principal que serviu de base para a elaboração das BPLDs. Destacamos ainda a tentativa de não deixar complexa a descrição das etapas e atividades para publicação de DACG, ao utilizar uma analogia com uma "receita para cozinhar algum alimento". Este ciclo de vida ainda apresenta opções de vocabulários e ontologias relevantes que podem ser utilizadas em diversos cenários de publicação de DACG, bem como apresenta recomendações relevantes para a maioria das BPLDs, por ter sido o principal trabalho de base.

No entanto, como limitação deste trabalho para o contexto atual da pesquisa em DACG, por ter sido desenvolvido anteriormente à publicação das BPLDs, não incorpora alguns conceitos relevantes propostos nas BPLDs como a "Preparação das Partes Interessadas" e "Seleção dos Conjuntos de Dados". Ademais, assim como Villazón-Terrazas et al. (2011) aparentemente este ciclo de vida considera que toda a publicação deva ser concluída com dados abertos conectados, não apresentado estágios intermediários que poderiam ser utilizados especialmente por instituições com menor experiência na publicação de dados abertos e dados abertos conectados.

3.4 Manual para Elaboração de Planos de Dados Abertos do Governo Brasileiro

Trata-se de um relevante documento técnico com o objetivo de orientar instituições brasileiras à desenvolver seus planos de dados abertos. Estabelece que os planos de dados abertos devem contemplar as seguintes seções e subseções:

- Objetivos (gerais e específicos);
- Legislação e demais normativos aplicáveis;
- Cenário institucional do órgão (demonstrar o alinhamento com compromissos assumidos internamente, perante a sociedade ou outros entes);
- Metodologia de construção e validação do documento;
- Definição dos dados a serem abertos;

- Critérios utilizados para priorização dos dados;
- Estratégia definida para abertura dos dados, com respectivo plano de ação;
- Modelo de Sustentação (de modo a perenizar o fluxo de atualização e manutenção dos dados);
- Estrutura de Governança, forma de monitoramento e controle;
- Canais de Comunicação e Participação Social;
- Metas de melhoria contínua;

Este trabalho tem como ponto forte uma grande preocupação com o estabelecimento de uma base institucional para as atividades de publicação de dados abertos. Complementarmente, apresenta recomendações relevantes para contemplar as BPLDs não contempladas no ciclo de vida de (HYLAND; WOOD, 2011) que são a "Preparação das Partes Interessadas" e a "Seleção dos Conjuntos de Dados". Destacamos a orientação para que o plano de dados abertos seja desenvolvido de forma alinhada à base legal, diretrizes institucionais e plano diretor de tecnologia da informação da instituição publicadora.

No entanto, como limitações deste trabalho, até pelo seu propósito, não apresenta um maior detalhamento de como deve ser desenvolvida a publicação de dados abertos. Consequentemente, não contempla nenhuma recomendação referente as BPLDs de natureza mais técnica. Além disto, por ser voltado a publicação apenas de dados abertos, não apresenta atividades que contemplem a publicação de dados abertos conectados governamentais. Por outro lado, algumas destas limitações são contempladas noutros documentos do Governo Brasileiro, como a "Cartilha para publicação de Dados Abertos" (BRASIL, 2014a) e o "Kit de Dados Abertos" (BRASIL, 2011a).

3.5 **Guía para la Apertura de Datos en Colombia**

Este documento técnico do governo colombiano visa orientar que instituições governamentais da Colômbia estabeleçam estruturas e procedimentos adequados para a publicação de dados abertos. O documento apresenta riqueza de detalhes no que tange ao alinhamento estratégico da iniciativa de publicação de dados abertos com as diretrizes institucionais, bem como quanto a seleção de conjuntos de dados que serão publicados e ainda, o estabelecimento de um padrão dos ambientes de oferta de dados a serem desenvolvidos pelos órgãos públicos. O guia apresenta a seguinte sequência de etapas para a publicação de dados abertos, conforme apresentado na Figura 27.

- Preparação para a abertura de dados
- Identificação das informações

- Análise das informações
- Priorização dos dados
- Documentação dos dados
- Estruturação, Carga e Publicação dos conjuntos de dados

Figura 27 – Ciclo de Vida de dados estabelecido em *Guía para la Apertura de Datos en Colombia*

Fonte: COLOMBIA (2012)

Assim como o documento brasileiro, este trabalho tem como ponto forte uma grande preocupação com o estabelecimento de diretrizes institucionais para as atividades de publicação de dados abertos. Também apresenta recomendações relevantes para contemplar as BPLDs não contempladas no ciclo de vida de Hyland e Wood (2011) que são a "Preparação das Partes Interessadas" e a "Seleção dos Conjuntos de Dados". Contribue ainda com orientações relevantes sobre a documentação das atividades de publicação bem como no estabelecimento de metadados detalhados para os conjuntos de dados a serem publicados.

No entanto, como limitações deste trabalho, até pelo seu propósito, não apresenta um maior detalhamento técnico (envolvendo ferramentas, por exemplo) de como deve ser desenvolvida a publicação de dados abertos. Consequentemente, não contempla nenhuma recomendação referente as BPLDs de natureza mais técnica. Além disto, por ser voltado a publicação apenas de dados abertos, não apresenta detalhamento que contemple a publicação de dados abertos conectados governamentais.

Além destes trabalhos relacionados, todos os demais processos de publicação de dados abertos investigados nesta pesquisa também possuem forte relação com esta pesquisa e terão diversas de suas características devidamente evidenciadas no próximo capítulo que apresenta uma revisão de literatura que analisou quinze processos de publicação de dados abertos extraíndo conhecimento relevante para o modelo de processo proposto nesta pesquisa.

4 RECOMENDAÇÕES PARA PUBLICAÇÃO DE DADOS ABERTOS E DADOS ABERTOS CONECTADOS

Este capítulo apresenta uma revisão de literatura que teve como objetivo responder questões relevantes à esta pesquisa de mestrado. Considerando o objetivo principal deste trabalho que visa guiar as instituições governamentais para publicarem Dados Abertos Conectados, foi identificado e estabelecido como um dos principais trabalhos relacionados para publicação de dados as *"Melhores Práticas para Publicação de Dados Conectados"* do W3C (*"W3C Best Practices for Publishing Linked Data"*).

Neste contexto é importante ressaltar que a revisão da literatura é uma parte vital do processo de investigação, envolvendo atividades de localização, análise, síntese e interpretação de investigações anteriores ao trabalho a ser desenvolvido. A revisão da literatura é indispensável não somente para definir bem o problema, mas também para obter uma ideia precisa sobre o estado atual dos conhecimentos sobre um dado tema, as suas lacunas e a contribuição da investigação para o desenvolvimento do conhecimento (BENTO, 2012).

Para o desenvolvimento da revisão de literatura desta pesquisa, foram estabelecidas as seguintes atividades:

1. Identificação e estabelecimento de um trabalho relacionado principal as atividades de publicação de dados conectados - "Melhores Práticas para Publicação de Dados Conectados" do W3C;

2. Identificação de processos de publicação de dados governamentais oficiais das nações da América do Sul e artigos científicos e documentos relacionados disponíveis na literatura;

3. Detecção e extração de recomendações que guiem as instituições governamentais a produzirem dados abertos e dados abertos conectados;

4. Associação das recomendações às "Melhores Práticas" estabelecidas pelo trabalho relacionado principal (vide item 1);

A principal contribuição desta revisão de literatura foi a extração de um conjunto de recomendações que poderão ser adotadas na publicação de dados abertos e dados abertos conectados no setor público. Tais recomendações foram incorporadas a proposta de modelo de processo desta pesquisa, cujo detalhamento está contido no capítulo 5.

As próximas seções detalharão as atividades desenvolvidas nesta revisão.

4.1 Revisão de Literatura

Visando extrair recomendações para publicação de dados abertos e dados abertos conectados, alguns documentos oficiais e artigos científicos foram analisados e comparados com as BPLDs do W3C (W3C, 2014), buscando contemplar o objetivo e responder o problema abaixo:

Tema: Uma revisão de literatura sobre processos de publicação de dados abertos e dados abertos conectados aplicáveis ao setor público

Objetivo: Fornecer uma visão geral de trabalhos sobre publicação de dados abertos que contemplam recomendações para a publicação de dados abertos conectados.

Problema (*high level question*): "Como os processos de publicação de dados abertos estão apresentando recomendações compatíveis com as melhores práticas para publicação de Dados Conectados[1]?"

4.1.1 Questões de pesquisa

Decorrente da questão de pesquisa principal, questões específicas foram elaboradas de acordo com os aspectos que foram analisados para cada processo, conforme descrito na Tabela 6:

Tabela 6 – Questões específicas utilizadas na investigação da revisão de literatura

Questões de pesquisa:	**Descrição:**
RQ 1: Os processos de publicação de dados abertos estão contemplando as BPLDs estabelecidas pelo W3C?	Esta questão provê uma visão geral para o entendimento de como as BPLDs estão sendo contempladas pelos processos de publicação de dados abertos, em que intensidade e ainda, se existem práticas que não estão sendo contempladas pelos processos.
RQ 2: Quais BPLDs estabelecidas pelo W3C estão sendo mais consideradas pelos processos de publicação de dados abertos?	Esta questão permite compreender qual(is) BPLDs estão sendo mais contempladas pelos processos. A sua resposta pode indicar a existência de recomendações que são muito relevantes para qualquer processo de publicação de dados.

[1] Este tipo de pergunta se caracteriza como exploratória, buscando o entendimento/esclarecimento das recomendações compatíveis com as Melhores Práticas para publicação de Dados Conectados.

RQ 3: Que recomendações para publicação de Dados Abertos e Dados Abertos Conectados podem ser extraídas dos processos analisados?	Esta questão visa, a partir dos processos, extrair recomendações e orientações para que o publicador de dados possa, de fato, atender as BPLDs ao longo do seu projeto de publicação de dados abertos conectados.
RQ 3.1: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Preparar Partes Interessadas?"	Identificar o que é recomendado para a preparação de partes interessadas nas atividades de publicação de dados abertos ou dados abertos conectados.
RQ 3.2: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Selecionar Conjuntos de Dados?"	Identificar o que é recomendado para a escolha de conjuntos de dados para serem publicados como dados abertos ou dados abertos conectados.
RQ 3.3: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Modelar os Dados?"	Identificar o que é recomendado para a modelagem de dados para serem publicados como dados abertos ou dados abertos conectados.
RQ 3.4: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Especificar uma licença apropriada?"	Identificar o que é recomendado para a utilização de vocabulários para dados que serão publicados como dados abertos ou dados abertos conectados.
RQ 3.5: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Estabelecer bons identificadores universais (URIs)?"	Identificar o que é recomendado para a especificação de identificadores universais em dados para serem publicados como dados abertos ou dados abertos conectados.
RQ 3.6: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Utilizar vocabulários padrão?"	Identificar o que é recomendado para a utilização de vocabulários para dados que serão publicados como dados abertos ou dados abertos conectados.

RQ 3.7: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Converter e enriquecer dados?"	Identificar quais técnicas ou recomendações devem ser adotadas para a conversão e o enriquecimento de dados que serão publicados como dados abertos ou dados abertos conectados.
RQ 3.8: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Prover acesso automatizado aos dados?"	Identificar o que é recomendado para o provimento de acesso automatizado aos dados que serão publicados como dados abertos ou dados abertos conectados.
RQ 3.9: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Anunciar os conjuntos de dados para o público para o público?"	Identificar o que é recomendado para a divulgação de dados que serão publicados como dados abertos ou dados abertos conectados.
RQ 3.10: O que os processos de publicação recomendam a ser feito para contemplar a prática de "Estabelecer um contrato social para os dados publicados?"	Identificar o que é recomendado para a manutenção e disponibilidade de dados que serão publicados como dados abertos ou dados abertos conectados.

Fonte: Autor desta dissertação, 2015.

4.1.2 Seleção de trabalhos

A partir do referencial comparativo estabelecido, por se tratar de uma pesquisa originada no Brasil, a investigação analisou os processos de publicação de dados abertos governamentais de países da América do Sul, sendo localizados apenas os documentos de cinco países da região (Brasil, Chile, Colômbia, Equador e Uruguai).

A escolha da América do Sul como recorte geográfico justifica-se por ser o continente de origem dos pesquisadores envolvidos, o que facilita a compreensão da aplicabilidade de políticas e iniciativas de publicação de dados governamentais. Além disso, os países da região estão associados a um contexto socioeconômico e geopolítico similar, o que facilita a comparabilidade dos países.

Complementarmente, para enriquecer a análise, foram identificados outros artigos científicos e documentos técnicos oriundos do Brasil e de países europeus que apresentam outros processos de publicação de dados abertos governamentais e dados abertos conectados.

Desta maneira, foram estabelecidos os processos de publicação de dados abertos expostos na Tabela 7 para servirem de objeto de estudo, relacionados abaixo:

Tabela 7 – Processos de publicação de dados abertos analisados

Cod.	**Origem**	**Documentos (processos)**	**Tipo**
P1	**Brasil**	Manual para Elaboração de Plano de Dados Abertos; Plano de Dados Abertos - Ministério do Planejamento, Orçamento e Gestão do Brasil; Kit de Dados Abertos; Cartilha Técnica para Publicação de Dados Abertos no Brasil v1.0 (BRASIL, 2014b; BRASIL, 2014c; BRASIL, 2014a; BRASIL, 2011a)	**Dados Abertos Governamentais**.
P2	**Chile**	Norma Técnica para Publicación de Datos Abiertos en Chile (CHILE, 2013b)	**Dados Abertos Governamentais**.
P3	**Colômbia**	*Guía para la Apertura de Datos en Colombia* (COLOMBIA, 2012)	**Dados Abertos Governamentais**.
P4	**Equador**	*Guía de Política Pública de Datos Abiertos - Ecuador* (ECUADOR, 2014)	**Dados Abertos Governamentais**.
P5	**Uruguai**	*Guía rápida de publicación em datos.gub.uy* (URUGUAY, 2012)	**Dados Abertos Governamentais**.
P6	**Itália**	*Geolinked Open Data for the Municipality of Catania* (CONSOLI et al., 2014)	**Dados Abertos Conectados Governamentais**.
P7	**Internacional**	*LOP - Capturing and Linking Open Provenance on LOD Cycle* (MENDONÇA et al., 2013)	**Dados Abertos Conectados**.
P8	**Internacional**	*TWC LOGD: A portal for linked open government data ecosystems* (DING et al., 2011)	**Dados Abertos Conectados Governamentais**.

P9	**Internacional**	*Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers* (BAUER; KALTENBÖCK, 2012)	**Dados Abertos Conectados**.
P10	**Internacional**	*The Joy of Data - A Cookbook for Publishing Linked Government Data on the Web* (HYLAND; WOOD, 2011)	**Dados Abertos Conectados**
P11	**Grécia**	*Applying Linked Data Technologies to Greek Open Government Data: A Case Study* (GALIOTOU; FRAGKOU, 2013)	**Dados Abertos Conectados Governamentais**
P12	**Internacional**	*Managing the Life-Cycle of Linked Data with the LOD2 Stack* (AUER et al., 2012)	**Dados Abertos Conectados**.
P13	**Espanha**	*Methodological Guidelines for Publishing Government Linked Data* (VILLAZÓN-TERRAZAS et al., 2011)	**Dados Abertos Conectados Governamentais**.
P14	**União Europeia**	*Methodology for publishing datasets as open data - COMSODE; Documents of Practice for Methodology for publishing datasets as open data - COMSODE* (COMSODE, 2014b; COMSODE, 2014a)	**Dados Abertos Governamentais**.
P15	**Internacional**	*Open Data Handbook* (OKF, 2015a)	**Dados Abertos**

Fonte: Autor desta dissertação, 2015.

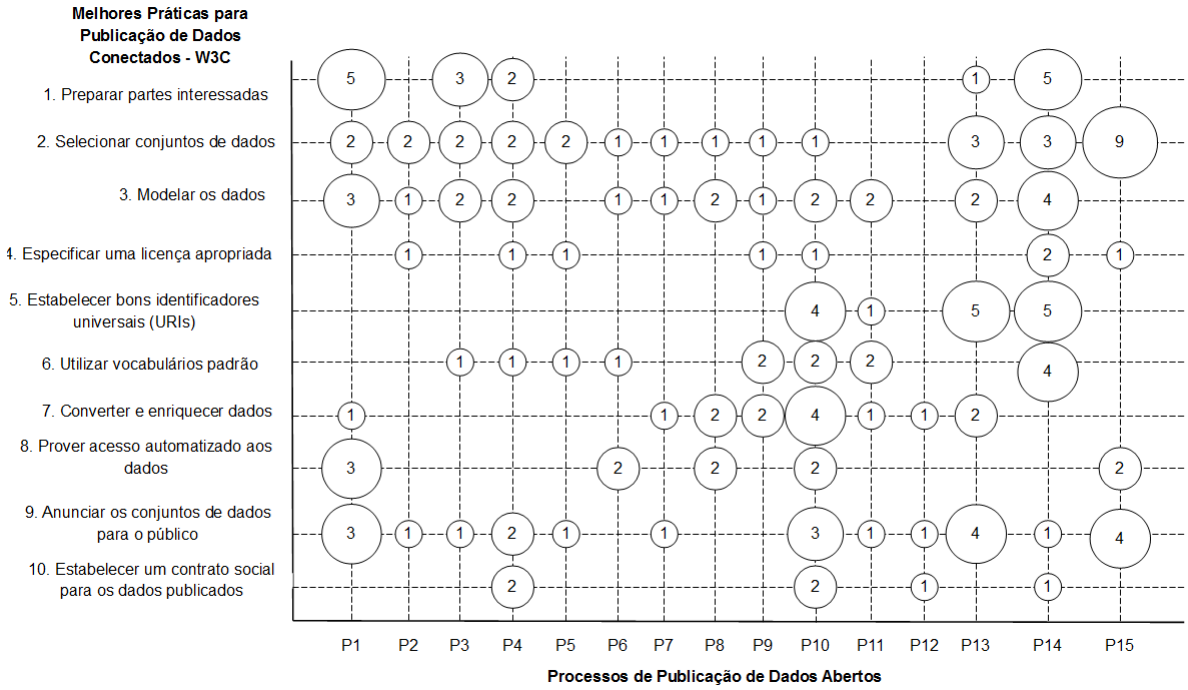
4.2 Análise dos Resultados

Considerando os documentos apresentados na Tabela 7 e as melhores práticas relacionadas, para cada processo de publicação de dados, foram identificadas em cada processo quais recomendações existentes para a implementação de cada melhor prática sugerida pelo W3C. Complementarmente, foi possível identificar quais recomendações por prática estavam presentes em mais de um processo.

Esta comparação permitiu identificar como cada processo considera tais práticas, para fins de estabelecimento de um quadro comparativo dos Processos de Publicação x Melhores Práticas.

A análise dos resultados teve inicio com a identificação das BPLDs do W3C contempladas em todos os processos de publicação. A Figura 28 apresenta a quantidade de recomendações extraídas por processo para uma das BPLDs.

Figura 28 – Ocorrência de recomendações na comparação dos 15 processos de publicação de dados abertos com as BPLDs



Fonte: Autor desta dissertação, 2015.

A seguir, serão apresentadas as respostas para cada questão de pesquisa a partir das informações identificadas.

4.2.1 RQ1: Processos de publicação de dados abertos que contemplam o maior número de BPLDs

Conforme verificado na Figura 03, nenhum dos processos analisados incorporou as 10 BPLDs. O processo P10 apresenta recomendações em 9 das 10 BPLDs, tendo os processos P14 (8/10), P13 (7/10), P4 (7/10) e P1 (6/10) respectivamente. Ademais, os outros processos apresentaram menos recomendações.

Por outro lado, há de se considerar ainda que: (i) alguns dos processos analisados foram propostos antes do estabelecimento das BPLDs; (ii) Foram analisados processos que não contemplam a publicação de dados conectados. Tais fatos justificam a resposta encontrada na revisão, em tempo que apresenta como oportunidade de pesquisa a replicação desta análise com um conjunto maior de processos de publicação ou ainda o desenvolvimento de novos processos de publicação que possam conter recomendações que contemplem as 10 BPLDs.

4.2.2 RQ 2: BPLDs que estão sendo mais consideradas pelos processos de publicação de dados abertos

Considerando a Figura 28, também observamos que nenhuma das 10 BPLDs foram contempladas em todos os processos de publicação. Complementarmente, as práticas que foram contempladas com maior intensidade foram: "Selecionar conjuntos de dados" (contemplada em 13 de 15 processos), "Modelar os dados" (12/15) e ""Anunciar os novos conjuntos de dados para o público" (12/15).

Por outro lado, práticas muito relevantes como "Estabelecer bons identificadores universais (URIs)" (4/15), "Prover acesso por máquina aos dados" (5/15) e "Estabelecer o contrato social do publicador" (5/15) foram contempladas por poucos processos. Em comparação com as práticas mais contempladas, percebe-se que alguns processos concentram suas atividades na rápida disponibilização dos dados, priorizando apenas as práticas indispensáveis à publicação de dados e, por outro lado, não contemplam práticas mais sofisticadas, o que caracteriza pouca preocupação na alta disponibilidade e qualidade dos dados (ex: por não estabelecer boas URIs nem o contrato social do publicador) e ainda, pouca priorização provimento do acesso por máquina.

4.2.3 RQ 3: Recomendações para publicação de Dados Conectados extraídas dos processos analisados

Para responder a esta questão de pesquisa, após a identificação da ocorrência das BPLDs nos processos analisados, a análise buscou extrair, para cada BPLD, o que os processos recomendavam a ser feito. Ao longo das 10 BPLDs, foram identificadas 70 recomendações, conforme descrito na Tabela 8:

Tabela 8 – Recomendações para publicação de dados abertos governamentais conectados extraídas dos processos de publicação

BPLD	**Recomendações**
1. Preparar partes interessadas (*stakeholders*)	(i) Identificar os benefícios para a abertura de dados (1A); (ii)Identificar as Partes Interessadas (1B); (iii)Definir perfis profissionais a serem envolvidos (1C); (iv) Definir grupos de usuários dos dados (1D); (v) Elaborar um plano de ações para publicação dos dados (1E); (vi) Capacitar os envolvidos (1F)

2. Selecionar conjuntos de dados	(i) Analisar a estrutura organizacional (2A); (ii) Estabelecer diretrizes que orientem a priorização de dados a serem abertos (2B); (iii) Realizar consultas aos usuários sobre a demanda de dados (2C); (iv) Identificar os dados que serão abertos (2D); (v) Definir nível de maturidade da abertura (1-5 estrelas) (2E); (vi) Analisar o nível de sigilo dos dados e informações (2F); (vii) Analisar relatórios anuais e documentação existente (2G); (viii) Analisar o esforço para abertura de dados (2H); (ix) Fazer e validar mapa de responsabilidades entre conjuntos de dados e unidades de negócio responsáveis (2I); (x) Identificar e analisar sistemas de informação que poderão ser objeto da abertura de dados (2J); (xi) Identificar dados que podem ser conectados (2K);
3. Modelar os dados	(i) Gerar cópias de segurança das bases de dados que serão abertas (3A); (ii) Higienizar os dados (3B); (iii) Estabelecer rotinas de conversão de dados para formatos legíveis por máquina (3C); (iv) Anonimizar dados sensíveis (3D); (v) Modelar rotinas automatizadas (ETL) (3E); (vi) Analisar se os dados serão conectados ou não; (vii) Estabelecer ou aprimorar documentação de dados (esquemas, vocabulários e ontologias);
4. Especificar uma licença apropriada	(i) Adotar Licenças Não restritivas (4A); (ii) Estabelecer de questões-chave para definição de licenças (4B); (iii) Apresentar opções de licenças a serem adotadas (4C)
5. Estabelecer bons identificadores universais (URIs) para dados conectados	(i) Utilizar URIs para conectar os dados (5A); (ii) Estabelecer URIS persistentes, que não se alterem em nenhum momento (5B); (iii) Proporcionar pelo menos um recurso de dados em formato que seja legível por máquina para cada URI; (iv) Usar URIS como nomes para as coisas; (v) Estabelecer Design simplificado de URIs (5E); (vi) Utilizar identificadores relacionados a informações do mundo real (5F); (vii) Usar URIs HTTP para que recursos de dados possam ser encontrados via *Web* por pessoas e máquinas (5G); (viii) Estabelecer URIs neutras (5H); (ix) Utilizar datas em URIs com moderação (5I); (x) Utilizar hashs (#) em URIs cautelosamente; (xi) URIs das entidades (conjuntos de dados ou recursos) sejam diferentes das URIs das páginas que apresentam estes recursos para a leitura feita por humanos;

6. Utilizar vocabulários padrão	(i) Estabelecer metadados obrigatórios (6A); (ii) Criar um esquema de dados para cada conjunto de dados (6B); (iii) Incentivar o reúso de vocabulários (6C); (iv) Publicar esquemas de dados em arquivos diferentes (6D); (v) Determinar linguagens para expressar esquemas de dados (6E); (vi) Estabelecer critérios de escolha de vocabulários (6F); (vii) Certificar que os dados estão conectados a outros conjuntos de dados (6G); (vii) Desenvolver ou utilizar ontologias para estruturar a semântica dos dados (6H);
7. Converter e enriquecer dados	(i) Converter dados para múltiplas finalidades e usos; (ii) Adotar rotinas ETL para enriquecimento de dados (7B); (iii) Estabelecer bons links com outros conjuntos de dados (7C); (iv) Permitir o envolvimento de várias pessoas na identificação de como os dados a serem convertidos se relacionam com outros dados (7D); (v) Utilizar rotinas automatizadas de conversão de dados, como a triplificação, quando possível (7E); (vi) Converter dados em várias serializações RDF (7F);
8. Prover acesso automatizado aos dados	(i) Disponibilizar bases completas para *download* (*dumps*) (8A); (ii) Estabelecer um Mapa de Decisões Tecnológicas (8B); (ii) Desenvolver uma *API* (8C); (iii) Desenvolver um *endpoint* SPARQL (8D);
9. Anunciar os conjuntos de dados para o público	(i) Publicar metadados junto aos dados (9A); (ii) Estabelecer dados tecnicamente e legalmente abertos (9B); (iii) Disponibilizar os dados com o menor custo possível ao usuário, preferencialmente de modo gratuito na internet (9C); (iv) Divulgar dados em meios complementares (Catálogos, FTP, Torrent) (9D); (v) Divulgar dados em seções destacadas de sítios de governo (9E); (vi) Estabelecer recursos de consulta parcial da base de dados como uma *API* ou *Webservice* (9F); (vii) Estabelecer visualizações e demais recursos de exploração dos dados (9G); (viii) Melhorar os dados para que sejam mais facilmente encontrados por máquinas (9H); (ix) Disponibilizar dados conectados em servidores de triplas (9I);

10.Estabelecer um contrato social para os dados publicados	(i) Estabelecer mecanismos de monitoramento e avaliação da oferta de dados disponibilizados ao público (10A); (ii) Estabelecer espaços para recebimento do feedback do usuário, preferencialmente publicando dados de uma pessoa e/ou telefone de contato para esclarecimento de dúvidas sobre o uso e disponibilidade dos dados (10B); (iii) Disponibilizar leis e atos normativos que explicitem aos usuários quanto às obrigações dos governos em publicarem dados com qualidade e disponibilidade (10C); (iv) Estabelecer com clareza que o processo de publicação contempla etapas de manutenção e atualização dos dados (10D); (v) Utilizar tecnologias que mantenham os dados conectados disponíveis, atualizados e abertos (10E);

Fonte: Autor desta dissertação, 2015.

Será descrito a seguir, para cada BPLD, as recomendações identificadas, sendo analisado o que os processos contribuem para a implementação de tais recomendações.

4.2.4 3.1: Recomendações para *"Preparar Partes Interessadas"*

De acordo com o W3C (2014) a preparação é crucial para o sucesso de um projeto de gestão da informação. Ao compartilhar com as partes interessadas (*stakeholders*) os benefícios esperados pelo projeto, as expectativas são niveladas entre as partes bem como pode ser realizado o alinhamento com a missão da instituição que está realizando o projeto. Além disso, os conceitos sobre produção e gestão de dados e informações passam a ser conhecidos para os técnicos e gestores envolvidos. Por outro lado, a não realização desta etapa pode inviabilizar um projeto, pois os gestores competentes à realização do projeto podem não apoiar nem disponibilizar os recursos necessários ao mesmo e consequentemente, os profissionais necessários ao sucesso do mesmo não estarão devidamente engajados. Cumpre destacar esta prática foi identificada com maior frequência nos planos e processos de abertura de dados com origem nos órgãos governamentais e consequentemente estando pouco presente nos artigos científicos.

Na sequência serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.4.1 Identificar os benefícios para a abertura de dados (1A)

A identificação de benefícios para os interessados é um elemento motivacional no desenvolvimento de qualquer projeto. Nesta direção o processo P14 apresenta um rico detalhamento de como esta atividade deve ser desenvolvida, sugerindo que sejam discutidas as motivações e benefícios gerais (para a organização) e específicos (para cada conjunto de

dados a ser aberto) decorrente de um processo de abertura e publicação de dados. Dentre os principais benefícios gerais, podemos listar (LBC, 2012): (i)Aumentar a transparência; (ii)Estimular o crescimento econômico; (iii)Melhorar os serviços governamentais e capacidade de resposta do governo; (iv)Reduzir as solicitações de dados, considerando que haverá uma oferta proativa de dados governamentais; (v)Incentivar a reutilização de dados; (vi)Melhorar as relações públicas e as atitudes relacionadas ao governo; e (vii)Melhorar os dados e processos do governo.

4.2.4.2 Identificar as partes interessadas (1B)

Outra recomendação identificada nos processos P1, P3 e P14 consiste na correta identificação dos atores interessados na publicação de dados.

O processo P3 recomenda, de um modo geral, a identificação das partes interessadas. O processo P1 acrescenta que esta identificação deve estar contida num plano de abertura de dados governamentais, devendo ser relacionadas e engajadas as principais partes interessadas. O processo P14 complementa que esta identificação deve ser feita dentro e fora da organização publicadora, visando registrar as percepções da ótica de quem publica e de quem consome os dados, gerando sinergia e engajamento entre publicadores e consumidores.

Ademais, negligenciar esta recomendação pode resultar num impedimento ao inicio do processo de abertura e publicação de dados. É necessário saber quem são os atores com poder de decisão para estimular ou inviabilizar o processo. Do ponto de vista técnico, deve-se identificar quem são os atores que possuem capacidade e disposição técnica para atuar no processo, e ainda, é recomendável a identificação dos principais consumidores dos dados, de tal maneira que a oferta de dados a ser gerada seja coerente com as expectativas e suas demandas.

É importante ressaltar que, especialmente quando da abertura de dados que ainda não foram publicados, poderá haver uma tarefa dispendiosa e tediosa que demandará um contato mais intenso e negociado com os "proprietários" dos dados para que seja concedido o acesso aos mesmos (VILLAZÓN-TERRAZAS et al., 2011). Tal situação enfatiza a importância do desenvolvimento da etapa de preparação das partes interessadas, especialmente para a publicação de dados que ainda não foram abertos.

4.2.4.3 Definir perfis profissionais a serem envolvidos (1C)

Alguns processos recomendam que sejam definidos os perfis profissionais a serem envolvidos com o processo de abertura e publicação.

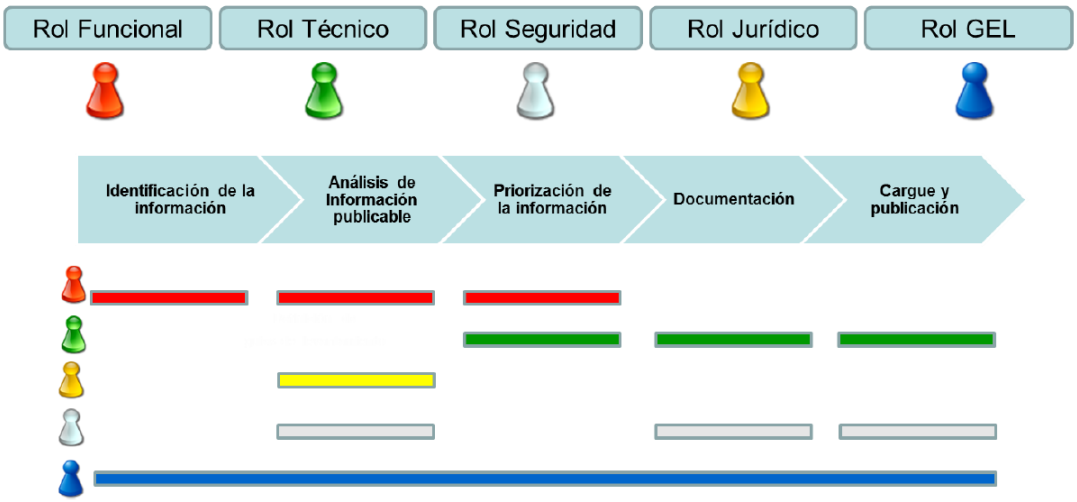
O processo P1 sugere que sejam envolvidos os gestores das áreas responsáveis pelas bases de dados que serão objetos da abertura, especialistas com conhecimento do negócio, o gestor da área de TI e os desenvolvedores e analistas de dados que tenham conhecimento da base de dados. O processo P4 (ECUADOR, 2014) recomenda que seja estabelecido um

"Comitê de Dados Abertos" para cada instituição publicadora, que deve ser composto pelo Diretor de Tecnologia da Informação, o Coordenador de Planejamento, o Coordenador Jurídico e representantes em nível técnico para atividades de apoio e assessoramento. Este comitê tem como atribuições a localização, identificação, catalogação, publicação e atualização de dados abertos publicados no Portal Equatoriano de Dados Abertos.

O processo P14 não apresenta especificamente quais profissionais devem ser envolvidos no processo de publicação, todavia, estabelece a existência de vários papéis profissionais que serão desempenhados pela equipe técnica envolvida e que possuirão alguma responsabilidade ao longo das etapas deste processo de abertura, que são: Publicador, Proprietário dos dados, Curador, Proprietário do Catálogo de Dados Abertos, Coordenador de Dados Abertos, Publicador do Catálogo de Dados Abertos, Profissional de TI, Gestor de Qualidade de Dados, Especialista de Qualidade de Dados e Especialista Jurídico.

O processo P3 estabelece a definição de uma equipe de trabalho, estabelecendo com clareza os perfis profissionais que devem ser envolvidos, suas atribuições com as atividades de publicação de dados e ainda, os momentos de atuação de cada perfil profissional ao longo de todo o processo, conforme esclarecido na Figura 29

Figura 29 – Papéis e atividades necessárias para desenvolver a publicação de dados no processo P3



Fonte: COLOMBIA (2012)

O processo P3 sugere o engajamento de profissionais com perfis de negócio (*Funcional*), de tecnologia da informação (*Técnico*), gestão da qualidade e riscos (*Seguridad*), jurídico e de comunicação (GEL), destacando que devem ser estabelecidos estes papeis ao longo do projeto, independente do cargo que tais profissionais ocupem, sendo permitido o acumulo de papéis por um mesmo profissional.

4.2.4.4 Definir grupos de usuários dos dados (1D)

Além da identificação das partes interessadas e dos perfis profissionais, os processos P1 e P14 também recomendam a identificação e envolvimento, quando possível, dos principais grupos de usuários dos dados. O processo P1 estabelece, de forma geral, que estes usuários costumam ser os próprios órgãos governamentais, empresas e especialistas. O processo P14 identifica estes como: Órgãos governamentais e seus servidores; Empresas; Organizações não governamentais e associações que atuam com controle social, acompanhamento e fiscalização das ações do setor público; Desenvolvedores de aplicativos; Jornalistas; Outras organizações, como bancos e instituições do mercado financeiro; Os próprios funcionários do órgão governamental publicador; Comunidade acadêmica; E ainda, cidadãos de um modo geral (COMSODE, 2014a).

4.2.4.5 Elaborar um plano de ações para publicação dos dados (1E)

Para organizar e estruturar as atividades ao longo do processo de abertura e publicação dos dados, vários processos recomendam a elaboração de um plano de ações. Os processos P1, P3 e P14 estabelecem que deva existir inicialmente um plano de abertura de dados governamentais recomendando ainda que este plano deve ser detalhado mediante o estabelecimento de uma matriz de responsabilidades pelo preparo e atualização dos dados e respectivo detalhamento com metas e prazos.

No processo P4, compete a um comitê de dados abertos criar e executar um planejamento para a publicação dos dados abertos da instituição publicadora, contemplando a conversão dos dados disponibilizados em seções de Transparência Pública para formatos abertos bem como identificar outros dados de interesse público que possam ser ofertados em formato aberto.

Complementarmente, é desejável que os usuários dos dados sejam engajados no desenvolvimento do plano de ações. A participação dos usuários permite a obtenção de informações a respeito da real demanda por dados. Esta identificação da demanda também poderá ser utilizada para subsidiar as atividades de seleção e priorização dos dados a serem abertos. Dentre as principais técnicas de engajamento que podem ser utilizadas são: enquetes, questionários, escolha e priorização online de uma lista pré-estabelecida de conjuntos de dados, workshops, audiências públicas, conferências, dentre outros (COMSODE, 2014a).

4.2.4.6 Capacitar os envolvidos (1F)

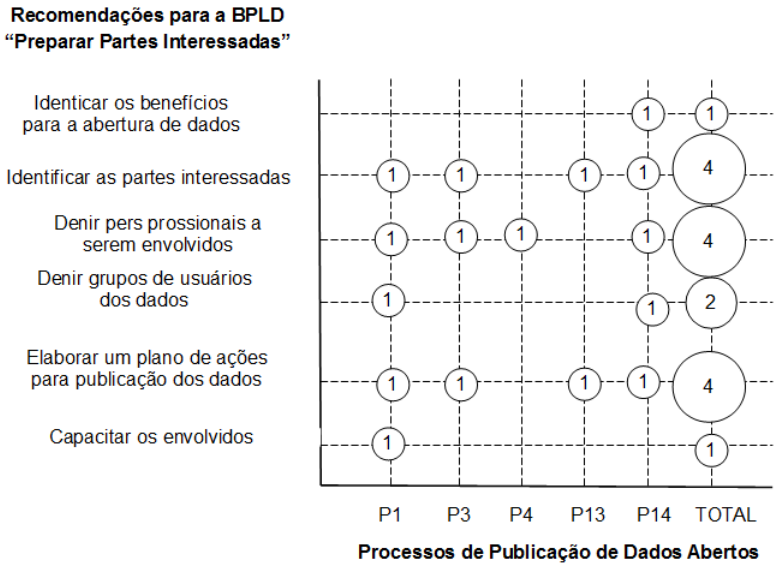
Considerando a identificação e engajamento prévio dos atores nas atividades de publicação de dados, o processo P1 destaca a importância de capacitar os técnicos e responsáveis nas áreas de negócio dos dados selecionados para abertura sobre alguns temas como: (1) O processo de publicação de dados abertos; (2) O processo de catalogação dos

metadados no Portal Nacional de Dados Abertos e a inclusão de dados em Infraestruturas Nacionais de Dados Espaciais, quando se tratarem de dados geoespaciais (BRASIL, 2014b).

4.2.4.7 Sumarização dos resultados

Da análise desta BPLD, foram extraídas 06 recomendações, conforme descrito nesta seção. Apesar de ser uma etapa fundamental para o inicio e o êxito de um processo de publicação de dados, apenas 05 dos 15 processos apresentaram ações de preparação de partes interessadas. A Figura 30 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram.

Figura 30 – Identificação de recomendações para a "Preparar Partes Interessadas" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Considerando a Figura 30, podemos observar a existência de recomendações mais relevantes a partir da ótica dos processos de publicação. Desta maneira, pode-se deduzir que, considerando cada recomendação identificada como uma atividade de um processo de publicação, as atividades de *"Identificar as partes interessadas"*, *"Definir perfis profissionais a serem envolvidos"* e *"Elaborar um plano de ações para publicação dos dados"* são fortemente recomendadas e que as demais atividades seriam desejáveis de serem cumpridas.

4.2.5 RQ 3.2: Recomendações para "Selecionar Conjuntos de Dados"

A segunda BPLD estabelecida consiste na seleção dos conjuntos de dados que serão publicados. Segundo o W3C (2014), devem ser selecionados apenas os dados que são catalogados ou criados pela instituição que está implementando o processo de abertura.

Preferencialmente, devem ser priorizados dados que, ao serem combinados com outros dados, produzam grande valor. Esta mensuração de valor deve ser guiada pelo potencial de reúso do dado e sua popularidade. Dados de natureza geoespacial, saúde, legislação, população e demografia costumam ser dados bem demandados e sua publicação como dados conectados deve ser priorizada.

A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.5.1 Analisar a estrutura organizacional (2A)

Para facilitar o entendimento da organização publicadora de dados, recomenda-se analisar a estrutura organizacional, visando identificar a sua complexidade, níveis hierárquicos, cultura organizacional, perfis profissionais, serviços que oferta ao público, principais clientes, dentre outros. O processo P14 recomenda que seja analisado, além da estrutura organizacional, a legislação envolvida e as regras e normas adotadas e estabelecidas pela instituição, bem como os documentos que descrevem o planejamento e agendas estratégicas. Sugere ainda a identificação prioritária das unidades organizacionais e respectivos líderes que estão envolvidos com a coleta, criação ou gestão de dados que poderão ser potenciais conjuntos de dados abertos COMSODE (2014b). Importante registrar tais dados numa relação de dados possíveis para serem abertos e conectados.

4.2.5.2 Estabelecer diretrizes que orientem a priorização de dados a serem abertos (2B)

Outra recomendação identificada visa o estabelecimento de diretrizes e questões-chave que orientem a priorização dos dados a serem abertos. O processo P14 contém um rico detalhamento sobre que tipos de dados devem ser priorizados durante um processo de abertura (COMSODE, 2014b). O processo P2 (CHILE, 2013a) estabelece que a priorização dos dados a serem publicados passe pela seleção do que é mais requisitado pelo cidadão.

Por outro lado, o processo P5 (URUGUAY, 2012) recomenda que sejam priorizados os dados que são de mais fácil transformação e acesso para serem publicados. Cumpre destacar que um processo de abertura de dados é interativo e por esta razão, o publicador pode retornar a esta etapa de escolha dos conjuntos de dados mesmo após ter desenvolvido outras etapas OKF (2015a).

Os processos P2 e P5 sugerem o estabelecimento de algumas perguntas-chave que ajudarão na identificação dos dados a serem priorizados na abertura, conforme relação a seguir CHILE (2013b), URUGUAY (2012):

- Que informação é entregue com maior frequência aos cidadãos através dos meios de solicitação de acesso às informações públicas?

- Que informações consideradas pela instituição como de interesse público são entregues à imprensa com maior frequência?
- Que informações são entregues para outras instituições regularmente e que podem ser ofertadas amplamente ao público?
- Que informações da sua instituição atendem aos requisitos de dados abertos e podem ser publicadas?
- Que informações da sua instituição ainda não atendem aos requisitos de dados abertos, mas que podem ser facilmente convertidas para dados abertos?
- Que informações são solicitadas habitualmente e que exigem um processamento de dados para serem entregues?

Complementarmente ao estabelecimento de perguntas-chave sobre quais dados abrir, o processo P15 sugere que o publicador faça uma lista curta de conjuntos de dados sobre os quais pode se haver retorno, onde esta lista também pode ser baseada noutros catálogos de dados existentes OKF (2015a). Não é essencial que essa lista coincida com as suas expectativas. O principal objetivo aqui é mensurar a demanda. Ela pode ser baseada nos catálogos de dados abertos de outros países. Este processo recomenda o estabelecimento de consultas públicas como elementos relevantes para se mensurar as demandas dos clientes das organizações publicadoras.

4.2.5.3 Realizar consultas aos usuários sobre a demanda de dados (2C)

Os processos P4, P14 e P15 recomendam que seja estabelecida uma sistemática periódica de se consultar a comunidade sobre quais dados são demandados para abertura, mediante uma consulta pública, disponível numa página da *Web*. Os processos sugerem que a consulta seja feita da forma mais acessível, mediante uma página *Web* e URL simples e que possa ser compartilhada em listas de e-mail, fóruns e em mídias sociais ECUADOR (2014), URUGUAY (2012), OKF (2015a). O processo P15 sugere ainda que a consulta deve facilitar ao máximo o envio de respostas, desencorajando a obrigatoriedade de identificação dos respondentes. Complementarmente pode ser realizada uma audiência pública para discutir os resultados da consulta pública e ainda, captar novas sugestões de dados a serem abertos OKF (2015a).

Por fim, é desejável o apoio explícito de algum agente político que anuncie esta intenção de abertura, pois dará maior abrangência e relevância para esta atividade junto ao público, conforme os processos P15 e P4, pois esta prática contribui para o estabelecimento de uma cultura de uso e reúso de dados motivando os usuários a consumirem os dados ofertados com maior frequência, bem como demandar novos dados ECUADOR (2014, p. 12).

Ademais, após analisar as demandas de informação oriundas dos usuários, sugere-se identificar quais conjuntos de dados da organização possuem alta relevância e múltiplos

usuários e comparar com a demanda dos usuários, conforme sugerido pelo processo P14 (COMSODE, 2014a).

4.2.5.4 Identificar os dados que serão abertos (2D)

Outra recomendação presente em vários processos consiste na identificação de quais dados serão abertos e publicados. Devem ser identificados os dados que ainda não foram abertos e publicados, bem como os dados que já foram publicados, mas que serão reusados, sendo publicados num formato mais enriquecido, conforme sugere o processo P13 (VILLAZÓN-TERRAZAS et al., 2011, p. 05).

Posterior à etapa de consulta pública, ao se identificar as informações candidatas a serem publicadas, devem ser selecionadas prioritariamente aquelas que se encontrem em condições imediatas de serem publicadas, conforme o processo P5, devendo ser considerado ainda os aspectos legais, de completude, capacidade para manter a informação atualizada, formatos, dentre outros. O processo P2 sugere que deve ser evitada a publicação de arquivos que possuam apenas parágrafos de texto em sua totalidade. Cumpre destacar que o processo P15 ressalta que, apesar de existirem abordagens que priorizem a publicação de dados que sejam mais fáceis de disponibilizar ao público, deve ser considerado se tais dados sejam relevantes, pois a publicação de dados que não tenham relevância podem prejudicar a credibilidade da iniciativa, dando a entender que a abertura de dados não considera o que é relevante para a sociedade, mas sim, o que é mais simples de se disponibilizar.

Para esta identificação dos conjuntos de dados, o processo P14 deve se registrar no mínimo, as seguintes informações COMSODE (2014b):

- Título e descrição
- Unidade organizacional responsável
- Pessoa de contato (Para consultas sobre o conjunto de dados)
- Formatos dos recursos de dados (Armazenado num banco de dados relacional ou não, armazenado como arquivos de dados tabulares em in XLS(X), ODS, XML, CSV ou ainda apenas em arquivos de texto não estruturados ou semi estruturados) e uma breve descrição de cada formato.

Caso o publicador deseje uma descrição mais detalhada, os processos P3 e P4 ainda sugerem outras informações adicionais relevantes que podem ser registradas para cada conjunto de dados identificado (COLOMBIA, 2012).

4.2.5.5 Definir nível de maturidade dos dados a serem publicados (1-5 estrelas) (2E)

Considerando o cumprimento das recomendações de se identificar os dados que serão abertos e publicados, e ainda, os que tem potencial para serem conectados, sugere-se que

seja definido previamente o nível de enriquecimento dos dados que serão publicados. A definição deste nível de maturidade servirá para nortear quais dados devem ser selecionados para publicação, considerando que, quanto maior o nível de maturidade, maior o esforço necessário para publicação. Desta maneira, a análise de custo-benefício entre conjunto de dado x nível de maturidade da publicação consiste de atividade relevante para o planejamento da publicação de dados abertos.

Nesta direção, para cada conjunto de dados, o processo P14 sugere seja definido o nível (alvo) da atividade de abertura, conforme o esquema 5 estrelas para dados abertos Berners-Lee (2006). O processo P14 destaca que, no mínimo, deve ser estabelecido o nível 3 de enriquecimento, sendo aceitável o nível 2 em casos especiais cujos dados existam exclusivamente em documentos não estruturados e que não seja possível a sua organização para convertê-los num documento estruturado COMSODE (2014b).

Para cada nível de abertura, este processo recomenda uma série de requisitos e providências a serem adotadas, sugerindo ainda atenção para que seja definido o período de atualização de cada conjunto de dados, de que maneira os dados serão disponibilizados (num único arquivo, particionado em vários arquivos, mediante uma *API Rest, endpoint SPARQL*, etc.), e ainda, para que sejam disponibilizadas as séries históricas de atualização do arquivo, quando for o caso.

4.2.5.6 Analisar o nível de sigilo dos dados e informações (2F)

É importante ressaltar que, apesar das leis vigentes nas nações recomendarem que o acesso à informação deva ser regra e o sigilo ser aplicado somente em casos excepcionais, há de se considerar que alguns dados governamentais não podem ser abertos por questões como ameaça a defesa nacional, ao desenvolvimento científico e tecnológico, dentre outros. Assim sendo, outra recomendação identificada consiste na análise do nível de sigilo dos dados e informações que podem ser considerados como objeto de abertura e publicação.

O processo P3 COLOMBIA (2012, p. 20) sugere que as organizações estabeleçam os níveis de sigilo dos dados e informações, classificando em:

- Dados e informações publicáveis: Aquelas que devem estar à disposição de qualquer pessoa e que o governo é obrigado a disponibilizar, considerando que não há obrigação legal para mantê-las como sigilosa;
- Dados e informações não publicáveis: Aquelas que se enquadram como reservadas ou sigilosas conforme os parâmetros legais vigentes;
- Dados e informações pessoais (semi-privadas): Dados e informações pessoais que não são de domínio público, mas foram obtidas por ordem de uma autoridade administrativa no exercício das suas funções ou no âmbito dos princípios de gestão de dados pessoais. Esta informação pode ou não pode estar sujeito ao sigilo, dependendo do caso.

O processo P1 tem como base legal a Lei Brasileira de Acesso a Informação (Lei Federal 12.527/2011) BRASIL (2011b) que aborda com clareza os casos e procedimentos para a adoção do sigilo das informações. Segundo o artigo 24 desta lei, a informação pública poderá ser classificada como reservada, secreta ou ultrassecreta e os respectivos prazos para se tornarem públicas são de 5 (cinco), 15 (quinze) e 25 (anos). O artigo 23 da lei estabelece os casos em que as informações devem ser classificadas como sigilosas.

4.2.5.7 Analisar relatórios anuais e documentação existente (2G)

O entendimento do contexto organizacional é fator crítico de sucesso num processo de abertura de dados e por este motivo, o processo P14 sugere que devem ser localizados e analisados os relatórios anuais (balanços financeiros, relatórios de gestão, avaliação de desempenho e de projetos, etc.) e ainda outros documentos relevantes, incluindo o portal *Web*, que informem sobre as atividades e principais resultados alcançados da organização publicadora. Devem ser identificadas tabelas e gráficos nestes documentos que orientem sobre potenciais conjuntos de dados para serem abertos e conectados e ainda, deve se identificar quais unidades organizacionais são responsáveis pela sua preparação. Tais informações devem ser devidamente registradas numa relação de dados possíveis para serem abertos e conectados (COMSODE, 2014b).

Complementarmente, o processo P13 devem ser identificados ainda a documentação e os esquemas dos dados que serão objetos de publicação (quando houverem), bem como suas propriedades, relações, sistemas de informação geradores destes dados e arquiteturas *Web* (VILLAZÓN-TERRAZAS et al., 2011).

4.2.5.8 Analisar o esforço para abertura de dados (2H)

Ademais, para que a publicação de dados seja exitosa e factível, o processo P14 ainda sugere que seja analisado o esforço de abertura de cada conjunto de dados. A análise de esforço deverá ser realizada a partir da priorização dos dados a serem publicados podendo, inclusive, resultar numa nova priorização, caso os dados inicialmente destacados para iniciarem o projeto de abertura tenham complexidade e demandem esforço muito alto (COMSODE, 2014b).

4.2.5.9 Fazer e validar mapa de responsabilidades entre conjuntos de dados e unidades de negócio responsáveis (2I)

Outra recomendação destacada pelo processo P14 consiste na criação um mapa de responsabilidades entre os conjuntos de dados a serem publicados e respectivas unidades de negócio responsáveis, sendo este mapa graficamente representado. Especialmente, caso a unidade de negócio dependa do apoio de outra unidade, deve ficar explicitado às responsabilidades. Por exemplo, uma unidade de negócio que realiza e publica informações

tabulares sobre o desempenho escolar pode depender de outra unidade de negócio para georreferenciar estes dados. Ademais, a relação de conjuntos de dados e o mapa de responsabilidades deve ser discutido e validado com pessoas que tenham poder de decisão sobre a publicação de dados (COMSODE, 2014a)

4.2.5.10 Identificar e analisar sistemas de informação que poderão ser objeto da abertura de dados (2J)

A análise e identificação dos principais sistemas de informação da organização e os principais conjuntos de dados gerenciados por estes sistemas representa outra recomendação relevante, de acordo com o processo P14 (COMSODE, 2014b). O processo P1 complementa sugerindo que seja estabelecida uma arquitetura para abertura de dados a partir de cada sistema de informação identificado (BRASIL, 2014c).

Outro item sugerido consiste na análise se as fontes dos dados são *Datawarehouses*. De acordo com o processo P6, quando for o caso, destacar a necessidade da construção de rotinas de extração/conversão destes dados ao iniciar o processo de publicação, devendo os descritores dos dados ser traduzidos para o idioma inglês ou o idioma nativo do país, quando aplicável (CONSOLI et al., 2014, pg. 03).

4.2.5.11 Identificar dados que podem ser conectados (2K)

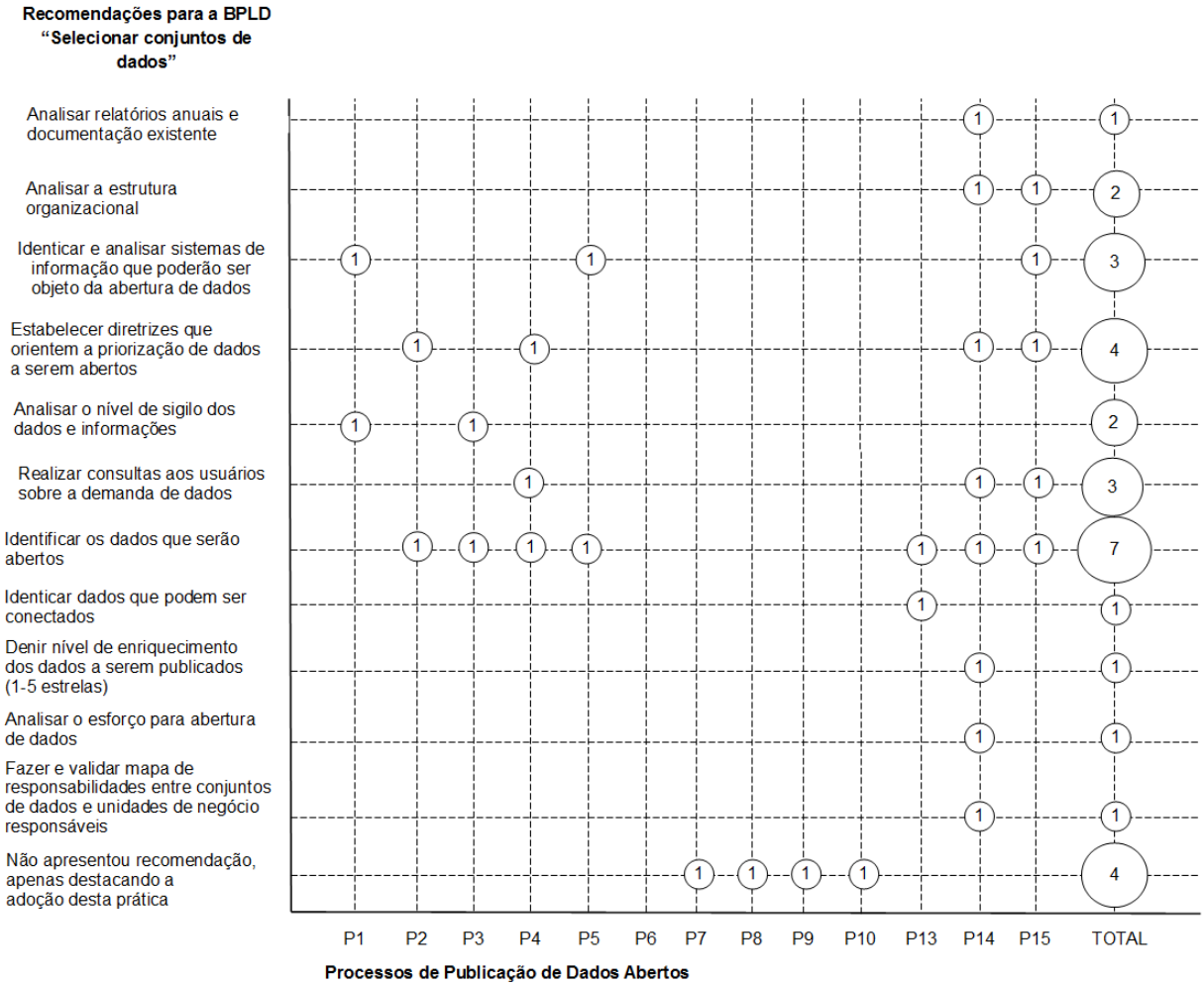
Além da identificação dos dados que serão abertos, sugere-se que sejam identificados os dados que tem potencial para serem conectados com menor esforço. Quando se tratar de um enriquecimento e reúso de dados já abertos e publicados por exemplo, devem ser procurados dados similares e relacionados (que possam se conectar ao dado que será publicado) em outros catálogos governamentais, como sugerido pelo processo P13 (VILLAZÓN-TERRAZAS et al., 2011). É destacado que o êxito esta atividade está relacionado com a recomendação de identificar a documentação e os esquemas dos dados que serão publicados, bem como suas propriedades, relações e arquiteturas *Web*. Outros autores, também destacam a identificação destes requisitos, entretanto, na etapa posterior, de modelagem dos dados.

4.2.5.12 Sumarização dos resultados

Da análise desta BPLD, foram extraídas onze recomendações, conforme descrito nesta seção. Onze dos quinze processos analisados estabelecem a BPLD *"selecionar conjuntos de dados"* como relevante, onde em alguns deles, esta etapa consta como a inicial. Todavia, foram extraídas recomendações de sete processos, pois outros quatro processos apresentam, num nível mais simplificado, atividades similares aos outros sete processos descritos nesta seção (MENDONÇA et al., 2013; DING et al., 2011; BAUER; KALTENBÖCK, 2012; HYLAND; WOOD, 2011).

A Figura 31 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram:

Figura 31 – Identificação de recomendações para a BPLD "Selecionar Conjuntos de Dados" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Considerando a Figura 31, para esta BPLD, observamos que as recomendações *"Identificar os dados que serão abertos" (7/13), "Estabelecer diretrizes que orientem a priorização de dados a serem abertos"(4/13)* e *"Realizar consultas aos usuários sobre a demanda de dados" e "Identificar e analisar sistemas de informação que poderão ser objeto da abertura de dados" (3/13)* foram contempladas em mais processos. As demais recomendações, apesar de serem relevantes, podem ser consideradas como recomendações desejáveis a serem cumpridas. Além disso, cumpre registrar que o processo P14 contemplou a maioria das recomendações extraídas nesta seção.

4.2.6 RQ 3.3: Recomendações para "Modelar os Dados"

Para esta BPLD, o W3C (2014) destaca a necessidade do envolvimento dos respon-

sáveis técnicos pelos dados, incluindo os gestores das bases de dados e os responsáveis por padrões e políticas de gestão da informação. Administradores de Bancos de Dados (DBAs) da organização devem ser envolvidos se os dados a serem publicados tiverem origem em bancos de dados relacionais utilizados por sistemas de informação.

Esta etapa também requer um estudo da documentação dos dados, podendo ser necessárias algumas reuniões de esclarecimento. Após o entendimento sobre os dados que serão publicados, devem ser explanados os conceitos de Dados Abertos e Dados Abertos Conectados e posteriormente, serem analisados os relacionamentos entre os conjuntos de dados. O W3C recomenda que seja evidenciado o maior número de outros conjuntos de dados que se relacionam com o conjunto de dados que será o objeto de publicação.

Os processos P1, P2 e P4 descreveram atividades aplicáveis a dados abertos que podem ser utilizadas também para dados abertos conectados. Os demais processos analisados (P6, P7, P8, P9, P10 e P13) são voltados à publicação de Dados Conectados. Todos os processos analisados, em sua maioria, contemplam também as atividades de modelagem de dados não-conectados e agregam novas atividades.

A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta boa prática em atividades de publicação de dados.

4.2.6.1 Gerar cópias de segurança das bases de dados que serão abertas (3A)

Considerando que a abertura de dados demanda integridade e confiabilidade das fontes originais, esta recomendação sugere que, inicialmente, sejam geradas cópias de segurança em todas as bases de dados que serão objeto das atividades de abertura e conexão de dados conforme sugerido pelos processos P8, P10 e P11. Complementarmente, esta recomendação deve ser adotada com maior prioridade quando se fizer necessária à adoção de técnicas de higienização em bases de dados (HYLAND; WOOD, 2011; GALIOTOU; FRAGKOU, 2013; DING et al., 2011).

4.2.6.2 Higienizar os dados (3B)

A melhoria da qualidade dos dados pode ser obtida mediante técnicas de higienização (limpeza dos dados), sendo possível à identificação de possíveis erros e inconsistências das bases de dados que serão posteriormente corrigidos conforme sugerido nos processos P1, P7, P8 e P13 (BRASIL, 2014b; MENDONÇA et al., 2013; DING et al., 2011; VILLAZÓN-TERRAZAS et al., 2011). No caso de dados e informações oriundas de muitas fontes de dados distintas, distribuídas e representadas em vários formatos diferentes (por exemplo, bases de dados, XML, CSV, dados geoespaciais, etc.) estas exigem um esforço adicional para assegurar modelagem fácil e eficiente. Isso inclui livrar seus dados e informações de qualquer informação adicional que não será incluída nos conjuntos de dados publicados conforme o processo P9 (BAUER; KALTENBÖCK, 2012).

4.2.6.3 Estabelecer rotinas de conversão de dados para formatos legíveis por máquina (3C)

Considerando que os dados publicados costumam ser utilizados por diversos públicos distintos e que fazem uso de tecnologias e formatos distintos, e ainda, que o volume de dados a serem publicados e mantidos costuma aumentar, outra recomendação relevante consiste no estabelecimento de rotinas de conversão de dados para vários formatos legíveis por máquina. Os processos P1, P2, P3 e P4 buscam detalhar esta etapa (COLOMBIA, 2012; ECUADOR, 2014). Recomendam que, posteriormente à modelagem, os dados sejam convertidos para formatos legíveis por máquina, como o XML, CSV, TXT, JSON, KML ou RDF. Devem ser eliminados conteúdos que não sejam relevantes ao usuário, como títulos, subtítulos e informações extra dos arquivos. O P3 enfatiza que as rotinas de conversão dos dados também contemplem a geração de metadados que detalhem a estruturação dos arquivos de dados.

4.2.6.4 Anonimizar dados sensíveis (3D)

Em que pese as políticas de dados abertos estimularem a publicização dos dados, este processo de abertura e publicação de dados deve ser feita com muita responsabilidade de maneira que não cause prejuízos a indivíduos e organizações. Assim, a recomendação de anonimização dos dados foi identificada como a técnica a ser adotada para não expor dados privados/particulares no arcabouço de uma oferta de dados públicos.

Janssen, Charalabidis e Zuiderwijk (2012) apresentam alguns motivos para que nem todos os dados sejam publicizados. Dentre eles, destacamos: (i) Dados podem permitir o rastreamento reverso chegando a identificação de indivíduos e resultando em violação de privacidade e direitos individuais; (ii) A abertura de dados inconsistentes podem gerar mais "confusões" do que benefícios, pois os cidadãos podem não obter as respostas que desejam e ainda, gerar questionamentos desnecessários as agências governamentais decorrentes de uma má interpretação dos dados; (iii) As legislações dos países apresentam casos explícitos em que certos dados devem ser restritos; e (iv) Certos dados são estratégicos e necessários a políticas de competitividade (por exemplo, dados sobre prospecção de recursos minerais são essenciais para a sustentabilidade de países e podem influenciar a disputa comercial e tecnológica entre empresas públicas que atuam neste setor).

A anonimização de dados é uma tarefa complexa, e se não for feita de forma eficaz, cria riscos a iniciativa de publicação de dados, especialmente por permitir a revelação de dados privados que não devem ser publicados. Apesar da importância desta atividade, apenas os processos P4 e P14 apresentaram recomendações e técnicas a serem adotadas, detalhadas abaixo (COMSODE, 2014b):

- Projeção (*projection*): Ocorre quando atributos particulares com dados privados são removidos do conjunto de dados. Por exemplo, no caso de arquivos tabulares, isto pode ser implementado mediante a remoção de colunas.

- Agregação (*aggregation*): Consiste na mesclagem de vários itens num único dado estatístico (por exemplo, a mesclagem de pessoas e suas idades numa região, publicando-se apenas a idade média das pessoas em cada região).

- Remoção de conexões (*removing links*): Providência que deve ser adotada especialmente quando se tratar de dados conectados, devendo ser analisado se as conexões com outros dados revelam dados privados. Caso isto ocorra é necessário remover os links antes de publicar o conjunto de dados.

Cumpre destacar que a anonimização consiste de uma estratégia de mitigação de riscos relacionados ao processo de publicação e caso for negligenciada, pode inviabilizar toda a estratégia de abertura decorrente dos impactos negativos da publicação de dados que não deveriam ser publicados.

4.2.6.5 Modelar rotinas automatizadas (ETL) (3E)

Além da oferta de dados em vários formatos, é recomendado que estas rotinas de publicação e manutenção dos dados sejam automatizadas, reduzindo o esforço humano com esta atividade e ainda, ofertando maiores garantias de disponibilidade e atualização dos dados para os usuários. Para esta atividade, serão apresentadas as recomendações de diversos processos com algum nível de detalhamento.

Para automatizar a publicação de dados, os processos P1, P5 e P14 recomendado o estabelecimento de rotinas de extração, tratamento e carga (ETL). A publicação manual de dados deve ser estabelecida apenas para dados que não possuem atualização periódica.

O processo P14 detalha tópicos relevantes que devem ser estabelecidos na modelagem de rotinas automatizadas. Para os extratores, deve ser fortemente considerada a origem dos dados a serem publicados. Dependendo desta origem, um extrator pode ser COMSODE (2014b):

- Um componente que faz *download* de um arquivo de dados a partir de uma dada URL;

- Um componente que copia 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).

Quanto aos transformadores estes podem ser:

- Componentes para transformar a estrutura e os formatos de dados, como:

  - Um componente para transformar formatos proprietários tabulares (XLS (x), ODS, DBF, etc.) e os resultados de consultas SQL para o formato CSV;
  - Um componente para transformar arquivos XML para outros arquivos XML na base de scripts XSLT;
  - Um componente para transformar arquivos JSON para outros arquivos JSON;
  - Um componente para transformar arquivos JSON para arquivos XML e vice-versa.
  - Um componente para transformar CSV, XML e JSON formatos de representação RDF. Em caso de XML, que pode ser baseada em scripts XSLT.
  - Um componente para transformar representação RDF usando a linguagem SPARQL.

- Ou ainda, componentes que transformem o conteúdo de um conjunto de dados aplicando técnicas de higienização ou anonimização de dados;

- Bem como, componentes para enriquecimento de dados associando-os ao conteúdo de outros conteúdos de dados decorrente de conexões pré-estabelecidas;

- Por fim, um transformador pode ser um componente de preenchimento automatizado e manual de metadados em conjuntos de dados de acordo com um esquema (ou vocabulário) de dados pré-estabelecido.

Quanto aos carregadores, consistem da etapa final antes da publicação do dado. São componentes que garantem que o dado exportado da origem estará armazenado num servidor de dados com a qualidade e os formatos adequados para serem publicados. O processo estabelece as seguintes recomendações para carregadores:

- Se o conjunto de dados estará disponível apenas para usuários que farão o *download* de dados em grandes volumes, a rotina ETL deve carregar os arquivos de dados para um local que pode ser acessado por usuários via protocolos HTTP ou FTP. Também é possível carregar os arquivos para um servidor *Git*, por exemplo, o *Github.com*.

- Se o conjunto de dados estará disponível através de uma *API*, a rotina ETL deve carregar os dados para um servidor de banco de dados.

  - Para a oferta de dados em 3 e 4 estrelas, a *API* deve ser um serviço REST que seja capaz de fornecer o acesso programático para os itens do conjunto de dados e retornar a representação dos itens em formatos JSON, CSV, ou XML. Os dados devem ser armazenados numa base de dados relacional ou numa base de dados noSQL.

– Para a oferta de dados em 5 estrelas, a *API* deve ser um *endpoint* SPARQL. Os dados devem ser armazenados em um banco de dados RDF ou em um banco de dados relacional com uma camada que permite visualizar os dados relacionais como dados RDF e avaliar consultas SPARQL.

O processo P1 ressalta que as rotinas automatizadas deve contemplar desde a extração inicial dos dados a partir do seu ambiente de produção até o local onde a base será disponibilizada como dados abertos. Por exemplo, se tiver sido decidido publicar os dados em arquivos csv, essa etapa contempla a obtenção dos dados, tratamento e hospedagem dos dados extraídos após conversão para o formato csv em um servidor de arquivos para a *Web* BRASIL (2014c). O processo P3 recomenda, que preferencialmente, a origem dos dados das rotinas ETL sejam sistemas de informações governamentais confiáveis e estruturados (COLOMBIA, 2012).

O processo P6 apresenta as diversas ferramentas para mineração e modelagem de dados utilizadas num experimento. Por se tratar de uso de dados geoespaciais, se fizeram necessários softwares de manipulação de ontologias, conversores de dados de bancos relacionais para servidores de triplas RDF e sistemas de informações geográficas (CONSOLI et al., 2014). Este processo, apesar de pouco detalhado, destaca-se pela utilização de dados geoespaciais, cuja complexidade para abertura e publicação é maior. Ademais, toda a rotina ETL deve ser exaustivamente testada antes de entrar em produção COMSODE (2014b).

4.2.6.6 Analisar se os dados serão conectados ou não (3F)

Considerando a relevância de se produzir dados conectados, o processo P14 sugere que devem ser estabelecidas atividades que visem identificar a conexão do conjunto de dados a ser aberto com outros conjuntos de dados (COMSODE, 2014b). Para desempenhar esta tarefa o processo P10 destaca que a modelagem de dados se compõe das atividades de identificação, modelagem, nomeação e teste. Na identificação deve-se obter uma cópia lógica e física do modelo do banco de dados, obter dicionários de dados e criar dados de forma que sejam replicáveis. Por último, deve ser observado no mundo real, objetos de interesse relacionados com os dados, como pessoas, locais, coisas e localidades (HYLAND; WOOD, 2011).

Segundo o P10, na modelagem em si, devem ser desenvolvidas as seguintes atividades:

- Esboçar ou desenhar os objetos em um quadro branco (ou similar) e desenhar linhas para expressar como eles estão relacionados uns com os outros;
- Investigar como os outros publicadores estão descrevendo dados semelhantes ou relacionados. Devem ser reutilizados vocabulários comuns para facilitar a fusão de dados e seu reúso; Devem ser analisadas ainda, ocorrências de duplicação de dados.

- Devem ser colocadas de lado às necessidades imediatas e/ou específicas de qualquer aplicação. O dado precisa ser útil para qualquer aplicação.
- Deve se utilizar finalmente, o bom senso para decidir o dado será ou não conectado a outro dado;
- Complementarmente, devem ser utilizadas URIs para referenciar os objetos de dados, devendo ser levado em consideração como os dados poderão ser atualizados ao longo do tempo. Por exemplo, devem ser evitados nas URIs, siglas ou nomes de instituições que podem sofrer alterações decorrentes de mudanças de natureza política ou de mudança governamental. Ademais, devem as hipóteses estabelecidas no esquema devem ser testadas com especialistas familiarizados com os dados.

Para uma maior familiarização com esta atividade, é recomendado que sejam modelados dois ou três objetos para se iniciar um rotina de abertura de dados com maior volume. Durante esta atividade, os especialistas envolvidos devem descobrir as relações e identificar como cada objeto se relaciona com o mundo real. Tal atividade pode ser apoiada com base em um grande quadro branco ou site wiki colaborativo. O processo P11 explana que foi adotada inicialmente a interconexão de dados de um mesmo domínio geográfico (neste caso um país, a Grécia) e após a conclusão deste domínio, foram buscados conjuntos de dados de outros países que poderiam ser interconectados (GALIOTOU; FRAGKOU, 2013). Nesta oportunidade, ontologias amplamente utilizadas como a DBPedia devem ser avaliadas como candidatas a se interconectar com o conjunto de dados a ser aberto.

4.2.6.7 Estabelecer ou aprimorar documentação de dados (esquemas, vocabulários e ontologias) (3G)

Para a melhoria da qualidade dos dados a serem publicados, durante a modelagem alguns processos recomendam o estabelecimento de uma boa documentação ou ainda, o aprimoramento da documentação existente, utilizando esquemas, vocabulários ou ainda, ontologias. Quanto houver a publicação de dados tabulares, por exemplo, deve ser realizada uma boa definição das colunas que contém as informações dos arquivos, definindo títulos compreensíveis pelos usuários, bem como quais são os tipos de dados aceitáveis em cada coluna de cada planilha. Em caso de publicação de arquivos XML ou JSON, definir a estrutura básica do documento. É possível, mas não necessário, estabelecer documentos XML Schema ou JSON Schema para validação. Para estas atividades, devem ser solicitados os esquemas e documentação das bases de dados, como modelos UML ou entidade-relacionamento, dicionário de dados, etc (BRASIL, 2014b; CHILE, 2013b; ECUADOR, 2014).

Para uma documentação mais aprimorada, após as etapas de identificação e seleção dos conjuntos de dados, é necessário estabelecer as ontologias que serão utilizadas para

modelar o domínio de uso destes conjuntos de dados. Nestes casos, orienta-se que seja reutilizado o maior número de vocabulários e outras ontologias, aproveitando o conhecimento produzido anteriormente. Esta abordagem baseada em reutilização acelera o desenvolvimento de ontologias, contribuindo para que os governos poupem tempo, esforço e recursos. O processo P13 descreve esta atividade nas seguintes tarefas (VILLAZÓN-TERRAZAS et al., 2011):

- Devem ser procurados os vocabulários mais adequados para serem reutilizados em repositórios populares como o *SchemaWeb*[2], *SchemaCache*[3], *Swoogle*[4], e LOV[5]. Para esta escolha, o P13 recomenda seguir um guia elaborado por (GÓMEZ-PÉREZ; SUÁREZ-FIGUEROA, 2009) que explica como reusar vocabulários de diferentes níveis de granularidade, como ontologias gerais, ontologias de domínio, bem como declarações de ontologias;

- No caso de não se encontrar nenhum vocabulário apropriado para o propósito, o vocabulário deve ser criado, tentando reusar recursos de dados existentes (como outros vocabulários) sempre que possível. Para esta atividade, o autor recomenda os guias propostos em Villazón-Terrazas, Suárez-Figueroa e Gómez-Pérez (2010) que demonstra como (1) procurar recursos de dados governamentais a partir de sites altamente confiáveis, sites relacionados com os domínios e catálogos de governo; (2) a seleção de recursos de dados governamentais mais adequadas; e (3) transformação destes vocabulários em ontologias;

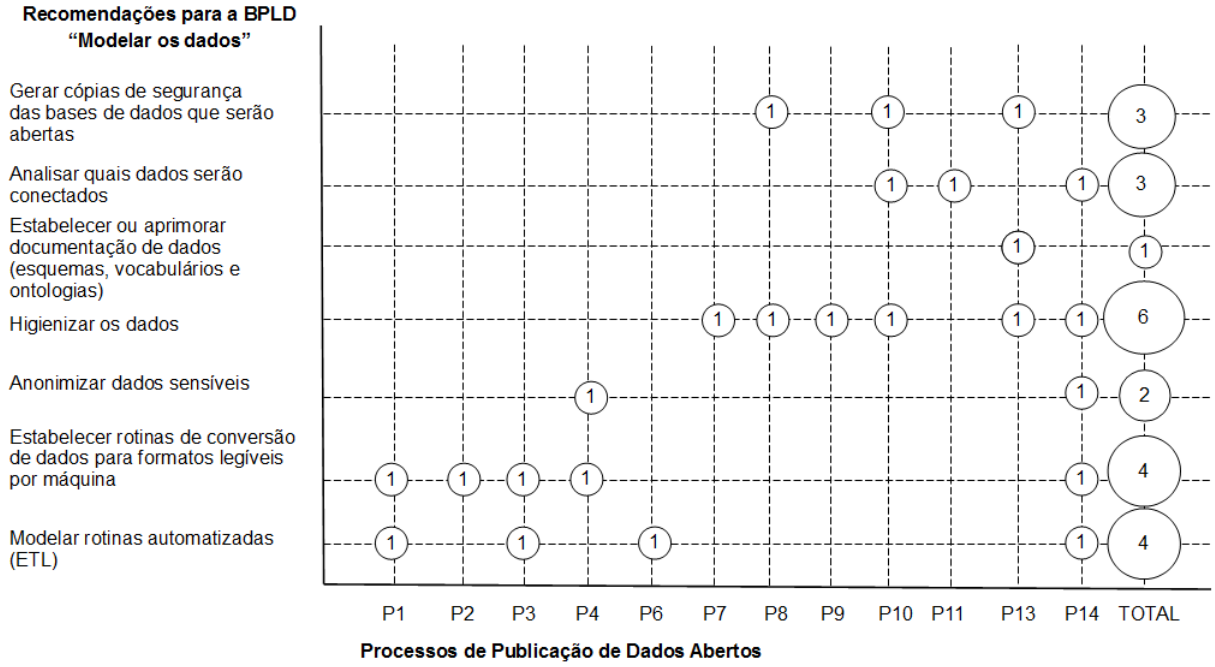
- Finalmente, se você não encontrou nenhum vocabulário ou recursos para construção de ontologias, você deve fazê-lo a partir do zero. Para este propósito, devem ser utilizadas metodologias para o desenvolvimento de ontologias, como a *NeOn Methodology* e ferramentas que fornecem suporte tecnológico para essa atividade como o *Protége*[6], NeOn Toolkit[7], *TopBraid Composer*[8] e *Altova SemanticWorks*[9].

4.2.6.8 Sumarização dos Resultados

Da análise desta BPLD, foram extraídas sete recomendações, conforme descrito nesta seção. Doze dos quinze processos contemplam esta BPLD, com a maioria das recomendações presentes em vários processos. A Figura 32 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram:

[2] Disponível em http://schemaweb.info
[3] Disponível em http://schemacache.com
[4] Disponível em http://swoogle.umbc.edu
[5] *Linked Open Vocabularies* - disponível em http://labs.mondeca.com/dataset/lov/index.html
[6] Disponível em http://protege.stanford.edu
[7] Disponível em http://www.neon-toolkit.org
[8] Disponível em http://www.topquadrant.com/products/TBComposer.html
[9] Disponível em http://www.altova.com/semanticworks.html

Figura 32 – Identificação de recomendações para a BPLD "Modelagem dos Dados" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Considerando a Figura 32, podemos notar que a preocupação com a *"higienização dos dados"*, e consequentemente com a qualidade do que será publicado foi identificado em 6 dos 12 processos que contemplaram esta BPLD. O *"estabelecimento de rotinas (procedimentos) para a conversão de dados para formatos legíveis por máquina"* e o *"estabelecimento de rotinas automatizadas"* foi considerada por 4 dos 12 processos. Por outro lado, atividades mais avançadas como o *"estabelecimento ou aprimoramento de documentações"* foi contemplado por apenas um processo. Ademais, as recomendações ficaram bem distribuídas entre os processos, não havendo, para esta BPLD, um processo que seja responsável pela grande maioria das recomendações extraídas.

4.2.7 RQ 3.4: Recomendações para "Especificar uma licença apropriada"

Segundo as BPLDs W3C (2014), após a modelagem dos dados, devem ser estabelecidas as licenças de uso. Segundo a OKF (2015b), na maioria das jurisdições existem direitos de propriedade intelectual em dados que impedem que o seu uso, reutilização e redistribuição terceiros sem permissão explícita. Mesmo em locais onde a existência de direitos é incerta, é importante aplicar uma licença para estabelecer clareza sobre suas condições de uso. Assim, se é planejada a disponibilização de dados, deve ser adotado uma licença sobre eles.

Uma licença de uso de dados estabelece as condições de uso do dado, por exemplo, se o mesmo pode ser utilizado para fins comerciais, se pode ser alterado, remixado com outros

dados ou se está disponível somente para leitura. O W3C ressalta a importância de que todos os dados disponíveis na *Web* possam ter sua licença de forma explicita, possibilitando que o uso dos mesmos seja feito conforme suas condições de uso estabelecidas.

Quanto à publicação de dados conectados, são recomendados a utilização de licenças abertas, como a *Creative Commons*[10], que estimulem o uso e o reúso dos dados. Destaca-se que os usuários ficarão mais motivados a utilizar tais dados quando suas licenças estiverem claramente definidas.

O estabelecimento de licenças foi ressaltado em oito processos analisados (CHILE, 2013b; ECUADOR, 2014; URUGUAY, 2012; BAUER; KALTENBÖCK, 2012; HYLAND; WOOD, 2011; VILLAZÓN-TERRAZAS et al., 2011; COMSODE, 2014b; OKF, 2015a). Os processos P4 e P15 ressaltam que a adoção de uma licença é essencial para que um dado publicado seja considerado aberto. O processo P9 ainda ressalta que a adoção de licenças promove segurança jurídica, habilitando a interoperabilidade com outros conjuntos de dados.

A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.7.1 Adotar Licenças Não restritivas (4A)

Para os dados abertos (conectados), as licenças a serem adotadas não podem estabelecer restrições de uso, pois caso o faça estará violando as três leis dos dados abertos (EAVES, 2009) e seus oito princípios (OGD, 2007).

Neste contexto, é recomendável que sejam atribuídas licenças não restritivas e atribuitivas(reconheçam a autoria da fonte original dos dados), como a CC-BY[11] (Creative Commons – BY) e por outro lado, desencoraja explicitamente a adoção de licenças que possuam algum tipo de restrição como a CC-BY-NC[12], CC-BY-ND[13] e CC-BY-SA[14], por entender que tais licenças desencorajam o reúso e o desenvolvimento de trabalhos derivados a partir destes dados. O processo P14 pondera ainda que o órgão publicador deve buscar a autorização legal para licenciar tais dados e caso não consiga, deve informar claramente os motivos pelo qual não pode estabelecer uma licença ao conjunto de dados COMSODE (2014b).

Novamente é reforçado a importância de se consultar a legislação vigente sobre licenças de uso, onde na experiência do processo P14 foram utilizadas como referência, algumas definições estabelecidas pela legislação Checa para a aplicação de proteções legais ou licenças restritivas. O processo apresenta ainda alguns casos onde o setor público, em

[10] Disponível em https://creativecommons.org
[11] Disponível em https://creativecommons.org/licenses/by/4.0/
[12] Disponível em https://creativecommons.org/licenses/by-nc/4.0/
[13] Disponível em https://creativecommons.org/licenses/by-nd/4.0/
[14] Disponível em https://creativecommons.org/licenses/by-sa/4.0/

especial o órgão publicador não teria autoridade para licenciar um conjunto de dados. São eles:

- Exista uma restrição explicitamente determinada pela legislação vigente de proteção de dados (como a ocorrência de dados pessoais ou dados que afetem a segurança nacional)
- Exista a obrigação de se manter tais informações como confidenciais;
- Envolvam direitos de terceiros (por exemplo, dos direitos dos empregados contidos em contratos de trabalho; restrições contratuais na produção do conjunto de dados que impeçam a distribuição ou reúso).

Além disso, é recomendado que o setor público não busque exercer direitos econômicos sobre os conjuntos de dados (ou de alguma parte dos mesmos), pois a existência de taxas ou *royalties* sobre o consumo dos dados provavelmente desencorajará o seu amplo uso.

4.2.7.2 Estabelecer questões-chave para definição de licenças (4B)

Para orientar a adoção de licenças, o processo P14 recomenda que haja um esforço para se responder três perguntas relacionadas aos direitos de uso do conjunto de dados, que são (COMSODE, 2014a):

- O respectivo conjunto de dados (ou alguns dos seus itens) é protegido por algum direito *sui generis* de banco de dados ou copyright, ou seja, ele pode ser considerado um trabalho ou um banco de dados, conforme definido pela legislação vigente?
- O órgão publicador tem autoridade para estabelecer licenças para o referido conjunto de dados, ou por ventura existem limitações para o publicador aplique licenças sobre um conjunto de dados?
- Qual licença específica deve ser escolhida para o referido conjunto de dados?
  - Nota: A fim de esclarecer algumas questões e/ou problemas particulares, o processo P14 utiliza termos e disposições contidas na Lei Checa de Copyright ("CCA"). Logo, a legislação vigente sobre o assunto da localidade da organização publicadora deve ser consultada.

Considerando o entendimento deste processo, com base nas respostas para as questões acima é então possível fazer uma conclusão sobre im/possibilidade de licenciar o referido conjunto de dados.

4.2.7.3 Apresentar opções de licenças a serem adotadas (4C)

Ademais, para que haja o estabelecimento de uma licença de uso, devem ser apresentadas diversas opções para que o órgão publicador adote a que for mais conveniente ao conjunto de dados a ser publicado. Esta investigação extraiu não apenas os argumentos para que o publicador estabeleça uma licença, mas também um conjunto de licenças que podem ser adotadas conforme sugerido nos processos P2, P4, P9 e P13. Busca-se distinguir a existência de licenças voltadas para conteúdos e informações, como a Creative Commons, das licenças orientadas para bases de dados, como a *Open Database Licence* (ODbL). Dentre as licenças recomendadas pelos processos analisados encontram-se:

- *Creative Commons*: Esta licença permite aos usuários dos conteúdos que possam distribuir, remixar, corrigir e construir novos conteúdos a partir de um conteúdo original, mesmo que para fins comerciais, desde que as fontes originais recebam o devido crédito.
- *UK Government License*[15]*:* Criada para permitir que qualquer detentor informação do setor público britânico para tornar a sua informação disponível para o uso e reúso sob seus termos.
- *Open Database Licence - ODbL*[16]: licença aberta para bancos de dados e dados que inclui atribuição explícita e requisitos *share-alike*.
- Public Domain Dedication and License (PDDL)[17]: é um documento que destina permitir que você um determinado dado seja compartilhado livremente, modificado e utilizado para qualquer finalidade e sem quaisquer restrições.
- Open Data Commons Attribution License[18]: Licença específica para bancos de dados que estabelece algumas atribuições para bancos de dados.

O processo P5 (URUGUAY, 2012, p. 13) também destaca a necessidade de adoção de uma licença e apresenta durante as atividades de publicação no seu catálogo nacional de dados abertos a *"Licencia de datos abiertos de gobierno de Uruguay"*, recomendada pela agência de Governo Eletrônico uruguaia. Entretanto, outras licenças abertas podem ser adotadas caso a instituição publicadora tenha requisitos específicos no dado a ser publicado que não sejam cobertas pela licença padrão.

4.2.7.4 Sumarização dos Resultados

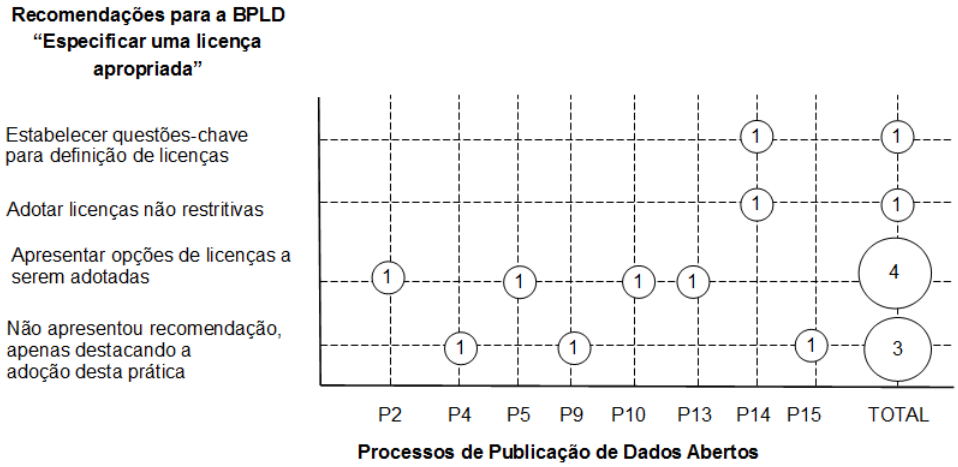
Da análise desta BPLD, foram extraídas apenas três recomendações, conforme descrito nesta seção. Oito dos quinze processos analisados estabelecem esta BPLD como relevante

[15] Disponível em http://www.nationalarchives.gov.uk/doc/open-government-licence/
[16] Disponível em http://opendatacommons.org/licenses/odbl/
[17] Disponível em http://opendatacommons.org/licenses/pddl/
[18] Disponível em http://opendatacommons.org/licenses/by/

para a abertura e publicação de dados. Foi possível extrair recomendações relevantes de cinco processos, pois os processos P4, P9 e P15 apenas citaram a importância desta prática, sem apresentar uma recomendação específica.

A Figura 33 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram.

Figura 33 – Identificação de recomendações para a BPLD "Especificar uma licença apropriada" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Pela Figura 33 observa-se que esta BPLD é pouco explorada pelos processos analisados, sendo possível à extração de apenas três recomendações. Apenas no processo P14 foi possível se identificar recomendações mais avançadas, como o estabelecimento de questões-chave para a definição de licenças. Por outro lado, a pouca exploração desta BPLD pode sugerir o desenvolvimento de novas pesquisas que aprimorem o estabelecimento de licenças para dados abertos e dados abertos conectados.

4.2.8 RQ 3.5: Recomendações para "Estabelecer bons identificadores universais (URIs)"

A oferta de dados abertos (conectados ou não) é provida através de páginas, sítios ou catálogos *Web* e por esta natureza, a definição dos endereços/identificadores eletrônicos (URIs) de acesso aos dados consiste de etapa muito relevante num processo de publicação, pois será através destes endereços que os dados serão encontrados pelos usuários.

A definição de identificadores de conjuntos de dados é uma forma de representar entidades do mundo real em arquivos digitais, seja mediante a inserção de linhas em tabelas, elementos em documentos XML ou objetos em documentos JSON. É importante saber o que o identificador de cada entidade representa. Por exemplo, em uma tabela, é necessário identificar quais colunas formam os das entidades representadas nesta tabela (COMSODE, 2014b).

Identificadores são muito importantes, especialmente para desenvolvedores de softwares. Eles os utilizam para identificar recursos de dados em seu código-fonte e para fundir informações sobre as entidades de diferentes fontes de dados (COMSODE, 2014b). Logo, esta etapa do processo é crucial e precisa ser desenvolvida considerando as melhores práticas disponíveis. Para a definição das estruturas dos catálogos (esquemas), o P14 recomenda a utilização do vocabulário DCAT[19], voltado a facilitar a interoperabilidade entre catálogos de dados publicados na *Web*. Villazón-Terrazas et al. (2011) complementa que o objetivo dos dados conectados é promover uma visão da *Web* como um banco de dados global, e a interligação dos dados da mesma forma que os documentos da *Web*. Neste banco de dados global é necessário para identificar os recursos de dados na Internet, e precisamente os URIs são estabelecidos para cumprir esta função.

Dos processos analisados, apenas três apresentam de forma explícita recomendações para o estabelecimento de URIs para publicação de dados. Tais recomendações serão analisadas em complementação ao que é recomendado por W3C (2014) e Wood et al. (2013). Galiotou e Fragkou (2013) destaca a importância do estabelecimento de boas URIs, sem apresentar uma recomendação específica.

A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.8.1 Utilizar URIs para conectar os dados (5A)

Ademais, outra recomendação muito importante e óbvia quando da existência de dados abertos conectados consiste na utilização de URIs para conectar tais dados. A inserção de links para outros URIs representa o quarto princípio dos dados conectados, de modo seja possível descobrir novos dados a partir de um determinado recurso ou conjunto de dados, seja no mesmo órgão publicador ou em conjuntos ou recursos de dados de outras organizações . Analogamente, esta conexão pode ser entendido como um URL de uma página *Web* que também é usado por outras páginas da *Web* para estabelecer um hiperlink para esta página (VILLAZÓN-TERRAZAS et al., 2011).

4.2.8.2 Estabelecer URIS persistentes, que não se alterem em nenhum momento (5B)

Uma URI, por ser um identificador universal, que potencialmente se conectará a diversos outros recursos de dados, não deve mudar. Por esta razão, este princípio recomenda que as URIs não contenham absolutamente nada que possam ser passíveis de mudanças, ou seja, as URIs precisam ser persistentes e estáveis. Hyland e Wood (2011), Wood et al. (2013), W3C (2014) sugerem ainda que devem ser utilizados domínios que a instituição publicadora tenha controle, de tal maneira que se elimine o risco de mudanças de domínio causadas por atores externos ao controle da instituição publicadora. O processo P14 re-

[19] Disponível em http://www.w3.org/TR/vocab-dcat/

força que uma boa URI não deve mudar durante todo o ciclo de vida do recurso de dados (COMSODE, 2014b).

4.2.8.3 Proporcionar pelo menos um recurso de dados em formato que seja legível por máquina para cada URI (5C)

Esta recomendação reforça outras recomendações anteriores de que sempre deve ser provido uma URI para recurso de dados legível por máquina para cada URI que resolva o recurso de dados legível para humanos (WOOD et al., 2013).

4.2.8.4 Usar URIs como nomes para as coisas (5D)

URIs são identificadores, mas também podem cumprir a função de nomear recursos, pois uma URI bem estabelecida ajudará o usuário tanto na localização quanto no reconhecimento do recurso de dado (W3C). Considerando que os cidadãos serão grandes usuários destes recursos, deve ser associada a URI o máximo de informações possíveis facilitando o entendimento do usuário sobre o que ele irá encontrar ao acessar este identificador (VILLAZÓN-TERRAZAS et al., 2011).

4.2.8.5 Estabelecer design simplificado de URIs (5E)

Considerando que os URIs são identificadores universais, devem ser de fácil compreensão para o usuário. Desta maneira, as URIs devem ser estabelecidas com simplicidade, estabilidade e capacidade de gerenciamento. Cumpre destacar que os URIs devem ser estabelecidos como identificadores e não apenas para nomear recursos da *Web* (VILLAZÓN-TERRAZAS et al., 2011). Hyland e Wood (2011) apresentam recomendações complementares, onde sugere a utilização de URIs limpas (de fácil entendimento).

4.2.8.6 Utilizar identificadores relacionados a informações do mundo real (5F)

Um identificador não deve ser um valor sem sentido artificialmente gerado para ser armazenado como uma mera chave primária no banco de dados. Deve ser um valor que será utilizado para compartilhar informações sobre a entidade no mundo real e pelos sistemas reais. Por outro lado, devem ser evitados, no estabelecimento de URIs, a geração de números ou chaves aleatórias, que não apresentem uma lógica de entendimento para o usuário.

Por exemplo, as organizações empresariais em um determinado país podem ser identificadas por um número de identificação exclusiva destas organizações (como o número do Cadastro Nacional de Pessoas Jurídicas (CNPJ) do Brasil). Este número é usado por diferentes autoridades públicas e os seus sistemas de informação para identificar esta entidade, neste caso à organização empresarial (COMSODE, 2014b).

4.2.8.7 Usar URIs HTTP para que recursos de dados possam ser encontrados via *Web* por pessoas e máquinas (5G)

Complementarmente a recomendação anterior, as URIs devem ser estabelecidas mediante uma estrutura lógica, que seja compreensível, ora por humanos, ora por máquinas com o objetivo que os recursos de dados possam ser encontrados na *Web* por ambos os tipos de usuários (HYLAND; WOOD, 2011; WOOD et al., 2013; W3C, 2014). Por exemplo, URIs que sejam formados por códigos identificadores numéricos podem ser facilmente entendidas por máquinas, mas dificilmente serão memorizadas por humanos, devendo ser utilizadas palavras-chave relacionadas ao mundo real como na recomendação anterior (HYLAND; WOOD, 2011).

No caso de dados abertos conectados, é necessário garantir que as URIs de entidades (conjuntos ou recursos de dados) sejam *dereferenciadas*. Isso significa que, se um cliente resolve uma URI de uma entidade, deve receber uma notação RDF legível por máquina desta entidade, em formatos como o *Turtle* ou JSON-LD mediante o seguinte detalhamento (COMSODE, 2014b)

- No acesso a URI do catálogo, o servidor retorna os metadados sobre o catálogo;
- No acesso a URI de um conjunto de dados, o servidor retorna o registro do conjunto de dados e os metadados sobre as distribuições (recursos) contidas no conjunto de dados;
- No acesso a URI de um recurso de dados, o servidor retorna os dados e os metadados sobre o respectivo recurso.

4.2.8.8 Estabelecer URIs neutras (5H)

Uma URI podem conter significados, estabelecidos por chaves naturais ou derivadas de sistemas. Todavia, deve-se ter o entendimento de que ao se estabelecer uma URI, ela deve existir para sempre, e por esta razão, não devem ser incluídos elementos passíveis de mudança futura, como números de versão ou siglas de tecnologias (Ex: Se uma URI de uma página desenvolvida na tecnologia ASP termina com .asp, se a tecnologia mudar para PHP é possível que a URI passe a terminar com .php). Além disso, URIs neutras também contribuem para a segurança do domínio e dos dados publicados, por não expor detalhes que possam comprometer a disponibilidade do sítio (W3C, 2014).

4.2.8.9 Utilizar datas em URIs com moderação (5I)

Datas devem ser utilizadas com moderação: O uso de datas em URIs deve ser utilizado apenas para casos onde os dados mudam ao longo do tempo e se faz necessário guardar e disponibilizar o seu histórico. Tal utilização é comum para publicação de dados estatísticos, regulamentos, especificações, dentre outros documentos que tenham atualização

periódica (mensal, trimestral, anual). O uso de datas deve ser utilizado apenas quando for realmente necessário, tendo uma justificativa plausível para seu uso (W3C, 2014).

4.2.8.10 Utilizar hashs (#) em URIs cautelosamente (5J)

Uma recomendação específica consiste no uso de hashs (#) em URIs pois, apesar de serem muito utilizados na *Web* para mapeamento de elementos de conteúdo, as hashs(#) não são enviadas para o servidor, ficando limitadas ao lado cliente. Desta maneira, a adoção de hashs não garante que o conteúdo identificado pela *hash* será devidamente processado por máquinas, podendo limitar a sua interpretação apenas para humanos (HYLAND; WOOD, 2011). Villazón-Terrazas et al. (2011) sugere a utilização de barras (*slashs*) sempre que possível em substituições ao *hash*.

4.2.8.11 URIs das entidades (conjuntos de dados ou recursos) sejam diferentes das URIs das páginas que apresentam estes recursos para a leitura feita por humanos (5K)

Foi extraída a recomendação que haja uma diferenciação entre as URIs com conteúdos acessíveis por humanos e por máquinas. Exemplificando: A URI (ou URL) de uma página que descreve um conjunto de dados (contendo seu título, descrição, fonte) é diferente da URI do conjunto de dados em si (o arquivo que contém os dados). O processo P14 (COMSODE, 2014b) apresenta ainda uma estrutura padronizada para estabelecimento de URIs para catálogos e respectivos dados e ainda recomenda que cada órgão publicador tenha seu próprio catálogo armazenando seus dados, pois desta maneira, terá maior controle sobre a definição de URIs adequadas, conforme as práticas apresentadas. É ressaltado ainda que para a publicação de dados de 1-3*, as convenções para estabelecimento de URIs são desejáveis. Entretanto, para publicação de dados 4 ou 5*, tais condições são obrigatórias. Ademais, uma URI sempre deve retornar informação útil e quando se tratar de dados abertos conectados, este retorno deverá ser utilizando os padrões RDF ou SPARQL (WOOD et al., 2013).

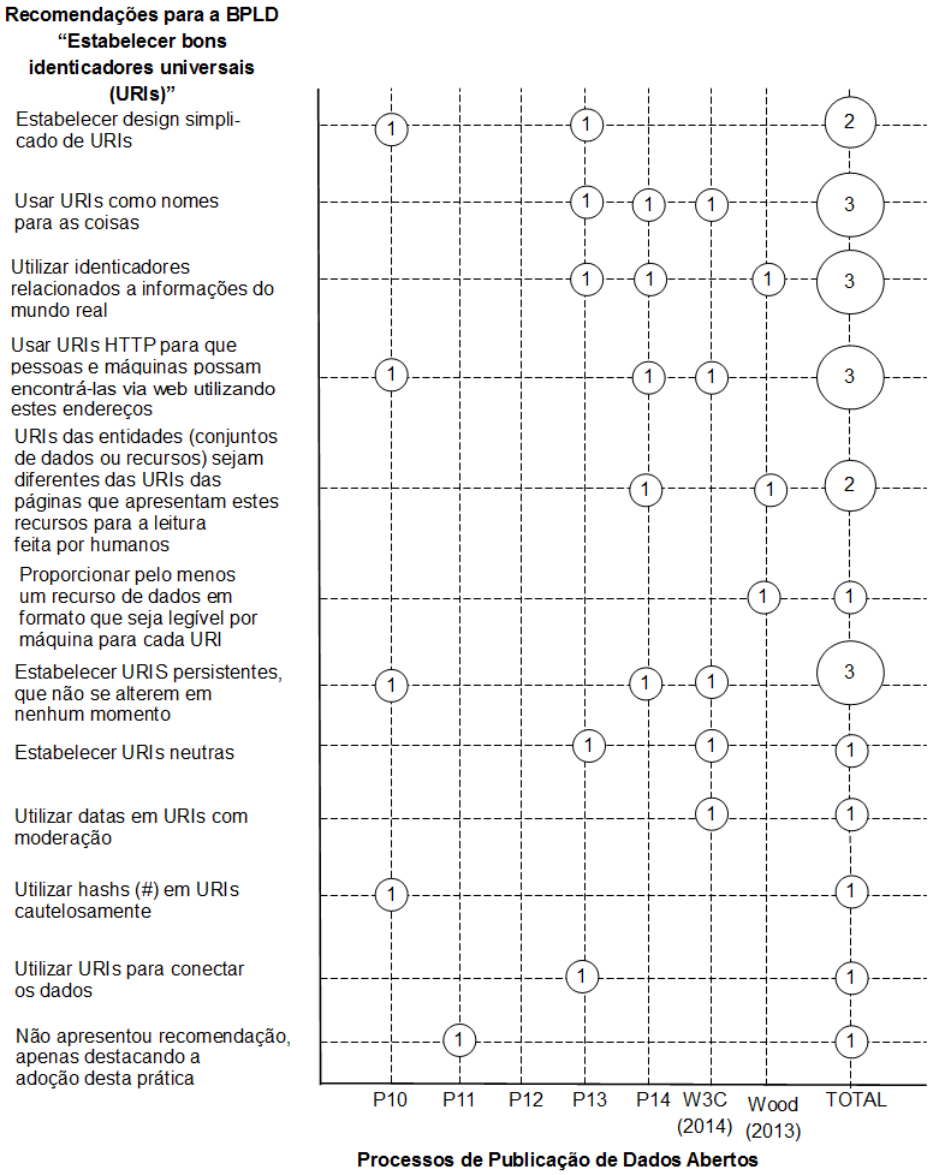
4.2.8.12 Sumarização dos resultados

Da análise desta BPLD, foram extraídas onze recomendações, conforme descrito nesta seção. Cinco dos quinze processos analisados estabelecem esta BPLD como relevante para a abertura, publicação e conexão de dados. Como esta BPLD é obrigatória para a produção de dados conectados, a investigação buscou extrair recomendações complementares estabelecidas pelo W3C e pelo livro "Linked Data", de David Wood.

Quatro processos apresentaram recomendações relevantes e apenas um citou a importância desta prática, sem apresentar uma recomendação específica (GALIOTOU; FRAGKOU, 2013). Cumpre destacar os processos desenvolvidos pelos governos da América do Sul não estabelecem recomendações para o desenvolvimento de URIs.

A Figura 34 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram:

Figura 34 – Identificação de recomendações para a BPLD "Estabelecer bons identificadores universais (URIs)" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Considerando a Figura 34, as recomendações encontram-se bem distribuídas entre os processos analisados, não sendo possível identificar recomendações mais ou menos frequentes. As recomendações estabelecidas por W3C (2014) e Wood et al. (2013) complementam ou orientam recomendações existentes nos processos. A única exceção consiste na recomendação de "Proporcionar pelo menos um recurso de dados em formato que seja legível por máquina para cada URI" proposta por Wood et al. (2013) e não identificada em nenhum dos processos analisados.

4.2.9 RQ 3.6: Recomendações para "Utilizar vocabulários padrão"

O uso de vocabulários consiste de elemento fundamental para a *Web* dos dados. Segundo o W3C, vocabulários devem ser reutilizados o máximo que for possível para facilitar a expansão e a inclusão de novos conjuntos de dados na *Web* dos dados. O W3C complementa que órgãos governamentais publicadores de dados são encorajados a utilizar vocabulários estabelecidos (padrões) evitando a "reinvenção da roda" e facilitando o entendimento destes dados, considerando a popularidade dos vocabulários.

Quanto aos processos analisados, apenas cinco abordam a importância do estabelecimento de esquemas de dados, vocabulários ou temas correlatos, como os metadados, onde **alguns processos apenas sugerem a adoção de metadados para descrição dos dados. Outros processos enfatizam a necessidade da adoção de vocabulários e, em nível mais avançado, de ontologias.** A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.9.1 Estabelecer metadados obrigatórios (6A)

Os processos P4 e P5 explanam que os metadados são campos que descrevem os dados, oferecendo meios suficientes para que localizar, recuperar, processar e descrever os dados ECUADOR (2014), URUGUAY (2012). O processo P3 apresenta cinco objetivos para a geração de metadados que são: (i) interoperabilidade com outros conjuntos de metadados, (ii) extensibilidade; (iii) modularidade; (iv) utilidade e; (v) simplicidade (COLOMBIA, 2012).

Os conjuntos de dados podem demandar metadados específicos que descrevam características particulares deste conjunto de dados (ECUADOR, 2014; COLOMBIA, 2012). Durante a produção dos metadados devem ser incluídos os formatos de arquivo em que serão armazenados os dados, para que o usuário tenha referência de quais softwares necessitará para manipular tais arquivos (COLOMBIA, 2012).

Buscando um maior esclarecimento sobre a adoção dos metadados, o processo P5 exemplifica que, para o caso de publicação de dados estatísticos, os metadados podem ser a descrição das colunas, ou para arquivos XML, os metadados pode ser o seu XML esquema (XSD)[20]. Os processos P3, P4 e P5 apresentam tabelas que descrevem os metadados obrigatórios para cada conjunto de dados e estabelecem campos que devem ser adotados para conjuntos de dados (ECUADOR, 2014; COLOMBIA, 2012; URUGUAY, 2012): (i) Título do dado; (ii) descrição; (iii) Tipo de dado (estatístico, geoespacial, descritivo, etc.); (iv) Recursos de dados relacionados (hiperlink para os recursos de dados); (v) Versão do dado; (vi) Fonte do dado; (vii) Idioma; (ix) Palavras-Chave; (x) Licenças; (xi) Nome e e-mail de contato do responsável pelo dado; (xii) URI; (xiii) Informações adicionais.

[20] Disponível em http://www.w3.org/TR/xmlschema11-1/

4.2.9.2 Criar um esquema de dados para cada conjunto de dados (6B)

Outra recomendação extraída consiste na criação de um esquema de dados para cada conjunto de dados e respectivos recursos de dados (e formatos) a serem publicados, tentando explicar o conteúdo a ser ofertado em forma de descrições e comentários e preferencialmente, de modo semântico. Caso haja um catálogo de dados da organização que forneça uma página *Web* para cada conjunto de dados, deve ser inserido um link para o esquema de dados relacionado a este conjunto de dados na página *Web* do catálogo conforme sugerido pelo processo P14 (COMSODE, 2014b).

4.2.9.3 Incentivar o reúso de vocabulários (6C)

Considerando as recomendações para utilização de vocabulários, é destacado a importância da escolha dos vocabulários, onde existem inúmeros para reúso, sendo necessário uma avaliação para identificar os vocabulários mais apropriados para a sua necessidade, sugerido pelo processo P9 (BAUER; KALTENBÖCK, 2012). Novos vocabulários devem ser criados apenas se não houver algum que atenda a demanda do publicador de dados. Vocabulários devem ser reusados sempre que possível, considerando que já existe uma quantidade de vocabulários que são utilizados rotineiramente para descrever pessoas, locais, coisas e localidades e que dados conectados precisam ser descritos e detalhados a partir de vocabulários existentes e preferencialmente, que sejam largamente utilizados. Como exemplo, podem ser reutilizados vocabulários como o *"Dublin Core"* (DC), que descreve os metadados sobre trabalhos publicados, ou o *"Friend-of-a-Friend"* (FOAF), usados para descrever as pessoas e suas relações com outras pessoas, ou ainda o GeoNames, uma base de dados geográfica abrange todos os países e contém mais de dez milhões de nomes geográficos, dentre outros vocabulários padrões existentes. Cumpre destacar que na comunidade de dados conectados, o reúso de vocabulários é algo presumido e que através do uso de URIs e de vocabulários que os curadores de dados e editores são capazes de publicar informações de forma mais rápida e reduzir os custos de integração de dados.

Com base na importância do reúso de vocabulários, o W3C (2014) recomenda algumas ferramentas para coletar, analisar e indexá-los como *Falcons*[21]*, Watson*[22] *e Swoogle*[23]. O W3C (2014) complementa que outro meio eficaz de encontrar bons vocabulários consiste na verificação de termos existentes em dados publicados em catálogos, preferencialmente os que contenham dados com temáticas relacionados ao do objeto de publicação.

4.2.9.4 Publicar esquemas de dados em arquivos diferentes (6D)

O processo P14 sugere que os esquemas de dados devem ser publicados em arquivos diferentes dos recursos de dados, mas devem ser conectados com os arquivos de dados

[21] Disponível em http://ws.nju.edu.cn/falcons/objectsearch/index.jsp
[22] Disponível em http://watson.kmi.open.ac.uk/WatsonWUI/
[23] Disponível em http://swoogle.umbc.edu

respectivos. Para estabelecer as conexões, utilizar os recursos disponíveis, seja uma URL ou URIs contidas em vocabulários ou ontologias quando cabível COMSODE (2014b).

4.2.9.5 Determinar linguagens para expressar esquemas de dados (6E)

O processo P14 estabelece uma recomendação para que sejam definidos os esquemas de dados, quando se tratar de dados nível 3 ou 4 estrelas, e que sejam definidos vocabulários e ontologias, para dados de nível 5 estrelas. Para estes casos, deve ser estabelecida a linguagem adequada para expressar os esquemas de dados, onde o processo orienta que sejam adotadas as seguintes premissas:

- Para o formato CSV, deve ser escolhido o *Metadata Vocabulary for Tabular Data*[24], estabelecido pelo W3C.
- Para o formato XML, escolher DTD ou *XML Schema*[25]
- Para o formato JSON, escolha *JSON Schema*[26]

4.2.9.6 Estabelecer critérios de escolha de vocabulários (6F)

Quanto aos critérios de escolha de vocabulários, o W3C (2014) os estabelece com maior clareza e detalhamento do que todos os processos analisados. São recomendadas as seguintes medidas para a seleção de bons vocabulários, certificando-se que:

- Devem ser documentados, contendo comentários que expliquem a sua estruturação, bem como as respectivas palavras-chave relacionadas com o tema. O publicador deve ainda providenciar páginas legíveis por humanos que descrevam o vocabulário e sua estrutura de classes e propriedades;
- Devem ser auto descritíveis, ou seja, cada propriedade ou termo em um vocabulário deve ter um título, descrição e comentários adicionais, permitindo a maior clareza possível quanto ao seu entendimento e uso;
- Devem ser descritos em mais de uma linguagem, pois isto contribui para seu uso e reúso de forma universal. Recomenda-se que todos os títulos, definições e comentários dos termos e propriedades sejam disponibilizados no mínimo em inglês, espanhol e no idioma oficial da entidade publicadora.
- Devem ser publicados por um grupo ou organização confiável, para que seus utilizadores tenham maiores garantias quanto a sua consistência, disponibilidade e atualização.

[24] Disponível em http://w3c.github.io/csvw/metadata/
[25] Disponível em http://www.w3.org/TR/xml/
[26] Disponível em http://tools.ietf.org/html/draft-zyp-json-schema?03

- Devem ser utilizados por outros conjuntos de dados e vocabulários, pois isto auxilia fortemente para que os dados se conectem através de vocabulários comuns. Por exemplo, o vocabulário FOAF é reutilizado por mais de 55 outros vocabulários. O processo P9 (BAUER; KALTENBÖCK, 2012) complementa que esta medida simplificará o entendimento comum deste vocabulário devido ao fato do mesmo já ser utilizado amplamente e consequentemente, conhecido por mais pessoas que contribuirão para um maior reúso do mesmo.

- Devem ser acessíveis por um longo período, tendo a garantia que estará sempre acessível por um longo período, idealmente para sempre. Este requisito segue o mesmo entendimento de disponibilização permanente já explorado para o estabelecimento de URIs, tendo URLs persistentes e política de controle de versão.

No caso da necessidade de se criar um novo vocabulário, é recomendado que além da adoção das medidas acima, devem ainda ser definidas uma boa URI para o vocabulário, atendendo as recomendações explanadas na seção respectiva e ainda, estabelecer URIs para as propriedades com sentidos verbais, de modo que possam facilitar o entendimento de triplas (Sujeito-Objeto-Predicado). Ex: temPropriedade

4.2.9.7 Certificar que os dados estão conectados a outros conjuntos de dados (6G)

Ademais, o processo P9 enfatiza a necessidade que, antes de se publicar os dados, deve se certificar que tais dados estão conectados a outros conjuntos de dados, contendo links para outros conjuntos de dados da mesma organização publicadora e conjuntos de dados de terceiros que tenham temática relacionada Hyland e Wood (2011). O processo P11 complementa este entendimento, descrevendo que na experiência do artigo, foram analisadas algumas ontologias que poderiam se conectar a ontologia utilizada (e-GIF), sendo a DBPedia como um repositório natural a ser consultado nestas ocasiões Galiotou e Fragkou (2013).

4.2.9.8 Desenvolver ou utilizar ontologias para estruturar a semântica dos dados (6H)

Em ambientes com maior maturidade na utilização de vocabulários, devem ser adotadas a estruturação ou reúso de ontologias visando ampliar a semântica dos dados a serem conectados com o suporte destes vocabulários. O processo P14 apresenta um conjunto de recomendações detalhadas para a criação e/ou reúso de novos vocabulários e ontologias, quando for o caso (COMSODE, 2014b). O reúso de uma ontologia foi descrito no processo P6 que apresenta na sua experiência que uma ontologia foi desenvolvida a partir da estrutura de tabelas do sistema de informação que serviu de base para a produção de dados conectados geoespaciais. Neste caso, as tabelas foram convertidas numa ontologia computacional (arquivo .owl), onde cada tabela foi representada como uma classe e cada

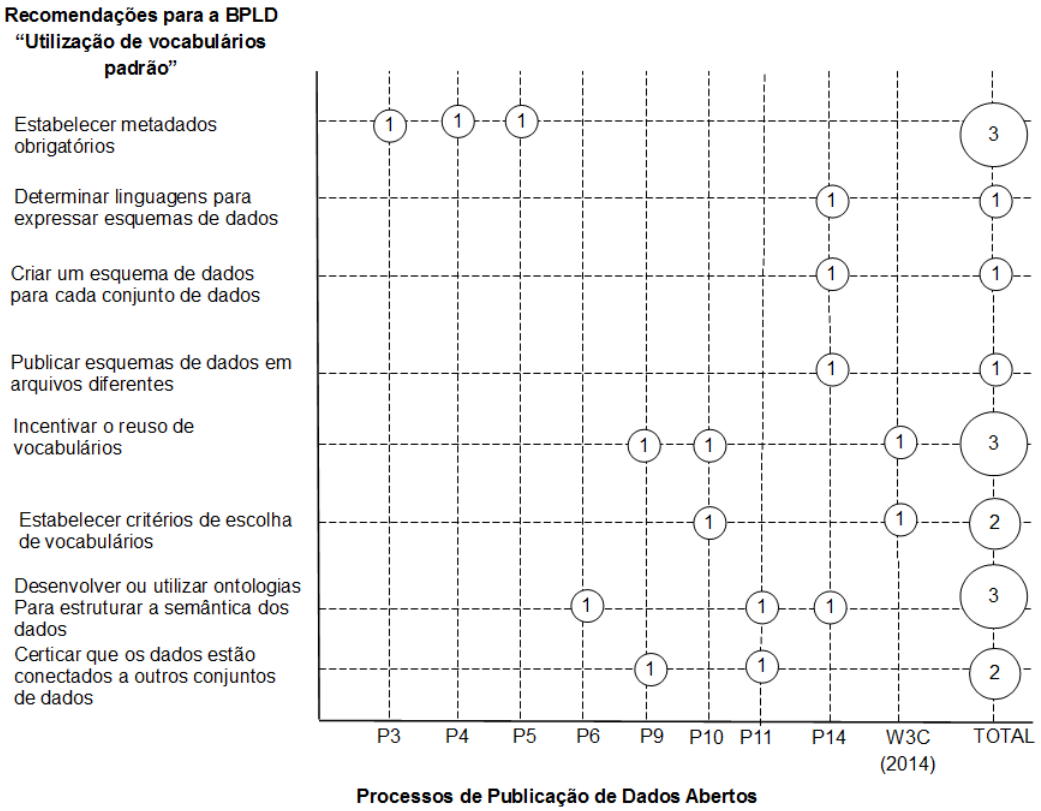
campo da tabela foi convertido para uma propriedade de dados. Ao final, o vocabulário ficou contido nesta ontologia sendo utilizada para o enriquecimento e publicação de dados conectados (CONSOLI et al., 2014).

4.2.9.9 Sumarização dos resultados

Da análise desta BPLD foram extraídas oito recomendações, conforme descrito nesta seção. Oito dos quinze processos analisados estabelecem esta BPLD como relevante para a abertura, publicação e conexão de dados. Como esta BPLD é obrigatória para a produção de dados conectados, a investigação buscou extrair recomendações complementares estabelecidas pelo W3C. Para esta BPLD, foi possível se extrair recomendações de todos os processos identificados.

A Figura 35 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram.

Figura 35 – Identificação de recomendações para a BPLD "Utilização de vocabulários padrão" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Como observado na Figura 35, as recomendações encontram-se bem distribuídas entre os processos analisados, não sendo possível identificar recomendações mais ou menos frequentes. Ademais, verifica-se que o processo P14 contemplou a maioria das recomen-

dações extraídas, podendo este processo ser mais bem analisado pelos órgãos publicadores quando houver a necessidade de estabelecer vocabulários.

4.2.10 RQ 3.7: Recomendações para "Converter e enriquecer dados"

Considerando as BPLDs todas as etapas até então envolveram atividades de planejamento e organização de meios para que a produção de dados conectados seja efetivada com êxito, contemplando desde a etapa de preparação de partes interessadas, modelagem de dados, definição de vocabulários e URIs, dentre outros. A partir desta etapa, são executadas tarefas que vão do enriquecimento de dados, ou seja, do melhoramento de dados existentes, passando pela geração de dados conectados até a sua disponibilização sustentável ao público. Cumpre destacar que, apesar desta investigação estar adotando como referencial comparativo as "Melhores Práticas para Publicação de Dados Conectados", também estão sendo analisados nos processos, recomendações para a publicação de dados não conectados, mas que que sejam correlatas com as recomendações para dados conectados. Por esta razão, as recomendações desta seção buscam ir além do que o W3C recomenda especificamente para dados conectados.

Diversos processos analisados implementam atividades de conversão e enriquecimento de dados de várias formas. MendonÇa et al. (2013) desenvolve esta atividade na etapa de conformidade, onde os dados são transformados para atender a finalidade desejada, onde devem ser tratados casos de duplicação de dados quando são necessários mesclar dados de várias fontes, especialmente as que compartilham vocabulários comuns.

A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.10.1 Converter dados para múltiplas finalidades e usos (7A)

Os processos também recomendam que os dados devem ser convertidos para múltiplas finalidades e usos. Deve se resistir a "tentação" de se converter os dados para um uso ou aplicação específica, afinal, dificilmente estes dados serão reutilizados e conectados com outros conjuntos de dados. Outra recomendação consiste em permitir o envolvimento de várias pessoas na identificação de como os dados a serem convertidos se relacionam com outros dados. Esta colaboração contribui para a geração de dados mais ricos e até em alguns casos, a construção de ontologias complexas que facilitarão o reúso dos mesmos (HYLAND; WOOD, 2011).

Uma boa etapa de conversão dos dados devem ser feita atendendo a requisitos como: (1) a conversão deve ser total, ou seja, que todas as consultas que são possíveis na fonte original, também deve ser possível na versão RDF; e (2) as instâncias geradas RDF deve refletir a estrutura da ontologia (ou vocabulário) base para a conversão, tanto quanto possível. O autor apresenta diversas ferramentas que podem ser adotadas para as atividades de conversão automática de dados (VILLAZÓN-TERRAZAS et al., 2011).

4.2.10.2 Adotar rotinas ETL para enriquecimento de dados (7B)

A implementação de rotinas automatizadas pode ser implementada através do conceito de ETL (Extração, Transformação e Carga), comumente utilizada em projetos de Datawarehousing e Business Intelligence (VILLAZÓN-TERRAZAS et al., 2011; BRASIL, 2014b). Ademais, para a geração de dados conectados, as rotinas ETL devem incorporar outras ferramentas de conversão de dados em RDF na fase de transformação (VILLAZÓN-TERRAZAS et al., 2011).

4.2.10.3 Conectar conjuntos de dados com outros dados relacionados (7C)

Antes da publicação de dados conectados, deve ser certificado de que os dados estão conectados a outros conjuntos de dados, pois links para outros conjuntos de dados do órgão publicador e conjuntos de dados a terceiros são úteis para a descoberta de conhecimento. Esses links contribuem para o processamento, integração e reúso de dados e permitem a criação de novos conhecimentos de seus conjuntos de dados, colocando-os em um novo contexto com outros dados. Avaliar e escolher cuidadosamente os conjuntos de dados mais relevantes a ser vinculado com o seu próprio (BAUER; KALTENBÖCK, 2012).

É importante ressaltar que as conexões devem ser feitas em diferentes níveis de granularidade do dado (DING et al., 2011), com atenção especial para a proveniência dos dados que serão conectados. Devem ser utilizados termos compartilhados e estabelecidos por ontologias e bases de conhecimento enriquecidas (AUER et al., 2012).

Além disso, devem ser evitadas a conversão de dados hierarquizados para RDF/XML com poucos ou nenhum link com outros dados, pois o valor da sua informação é relacionado com o conteúdo em si disponibilizado em páginas *Web* somado as conexões que podem ser associadas a este conteúdo (HYLAND; WOOD, 2011).

Villazón-Terrazas et al. (2011) destaca que os links para outros conjuntos de dados podem ser criados manualmente, consumindo maior tempo, ou ainda, mediante o uso de ferramentas automáticas ou supervisionadas. Para isto devem ser seguidos os passos abaixo:

- Devem ser identificados os conjuntos de dados que podem ser conectados. Para isto, devem ser localizados temáticas semelhantes em repositórios de dados disponíveis na *Web*.
- Devem ser identificadas as relações do conjunto de dados com dados ofertados em repositórios governamentais. Algumas ferramentas para apoiar esta tarefa estão disponíveis como o framework SILK, ou o LIMES.
- Finalmente, devem ser validadas as relações identificadas
- Para validar as relações que foram descobertas na etapa anterior. Normalmente esta é uma tarefa manual, feita por especialistas. Nesta etapa, é possível utilizar

ferramentas como o *sameAs link Validator*[27] que tem como objetivo fornecer uma interface amigável para validar links do tipo sameAs.

4.2.10.4 Permitir o envolvimento de várias pessoas na identificação de como os dados a serem convertidos se relacionam com outros dados (7D)

Outra recomendação relevante visa, sempre que possível, envolver o conhecimento de profissionais experientes e usuários no estabelecimento da conexão entre os dados, onde são envolvidos especialistas no domínio dos dados a serem convertidos, analisando os próprios dados, suas lógicas e esquemas. Seria algo comparável às atividades tradicionais de modelagem de banco de dados, agregando requisitos inerentes a dados conectados como o estabelecimento de URIs, vocabulários e a preocupação de se publicar dados e esquemas legíveis por humanos e máquinas. Esta colaboração de especialistas não envolve apenas profissionais técnicos, mas também indivíduos que conheçam do domínio e principalmente, saibam apresentar possibilidades concretas de reutilização dos dados a serem convertidos (HYLAND; WOOD, 2011).

4.2.10.5 Utilizar rotinas automatizadas de conversão de dados, como a triplificação, quando possível (7E)

Especialmente quando a publicação de dados conectados for feita em grandes volumes, como dados geoespaciais, podem ser adotadas rotinas automatizadas de conversão e enriquecimento de dados (HYLAND; WOOD, 2011). Existem três abordagens utilizadas para a conversão de dados conectados que são a conversão automática, conhecida como **triplificação**, a **conversão parcial** com parte suportada por scripts automatizados e finalização mediante trabalho manual e a **modelagem**, que se baseia na elaboração de um modelo de conhecimento desenvolvido por especialistas humanos com posterior conversão automatizada dos dados a partir deste modelo (VILLAZÓN-TERRAZAS et al., 2011; HYLAND; WOOD, 2011). O resultado esperado do uso de uma das três técnicas é a geração de triplas RDF (GALIOTOU; FRAGKOU, 2013). Todavia, a triplificação é mais utilizada para conversão de dados em grandes volumes, com menor compromisso com a qualidade, enquanto as outras duas abordagens garantem maior qualidade, entretanto, com uma produtividade menor (HYLAND; WOOD, 2011).

4.2.10.6 Converter dados em várias serializações RDF (7F)

Os dados conectados ao serem convertidos, devem ser disponíveis em mais de uma serialização, visando atender a diversos públicos que possuem mais facilidade de consumir uma serialização do que outra, e ainda, que ao se publicar em serializações diferentes, esta prática ajuda a validar a qualidade dos dados convertidos pois, se tudo ocorrer de forma

[27] Disponível em http://oegdev.dia.fi.upm.es:8080/sameAs

exitosa, serão mais de um arquivo final contendo os mesmos dados convertidos (HYLAND; WOOD, 2011). O W3C (2014) pontua que existem alguns caminhos para a geração de dados conectados incluindo scripts, linguagens de mapeamento declarativo, linguagens que realizam consulta aos dados conectados ao invés de conversão (ex: R2RML) e independente da abordagem utilizada, a geração de dados conectados é implementada através da estruturação destes dados mediante uma descrição RDF, que são disponibilizados mediante serializações como:

- RDF/XML – Arquivo descrito em RDF em formato XML. É a serialização padrão de RDFs para o W3C.
- RDFa - Serialização RDF que é representada através de sintaxe codificada em páginas HTML. RDFa provê um conjunto de marcações que tornam dados de contidos em páginas HTML legíveis para humanos também legíveis por máquina.
- JSON-LD - Serialização RDF implementada baseada no JSON, que é uma notação de objetos em javascript. A sintaxe JSON-LD é fácil para que humanos a leiam e produzam e por se inspirar em javascript, se beneficia de diversos recursos desta poderosa linguagem *Web*. Foi desenvolvido também como um meio de popularizar dados descritos em RDF para desenvolvedores *Web* já habituados com o uso de javascript e JSON.
- Turtle - Uma serialização projetada para ser facilmente legível por humanos, permitindo que a descrição RDF seja realizada em linguagem próxima a natural, com abreviaturas para padrões de uso comum e tipos de dados.

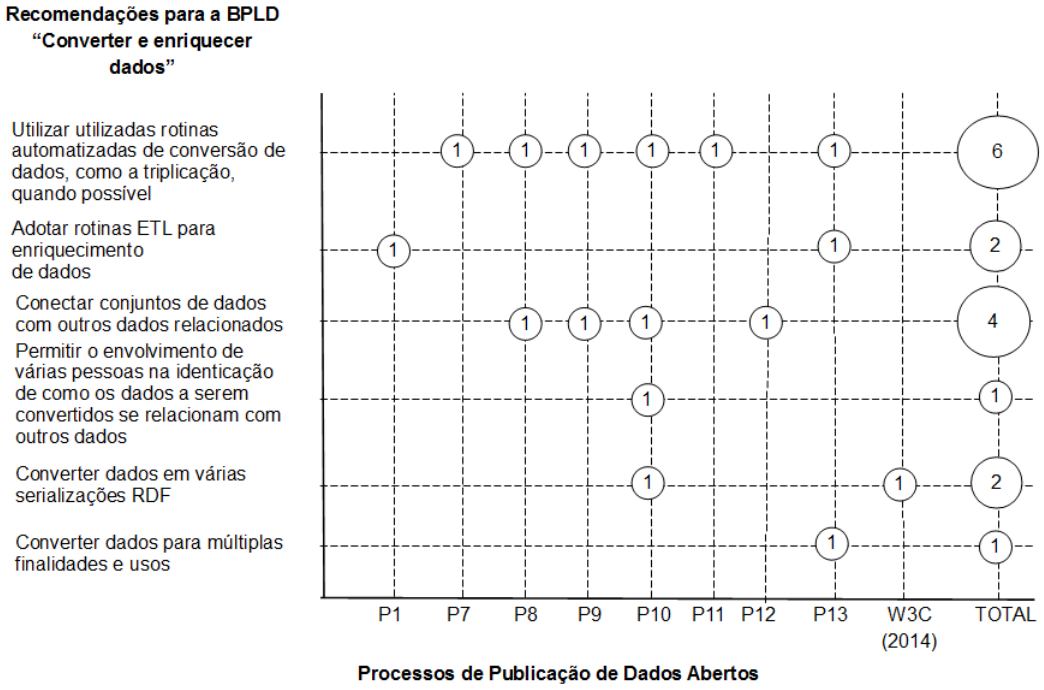
Conforme explanado, tais serializações possuem algumas particularidades que beneficiam o seu consumo por determinado grupo de usuários. O W3C entende que não existe uma serialização melhor do que a outra, mas sim algumas vantagens para o consumo de determinados públicos.

4.2.10.7 Sumarização dos resultados

Da análise desta BPLD foram extraídas seis recomendações, conforme descrito nesta seção. Oito dos quinze processos analisados estabelecem esta BPLD como relevante para a abertura, publicação e conexão de dados. Como esta BPLD é obrigatória para a produção de dados conectados, a investigação buscou extrair recomendações complementares estabelecidas pelo W3C. Para esta BPLD, foi possível se extrair recomendações de todos os processos identificados.

A Figura 36 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram.

Figura 36 – Identificação de recomendações para a BPLD "Converter e enriquecer dados" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Considerando a Figura 36, podemos notar que a existência de recomendações mais e menos significativas a partir da ótica dos processos de publicação. Desta maneira, poderíamos deduzir que, considerando cada recomendação identificada como uma atividade de um processo de publicação, as atividades "Utilizar rotinas automatizadas de conversão de dados, como a triplificação, quando possível" e "Estabelecer bons links com outros conjuntos de dados" foram as que tiveram maior número de recomendações. As demais recomendações, apesar de serem relevantes, podem ser consideradas como recomendações desejáveis a serem cumpridas. As recomendações estabelecidas pelo W3C complementam ou orientam recomendações existentes nos processos. Além disso, cumpre registrar que o processo (HYLAND; WOOD, 2011) contemplou a maioria das recomendações extraídas nesta seção.

4.2.11 RQ 3.8: Recomendações para "Prover acesso automatizado aos dados"

Segundo o (W3C, 2014), o maior benefício dos dados conectados consiste no provimento do acesso a estes dados por máquinas, onde tais máquinas podem utilizar uma variedade de métodos para consumir dados, como: (i) resolver o conteúdo de uma URI; (ii) uma *API* RESTFUL; (iii) um *endpoint* SPARQL; (iv) *download* direto de arquivos, dentre outros.

Para esta BPLD, a extração de recomendações priorizou o provimento do acesso automatizado aos dados conectados, mas também buscou extrair técnicas para acesso aos

dados abertos de uma forma geral.

A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.11.1 Disponibilizar bases completas para *download* (*dumps*) (8A)

Especialmente nos casos onde houver previsão de acesso de conjuntos e recursos de dados de forma integral e com regularidade, são recomendáveis a oferta dos arquivos completos (*dumps*), ora como forma de reduzir o esforço e intensidade de consulta dos *Webcrawlers* as *API*S e *endpoint*s SPARQL, bem como para não prejudicar a performance dos servidores de dados (HYLAND; WOOD, 2011; CONSOLI et al., 2014; DING et al., 2011).

Um "dump" consiste de uma descarga de todo o conteúdo de uma base de dados, estruturada de uma forma que possa ser novamente carregada em um sistema gerenciador de banco de dados (SGBD) idêntico ou compatível, produzindo-se por esse processo uma base de dados que é uma cópia fidedigna da original (BRASIL, 2014a). A utilização de SGBDs facilita a disponibilização de *dumps*, por possuírem recursos naturais para a geração destes arquivos.

Ademais, é importante considerar os formatos de dados necessários à disponibilização via dump, seja em CSV, JSON, XML, formatos geoespaciais como o GeoJSON e KML e ainda as diversas serializações RDF (BRASIL, 2014a).

4.2.11.2 Estabelecer um Mapa de Decisões Tecnológicas (8B)

O processo P1 apresenta uma recomendação muito relevante que consiste no estabelecimento de um mapa de decisões tecnológicas. Este mapa, conforme a Tabela 9 abaixo, permite que a instituição publicadora tenha a dimensão do esforço e complexidade para cada solução tecnológica que for adotar para a oferta de dados.

Cumpre destacar que o mapa elaborado pelo processo brasileiro pode ser amplamente reusado por qualquer iniciativa de publicação de dados abertos e consiste de ferramenta relevante para discussões em nível técnico sobre qual será a melhor solução adotada, considerando a complexidade do projeto, esforço desejado, equipe disponível, dentre outros fatores.

4.2.11.3 Desenvolver uma *API* (8C)

A oferta de uma interface de programação (API) para acesso aos dados vem se tornando requisito comum nos catálogos de dados abertos disponíveis mundialmente e o seu desenvolvimento é fortemente recomendado pois as APIs proporcionam o acesso a partes específicas dos conjuntos de dados ofertados, ao invés de se acessar todo o arquivo.

Tabela 9 – Opções tecnológicas para disponibilização de dados abertos estabelecido pelo processo do Brasil

Solução	Pré-requisitos	Prazo
Publicar *dump* da base de dados	Acesso à base de dados; Servidor *Web* para arquivos	Curto
Publicar dados em arquivos CSV	Mecanismo de ETL (caso esteja em banco relacional); Servidor *Web* para arquivos	Curto
Publicar dados em arquivos JSON / XML	Mecanismo de ETL (caso esteja em banco relacional); Serviço de desenvolvimento; Servidor *Web* para arquivos	Médio
Desenvolver módulo de dados abertos em sistema existente	Serviço de desenvolvimento; Servidor *Web* para *deploy* da nova solução	Longo
Desenvolver *API RESTful* de dados abertos desacoplada da solução	Mecanismo de ETL; Serviço de desenvolvimento; Servidor *Web* para *deploy* da nova solução	Longo
Novo sistema, com a gestão de dados incorporados em sua arquitetura	Mecanismo de ETL; Serviço de desenvolvimento; Servidor *Web* para deploy da nova solução	Longo
Publicar dados em arquivos RDF	Ontologia da área do conhecimento do sistema; Mecanismo de ETL; Servidor *Web* para arquivos	Longo
Disponibilizar dados por *endpoint* SPARQL	Ontologia da área do conhecimento do sistema; Mecanismo de ETL; Banco de dados de triplas	Mais Longo
Publicar dados em *API* de dados conectados (Linked Data)	Ontologia da área do conhecimento do sistema; Banco de dados de triplas; Serviço de desenvolvimento; Mecanismo de ETL; Servidor *Web* para deploy da nova solução	Mais Longo

Fonte: BRASIL (2014c)

Além disso, é um recurso importante para popularizar a oferta de dados abertos junto aos desenvolvedores de software (OKF, 2015a).

Ademais, é recomendado que a *API* em desenvolvimento fique disponível em repositórios públicos, de forma a reduzir as barreiras para que eventuais interessados em testar a solução (por exemplo, outro setor ou organização que tenham interesse em consumir os dados abertos) possam oferecer feedback durante o desenvolvimento e, assim, aprimorar a solução e garantir que quando ela esteja pronta atenda ao seu propósito de ser útil para os consumidores de dados. Desta maneira, a *API* do catálogo de dados estará aderente às demandas dos futuros usuários e provavelmente será muito melhor utilizada (BRASIL, 2014b).

4.2.11.4 Desenvolver um *endpoint* SPARQL (8D)

No caso de oferta de dados conectados, é recomendado que instituições publicadoras o façam através de um *endpoint* SPARQL (HYLAND; WOOD, 2011; CONSOLI et al., 2014; DING et al., 2011), pois tal recurso permite o acesso controlado e específico aos conjuntos de dados descritos em RDF.

Cumpre destacar que é possível prover também o acesso irrestrito aos conjuntos de dados através de um *endpoint* SPARQL, todavia, uma consulta mal elaborada por derrubar um servidor, assim como consultas SQL podem criar problemas a um servidor de banco de dados relacional. Um *endpoint* SPARQL pode estar disponível mediante autenticação ou ainda, pode ser conFigurado determinados limites às consultas através deste recurso, como o tamanho do conjunto de resultados. Normalmente, a resposta de uma consulta a um *endpoint* SPARQL consiste de um resultado SPARQL disponível mediante um link para acesso a um arquivo no formato XML.

Algumas agências governamentais têm um ou mais terminais SPARQL, permitindo que as pessoas para realizar pesquisas através de seus dados. Por exemplo, o Governo do Reino Unido permite o acesso via data.gov.uk/sparql. Eles fornecem dados de referência que cobre o trabalho central do governo, incluindo estruturas organizacionais, disponibilizando todos os conteúdos em RDF. Para os consumidores, *endpoints* SPARQL permitem que sejam construídas consultas integradas acessando vários *endpoints* na *Web* como se o desenvolvedor tivesse consultando dados de diversas tabelas de um mesmo banco de dados.

Ademais, ao desenvolver o *endpoint* SPARQL é recomendável considerar requisitos de desempenho associados às prováveis consultas que serão feitas. Um especialista neste assunto deve ser consultado e ainda, pode ser necessário à oferta de dados brutos completos para aqueles que necessitarem de consultas muito grandes (HYLAND; WOOD, 2011).

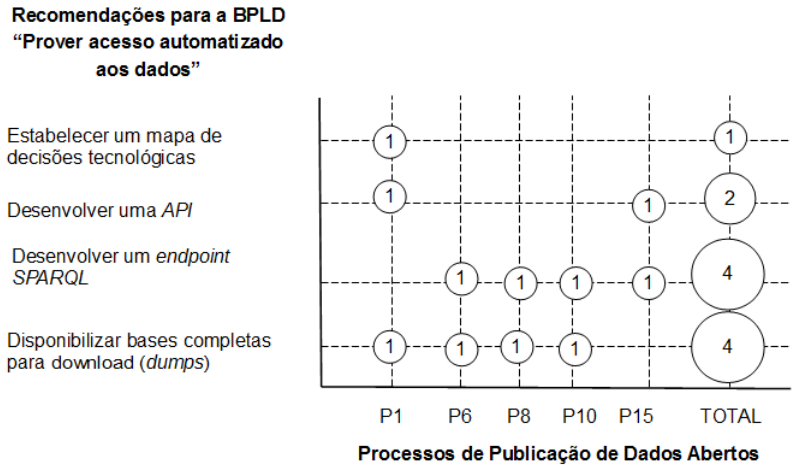
4.2.11.5 Sumarização dos resultados

Da análise desta BPLD, foram extraídas quatro recomendações, conforme descrito nesta seção. Cinco dos quinze processos analisados estabelecem recomendações para esta BPLD, que é fundamental para a oferta de dados abertos para consumo em grande volume e escala. Para esta BPLD, foi possível se extrair recomendações de todos os processos identificados.

A Figura 37 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram:

Considerando a Figura 37, podemos notar que a existência de recomendações mais e menos significativas a partir da ótica dos processos de publicação. Desta maneira, apresentam-se como mais relevantes as recomendações "Desenvolver um *endpoint* SPARQL" e "Disponibilizar bases completas para download (*dumps*)". As demais recomendações podem serem consideradas em caráter desejável, todavia, a recomendação de "Estabelecer

Figura 37 – Identificação de recomendações para a BPLD "Prover acesso automatizado aos dados" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

um Mapa de Decisões Tecnológicas" é extremamente relevante para o planejamento deste esforço para o provimento de acesso automatizado aos dados. Além disso, cumpre registrar que o processo P1 contemplou a maioria das recomendações extraídas nesta seção.

4.2.12 RQ 3.9: Recomendações para "Anunciar os novos conjuntos de dados para o público"

Após todo o esforço para planejamento, preparação e produção dos dados abertos (conectados), fazem-se necessárias medidas para apresentar esta oferta de dados aos usuários. Nesta direção, o W3C (2014) estabeleceu como uma BPLD o "Anuncio dos novos conjuntos de dados ao público".

São sugeridas a adoção de medidas, preferencialmente a serem estruturadas num formato de check-list, para uma boa divulgação dos dados, dentre elas: (i) Adotar um domínio governamental oficial para transmitir maior confiança aos usuários; (ii) Divulgar os dados em diversos canais, como listas, boletins, etc.; (iii) Publicar descrições para os conjuntos de dados; (iv) Definir e disponibilizar a frequência de atualização dos dados; (v) Apresentar garantias de precisão e confiabilidade dos dados; (vi) Disponibilizar espaço para feedback sobre os dados divulgados, dentre outras medidas.

A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.12.1 Publicar metadados junto aos dados (9A)

A publicação de metadados associado aos dados consiste de recomendação relevante. Com ênfase na divulgação dos dados para o público, a publicação e divulgação dos metadados junto aos dados são recomendações presentes nos processos P1, P2, P4, P5, P9

e P13. Para este propósito, existem vocabulários como o VoID [28] que permite expressar metadados sobre conjuntos de dados RDF, contemplando metadados gerais, estruturais e as conexões entre conjuntos de dados.

4.2.12.2 Estabelecer dados tecnicamente e legalmente abertos (9B)

Conforme diversas recomendações apresentadas noutras boas práticas, os dados ofertados precisam ser técnica e legalmente abertos, ou seja não devem dispor de restrições de ordem jurídica ou técnica que impeçam seu uso (OKF, 2015a). Ou seja, do ponto de vista técnico, os dados devem ser disponibilizados em formatos não-proprietários permitindo que qualquer usuário possa acessá-lo, sem a obrigação de fazê-lo mediante a compra de uma licença de software. Do ponto de vista jurídico, devem ser estabelecidas licenças que permitam o acesso, uso e reúso dos dados de forma livre e irrestrita, inclusive para fins comerciais, onde as restrições devem ser aplicáveis somente mediante uma justificativa relevante.

4.2.12.3 Disponibilizar os dados com o menor custo possível ao usuário, preferencialmente de modo gratuito na internet (9C)

Devem ser evitados a cobrança de qualquer taxa para o acesso aos dados públicos. Estes devem ser disponibilizados com o menor custo possível ao usuário e de forma integral (com completude), deixando o usuário decidir qual parte da base de dados irá utilizar (OKF, 2015a). Isto inclui a disponibilização dos dados em formatos abertos, considerando que os formatos proprietários podem impor o custo de aquisição de licenças de software para acessar e usar os dados.

4.2.12.4 Divulgar dados em meios complementares (Catálogos, FTP, Torrent) (9D)

Ademais, todos os recursos complementares para a divulgação dos dados deve ser adotado. Uma medida importante consiste na inclusão dos catálogos e respectivos conjuntos de dados da instituição publicadora em ferramentas de indexação de catálogos como o DataPortals [29] e Open Government Data Catalog [30] (VILLAZÓN-TERRAZAS et al., 2011). Quanto ao estabelecimento de catálogos de dados como instrumentos de publicação e disseminação dos dados, o processo P14 apresenta uma escala que pode ser adotada a depender do nível de recursos e maturidade que a organização publicadora possua (COMSODE, 2014b):

- Catálogo de dados primário: Página HTML simples que informa ao público que a organização publica seus dados como dados abertos. Deve habilitar o download de

[28] Disponível em http://vocab.deri.ie/void
[29] Disponível em http://dataportals.org
[30] Disponível em http://opengovernmentdata.org/data/catalogues/

um arquivo de dados com registros de catálogos exportados numa notação legível por máquina;

- Catálogo de dados básico: Amplia a página HTML simples do catálogo de dados primitivo com uma lista de links para páginas HTML dedicados a conjuntos de dados específicos publicados pela organização. Cada página HTML descreve de forma legível os metadados a partir do registro do catálogo do conjunto de dados, fornecendo links para baixar os recursos do conjunto de dados. Também permite fazer o download do registro do catálogo do conjunto de dados em uma notação legível por máquina.

- Catálogo de dados completa: Amplia as funcionalidades do catálogo, incluindo recursos de pesquisa e outras funções (por exemplo, previews de distribuições de conjunto de dados, discussões de usuários, etc.).

Complementarmente, outros recursos de disseminação de podem ser adotados, conforme descrição abaixo:

- Divulgação via FTP: Apesar de ser um método menos elegante, a divulgação via protocolo FTP pode ser utilizada para um público mais técnico, como desenvolvedores, costumando ser utilizado para a transferência de arquivos muito grandes, que é uma das finalidades do protocolo FTP.

- Divulgação na forma de *torrent*: A divulgação via *torrent* pode ser utilizada quando da existência de servidores distribuídos para armazenamento dos dados. É recomendável para oferta de dados que são muito demandados e que precisam de estrutura distribuída de disponibilização que suporte tal demanda.

Destaca-se ainda que as atividades de publicação de conjuntos e recursos de dados em catálogos de dados são consideradas como atividades preliminares à divulgação para o público em alguns processos (P2, P4 e P5).

4.2.12.5 Divulgar dados em seções destacadas de sítios de governo (9E)

Complementarmente, os processos P1 e P4 sugerem que os dados publicados devem ser divulgados em seções específicas e destacadas de sítios institucionais de governo (BRASIL, 2014c; ECUADOR, 2014). O processo P4, por exemplo, estabelece os procedimentos obrigatórios para divulgação dos dados abertos nos sítios institucionais.

Além dos catálogos e portais centrais, outra medida relevante a ser adotada consiste na divulgação dos dados em sítios de terceiros, devendo ser disponibilizados recursos para divulgação dos dados dos catálogos (todos ou um subconjunto) em sítios externos. Por exemplo, um sítio relacionado à área de saúde se interessará em divulgar para os

seus usuários quais dados sobre saúde estão disponíveis para consulta num determinado catálogo. A divulgação por terceiros estimula o uso dos dados em comunidades de interesse específico.

4.2.12.6 Estabelecer recursos de consulta parcial da base de dados como uma *API* ou *Web-service* (9F)

Conforme explanado nas recomendações para *"Prover acesso automatizado aos dados"* no item 4.2.11, a adoção de recursos de consulta parcial da base de dados como uma *API*, *Webservice* ou um *endpoint* SPARQL devem ser ofertados complementarmente a disponibilização da base integral. Cumpre destacar que o usuário preferencialmente não deve ser obrigado a baixar um arquivo muito grande para ter acesso a um ou alguns dados. Como exemplo, caso o usuário deseje consultar o nome de algumas ruas da Cidade de São Paulo, será necessário baixar um arquivo de 8GB no sítio do Instituto Brasileiro de Geografia e Estatística – IBGE.

4.2.12.7 Estabelecer visualizações e demais recursos de exploração dos dados (9G)

Ademais, há de se considerar que para muitos usuários, os dados na *Web* ainda estão "invisíveis abaixo da superfície", sendo necessário o desenvolvimento de recursos de busca, exploração e visualizações para os diferentes tipos de dados, especialmente os conectados (geoespaciais, temporais, estatísticos) que tornem os dados relevantes e tangíveis para os usuários (AUER et al., 2012).

4.2.12.8 Melhorar os dados para que sejam mais facilmente encontrados por máquinas (9H)

Uma das etapas finais da publicação de dados consiste na adoção de medidas eficazes que proporcionem a descoberta e sincronização dos conjuntos de dados (VILLAZÓN-TERRAZAS et al., 2011). No que tange a divulgação de dados conectados, é ressaltado que tais atividades devem ser ainda mais amplas que as atividades comumente utilizados para dados não conectados (e legíveis por máquina), pois dados conectados devem ser legíveis por máquina e consequentemente, encontráveis por máquinas (HYLAND; WOOD, 2011).

A seguir serão apresentados alguns critérios estabelecidos por Richard Cyganiak para fins de adesão de dados ao projeto *Linked Data Cloud*, onde a adoção de tais critérios poderão garantir uma divulgação mais ampla dos dados.

1. Os dados devem ser acessíveis através de uma URI http:// (ou https://);
2. As URIs devem resolver, com ou sem a negociação de conteúdo, para dados RDF em um dos formatos populares RDF (*RDFa, RDF / XML, Turtle, N-Triple*).
3. O conjunto de dados deve conter, pelo menos, 1.000 triplas.

4. O conjunto de dados deve ser conectado através de links RDF para um conjunto de dados no diagrama LOD (*Linked Open Data Diagram*)[31]. Isto significa que o seu conjunto de dados deve usar URIs de outros conjuntos de dados, e vice-versa. Sugere-se que cada conjunto de dados tenha, pelo menos, 50 destas conexões.

5. Deve ser viabilizado o acesso a todo o conjunto de dados mediante RDF *crawling*, através de um *dump* RDF ou um *endpoint* SPARQL.

Complementarmente, outras medidas podem ser adotadas para tornar os dados ainda mais localizáveis como anunciar os conjuntos de dados para o público para ferramentas de busca, adicionando tags em RDFa. Villazón-Terrazas et al. (2011) recomenda a adoção do *sitemaps* como um padrão para que os *crawlers* tomem conhecimento da existência de páginas e conjuntos de dados publicados nos sites e/ou catálogos de dados. Desta maneira, os motores de busca saberão quando existirem atualização de dados ou publicação de novos. Para isto, é necessária (1) a geração de um conjunto de arquivos sitemap.xml referente à oferta de dados ou do *endpoint* SPARQL, e (2) enviar os arquivos para motores de busca (semânticos) na *Web*, tais como *Google* [32]. Nesta etapa, podemos contar com ferramentas automáticas como *sitemap4rdf* [33].

Por último, deve ser informada a existência de novos dados (ou atualizações) para os *mailing lists* das comunidades interessadas, organizações da sociedade civil, imprensa, instituições acadêmicas e *hackers* cívicos (HYLAND; WOOD, 2011; BRASIL, 2014c).

4.2.12.9 Disponibilizar dados conectados em servidores de triplas (9I)

Para dados conectados, devem ser estabelecidos servidores de triplas que permitam a publicação dos dados na *Web*, sendo estes dados conectados com outros dados através de URIs bem estabelecidas. Ferramentas como o *Virtuoso* [34] são recomendadas para a implementação desta atividade (BAUER; KALTENBÖCK, 2012; VILLAZÓN-TERRAZAS et al., 2011; GALIOTOU; FRAGKOU, 2013; MENDONÇA et al., 2013)

4.2.12.10 Sumarização dos resultados

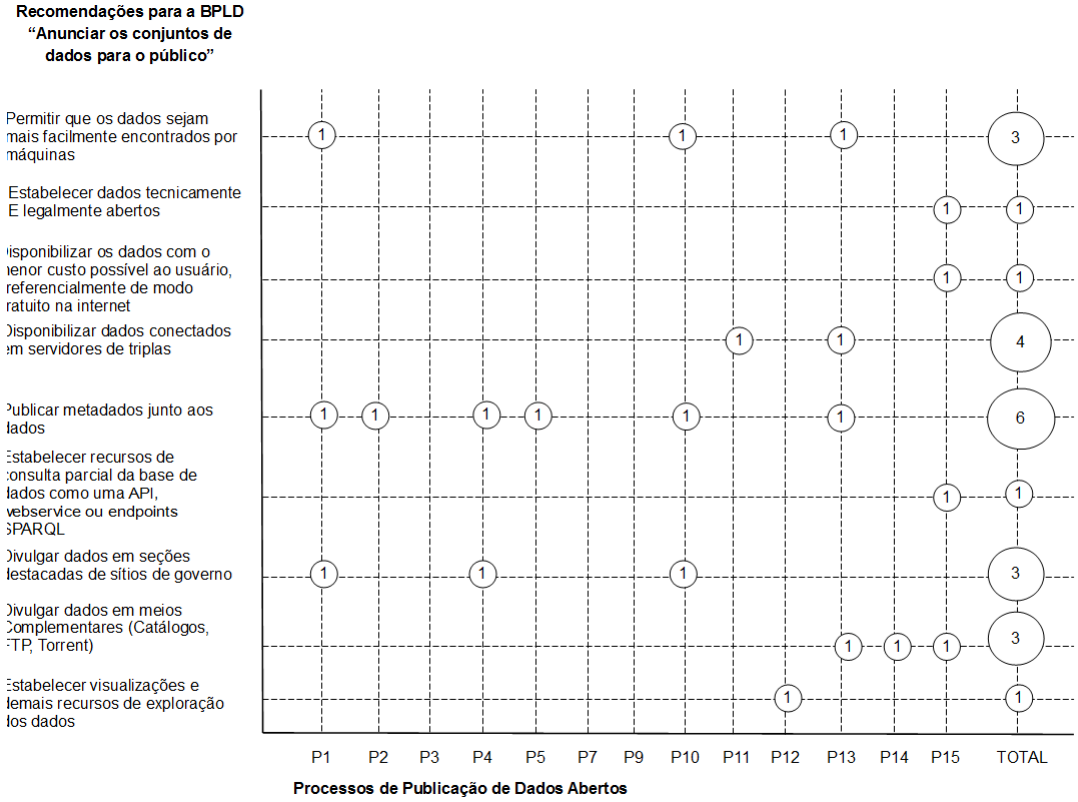
Considerando os processos analisados, praticamente houve unanimidade quanto à existência de uma etapa de divulgação dos dados para o público, conforme recomenda o W3C. Em apenas dois processos não foi extraída alguma recomendação para esta BPLD, pois nestes casos, o processo se encerra na disponibilização de uma ferramenta de consumo de dados (como uma *API* ou *endpoint* SPARQL) não havendo esforços adicionais para

[31] Disponível em http://lod-cloud.net
[32] Disponível em http://www.google.com
[33] Disponível em http://lab.linkeddata.deri.ie/2010/sitemap4rdf/
[34] Disponível em http://virtuoso.openlinksw.com/

a divulgação desta oferta de dados. Ademais, cumpre destacar que os métodos de divulgação de dados são muito variados o que permitiu a extração de diversas recomendações. Para esta BPLD, foi possível extrair recomendações de todos os processos identificados.

A Figura 38 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram:

Figura 38 – Identificação de recomendações para a BPLD "Anunciar os novos conjuntos de dados para o público" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Considerando a Figura 38, as recomendações encontram-se bem distribuídas entre os processos analisados, não sendo possível identificar recomendações mais ou menos frequentes.

4.2.13 RQ 3.10: Recomendações para "Estabelecer um contrato social para os dados publicados"

Dados públicos, especialmente os de natureza conectada, por naturalmente estarem conectados com outros inúmeros conjuntos de dados, precisam dispor de condições e compromissos que garantam a sua alta disponibilidade e atualização. A indisponibilidade de um dado conectado pode inviabilizar o funcionamento de todos os demais dados que estejam conectados a este dado "indisponível", prejudicando consequentemente as aplicações que dependem dele.

Devido a este especial requisito, o W3C (2014) estabelece como BPLD para a publicação de dados conectados a existência de um "contrato" que registre a responsabilidade social da instituição publicadora com os usuários dos dados. Este contrato deve estar disponível publicamente no repositório de dados governamentais proporcionando maiores garantias para que os clientes reutilizem estes dados sabendo exatamente como e onde os dados estarão disponíveis permanentemente.

A seguir serão apresentadas as recomendações identificadas nos processos que poderão auxiliar a incorporação desta BPLD em atividades de publicação de dados.

4.2.13.1 Estabelecer com clareza que o processo de publicação contempla etapas de manutenção e atualização dos dados (10A)

Considerando que os dados na *Web* são dinâmicos e que a evolução dos dados deve ser descrita e programada, demonstrando a estabilidade no serviço de dados. Mudanças e modificações nas bases de conhecimento, vocabulários e ontologias devem ser transparentes e auditáveis. O processo apresenta ferramentas que permitem visualizar a evolução e a interligação das bases de conhecimento, especialmente os vocabulários (AUER et al., 2012). Para demonstrar maior clareza quanto à comunicação destas atividades, preferencialmente deve ser descrito, em linhas gerais, como a manutenção e atualização dos dados é feita. Por exemplo, podem ser explicados quais conjuntos de dados são atualizados automaticamente e quais não são, explicando os razões em cada caso (COMSODE, 2014b).

Complementarmente, outras medidas podem ser adotadas para comunicar a atualização de dados. No geral, estas atividades consistem em:

1. Disponibilizar a relação de versões disponíveis do conjunto de dados;
2. Disponibilizar a última versão exata do conjunto de dados como um "dump";
3. Disponibilizar uma informação sobre o *"dif."* entre a versão atual e uma determinada versão anterior de um conjunto de dados

4.2.13.2 Estabelecer mecanismos de monitoramento e avaliação da oferta de dados disponibilizados ao público (10B)

O monitoramento e avaliação da qualidade do serviço de dados consiste de uma recomendação muito relevante para disponibilizar ao usuário um compromisso do órgão publicador com sua manutenção e atualização. Para isto, destacamos as determinações do Governo do Equador, que estabelece um procedimento de monitoramento e avaliação dos dados abertos ofertados, conforme detalhamento abaixo (ECUADOR, 2014)

- A instituição deve comunicar a sociedade a taxa de cumprimento do seu plano de divulgação de dados, de acordo com o artigo 7º da *La Ley Orgánica de Transparencia y Acceso a la Información (LOTAIP)*;
- Disponibilizar a relação atualizada dos conjuntos de dados publicados nos portais institucionais;
- Disponibilizar a relação de serviços on-line que ofertam conjuntos de dados abertos;
- Estabelecer um serviço de classificação do dado ofertado, com um sistema de pontuação de 1 a 5, onde 1 é a classificação mais baixa e 5 é a mais alta;
- Disponibilizar o número de visitas ao catálogo (ou página) que oferta dados abertos;
- Disponibilizar as estatísticas de acesso e *downloads* de conjuntos de dados.

4.2.13.3 Disponibilizar leis e atos normativos que explicitem aos usuários quanto às obrigações dos governos em publicarem dados com qualidade e disponibilidade (10C)

Outra medida importante a ser adotada visa aumentar a segurança jurídica da oferta de dados a ser disponibilizada. Especialmente quanto aos dados governamentais, que são sujeitos a mudanças de governantes periodicamente, se não houver elementos que garantam a disponibilidade e atualização dos dados independente das mudanças de diretrizes governamentais, os usuários podem não obter confiança em utilizar tais dados, especialmente se forem conectados, por requererem uma alta disponibilidade.

Por esta razão, é fortemente recomendado que sejam disponibilizadas as leis e normas que explicitem aos usuários quanto às obrigações dos governos em publicarem dados com qualidade e disponibilidade. Esta medida inclusive servirá para que os usuários exijam dos governos a manutenção do serviço especialmente nos momentos de mudanças, após eleições (HYLAND; WOOD, 2011). Por exemplo, o Governo do Equador determina que sejam informados nos catálogos de dados que a oferta de dados estará disponível ao público durante as 24 horas do dia, e em todos os 365 dias do ano (ECUADOR, 2014).

4.2.13.4 Estabelecer espaços para recebimento do feedback do usuário, preferencialmente publicando dados de uma pessoa e/ou telefone de contato para esclarecimento de dúvidas sobre o uso e disponibilidade dos dados (10D)

Considerando que a oferta de dados deve ser orientada a demanda dos usuários, é fortemente recomendado que sejam estabelecidos canais de contato e relacionamento entre a instituição publicadora e seus clientes. O feedback do usuário pode ser uma contribuição muito valiosa para a manutenção e aprimoramento do conjunto de dados, pois eles podem ser capazes de identificar e reportar erros nos conjuntos de dados. Erros detectados em conjuntos de dados e metadados ou outras questões relacionadas com o fornecimento

de dados pode desencadear atividades de manutenção de conjuntos de dados, por exemplo, liberação do arquivo de dados corrigidos e atualização para o respectivo registro do catálogo (COMSODE, 2014b).

Por esta razão, devem ser disponibilizados, no mínimo, os dados de um responsável e seus contatos (telefone e e-mail) para que os usuários possam enviar seus feedbacks. Obviamente, outros recursos de interação mais avançados podem ser estabelecidos (HYLAND; WOOD, 2011).

4.2.13.5 Utilizar tecnologias que mantenham os dados conectados disponíveis, atualizados e abertos (10E)

Outra forma de comunicar aos usuários o compromisso com a alta disponibilidade dos serviços de dados está relacionada com as tecnologias adotadas para suportar a oferta de dados conectados ou não, atualizados e disponíveis (HYLAND; WOOD, 2011). Nesta direção, o Governo do Equador estabelece que as instituições públicas devem garantir o acesso livre aos seus dados através da internet utilizando tecnologias livres, que não criem barreiras ao acesso aos dados e ainda, estabeleçam as condições necessárias para a proteção dos direitos da instituição publicadora e dos autores dos dados (ECUADOR, 2014).

4.2.13.6 Sumarização dos Resultados

Da análise desta BPLD, foram extraídas cinco recomendações, conforme descrito nesta seção. Apenas quatro dos quinze processos analisados estabelecem esta BPLD como relevante para a abertura, publicação e conexão de dados. Para esta BPLD, foi possível se extrair recomendações de todos os processos identificados.

A Figura 39 apresenta a relação entre as recomendações identificadas e a presença nos processos que a contemplaram.

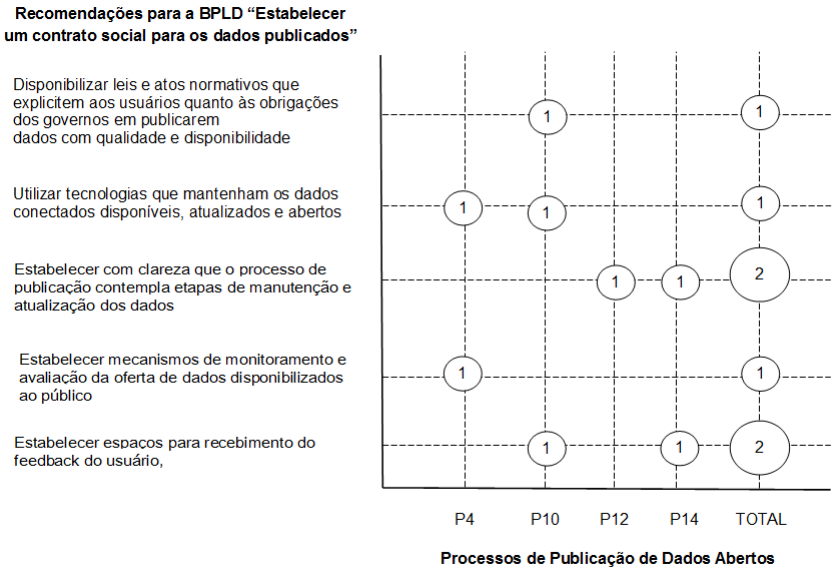
Considerando a Figura 39, as recomendações encontram-se bem distribuídas entre os processos analisados, não sendo possível identificar recomendações mais ou menos frequentes. Ademais, verifica-se que o processo P10 contemplou três das cinco das recomendações extraídas, podendo este processo ser melhor analisado pelos órgãos publicadores quando houver a necessidade de estabelecer esta BPLD.

4.3 **Sumarização geral**

Conforme exposto, esta revisão de literatura permitiu a identificação de recomendações relevantes para a implementação de cada uma das BPLDs. O gráfico representado na Figura 40 apresenta como sumarização, o total de recomendações estabelecidas.

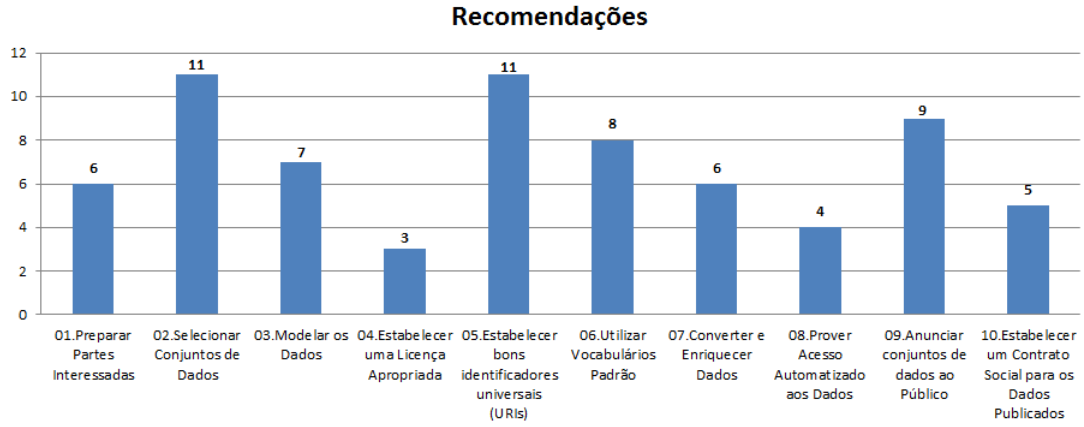
Ademais, como cada uma destas recomendações visam implementar as 10 BPLDs, o modelo proposto nesta pesquisa apresenta as recomendações identificadas como atividade-

Figura 39 – Identificação de recomendações para a BPLD "Estabelecer um contrato social para os dados publicados" nos processos de publicação de dados abertos analisados



Fonte: Autor desta dissertação, 2015.

Figura 40 – Sumarização das recomendações identificadas na revisão de literatura



Fonte: Autor desta dissertação, 2015.

des que poderão ser desenvolvidas para a publicação de dados abertos e dados abertos conectados, conforme descrito no próximo capítulo.

5 MODELO DE PROCESSO "PIECE OF CAKE"

Este capítulo apresenta a principal contribuição desenvolvida nesta dissertação, que consiste numa proposta de modelo de processo iterativo e incremental para publicação de Dados Abertos Conectados Governamentais (DACG). Aqui são apresentadas as premissas do modelo, seus objetivos, sua visão geral e específica, bem como seus passos e como cada um foi construído. Na visão específica, descrita nas seções e subseções deste capítulo, são disponibilizados um conjunto de atividades obrigatórias e desejáveis sugeridas pelo modelo para serem aplicadas a depender da maturidade da instituição publicadora e do nível de maturidade desejado para a publicação de dados abertos e dados abertos conectados.

As características gerais partem do **Esquema 5-Estrelas dos Dados Abertos** (BERNERS-LEE, 2006). Complementarmente, foram extraídas características extraídas dos **modelos de processo de software iterativos e do PMBoK**. Quanto às características específicas, são apresentadas atividades de implementação para cada uma das BPLDs, propostas a partir das recomendações extraídas da revisão de literatura.

5.1 Visão geral do modelo

Esta seção apresenta uma visão geral do modelo de processo para publicação de DACG, bem como os requisitos estabelecidos para o seu desenvolvimento.

5.1.1 Características extraídas do Esquema 5-Estrelas dos Dados Abertos

Conforme explorado nos capítulos anteriores, os dados abertos conectados governamentais proporcionam inúmeros benefícios para publicadores e consumidores destes dados. Entretanto, conforme estabelecido pelo Esquema 5-Estrelas, um dado para ser considerado conectado precisa estar no 5º nível deste esquema, e para isto, é necessário atender aos requisitos dos níveis anteriores.

Por haver este pré-requisito para que um dado seja considerado conectado, e considerando as características de cada nível do Esquema 5-Estrelas, é possível deduzir que a publicação de um dado aberto conectado deve, preferencialmente, ser a consequência da evolução de um dado aberto que se aprimorará mediante a incorporação de novas atividades em torno deste determinado dado ou conjuntos de dados.

Entretanto, o Esquema 5-Estrelas considera, nos níveis 1 e 2, que dados são abertos desde que sejam publicados na *Web* com uma licença aberta contemplando formatos proprietários como o PDF. Neste contexto, Alcantara et al. (2015) pontua que formatos proprietários, como o PDF, não incentivam a sua reutilização, pois restringe o seu acesso somente às pessoas que possam adquirir softwares proprietários que possibilitem o manuseio deste tipo de arquivo. Por outro lado, a OKF (2015c) define que um dado para

ser aberto deve permitir que qualquer pessoa possa utilizá-lo, reutilizá-lo e redistribuí-lo livremente. Pelo nosso entendimento, a definição do Esquema 5-Estrelas contempla a utilização de formatos de dados proprietários (Nos níveis 1 e 2) enquanto a definição da OKF contempla apenas formatos abertos.

Desta maneira, comparando as duas definições, existe convergência a partir do terceiro nível do esquema 5-Estrelas. Visando pacificar esta conceituação, este modelo adotará o entendimento de Isotani e Bittencourt (2015), que aconselham que os dados sejam abertos considerando no mínimo 3 estrelas. Assim, o modelo se propõe a guiar a publicação de dados que atendam os requisitos dos níveis 3, 4 e 5 do Esquema 5-Estrelas.

5.1.2 Características extraídas dos modelos de processo de software iterativos

Segundo Sommerville (2007), em modelos de processos de software incrementais, na medida que novos incrementos são concluídos, estes são integrados aos já existentes, de tal forma que o software é aprimorado a cada incremento entregue. Nesta subseção são apresentadas as características destes modelos de processo que foram absorvidas pelo *"Piece of Cake"*.

O modelo de processo de software Entrega Incremental proporciona que a solução (neste caso um software), possa ser desenvolvida em incrementos, iniciando pelas atividades e requisitos mais prioritários e, gradativamente, incorporando requisitos adicionais. Os novos requisitos são adicionados aos existentes, aprimorando a solução final.

Com base no Esquema 5-Estrelas, podemos deduzir que tal lógica é aplicável à publicação de dados abertos e dados abertos conectados. Desta maneira, utilizando o conceito de entrega incremental, um dado a ser publicado pode ser desenvolvido incrementalmente, mediante a execução de atividades prioritárias inicialmente com a incorporação de novas atividades.

Neste raciocínio, como um dado é considerado conectado no nível máximo do Esquema 5-Estrelas, é necessário que um modelo de processo para esta finalidade possa guiar o publicador à, inicialmente, desenvolver atividades prioritárias para publicação de dados abertos (3-Estrelas) para posteriormente incorporar novas atividades que evoluam a publicação para os níveis de 4 e 5 estrelas.

De forma complementar, o modelo de processo de software espiral proporciona que a solução (neste caso um software), possa ser em ciclos e fases que organizam as etapas de desenvolvimento de uma solução. Conforme as visões de Boehm (1986) e (PRESSMAN, 1995) a espiral de software é composta de 4 grandes quadrantes (fases) com atividades distribuídas do longo de cada quadrante conforme a Tabela 10:

Tabela 10 – Comparativo entre as visões de Boehm (1986) e Pressman (1995) sobre a espiral de software

Etapas:	**Visão de Boehm (1986)**	**Visão de Pressman (1995)**
1º Quadrante:	Determinar objetivos, alternativas e restrições	Planejamento
2º Quadrante:	Avaliação e redução de riscos	Análise de Riscos
3º Quadrante:	Desenvolvimento e validação	Engenharia
4º Quadrante:	Planejamento da próxima fase	Avaliação dos resultados

Fonte: Autor desta dissertação, 2015.

5.1.3 *Características extraídas do Project Management Body of Knowledge (PMBoK)*

Para o gerenciamento de projetos, o PMBoK considera cinco grupos de processos. A iniciação corresponde às atividades para definir um novo projeto ou nova etapa de um projeto. O Planejamento corresponde ao escopo do projeto, objetivos e linha de ação a ser desenvolvida. No grupo de processos de Execução são estabelecidas as atividades para o desenvolvimento do projeto atendendo ao escopo, cronograma, orçamento e riscos. O Monitoramento e Controle corresponde ao acompanhamento, análise e verificação de riscos e necessidades de mudança e o Encerramento visa a conclusão de um determinado projeto ou fase.

Aprimorando a Tabela 10 mediante o comparativo das fases da espiral com os grupos de processos do PMBoK, deparamo-nos com a Tabela 11:

Tabela 11 – Comparativo sobre os quadrantes da espiral de software conforme as visões de Boehm (1986) e Pressman (1995) com os grupos de processos propostos pelo PMBoK

Fases:	PMBoK(PMI, 2013)	**Visão de Boehm (1986)**	**Visão de Pressman (1995)**
1a Fase:	Iniciação	-	-
2a Fase:	Planejamento	Determinar objetivos, alternativas e restrições	Planejamento
3ª Fase:	Execução	Desenvolvimento e validação + Avaliação e redução de riscos	Engenharia
4ª Fase:	Monitoramento e Controle	Desenvolvimento e validação + Avaliação e redução de riscos	Análise de Riscos
5ª Fase:	Encerramento	Planejamento da próxima fase	Avaliação dos resultados

Fonte: Autor desta dissertação, 2015.

Conforme observado, os grupos de processos do PMBoK apresentam correlação com as fases da espiral propostas por Boehm e Pressman, acrescentando a fase de Iniciação.

Neste raciocínio, considerando que a publicação de dados é um projeto e que deve ser desenvolvido de forma iterativa e incremental pelas justificativas anteriores, o modelo de processo desta pesquisa propõe a integração do conceito estabelecido pelos 5 grupos de processos do PMBoK com os quadrantes do processo de desenvolvimento em espiral. A próxima subseção apresentará a visão geral do modelo de processo que considera as características apresentadas nesta seção.

5.1.4 Visão global do modelo

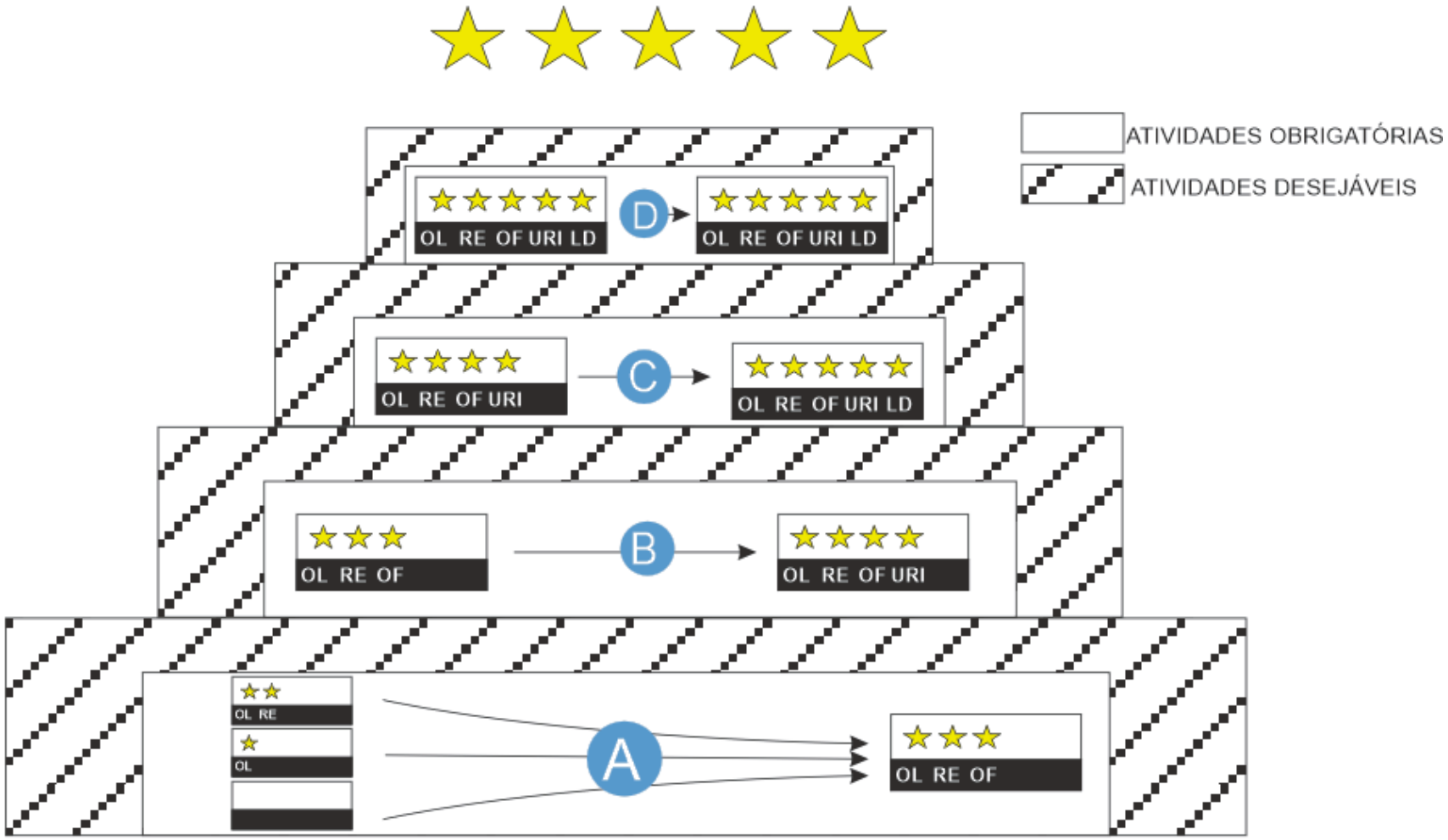
Considerando as características acima, esta proposta de modelo de processo tem como objetivo guiar instituições governamentais, agentes públicos e demais indivíduos e instituições interessadas para que possam publicar dados abertos conectados governamentais respeitando, principalmente, as características dos níveis 3, 4 e 5 do esquema 5-Estrelas dos Dados Abertos. O modelo tem esta denominação por sugerir o acréscimo de atividades em camadas (como pedaços de um bolo) conforme os níveis do Esquema 5-Estrelas, onde cada camada do "bolo" é um processo a ser desenvolvido. Cada processo é composto por um conjunto de atividades que são distribuidas ao longo de etapas e fases conforme será explorado mais adiante.

Neste contexto, os processos do modelo visam:

- Inicialmente guiar a publicação de um dado (ou conjunto de dado) em formato aberto no nível 3 estrelas (Processo A);
- Posteriormente, evoluí-lo para os níveis 4 e 5 estrelas (Processos B e C); e
- Por último, propõe-se um último estágio de enriquecimento dos dados, denominado preliminarmente de nível 5-estrelas aprimorado (Processo D).

A Figura 41 ilustra a visão global do modelo.

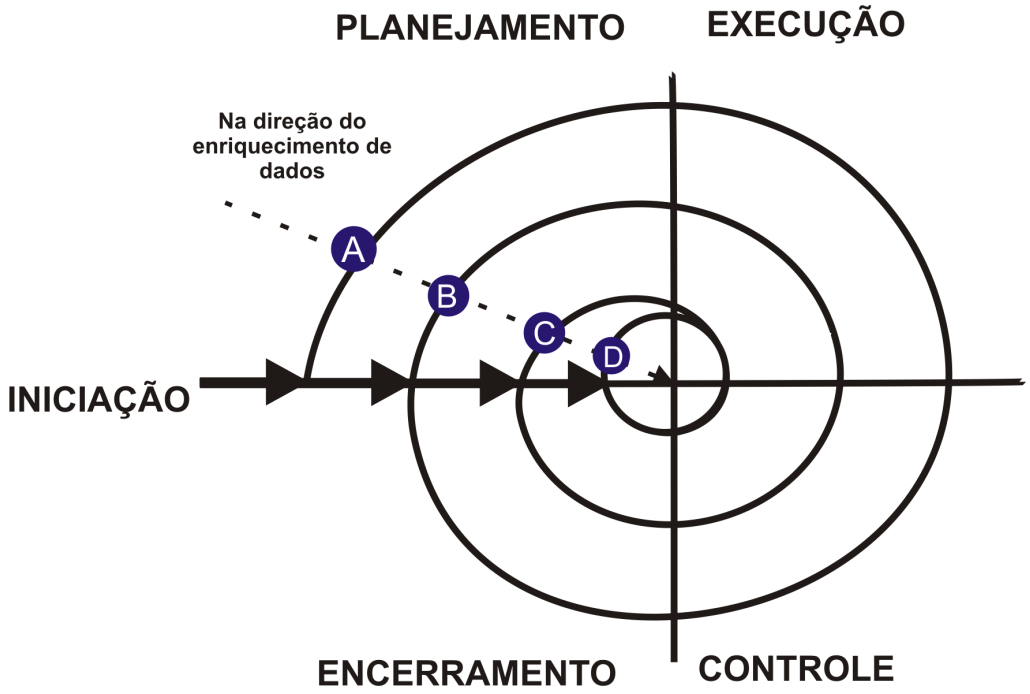
Figura 41 – Visão geral do modelo de processo *"Piece of Cake"*



Fonte: Autor desta dissertação, 2015.

A evolução entre as camadas será feita de forma iterativa e incremental, podendo percorrer cada um dos quatro processos (A, B, C e D) conforme apresentado na Figura 41. As iterações tem como referência o modelo de processo de desenvolvimento de software "espiral", onde os quatro processos serão apresentados e detalhados na próxima subseção. A Figura 42 apresenta um detalhamento de cada ciclo evolutivo dos dados abertos a partir desta proposta.

Figura 42 – Visão evolutiva do modelo de processo *"Piece of Cake"*



Fonte: Autor desta dissertação, 2015.

Os processos do modelo serão desenvolvidos mediante grandes fases, compostas de etapas associadas as BPLDs e atividades complementares. Abaixo descrevemos as cinco fases:

1. **Iniciação:** Atividades iniciais do modelo de processo. Visa principalmente identificar a maturidade da instituição governamental em atividades de publicação de dados bem como a maturidade da publicação de dados desejada;

2. **Planejamento:** Atividades relacionadas à definição dos objetivos, benefícios da publicação de dados, atores a serem envolvidos, usuários dos dados, seleção dos conjuntos de dados que serão publicados, identificação de necessidades de sigilo de dados, mensuração do esforço da tarefa de publicação, dentre outras.

3. **Execução:** Atividades relacionadas ao desenvolvimento da publicação de dados. Contempla a modelagem dos dados, especificação de licenças e provimento de acesso automatizado. Para os níveis finais do esquema 5-Estrelas, também contemplará

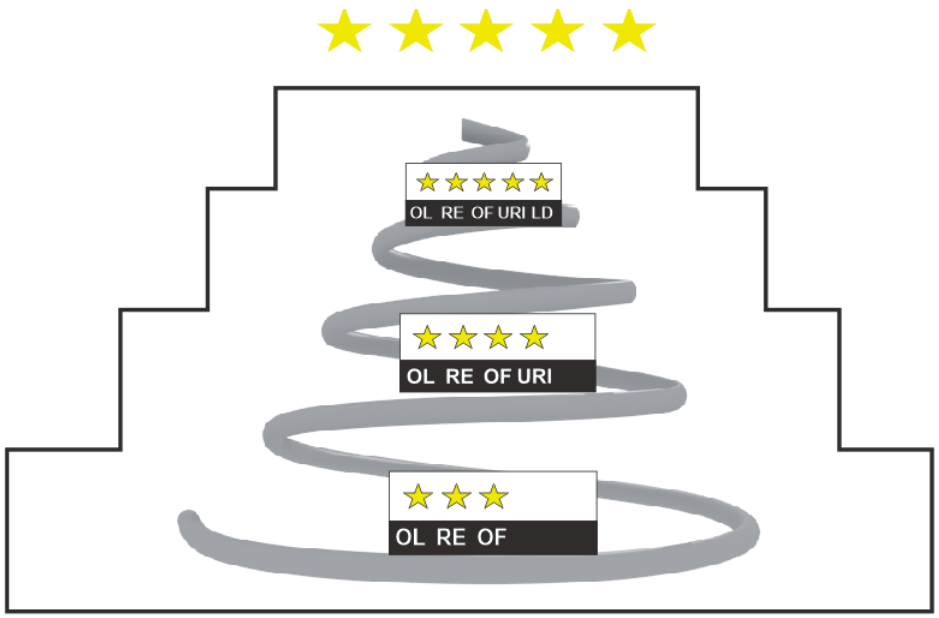
atividades como definição de boas URIs, vocabulários e conversão e enriquecimento de dados.

4. **Controle:** Consiste na verificação da eficácia e da completude das atividades relacionadas ao desenvolvimento da publicação de dados. Nesta fase deverá ser realizada a identificação de problemas ocorridas na fase anterior (e possíveis causas). Esta fase gera subsídios pra primeira etapa da fase de encerramento, a retrospectiva.

5. **Encerramento:** Composta de uma fase inicial, onde deve ser feito uma retrospectiva das fases e etapas anteriores, visando decidir se haverá necessidade de melhorar a qualidade da publicação (repetindo o processo corrente) ou se será possível avançar para o novo processo. Caso seja possível avançar, devem ser desenvolvidas as atividades relacionas à conclusão do projeto de publicação de dados ou ainda, uma parte do mesmo. Contempla a divulgação dos dados para o público e o estabelecimento de regras e diretrizes onde a instituição se compromete a manter disponível, atualizada e aberta esta oferta de dados que está sendo publicada.

Por fim, espera-se que o modelo possa guiar instituições governamentais a publicarem seus dados até o nível de dados abertos conectados. A Figura 43 apresenta uma abstração do que o objetivo desta proposta de modelo de processo pretende alcançar.

Figura 43 – Evolução do modelo de processo *"Piece of Cake"* dentre os níveis do esquema 5-Estrelas



Fonte: Autor desta dissertação, 2015.

A próxima seção apresentará o detalhamento do modelo de processo *"Piece of Cake"*.

5.2 **Visão detalhada do modelo de processo "Piece of Cake"**

Esta seção apresenta o detalhamento do modelo de processo desta pesquisa, bem como as atividades estabelecidas para o seu desenvolvimento.

5.2.1 Etapas e atividades extraídas das BPLDs

Conforme explanado, este modelo visa induzir a publicação de dados abertos conectados. Neste contexto, o W3C elaborou as 10 BPLDs que consideram o nível de envolvimento de partes interessadas na publicação de dados, cuidados com a legalidade, modelagem, processamento e organização dos dados, semântica, atualização, alta disponibilidade dos dados e ainda, como deve ser feita a sua divulgação para o público e sua manutenção.

Assim sendo, considerando o propósito, relevância e atualidade desse trabalho, estas BPLDs são incorporadas ao núcleo desta pesquisa como etapas que poderão ser percorridas pelos processos de publicação de dados estabelecidos no modelo.

5.2.2 Etapas e atividades complementares às BPLDs

Visando incorporar novas etapas e atividades necessárias ao modelo mas não contempladas pelas BPLDs, adicionalmente foram incorporadas mais duas etapas que visam desenvolver, na fase de **iniciação**, o estabelecimento das informações necessárias à escolha de qual processo de publicação deve ser escolhido a partir das opções apresentadas no modelo a depender da maturidade da instituição publicadora.

Além disso, na fase de **encerramento**, inspirado na atividade de "Planejamento da Próxima Fase" da espiral de Boehm (1986) é proposta uma etapa de retrospectiva e avaliação para as próximas atividades a serem adotadas no aprimoramento dos dados abertos que estão sendo publicados, conforme descrito abaixo:

- Fase – Iniciação
  - **Etapa - Identificar maturidade da instituição publicadora de dados abertos**
- Fase – Encerramento
  - **Etapa – Fazer Retrospectiva e avaliar a continuidade das atividades de publicação de dados**

5.2.2.1 Etapa: "Identificar maturidade da instituição publicadora de dados abertos"

A nova etapa **"Identificar maturidade da instituição publicadora de dados abertos"** conterá uma atividade inicial do modelo de processo com o objetivo de identificar qual a experiência da instituição governamental bem como da sua equipe em publicação de dados. Devem ser identificados se a instituição publica dados periodicamente,

se publica em diversos formatos, que tipos de dados e informações costuma publicar, se possui um setor específico para produção e gestão de dados e informações. Para esta atividade, pode ser adotado um questionário ou entrevista conforme proposta apresentada no Apêndice A.

5.2.2.2 Etapa: "Fazer Retrospectiva e avaliar a continuidade das atividades de publicação de dados"

Para esta nova etapa são propostas duas novas atividades: (i) Fazer retrospectiva e (ii) Tomar decisão sobre a continuidade do processo.

Para a atividade **"Fazer Retrospectiva"**, como referência, podem ser desenvolvidas algumas tarefas com características similares a *"Sprint Retrospective"* utilizada no método de desenvolvimento ágil de software *Scrum*. Segundo Varaschim (2009), no Scrum, esta atividade consiste de uma reunião realizada após cada *"Sprint de desenvolvimento"* utilizada para a correção dos problemas detectados pelo time no processo de desenvolvimento, bem como para a implementação de melhorias a partir de sugestões.

Para a atividade de **Tomar decisão sobre a continuidade do processo**, o publicador de dados, após a realização da retrospectiva, deverá dispor de informações sobre a qualidade das atividades, dos problemas e respectivas causas, bem como o produto gerado no ciclo, devendo desenvolver as tarefas abaixo:

- Analisar todos os possíveis problemas identificados ao longo das fases de planejamento, execução e controle;
- Buscar a identificação das causas destes problemas, bem como as possíveis soluções.
- Mediante esta análise:
  - Encerrar o trabalho de publicação neste ciclo ou;
  - Reiniciar o ciclo, desenvolvendo novamente as atividades do processo relacionado ao ciclo em que o trabalho de publicação se encontra, ou;
  - Avançar para o próximo ciclo, desenvolvendo as atividades referentes ao próximo processo estabelecido pelo modelo de processo.

Desta maneira, considerando a espiral apresentada na Figura 42, é proposta a distribuição destas 12 etapas (10 BPLDs + 2 novas etapas propostas) dentre as grandes fases desta espiral conforme apresentado na Tabela 12.

OBS.: Para facilitar a correlação das etapas do modelo *"Piece of Cake"*, foram mantidas a numeração ordinal das BPLDs. Por este motivo, a etapa "Identificar nível de maturidade da organização em publicação de dados" recebeu o número de ordem 0 (predecessor às BPLDs) e a etapa de "Fazer retrospectiva e avaliar a continuidade das atividades de publicação de dados" a ordem 11 (sucessora às BPLDs).

Tabela 12 – Sequenciamento de fases e etapas para publicação de dados abertos e dados abertos conectados governamentais

Fases:	**Etapas:**
I.Iniciação	0. Identificar nível de maturidade da organização em publicação de dados
II.Planejamento	1. Preparar partes interessadas 2. Selecionar conjuntos de dados
III. Execução e IV. Controle	3. Modelar os dados 4. Especificar licenças apropriadas 5. Estabelecer bons identificadores universais (URIs) para dados conectados 6. Utilizar vocabulários padrão 7. Converter e enriquecer dados 8. Prover acesso automatizado aos dados
V. Encerramento	9. Anunciar os novos conjuntos de dados para o público 10. Estabelecer um contrato social para os dados publicados 11. Fazer retrospectiva e avaliar a continuidade das atividades de publicação de dados

Fonte: Autor desta dissertação, 2015.

Assim, considerando a incorporação das 3 novas atividades complementares propostas às 70 atividades decorrentes da revisão de literatura, o modelo de processo *"Piece of Cake"* passou a dispor do seguinte conjunto de atividades distribuídas entre 12 etapas e agrupadas por cinco grandes fases conforme apresentado pela Tabela 13:

Tabela 13 – Atividades propostas pelo modelo de processo *"Piece of Cake"*

Etapas:	**Atividades:**
0. Identificar nível de maturidade da organização em publicação de dados	Identificar nível de maturidade da organização em publicação de dados (0A)
1. Preparar partes interessadas (*stakeholders*)	Identificar as partes interessadas (1A) Identificar os benefícios para a abertura de dados (1B) Definir perfis profissionais a serem envolvidos (1C) Definir grupos de usuários dos dados (1D) Elaborar um plano de ações para publicação dos dados (1E) Capacitar os envolvidos na publicação dos dados (1F)

2. Selecionar conjuntos de dados	Analisar a estrutura organizacional da instituição publicadora (2A) Estabelecer diretrizes que orientem a priorização da publicação de dados abertos (2B) Realizar consultas aos usuários sobre a demanda de dados (2C) Identificar os dados que serão abertos (2D) Definir nível de maturidade dos dados a serem publicados (1-5 estrelas) (2E) Analisar o nível de sigilo dos dados e informações (2F) Analisar relatórios anuais e documentações da instituição publicadora (2G) Analisar o esforço para abertura de dados (2H) Fazer e validar mapa de responsabilidades entre conjuntos de dados e unidades de negócio responsáveis (2I) Identificar e analisar sistemas de informação que poderão ser objeto da abertura de dados (2J) Identificar dados que podem ser conectados (2K)
3. Modelar os dados	Gerar cópias de segurança das bases de dados que serão abertas (3A) Higienizar os dados (3B) Estabelecer rotinas de conversão de dados para formatos legíveis por máquina (3C) Anonimizar dados sensíveis (3D) Modelar rotinas automatizadas (ETL) (3E) Analisar se os dados serão conectados ou não (3F) Estabelecer ou aprimorar documentação de dados (esquemas, vocabulários e ontologias) (3G)
4. Especificar uma licença apropriada	Adotar licenças de uso dos dados não restritivas (4A) Apresentar opções de licenças de dados a serem adotadas (4B) Estabelecer questões-chave para definição de licenças (4C)

5. Estabelecer bons identificadores universais (URIs) para dados conectados	Utilizar URIs para conectar os dados (5A) Estabelecer URIs persistentes, que não se alterem em nenhum momento (5B) Proporcionar pelo menos um recurso de dados em formato que seja legível por máquina para cada URI (5C) Usar URIs como nomes para as coisas (5D) Estabelecer design simplificado de URIs (5E) Utilizar identificadores relacionados a informações do mundo real (5F) Usar URIs HTTP para que recursos de dados possam ser encontrados via *Web* por pessoas e máquinas (5G) Estabelecer URIs neutras (5H) Utilizar datas em URIs com moderação (5I) Utilizar hashs (#) em URIs cautelosamente (5J) URIs das entidades (conjuntos de dados ou recursos) sejam diferentes das URIs das páginas que apresentam estes recursos para a leitura feita por humanos (5K)
6. Utilizar vocabulários padrão	Estabelecer os metadados obrigatórios (6A) Criar um esquema de dados para cada conjunto de dados (6B) Incentivar o reúso de vocabulários (6C) Publicar esquemas de dados em arquivos diferentes (6D) Determinar linguagens para expressar esquemas de dados (6E) Estabelecer critérios de escolha de vocabulários (6F) Certificar que os dados estão conectados a outros conjuntos de dados (6G) Desenvolver ou utilizar ontologias para estruturar a semântica dos dados (6H)

7. Converter e enriquecer dados	Converter dados para múltiplas finalidades e usos (7A) Adotar rotinas ETL para enriquecimento de dados (7B) Conectar conjuntos de dados com outros dados relacionados (7C) Permitir o envolvimento de várias pessoas na identificação de como os dados a serem convertidos se relacionam com outros dados (7D) Utilizar rotinas automatizadas de conversão de dados, como a triplificação, quando possível (7E) Converter dados em várias serializações RDF (7F)
8. Prover acesso automatizado aos dados	Disponibilizar bases completas para *download* (*dumps*) (8A) Estabelecer um Mapa de Decisões Tecnológicas (8B) Desenvolver uma API (8C) Desenvolver um *endpoint* SPARQL (8D)
9. Anunciar os conjuntos de dados para o público	Estabelecer dados tecnicamente e legalmente abertos (9A) Publicar metadados junto aos dados (9B) Disponibilizar os dados com o menor custo possível ao usuário, preferencialmente de modo gratuito na internet (9C) Divulgar dados em meios complementares (Catálogos, FTP, Torrent) (9D) Divulgar dados em seções destacadas de sítios de governo (9E) Estabelecer recursos de consulta parcial da base de dados como uma API ou *Web*service (9F) Estabelecer visualizações e demais recursos de exploração dos dados (9G) Melhorar os dados para serem melhor divulgados e encontrados por máquinas (9H) Disponibilizar dados conectados em servidores de triplas (9I)

10.Estabelecer um contrato social para os dados publicados	Estabelecer com clareza que o processo de publicação contempla etapas de manutenção e atualização dos dados (10A) Estabelecer mecanismos de monitoramento e avaliação da oferta de dados disponibilizados ao público (10B) Disponibilizar leis e atos normativos que explicitem aos usuários quanto às obrigações dos governos em publicarem dados com qualidade e disponibilidade (10C) Estabelecer espaços para recebimento do feedback do usuário, preferencialmente publicando dados de uma pessoa e/ou telefone de contato para esclarecimento de dúvidas sobre o uso e disponibilidade dos dados (10D) Utilizar tecnologias que mantenham os dados conectados disponíveis, atualizados e abertos (10E)
11. Fazer Retrospectiva e avaliar a continuidade das atividades de publicação de dados	Fazer retrospectiva (11A) Tomar decisão sobre a continuidade do processo (11B)

Fonte: Autor desta dissertação, 2015.

A próxima subseção descreverá o detalhamento dos processos do modelo, onde serão apresentados os conjuntos de atividades a serem desenvolvidos de acordo com o nível de maturidade da instituição publicadora ou a maturidade desejada para a publicação de dados abertos governamentais.

5.2.3 Detalhamento do modelo de processo *"Piece of Cake"*

Considerando as atividades estabelecidas conforme a Tabela 13, para a proposição do modelo de processo foi necessário desenvolver duas atividades de classificação, com os objetivos de:

1. Propor uma distribuição das atividades identificadas entre os processos A, B, C e D estabelecidos pelo modelo;

2. Definir, mediante a avaliação de especialistas, as atividades obrigatórias e desejáveis para cada um dos quatro processos estabelecidos.

5.2.4 Distribuição de atividades por processos do modelo (ciclos evolutivos)

Visando contemplar o requisito de adequar às atividades de publicação de dados abertos e dados abertos conectados aos três níveis finais do Esquema 5-Estrelas dos Dados Abertos, o modelo *"Piece of Cake"* buscou distribuir as atividades entre os níveis 3,4,5 deste Esquema. Ainda foi proposto, por decisão do pesquisador, duas atividades para o nivel 5-Estrelas aprimorado, como segue:

- Nível 3-Estrelas: Atividades necessárias para a publicação de dados em formato aberto, atendendo as características do nível 3-Estrelas;
- Nível 4-Estrelas: Atividades necessárias para o enriquecimento de dados do nível 3 para 4-Estrelas. É composto principalmente de atividades sobre estabelecimento de URIs e uso e reúso de vocabulários.
- Nível 5-Estrelas: Atividades necessárias para a produção de dados conectados (nível 5-Estrelas). Apresenta atividades específicas para dados conectados, especialmente no que tange a conversão de dados para dados conectados bem como o provimento de acesso automatizado aos dados;
- Nível 5-Estrelas aprimorado: Atividades voltadas ao aprimoramento da oferta de dados conectados, mediante o uso de ontologias e vários tipos de serializações RDF.

Esta classificação, que será apresentada na seção 6.1 do capítulo sobre a validação, foi decorrente da aplicação de um questionário simplificado com especialistas na temática. A sumarização dos resultados desta classificação será apresentada conjuntamente na próxima subseção, na Tabela 14.

5.2.5 Distribuição de atividades por nível de maturidade da instituição

Visando contemplar o requisito de adequar às atividades de publicação de dados abertos e dados abertos conectados ao nível de maturidade da instituição governamental, o modelo *"Piece of Cake"* classificou as atividades em dois níveis de maturidade:

- Obrigatórias: Atividades que devem ser executadas por qualquer instituição publicadora de dados abertos (conectados) governamentais, independente do nível de maturidade. Aplicáveis principalmente para instituições que possuem baixa maturidade em publicação de dados e desejam um conjunto sucinto de atividades para disponibilizar seus dados abertos aos seus usuários.
- Desejáveis: Atividades que devem ser executadas por instituições com nível mais alto de maturidade. Este grupo de atividades busca melhorar significativamente a oferta de dados mediante a incorporação de requisitos de maior qualidade, disponibilidade e conformidade dos dados que serão disponibilizados.

Esta classificação, que será apresentada na seção 6.2 do capítulo sobre a validação, foi um resultado integrado dos dois instrumentos de validação empírica. Considerando apenas as atividades extraídas da revisão de literatura, no modelo de regressão, foram identificadas 33 atividades obrigatórias e 37 desejáveis. Já no estudo empírico, foram identificadas 12 atividades obrigatórias. O quantitativo de atividades por nível de maturidade está disponível na tabela 14.

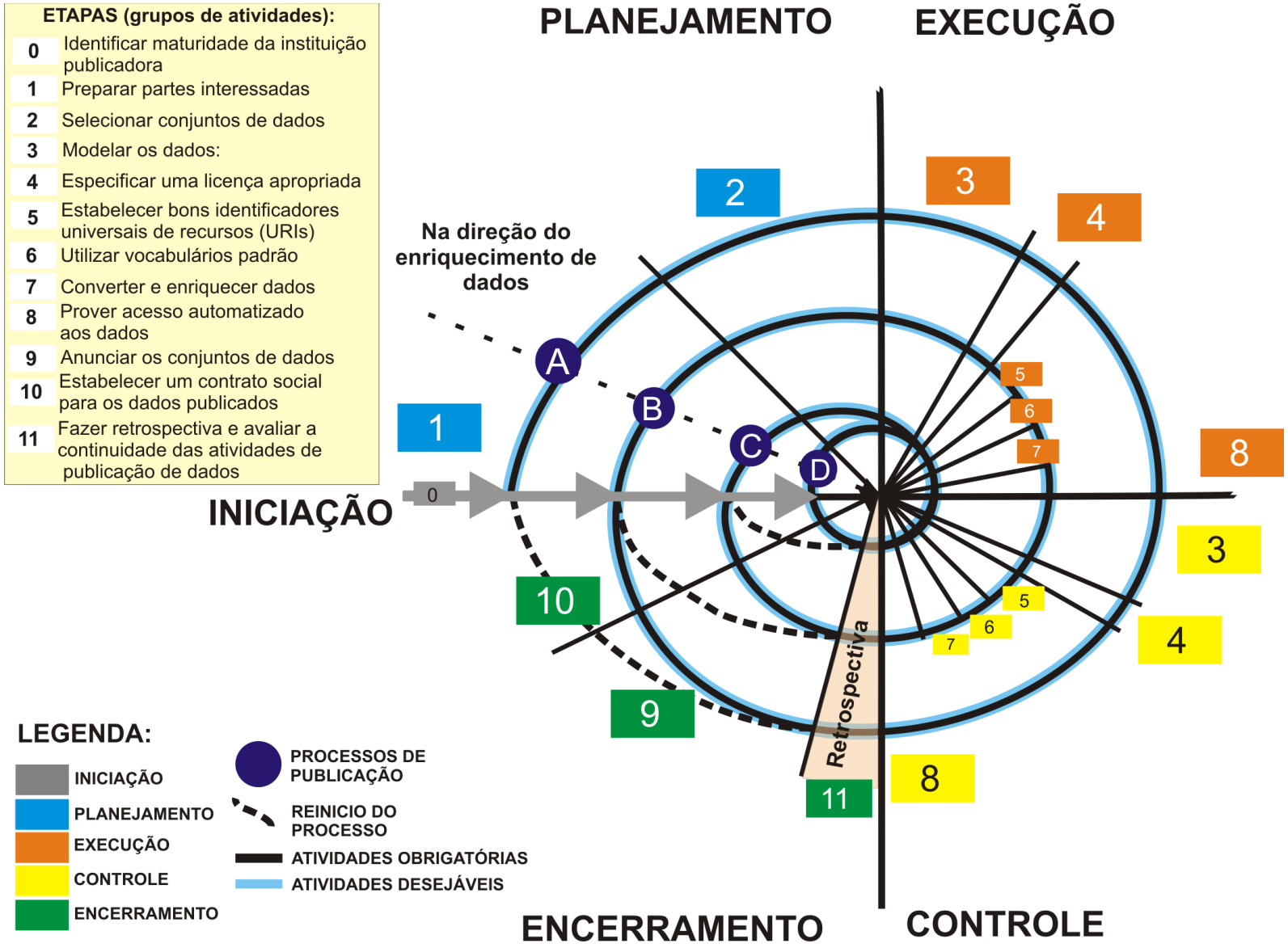
Tabela 14 – Total das atividades obrigatórias e desejáveis entre os processos do modelo

	Obrigatórias	**Desejáveis**	**Total**
Processo A (Para 3 estrelas):	21	18	**39**
Processo B (Para 4 estrelas):	12	10	**22**
Processo C (Para 5 estrelas):	3	4	**7**
Processo D (Para 5 estrelas-aprimorado):	1	1	**2**
Subtotal:	**37**	**33**	**70**
Comum para todos os processos:[1]	3	0	3
Total:	**39**	**34**	**73**

Fonte: Autor desta dissertação, 2015.

Assim, foi possível desenvolver a Figura 44 que apresenta uma visão integrada e detalhada da Espiral contendo as 12 etapas do modelo *"Piece of Cake"* conforme exposto na Tabela 12. A próxima subseção descreverá os passos genéricos para a utilização do modelo.

Figura 44 – Distribuição das etapas nas fases da espiral do modelo *"Piece of Cake"*



Fonte: Autor desta dissertação, 2015.

5.2.6 Passos genéricos do modelo

Entendemos como passos genéricos do modelo as macroatividades que devem ser desenvolvidas para a sua execução. Considerando a classificação de atividades por nível de maturidade apresentada, caso **a instituição tenha baixa maturidade em publicação de dados abertos, recomenda-se apenas o desenvolvimento das atividades obrigatórias**. As instituições com maior nível de maturidade poderão desenvolver todas as atividades (obrigatórias e desejáveis) visando uma melhor qualidade do resultado final da publicação de dados.

Estabelecida esta distribuição de atividades, os processos devem ser desenvolvidos conforme os seguintes passos genéricos:

- **Passo 1 - Identificar o nível de maturidade da instituição publicadora:** É o passo inicial, que visa avaliar a maturidade em publicação de dados da instituição que utilizará o modelo. Decorrente desta avaliação o modelo irá estabelecer se a instituição deverá desenvolver apenas as atividades obrigatórias ou se também deverá desenvolver as atividades desejáveis.

- **Passo 2 – Verificar as atividades a serem implementadas:** Dependendo do nível de maturidade da instituição, deverão ser implementadas atividades obrigatórias (atividades de menor complexidade, voltadas para instituições de baixa maturidade) e desejáveis (atividades de maior complexidade, voltadas para instituições de maior maturidade em publicação de dados)

- **Passo 3 – Seleção do processo de publicação:** Escolher o processo de publicação conforme o nível de maturidade desejada para a publicação de dados, conforme os níveis do Esquema 5-Estrelas.

Dependendo do nível de publicação de dados desejado, deverá ser escolhido um dos processos abaixo:

5.2.6.1 Processo A: Voltado à publicação de novos dados abertos no nível três estrelas

- Na fase de **planejamento**, devem ser desenvolvidas atividades específicas para este processo conforme proposto na Tabela 15, especialmente a atividade **"Definir nível de maturidade dos dados a serem publicados (1-5 estrelas) (2E)"**, que irá estabelecer até qual nível do Esquema 5-Estrelas o modelo de processo deve guiar a instituição publicadora.

- Na fase de **execução**, devem ser desenvolvidas atividades específicas para este processo conforme proposto na Tabela 15. Caso aconteça alguma excepcionalidade ou impedimento na execução de alguma atividade, esta deverá ser devidamente registrada.

- Na fase de **controle**, verificar se as atividades desenvolvidas na fase de execução deste processo foram desenvolvidas adequadamente. Em caso de excepcionalidades ou impedimentos, deverão ser analisadas as respectivas causas, com atenção especial se existem causas comuns para vários problemas.

- Na fase de **encerramento**, desenvolver inicialmente as atividades da etapa de "Fazer retrospectiva e avaliar a continuidade das atividades de publicação de dados". Caso existam problemas não-solucionáveis nesta execução do processo, recomenda-se o reinicio do mesmo. Caso seja decidido aprimorar o nível dos dados que serão publicados para 4-Estrelas, desenvolver as atividades relacionadas a "Anunciar os novos conjuntos de dados para o público", "Estabelecer um contrato social para os dados publicados" e avançar para o Processo B.

O conjunto de atividades obrigatórias e desejáveis para este processo está disponível na Tabela 15.

Tabela 15 – Atividades propostas para o processo A (Publicação de Dados Abertos - 3 estrelas)

Etapas:	**Obrigatórias**	**Desejáveis**
ETAPA 1:	1A, 1B, 1C, 1D	1E,1F
ETAPA 2:	2A, 2B, 2C, 2D, 2E, 2H	2F, 2G, 2I, 2J
ETAPA 3:	3A, 3B, 3C	3D, 3E
ETAPA 4:	4A	4B, 4C
ETAPA 8:	8A, 8B	8C
ETAPA 9:	9A, 9B, 9C	9D, 9E, 9F, 9G
ETAPA 10:	10A, 10B	10C, 10D, 10E
ETAPA 11:	11A, 11B	-

Fonte: Autor desta dissertação, 2015.

5.2.6.2 Processo B: Voltado à publicação de novos dados abertos no nível quatro estrelas

- Caso seja uma publicação de novos dados abertos, inicialmente devem ser desenvolvidas as atividades do Processo A. Caso seja um enriquecimento de dados abertos, sugere-se a verificação se as atividades do Processo A foram desenvolvidas para os dados que serão enriquecidos;

- Na fase de **planejamento**, recomenda-se verificar se haverá necessidade de aplicar alguma atividade do ciclo anterior (Processo A) neste ciclo (Processo B). Posteriomente devem ser desenvolvidas atividades específicas para este processo conforme proposto na Tabela 16;

- Na fase de **execução**, recomenda-se verificar se haverá necessidade de aplicar alguma atividade do ciclo anterior (Processo A) neste ciclo (Processo B). Posteriomente

devem ser desenvolvidas atividades específicas para este processo conforme proposto na Tabela 16;

- Na fase de **controle**, verificar se as atividades desenvolvidas na fase de execução deste processo foram desenvolvidas adequadamente. Em caso de excepcionalidades ou impedimentos, deverão ser analisadas as respectivas causas, com atenção especial se existem causas comuns para vários problemas;

- Na fase de **encerramento**, desenvolver inicialmente as atividades da etapa de "Fazer retrospectiva e avaliar a continuidade das atividades de publicação de dados". Caso existam problemas não-solucionáveis nesta execução do processo, recomenda-se o reinicio do mesmo. Caso seja decidido aprimorar o nível dos dados que serão publicados para 5-Estrelas, desenvolver as atividades relacionadas a "Anunciar os novos conjuntos de dados para o público", "Estabelecer um contrato social para os dados publicados" e avançar para o Processo C.

O conjunto de atividades obrigatórias e desejáveis para este processo está disponível na Tabela 16.

Tabela 16 – Atividades propostas para o processo B (Publicação de Dados Abertos - 4 estrelas)

Etapas:	**Obrigatórias**	**Desejáveis**
ETAPA 2:	2K	-
ETAPA 3:	3F, 3G	-
ETAPA 5:	5A, 5B, 5C, 5D	5E, 5F, 5G, 5H, 5I, 5J, 5K
ETAPA 6:	6A, 6B, 6C, 6D, 6F	6E
ETAPA 7:	7A	7B
ETAPA 11:	11A, 11B	-

Fonte: Autor desta dissertação, 2015.

5.2.6.3 Processo C: Voltado à publicação de novos dados abertos no nível cinco estrelas

- Caso seja uma publicação de novos dados abertos, inicialmente devem ser desenvolvidas as atividades do Processo A e B. Caso seja um enriquecimento de dados abertos, sugere-se a verificação se as atividades do Processo A e/ou B foram desenvolvidas para os dados que serão enriquecidos;

- Na fase de **planejamento**, recomenda-se verificar se haverá necessidade de aplicar alguma atividade dos ciclos anteriores (Processos A e B) neste ciclo (Processo C);

- Na fase de **execução**, recomenda-se verificar se haverá necessidade de aplicar alguma atividade dos ciclos anteriores (Processos A e B) neste ciclo (Processo C). Posteriomente devem ser desenvolvidas atividades específicas para este processo conforme proposto na Tabela 17;

- Na fase de **controle**, verificar se as atividades desenvolvidas na fase de execução deste processo foram desenvolvidas adequadamente. Em caso de excepcionalidades ou impedimentos, deverão ser analisadas as respectivas causas, com atenção especial se existem causas comuns para vários problemas;

- Na fase de **encerramento**, desenvolver inicialmente as atividades da etapa de "Fazer retrospectiva e avaliar a continuidade das atividades de publicação de dados". Caso existam problemas não-solucionáveis nesta execução do processo, recomenda-se o reinicio do mesmo. Caso seja decidido aprimorar o nível dos dados que serão publicados para 5-Estrelas aprimorado, desenvolver as atividades relacionadas a "Anunciar os novos conjuntos de dados para o público", "Estabelecer um contrato social para os dados publicados" e avançar para o Processo D.

O conjunto de atividades obrigatórias e desejáveis para este processo está disponível na Tabela 17.

Tabela 17 – Atividades propostas para o processo C (Publicação de Dados Abertos Conectados - 5 estrelas)

Etapas:	**Obrigatórias**	**Desejáveis**
ETAPA 6:	-	6G
ETAPA 7:	7C, 7D	7E
ETAPA 8:	-	8D
ETAPA 9:	9I	9H
ETAPA 11:	11A, 11B	-

Fonte: Autor desta dissertação, 2015.

5.2.6.4 Processo D: Voltado à publicação de novos dados abertos no nível cinco estrelas aprimorado

- Caso seja uma publicação de novos dados abertos, inicialmente devem ser desenvolvidas as atividades do Processo A, B e C. Caso seja um enriquecimento de dados abertos, sugere-se a verificação se as atividades do Processo A e/ou B e/ou C foram desenvolvidas para os dados que serão enriquecidos;

- Na fase de **planejamento**, recomenda-se verificar se haverá necessidade de aplicar alguma atividade dos ciclos anteriores (Processos A, B e C) neste ciclo (Processo D);

- Na fase de **execução**, recomenda-se verificar se haverá necessidade de aplicar alguma atividade dos ciclos anteriores (Processos A, B e C) neste ciclo (Processo D). Posteriomente devem ser desenvolvidas atividades específicas para este processo conforme proposto na Tabela 18;

- Na fase de **controle**, verificar se as atividades desenvolvidas na fase de execução deste processo foram desenvolvidas adequadamente. Em caso de excepcionalidades ou impedimentos, deverão ser analisadas as respectivas causas, com atenção especial se existem causas comuns para vários problemas;

- Na fase de **encerramento**, desenvolver inicialmente as atividades da etapa de "Fazer retrospectiva e avaliar a continuidade das atividades de publicação de dados". Caso existam problemas não-solucionáveis nesta execução do processo, recomenda-se o reinicio do processo sem adicionar novos requisitos de melhoria. Caso seja decidido aprimorar o nível dos dados que serão publicados para 5-Estrelas aprimorado, desenvolver as atividades relacionadas a "Anunciar os novos conjuntos de dados para o público", "Estabelecer um contrato social para os dados publicados" e repetir para o Processo D com os novos requisitos de melhoria.

O conjunto de atividades obrigatórias e desejáveis para este processo está disponível na Tabela 18.

Tabela 18 – Atividades propostas para o processo D (Publicação de Dados Abertos Conectados - 5 estrelas aprimorado)

Etapas:	**Obrigatórias**	**Desejáveis**
ETAPA 6:	6H	-
ETAPA 7:	-	7F

Fonte: Autor desta dissertação, 2015.

Neste capítulo apresentamos a estruturação do modelo *"Piece of Cake"*, apresentando a sua visão geral e respectivos detalhamentos. No próximo capítulo serão apresentados os procedimentos adotados para a validação do modelo, especialmente quanto aos procedimentos adotados para validar a sua estruturação (fases, etapas e atividades).

6 VALIDAÇÃO

Este capítulo apresenta a estratégia de validação da pesquisa, mediante a aplicação de duas técnicas de investigação empírica, sendo (i) um modelo de regressão estabelecido através da aplicação de um questionário visando a validação da proposta de atividades obrigatórias e opcionais (desejáveis) para o modelo, e ainda (ii) um estudo empírico com o objetivo de validar a eficácia do modelo de processo proposto. Segundo Molina (2002), a validação empírica de uma teoria é uma componente fundamental de todas as teorias que aspirem a ser científicas, logo, é uma etapa muito relevante da pesquisa desenvolvida.

6.1 Classificação das atividades dentre os níveis do Esquema 5-Estrelas

Para estabelecer uma proposta de distribuição das atividades entre os 4 ciclos evolutivos do modelo de processo, foram reunidos 5 especialistas em publicação e consumo de dados abertos e dados abertos conectados e conhecimento sobre o Esquema 5-Estrelas dos Dados Abertos, que responderam um questionário online, onde, para cada uma das atividades propostas, deveria responder a partir de que nível do Esquema 5-Estrelas esta atividade é aplicável. Nesta classificação, foram avaliadas apenas as atividades derivadas das recomendações extraídas da revisão de literatura.

A seguir é apresentado o planejamento deste questionário e análise dos seus resultados.

6.1.1 Planejamento do questionário de distribuição de atividades por processos do modelo

Para esta atividade o questionário foi estruturado conforme o seguinte *template*:

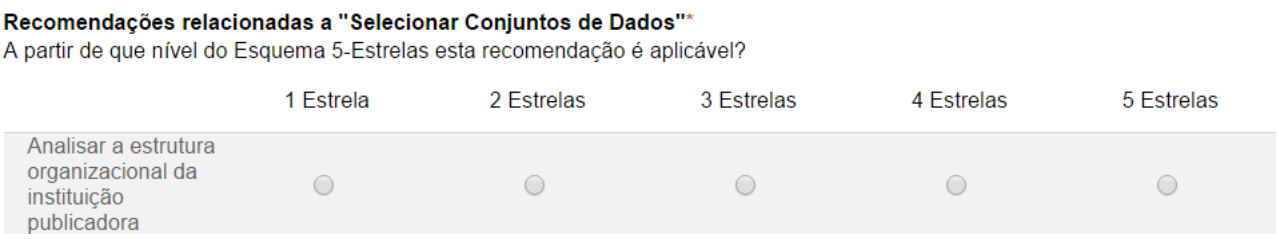
- **Objetivo do questionário:** Propor a distribuição das 70 atividades do modelo de processo ao longo dos seus ciclos evolutivos
- **Tipo de questionário:** Questionário não supervisionado – realizado na *Web* (online, utilizando a ferramenta Google Forms)
- **Pesquisa da literatura relevante:** Atividades a serem classificadas são oriundas da revisão de literatura desta pesquisa;
- **Construção do Questionário:** Optou-se por desenvolver um novo questionário que contemplasse todas as recomendações extraídas da revisão
- **Método de Amostragem:** Foi adotado o método não-probabilístico de amostragem "Grupo Focal". Este método foi escolhido devido à população apta a responder o questionário ser muito específica, neste caso, pesquisadores sobre dados abertos, dados abertos conectados e conhecedores do Esquema 5-Estrelas dos Dados Abertos.

Além disso, devido ao prazo necessário para esta atividade de classificação, para o resultado ser alcançado foi necessário solicitar que pesquisadores da rede de contatos respondessem o questionário.

- **Tamanho da Amostra:** Foram selecionados 5 pesquisadores que possuem experiência teórica e/ou prática no assunto dados abertos e dados abertos conectados
- **Demais informações relevantes:**
  - Questionário composto por 70 itens de avaliação. Cada item foi avaliado mediante uma mesma pergunta ("A partir de que nível do Esquema 5- Estrelas esta recomendação é aplicável?"). Todas as respostas foram fechadas, com o mesmo padrão de resposta mediante uma escala likert conforme apresentado na Figura 45.
  - Apesar do número grande de perguntas, tratou-se de um questionário aderente a atuação dos pesquisadores
  - O questionário foi padronizado, sendo estabelecido, para o objeto principal de avaliação (recomendações)
  - Para motivar os avaliadores, houve uma explicação prévia dos objetivos do questionário, apresentando a relevância dos resultados para a área de pesquisa.
  - O questionário deveria ser respondido de forma anônima. Entretanto, ao final, o avaliador respondeu algumas questões relacionadas à qualidade e eficácia do questionário, bem como uma auto-avaliação do seu conhecimento com as temáticas abordadas.
  - O questionário utilizado nesta atividade está disponível no Apêndice B.

A Figura 45 apresenta um exemplo de questão disposta neste questionário.

Figura 45 – Exemplo de questão utilizada para classificar as atividades nos ciclos evolutivos do modelo de processo



Fonte: Autor desta dissertação, 2015.

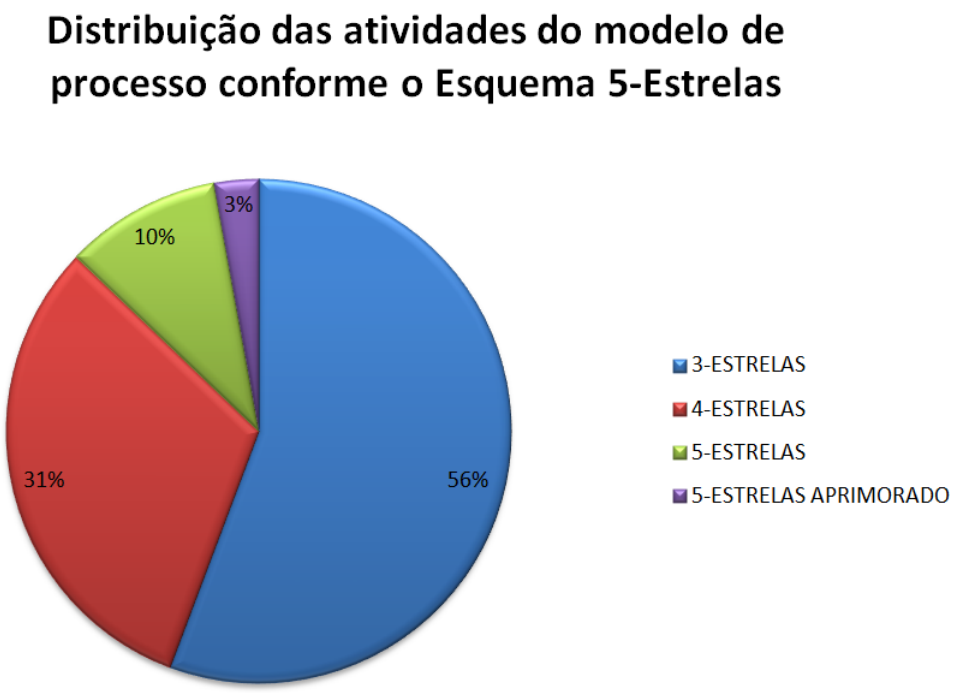
6.1.1.1 Análise dos resultados

A consolidação dos resultados do questionário foi desenvolvida mediante os seguintes critérios:

- Considerando o requisito deste modelo de processo, que considera um dado aberto a partir do nível 3 do Esquema 5-Estrelas, as respostas para os níveis 1, 2 ou 3 estrelas foram classificadas como uma atividade aplicável a partir do nível 3, que é o nível mínimo do modelo de processo;
- Cada atividade foi associada ao nível que recebeu o maior percentual de respostas. No caso de empates entre respostas de mais de um nível, a atividade foi associada ao nível mais baixo;
- Duas atividades foram selecionadas pelo pesquisador para integrar o processo D, que permite o aprimoramento de dados que já estão no nível 5 estrelas;

Como resultado desta consolidação as atividades foram distribuídas nos níveis 3, 4, 5 estrelas além das atividades que foram selecionadas pelo pesquisador para o nível 5-estrelas aprimorado. O gráfico abaixo mostra o total de atividades classificadas em cada nível:

Figura 46 – Percentual de distribuição das atividades por ciclo evolutivo do modelo de processo



Fonte: Autor desta dissertação, 2015.

Esta classificação permitiu o aprimoramento do modelo de processo, mediante a distribuição das etapas (incluindo as BPLDs) dentre as grandes fases e os quatro processos (A, B, C e D) da espiral. A próxima seção apresentará os critérios adotados para a distribuição das atividades nos dois níveis de maturidade estabelecidos pelo modelo de processo.

6.2 Modelo de Regressão e questionário para classificação do nível de dificuldade e relevância associado às atividades do modelo de processo

Para obtenção da validade de atividades que o modelo *"Piece of Cake"* deveria propor como obrigatórias, foi desenvolvido um modelo de regressão a partir de dados coletados da aplicação de um questionário. As próximas subseções explicarão este procedimento.

6.2.1 Questionário para classificação do nível de dificuldade e relevância das recomendações extraídas da revisão de literatura

Conforme exposto na fundamentação teórica e nos trabalhos relacionados, os processos de publicação de dados abertos (conectados) governamentais propostos costumam aparentemente não consideram o nível de maturidade das instituições publicadoras, ou seja, o mesmo processo deve ser aplicado independente da experiência da instituição e da equipe envolvida nas atividades de publicação.

Buscando propor uma solução para esta problemática, esta pesquisa buscou estabelecer uma técnica para classificação, junto a especialistas, do nível de dificuldade e relevância do conjunto de recomendações para publicação de dados abertos e dados abertos conectados que foram extraídos da revisão de literatura. Para atingir este objetivo, um questionário foi escolhido como ferramenta.

Segundo Encinosa (2006), os questionários são ferramenas muito úteis para obtenção de informações. São utilizados especialmente quando a quantidade de pessoas a se investigar é muito grande e/ou quando existe uma grande dispersão geográfica dos respondentes impedindo, fisica e economicamente, a realização de entrevistas.

Para proporcionar a confiança necessária à esta atividade de classificação, fez-se necessária a aplicação do questionário com um número amplo de especialistas em pesquisa, publicação e consumo de dados abertos e dados abertos conectados. Este requisito justifica a adoção desta ferramenta, por haver dispersão geográfica dos especialistas, bem como pela necessidade de um número razoável de respondentes.

Encinosa (2006) complementa que um questionário deve possuir cinco atributos básicos que são: idoneidade, ser processável automaticamente, ser objetivo, ter um alcance razoável e ainda, possuir um design agradável. Este questionário buscou atender estes cinco atributos conforme será explanado na subseção de planejamento do questionário.

Ademais, o autor destaca a importância do planejamento das etapas para execução do questionário, devendo haver uma preparação, elaboração, prova, aplicação e análise dos resultados.

Desta maneira, a pesquisa estabeleceu como ferramenta de coleta de dados um questionário contendo o conjunto de recomendações para publicação de dados abertos e dados abertos conectados extraídas na revisão da literatura, estruturado conforme a subseção a seguir. Após propor uma versão inicial do questionário, para as atividades de pla-

nejamento, elaboração e análise dos resultados deste questionário, o pesquisador contou com o apoio de outros três pesquisadores, sendo uma pesquisadora com experiência no planejamento e desenvolvimento de questionários, uma outra pesquisadora com amplo conhecimento em análise estatística e um pesquisador com ótimo domínio do idioma inglês que atuou na revisão da tradução das questões para este idioma.

6.2.1.1 Planejamento do questionário

Para esta atividade o questionário foi estruturado inicialmente conforme a seguinte estrutura:

- **Objetivo do questionário:** Classificar, quanto a sua dificuldade e relevância, 70 recomendações para publicação de dados abertos e dados abertos conectados no setor público;
- **Tipo de questionário:** Questionário não supervisionado – disponível na *Web* (*online*, utilizando a ferramenta *Google Forms*). Esta ferramenta permite a tabulação e alguns resultados preliminares gerados de forma automática;
- **Pesquisa da literatura relevante:** Recomendações classificadas são oriundas da revisão de literatura da pesquisa;
- **Construção do Questionário:** Optou-se por desenvolver um novo questionário que contemplasse todas as recomendações extraídas da revisão de literatura. Não havia questionário similar existente. O questionário é inédito e foi desenvolvido pelo pesquisador. O questionário foi desenvolvido inicialmente no idioma português. Após passar pelo pré-teste, foi elaborada uma versão no idioma inglês, já considerando as sugestões dos avaliadores do pré-teste.
- **Pré-Teste:** Uma versão inicial do questionário foi aplicada num estudo-piloto com representantes do público-alvo desejado pelo pesquisador para que respondesse o questionário. Fizeram o pré-teste:
  - 1 publicador de Dados Abertos Governamentais
  - 1 publicador de Dados Abertos Conectados
  - 1 consumidor de Dados Abertos Governamentais
  - 1 pesquisador sobre Dados Abertos e Dados Abertos Conectados
- **População do Questionário:** Foram estabelecidos como integrantes da população deste questionário os integrantes das listas de discussão da Infraestrutura Nacional de Dados Abertos (INDA-Br), Parceria para o Governo Aberto – Brasil, Participantes do I Concurso de Aplicativos Apps.Gov - SBTI 2014, além de todos

os pesquisadores autores dos artigos que foram utilizados na revisão de literatura bem como alguns contatos do orientador que desenvolvem pesquisas relacionadas a temática da investigação;

- **Método de Amostragem:** Foi adotado o método probabilístico de amostragem *"Simple Random Sample"* [1], onde cada membro da população tem a mesma probabilidade de ser escolhido. O questionário foi enviado via formulário online, onde qualquer membro desta população poderia respondê-lo;

Para motivar os avaliadores, houve uma explicação prévia dos objetivos do questionário, apresentando a relevância dos resultados para a área de pesquisa. Ademais, o questionário deveria ser respondido de forma anônima.

6.2.1.2 Elaboração e Execução do questionário

A partir do objetivo do questionário, o pesquisador elaborou uma versão inicial composta por três seções: (i) perfil etário e demográfico do avaliador, (ii) nível de entendimento do avaliador sobre os temas dados abertos e dados abertos conectados e (iii) a classificação das recomendações. Neste último, o objetivo específico consistiu na classificação de 70 recomendações (itens de avaliação) quanto a sua dificuldade e relevância.

Inicialmente, na seção (iii) do questionário, cada um dos 70 itens de avaliação deveriam ser avaliados quanto a sua complexidade, mediante uma escala *likert* de 4 níveis (baixa, média, alta e muito alta). Para cada seção foi apresentada uma página, onde, a página da seção (iii) ficou longa devido ao quantitativo de recomendações.

A aplicação do pré-teste permitiu a implementação de melhorias relevantes no questionário como:

- Agrupamento das recomendações relacionadas a Dados Abertos e a Dados Abertos Conectados;
- Paginação da seção referente à classificação das recomendações, passando a dispor entre 15 a 20 recomendações por página;
- Melhoria do enunciado de algumas questões muito específicas;
- Os avaliadores registraram dificuldade quanto à clareza do entendimento sobre a complexidade do item de avaliação.

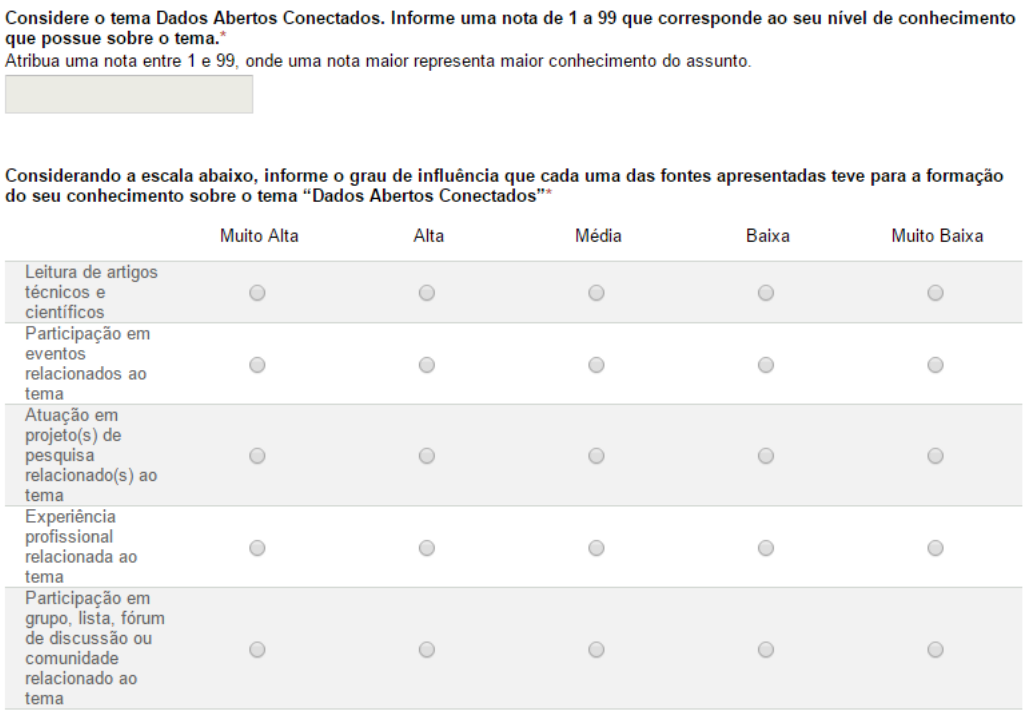
Além das sugestões do pré-teste, o pesquisador se reuniu com as pesquisadoras especialistas em questionário e análise estatística e foram estipuladas novas melhorias como:

[1] Mais informações disponíveis em: https://en.wikipedia.org/wiki/Simple_random_sample

- Ajuste no design do questionário, permitindo que o avaliador optasse por avaliar as recomendações de cada temática (Dados Abertos ou Dados Abertos Conectados). Caso desejasse, o avaliador poderia avaliar os dois subconjuntos. Desta maneira, os avaliadores foram exigidos a uma dedicação menor de tempo minimizando o risco de respostas enviesadas decorrentes do cansaço do avaliador;

- Aprimoramento das questões sobre o entendimento do avaliador sobre o tema. Foram adicionadas duas questões, sendo a primeira onde os avaliadores qualificam o seu conhecimento geral no assunto adotando um escore de 1 à 99 e uma outra questão multidimensional referente a origem do conhecimento do avaliador na temática, conforme apresentado na Figura 47.

Figura 47 – Itens de avaliação da experiência do avaliador



Considere o tema Dados Abertos Conectados. Informe uma nota de 1 a 99 que corresponde ao seu nível de conhecimento que possue sobre o tema.*

Atribua uma nota entre 1 e 99, onde uma nota maior representa maior conhecimento do assunto.

Considerando a escala abaixo, informe o grau de influência que cada uma das fontes apresentadas teve para a formação do seu conhecimento sobre o tema "Dados Abertos Conectados"*

	Muito Alta	Alta	Média	Baixa	Muito Baixa
Leitura de artigos técnicos e científicos	○	○	○	○	○
Participação em eventos relacionados ao tema	○	○	○	○	○
Atuação em projeto(s) de pesquisa relacionado(s) ao tema	○	○	○	○	○
Experiência profissional relacionada ao tema	○	○	○	○	○
Participação em grupo, lista, fórum de discussão ou comunidade relacionado ao tema	○	○	○	○	○

Fonte: Autor desta dissertação, 2015.

- Para cada temática, houve uma pré-seleção de um subconjunto de recomendações estipuladas pelo pesquisador como obrigatórias, onde os avaliadores deveriam atribuir um escore de 1 à 99 qualificando estes subconjuntos conforme descrito a seguir:

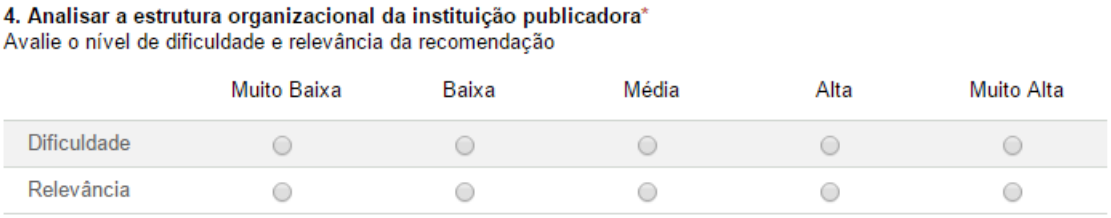
  - **Recomendações obrigatórias - Dados Abertos:** 1A,1B, 1C, 2B, 2D, 2E, 2H, 3A, 3B, 4A, 6A, 7A, 8A, 9A, 9B, 9H;
  - **Recomendações obrigatórias - Dados Abertos Conectados:** 3F, 5A, 5C, 5D, 7C.

- Inserção de uma questão final que solicita ao avaliador que informe um escore de 1 a 99 para avaliar o total de recomendações para dados abertos e o total de recomendações para dados abertos conectados;

- Aprimoramento da escala *likert*, de 4 para 5 níveis, sendo possível atribuir escores para cada nível da escala: Muito baixo = 0; Baixo = 25; Médio = 50; Alto = 75; Muito Alto = 100;

- Para cada item de avaliação, passou a ser avaliada a sua dificuldade e a relevância em substituição à complexidade, sendo possível obter mais dados para fins de análise estatística. A Figura 48 apresenta um exemplo de item de avaliação bem como a escala *likert* utilizada.

Figura 48 – Itens de avaliação quanto à dificuldade e relevância das recomendações



Fonte: Autor desta dissertação, 2015.

- O questionário utilizado nesta atividade está disponível no Apêndice C.

6.2.1.3 Análise de Ameaças à validade

Embora este questionário tenha sido realizado com o devido planejamento e testes visando minimizar possíveis ameaças a sua validade, as quais possam comprometer as conclusões, existem algumas que devem ser mencionadas:

- Cansaço dos avaliadores ao responder: O questionário, por ser extenso, pode ter levado ao cansaço de alguns avaliadores ao longo de sua resposta, o que pode ter prejudicado a qualidade de algumas respostas;

- Respostas emitidas por avaliadores inexperientes: Devido ao método de seleção dos avaliadores, é possível que um avaliador inexperiente responda o questionário. Entretanto, tal ameaça consegue ser tratada pelo escore que avalia o nível de conhecimento dos avaliadores (ScoreConhecimento) utilizado no modelo de regressão, que considera níveis mais altos de conhecimento dos avaliadores para sugerir as atividades mais relevantes ao modelo.

6.2.2 Análise Estatística

Para a realização desta análise foi possível obter 27 respostas referentes às recomendações para dados abertos e 24 respostas referentes às recomendações para dados abertos conectados. Para atingir o objetivo de classificar as recomendações mais relevantes, que serão estabelecidas como atividades obrigatórias para o modelo de processo *"Piece of Cake"*, foi desenvolvido um modelo de regressão conforme apresentado nas próximas subseções.

O modelo de regressão tem como objetivo apresentar um subconjunto das recomendações mais relevantes para a publicação de dados abertos e dados abertos conectados, levando em consideração a dificuldade e a relevância para implementar cada recomendação. Para alcançar tal objetivo é utilizado o modelo de regressão beta Ferrari e Cribari-Neto (2004). Desta maneira, foi fundamental a discussão entre o pesquisador e as pesquisadoras especialistas em estatística e elaboração de questionários em diversos sentidos, a saber:

- Inserção de uma variável capaz de medir conjuntamente a relevância e a dificuldade de cada recomendação. A partir desta variável foi possível obter o escore global de qualificação das recomendações sugeridas pelo modelo de regressão; Para essa variável foram atribuídos valores no intervalo $(1, 99)$, intervalo que foi transformado para o $(0, 1)$ com o objetivo da utilização do modelo de regressão beta;

- Definir uma escala *likert* de cinco níveis associada as variáveis **dificuldade** e **relevância** das recomendações. Em seguida foi possível transformar essas categorias em valores numéricos, de tal forma que os mesmos pudessem ser computados pelo modelo de regressão;

- Desta maneira, as variáveis **dificuldade** e **relevância** podem assumir valores entre 0 e 100. No caso da **dificuldade**, quanto menor a nota menor a dificuldade de implementar a recomendação e quanto maior a nota de **relevância** mais necessária é a recomendação para a publicação dos dados;

- Além disto, por recomendação estatística, foi inserido uma variável escore entre $(1, 99)$, onde o próprio avaliador do modelo de recomendações pode informar o seu grau de experiência em cada uma das duas temáticas avaliadas;

- Adicionalmente, visto que no questionário existem dois subconjuntos de recomendações obrigatórias pré-selecionadas pelo pesquisador para avaliar de forma global a qualidade destes subconjuntos, foi inserido no questionário, uma variável escore de qualidade para as recomendações obrigatórias de cada temática, variando também entre $(1, 99)$;

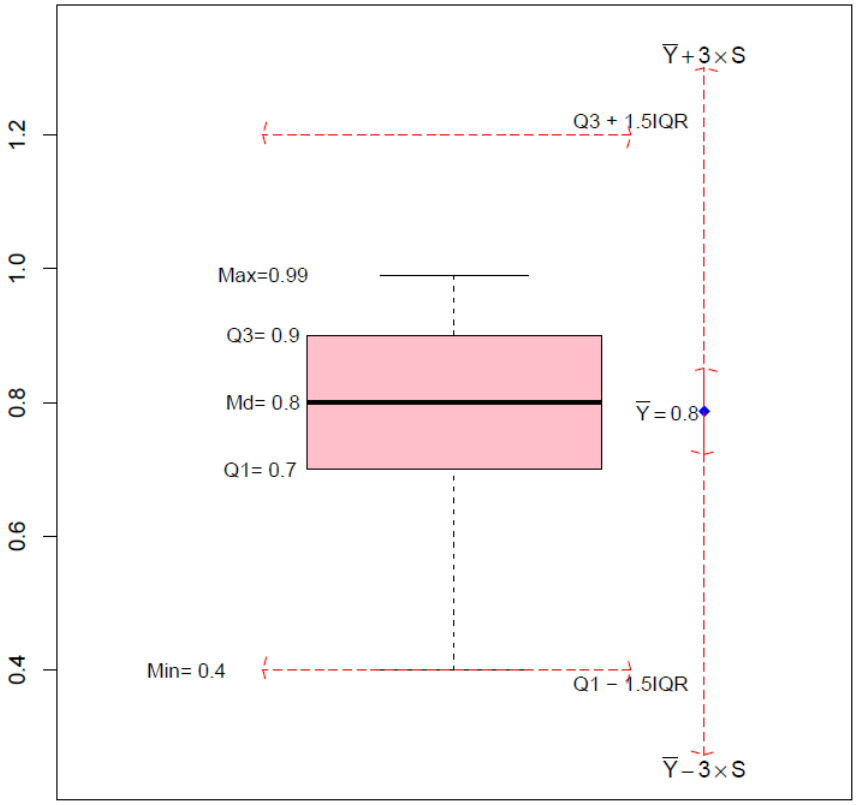
- Finalmente, a construção do questionário sofreu intervenção estatística em todos os seus aspectos, desde a forma mais adequada de arguir os avaliadores, inserção de novas variáveis relevantes e orientação de como mensurar adequadamente todas as

variáveis contidas no questionário, com o objetivo de permitir uma futura análise estatística ampla, robusta e confiável.

6.2.3 Estatística descritiva

Para a obtenção de respostas iniciais, foram construídos alguns gráficos referentes às variáveis envolvidas no modelo estatístico. As Figuras 49 e 50 apresentam, respectivamente, gráficos *boxplot* com os escores referentes as avaliações para as recomendações obrigatórias para dados abertos e para dados abertos conectados. A partir destes gráficos foi possível detectar que as recomendações obrigatórias foram muito bem avaliadas, recebendo escore médio de 0.8 para dados abertos e 0.9 para dados abertos conectados, numa escala de 0(zero) à 1(um).

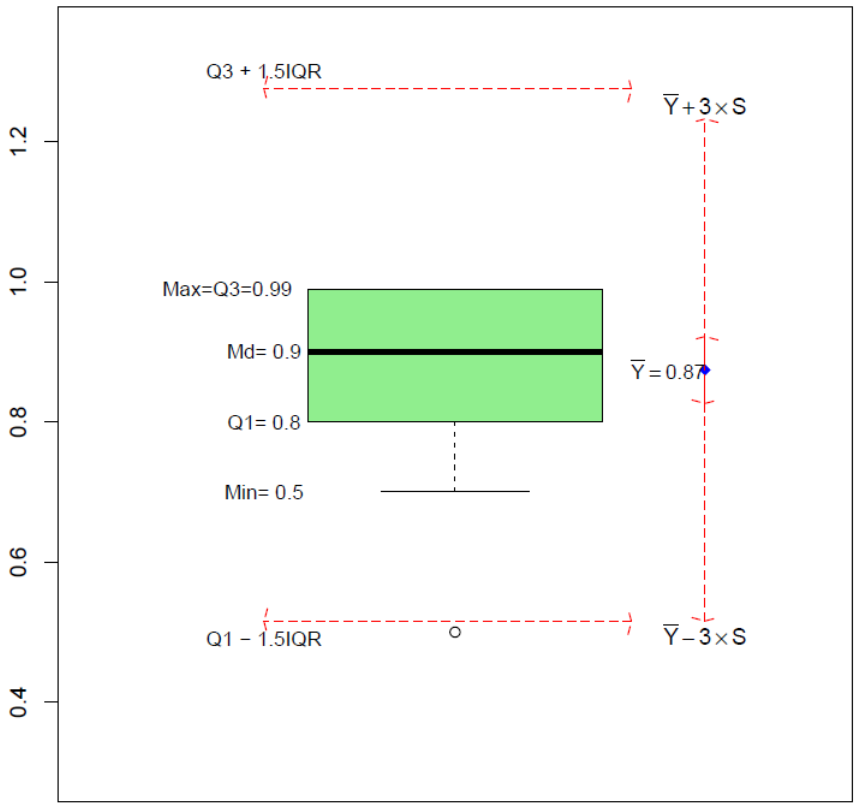
Figura 49 – Boxplot dos escores de recomendações obrigatórias para dados abertos



Fonte: Autor desta dissertação, a partir dos dados obtidos com o questionário, 2015.

As recomendações para dados abertos conectados tiveram qualificação superior, considerando a grande concentração de valores próximo ao valor máximo (0.99), onde destacamos que o valor máximo e o terceiro quartil são iguais ao valor máximo possível. Ou seja, a partir do *boxplot* 50 observa-se que apenas 25% dos avaliadores qualificaram as recomendações obrigatórias para dados abertos conectados com escores inferiores a 0.8 e que 50% dos escores estão entre 0.8 e 0.99.

Figura 50 – Boxplot dos escores de recomendações obrigatórias para dados abertos conectados



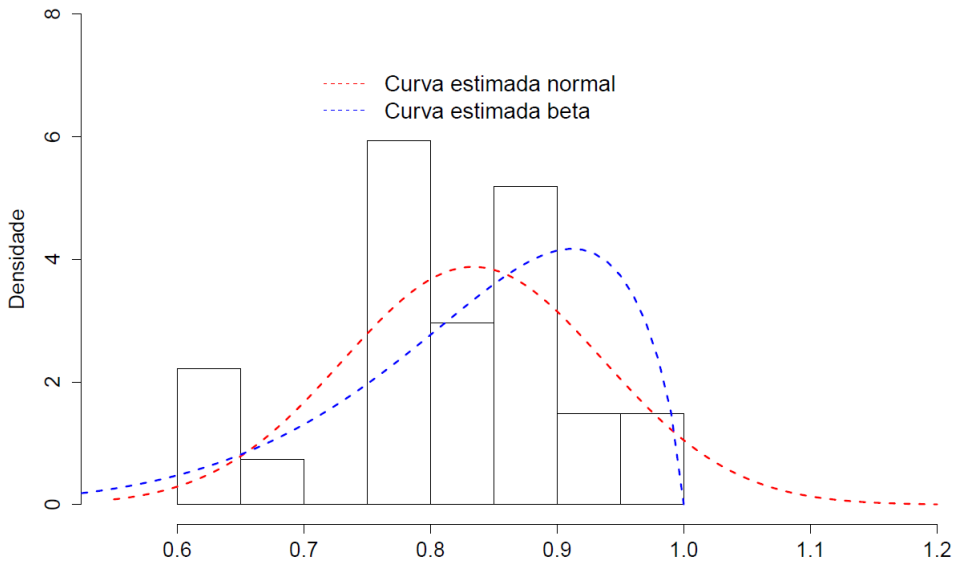
Fonte: Autor desta dissertação, a partir dos dados obtidos com o questionário, 2015.

Quanto aos dados abertos, também percebemos que a maioria das avaliações está acima do escore 0.7 (equivalente a *alto* na escala *likert* do questionário). Desta maneira é possível concluir que os dois subconjuntos de recomendações obrigatórias foram validadas pelos avaliadores.

Em seguida, avaliamos a distribuição de probabilidade das variáveis: escore global de qualificação das recomendações sugeridas para dados abertos e para dados abertos conectados conforme as Figuras 51 e 52, respectivamente.

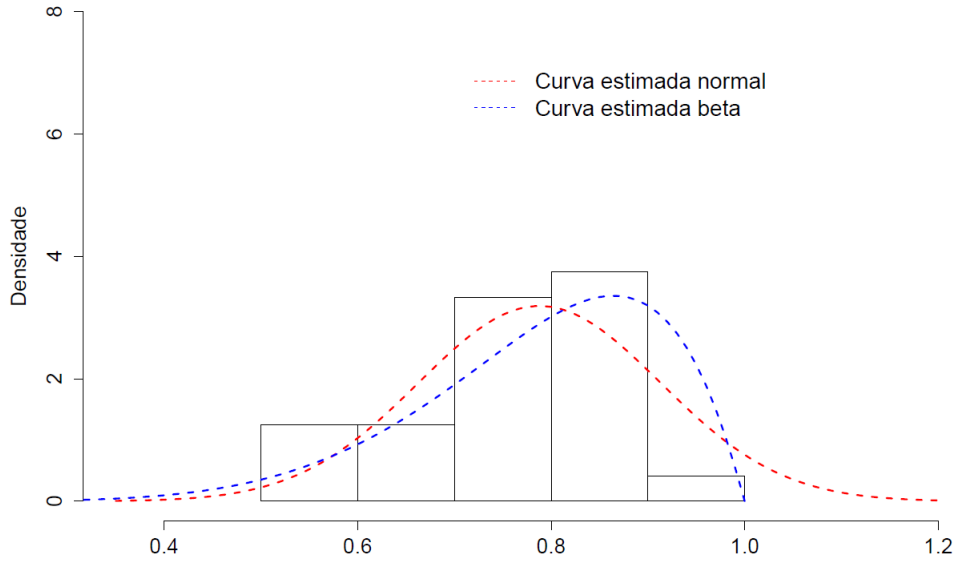
Para isto, foi ajustado aos dois histogramas, duas possibilidades de densidades, sendo uma considerando a distribuição normal e outra considerando a distribuição beta. Com base nestas figuras, percebe-se que a distribuição beta (FERRARI; CRIBARI-NETO, 2004) ajusta-se melhor aos dados, pois caso seja considerada a distribuição normal seriam estimados valores maiores que 1, algo impossível para os dados analisados. Adicionalmente, a curva beta descreve melhor o comportamento probabilístico da população de onde os dados foram amostrados, visto que o ajuste da densidade beta ao histograma amostral é melhor que se a população considerada fosse a distribuição normal padrão.

Figura 51 – Histograma dos escores para qualificação das recomendações propostas para dados abertos



Fonte: Autor desta dissertação, 2015.

Figura 52 – Histograma dos escores para qualificação das recomendações propostas para dados abertos conectados



Fonte: Autor desta dissertação, 2015.

6.2.4 Modelos de regressão

Em muitos casos, temos conhecimento limitado sobre a relação entre variáveis envolvidas em um determinado problema de interesse. Ao visualizar os valores observados destas variáveis como os resultados de um experimento devemos ter, então, uma ferramenta teórica, um modelo matemático, através do qual estas variáveis estejam relacionadas, para atuar como base do processo gerador de dados.

Entretanto, é importante destacar que todos os modelos representam simplificações da realidade e, além das variáveis possíveis de serem mensuradas, existem fatores que não podem ser controlados, ou são desconhecidos. Tais fatores podem ser considerados através de uma componente casual, representada pelos erros aleatórios. Neste contexto, o modelo de regressão linear se inclui assim como o que é apresentado a seguir:

$$y = X\beta + \epsilon =$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}}_{\begin{pmatrix} x_1 & x_2 & \dots & x_k \end{pmatrix}} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Neste modelo observamos que:

- y é um vetor de n observações da variável aleatória dependente, ou ainda, $y_1, \dots, y_n$ é uma amostra da população de interesse;
- X é uma matriz $n \times k$ formada pelas covariadas, onde cada coluna X é um conjunto de n observações da covariada x_t, $t = 1, \dots, k$. Assim temos k covariadas. Ademais, destacamos que X não é uma variável aleatória. Ela é uma variável observada e fixa;
- Além disso, β é um vetor de k parâmetros também fixos e desconhecidos (não são variáveis aleatórias);
- E finalmente, ϵ é um vetor de n erros aleatórios X com média zero $(\mathrm{E})(\epsilon_{\mathrm{i}}) = 0$ e variância constante ao longo das observações, isto é, $\mathrm{var}(\epsilon_i) = \sigma^2$ para todo $i = 1, \dots, n$.

Em um modelo de regressão, busca-se que o modelo matématico explique o máximo possível do caráter aleatório da resposta de forma que, o que não for possível explicar esteja contido no erro aleatório ϵ. Assim, pela necessidade de se explicar o máximo com base no modelo que envolve as covariadas e os parâmetros desconhecidos, o erro deve ser

pequeno, ou seja, que seja **zero**. Assim, uma das principais suposições de modelos lineares de regressão consiste que:

$$\mathrm{E}(\epsilon) = \mu_\epsilon = 0.$$

Esta premissa nos apresenta consequências importantes. Desta maneira, podemos calcular o valor esperado da seguinte expressão:

$$\mathrm{E}(y) = \mathrm{E}(X\beta) + \mathrm{E}(\epsilon) \Leftrightarrow \mathrm{E}(y) = X\beta \Leftrightarrow \mu = X\beta.$$

Consequentemente, o modelo final resultante é:

$$\mu = X\beta.$$

Assim, é possível observar que $\mathrm{E}(X\beta) = X\beta$ pois, nem X nem β são variáveis aleatórias (considerando que o valor esperado de uma constante é uma constante). Desta maneira, o comportamento da variância pode ser analisado considerando a seguinte equação:

$$\mathrm{var}(y) = \underbrace{\mathrm{var}(X\beta)}_{0} + \underbrace{\mathrm{var}(\epsilon)}_{\sigma^2} \Leftrightarrow \mathrm{var}(y) = \sigma^2.$$

Ou seja, $\mathrm{var}(X\beta) = 0$, pois nem X nem β são variáveis aleatórias (a variância de uma constante é zero. Se a variância é igual zero isto se dá porque estamos tratando de uma constante, não de um variável aleatória).

Desta maneira, é possível representar este modelo considerando a i-ésima observação como

$$\mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n.$$

Para que o modelo acima seja melhor conhecido precisamos estimar $\beta_1, \beta_2, \ldots, \beta_k$. Esta estimativa é feita utilizando o método de máxima verossimilhança que será discutido adiante. Assim, é obtido $\widehat{m}u_i$ quando obtemos $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$, tal que:

$$\widehat{\mu}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \ldots + \widehat{\beta}_k x_{ik}, \quad i = 1, \ldots, n.$$

Assim, é possível obter estimativas para μ_i.

Diante do que foi apresentado, percebe-se que o mais relevante nesta análise é a distribuição de y, sua média, sua variância e o tipo de distribuição de probabilidades que a variável aleatória y (resposta desejada) segue.

É importante ressaltar que a distribuição mais conhecida é a distribuição normal. No entanto, na prática essa distribuição não é adequada para diversos tipos de variáveis aleatórias, como segue:

- Se μ, a média da variável resposta, pode assumir tanto valores positivos quanto valores negativos e a curva de densidade de y é próxima da forma de sino, então se justifica pensar na distribuição normal [2];
- Se μ só pode assumir valor positivo e a curva de densidade de y é simétrica positiva, devemos considerar a distribuição gama [3];
- Se a variável aleatória y (que é o tipo de resposta que este modelo busca) representa dados de contagens, por exemplo $y = 0, 1, 2, 3, \ldots,$, variável aleatória discreta, podemos pensar na distribuição binomial [4];
- Se $y \in (0, 1)$ podemos pensar na distribuição beta ou na distribuição simplex.

Este último exemplo é particulamente útil, pois são tratados alguns escores, notas, desempenhos e variáveis correlatas. Tais elementos podem ser taxas, proporções ou índices que se encontram em um intervalo do tipo (a, b) que podem ser transformados para um intervalo do tipo $(0, 1)$. Neste ponto, é necessário a generalização do modelo linear com o objetivo de permitir o uso de diversas distribuições além da normal. Isto é possível considerando a expressão abaixo.

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}.$$

Onde $g(\mu_i)$ é uma função de ligação, que conecta a média da variável resposta e o modelo envolvendo as covariadas e os β's.

Para a utilização de $g(\mu_i)$ é necessário estabelecer a premissa que, quando $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, temos que $g(\mu_i) = \mu_i$. Quando isto ocorre, g desempenha o que chamamos de função identidade. Isto acontece no modelo normal porque assim como a resposta que pertence a todos os reais, isto é, $y \in (-\infty, +\infty)$, o mesmo ocorre com sua média $\mu \in (-\infty, +\infty) = IR$. Ao ser obtido $\widehat{\mu_i}$ o mesmo fenômeno deve acontecer, ou seja, $\widehat{\mu_i} \in (-\infty, +\infty) = IR$. Assim, é possível observar que $\widehat{\mu_i}$ pode assumir qualquer valor Real. Desta maneira, $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ poderão também assumir qualquer valor.

Este fenômeno não acontece se, por exemplo, a variável resposta seguir uma distribuição gama. Porque, assim como a resposta que pertence assume apenas valores reais

[2] Maiores informações em https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_normal
[3] Maiores informações em: https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_gama
[4] Maiores informações em: https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_binomial

positivos, isto é, $y \in (0, +\infty)$, o mesmo ocorre com sua média $\mu \in (0, +\infty) = IR^+$ e deve acontecer com $\widehat{\mu}$. Desta maneira, tem que acontecer: $\widehat{\mu} \in (0, +\infty) = IR^+$.

Com esta restrição os $\widehat{\beta}$'s não estão livres, pois é necessário garantir que $\widehat{\beta}X$ assuma apenas valores reais positivos. Em X não é possível haver alteração (fixa e conhecida), sendo necessário haver um um processo de estimação do β's com restrição para garantir que $\widehat{\beta}X \in (0, +\infty) = IR^+$, ou seja, uma operação que pode ser bastante complicada.

Desta maneira, a alternativa é aplicar uma função g em μ_i de forma que $g(\mu_i) \in (-\infty, +\infty) = IR$. Assim, os *$\widehat{\beta}$'s estão liberados.* Tomando a gama como exemplo, seja y uma variável aleatória com distribuição gama, aqui denotada por $Y \sim G(\mu, \phi)$, tal que

$$f_{(\mu,\phi)}(y) = \frac{1}{\Gamma(\phi)} \left(\frac{\phi x}{\mu}\right)^{\phi} \exp\left(-\frac{\phi y}{\mu}\right) \frac{1}{y},\ y \geq 0, \mu > 0, \phi > 0, \Gamma(\phi) = \int_0^{\infty} t^{(\phi-1)} e^{-t} dt.$$

Se a variável resposta pertence aos reais positivos e decide-se a distribuição gama, um modelo de regressão adequado seria

$$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, i = 1, \ldots, n.$$

pois, o logaritmo só pode ser calculado para valores reais positivos e o seu resultado assume valores em todos os reais, o que libera os $\widehat{\beta}$'s. É importante observar que $\log(u) \in (-\infty, +\infty)$ para todo $u \in (0, +\infty) IR^+$.

6.2.5 Exemplo Final - Modelo de regressão beta

A distribuição beta costuma ser usada para modelar variáveis aleatórias que assumem valores no intervalo de $(0, 1)$, tais como taxas, porcentagens e proporções como neste caso em análise. A densidade beta pode assumir formas diferentes dependendo da combinação de valores de parâmetros.

Seja $y_1, \ldots, y_n$uma amostra de variáveis independentes tal que cada y_i, $i = 1, \ldots, n$, segue a distribuição beta com densidade:

$$f(y; \mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma((1-\mu_i)\phi_i)} y^{\mu_i\phi_i - 1}(1-y)^{(1-\mu_i)\phi_i - 1}, \quad 0 < y < 1,$$

onde $0 < \mu_i < 1$ and $\phi_i > 0$. Aqui, $\mathrm{E}(y_i) = \mu_i$ e $\mathrm{var}(y_i) = (\mu_i(1-\mu_i))/(1+\phi_i)$.

Desta maneira, observa-se que ϕ pode ser visto como uma parâmetro de precisão, pois quanto maior ϕ menor a variância de y_i, por outro lado, ϕ^{-1} é um parâmetro de dispersão. Ferrari e Cribari-Neto (2004) propoem que a média da variável resposta y_i, i.e., μ_i seja escrita como

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n.$$

Assim, é possível entender porque utilizar a função de ligação para liberar os possíveis valores que os β_t's podem assumir. Neste caso como $\mu_i \in (0, 1)$ uma função de ligação

que conduz essa média à todos os Reais é a função de ligação logito, dada por

$$\mu_i \in (0,1) \leftrightarrow \log\left\{\frac{\mu_i}{(1-\mu_i)}\right\} \in (-\infty,+\infty) = IR.$$

Deste modo, como houve a tentativa de explicação da média da variável resposta μ_i, que se trata de um parâmetro desconhecido, é cabível a tentativa de modelagem da sua Variância. Neste caso, isto implica em modelar o parâmetro ϕ, pois a variância depende de μ_i (que já está sendo modelada) e de phi_i. Assim, é sugerido um modelo para ϕ quando suspeitamos que a dispersão não é constante para os dados, ou que é possível haver grupos com dispersões diferentes. Assim, Smithson e Verkuilen (2006) propõem um modelo de regressão beta em que:

$$h(\phi_i) = \gamma_1 + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \ldots + \gamma_q z_{iq}, \quad i = 1, \ldots, n.$$

Neste caso, como $\phi > 0$, uma função de ligação adequada é $\log(\phi) \in IR$. Para isto, é necessário estimar os β_t's e os γ_j'a para estimarmos ϕ_i e μ_i. Isto é feito utilizando o método de máxima verossimilhança [5].

A distribuição beta é muito flexível. De fato, a densidade da beta, a depender dos valores de seus parâmetros pode representar populações simétricas, assimétricas positivas e negativas, distribuições forma de "J" de "J" invertido, forma de "U" conforme exemplificado na Figura 53.

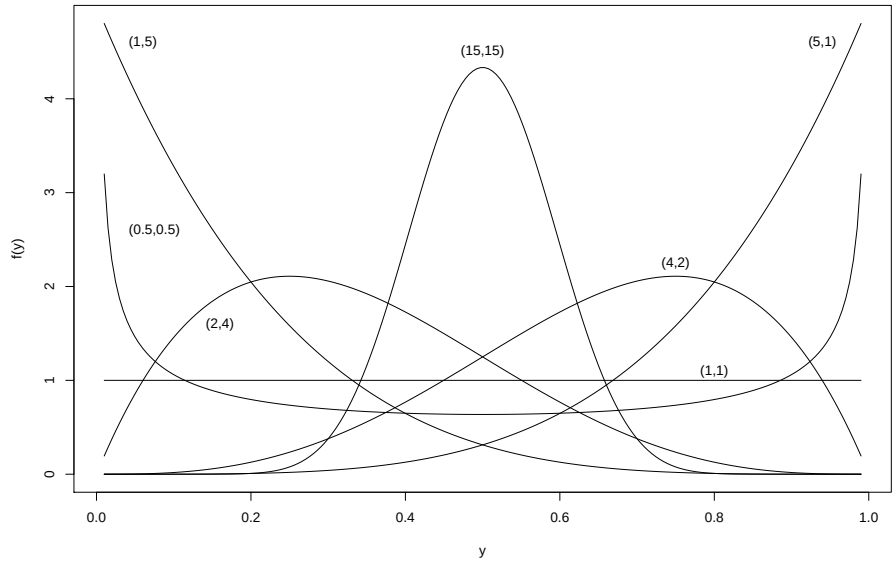
6.3 Análise dos Resultados do modelo de regressão

O objetivo da modelagem estatística é explicar o escore global de qualidade das recomendações sugeridas baseadas nos escores de dificuldade e relevância de cada recomendação, baseado também no escore global das recomendações obrigatórias e no escore de experiência do avaliador. Esse modelo estatístico deverá apontar para as recomendações que devem fazer parte do modelo de recomendações e ainda classificar os níveis de relevância das recomendações.

Assim a variável resposta do modelo é o escore de qualidade global (y) e as recomendações sugeridas são as variáveis explicativas. Para uma recomendação ser considerada estatisticamente importante para a explicação do escore de qualidade global seu p_Valor tem que ser inferior ou próximo de 5%. Em isto acontecendo, essa recomendação é escolhida para compor o modelo de recomendações para a publicação de dados abertos ou de dados abertos conectados. É importante ressaltar que o modelo de regressão é capaz de capturar as recomendações com alto escore de relevância associado com um grau baixo ou mediano de dificuldade de implementação, quando todas as recomendações são

[5] Disponível em https://www.ime.usp.br/ giapaula/cursospos.htm. Pesquisadores podem usar o pacote `betareg` o qual está disponível no software estatístico `R`

Figura 53 – Diferentes formas da distribuição beta



Fonte: Autor desta dissertação, 2015.

consideradas. Isto também só foi possível porque o questionário foi construído com este objetivo.

Nas Tabelas 19 e 20 estão, respectivamente, as estimativas e p_Valores das variáveis referentes as recomendações para dados abertos e dados abertos conectados propostas pelo modelo de regressão. As tabelas contemplam as recomendações com p_Valores próximos ou abaixo de 0.1.

Tabela 19 – Valores das estimativas dos coeficientes e p_Valores das variáveis referentes as recomendações para dados abertos propostas pelo modelo de regressão

Variáveis	**Estimativas**	**p_Valor**
ScoreConhecimento	-0.10	0.06
Recomendações		
1A,1B, 1C, 2B, 2D, 2E, 2H, 3A, 3B, 4A, 6A, 7D, 8A, 9A, 9B, 9H	1.7	0.03
1D	0.10	0.06
2A	0.25	0.00
2C	0.21	0.01
8B	0.12	0.10
9C	0.02	0.11
10A	0.15	0.03
10B	0.14	0.00

Fonte: Autor desta dissertação, 2015.

Tabela 20 – Valores das estimativas dos coeficientes e p_Valores das variáveis referentes as recomendações para dados abertos conectados propostas pelo modelo de regressão

Variáveis	**Estimativas**	**p_Valor**
ScoreConhecimento	-0.80	0.03
Recomendações		
3F, 5A, 5C, 5D, 7C	0.53	0.00
2K	1.17	0.00
5B	0.70	0.06
6B	0.02	0.08
6D	0.72	0.04
6F	0.15	0.07
9I	1.20	0.00

Fonte: Autor desta dissertação, 2015.

Os valores das estimativas dos coeficientes indicam o quanto cada recomendação impacta o escore global de qualidade das recomendações sugeridas. Adicionalmente, o sinal destas estimativas apontam em que sentido ocorre o grau de interferências das recomendações sugeridas para o escore de qualidade global.

Por exemplo, tanto para o modelo estatístico de dados abertos quanto para o modelo de dados abertos conectados, nota-se que quanto maior o escore de qualificação do avaliador mais rigorosa é sua nota, em geral o avaliador com maior escore de experiência tende a atribuir escores menores, quando comparado aos demais avaliadores, isto é observado pelas estimativas dos coeficientes das variáveis ScoresConhecimento, iguais a -0.10 e -0.8, para dados abertos e dados abertos conectados, respectivamente.

Também é importante destacar que o grupo de recomendações obrigatórias foi de grande importância para a qualificação global das recomendações sugeridas, coeficientes estimados iguais a 1.70 e 0.53, respectivamente, para dados abertos e dados abertos conectados.

Assim, caso seja necessário elencar entre as recomendações escolhidas pelo modelo estatístico, aquelas com maior impacto no escore de qualificação global, deve-se associar em grau crescente de importância, aquelas com maiores valores de estimativas e menores p_Valores.

Assim, a intervenção estatística na construção do questionário foi de grande importância para alcançar o objetivo desejado: selecionar empiricamente e com embasamento científico as recomendações relevantes e ainda possibilitando a classificação destas recomendações em níveis de prioridades ou de excelência.

6.3.1 Discussão dos Resultados

De um modo geral, os resultados apresentados pelo modelo de regressão permitiram validar os subconjuntos de recomendações altamente relevantes para serem utilizadas como

atividades do modelo de processo *"Piece of Cake"*. Neste contexto é importante destacar que tais recomendações foram validadas por um grupo de avaliadores com experiência média alta, conforme apresentado pelo modelo de regressão. Logo, para este contexto, a pesquisa pode considerar estas recomendações como imprescindíveis para qualquer atividade de publicação de dados abertos e dados abertos conectados, conforme apresentado nas Tabelas 15, 16, 17 e 18 no capítulo que explica a proposta do modelo *"Piece of Cake"*.

Por outro lado, é importante admitir que a amostra de avaliadores poderia ter sido maior, possibilitando um resultado mais acurado. Por este motivo, o questionário continua disponível para novas respostas da comunidade especializada em dados abertos e dados abertos conectados para que seja gerada uma atualização deste modelo de regressão num momento adiante, sendo posteriormente publicado num artigo científico.

Outro ponto relevante para discussão consiste na seleção das recomendações obrigatórias. Tratou-se de uma escolha do pesquisador a partir de sua experiência teórica e prática com os temas da pesquisa. Todavia, é aceitável uma discussão mais ampla neste sentido, que pode resultar num novo subconjunto de recomendações obrigatórias, bem como a incorporação de novas recomendações em todo o modelo de processo *"Piece of Cake"*.

6.4 Estudo Empírico para publicação de Dados Abertos Conectados Governamentais

O estudo empírico foi desenvolvido através do método GQM (*Goals, Questions, Metrics*) conforme conceituação apresentada anteriormente (BASILI, 1993). Nesta seção serão detalhados o planejamento do estudo empírico, a definição da amostra, a execução do estudo, análise dos resultados, discussões e ameaças à validade.

6.4.1 Planejamento do Estudo

Segundo Sauvé (2012), um estudo empírico necessita da definição de seus objetivos. A partir da definição de uma questão principal de investigação, deve ser estabelecido um modelo (*template*) que responda as seguintes perguntas:

- Qual será o objeto de estudo?
- Qual a finalidade do estudo?
- Com respeito a que foco de qualidade?
- Com qual ponto de vista?
- E em que contexto?

Para este estudo foi estabelecido o seguinte *template*:

- **Questão principal de investigação (*High Level Question*):** O modelo de processo apoiou a publicação de Dados Abertos Conectados Governamentais (DACG)?
- **Objeto de estudo:** O objeto de estudo é o processo de publicação de dados abertos conectados governamentais
- **Finalidade:** Avaliar a publicação de dados abertos governamentais (DAG)
- **Foco da qualidade:** Verificar a eficácia na publicação de DACG
- **Perspectiva:** A perspectiva é do ponto de vista do publicador de dados (e/ou governo).
- **Contexto:** Alunos de Pós-Graduação e Graduação com níveis de experiência distintos em publicação de dados governamentais
- **Aspectos:** Nível de maturidade do dado publicado; Tempo utilizado para publicação.

Assim, considerando o método GQM estabelecido, para verificar a eficácia do modelo quanto a guiar publicadores para disponibilizarem Dados Abertos Governamentais (DAG) e Dados Abertos Conectados Governamentais (DACG) foram estabelecidos 3 objetivos (*Goals*), 9 questões (*Questions*) e 15 métricas (*Metrics*), que em parte, consideram requisitos do modelo, como a utilização de BPLDs e atividades de publicação de dados, conforme apresentado na Tabela 21.

Tabela 21 – Template GQM utilizado no estudo empírico

Objetivo:	**G1: Publicar Dados Abertos Governamentais**
Questões:	**Métricas:**
Q1: Foi possível publicar DAG com 3-estrelas?	M1: Tempo de publicação do DAG com 3 estrelas
Q2: Foi possível publicar DAG com 4-estrelas?	M2: Tempo de publicação do DAG com 4 estrelas
Q3: Foi possível publicar DAG com 5-estrelas?	M3: Tempo de publicação do DAG com 5 estrelas
Objetivo:	**G2: Publicar Dados Abertos Governamentais com BPLDs**
Q4: Quantas BPLDs foram utilizadas na publicação de DAG?	M4: Percentual de BPLDs utilizadas na publicação de DAG com 3 estrelas
	M5: Percentual de BPLDs utilizadas na publicação de DAG com 4 estrelas

	M6: Percentual de BPLDs utilizadas na publicação de DAG com 5 estrelas
Objetivo:	**G3: Publicar Dados Abertos Governamentais seguindo o modelo de processo *"Piece of Cake"*?**
Q5: Quantas atividades do modelo de processo foram utilizadas na publicação de DAG?	M7: Percentual de atividades do modelo utilizadas na publicação de DAG com 3 estrelas
	M8: Percentual de atividades do modelo utilizadas na publicação de DAG com 4 estrelas
	M9: Percentual de atividades do modelo utilizadas na publicação de DAG com 5 estrelas
Q6: Quantas atividades do modelo de processo foram utilizadas por todas as equipes?	M10: Atividades do modelo utilizadas pelas 3 equipes
Q7: Quantas atividades do modelo de processo não foram utilizadas por todas as equipes?	M11: Atividades utilizadas por 2 equipes
	M12: Atividades do modelo utilizadas por 1 equipe
	M13: Atividades não utilizadas por nenhuma equipe
Q8: O modelo de processo apoia a publicação de DACG?	M14: Avaliação dos publicadores sobre o apoio proporcionado pelo modelo de processo às atividades de publicação de DACG
Q9: Publicadores estão satisfeitos com o modelo de processo?	M15: Satisfação dos publicadores com o modelo de processo

Fonte: Autor desta dissertação, 2015.

6.4.1.1 Definição da Amostra

Entende-se por população o conjunto de elementos que tem, em comum, determinada característica. Consequentemente, todo o subconjunto não vazio e com menor quantidade de elementos do que a população considerada, constitui uma amostra. A informação recolhida para uma amostra pode, posteriormente, ser generalizada a toda a população.

Para a definição da população, foi adotado o método não-probabilístico de amostragem "Grupo Focal". Este método foi adotado por que era necessário avaliar a execução do estudo com alguns indivíduos com experiência na publicação de DAG e DACG, bem como com indivíduos que não possuíssem experiência. Como não houve disponibilidade de instituições governamentais para se aplicar o estudo, foi necessário o convite a pesquisadores da rede de contatos do pesquisador.

Inicialmente, a população foi composta por 7 pesquisadores integrantes do Núcleo de Excelência em Tecnologias Sociais – NEES do Instituto de Computação da UFAL. Todavia, um dos pesquisadores foi considerado inapto para participar do estudo por ter tido contato com versões preliminares da proposta de modelo da pesquisa. Assim sendo, foram selecionados 6 pesquisadores, segmentados em 3 duplas, sendo:

- Equipe A: Formada com por dois pesquisadores com experiência teórica e prática (1 Mestre em Modelagem Computacional do Conhecimento com Bacharelado em Ciência da Computação e 1 Bacharelando em Ciência da Computação);
- Equipe B: Formada por um pesquisador com experiência teórica e prática e outro com experiência teórica (1 Mestranda em Informática com Bacharelado em Ciência da Computação e 1 Bacharelando em Engenharia da Computação);
- Equipe C: Formada por dois pesquisadores sem nenhuma experiência em publicação de dados governamentais (2 Bacharelandos em Ciência da Computação).

Justifica-se a formação de duplas pelos seguintes motivos: o modelo de processo possui atividades que necessitam de discussão entre os participantes; equipes com o mesmo número de integrantes facilitam a comparação dos resultados desenvolvidos; e ainda, caso as equipes tivessem mais de dois integrantes, reduziria a quantidade de vezes que o modelo de processo seria executado, prejudicando os resultados do estudo.

A amostra deve ser considerada como pequena, entretanto, viabilizou a execução do estudo. Importante ressaltar que não foi possível a incorporação de outros pesquisadores, por motivo de indisponibilidade de agenda dos mesmos.

6.4.1.2 Insumos para o estudo

Para subsidiar a execução do estudo, foram disponibilizados às três equipes alguns conhecimentos relevantes como:

- O pesquisador realizou uma apresentação de uma hora, esclarecendo a motivação e a problemática da pesquisa de mestrado, bem como a estruturação do estudo, e a composição do modelo de processo, mediante a apresentação de diagramas explicativos e informações complementares;
- Foram disponibilizados ainda, para cada equipe:

  - Dois arquivos no formato PDF (*Portable Document Format*) contendo dados governamentais, que deveriam ser convertidos para dados abertos conectados. Um arquivo é referente a algumas empresas contidas no Cadastro de Empresas Inidôneas e Suspensas – CEIS do Estado de Alagoas. O outro arquivo refere-se à relação de servidores e folha de pagamento de um órgão do Governo de Alagoas. Tais arquivos foram escolhidos por se tratarem de dados e informações bastante consumidas pela sociedade para fins de controle social do governo;
  - Os slides utilizados na apresentação;
  - Um documento contendo todas as atividades propostas na versão inicial do modelo de processo, estabelecidas em decorrência da revisão de literatura.

6.4.1.3 Instrumentos de coleta de dados

Complementar aos insumos disponibilizados para o estudo, a partir do template GQM, foram disponibilizados dois instrumentos para registro de dados relevantes para contabilização das métricas sendo:

- Uma planilha contendo a relação de 65 atividades classificadas para os quatro processos do modelo. Os avaliadores foram orientados a registrar nesta planilha o tempo gasto para executar o processo, quais recomendações foram utilizadas e ainda, comentários adicionais. A planilha foi preenchida pelas equipes;
- Um questionário de avaliação do estudo empírico contendo questões como:
  - Avaliação da metodologia do estudo;
  - Avaliação do modelo de processo *"Piece of Cake"*;
  - Avaliação tanto quanto à eficácia do modelo de processo quanto à publicação da DACG;
  - Avaliação da satisfação dos pesquisadores em ter participado do estudo;
  - Avaliação da experiência dos avaliadores com a temática do estudo.

Além destes instrumentos, foram solicitados aos avaliadores que caso fossem produzidos registros adicionais, deveriam ser disponibilizados ao pesquisador. Foi solicitado ainda os artefatos gerados em decorrência do estudo, bem como o endereço eletrônico (quando houvesse) de onde os dados abertos conectados governamentais publicados foram disponibilizados na *Web*.

6.4.1.4 Análise de Ameaças à validade

Embora este estudo empírico tenha sido realizado com planejamento e cautela para minimizar possíveis ameaças a sua validade, as quais possam comprometer as conclusões, existem algumas que devem ser mencionadas:

- Conjuntos de dados: A utilização dos mesmos conjuntos de dados pelas equipes participantes pode ser considerada uma ameaça à validade, pois tais conjuntos de dados poderiam ser de conhecimento maior de parte dos avaliadores do que outros. Uma diversidade de conjuntos de dados poderia minimizar esta ameaça.

- Quantidade de conjuntos de dados: Foram utilizados apenas dois conjuntos de dados por equipe. A utilização de uma maior quantidade de conjuntos de dados permitiria uma mensuração mais acurada dos resultados do estudo;

- Avaliadores escolhidos subjetivamente: Os avaliadores foram escolhidos avulsamente observando o nível de experiência em relação a temática do estudo. Nada foi medido para classificar corretamente os participantes, o que pode comprometer os resultados;

- Cansaço dos avaliadores durante o estudo: O estudo foi composto de uma preparação inicial, análise de um documento extenso com as recomendações/atividades para publicação de DACG, e ainda, o desenvolvimento das atividades para a publicação dos dados. Assim, o volume de trabalho foi significativo o que pode ter influenciado no desempenho de algumas equipes;

- Formação dos participantes do estudo: Todos os participantes do estudo tem formação em computação, o que pode ter reduzido a complexidade para a execução do estudo. Uma maior variação na formação dos participantes seria desejável para uma maior aproximação de um cenário real;

- Ambiente de desenvolvimento do estudo: Por não ter havido disponibilidade de um cenário real de aplicação do estudo, as atividades foram realizadas no laboratório, buscando similar uma situação real. Todavia, esta simulação não reproduz fidedignamente a rotina de um órgão público, o que pode ter influenciado nos resultados do estudo empírico.

6.4.2 Execução do estudo

O estudo foi executado no laboratório do Núcleo de Excelência em Tecnologias Sociais – NEES. Além dos insumos fornecidos, as duplas tiveram disponibilidade de computadores com acesso à internet e poderiam utilizar outros insumos e ferramentas que necessitassem. Foi estipulado um período de 2 dias (16 horas), entretanto, foi permitido que as equipes que desejassem, poderiam concluir o estudo em mais tempo.

O pesquisador observou o trabalho das equipes sem realizar interferência. A observação foi necessária para registrar fatos relevantes do estudo que precisavam ser capturados para melhor embasar a análise dos resultados.

As três equipes conseguiram executar o estudo. A equipe A, de maior experiência, desenvolveu o estudo em menor tempo, seguido respectivamente pela equipe B (de experiência intermediária) e pela equipe C (de nenhuma experiência). Os resultados deste processo de validação empírica são descritos na subseção seguinte.

6.4.3 Análise dos Resultados

Em subseções anteriores, foram apresentados as atividades adotadas no planejamento e execução do estudo, incluindo os objetivos e métricas de avaliação. Nesta subseção são detalhados os resultados obtidos, a seguir.

6.4.3.1 Questões e Métricas para o objetivo G1

Para verificar se o estudo resultaria na publicação de dados abertos (conectados) governamentais, foram estabelecidas as seguintes questões e métricas conforme a Tabela 22.

Tabela 22 – Questões do modelo GQM utilizado no estudo empírico

Questões:	**Métricas:**
Q1: Foi possível publicar DAG com 3-estrelas?	M1: Tempo de publicação do DAG com 3 estrelas
Q2: Foi possível publicar DAG com 4-estrelas?	M1: Tempo de publicação do DAG com 4 estrelas
Q3: Foi possível publicar DAG com 5-estrelas?	M1: Tempo de publicação do DAG com 5 estrelas

Fonte: Autor desta dissertação, 2015.

Em resposta a Q1, todas as três equipes conseguiram publicar os arquivos fornecidos como insumo no formato CSV (*Comma Separated Values*), compatível com o nível 3 estrelas. A Tabela 23 sumariza o progresso desta atividade pelas três equipes:

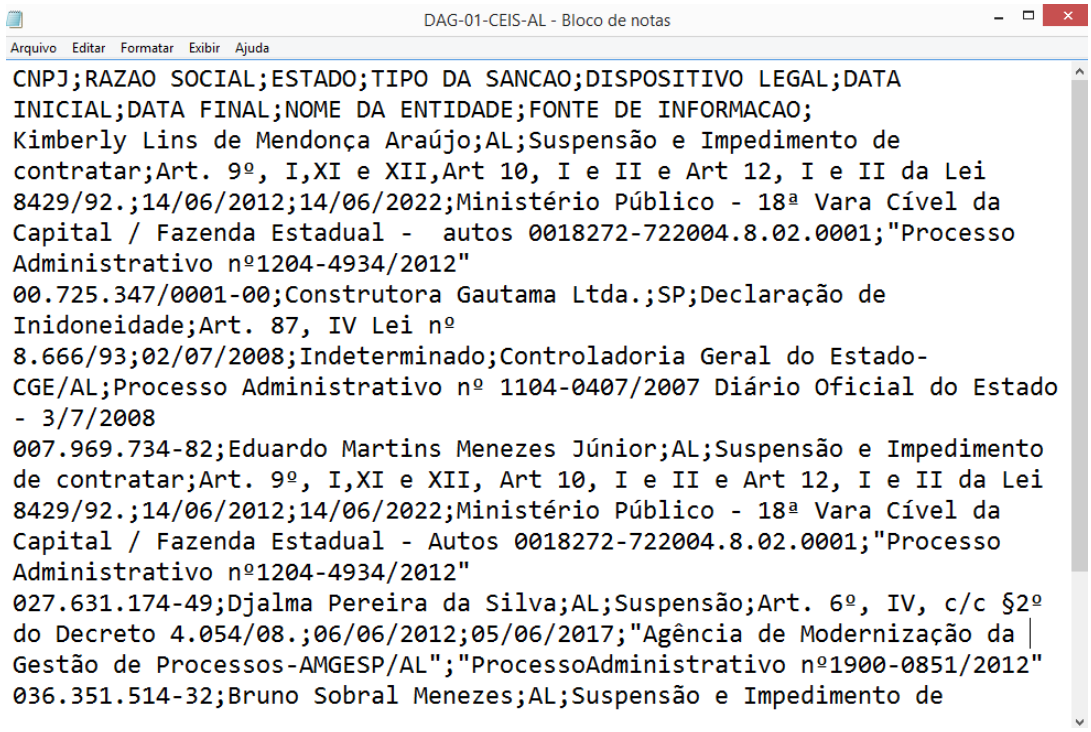
Tabela 23 – Métricas relacionadas à questão Q1 no template GQM

	Publicou Arquivo 01?	**Publicou Arquivo 02?**	**Disponíveis em:**	**M1**
Equipe A:	Sim	Sim	https://sites.google.com/site/catalogothiago/	276 minutos
Equipe B:	Sim	Sim	http://nees.com.br/experimento/thiago/	-
Equipe C:	Sim	Sim	http://dadosabertosconectados.blogspot.com.br/	300 minutos

Fonte: Autor desta dissertação, 2015.

As Figuras 54 e 55 evidenciam o êxito desta atividade por uma das equipes participantes do estudo empírico.

Figura 54 – Captura de tela referente ao arquivo no formato CSV do CEIS-Alagoas



Fonte: Autor desta dissertação, a partir das evidências do estudo empírico, 2015.

Em resposta a Q2 e Q3, as equipes A e B conseguiram publicar os dois arquivos em formato RDF/XML acompanhado de uma ontologia, salva em arquivos OWL. A equipe C conseguiu apenas publicar um dos arquivos. A Tabela 24 sumariza o progresso desta atividade pelas três equipes:

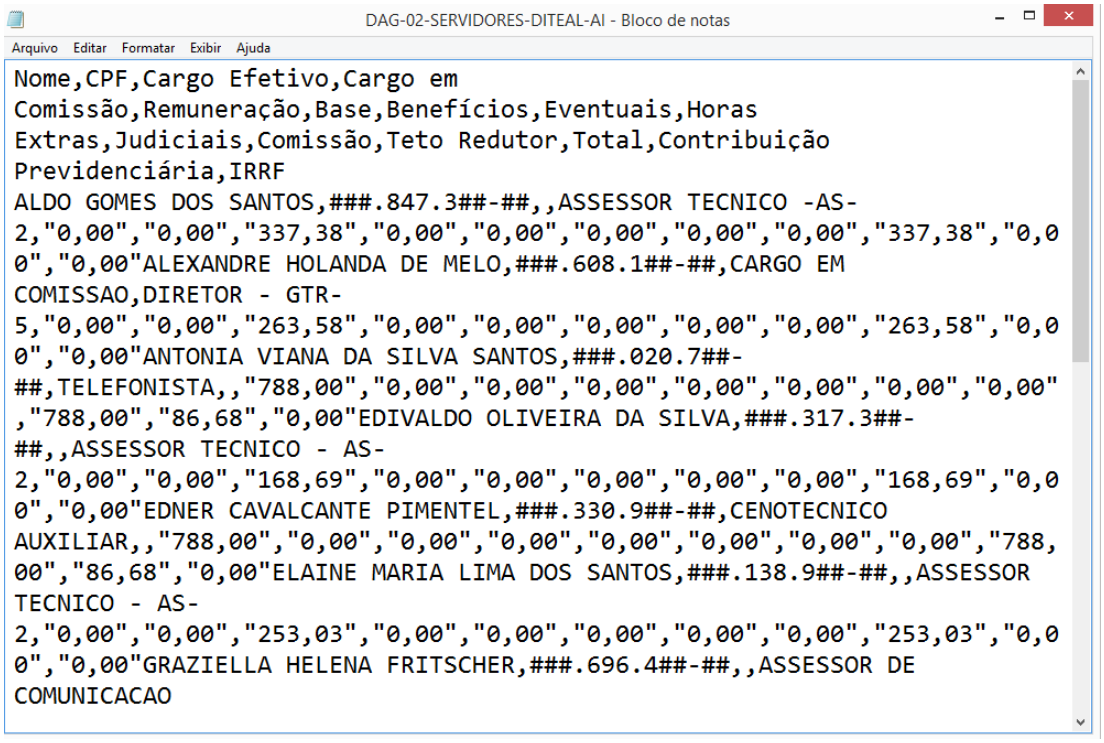
Tabela 24 – Métricas relacionadas às questões Q2 e Q3 no template GQM

	Publicou Arquivo 01?	**Publicou Arquivo 02?**	**Disponíveis em:**	**M2**	**M3**
Equipe A:	Sim	Sim	https://sites.google.com/site/catalogothiago/	115 min.	70 min.
Equipe B:	Sim	Sim	http://nees.com.br/experimento/thiago/	219 min.	60 min.
Equipe C:	Sim	Não	http://dadosabertosconectados.blogspot.com.br	-	30 min.

Fonte: Autor desta dissertação, 2015.

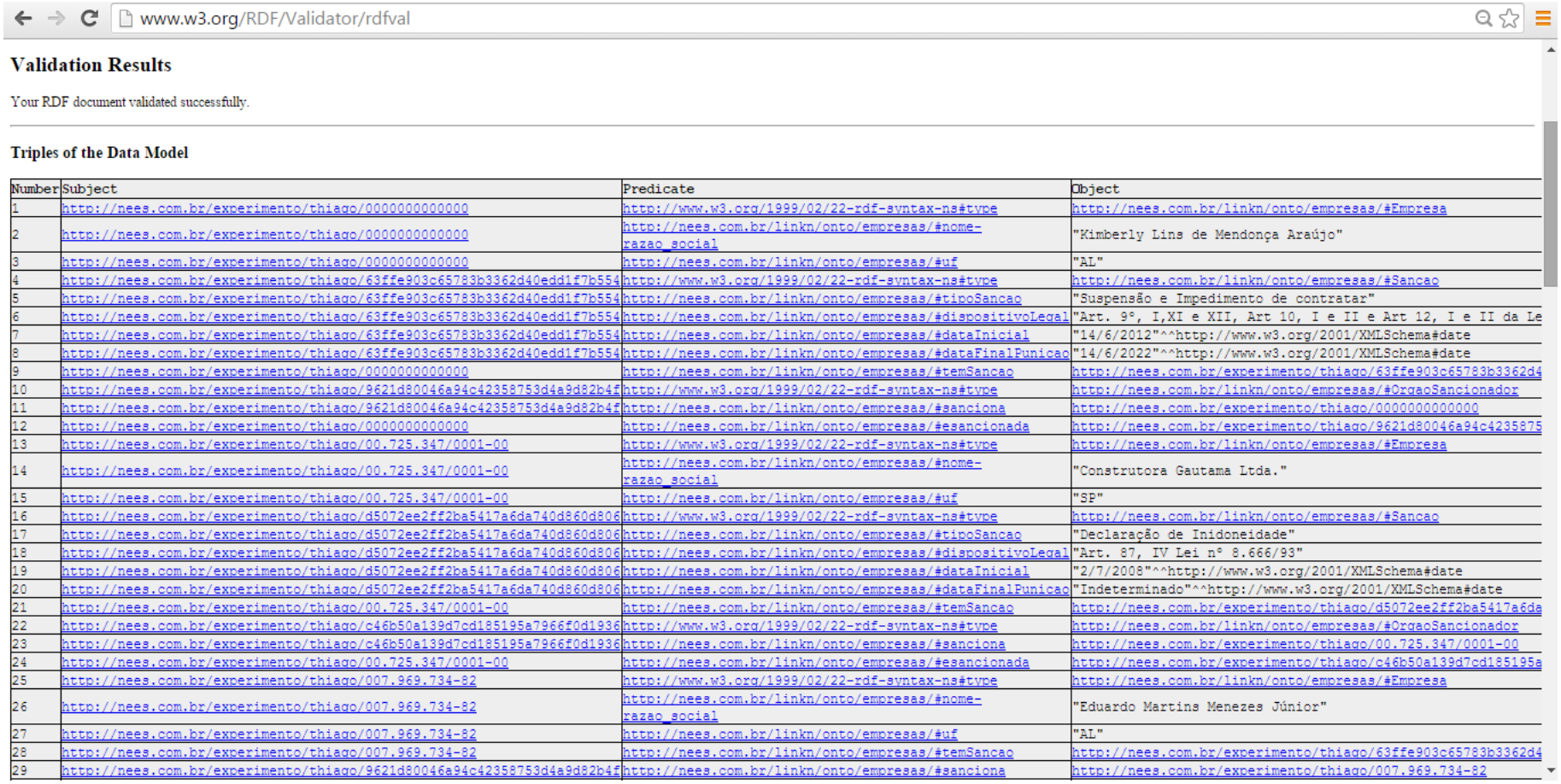
A Figura 56 evidencia o êxito desta atividade por uma das equipes participantes do estudo empírico.

Figura 55 – Captura de tela referente ao arquivo no formato CSV da folha de servidores do DITEAL



Fonte: Autor desta dissertação, a partir das evidências do estudo empírico, 2015.

Figura 56 – Captura de tela referente ao arquivo no formato RDF do CEIS-Alagoas



www.w3.org/RDF/Validator/rdfval

Validation Results

Your RDF document validated successfully.

Triples of the Data Model

Number	Subject	Predicate	Object
1	http://nees.com.br/experimento/thiago/0000000000000	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://nees.com.br/linkn/onto/empresas/#Empresa
2	http://nees.com.br/experimento/thiago/0000000000000	http://nees.com.br/linkn/onto/empresas/#nome-razao_social	"Kimberly Lins de Mendonça Araújo"
3	http://nees.com.br/experimento/thiago/0000000000000	http://nees.com.br/linkn/onto/empresas/#uf	"AL"
4	http://nees.com.br/experimento/thiago/63ffe903c65783b3362d40edd1f7b554	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://nees.com.br/linkn/onto/empresas/#Sancao
5	http://nees.com.br/experimento/thiago/63ffe903c65783b3362d40edd1f7b554	http://nees.com.br/linkn/onto/empresas/#tipoSancao	"Suspensão e Impedimento de contratar"
6	http://nees.com.br/experimento/thiago/63ffe903c65783b3362d40edd1f7b554	http://nees.com.br/linkn/onto/empresas/#dispositivoLegal	"Art. 9º, I,XI e XII, Art 10, I e II e Art 12, I e II da Le
7	http://nees.com.br/experimento/thiago/63ffe903c65783b3362d40edd1f7b554	http://nees.com.br/linkn/onto/empresas/#dataInicial	"14/6/2012"^^http://www.w3.org/2001/XMLSchema#date
8	http://nees.com.br/experimento/thiago/63ffe903c65783b3362d40edd1f7b554	http://nees.com.br/linkn/onto/empresas/#dataFinalPunicao	"14/6/2022"^^http://www.w3.org/2001/XMLSchema#date
9	http://nees.com.br/experimento/thiago/0000000000000	http://nees.com.br/linkn/onto/empresas/#temSancao	http://nees.com.br/experimento/thiago/63ffe903c65783b3362d4
10	http://nees.com.br/experimento/thiago/9621d80046a94c42358753d4a9d82b4f	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://nees.com.br/linkn/onto/empresas/#OrgaoSancionador
11	http://nees.com.br/experimento/thiago/9621d80046a94c42358753d4a9d82b4f	http://nees.com.br/linkn/onto/empresas/#sanciona	http://nees.com.br/experimento/thiago/0000000000000
12	http://nees.com.br/experimento/thiago/0000000000000	http://nees.com.br/linkn/onto/empresas/#esancionada	http://nees.com.br/experimento/thiago/9621d80046a94c4235875
13	http://nees.com.br/experimento/thiago/00.725.347/0001-00	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://nees.com.br/linkn/onto/empresas/#Empresa
14	http://nees.com.br/experimento/thiago/00.725.347/0001-00	http://nees.com.br/linkn/onto/empresas/#nome-razao_social	"Construtora Gautama Ltda."
15	http://nees.com.br/experimento/thiago/00.725.347/0001-00	http://nees.com.br/linkn/onto/empresas/#uf	"SP"
16	http://nees.com.br/experimento/thiago/d5072ee2ff2ba5417a6da740d860d806	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://nees.com.br/linkn/onto/empresas/#Sancao
17	http://nees.com.br/experimento/thiago/d5072ee2ff2ba5417a6da740d860d806	http://nees.com.br/linkn/onto/empresas/#tipoSancao	"Declaração de Inidoneidade"
18	http://nees.com.br/experimento/thiago/d5072ee2ff2ba5417a6da740d860d806	http://nees.com.br/linkn/onto/empresas/#dispositivoLegal	"Art. 87, IV Lei nº 8.666/93"
19	http://nees.com.br/experimento/thiago/d5072ee2ff2ba5417a6da740d860d806	http://nees.com.br/linkn/onto/empresas/#dataInicial	"2/7/2008"^^http://www.w3.org/2001/XMLSchema#date
20	http://nees.com.br/experimento/thiago/d5072ee2ff2ba5417a6da740d860d806	http://nees.com.br/linkn/onto/empresas/#dataFinalPunicao	"Indeterminado"^^http://www.w3.org/2001/XMLSchema#date
21	http://nees.com.br/experimento/thiago/00.725.347/0001-00	http://nees.com.br/linkn/onto/empresas/#temSancao	http://nees.com.br/experimento/thiago/d5072ee2ff2ba5417a6da
22	http://nees.com.br/experimento/thiago/c46b50a139d7cd185195a7966f0d1936	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://nees.com.br/linkn/onto/empresas/#OrgaoSancionador
23	http://nees.com.br/experimento/thiago/c46b50a139d7cd185195a7966f0d1936	http://nees.com.br/linkn/onto/empresas/#sanciona	http://nees.com.br/experimento/thiago/00.725.347/0001-00
24	http://nees.com.br/experimento/thiago/00.725.347/0001-00	http://nees.com.br/linkn/onto/empresas/#esancionada	http://nees.com.br/experimento/thiago/c46b50a139d7cd185195a
25	http://nees.com.br/experimento/thiago/007.969.734-82	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://nees.com.br/linkn/onto/empresas/#Empresa
26	http://nees.com.br/experimento/thiago/007.969.734-82	http://nees.com.br/linkn/onto/empresas/#nome-razao_social	"Eduardo Martins Menezes Júnior"
27	http://nees.com.br/experimento/thiago/007.969.734-82	http://nees.com.br/linkn/onto/empresas/#uf	"AL"
28	http://nees.com.br/experimento/thiago/007.969.734-82	http://nees.com.br/linkn/onto/empresas/#temSancao	http://nees.com.br/experimento/thiago/63ffe903c65783b3362d4
29	http://nees.com.br/experimento/thiago/9621d80046a94c42358753d4a9d82b4f	http://nees.com.br/linkn/onto/empresas/#sanciona	http://nees.com.br/experimento/thiago/007.969.734-82

Fonte: Autor desta dissertação, a partir das evidências do estudo empírico, 2015.

Esta questão evidencia que o estudo empírico teve seu objeto nos três cenários de publicação de dados (envolvendo as 3 equipes com experiências diferentes). Destaca-se o êxito da equipe C, que mesmo sem possuir nenhuma experiência anterior com publicação de dados abertos conectados governamentais, conseguiu publicar um dos arquivos no nível 5-estrelas.

Entretanto, conforme já exposto, os pesquisadores envolvidos tem formação na área de computação e podem ter desenvolvido atividades de publicação de outros tipos de dados em atividades de caráter acadêmico ou experiência prática. Logo, somente com a resposta de Q1 não é possível afirmar que o modelo de processo *"Piece of Cake"* corroborou com o êxito deste estudo, sendo necessário analisar as demais questões, descritas nas próximas subseções.

6.4.3.2 Questões e Métricas para o objetivo G2

Para verificar se o estudo resultaria na publicação de dados abertos governamentais utilizando as BPLDs, foram as seguintes Questões e Métricas apresentadas na Tabela 25.

Tabela 25 – Métricas relacionadas à questão Q4 no template GQM

Questões:	**Métricas:**
Q4: Quantas BPLDs foram utilizadas na publicação de DAG?	M4: Percentual de BPLDs utilizadas na publicação de DAG com 3 estrelas
	M5: Percentual de BPLDs utilizadas na publicação de DAG com 4 estrelas
	M6: Percentual de BPLDs utilizadas na publicação de DAG com 5 estrelas

Fonte: Autor desta dissertação, 2015.

Para responder a Q4, as três equipes registraram na planilha de atividades fornecida no estudo, quais BPLDs utilizaram para publicação de dados em 3, 4 e 5 estrelas. As três equipes utilizaram 6 de 7 BPLDs (86%) ao longo do processo A. Para o processo B, duas equipes utilizaram todas as 5 BPLDs (100%), e a equipe B apenas 4 de 5 (80%). No processo C, duas equipes utilizaram 2 de 4 BPLDs (50%) e a equipe B, 3 de 4 (75%). A Tabela 26 apresenta os resultados para M4, M5 e M6 por equipe:

Tabela 26 – Resultados das métricas relacionadas à questão Q4 no template GQM

	M4	**M5**	**M6**
Equipe A:	86%	100%	50%
Equipe B:	86%	80%	75%
Equipe C:	86%	100%	50%

Fonte: Autor desta dissertação, 2015.

A partir dos resultados de M4, M5 e M6, já é possível detectar algum nível de contribuição do modelo de processo à atividade de publicação, pois apesar da disponibilidade do modelo, as equipes poderiam ter optado por não segui-lo, publicando os dados de outra maneira. Conforme os dados da Tabela 26, especialmente para os processos de publicação nos níveis 3 e 4 estrelas, o percentual de BPLDs utilizado foi significativo, o que nos permite deduzir que o modelo de processo guiou as três equipes ao longo destas melhores práticas. Novamente, destacamos a desempenho da equipe C que, mesmo sem experiência prévia, informou ter seguido a maioria das BPLDs associadas da cada processo do modelo. A partir desta questão, é possível afirmar que o modelo de processo *"Piece of Cake"* teve alguma influencia no êxito deste estudo. Para uma comprovação mais aprofundada, é necessário analisar as demais questões.

6.4.3.3 Questões e Métricas para o objetivo G3

Para verificar se o estudo resultaria na publicação de dados abertos governamentais utilizando o modelo de processo *"Piece of Cake"*, foram estabelecidas as Questões e Métricas descritas na Tabela 27.

Para responder a Q5, as três equipes registraram na planilha de atividades fornecida no estudo, quais atividades utilizaram do modelo de processo *"Piece of Cake"* para realizar a publicação de dados em 3, 4 e 5 estrelas. Para publicar em 3-estrelas, o modelo apresenta 37 atividades. Para 4-Estrelas, mais 23 atividades e para 5-Estrelas, mais 5 atividades. Ainda existem 2 atividades recomendadas para o nível 5-Estrelas aprimorado, onde 1 das atividades foi desenvolvida como atividade voltada à publicação em 5-estrelas. Entretanto, não foi contabilizada para esta métrica. A Tabela 28 apresenta os resultados para M7, M8 e M9 por equipe:

Os resultados de M7, M8 e M9 possibilitam inferir que o modelo de processo *"Piece of Cake"*, neste estudo, atinge o seu objetivo. Entretanto, a partir dos dados apresentados, detecta-se que a sua utilização é inversamente proporcional ao nível de experiência da equipe. Somando-se as 70 atividades utilizadas para percorrer os processos A, B e C, a equipe A utilizou 28 (40%). A equipe B utilizou 38 (54%) e a equipe C, menos experiente, utilizou 42 (60%).

Entretanto, do ponto de vista percentual considerando M7, M8 e M9, verifica-se que um conjunto relevante de atividades propostas pelo modelo não foi utilizado. A partir das observações registradas pelas equipes e disponibilizadas ao pesquisador, serão apresentados na subseção de discussões questões positivas e oportunidades de melhoria do modelo de processo. Há de se considerar que, a proposta original do modelo de processo *"Piece of Cake"*, utilizada neste estudo empírico, dispunha de um conjunto grande de atividades que pode ser a causa do não uso da totalidade de atividades propostas. Exemplificando, com base na Tabela 28, a equipe A não desenvolveu 42 atividades do modelo (60%).

Para responder a Q6 e Q7, a partir da planilha de atividades fornecida no estudo,

Tabela 27 – Métricas relacionadas à questão Q5 no template GQM

Questões:	**Métricas:**
Q5: Quantas Atividades do modelo de processo foram utilizadas?	M7: Percentual de atividades utilizadas na publicação de DAG com 3 estrelas
	M8: Percentual de atividades utilizadas na publicação de DAG com 4 estrelas
	M9: Percentual de atividades utilizadas na publicação de DAG com 5 estrelas
Q6: Quantas atividades do modelo de processo foram utilizadas por todas as equipes?	M10: Atividades do modelo utilizadas pelas 3 equipes
Q7: Quantas atividades do modelo de processo não foram utilizadas por todas as equipes?	M11: Atividades utilizadas por 2 equipes
	M12: Atividades do modelo utilizadas por 1 equipe
	M13: Atividades não utilizadas por nenhuma equipe
Q8: O modelo de processo apoia a publicação de DACG?	M14: Avaliação dos publicadores sobre o apoio proporcionado pelo modelo de processo às atividades de publicação de DACG
Q9: Publicadores estão satisfeitos com o modelo de processo?	M15: Satisfação dos publicadores com o modelo de processo

Fonte: Autor desta dissertação, 2015.

Tabela 28 – Resultados das Métricas relacionadas à questão Q5 no template GQM

	M7	**M8**	**M9**
Equipe A:	46%	38%	40%
Equipe B:	54%	71%	60%
Equipe C:	53%	63%	40%

Fonte: Autor desta dissertação, 2015.

foram contabilizadas quais atividades foram utilizadas por todas as equipes (M10), por pelo menos 2 equipes (M11), por apenas uma equipe (M12) e as atividades não utilizadas durante o estudo (M13). A Tabela 29 apresenta a sumarização dos resultados destas métricas associadas às questões Q6 e Q7.

Para o desenvolvimento desta tarefa, foi entregue para cada equipe uma planilha com o conteúdo das Tabelas 15, 16, 17 e 18 sendo o uso das atividades utilizadas como elemento de mensuração das métricas M10, M11, M12, e M13.

A partir das respostas das questões Q6 e Q7, é possível obter alguns indícios para propor uma classificação entre atividades obrigatórias e desejáveis para o modelo de pro-

Tabela 29 – Métricas relacionados às questões Q6 e Q7 no template GQM

Métricas:	**Processo A**	**Processo B**	**Processo C**	**Processo D**	**Total (Ativ.)**	**Acumul. (Ativ.)**
Utilizadas pelas 3 equipes (M10)	**7 Atividades** (1A, 1B, 1C, 3A, 3C, 4A, 9B)	**5 Atividades** (3G, 5E, 6B, 6D, 7A)	0	0	**12**	**12**
Utilizadas por 2 equipes (M11)	**14 Atividades** (1D, 1E, 2D, 2F, 2G, 3B, 8B, 9A, 9C, 9D, 10A, 10B, 10C, 10D)	**10 Atividades** (2K, 3F, 5A, 5D, 5F, 5G, 5H, 5I, 5J, 6E)	**5 Atividades** (6G, 7D, 8D, 9H, 9I)	**1 Atividade** (6H)	**30**	**42**
Utilizadas por 1 equipe (M12)	**11 Atividades** (1F, 2A, 2B, 2I, 2J, 3D, 3E, 4B, 4C, 8A, 9G)	**2 Atividades** (5B, 5C)	0	0	**13**	**55**
Não-Utilizada (M13)	**7 Atividades** (2C, 2E, 2H, 8C, 9E, 9F, 10E)	**5 Atividades** (5K, 6A, 6C, 6F, 7B)	**2 Atividades** (7C, 7E)	**1 Atividade** (7F)	**15**	**70**

Fonte: Autor desta dissertação, 2015.

cesso. Considerando o conjunto de 12 atividades que foram desenvolvidas por todas as equipes, podemos inferir que tais atividades tiveram maior relevância na execução do modelo independente do seu nível de dificuldade, afinal, foram utilizadas de forma unânime para apoiar a publicação, podendo ser propostas como atividades obrigatórias. Quanto às demais 58 atividades, por terem sido menos utilizadas no estudo, a priori podem ser propostas como atividades desejáveis, ou em alguns casos específicos, também como obrigatórias a depender do contexto de publicação de dados.

Cumpre destacar que o modelo de regressão já apresentou um conjunto de atividades obrigatórias para o modelo *"Piece of Cake"*. Nas discussões gerais deste capítulo será apresentado o conjunto final de atividades obrigatórias considerando os insumos gerados pelos dois instrumentos de validação empírica bem como a experiência do pesquisador na

temática da pesquisa.

Para responder a Q8 e Q9, os avaliadores informaram individualmente, mediante um questionário aplicado no término do experimento, uma nota de 1 a 99 em resposta às questões. As Tabelas 30 e 31 apresentam as respostas fornecidas para M14 e M15 mediante coleta de dados neste questionário:

Tabela 30 – Métricas relacionados à questão Q8 no template GQM

Métrica:	**Avaliador 1**	**Avaliador 2**	**Avaliador 3**	**Avaliador 4**	**Avaliador 5**	**Média**
M14	80	25	90	99	99	78,6

Fonte: Autor desta dissertação, 2015.

Tabela 31 – Métricas relacionados à questão Q9 no template GQM

Métrica:	**Avaliador 1**	**Avaliador 2**	**Avaliador 3**	**Avaliador 4**	**Avaliador 5**	**Média**
M15	90	10	80	80	99	71,8

Fonte: Autor desta dissertação, 2015.

O valor médio das notas atribuídas em M14 reforça o entendimento que o modelo *"Piece of Cake"* atinge seu propósito e apoia a publicação de DACG. Além disso, os valores atribuídos à M15 permitem responder que os publicadores envolvidos no estudo ficaram satisfeitos com o modelo de processo apresentado. Destacamos ainda que todos os artefatos utilizados neste estudo empírico estão disponíveis para consulta no endereço eletrônico http://www.thiagoavila.net/research/masterThesis/estudoEmpirico.

Na próxima subseção serão desenvolvidas algumas discussões sobre o estudo apresentado e as observações registradas e remetidas pelos avaliadores.

6.4.4 Discussão dos Resultados

De uma forma geral, pelos resultados apresentados para as todas as questões do estudo empírico, percebe-se que a abordagem proposta (modelo de processo para publicação de dados abertos conectados governamentais) possui eficácia, especialmente se considerar o desempenho da equipe C (não possuía experiência). Para uma avaliação geral, o questionário aplicado ao final do estudo empírico obteve também uma nota de 1 a 99 em resposta à avaliação geral feita por cada participante sobre o modelo de processo *"Piece of Cake"*, conforme exposto na Tabela 32.

Desta feita, observa-se que, considerando os resultados deste estudo empírico, há indícios de que o modelo de processo *"Piece of Cake"* guia publicadores de dados abertos conectados governamentais para disponibilizar tal oferta de dados. Entretanto, conforme

Tabela 32 – Notas emitidas para avaliação geral do modelo de processo *"Piece of Cake"*

Avaliador 1	**Avaliador 2**	**Avaliador 3**	**Avaliador 4**	**Avaliador 5**	**Média**
60	40	70	90	99	71,8

Fonte: Autor desta dissertação, 2015.

exposto, a atual versão do modelo possui alguns itens que merecem ser aprimorados. Reforçando este entendimento, na avaliação geral, apesar do modelo ter obtido uma média satisfatória, a existência de duas notas com valores mais desfavoráveis indicam uma maior reflexão sobre as deficiências do modelo. Em observação à documentação produzida pelas equipes, alguns tópicos relevantes foram destacados por todas ou pelo menos uma das equipes e serão discutidos nas subseções a seguir.

6.4.4.1 Detalhamento do Modelo

Este item foi objeto de diversas observações, especialmente feitas pelas equipes de maior experiência (A e B). Em várias atividades, foi registrado que foi apresentado uma descrição do que se trata a atividade (o que fazer?) sem haver o devido detalhamento de como se executar a atividade (como fazer?). Uma das equipes sugeriu que cada atividade fosse estruturada mediante o seguinte template disposto na Tabela **??**.

Tabela 33 – Modelo para detalhamento das atividades do modelo de processo proposto pelos avaliadores

PROCESSO (A, B, C ou D):	
Atividade (Código da atividade) (Obrigatória / Desejável):	
Descrição da atividade:	
Como executar a atividade:	

Fonte: Autor desta dissertação, 2015.

Por se tratar de um modelo genérico de processo, a pesquisa optou por não aprofundar o nível de detalhamento das atividades. O objetivo principal do modelo consiste em guiar o publicador governamental de dados até a disponibilização da sua oferta de dados de forma aberta e conectada ou ainda, pelo menos de forma aberta.

Neste contexto, Sommerville (2007) complementa que, no caso de software, modelos genéricos não são descrições definitivas de processos de software. São abstrações do processo que podem ser utilizadas para diferentes abordagens do processo de software, podendo ser adaptadas e ampliadas para criar processos mais específicos. Para o modelo de processo *"Piece of Cake"* a pesquisa adota o mesmo entendimento que este autor, ou seja, trata-se de um modelo genérico sem descrições definitivas de processos.

Para o detalhamento de cada uma das atividades, existem opções de implementação que já foram estabelecidas noutros processos de publicação de dados abertos incluindo aqueles que foram citados na revisão de literatura. Desta maneira, o modelo *"Piece of Cake"* permite que o publicador de dados implemente cada atividade da melhor forma que for possível no seu contexto de aplicação.

Como o *"Piece of Cake"* ainda é uma proposta de modelo e ainda, considerando os inúmeros cenários possíveis de publicação de DACG, a pesquisa entende que um maior detalhamento de cada uma das atividades poderia inviabilizar a implementação do modelo em cenários específicos de publicação de dados abertos.

Entretanto, em aproveitamento à esta contribuição, a proposta de modelo passou a evidenciar a importância de que cada instituição publicadora, ao analisar as atividades propostas, busque identificar na literatura disponível qual a melhor forma de implementar cada atividade a depender do seu contexto de aplicação. Ademais, o detalhamento, bem como a incorporação de novas atividades ao modelo e a cada um dos seus processos, deve ser objeto de trabalhos futuros e derivados desta pesquisa.

6.4.4.2 Classificação de Atividades Obrigatórias e Desejáveis

Outra questão importante destacada pelos avaliadores foi que, a visão geral do modelo *"Piece of Cake"*, conforme a Figura 41 apresenta a existência de atividades obrigatórias e desejáveis, entretanto, para o estudo empírico, as equipes tiveram que analisar todas as atividades.

A proposta geral do modelo prevê a classificação de atividades obrigatórias e desejáveis. Para isto, os dois instrumentos de validação empírica foram desenvolvidos para gerarem subsídios a esta classificação.

Conforme exposto, da forma que o modelo foi implementado pelas equipes permitiu mensurar as atividades mais relevantes e em consonância com as respostas das questões Q6 e Q7, que, aliado aos resultados do modelo de regressão foi possível haver fundamentos para propor esta classificação.

6.4.4.3 Distribuição das Atividades dentre os Processos

Algumas atividades dispostas em determinado processo foram passíveis de sugestões para que sejam realocadas noutros processos. Os avaliadores manifestaram tais sugestões explicitamente, registrando na planilha fornecida para o estudo, bem como informando que determinada atividade foi executada num processo anterior. A Tabela 34 apresenta as sugestões de realocação de atividades:

As justificativas apresentadas pelos avaliadores não foram consideradas como altamente relevantes para fundamentar uma realocação destas atividades no âmbito desta pesquisa. Todavia, é reconhecido que, em trabalhos futuros, uma discussão mais am-

Tabela 34 – Comentários e sugestões dos avaliadores propondo realocação de atividades entre os processos do modelo

Processo	**Atividade(s)**	**Comentário dos avaliadores**	**Sugestão dos avaliadores**
B	"Estabelecer URIs neutras" e "Estabelecer URIs persistentes, que não se alterem em nenhum momento"	As atividades foram desenvolvidas no processo A	Realocar a atividade para o processo A
	"Proporcionar pelo menos um recurso de dados em formato que seja legível por máquina para cada URI"	A criação de URIs pode ser desenvolvida já a partir da criação de dados abertos 3-estrelas	A atividade poderia ser incorporada como um requisito de qualidade de alguma atividade relacionada a etapa "Anunciar os conjuntos de dados ao público"
	"Usar URIs HTTP para que pessoas e máquinas possam encontrá-las via *Web* utilizando estes endereços"	Ao se criar o repositório de disponibilização dos dados, esta atividade foi desenvolvida conjuntamente	A atividade poderia ser incorporada como um requisito de qualidade de alguma atividade relacionada a etapa "Anunciar os conjuntos de dados ao público"
	"Utilizar datas em URIs com moderação" e "Utilizar hashs (#) em URIs cautelosamente"	Atividade desenvolvida conjuntamente com outras atividades da etapa "Modelar os Dados"	Atividade pode ser incorporada como requisito de qualidade de alguma atividade da etapa "Modelar os dados"
D	"Desenvolver ou utilizar ontologias para estruturar a semântica dos dados"	A atividade foi feita no processo C	Realocar a atividade para o processo C

Fonte: Autor desta dissertação, 2015.

pla bem como a execução de novos estudos empíricos com populações mais significativas podem gerar mais subsídios para uma nova proposta de distribuição destas atividades.

6.4.4.4 Ferramentas Adicionais

Uma contribuição relevante (e não prevista no escopo original da pesquisa) consistiu na apresentação de ferramentas para apoiar o processo de publicação de DACG. Ao longo do estudo empírico, as equipes utilizaram um conjunto de softwares e recursos adicionais para apoiar as atividades de conversão. A Tabela 35 apresenta uma sumarização das ferramentas utilizadas pelos avaliadores no estudo empírico:

É importante registrar que, conforme a Tabela **??**, uma grande variedade de softwa-

Tabela 35 – Ferramentas e softwares utilizados pelos avaliadores para apoiar a execução de atividades do modelo

Etapa:	**Ferramenta:**	**Funcionalidade:**
Modelar os Dados	Protegé	Editor de Ontologias
Converter e Enriquecer Dados	ConvertCSV [6]	- Conversão de arquivos csv para diversos formatos
	Bloco de notas (*notepad*)	- Conversão de arquivos csv para formato txt
	Microsoft Excel	- Conversão de arquivos csv para formatos xls e xlsx - Gerenciamento de atividades desenvolvidas
	Libre Office Calc	- Conversão de arquivos csv para formatos xls e xlsx
	Google Refine [7], RDF123 [8]	- Conversão de arquivos em formatos estruturados para arquivos descritos em RDF
Anunciar Conjuntos de Dados ao Público	Google Sites[9], Blogger[10], DataHub [11] e Google Drive[12]	- Publicação de conjuntos de dados na *Web*
Diversas etapas	Google Docs [13]	- Registro de atividades
	Slack[14]	- Gerenciamento de atividades desenvolvidas

Fonte: Autor desta dissertação, 2015.

res foi utilizada durante o estudo empírico. Para o estágio atual da pesquisa, isto pode representar um elemento dificultador para a utilização do modelo de processo noutras experiências de publicação de dados abertos, considerando a necessidade de conhecimento em tais ferramentas. Por outro lado, trata-se de uma grande oportunidade de trabalhos futuros voltados ao desenvolvimento de ferramentas que simplifiquem atividades e preferencialmente integrem tarefas otimizando e reduzindo o esforço e a complexidade dos processos de publicação de dados abertos e dados abertos conectados.

6.4.4.5 Atividades do modelo não-desenvolvidas no estudo

Conforme exposto na subseção relacionada ao objetivo G3, diversas atividades do modelo de processo não foram executadas. Dentre os principais motivos mencionados

pelos avaliadores temos:

- O modelo de processo não apresentou detalhamento suficiente para a execução da atividade;
- A atividade não foi considerada necessária para o estudo empírico;
- A atividade foi desenvolvida conjuntamente com outra atividade;
- A descrição da atividade não foi útil, pois a equipe sabia como desenvolvê-la de outra maneira mais eficiente;
- A equipe não teve conhecimento técnico suficiente para desenvolver a atividade.

A pesquisa considera que todos estes motivos relacionados representam oportunidades para desenvolvimento de trabalhos futuros. É possível avaliar que a consolidação desta proposta num modelo para que seja adotado em larga escala requer o desenvolvimento de novos estudos empíricos complementares com objetos mais focados, bem como contemplando populações maiores. Para o escopo desta pesquisa, não foram desenvolvidas ações complementares para tratar os motivos mencionados.

6.5 Discussões gerais

Nesta seção apresentamos algumas discussões relevantes a alguns pontos de intersecção dos dois métodos de validação empírica utilizados na pesquisa.

6.5.1 Definição das atividades obrigatórias e desejáveis para os processos do modelo *"Piece of Cake"*

Um ponto forte desta pesquisa consiste na apresentação de conjuntos distintos de atividades para publicação de dados abertos e dados abertos conectados para o setor público que possam ser utilizados por instituições publicadoras com nível de maturidade baixo ou avançado. Assim, os dois instrumentos de validação empírica contemplaram métodos para a classificação das atividades do modelo de processo *"Piece of Cake"* em atividades obrigatórias e desejáveis.

O modelo de regressão, conforme explanado, validou um subconjunto de recomendações escolhidas pelo pesquisador e incorporou novas recomendações a partir dos dados obtidos pelo questionário de avaliação da dificuldade e relevância destas recomendações. Por outro lado, para o estudo empírico, a pesquisa atribuiu o seguinte critério de classificação:

- **Obrigatórias:** Atividades utilizadas pelas três das equipes que executaram o estudo empírico. Este critério é justificado porquê tais atividades foram necessárias para a

publicação dos dados abertos conectados governamentais por todas ou pela maioria das equipes.

- **Desejáveis:** Atividades utilizadas por duas, uma ou nenhuma das equipes que executaram o estudo empírico. Neste caso, estas atividades não foram utilizadas por todas as equipes, o que nos permite inferir que podem ser facultativas.

Além disto, o pesquisador utilizou a sua experiência teórica e prática no tema da pesquisa para estabelecer como obrigatórias algumas atividades que não foram sugeridas pelos instrumentos de validação empírica.

Assim, a Tabela 36 apresenta as atividades obrigatórias do modelo *"Piece of Cake"* com o respectivo instrumento que a classificou como tal. A maioria das atividades obrigatórias tiveram como origem o modelo de regressão, sendo complementada por atividades desenvolvidas pelas 3 equipes que participaram do estudo empírico. Ademais, a atividade 6H, referente ao processo D, bem como as atividades 0A, 11A e 11B, referentes as etapas de identificação da maturidade da instituição publicadora e de Retrospectiva, foram incorporadas por decisão do pesquisador.

Tabela 36 – Origem dos subsídios para classificação das atividades obrigatórias do modelo de processo *"Piece of Cake"*

Atividades:	Origem		
	Modelo de Regressão	Estudo Empírico	Decisão do Pesquisador
Identificar nível de maturidade da organização em publicação de dados (0A)			X
Identificar as partes interessadas (1A)	X	X	
Identificar os benefícios para a abertura de dados (1B)	X	X	
Definir perfis profissionais a serem envolvidos (1C)	X	X	
Definir grupos de usuários dos dados (1D)	X		
Analisar a estrutura organizacional da instituição publicadora (2A)	X		
Estabelecer diretrizes que orientem a priorização da publicação de dados abertos (2B)	X		

Realizar consultas aos usuários sobre a demanda de dados (2C)	X		
Identificar os dados que serão abertos (2D)	X		
Definir nível de maturidade dos dados a serem publicados (1-5 estrelas) (2E)	X		
Analisar o esforço para abertura de dados (2H)	X		
Identificar dados que podem ser conectados (2K)	X		
Gerar cópias de segurança das bases de dados que serão abertas (3A)	X	X	
Higienizar os dados (3B)	X		
Estabelecer rotinas de conversão de dados para formatos legíveis por máquina (3C)		X	
Analisar se os dados serão conectados ou não (3F)	X		
Estabelecer ou aprimorar documentação de dados (esquemas, vocabulários e ontologias) (3G)		X	
Adotar licenças de uso dos dados não restritivas (4A)	X	X	
Utilizar URIs para conectar os dados (5A)	X		
Estabelecer URIs persistentes, que não se alterem em nenhum momento (5B)	X		
Proporcionar pelo menos um recurso de dados em formato que seja legível por máquina para cada URI (5C)	X		
Usar URIs como nomes para as coisas (5D)	X		
Estabelecer os metadados obrigatórios (6A)	X		
Criar um esquema de dados para cada conjunto de dados (6B)	X	X	

Publicar esquemas de dados em arquivos diferentes (6D)	X	X	
Estabelecer critérios de escolha de vocabulários (6F)	X		
Desenvolver ou utilizar ontologias para estruturar a semântica dos dados (6H)			X
Converter dados para múltiplas finalidades e usos (7A)	X	X	
Conectar conjuntos de dados com outros dados relacionados (7C)	X		
Permitir o envolvimento de várias pessoas na identificação de como os dados a serem convertidos se relacionam com outros dados (7D)	X		
Disponibilizar bases completas para *download* (*dumps*) (8A)	X		
Estabelecer um Mapa de Decisões Tecnológicas (8B)	X		
Estabelecer dados tecnicamente e legalmente abertos (9A)	X		
Publicar metadados junto aos dados (9B)	X	X	
Disponibilizar os dados com o menor custo possível ao usuário, preferencialmente de modo gratuito na internet (9C)	X		
Melhorar os dados para serem melhor divulgados e encontrados por máquinas (9H)	X		
Disponibilizar dados conectados em servidores de triplas (9I)	X		
Estabelecer com clareza que o processo de publicação contempla etapas de manutenção e atualização dos dados (10A)	X		

Estabelecer mecanismos de monitoramento e avaliação da oferta de dados disponibilizados ao público (10B	X		
Fazer retrospectiva (11A)			X
Tomar decisão sobre continuidade do processo (11B)			X

Fonte: Autor desta dissertação, 2015.

Assim, as demais atividades que não foram classificadas como obrigatórias serviram para compor o conjunto de atividades que devem ser implementadas opcionalmente, voltadas a instituições publicadoras com maior maturidade.

Importante destacar que esta classificação é uma proposta com base nos critérios acima estabelecidos. Este item pode ser passível de aprimoramentos mediante o desenvolvimento de novas validações empíricas, através da incorporação de novas atividades ao modelo de processo bem como, a fusão de atividades.

7 CONCLUSÕES E TRABALHOS FUTUROS

Conforme exposto ao longo desta pesquisa, a publicação de dados abertos governamentais de forma conectada representa uma grande contribuição a produção de conhecimento sobre o setor público. Do ponto de vista computacional, esta contribuição é ainda mais relevante, pois os dados produzidos neste formato dispõem de melhor qualidade computacional bem como de recursos semânticos que permitem a produção de conhecimento e geração de subsídios para novas pesquisas, projetos e até novos negócios baseados na economia do conhecimento.

No capítulo 2 buscamos apresentar um rico referencial teórico que permite ao pesquisador que ler este trabalho ter uma visão ampla a respeito do contexto de publicação de dados abertos na esfera governamental, bem como a conceituação sobre dados abertos conectados e sua aplicação no setor público. Além disso, são apresentados modelos de maturidade e de processo de software fundamentais a concepção e proposição do modelo *"Piece of Cake"*. O capítulo 3 apresenta trabalhos relacionados de grande relevância para a pesquisa e o capítulo 4 apresenta uma sistematização das atividades para publicação de dados abertos (governamentais) e dados abertos conectados (governamentais) a partir de 15 processos de publicação desenvolvidos pela comunidade científica e instituições governamentais.

O capítulo 5 detalha as principais contribuições decorrentes desta pesquisa de mestrado, que esperamos fornecer para o setor público, bem como para a academia e para a indústria. Trata-se de uma área ampla e muito recente, com forte predominância na União Europeia e que deve se expandir para outros continentes e nações mundiais nos próximos anos. Esta pesquisa, por ter sido desenvolvida no Brasil, pode contribuir para fortalecer o protagonismo do nosso país no que tange ao Governo Aberto, Ecossistema de Dados Abertos bem como obter liderança na pesquisa e produção de produtos e serviços baseados em dados abertos conectados. A proposta teve sua validação empírica descrita no capítulo 6, conforme apresentado.

Considerando a validação apresentadas, a pesquisa atingiu o objetivo proposto neste trabalho, apoiando a publicação de DACG inclusive por atores sem nenhuma experiência com o assunto, mediante um modelo que os guiou dentre as atividades necessárias à publicação de tais dados. Entretanto, esta proposta possui algumas limitações e oportunidades de melhoria, como o seu nível de detalhamento, a indisponibilidade de ferramentas que apoiem o modelo de processo, bem como por não ter apresentado uma proposta mais robusta para o processo D, voltado a publicação de dados 5-Estrelas de forma aprimorada.

As principais dificuldades da pesquisa estão relacionados ao seu aparente ineditismo, já que a revisão de literatura pesquisada não identificou proposta semelhante, o que motivou um esforço significativo para incorporar à pesquisa conceitos relevantes da Engenharia de

Software como os modelos de processo. Além disso, a pesquisa teve restrições para a sua validação, especialmente por não ter sido possível a execução de estudos de caso em ambientes reais.

7.1 Principais contribuições da pesquisa

O levantamento do estado da arte nos mostrou que ainda existem algumas limitações nas abordagens existentes para publicação de dados abertos governamentais e dados abertos conectados governamentais. Além disso, é importante destacar que as pesquisas consultadas foram desenvolvidas antes da publicação das *"Melhores Práticas para publicação de Dados Conectados" (BPLDs)* do W3C e que, não foram identificadas pesquisas recentes que incorporassem tais práticas. Desta maneira, foi possível identificar quais ações os processos disponíveis já desenvolviam que são compatíveis com as práticas do W3C, integrando-as e associando-as. Desta maneira, esta pesquisa apresenta uma contribuição direta à disseminação das BPLDs com uma proposta estruturada de implementação em qualquer atividade de publicação de dados governamentais.

Outra contribuição interessante consiste na proposta de um modelo de processo. Ou seja, um modelo genérico que permite que instituições em estágios variados de maturidade em produção de dados possam utilizá-lo para guiar suas tarefas de publicação de dados governamentais. Entretanto, para que este modelo seja utilizado em larga escala, é necessário avançar na otimização de atividades necessárias à publicação de DACG bem com no desenvolvimento de ferramentas e softwares que auxiliem esta atividade.

7.2 Trabalhos futuros

Como trabalhos futuros esta pesquisa propõe:

1. Aprimorar a proposição de novas atividades para os processos do modelo, mediante a execução de investigações empíricas que permitam uma melhor distribuição das atividades bem como o seu detalhamento, quando for o caso;

2. Discussões sobre o desenvolvimento de processos específicos às necessidades particulares do setor público. Poderão ser desenvolvidos processos de publicação de DACG aplicáveis a contextos particulares como a saúde, educação, energia e recursos hídricos, geociências, dentre outros, caso hajam elementos que justifiquem tal nível de especificidade;

3. Pesquisas e produtos que visem a simplificação das atividades de publicação de DACG bem como a sua popularização dentre as instituições governamentais;

4. Desenvolvimento de ferramentas e softwares que simplifiquem a execução das atividades necessárias a publicação de dados abertos conectados governamentais;

5. Validar o modelo de processo atual com estudos de caso reais, preferencialmente em cenários e países distintos para verificar a validade do modelo em contextos técnicos, organizacionais e políticos distintos;

6. Eliminar as limitações apresentadas por esta pesquisa.

7. Publicar artigos em periódicos relevantes para a área de pesquisa visando disseminar o conhecimento produzido por esta pesquisa para a comunidade científica.

8. Apresentar trabalhos em congressos e eventos técnicos e científicos para disseminar este trabalho junto às instituições governamentais e demais interessados na melhoria da qualidade da oferta de dados públicos.

REFERÊNCIAS

ABIB, J. C.; KIRNER, T. G. GQM-PLAN: ferramenta para apoiar avaliações de qualidade de software. In: CONFERÊNCIA INTERNACIONAL DE TECNOLOGIA DE SOFTWARE, 9., 1998, Curitiba, Brasil. *Anais ...* Curitiba, 1998. p. 119–130.

ACM. *ACM recommendation on open government*. 2009. Acesso em: 12 out. 2015. Disponível em: <http://www.acm.org/public-policy/open-government>.

ALCANTARA, W. et al. Desafios no uso de dados abertos conectados na educação brasileira. In: WORKSHOP DE DESAFIOS DA COMPUTAÇÃO APLICADA À EDUCAÇÃO, 4., 2015, Recife, Brasil. *Anais ...* Recife: CSBC, 2015.

ARAÚJO, L. S. d. O. et al. *Uma ontologia das classificações da despesa do orçamento federal*. Brasília, Brasil, 2013. Acesso em: 15 out. 2015. Disponível em: <http://www.orcamentofederal.gov.br/biblioteca/estudos_e_pesquisas/Artigo_Ontobras_Orcamento_Aberto.pdf>.

AUER, S. et al. Managing the life-cycle of linked data with the LOD2 stack. *The Semantic Web - Lecture Notes in Computer Science*, v. 7650, n. 257943, p. 16, 2012.

BANDEIRA, J. et al. Dados abertos conectados. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO, 3., 2014, Maceió, Brasil. *Anais ...* Maceió: SBTI, 2014. Acesso em: 15 out. 2015. Disponível em: <http://nees.com.br/ojs/index.php/SBTI3/article/view/7/10>.

BANDEIRA, J. et al. Dados abertos conectados para a educação. *Jornada de Atualização em Informática na Educação*, Revista Brasileira de Informática na Educação, v. 4, n. 1, p. 47–69, 2015. Acesso em: 20 out. 2015. Disponível em: <http://www.br-ie.org/pub/index.php/pie/article/view/3551>.

BASILI, V. R. The experimental paradigm in software engineering. In: INTERNATIONAL WORKSHOP ON EXPERIMENTAL SOFTWARE ENGINEERING ISSUES: CRITICAL ASSESSMENT AND FUTURE DIRECTIONS, London, UK. *Proceedings ...* London: Springer – Verlag, 1993. p. 3–12. ISBN 3-540-57092-6.

BASILI, V. R.; CALDIERA, G.; ROMBACH, H. D. The goal question metric approach. In: *Encyclopedia of Software Engineering*. [S.l.]: John Wiley & Sons, 1994.

BAUER, F.; KALTENBÖCK, M. *Linked open data: the essentials - a quick start guide for decision makers*. [S.l.]: Semantic Web Company, 2012. 59 p. ISBN 9783902796059.

BENTO, A. V. Como fazer uma revisão da literatura: considerações teóricas e práticas. *Revista JA (Associação Acadêmica da Universidade da Madeira)*, p. 42–44, 2012. ISSN 0363-6127. Acesso em: 23 out. 2015. Disponível em: <http://www3.uma.pt/bento/Repositorio/Revisaodaliteratura.pdf>.

BERNERS-LEE, T. *Linked data*. 2006. Acesso em: 18 out. 2015. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>.

BOEHM, B. A spiral model of software development and enhancement. *SIGSOFT Softw. Eng. Notes*, ACM, New York, NY, USA, v. 11, n. 4, p. 14–24, ago. 1986. ISSN 0163-5948.

BRASIL. *Cartilha técnica para publicação de dados abertos no Brasil*. Brasília, Brasil, 2011. 10 p. Acesso em: 25 set. 2015. Disponível em: <http://dados.gov.br/cartilha-publicacao-dados-abertos/>.

BRASIL. *Lei nº. 12.527, de 18 de Novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências.* Diário Oficial [da] República Federativa do Brasil, Brasília, DF, Brasil, 19 nov, 2011. Acesso em: 25 set. 2015. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>.

BRASIL. *Kit de dados abertos*. Brasília, Brasil, 2014. 3–5 p. Acesso em: 25 set. 2015. Disponível em: <http://kit.dados.gov.br>.

BRASIL. *Manual para elaboração de plano de dados abertos*. Brasília, Brasil, 2014. v. 7, 38 p. Acesso em: 25 set. 2015. Disponível em: <http://www.planejamento.gov.br/secretarias/upload/Arquivos/governo_aberto/manual_elaboracao_plano_dados_abertos.pdf>.

BRASIL. *Plano de dados abertos - Ministério do Planejamento, Orçamento e Gestão do Brasil*. Brasília, Brasil, 2014. 1–15 p. Acesso em: 25 set. 2015. Disponível em: <http://www.planejamento.gov.br/secretarias/upload/Arquivos/governo_aberto/plano_dados_abertos.pdf>.

CARRION, A. P.; WERNER, C. *Modelos de processo – técnicas inteligentes que apóiam a construção de um software*. 2013. Acesso em: 10 out. 2015. Disponível em: <http://web.unipar.br/~seinpar/2013/artigos/Ana%20Paula%20Carrion%202.pdf>.

CEWEB.BR. *Dados conectados*. 2015. Acesso em: 06 nov. 2015. Disponível em: <http://ceweb.br/guias/web-semantica/capitulo-5/>.

CHILE. *Datos abiertos para chile - guía rápida de publicación*. Santiago, 2013. 8 p. Acesso em: 27 set. 2015. Disponível em: <http://datos.gob.cl/assets/files/GuiaRapidaPublicacionDatosAbietos_v02.1.pdf>.

CHILE. *Norma técnica para publicación de datos abiertos en chile*. Santiago, Chile, 2013. 1–28 p. Acesso em: 27 set. 2015. Disponível em: <http://instituciones.gobiernoabierto.cl/NormaTecnicaPublicacionDatosChile_v2-1.pdf>.

COLOMBIA. *Guía para la apertura de datos en Colombia*. Bogotá, Colômbia, 2012. 67 p. Acesso em: 27 set. 2015. Disponível em: <http://programa.gobiernoenlinea.gov.co/apc-aa-files/da4567033d075590cd3050598756222c/Datos_Abiertos_Guia_v2_0.pdf>.

COMSODE. *Documents of practice for methodology for publishing datasets as open data - COMSODE*. [S.l.], 2014. 1–58 p. Acesso em: 28 set. 2015. Disponível em: <http://www.comsode.eu/wp-content/uploads/Annex1_D5.1-Documentation_of_practices.pdf>.

COMSODE. *Methodology for publishing datasets as open data - COMSODE*. [S.l.], 2014. 1–31 p. Acesso em: 28 set. 2015. Disponível em: <http://www.comsode.eu/index.php/deliverables/>.

CONSOLI, S. et al. Geolinked Open Data for the Municipality of Catania. In: INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS, 4., 2014, Thessaloniki, Greece. *Proceedings ...* New York, USA: ACM, 2014. (WIMS '14), p. 1–8. ISBN 978-1-4503-2538-7. Disponível em: <http://doi.acm.org/10.1145/2611040.2611092>.

DBPEDIA. *About dbpedia*. 2015. Acesso em: 09 nov. 2015. Disponível em: <http://wiki.dbpedia.org/about>.

DING, L. et al. TWC LOGD: a portal for linked open government data ecosystems. *Journal of Web Semantics*, Elsevier B.V., v. 9, n. 3, p. 325–333, 2011. ISSN 15708268.

DING, L.; PERISTERAS, V.; HAUSENBLAS, M. Linked open government data. *IEEE Intelligent Systems*, v. 27, n. January 2010, p. 11–15, 2012. ISSN 15411672.

EAVES, D. *The three laws of open government data*. 2009. Acesso em: 25 set. 2015. Disponível em: <http://eaves.ca/2009/09/30/three-law-of-open-government-data/>.

ECUADOR. *Guia de política pública de datos abiertos*. Quito, Ecuador, 2014. 21 p. Acesso em: 28 set. 2015. Disponível em: <http://www.gobiernoelectronico.gob.ec/wp-content/uploads/2014/12/GPP-DA-v01-20141128-SNAP-SGE.pdf>.

EMC. *The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east*. EMC Corporation, 2012. Acesso em: 13 out. 2015. Disponível em: <http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf>.

ENCINOSA, J. B. Cuestionarios. In: _____. La Habana, Cuba: [s.n.], 2006. cap. 15. Sistemas de información para el economista y el contador.

ENGLISH, L. P. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. [S.l.]: Wiley, 1999. 544 p. ISBN 978-0-471-25383-9.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, v. 31, n. 7, p. 799–815, 2004. Disponível em: <http://dx.doi.org/10.1080/0266476042000214501>.

GALIOTOU, E.; FRAGKOU, P. Applying linked data technologies to greek open government data: a case study. *Procedia - Social and Behavioral Sciences*, v. 73, p. 479–486, 2013. ISSN 18770428.

GÓMEZ-PÉREZ, A.; SUÁREZ-FIGUEROA, M. C. NeOn methodology for building ontology networks: a scenario-based methodology. In: INTERNATIONAL CONFERENCE ON SOFTWARE, SERVICES & SEMANTIC TECHNOLOGIES, 1., 2009, Sofia, Bulgaria. *Proceedings ...* Sofia, Bulgaria, 2009, (S3T 2009). ISBN 978-954-9526-62-2.

GOVERNO DO ESTADO DE SÃO PAULO. *SPUK - Melhorando o ambiente de negócios por meio da transparência no Governo de São Paulo*. 2015. Acesso em: 06 nov. 2015. Disponível em: <http://pt.slideshare.net/SPUK/spuk-visao-geral>.

GUIMARÃES, C. B. d. S.; DINIZ, N. V. Como os dados abertos contribuem para um governo aberto. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO, 3., 2014, Maceió, Brasil. *Anais ...* Maceió: SBTI, 2014. Acesso em: 15 out. 2015. Disponível em: <http://nees.com.br/ojs/index.php/AnaisSBTI/article/view/5/1>.

HAND, D. J. Data not dogma: big data, open data and the opportunities ahead. In: INTERNATIONAL SYMPOSIUM ADVANCES IN INTELLIGENT DATA ANALYSIS, 12., 2013, London, UK. *Proceedings ...* London, 2013. (Lecture Notes in Computer Science), p. 1–12. ISBN 978-3-642-41397-1.

HEATH, T. Linked data - welcome to the data network. *Internet Computing, IEEE*, v. 15, n. 6, p. 70–73, Nov 2011. ISSN 1089-7801.

HEATH, T.; BIZER, C. *Linked data: evolving the web into a global data space*. [S.l.]: Morgan & Claypool, 2011. 136 p. (Synthesis Lectures on Web Engineering Series). ISBN 9781608454303.

HENDLER, J. et al. US government linked open data: semantic.data.gov. *IEEE Intelligent Systems*, v. 27, n. June, p. 25–31, 2012. ISSN 15411672.

HÜNER, K. M.; OFNER, M.; OTTO, B. Towards a maturity model for corporate data quality management. In: SYMPOSIUM ON APPLIED COMPUTING, 24., 2009, Honolulu, Hawaii. *Proceedings ...* New York, NY, USA: ACM, 2009. (SAC '09), p. 231–238. ISBN 978-1-60558-166-8.

HYLAND, B.; WOOD, D. The joy of data - a cookbook for publishing linked government data on the Web. In: WOOD, DAVID. [S.l.], 2011. cap. 1, p. 3–25. Linking Government Data.

ISA. *Study on business models for linked open government data*. [S.l.], 2013. Acesso em: 11 nov. 2015. Disponível em: <http://ec.europa.eu/isa/documents/study-on-business-models-open-government_en.pdf>.

ISA. *How linked data is transforming egovernment*. [S.l.], 2014. Acesso em: 13 nov. 2015. Disponível em: <http://ec.europa.eu/isa/documents/publications/how-linked-data-20140711_en.pdf>.

ISOTANI, S.; BITTENCOURT, I. I. *Dados abertos conectados*. São Paulo, Brasil: Novatec, 2015. 175 p. ISBN 978-85-7522-449-6.

JANNUZZI, P. d. M. *Indicadores de monitoramento e avaliação de políticas e programas sociais*. Brasília, Brasil: [s.n.], 2012.

JANSSEN, K. The influence of the psi directive on open government data: an overview of recent developments. *Government Information Quarterly*, v. 28, n. 4, p. 446–456, 2011. ISSN 0740-624X.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, Taylor & Francis, v. 29, n. 4, p. 258–268, 2012.

KOBILAROV, G. et al. Media meets semantic web – how the bbc uses dbpedia and linked data to make connections. In: AROYO, L. et al. (Ed.). *The Semantic Web: Research and Applications*. [S.l.]: Springer Berlin Heidelberg, 2009, (Lecture Notes in Computer Science, v. 5554). p. 723–737. ISBN 978-3-642-02120-6.

LBC. *Open data and use of standards: towards a better supply and distribution process for open Data.* [S.l.], 2012. Acesso em: 15 nov. 2015. Disponível em: <https://www.forumstandaardisatie.nl/fileadmin/os/documenten/Internationale_benchmark_v1_03_final.pdf>.

LEE, G.; KWAK, Y. H. An open government maturity model for social media-based public engagement. *Government Information Quarterly*, v. 29, n. 4, p. 492 – 503, 2012. ISSN 0740-624X. Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011).

LEHMANN, J. et al. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, v. 5, p. 1–29, 2014.

LÓSCIO, B. F. Ciclo de vida e ecossistema de dados abertos: etapas e atores envolvidos na abertura de dados. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO, 3., 2014, Maceió, Brasil. *Anais ...* Maceió: SBTI, 2014. Acesso em: 13 out. 2015. Disponível em: <http://nees.com.br/ojs/index.php/AnaisSBTI/article/view/6/2>.

MCKINSEY. *Big data: the next frontier for innovation, competition, and productivity.* [S.l.], 2011.

MEIJER, A.; THAENS; M. Public information strategies: making government information available to citizens. *Information Polity*, n. 14, p. 31–45, 2009.

MENDONÇA, R. R. d. et al. LOP - capturing and linking open provenance on LOD cycle. In: WORKSHOP ON SEMANTIC WEB INFORMATION MANAGEMENT, 5., 2013, New York, USA. *Proceedings ...* New York, USA: ACM Press, 2013. p. 1–8.

MOLINA, M. J. T. *Teoria científica e investigação.* 2002. Acesso em: 23 out. 2015. Disponível em: <http://www.molwick.com/pt/evolucao/560-teorias-cientificas.html>.

MUREDDU, F. et al. A new roadmap for next-generation policy-making. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF ELECTRONIC GOVERNANCE, 6., 2012,Albany, USA. *Proceedings ...* Albany, USA: ACM Press, 2012. (ICEGOV '12), p. 62–66. ISBN 978-1-4503-1200-4.

NEVES, O. M. d. C. Evolução das políticas de governo aberto no brasil. In: CONGRESSO CONSAD DE GESTÃO PÚBLICA, 6., 2013, Brasília, Brasil. *Anais ...* Brasília: CONSAD, 2013. Acesso em: 13 out. 2015. Disponível em: <http://consadnacional.org.br/wp-content/uploads/2013/05/092-EVOLU%C3%87%C3%83O-DAS-POL%C3%8DTICAS-DE-GOVERNO-ABERTO-NO-BRASIL.pdf>.

OBAMA, B. *Open government directive.* Washington DC, USA, 2009. 1–11 p. Acesso em: 10 out. 2015. Disponível em: <http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf>.

OBAMA, B. *Transparency and open government.* Washington DC, USA, 2009. 1–2 p. Acesso em: 10 out. 2015. Disponível em: <http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment>.

OGD. *The 8 principles of open government data.* 2007. Acesso em: 10 out. 2015. Disponível em: <http://www.opengovdata.org>.

OGP. *OGP participating countries*. 2015. Acesso em: 10 out. 2015. Disponível em: <http://www.opengovpartnership.org/countries>.

OGP. *What is the open government partnership?* 2015. Acesso em: 10 out. 2015. Disponível em: <http://www.opengovpartnership.org/about>.

OKF. *Guia de dados abertos*. 2015. Acesso em: 10 out. 2015. Disponível em: <http://opendatahandbook.org/guide/pt_BR>.

OKF. *Guide to open data Licensing*. 2015. Acesso em: 10 out. 2015. Disponível em: <http://opendefinition.org/guide/data/>.

OKF. *O que são dados abertos?* 2015. Acesso em: 10 out. 2015. Disponível em: <http://opendatahandbook.org/guide/pt_BR/what-is-open-data/>.

PMI. *Um guia do conhecimento em gerenciamento de projetos (guia PMBOK)*. 5ª. ed. Pennsylvania, USA: Project Management Institute, 2013. 595 p. ISSN 0389-4991. ISBN 9781628250091.

PRESSMAN, R. S. *Engenharia de software*. [S.l.]: Makron Books, 1995. 1088 p. ISBN 8534602379.

REILLY, T. O. Government as a platform. In: *Open Government: Collaboration, Transparency, and Participation in Practice*. [S.l.]: O'Reilly Media, Inc., 2010. v. 6, n. 1, cap. 2, p. 13–40.

ROCHA, Á.; VASCONCELOS, J. Os modelos de maturidade na gestão de sistemas de informação. Edições Universidade Fernando Pessoa, 2004.

SAUVÉ, J. *Identificação dos objetivos do estudo*. 2012. Acesso em: 25 out. 2015. Disponível em: <https://docs.google.com/presentation/d/1RQ0xXogJ2QlWjIIDlJNktpjPRAYv_T3OtfUxy0bzq1g/edit#slide=id.i19>.

SILVA, C. V. P. d. et al. *GQM - Goal–Question-Metric*. Recife, Brasil, 2009. Acesso em: 25 out. 2015. Disponível em: <http://www.cin.ufpe.br/~scbs/metricas/seminarios/GQM_texto.pdf>.

SMITHSON, M.; VERKUILEN, J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, American Psychological Association, v. 11, n. 1, p. 54, 2006.

SOMMERVILLE, I. *Engenharia de software*. [S.l.]: Pearson - Addison Wesley, 2007. 568 p. ISBN 978-85-88639-28-7.

URUGUAY. *Guía rápida de publicación em datos.gub.uy*. Montevideo, Uruguay, 2012. 17 p. Disponível em: <http://www.agesic.gub.uy/innovaportal/file/2478/1/guia_publicacion_datos_abiertos.pdf>.

VARASCHIM, J. D. *Implantando o SCRUM em um ambiente de desenvolvimento de produtos para internet*. Rio de Janeiro, Brazil, 2009.

ÁVILA, T. et al. A importância das plataformas de informação para apoio ao processo de planejamento: O caso do portal alagoas em dados e informações. In: CONGRESSO BRASILEIRO DE GESTÃO DO CONHECIMENTO, 11., 2012, São Paulo, Brasil. *Anais ...* São Paulo: SBGC, 2012. (KMBRASIL 2012). Acesso em: 15 out. 2015. Disponível em: <http://www.sbgc.org.br/sbgc/kmbrasil-2012/anais/pdf/TC32.pdf>.

VILLAZÓN-TERRAZAS, B.; SUÁREZ-FIGUEROA, M. C.; GÓMEZ-PÉREZ, A. A pattern-based method for re-engineering non-ontological resources into ontologies. *International Journal on Semantic Web and Information Systems (IJSWIS)*, v. 6, n. 4, p. 27 – 63, 2010.

VILLAZÓN-TERRAZAS, B. et al. Methodological guidelines for publishing government linked data. *Linking Government Data*, p. 27–49, 2011.

W3C. *Government linked data working group charter*. 2011. Acesso em: 13 out. 2015. Disponível em: <http://www.w3.org/2011/gld/charter.htm>.

W3C. *Data on the Web best practices working group*. 2013. Acesso em: 13 out. 2015. Disponível em: <http://www.w3.org/2013/dwbp>.

W3C. *W3C data activity building the Web of data*. 2013. Acesso em: 13 out. 2015. Disponível em: <http://www.w3.org/2013/data/>.

W3C. *Best practices for publishing linked data*. 2014. Acesso em: 12 out. 2015. Disponível em: <http://www.w3.org/TR/ld-bp/>.

W3C Brasil. *Manual dos dados abertos: governo*. 2011. Acesso em: 10 out. 2015. Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>.

W3C Brasil. *Publicação de dados em formato aberto*. 2013. Acesso em: 10 out. 2015. Disponível em: <http://cursos.ep.org.br/course/view.php?id=25&section=1>.

WOOD, D. et al. *Linked data: structured data on the Web*. [S.l.]: Manning Publications, 2013. 336 p. ISBN 9781617290398.

APÊNDICE A – PROPOSTA DE QUESTIONÁRIO PARA ORIENTAR A IDENTIFICAÇÃO DA MATURIDADE DA INSTITUIÇÃO PUBLICADORA DE DADOS GOVERNAMENTAIS

Apresentamos um conjunto de questões que podem ser utilizadas para identificar a maturidade institucional em publicação de dados governamentais.

1. Qual a área de atuação da instituição no setor público?
2. A instituição possui algum setor específico para produção e gestão de dados e informações?
3. Caso a resposta anterior for sim, quantas pessoas deste setor trabalham diretamente com atividades técnicas de produção e gestão de dados e informações?
    - 1-2
    - 3-5
    - 6-8
    - Mais que 8
4. Quais tecnologias de armazenamento de dados são utilizadas com maior frequência na instituição?
    - Planilhas Eletrônicas
    - Arquivos de Texto
    - Bancos de Dados Setoriais (arquivos do MS-Access, por exemplo)
    - Sistemas Gerenciadores de Bancos de Dados (SGBDs)
    - Bancos de Dados Geoespaciais
    - Outros
5. A instituição publica dados em formato aberto?
6. A instituição publica dados em formato aberto e conectado?
7. Os dados são disponibilizados aos usuários mediante uma API?
8. Os dados são disponibilizados aos usuários mediante um endpoint SPARQL?
9. Quais formatos de dados são disponibilizados com maior regularidade?
    - Documentos (formatos PDF, DOC, DOCX, ODT)

- Planilhas (formatos XLS, XLSX, ODS)
- Imagens (formatos JPG, PNG, GIF, etc.)
- Hipertexto (formatos HTML, XML, etc.)
- Arquivos compactados (formatos ZIP, RAR, etc)
- Geoespaciais (formatos KML, SHP, GeoJSON, etc)
- Descrição de Recursos (RDF/XML, RDFa, Turtle, JSON-LD, etc)
- Modelos de conhecimento (OWL, etc.)
- Outros

10. Avalie o nível de entendimento da instituição sobre a legislação vigente sobre acesso a informação:

    - Muito Baixo
    - Baixo
    - Médio
    - Alto
    - Muito Alto

11. A instituição possui um setor específico para atender as demandas de informações da sociedade?

12. A instituição publica dados periodicamente num catálogo de dados governamentais?

13. Caso sim, em que periodicidade?

    - Diariamente
    - Semanalmente
    - Quinzenalmente
    - Mensalmente
    - Trimestralmente
    - Semestralmente
    - Periodicidades mais longas (anual, bianual, etc.)

APÊNDICE B – QUESTIONÁRIO UTILIZADO PARA CLASSIFICAÇÃO DA DIFICULDADE E RELEVÂNCIA DAS RECOMENDAÇÕES EXTRAÍDAS DA REVISÃO DE LITERATURA

A versão online do questionário utilizado para esta atividade está disponível neste endereço: endereço eletrônico:
http://www.thiagoavila.net/research/masterThesis/formOnline01.txt.

Também está disponível um arquivo PDF com todas as questões no endereço eletrônico:
http://www.thiagoavila.net/research/masterThesis/formDificuldadeRelevancia.pdf

APÊNDICE C – QUESTIONÁRIO UTILIZADO PARA CLASSIFICAÇÃO DAS RECOMENDAÇÕES EXTRAÍDAS DA REVISÃO DE LITERATURA DE ACORDO COM O ESQUEMA 5-ESTRELAS DOS DADOS ABERTOS

A versão online do questionário utilizado para esta atividade está disponível no endereço eletrônico:
http://www.thiagoavila.net/research/masterThesis/formOnline02.txt.

Também está disponível baixar um arquivo PDF com todas as questões no endereço eletrônico:
http://www.thiagoavila.net/research/masterThesis/formEsquema5Estrelas.pdf.