Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Filosofia e Ciências

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

LARISSA PAVARINI DA LUZ

FRAMEWORK PARA PUBLICAÇÃO DE DADOS COM ÊNFASE EM ENRIQUECIMENTO E MAPEAMENTO SEMÂNTICO

Marília
2021

LARISSA PAVARINI DA LUZ

FRAMEWORK PARA PUBLICAÇÃO DE DADOS COM ÊNFASE EM ENRIQUECIMENTO E MAPEAMENTO SEMÂNTICO

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação, como requisito para obtenção de título de Doutora em Ciência da Informação, pela Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, campus de Marília.

Orientador: Prof. Dr. José Eduardo Santarem Segundo

Linha de Pesquisa: Informação e Tecnologia

Marília
2021

LARISSA PAVARINI DA LUZ

***FRAMEWORK* PARA PUBLICAÇÃO DE DADOS COM ÊNFASE EM ENRIQUECIMENTO E MAPEAMENTO SEMÂNTICO**

BANCA EXAMINADORA:

Prof. Dr. José Eduardo Santarem Segundo
Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP/Marília)
Programa de Pós-Graduação em Ciência da Informação Faculdade de Filosofia

Prof. Dr. Leonardo de Castro Botega
Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP/Marília)
Programa de Pós-Graduação em Ciência da Informação Faculdade de Filosofia

Prof. Dr. Edberto Ferneda
Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP/Marília)
Programa de Pós-Graduação em Ciência da Informação Faculdade de Filosofia

Prof. Dra. Ana Carolina Simionato Arakaki
Universidade Federal de São Carlos (UFSCar)
Programa de Pós-Graduação em Ciência da Informação - PPCGI/UFSCar

Prof. Dra. Márcia Regina da Silva
Universidade de São Paulo - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (USP)
Departamento de Educação, Informação e Comunicação

Marília, 23 de fevereiro de 2021.

larissapavarinidaluz@gmail.com

Luz, Larissa Pavarini da
L979f Framework para publicação de dados com ênfase em enriquecimento e mapeamento semântico / Larissa Pavarini da. – Marília, 2021.
278 f. : il. ; 30 cm.

Orientador: Prof. Dr. José Eduardo Santarém Segundo.
Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, 2021.
Bibliografia: f. 250-278.

1. Mapeamento semântico. 2. Enriquecimento de dados conectados. 3. Framework de dados conectados. 4. Abordagem de dados conectados. 5. Web semântica. 6. Web de dados. I. Santarém Segundo, José Eduardo (orientador). II. Título.

CDD 004.678

Ficha catalográfica elaborada por
Liliana Giusti Serra
CRB 8/5695

Dedico este trabalho à minha filha,

Helena Pavarini da Luz,

Minha razão de felicidade, presente de Deus, Luz da minha vida.

AGRADECIMENTO

Primeiramente a Deus, por ter me iluminado em mais uma jornada da minha carreira acadêmica e cumprindo o desejo do meu coração de fazer um doutorado, pelo seu infinito amor, carinho e saber que EM TODOS OS MOMENTOS esteve do meu lado.

À Unesp, ao Centro Paula Souza (CPS) e à Fatec Garça pela oportunidade de fazer o meu doutorado, tão sonhado e esperado, através de um programa tão especial entre essas unidades. Agradeço imensamente à Fatec Garça por ter permitido meu afastamento para a realização do programa de pós-graduação e, principalmente, ter confiado em mim e no meu trabalho.

À minha Família, André e Helena, por todo amor, apoio, compreensão e, principalmente, por entender as minhas ausências. Souberam, com amor, tolerar e compreender o meu estranho mau humor em determinados momentos desta pesquisa, com paciência e muita sabedoria. Meu imenso amor, agradecimentos e carinho.

Aos meus pais, Carlos e Juraci, e minhas irmãs, Carla e Fernanda, que sempre acreditaram em mim e me apoiaram incondicionalmente em toda minha jornada, que só tinham palavras de incentivo e zelo nesse momento tão complexo da minha vida.

Ao meu orientador e amigo, José Eduardo Santarem Segundo que se portou como só o fazem os mestres. Acreditando no meu trabalho, deu-me a liberdade necessária dividindo comigo as expectativas, conduziu-me a maiores reflexões e, desta forma, abrilhantou muito meu trabalho. Minha especial admiração e gratidão.

A todos os professores do programa de CI da Unesp de Marília que participaram desta jornada, sempre solícitos, até mesmo fora do horário do curso, porque sem eles não haveria enriquecedoras ideias. Meus sinceros agradecimentos.

Às minhas amigas do Clube da Luluzinha Unesp (Diana, Elaine, Maira, Liliana, Jéssica, Jacquy, Luciana, Mariana e Bete) que são especiais e queridas, pela divertida companhia nas disciplinas, qualificações e defesas, sempre torcendo umas pelas outras em tudo!

Às amigas, Ivania, Mirelys, Diana, Eliane Nascimento, Maria Fernanda e Tassiana, amigas, irmãs, amadas. Meu carinho imenso pelas conversas, orações, convivência e troca de experiências em momentos difíceis, dias de choro, risadas, mas principalmente em momentos únicos e especiais. Não tenho palavras para agradecer o quanto vocês são amadas e especiais na minha vida e por tudo que me ajudaram.

A minha amada Marta, essencial para tudo e em tudo, você é um presente de Deus na minha vida nesse ano de 2020 que foi tão difícil e você me ajudou a levantar e me reerguer mesmo não acreditando que poderia.

A minha teraupeta, incentivadora e acima de tudo amiga para tudo, Daniela Rosas, que me fez enxergar coisas lindas em mim mesmo, você é espetacular.

Aos meus queridos André Luis e João Baptista por estarem sempre presentes, mesmo que, às vezes, a distância falou mais alto, porém não a constância de serem tão importantes para mim e fazerem toda diferença nesse processo da minha vida, meu amor e carinho por vocês.

Aos melhores vizinhos da vida, Rogério e Fabiana por me ajudarem pacientemente nas horas difíceis. Meus agradecimentos com carinho.

À minha querida Haydee, amiga, vizinha e professora de português, pela correção da tese, sua alegria e motivação sempre comigo.

Aos meus primos irmãos José Luis Fiorelli Júnior e Wagner Pavarini Assen, e minha prima e melhor amiga Débora Pavarini, amo vocês demais e obrigada por tudo que fizeram por mim neste período.

Ao Pessoal do Laboratório (Laborartório de Desenvolvimento e Aplicação Multimídia) Clayton, Caio, Ana Maria, Jean e Sandra obrigada por serem especiais e companheiros, e por serem pessoas especiais e que Deus colocou na minha vida.

A todos professores e colegas de trabalho da FATEC de Garça, em especial aos do núcleo de Iniciação e Científica e Tecnológica da FATEC de Garça, Rafael, Maíza, Dener, Gustavo Camossi, Giovani, Lucas Carmichael e e em memória Alex Gomes. Sou-lhes muito grata por tudo e em tudo.

A todos os meus Alunos da FATEC de Garça, que me incetivaram... muito obrigada.

A tantos e todos que contribuíram de alguma forma na realização deste trabalho, nem todos os nomes estão citados, mas saibam da importância que fizeram nesse momento da minha vida, por terem acreditado, me motivado, enxugado minhas lágrimas, dado conselhos, me motivarem com muito amor. Levarei pela vida toda.

Os sonhos não determinam o lugar aonde vocês vão chegar, mas produzem a força necessária para tirá-los do lugar em que vocês estão. Sonhem com as estrelas para que vocês possam pisar pelo menos na Lua. Sonhem com a Lua para que vocês possam pisar pelo menos nos altos montes. Sonhem com os altos montes para que vocês possam ter dignidade quando atravessarem os vales das perdas e das frustrações.

Bons alunos aprendem a matemática numérica, alunos fascinantes vão além, aprendem a matemática da emoção, que não tem conta exata e que rompe a regra da lógica. Nessa matemática você só aprende a multiplicar quando aprende a dividir, só consegue ganhar quando aprende a perder, só consegue receber, quando aprende a se doar.

Augusto Cury (2003)

LUZ, Larissa Pavarini da. **Framework para publicação de dados com ênfase em enriquecimento e mapeamento semântico**. Orientador: José Eduardo Santarém Segundo. 2020. 278 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2021. Versões impressa e eletrônica.

RESUMO

Em um ambiente de dados como a *Web* a publicação de dados ainda é vista como um contexto de processo de descrição sem padronização, uma vez que, para cada domínio, criam-se diferentes formatos e modelos de dados. Para melhorar a interoperabilidade semântica sobre esses *datasets* ou repositórios heterogêneos, a *Web* Semântica tem sido adotada, permitindo, assim, a troca reuso ou a coleta de recursos digitais. O objetivo desse trabalho é propor um *Framework* com a finalidade de guiar, por meio de diretrizes e recomendações, o processo que gera todo o contexto necessário ao processo de extração, limpeza, enriquecimento e mapeamento de dados, descrevendo seus passos de forma organizada e sequencial, com o objetivo de seguir as melhores práticas de publicação de dados na *Web*. A metodologia define-se como pesquisa descritiva e técnica de análise dos dados e de conteúdo, em que foram aplicados métodos como revisão bibliográfica - definido para o desenvolvimento da investigação -, bem como os resultados de uma revisão sistemática da literatura, que evidencia a singularidade da proposta. Este trabalho determinou cinco diretrizes que geram o estado em questão, analisando alguns projetos correlatos e todo o contexto envolvido para que se chegue ao processo de enriquecimento e mapeamento, dando ênfase aos processos necessários que intervêm durante a proposta da tese, com o objetivo de estabeler diretrizes generalizadas. Para testar o *Framework* proposto foi realizada uma prova de conceito com as diretrizes identificadas em conjunto de dados do Metrô de São Paulo enriquecidos com outros conjuntos de dados disponíveis. Esta prova de conceito proporcionou o desenvolvimento de uma aplicação, com o propósito de conhecer todo o processo relacionado com enriquecimento e mapeamento de dados vinculados às boas práticas para publicação.

Palavras-Chave: Mapeamento semântico. Enriquecimento de dados conectados. *Framework* de dados conectados. Abordagem de dados conectados. Framework. Web semântica. Web de dados.

ABSTRACT

In a data environment such as the web, data publication is still seen as a context of a non-standardized description process, since for each domain different data formats and models are created. To improve semantic interoperability over these datasets or heterogeneous repositories, the Semantic Web has been adopted, thus allowing the exchange, reuse, or collection of digital resources. This work aims to propose a *Framework* to guide through orientations and recommendations the process needed to manage all the necessary context in some concerns in the process of extraction, cleaning, enrichment, and mapping of data, describing its steps in an organized and sequential manner, with the objective of following the best practices of data publishing on the web. The methodology is defined as a descriptive research with data and content analysis technique, with application of methods, such as bibliographic review during the development of the investigation, as well as the results of a systematic literature review, highlighting the uniqueness of the proposal. This work determined five guidelines that generate the state in question, analyzing some related projects, and the whole context that it involves in order to achieve the enrichment and mapping process, emphasizing the necessary processes that intervene during the thesis proposal, establishing generalized guidelines, and carrying out a proof of concept, using these guidelines, and map data from the São Paulo subway into linked data, enriching them through other data sets and making them available to all. Thus, a proof of concept was applied, creating an application that is based on the proposed guidelines of the *Framework* to know the whole process that applies when talking about data enrichment and mapping, linked to the good practices for publishing it.

Keywords: Semantic mapping. Linked data enrichment. Linked data Framework. Linked data approach. Semantic Web Framework. Web of data.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

ArchMS *Archetype Management System*

ARNDT *Agile Knowlegde and Semantic Web*

ARPAnet *Advanced Research Projects Agency Network*

CI Ciência da Informação

CC Ciência da Computação

CKAN *Comprehensive Knowledge Archive Network*

CERN *Organisation Européenne pour la Recherche Nucléaire*

CRIS *Community Research and Development Information Service for Science Research and Development*

D2QR Accessing Relational Databases as Virtual RDF Graphs

DCAT *Data Catalog Vocabulary*

DOSE *Distributed Open Semantic Platform*

Dublin Core *Dublin Core Metadata Initiative*

DWBP *Data on the Web Best Practices*

EHR *Electronic Healthcare Records*

FERMI *Formalization and Experimentation on the Retrieval Multimedia Information*

FP7 *Collaborative Manufacturing Network for Competitive Advantage*

FOAF *Friend of a Friend Project*

HTML *Hypertext Markup Language*

HTTP *HyperText Transfer Protocol*

IHC Interação Homem-Computador

IE *Internet Explorer*

IFF *Information Flow Framework*

ILOD *Intelligent Linked Open Data*

IMP *Information Manifold Project*

IOT *Internet of Things*

IR *Information Retrieval*

JSON *JavaScript Object Notation*

JSPJava *Java Server Pages Java*

KSE *Knowledge Sharing Effort*

LISA *Library and Information Science Abstract*

LOD *Linked Open Data*

LOV *Linked Open Vocabulary*

Luzzu *A Framework for Linked Data Quality Assessment*

MnM *M-OntoMat-Annotizer*

N3 *Notation3*

NoSQL *Not Only Structure Query Language*

NPI's *National Integrated Indicator*

NSFnet *National Science Foundation's Network*

OAI-PMH *Open Archives Initiative Protocol for Metadata Harvesting*

OLAP *Online Analytical Processing*

OAM *An Ontology Application Management Framework for Simplifying Ontology-Based Semantic Web Application Development*

OIL *Ontology Interchange Language*

OKBC *Open Knowledge BaseConnectivity*

OpLiDaF *Open Linked Data Framework*

OWL *Web Ontology Language*

PHP *Program Hypertext Preprocessor*

PowerMap *Mapping the Real Semantic Web on the Fly*

PPGCI Programa de Pós-Graduação em Ciência da Informação

RDF *Resource Description Framework*

RSS *Rich Site Summary*

Scopus *SciVerse Scopus*

SDL *Sapienza digital sistema*

SDW *Semantic Data Warehouse*

SGBD Sistema Gerenciador de Banco de Dados

SGML *Standard Generalized Markup Language*

SHOE *Simple HTML Extensions*

SKOS *Simple Knowledge Organization System*

SPARQL *Protocol and RDF Query Language*

SQL *Structura Query Language*

TCP/IP *Transmission Control Protocol/Internet Protocol*

TREC *Text Retrieval Conference*

UMLS *Unified Medical Lenguage System*

URIs *Uniform Resource Identifier*

URLs	*Uniform Resource Locator*
VANN	*Vocabulary for Annotations*
VOAF	*Vocabulary Of A Friend*
VoID	*Vocabulary of Interlinked Datasets*
XML	*Extensible Markup Language*
XOL	*Ontology eXchange Language*
W3C	*World Wide Web Consortium*
Webcrawler	Rastreador Web
WoS	*Web of Science*
WSML	*Web Service Modeling Language*
WWW	*World Wide Web*

SUMÁRIO

1 INTRODUÇÃO

Para possibilitar uma recuperação da informação com eficácia a fim de suprir as necessidades do usuário em qualquer contexto é fundamental que os dados estejam disponibilizados de forma organizada. No contexto da *Web* Semântica**,** tal ação significa compartilhar dados e fatos em vez de somente o texto de uma página.

Para auxiliar no debate, apresenta-se a *Web* Semântica, a qual foi proposta por Berners-Lee, Hendler e Lassila (2001), que a define como uma base tecnológica que apoia uma *Web* de Dados. O objetivo da *Web* de Dados é permitir que, baseado em algum ambiente ou domínio, sejam identificadas as necessidades de informação a serem buscadas, de forma que a máquina assuma um papel que antes não fazia na *Web* de Documentos, ou seja, o processo de agir na busca, um passo além da recuperação da informação, que pode levar ao desenvolvimento de sistemas que suportam dados confiáveis e com uma maior interatividade na *Web*.

A *Web* Semântica pode ser classificada como uma extensão *Web* de documentos, uma vez que devido ao exponencial crescimento e às diversas possibilidades de utilização de todas as suas tecnologias implícitas, é gerada uma infraestrutura tecnológica de comunicação com formatos variados. Nesse cenário, Berners-Lee, Hendler e Lassila (2001, p. 1) afirmam que “[...] a informação é dada com um significado bem definido, permitindo melhor interação entre os computadores e as pessoas”.

Assim, a *Web* tem passado por uma verdadeira revolução, e que conceitos como ontologias, vocabulários, metadados e marcadores semânticos se fazem presentes no contexto, transformando-se em uma *Web* de Dados, o que não descaracteriza as formas tradicionais de navegação, porém com uma nova forma de resultados.

Verifica-se, ainda, que o formato semântico, uma vez que sua finalidade é trazer resultados mais concisos e que possam estar próximos das necessidades dos usuários, sejam eles máquinas ou humanos, expresse uma forma de comunicação de um domínio e suas particularidades.

Os pontos citados, atrelados à *Web* Semântica, como os dados abertos e conectados, servem de base para estudos que podem variar desde a área

filosófica até mesmo às áreas técnicas, de modelos, de linguagens e mecanismos para descrever e gerenciar essa publicação de dados na *Web*.

Sua utilidade, segundo Berners-Lee, Hendler e Lassila (2001), não está relacionada somente à produção, leitura, disseminação e publicação de conteúdo na Internet, mas ele afirma que as tecnologias podem contribuir e que recursos podem ser usados e oferecidos no processo, uma vez que não se encontra uma padronização informacional do uso tecnológico em sua descrição digital. Isso está relacionado a como os dados estão conectados, e a como e onde as pessoas ou máquinas podem explorar essa nova *Web* de Dados, e ainda ser possível encontrar outros dados publicados que sejam relacionados a estes.

Neste contexto, Berners-Lee, Hendler e Lassila (2001) afirmam que a *Web* Semântica é um projeto que tem como objetivo dar significação ao dado, de modo a utilizar-se de tecnologias padronizadas e dados estruturados, a fim de gerar um ambiente provedor de resultados que enriqueçam e ampliem as possibilidades de busca do usuário, sejam eles uma máquina ou um humano. Assim, espera-se que o estudo seja uma vantagem para se obter uma diversidade de informação no contexto da publicação de dados na Web.

Dados heterogêneos ou de diferentes formatos (textos, mídias sociais, imagens, vídeos, entre outros) são produzidos, publicados e compartilhados diariamente na Internet. Verifica-se que sua estrutura semântica ainda não é precisa, uma vez que a sua heterogeneidade está relacionada a documentos dispostos em diferentes formatos na *Web*.

Ao refletir sobre toda a diferença que denota de tais necessidades de informação, nota-se que essas são tratadas em diferentes ambientes e áreas, e em algumas ocasiões as bases de dados não possuem interoperabilidade entre si. É neste momento que os conceitos e tecnologias da *Web* Semântica passam a fazer sentido, pois tornam possíveis que dados antes desconectados passem a operar de forma integrada e interoperável, de forma conectada (PATEL; JAIN, 2019). Assim, para que a Web Semântica seja aplicada e tenha exito, existe a necessidade das áreas se complementarem (HENDLER, 2001), ou seja, ela não alcançará sua completude se o conteúdo (metadados) e suas ferramentas não forem amplamente compartilhados.

Ao observar o início da *World Wide Web* (WWW), relaciona-se sua origem à expansão em massa de código aberto e de *softwares* livres, a disseminação de tecnologias de computação de baixo custo. No cenário da *Web* Semântica não é diferente, porém, além do olhar voltado às tecnologias para a disseminação de maneira a propiciar a interação entre o conhecimento humano e as máquinas, faz-se indispensável o desenvolvimento das competências do profissional da informação que irá atuar para atender às demandas do campo que estão em constante evolução, tanto no que se refere à delimitação da informação, quanto à informação propriamente dita, faz-se necessário controlar e gerenciar o grande volume de informação de acordo com diferentes comunidades (BARRETO, 1997).

As tecnologias da *Web* Semântica incluem diferentes formatos para troca de dados, como por exemplo, *Turtle*, *Resource Description Framework* (RDF), *Extensible Markup Language* (XML), *Notation3* (N3) (2011), *NTriples*, linguagens de consulta [SPARQL (*Protocol and RDF Query Language*)], ontologias e notações como RDF *Schema* e *Web Ontology Language* (OWL), que se fornecem uma descrição formal de entidades e correspondências dentro de um determinado domínio do conhecimento. As tecnologias são úteis para alcançar o principal objetivo de desempenho da *Web* Semântica, uma vez que os dados conectados (*Linked Data*) podem ser entendidos como “[...] um conjunto de boas práticas para publicação e conexão de dados estruturados na *Web*, usando padrões internacionais recomendados pelo W3C [...]”, que fornecem informações em larga escala para a integração e raciocínio sobre os dados (ISOTANI; BITTENCOURT, 2015, p. 14).

Os dados conectados são essencias para tecnologias como SPARQL, RDF, OWL e *Simple Knowledge Organization System* (SKOS). Entretanto, existem desafios para que os dados conectados possam atingir a funcionalidade completa da *Web* Semântica. As tecnologias estruturam os dados conectados e atuam como peças fundamentais na definição de ligações, no que diz respeito dentro de um conjunto de dados, e entre conjuntos de dados para outros dados conectados.

A publicação de dados na *Web* é uma realidade, ou seja, não se trata de uma ideia nova, pelo contrário, verifica-se um aumento em sua aplicação nos últimos anos, principalmente devido à importância que vem a adquirir.

Uma vez consolidada a *Web* de Dados, na qual os dados tramitam na rede em alta velocidade e em larga escala, os conteúdos precisam estar conectados, fazendo-se necessário o estudo aprofundado de todo o processo da informação, a fim de permitir e estimular a reutilização de dados.

Autores como Berners-Lee; Connolly; Swick (1999) e Abiteboul; Buneman; Suciu (2000) já mostraram em suas pesquisas a importância da distribuição e reuso de dados, que permitem o consumo e a produção de novas informações úteis e relevantes, de forma que se tenha uma recuperação baseada em regras, em que o dado que seja aberto ou controlado deve ser aplicado tanto para agentes de máquina quanto para usuários finais.

Nesse contexto, o programa de Pós-Graduação em Ciência da Informação (PPGCI) da Unesp de Marília estuda processos relativos à produção, organização, gestão, mediação, apropriação, recuperação e ao uso da informação, utilizando-se de aportes interdisciplinares oriundos de diversas áreas, cunhando-se como área de conhecimento, a fim de nortear o estudo crítico de teorias, metodologias e práticas voltadas à informação com especial ênfase na recuperação e ao uso da informação, em que as tecnologias de informação e comunicação ocupam importante papel para o desenvolvimento científico, tecnológico e social da sociedade.

A área de concentração 'Informação, Tecnologia e Conhecimento' está alicerçada na linha 'Informação e Tecnologia' que objetiva contribuir com o fortalecimento das pesquisas científicas, e um dos focos de interesse a que o presente estudo está relacionado - a disseminação de dados na Web -, uma vez que a Internet se faz tão presente, e autores como Borko (1968) e Saracevic (1996) a apresentam como ciência e tecnologia provedora de informação e acessibilidade em um contexto de uma sociedade variada, que gera elementos necessários para amparar e suprir seu desenvolvimento.

> CIÊNCIA DA INFORMAÇÃO é um campo dedicado à investigação científica e à prática profissional que aborda os problemas da comunicação eficaz de registros de conhecimento e conhecimento entre seres humanos no contexto de usos sociais, institucionais e / ou individuais e necessidades de informação. Ao abordar esses problemas de interesse particular, é tirar o máximo proveito possível da

> moderna tecnologia da informação (SARACEVIC, 1996, p. 43, tradução nossa, grifo da autora)[1].

Atualmente a Ciência da Informação (CI) pode se organizar em diversas subáreas que tratam da incorporação de tecnologias as quais facilitam a transmissão, organização e recuperação eficiente de dados para esse universo de usuários (humanos e máquinas).

Visto isso, e por se tratar de um trabalho relacionado à tecnologia no âmbito da CI, aportes teóricos relevantes às temáticas apoiam o presente estudo, em que todo o desenvolvimento e os instrumentos para a realização da pesquisa relacionam-se aos termos: *Web* Semântica, Ontologias, Dados Ligados ou Dados Conectados, Ciência dos Dados, *Web* de Dados, Mapeamento e Enriquecimento de Dados Semânticos.

Assim, visualizam-se algumas questões que se referem ao consumo final das informações, não no tocante às necessidades dos usuários humanos, mas às voltadas aos usuários agentes (máquina), como afirmam Wood et al. (2013), pois para que os dados publicados na *Web* tragam significados sobre as informações e possibilitem o acesso, consumo e interpretação de forma automática por esses tipos de usuários, de forma a gerar novas informações, fazem-se necessárias regras e padronizações na publicação de dados (BERNERS-LEE; CONNOLLY; SWICK, 1999).

Diante de tal realidade, questiona-se: Como tornar dados, que são por muitas vezes heterogêneos e desconexos, em informações relevantes e que possam ser empregadas na geração de novas informações e conhecimentos?

Para discutir sobre a temática que perpassa a questão anterior, insere-se o termo *Linked Data,* ou dados conectados, o qual foi criado por Berners-Lee em 2006, e segue os princípios da *Web* Semântica. Berners-Lee (2006a) refere-se a um estilo de publicar e interligar dados estruturados na Web e que tem como objetivo permitir que pessoas compartilhem dados de forma tão fácil quanto os documentos são compartilhados na *Web*.

---

[1] *"INFORMATION SCIENCE is a field devoted to scientific inquiry and professional practice addressing the problems of effective communication of knowledge and knowledge records among humans in the context of social, institutional and/or individual uses of and needs for information. In addressing these problems of particular interest is taking as much advantage as possible of the modern information technology"* (SARACEVIC, 1996, p. 43).

O *Linked Data* contempla essencialmente as diretrizes para a disponibilização de dados na *Web*, de maneira a interligar variados conjuntos de dados segundo os princípios da *Web* de Dados, em que as informações estejam contextualizadas e relacionadas a outros recursos a fim de possibilitar que agentes computacionais sejam capazes de compreender o domínio de um dado e aprimorar, desta forma, a recuperação da informação.

Com isso, vislumbra-se a transformação da *Web* de Documentos em uma *Web* de Dados, na qual o *Linked Data* tem crescido e se popularizado. Porém, para que a expansão se mantenha, regras e diretrizes foram apresentadas em 2017 por um grupo do *World Wide Web Consortium* (W3C) que publicou as “35 boas práticas para a publicação de dados na *Web*”, partindo dos princípios do *Linked Data* (LÓSCIO et al., 2018).

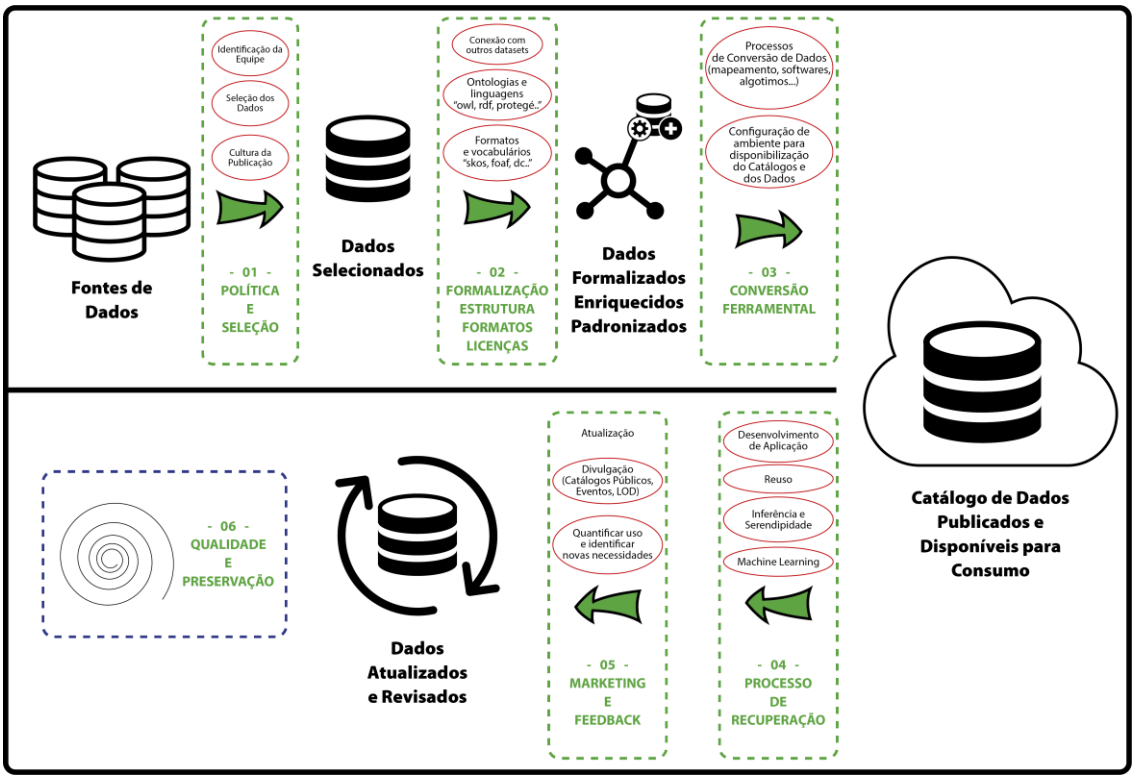
Todas as diretrizes têm grande impacto na publicação de dados na *Web*, uma vez que é de extrema importância aperfeiçoar o processo e melhorar a qualidade dos dados publicados.

Uma maneira de visualizar novos contextos relacionados à disponibilização de dados na *Web* é por meio da anotação semântica dos dados, para que estes possam ser mais bem utilizados e disponibilizados mediante uma enorme gama de aplicações, como por exemplo, por meio de sua publicação na *Web*.

Lóscio, Burle e Calegari (2018), quando apresentaram as “Boas Práticas de Dados na *Web*”, descritas nas recomendações da W3C, apontaram que uma das boas práticas é o processo de enriquecer dados por meio da geração de novos dados, que visa formar um perfil capaz de determinar os interesses de um usuário, relativos à área de conhecimento daquela informação a qual está sendo trabalhada.

Nesta linha, Santarem Segundo (2018) propôs um fluxo segmentado em 6 (seis) fases, que descrevem as atividades desenvolvidas no macroprocesso de publicação de dados na *Web* em formato aberto e semântico, com a finalidade de seguir as melhores práticas de dados conectados (*Linked Data*), o qual é apresentado na **Figura 1**.

Figura 1- Fluxo Organizacional para Publicação de Dados



Fonte: SANTAREM SEGUNDO (2018, p. 125)

Das 6 (seis) fases do fluxo proposto por Santarém Segundo (2018), atenta-se ao processo de heterogeneidade dos dados e de seu mapeamento e enriquecimento semântico, o qual é discutido e apresentado neste trabalho, e que se refere à expansão do dado como semântico, despertando um grande interesse tanto na CI quanto na Ciência da Computação (CC).

Tal processo se refere a como os dados podem ser formalizados, enriquecidos e padronizados para que possam alcançar uma interoperabilidade a ponto de permitir, por exemplo, que partes específicas de um texto tenham seu significado interpretado por máquina, a fim de proporcionar ao usuário uma recuperação mais eficaz.

Conforme exibido na **Figura 1,** as fases classificadas como 02 e 03 do fluxo organizacional são denominadas como “Formalização, Estrutura, Formatos e Licenças” e “Conversão Ferramental”, respectivamente. A literatura define esse último processo como “Mapeamento Semântico de Dados”, no qual se encontram os critérios de tecnologia e de estruturação dos dados, determinados por Berners-Lee nos princípios do *Linked Data*.

Trata-se de uma fase importante, pela qual o dado deve passar para garantir os aspectos discutidos, que vão desde o enriquecimento semântico até

sua conversão em formatos de uso na *Web* (uso de *softwares* e algoritmos), bem como a configuração de ambientes para que possam ser disponibilizados e, então, consumidos por um usuário, quer seja máquina ou humano, por meio de alguma aplicação baseada na estrutura de dados conectados.

Para que os dados publicados sejam consumidos, é necessário primeiramente entender que existe todo um processo, que abarca desde a coleta até a publicação dos dados de forma conectada na *Web*. Tais contextos são debatidos na literatura, indo desde o Ciclo de Vida dos Dados (SANT'ANNA, 2016) até a publicação dos dados na *Web* e suas diretrizes (ROZSA; DUTRA; NHACUONGUE, 2017), que são apresentadas e estudadas com frequência, de forma a gerar uma padronização coerente e assertiva.

A partir deste cenário e da relevância de se estudar todas as diretrizes conceituais sobre o processo de mapeamento semântico, também se faz necessário entender como é estruturado o processo, quais as tecnologias envolvidas, e ainda como os dados são enriquecidos semanticamente e seus formatos, criando diretrizes baseadas em boas práticas, que sejam padronizadas para os dados conectados. Dessa forma, justifica-se a proposição de um *Framework* para Enriquecimento e Mapeamento Semântico.

O processo de mapeamento está ligado com o processo de extração dos Dados de Fonte de Dados Heterogêneos que se classificam como estruturados, semiestruturados e não estruturados. A importação dos arquivos utiliza-se de ferramentas para limpar e reparar os dados para criar colunas de dados para os URIs, que representam os dados em vocabulários no LOD e exportar o conjunto de dados no formato RDF.

Neste contexto, é necessário criar uma base estruturada, utilizando-se de um vocabulário e da abordagem de enriquecimento de Metadados.

Pontuamos que para ocorrer o processo de Enriquecimento Semântico fazem-se necessários 3 (três) recursos:

- 1 - A Anotação Semântica de Dados, que é a atribuição de significado aos elementos de um domínio gerado por um esquema de origem, que consiste em ser manual ou automatizado por meio da adição de informação semântica (UREN *et al.*, 2005).
- 2 - A Vinculação e Mapeamento de Recursos, que segundo

Sorrentino *et al.* (2013), descobrir *links* (URIs) entre as combinações semânticas dos dados e metadados com outros recursos na *Web* de Dados, interligando recursos na nuvem LOD.

- 3 - E o Mapeamento para modelos de dados semânticos (RDF/XML) chamado de conversão ou processo de modelar os dados em um formato semântico e estruturado além da manipulação por aplicações que consomem esses modelos de dados.

Segundo Lira (2014), o Enriquecimento Semântico trabalha como um processo para atribuir maior significado aos metadados e dados por meio da aplicação de recursos auxiliares com o objetivo de facilitar a compreensão, a integração e o processamento dos dados por pessoas e máquinas. Processo que torna os dados e metadados mais qualificados e descritivos, mediante ao uso da semântica atribuída por recursos semanticamente bem-descritos, como recursos de bases de conhecimento, ontologias, vocabulários existentes, entre outros.

Para a presente pesquisa têm-se os seguintes pressupostos iniciais oriundos da problemática proposta:

Pressuposto 1: A literatura não disponibiliza a melhor forma, ou a mais eficiente e efetiva para os processos desde a extração e limpeza de dados para mostrar a melhor forma de se executar o mapeamento de dados estruturados, semiestruturados ou desestruturados, a fim de enriquecê-los no contexto informacional, até se transformarem em um conjunto de dados semânticos que possam ser utilizados em forma de aplicativos, ou através de buscas por máquinas de forma semântica.

Pressuposto 2: A conexão com as bases de dados (*datasets*) não apresentam as diretrizes informacionais de como isso deve ser feito e ainda quais as diretrizes para tal processo.

Pressuposto 3: A informação é o canal de acesso, compartilhamento, recuperação e uso, fato esse que mostra a diversidade de tipos de *softwares* e maneiras para resolver a estruturação dessa informação no âmbito da *Web* Semântica.

Pressuposto 4: O saber como fazer e, consequentemente, acessar, usar e reusar essa informação enriquecida de forma que os dados possam ser

publicados na *Web*, é o contexto de que tanto os agentes de máquina quantos usuários poderão se apropriar da informação. Isso gera conexões e novas descobertas que se aplicam ao contexto do *Linked Data* no âmbito da *Web* Semântica.

1.1 PROBLEMÁTICA DA PESQUISA

A Publicação de Dados conectados na *Web* se fundamenta em 4 (quatro) princípios básicos propostos por Berners-Lee (2006b). O primeiro se relaciona a usar *Uniform Resource Identifier* (URIs) para identificar os recursos da Web; o segundo é usar URIs- *HyperText Transfer Protocol* (HTTP) para que os usuários possam localizar e consultar estes recursos; o terceiro visa proporcionar informação útil em relação ao recurso quando a URI for requisitada, utilizando RDF para descrever os recursos e *SPARQL Protocol and RDF Query Language* (SPARQL) para permitir as consultas, e por fim o quarto, incluir ligações a outras URIs relacionadas com dados que estão contidos em um recurso de forma que enriqueça o descobrimento da informação na *Web*.

A abordagem de dados conectados oferece vantagens significativas sobre as práticas atuais na publicação de dados na *Web*, uma vez que, por meio do uso de URIs, os domínios que os utilizarão, por exemplo bibliotecas, repositórios, órgãos públicos entre outros, permitem que os recursos sejam mais facilmente acessíveis.

Nesse contexto, a motivação do Trabalho de Pesquisa é estudar, no âmbito da CI, o processo de mapeamento e enriquecimento semântico na publicação de dados abertos conectados na *Web*, baseando-se nas melhores Práticas determinadas por Lóscio et al. (2018), no que tangem interligar dados estruturados seguindo os princípios do *Linked Data*.

Verifica-se no processo um problema a ser estudado e analisado, o qual se refere à inexistência de diretrizes padronizadas e específicas que orientem todo o processo de mapeamento e enriquecimento semântico de dados abertos para publicação de forma ligada ao contexto informacional.

Assim, o presente trabalho visa responder ao seguinte questionamento: Como guiar o processo de mapeamento e enriquecimento semântico de dados

para publicação na *Web*, a partir de diretrizes e recomendações de boas práticas para o profissional da Informação?

O enriquecimento semântico, segundo Thakker *et al.* (2012), é o processo de acrescentar conceitos semânticos aos dados na *Web*, ou seja, de prover uma compreensão para alcançar um determinado contexto de forma interligada, com o objetivo de gerar resultados para os usuários de forma mais abrangente, com uma coleção diversa que possui essas mesmas características.

1.2 OBJETIVOS

O presente trabalho tem como **objetivo geral** a formalização de um ***Framework*** que oriente, por meio de diretrizes, o processo de Publlicação de dados, com ênfase ao enriquecimento e mapeamento semântico baseadas nas Boas Práticas de Publicações de Dados na Web.

A solução proposta não é apenas contextualizar um resultado com uso de um determinado vocabulário, ontologia e ferramenta adequada, mas sim abstrair e descrever o contexto do processo informacional desde a sua extração e limpeza, até que se contemple o processo do enriquecimento e mapeamento semântico para que se chegue à condição da publicação dos dados na *Web* de forma sistemática, em que uma das possibilidades de apresentação desses dados esteja em formato de um grafo *Resource Description Framework* (RDF) em consonância com os princípios *Linked Data*.

A partir do objetivo geral deste trabalho serão desenvolvidos os seguintes objetivos específicos:

1. Apresentar uma revisão da literatura sobre a transição da atual *Web* de documentos para a *Web* de Dados, caracterizando a natureza da ecologia de dados na *Web* Semântica;
2. Contextualizar o processo da publicação de dados abertos na *Web* Semântica.
3. Abordar as melhores práticas para transformação de formatos e o enriquecimento semântico de dados abertos.
4. Segmentar o processo de tratamento dos dados em relação a extração e limpeza, fator esse que permite a estruturação do dado, independente

do seu formato, até chegar na sequência do enriquecimento dos dados (a partir da aplicação de vocabulários e ontologias) até o mapeamento (relacionado às possibilidades de tecnologias existentes para isso) categorizando assim o dado em um modelo semântico.

5. Apresentar, em forma de grafo, um modelo que utilize essas diretrizes de tratamento de dados, principalmente no que se diz respeito ao processo de enriquecimento e mapeamento semântico, no contexto da Publicação de Dados na *Web* em contexto informacional.

1.3 JUSTIFICATIVA

A mudança contínua da *Web* atual permite que seja experimentado um momento de transição no qual processos documentais sejam visualizados, recuperados e usados como dados. E, com o grande crescimento da publicação de dados conectados na *Web*, se intensifica mais a importância da participação do profissional da Informação no processo (MARTINS; LUZ; SANTAREM SEGUNDO, 2019).

Carey, Onose e Petropoulos (2012) afirmam que o número de servidores *Web*, mais conhecidos como serviço de dados (*Web Data Services*), têm aumentado consideravelmente, uma vez que espalhados em redes públicas ou privadas criam uma procura e busca de fornecimento de dados muito grande, por meio de interfaces *Web* que permitem acesso e manipulação à fonte de dados, e estão em formatos distintos, classificados na literatura como heterogêneos, o que dificulta sua acessibilidade e reuso.

Berners-Lee (2006b) enfatiza que a *Web* Semântica e suas tecnologias surgem como um ponto agregador para permitir a disponibilidade dos dados tanto para humanos como para máquinas, numa compreensão que se estende a esses serviços de dados, ao empregar tecnologias para transformar dados heterogêneos em um formato mais refinado por meio de estudos de mapeamento e enriquecimento semântico, de forma que o resultado das buscas por um computador facilite o acesso e o reuso das aplicações *Web*, que são tão complexas em um âmbito informacional.

Quando abordada a temática de Mapeamento e Enriquecimento Semântico de dados, pesquisadores como Oliveira *et al.* (2016) e Salvadori *et*

al. (2017) apontam vantagens como a interoperabilidade dos dados e sua inferência nos processos de busca, e na construção de serviços como, por exemplo, a publicação de dados na *Web*.

Justifica-se a importância da presente pesquisa de doutorado uma vez que o levantamento de informação sobre o tema apresenta *Frameworks,* arquiteturas e metodologias que se relacionam a finalidades e domínios específicos, tornando-se uma tarefa onerosa para os profissionais da computação e da informação que atuam no processo do enriquecimento semântico de dados conectados, porém com tarefas distintas e sem haver necessariamente uma conexão entre eles.

Neste sentido, a pesquisa avança ao prover diretrizes para o mapeamento e enriquecimento semântico de dados baseadas nas boas práticas de publicação de dados na *Web*, para que tal processo seja visto de forma mais informacional e com conteúdo enriquecido para o desenvolvimento na área computacional.

O tema apresentado ajuda a visão informacional, visto que os estudos sobre Enriquecimento e Mapeamaneto semântico de dados não indicam um guia para a publicação de Dados na *Web*, deixando a responsabilidade do processo somente para o profissional de tecnologia envolvido. Nessa perspectiva, o trabalho poderá contribuir para a construção de conceitos, definições e aplicação geral, para qualquer domínio de dados a ser mapeado e enriquecido, destinado ao público não computacional, com foco nos processos informacionais, que agreguem avanço ao estado da arte da área.

A pesquisa visa contribuir para a Linha de Pesquisa "Informação e Tecnologia” do Programa de Pós-Graduação em Ciência da Informação tanto em relação ao avanço teórico sobre o mapeamento e enriquecimento semântico no processo de publicação de dados na *Web*, como também no contexto da construção e desmistificação da problemática de se saber as diretrizes corretas para que o dado seja semântico, enriquecido e ligado.

A comunidade científica brasileira vem criando alternativas de melhor posicionamento na investigação. No âmbito da pesquisa proposta, mapeamento e enriquecimento semântico direcionam pesquisas relevantes no âmbito da *Web* Semântica e geram investigações e possíveis soluções relacionadas à *Web* de Dados. Por isso, a necessidade de buscas afinadas e

de publicações relacionadas ao contexto de enriquecimento, uma vez que se verificou um escasso número de publicações relacionadas ao tema.

O impacto social da tese proposta traz uma vertente na possibilidade de proporcionar maior interoperabilidade entre sistemas e aplicações da *Web*, que visam uma maior reutilização de dados abertos, principalmente os governamentais, pela população.

Karger (2014, p. 64) afirma não ter sido realizado, nesse sentido, o uso do potencial da Web Semântica aos usuários finais.

> O potencial da Web Semântica para entregar ferramentas que ajudem os usuários finais a capturar, comunicar e gerenciar informações ainda deve ser realizado e muito pouco tem sido pesquisado nesse sentido (tradução nossa)[2].

Esse *Framework*, quando utilizado, pode contribuir socialmente a partir do momento que os dados podem ser resgatados ou trabalhados dentro de uma comunidade. O aumento do controle social sobre dados de determinado domínio incentiva a produção de aplicativos e agentes inteligentes na *Web*, e amplia as oportunidades para profissionais da informação e de demais campos, entre outros aspectos.

1.4 ESTRUTURA DA TESE

O trabalho contempla, além dessa '**Introdução**', a seguinte estrutura:

- No Segundo Capítulo são apresentados os **Trabalhos Relacionados** ao tema da pesquisa, que contribuem com o debate sobre enriquecimento e mapeamento semântico de dados publicados nos últimos anos.

- No Terceiro Capítulo, **Metodologia,** são abordados os aspectos metodológicos que indicam seus pressupostos e possíveis classificações, além de um indicativo de percurso pretendido para

[2] *The Semantic Web's potential to deliver tools that help end users capture, communicate, and manage information has yet to be fulfilled, and far too little research is going into doing so.* KARGER (2014, p. 64).

alcançar os objetivos propostos, trazendo também uma contextualização de '***Framework*: Definições**', que discute as abordagens, conceitos e definições que desdobram na apresentação dos modelos de *Frameworks* e, dessa forma, busca-se evidenciar a sua relevância para escrita de diretrizes para o Mapeamento dos dados, uma vez que quer transformá-los em semânticos para um uso posterior nas publicações dos Dados na Web.

- No Quarto Capítulo, '**Revisão da Literatura**', discorre-se sobre as abordagens, conceitos e definições que têm como foco a Web Semântica e a Web de Dados, apresentando o contexto de dados que tramitam em tal contexto, além de evidenciar a relevância do *Linked Data* à materialização da Web Semântica. Descreve-se o '**Macroprocesso da Publicação de Dados na Web**', um cenário conceitual, apresentado pela W3C (2020), demonstrando cada uma das fases para contextualizar o estudo de mapeamento envolvido. A seção aborda contextos como o ciclo de vida dos dados na Web, que envolve desde a questão da desestruturação dos dados e a sua busca, de forma simples, à aplicação das Boas práticas (Lóscio et al., 2018) para Publicação de Dados dos dados na Web. Ainda serão apresentadas algumas técnicas e desafios para se chegar à publicação.

- No Quinto Capítuo, apresenta-se o processo de '**Enriquecimento e Mapeamento semântico dos Dados para Web**'. As teorias relacionadas a estes aspectos são apresentadas, desde a Limpeza e Modelagem dos dados, incluindo seus aspectos de construção e adaptação em Vocabulários e Ontologias, e Ferramentas relacionadas ao Mapeamento. Também é apresentado o Enriquecimento, dando ênfase nos problemas encontrados e o processo do uso dos metadados enriquecidos.

- O Sexto Capítulo apresenta os detalhes do '***Framework***' que apresentam passos e diretrizes do que foi proposto nos objetivos específicos e um estudo de caso dos dados do metrô de São Paulo,

aplicando as diretrizes e a contextualização das Boas Práticas de Publicação de dados na Web, apresentadas pelo *Framework*.

- As Considerações Finais apresentam os resultados obtidos e a discussão acerca desses resultados. As conclusões obtidas e a citação de trabalhos que podem contribuir para o desenvolvimento de projetos futuros relacionado ao '***Framework***' e as diretrizes propostas.

- Ao final, são elencadas as **'Referências'** que foram utilizadas para a construção do texto de defesa da Tese proposta.

A seguir, a **Figura 2** apresenta o mapa conceitual da estrutura da Tese.

Figura 2- Mapa Conceitual do Conteúdo da Tese

Fonte: Elaborada pela autora (2021)

2 TRABALHOS RELACIONADOS

Mapeamento e Enriquecimento de dados conectados é o nome dado ao processo que gera recursos semânticos, implícitos ou ocultos, incorporados a um conjunto de dados, e explícito a ferramentas que utilizam o padrão RDF.

Diversos estudos anteriores no campo, focados principalmente na geração de ligações semânticas relacionadas a esses recursos, já foram realizados, contribuindo com o debate proposto na pesquisa.

São apresentadas as técnicas e formas de mapeamento e enriquecimento semântico, de modo a trazer pontos importantes a serem discutidos no presente trabalho.

Para tal, foram realizadas buscas de trabalhos publicados em periódicos indexados nas bases de dados internacionais: *Core Collection* da *Web of Science* (WoS), *Scopus* e *Library and Information Science Abstract* (LISA), utilizando de termos como: "*semantic mapping*" *or "linked data enrichment" or "linked data Framework" or "linked data approach" or "semantic web approach" or "semantic web Framework" or "web of data".*

Os respectivos termos em português não retornaram quase nada, então foram desconsiderados nesta pesquisa. Termos gerais como *Web Semantic*, *Linked Data* e *Linked Open Data* também não foram usados para a busca, uma vez que são termos que abrem a amplitude da busca e retornaram muitos resultados que não têm relevância com o tema proposto de mapeamento e enriquecimento de dados semânticos.

A busca na base WoS retornou 597 referências. Já na Scopus foram encontradas 1306 referências. Por fim, na LISA foram encontradas 132 referências, porém todas constavam nos resultados retornados pela WoS e pela Scopus. As buscas nestas bases de dados se deram nos campos de título, resumo e palavras-chaves, limitando-se ao período de 2014 a 2019, e o período se justifica devido a grande quantidade de artigos publicados referentes ao assunto "mapeamento semântico e enriquecimento de dados".

A partir das discussões fomentadas elaborou-se uma tabela que contém os elementos norteadores ao desenvolvimento da pesquisa, ferramentas utilizadas e processos necessários para que possa dar uma linha de raciocínio para o desenvolvimento das diretrizes do *Framework* que será proposto. Os

pontos estão relacionados às seguintes informações: Trabalhos / Objetivos / Uso de ontologias ou vocabulários / Nível de automação / Repositório / e como eles se enquadram: metodologia, *Framework*.

A partir do levantamento do cenário decorrente nos últimos treze anos (2006 até 2019), o próximo passo foi analisar o processo evolutivo da construção de ferramentas para o mapeamento e que, por consequência, geram enriquecimento dos dados, tornando-os semânticos, e as medidas tomadas nesses últimos 13 anos para validá-las e verificá-las corretamente, e como isso vem acontecendo para que, de alguma forma, essas informações sejam visualizadas e entendidas nesta pesquisa.

A análise avalia e determina algumas diretrizes tomadas para permitir o enriquecimento e o mapeamento dos dados tornando-os semânticos para que, assim, possam ser usados e determinados como um dos passos para entender e vislumbrar as possíveis técnicas para a construção de um *Framework* conceitual que gere essas diretrizes baseadas nas Boas Práticas para publicação de Dados na *Web*.

Mapeamento e Enriquecimento de dados conectados são processos que geram recursos semânticos implícitos ou ocultos incorporados a um conjunto de dados, e explícitos a ferramentas que utilizam o padrão RDF. Diversos estudos anteriores neste campo, focados principalmente na geração de ligações semânticas relacionadas a esses recursos, foram feitos.

Também foi gerada uma tabela, que contém os pontos principais para o desenvolvimento desse estudo, ferramentas utilizadas e processos necessários a fim de que possam dar uma linha de raciocínio em relação ao desenvolvimento das diretrizes do *Framework* que será proposto.

2.1 TRABALHOS

Os trabalhos discorridos proporcionaram a elaboração de uma ilustração do cenário de enriquecimento e mapeamento semântico. Para isso, segue-se a discussão fomentada em cada trabalho.

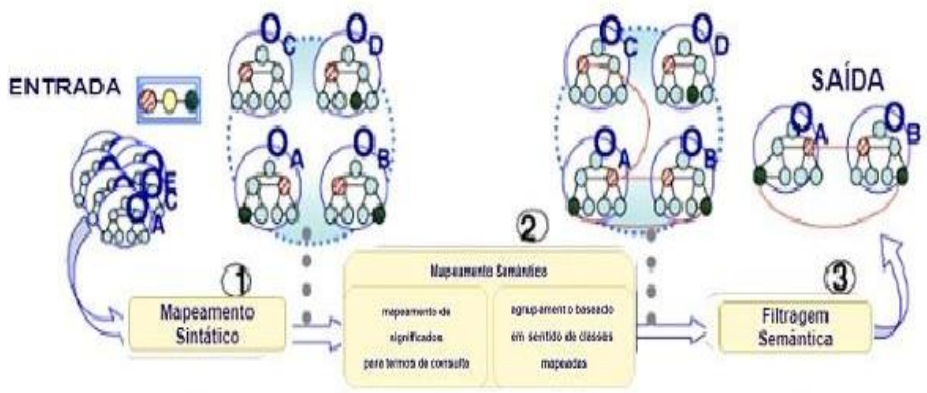
2.1.1 *PowerMap: Mapping the Real Semantic Web on the Fly*

Lopez, Motta e Sabou (2006) propuseram um projeto denominado como PowerMap: Mapeando Real da Web Semântica em Tempo Real, que traz diretrizes nas quais o mapeamento, ao invés de ser realizado no tempo de *design* da ferramenta, verifica se novas ferramentas que requerem técnicas de mapeamento podem ser executadas no tempo de execução.

O mapeamento de ontologias desempenha um papel importante na ponte entre as fontes de dados distribuídas e heterogêneas. À medida que a *Web* Semântica se torna lentamente real e a quantidade de dados semânticos *online* aumenta, uma nova geração de ferramentas é desenvolvida para encontrar e integrar automaticamente esses dados.

A investigação sobre os requisitos gerais para técnicas de mapeamento em tempo de execução foi um dos levantamentos como resultado, além de descrever um algoritmo de mapeamento denominado como *PowerMap*, desenvolvido para ser usado em tempo de execução por uma ferramenta de resposta a perguntas baseadas em ontologia. O *PowerMap* é um algoritmo de correspondência híbrida que compreende técnicas de correspondência de esquemas estruturais e terminológicos com a assistência de recursos ontológicos ou lexicais em larga escala. A **Figura 3** mostra as três principais fases do *PowerMap.*

Figura 3 - Exemplo de processo de Mapeamento para obter o potencial de mapeamento de triplas da ontologia



Fonte: Adaptado pela autora a partir de LOPES; SABOU; MOTTA (2006, p. 419)

O Mapeamento Sintático identifica os mapeamentos candidatos para todos os termos de consulta em diferentes ontologias *online* (identifica ontologias potencialmente relevantes para essa consulta específica). Essa é a fase mais simples, pois considera apenas rótulos conceituais, ou seja, ignora a estrutura das ontologias. Ele se baseia em métodos simples de comparação baseados em cadeias de caracteres (por exemplo, editar métricas de distância) e no *WordNet* para pesquisar palavras relacionadas lexicamente (sinônimos, *hypernyms* e *hyponyms)*.

Já o Mapeamento Semântico opera no conjunto reduzido de ontologias identificadas na fase anterior, que verifica os mapeamentos sintáticos identificados anteriormente e exclui aqueles que não fazem sentido de uma perspectiva semântica. Ao contrário da fase anterior, esta fase se baseia em métodos mais complexos.

Primeiro, explora a estrutura hierárquica das ontologias candidatas para extrair o sentido dos conceitos de candidatos. Segundo Lopez, Motta e Sabou (2006), o autor supracitado usa métodos baseados no *WordNet* para calcular a semelhança semântica entre os termos da consulta e as classes de ontologia.

Os mapeamentos filtrados pela fase anterior estão espalhados por várias ontologias. O objetivo da filtragem semântica é filtrar os mapeamentos significativos que melhor representam o domínio da consulta, determinando aquelas ontologias que cobrem triplas inteiras e não apenas os termos individuais das triplas e estuda a relação da ontologia para determinar a interpretação semântica válida.

No caso deste trabalho, *PowerMap*, verifica-se o uso de uma metodologia para mostrar o processo de enriquecimento e mapeamento semântico a partir de um algoritmo, uma vez que é gerado a partir de uma lógica que se ocupa dos métodos computacionais, que gera uma analogia híbrida com técnicas de correspondência e recursos estruturais, em que os recursos ontológicos ou lexicais em larga escala possuem terminologias que geram como uma assistência para outros projetos de recuperação de Informação.

2.1.2 *Semantically Enhanced Information Retrieval: An Ontology-Based Approach*

Recuperação de informações aprimorada semântica: uma abordagem baseada em ontologia, desenvolvida por Fernández et al. (2011). Os autores afirmam que, atualmente, as técnicas para descrição de conteúdo e processamento de consulta em Recuperação de informação (IR - *Information Retrieval*) são baseadas em palavras-chave e, portanto, fornecem recursos limitados para capturar as conceituações associadas às necessidades e conteúdo do usuário (FERNÁNDEZ *et al.*, 2011).

Com o objetivo de resolver as limitações dos modelos baseados em palavras-chave, a ideia de busca conceitual, entendida como busca por significados e não por cadeias literais, tem sido o foco de um amplo corpo de pesquisa na área de IR (*Information Retrieval* – Recuperação da Informação). Mais recentemente, ele tem sido usado como um cenário de protótipo (considerado como um ''aplicativo potencial'') na visão da *Web* Semântica, desde o seu surgimento no final dos anos noventa.

No entanto, as abordagens atuais de busca semântica desenvolvidas na área de *Web* Semântica ainda não aproveitaram ao máximo o conhecimento adquirido, a experiência acumulada e a sofisticação tecnológica alcançada por várias décadas de trabalho na área de IR. Partindo desta posição, Fernández *et al.* (2011) investigam a definição de um modelo de IR fundamentado em ontologia, orientado à exploração de Bases de Conhecimento de domínio para suportar recursos de pesquisa semântica em grandes repositórios de documentos, de maneira a enfatizar, por um lado, o uso de ontologias de pleno direito na semântica, apoiado nessa perspectiva e, por outro lado, a consideração de conteúdo não estruturado como o espaço de pesquisa de destino.

A principal contribuição do trabalho é um modelo de pesquisa semântica inovador e abrangente, que estende o modelo de RI clássico, aborda os desafios do ambiente *Web* massivo e heterogêneo e integra os benefícios da pesquisa por palavras-chave e semântica.

Contribuições adicionais incluem: uma inovadora técnica de fusão de *rank* que minimiza os efeitos indesejados da escassez de conhecimento sobre

a *Web* Semântica ainda juvenil e a criação de um *benchmark* de avaliação em larga escala, com base nos padrões de avaliação de IR, a TREC (*Text Retrieval Conference*) que permite uma comparação rigorosa entre as abordagens de IR e da *Web* Semântica.

As experiências conduzidas no trabalho mostram que o modelo de busca semântica obteve resultados comparáveis e de melhor desempenho do que o melhor sistema automático, TREC.

Visualiza-se, neste projeto, um processo de diretrizes as quais guiam e definem uma abordagem baseada em ontologia e regulam um caminho a seguir referente ao contexto da recuperação de informações aprimoradas no contexto da semântica, diferente de um *Framework* que pode atingir uma funcionalidade específica, por configuração, no processo de uma aplicação.

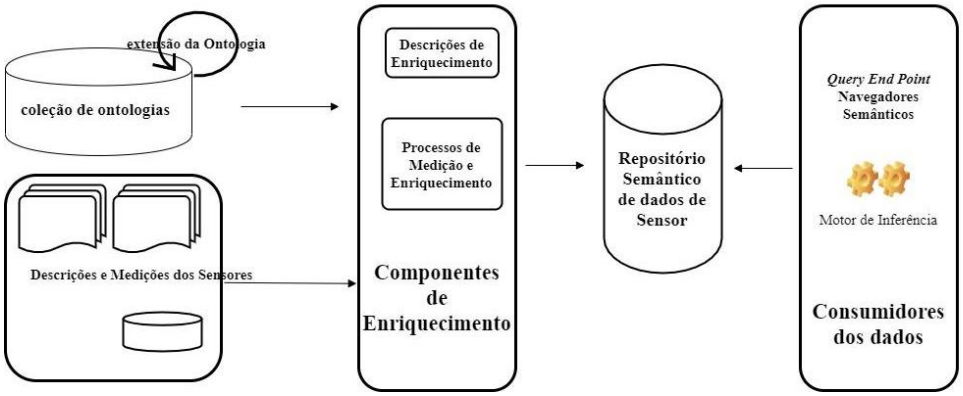
2.1.3 *A Framework for Semantic Enrichment of Sensor Data*

O *Framework* proposto por Mararu e Mladenic (2012) tem como proposta a criação de uma estrutura que traga solução para o enriquecimento semântico de dados em sensores. Baseado no interesse crescente de saber o ambiente em que se vive e de tornar cidades com *status* de cidades inteligentes, a implantação de milhares de sensores é um fato, pois podem medir e relatar o *status* das pessoas.

Neste contexto, o *Framework* proposto por Mararu e Mladenic (2012) traz o enriquecimento semântico de descrições e medições de sensores, a fim de fornecer dados mais ricos, isto é, dados que contêm mais significado, aumentam o impacto e como as redes de sensores e podendo melhorar a usabilidade, a acessibilidade dos dados.

No trabalho, foi proposto um *Framework* que automatiza o processo de enriquecer semanticamente as descrições e medições dos sensores, com o objetivo de melhorar a usabilidade e acessibilidade dos dados, que geram um repositório semântico de sensores de dados, permitindo assim uma inferência para os consumidores, como demonstrados na **Figura 4**, o processo do *Framework* proposto.

Figura 4- *Framework* conceitual. Ilustração dos principais componentes que constituem a estrutura proposta



Fonte: Adaptado pela autora de MARARU; MLADENIC (2012, p.171)

Verifica-se que os *Frameworks* têm como finalidade sempre atingir uma funcionalidade específica, como é o caso proposto por Mararu e Mladenic (2012) que traz o enriquecimento semântico de descrições e medições de sensores.

2.1.4 *OpLiDaF Open Linked Data Framework: a platform for the creation and publication of Linked Data*

O *Framework* OpLiDaF relacionado a dados abertos ligados apresenta uma plataforma para criação e publicação de dados conectados, descritos por Possemato (2013).

É uma estrutura para a criação, estruturação e visualização de dados no formato RDF/XML (*eXtensible Markup Language*). Configura-se como uma plataforma especializada para o tratamento (que integra passos de mapeamento, conversão, limpeza e publicação) dos dados conectados para dados formatados de forma heterogênea, por meio de ferramentas e procedimentos *ad hoc* ou sistemas integrados de código aberto, e pelo uso de padrões e idiomas reconhecidos pela Web Semântica.

As principais funções da plataforma OpLiDaF estão relacionadas à seleção de ontologias; mapeamento entre os dados de origem e ontologia ou selecionados de uma determinada ontologia; criação de ontologias específicas a partir de um conjunto de dados; a produção de arquivos RDF/XML e a limpeza de dados.

O sistema OpLiDaF baseia-se na observação da composição e tipologia, apesar das diferenças de conteúdo e formato, dos dados que formam o corpo de informações de bibliotecas, arquivos, museus, organizações turísticas e regionais e outras instituições (POSSEMATO, 2013).

Além do processo, o trabalho apresenta, a partir do ciclo de vida de dados conectados, que podem ser subdivididos em várias etapas e foram divididas em sete etapas, classificadas por meio de “Diretrizes Metodológicas para Publicação de dados conectados governamentais, a plataforma OpLiDaF é capaz de cobrir a seguintes soluções:

1. Identificação da fonte de dados;

2. Modelagem de vocabulário;

3. Geração de dados no formato RDF, através das diferentes linguagens de mapeamento;

4. Publicação dos dados em RDF;

5. Limpeza dos dados produzidos;

6. Criação de links entre diferentes conjuntos de dados; e

7. Disponibilizar dados, com diferentes etapas, incluindo a publicação do conjunto de dados obtido pelo processo no Registro CKAN[3] (rede abrangente de arquivos do conhecimento).

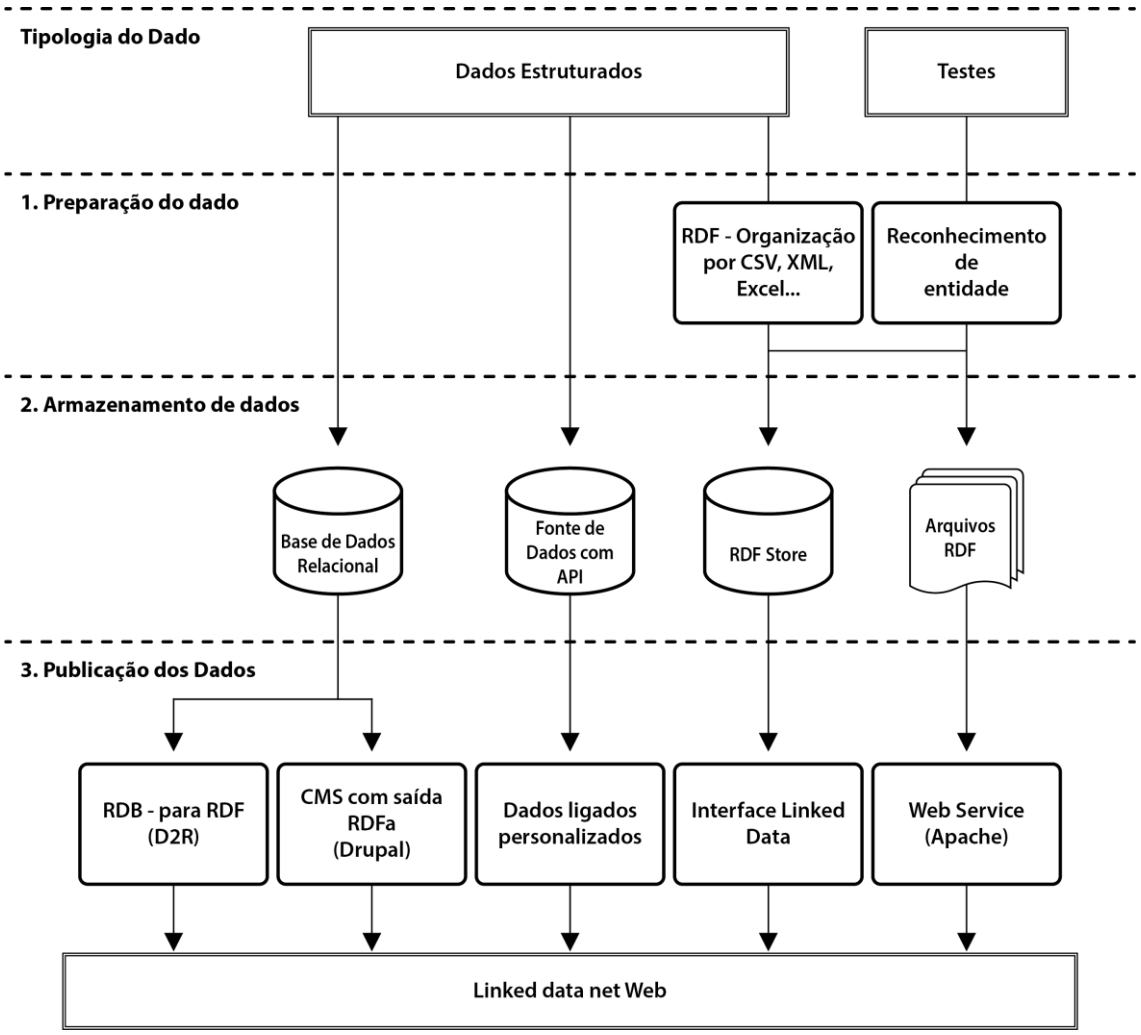
Verifica-se mais um *Framework* e, para atingir uma aplicação específica, foi proposta uma plataforma para criação e publicação de dados conectados que possui a funcionalidade específica de criação, estruturação e visualização de dados no formato RDF/XML, como descrito na **Error! Not a valid bookmark self-reference.**.

Possemato (2013) afirma que a plataforma é capaz de satisfazer completamente as etapas apresentadas e constitui uma ferramenta útil para quem deseja produzir dados (independentemente do sistema de gerenciamento, formato dos dados, tamanho do conjunto de dados e do modo

[3] *Comprehensive Knowledge Archive Network*, segundo o site oficial é uma aplicação Web destinado a editores de dados (governos, empresas e organizações nacionais e regionais) que desejam tornar seus dados abertos e disponíveis (CKAN, 2020).

e frequência das atualizações), resolvendo parte dos obstáculos e problemas que a passagem da *Web* Tradicional para a *Web* Semântica pode representar.

Figura 5 - Fluxo de trabalho da publicação de dados heterogêneos em dados conectados (OpLiDaF)



Fonte: Adaptado pela autora a partir de POSSEMATO (2013, p. 275)

2.1.5 *Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling*

A escolha de qual vocabulário reutilizar ao modelar e publicar no LOD (*Linked Open Data*) está longe de ser trivial. Não há estudo que investiga as diferentes estratégias de reutilização de vocabulários para modelagem e publicação de LOD.

Neste contexto, os autores Schaible, Gottron e Scherp (2014) publicaram um artigo apresentando os resultados de uma pesquisa com 79

participantes que examinaram as estratégias de reutilização de vocabulário mais utilizado na modelagem de LOD.

Os participantes, editores e profissionais de LOD, foram convidados a avaliar diferentes estratégias de reutilização de vocabulário e explicar sua decisão de classificação. Foram encontradas diferenças significativas entre as estratégias de modelagem, que variam de reuso, com o objetivo de minimizar o número e permitir que permaneça no vocabulário de um domínio.

Um *insight* muito interessante é que a popularidade no significado da frequência com que um vocabulário é usado em uma fonte de dados é mais importante do que a frequência de classes e propriedades individuais como são usadas na nuvem LOD. No geral, os resultados desta pesquisa ajudaram a entender melhor as estratégias e diretrizes de como os engenheiros de dados utilizam vocabulários e podem ser usados para desenvolver futuras ferramentas de engenharia de vocabulário.

2.1.6 *Towards an Intelligent Framework for Multimodal Affective Data Analysis*

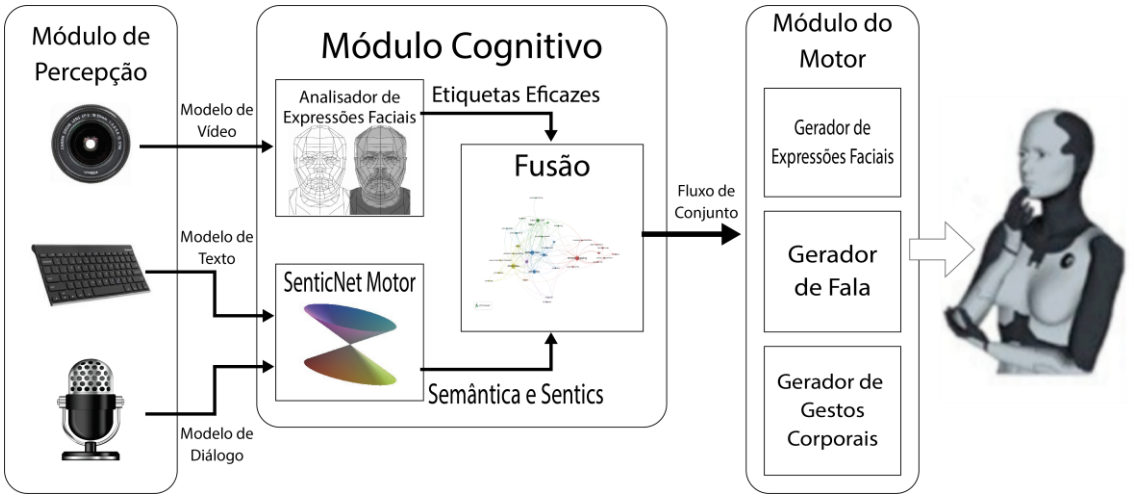
Segundo Poria *et al.* (2015), uma quantidade cada vez maior de conteúdo desestruturado é postado em *sites* de mídia social como *YouTube* e *Facebook* todos os dias. Para lidar com o crescimento dos dados, os autores viram a necessidade de desenvolver uma estrutura de análise multimodal inteligente que viesse efetivamente extrair informações de várias modalidades.

O trabalho, que tem como tradução “Uma estrutura inteligente para análise de dados afetivos multimodais” dos autores supracitados, teve como objetivo criar um agente de extração de informações multimodais, que infere e agrega as informações semânticas e afetivas associadas aos dados multimodais gerados pelos usuários em contextos como *e-learning*, saúde, eletrônica, marcação automática de conteúdo de vídeo e Interação Homem-Computador (IHC).

No caso, o agente inteligente desenvolvido adota uma abordagem de extração de conjunto de recursos ou diretrizes, de maneira a explorar o uso conjunto de recursos tri-modais (texto, áudio e vídeo) para aprimorar o processo de extração de informações multimodais.

Em experimentos preliminares usando o conjunto de dados e interface, o sistema multimodal proposto atingiu com uma precisão de 87,95%, que supera o melhor sistema de ponta em mais de 10%, ou em termos relativos, uma redução de 56% na taxa de erro. A **Figura 6** representa o esquema de funcionamento do Sistema de análise de Emoção em tempo real:

Figura 6 - Sistema de análise de emoções em tempo real



Fonte: Adaptado pela autora de PORIA *et al.* (2015, p.19)

A **Figura 6**, adaptada de Poria *et al.* (2015), classifica o módulo de percepção em três modalidades: informações visuais, trilha sonora (fala) e legendas (texto).

A obtenção da extração das informações afetivas foi feita a partir de dados multimodais, que apresentam um módulo cognitivo que permite uma fusão de resultados em diferentes modalidades, a fim de envolver todas as modalidades no processo de análise emocional.

O algoritmo desenvolvido por Poria *et al.* (2015) contou com características de processamento, permitindo que os dados para cada modalidade fossem processados como extração de recursos, e esses recursos construíssem modelos de treinamento.

A partir da extração dos dados para cada modalidade, os dados visuais e o processo de extração de recursos incluíram a etapa de classificação que gera o próprio treinamento.

Na modalidade de fusão, os dados estão relacionados com as saídas dos classificadores para todas as modalidades e como elas foram fundidas, usando a técnica de fusão baseada em recursos.

O módulo motor está relacionado com o treinamento usado pelos recursos dispostos pelas modalidades apresentadas, que mostram, assim, um modelo multimodal.

Segundo os Portia *et al.* (2015), o método trouxe uma simplicidade vantajosa na forma de adquirir os dados, mas é necessária uma precisão significativamente alta para o seu processamento, uma vez que foi usado um conjunto de dados para detectar emoções de conteúdos multimodais.

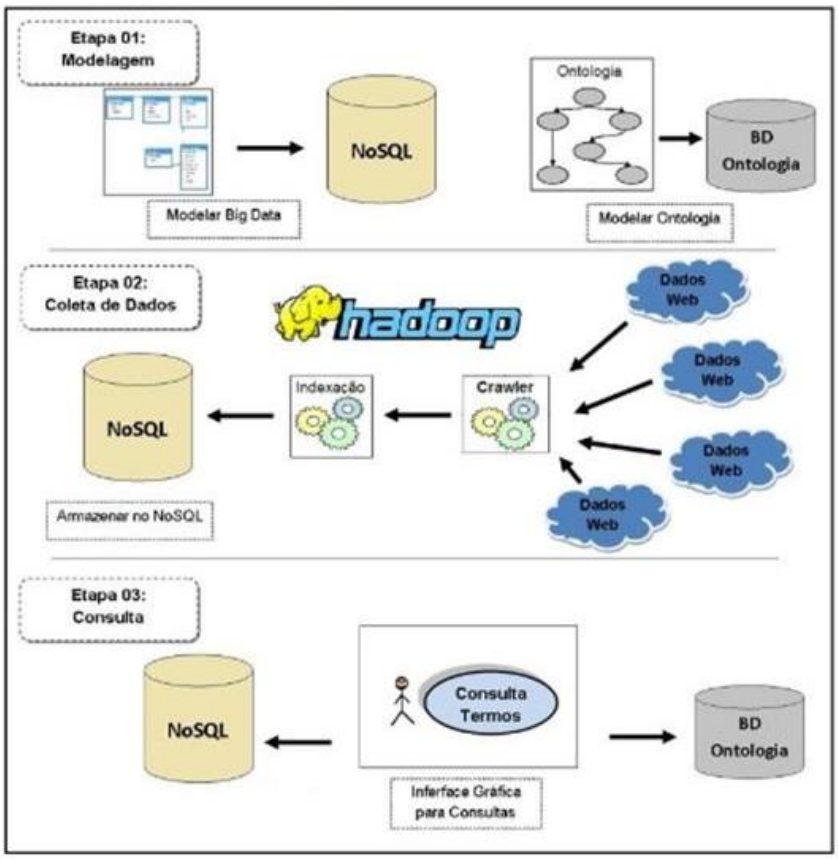
Seu resultado a partir da extração dos principais recursos dos dados de áudio e vídeo permitiu a análise da relação entre os recursos de áudio e visuais, e permitiu que o algoritmo do HMM fosse usado como um classificador para entender a emoção e para medir a dependência estatística nos sucessivos segmentos de tempo.

2.1.7 Ferramenta para Extração de Dados Semiestruturados para Carga de um *Big Data*

Furtado *et al.* (2015) apresentam uma ferramenta para extração de dados semiestruturados para carga de um *Big Data*. O principal objetivo do trabalho foi desenvolver uma ferramenta que realizasse a extração de dados de diversas fontes da *Web*, para armazená-los em um *Big Data* (FURTADO *et al.*, 2015).

Por meio da ferramenta proposta, os autores afirmam a possibilidade de analisar informações que antes não estavam disponíveis, de maneira a auxiliar as empresas no processo de tomada de decisões com maior velocidade. A **Figura 7** mostra a estrutura adotada por Furtado *et al.* (2015) para o desenvolvimento da ferramenta.

Figura 7 - Visão Geral do ambiente de Trabalho



Fonte: FURTADO *et al.* (2015, p. 48)

Furtado *et al.* (2015) apresentam as etapas da seguinte forma: a primeira etapa (modelagem) consiste em desenvolver um ambiente para identificação e modelagem de documentos semiestruturados e não estruturados provenientes da Web.

Os autores afirmam que, para armazenar as informações extraídas da Web, foi modelado um banco de dados NoSQL (*Not Only Structure Query Language*) *Apache Cassandra*, e utilizaram a modelagem de uma ontologia para representar o conhecimento de acordo com necessidades apontadas por uma situação real.

É uma etapa que deve ser realizada primeiro, antecedendo a etapa de coleta de dados, pois compreende a prévia identificação e expressão de uma área do conhecimento, em que será indexado o conteúdo buscado por um Rastreador *Web* (*webcrawler*).

A construção da ontologia utilizada no projeto foi realizada manualmente e armazenada em um banco de dados relacional. As *Uniform Resource Locator* (URLs) dos *sites* iniciais, as quais foram determinadas pelo rastreador

(*Crawler*) no início da busca, foram armazenadas em um sistema gerenciador de banco de dados MySQL e definidas antes de realizar a execução do motor de busca.

Na segunda etapa, ocorre a coleta de dados através de um motor de busca, que realiza a indexação e o armazenamento em um banco de dados. Para a coleta de dados da *Web*, o robô de captura (*webcrawler*) recupera as informações alimentando a base de dados do *Apache Cassandra*, indexando o conteúdo coletado pela *webcrawler*.

Todo o processo de indexação é executado e controlado pelo *Apache Lucene* e, ainda durante esta etapa de coleta, entra em funcionamento do *Framework Hado*op, *Framework* usado para computação distribuída e que utiliza o modelo de programação *MapReduce*, que foi usado nesse projeto.

Furtado *et al.* (2015) finalizam a ideia do projeto com a apresentação da terceira etapa, classificando-a como uma consulta. Foi proposto e apresentado seu desenvolvimento através de uma interface gráfica amigável para visualização dos arquivos extraídos da *Web*, e armazenados no banco de dados *Cassandra* conforme a ontologia especificada na primeira etapa, e possibilitou uma visualização da busca em forma de *ranking*, semelhante a um *site* de busca.

Visualiza-se, então, que o trabalho proposto só gerencia a necessidade aplicada naquele momento, não determinando as possibilidades de dados diversos no meio do processo e ainda não usa ferramentas tecnológicas para criar o processo de ontologia, demorando assim a possibilidade de uso.No entanto gera uma identificação necessária em um processo antes do mapeamento, que é a descoberta do domínio e ainda a possibilidade de buscar esses dados, mesmo que se caracterize como site de busca e não um repositório semântico.

2.1.8 *Semantic Representation and Enrichment of Information Retrieval Experimental Data*

A avaliação experimental realizada em campanhas internacionais de grande escala é um pilar fundamental do avanço científico e tecnológico dos sistemas de recuperação de informações (RI). Essas atividades de avaliação produzem uma grande quantidade de dados científicos e experimentais, que

são a base para toda a produção científica subsequente e o desenvolvimento de novos sistemas.

Silvello *et al.* (2017) apresentam o trabalho que anota e interliga semanticamente os dados, com o objetivo de aprimorar sua interpretação, compartilhamento e reutilização.

Um fluxo de trabalho de avaliação subjacente foi proposto como um modelo de estrutura de descrição de recursos para as artes de fluxo de trabalho e utilizou-se a recuperação de conhecimento como um estudo de caso para demonstrar os benefícios de abordagem de representação semântica.

Foi empregado também no modelo como um meio de expor dados experimentais como Dados Abertos Conectados (LOD) na *Web* e como base para enriquecer e conectar automaticamente os dados a tópicos e perfis de especialistas.

Neste contexto, o trabalho é uma abordagem centrada em tópicos para pesquisa de especialistas. Discorre sobre a extração de tópicos de especialistas, sua fundamentação semântica com a nuvem LOD e sua conexão com dados experimentais de RI, como mostra a **Figura 8**.

Vários métodos para criação de perfil e descoberta de especialistas foram analisados e avaliados, trazendo como resultados a possibilidade de construir perfis a partir de diretrizes extraídas de dados de forma automática com abordagens centradas em tópicos que mostram uma modelagem de linguagem de alto nível para encontrar dados especialistas.

Figura 8 - O fluxo de trabalho típico de avaliação experimental de IR e os dados produzidos

Fonte: Adaptado pela autora de SILVELLO *et al.* (2015, p. 147)

O trabalho apresenta a importância de se disponibilizarem os dados semanticamente na nuvem (LOD), gerando assim repositórios de uso e permitir a criação de perfis específicos de usabilidade.

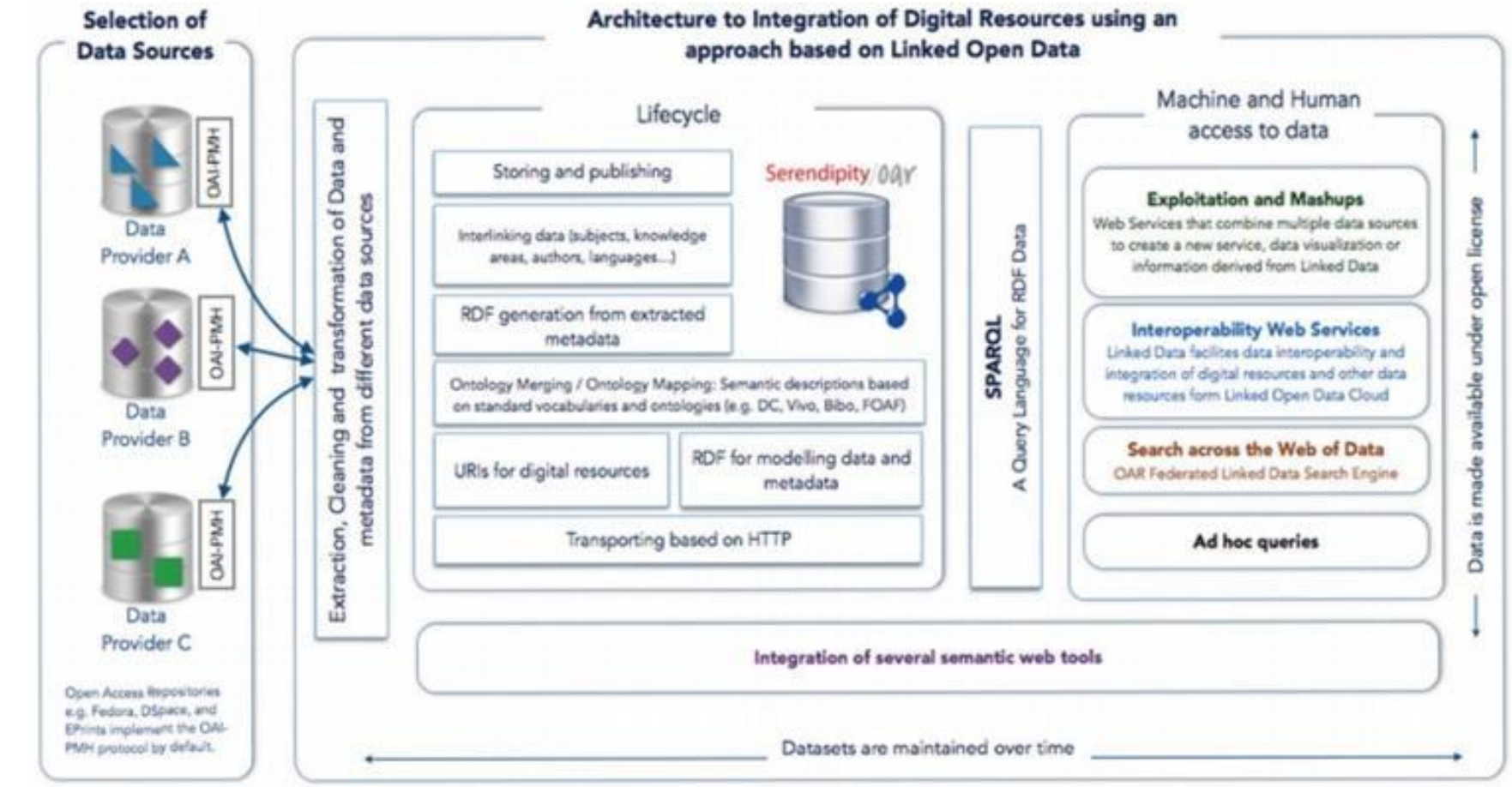
2.1.9 *Marco de Trabajo para la Integración de Recursos Digitales Basado en un Enfoque de Web Semántica*

Estrutura para a integração de recursos digitais com base em uma abordagem da *Web* Semântica desenvolvida por Piedra *et al.* (2015) trazem como proposta um ciclo que abrange os processos de extração de metadados dos repositórios OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*) e geração e publicação de dados conectados, com o objetivo de

melhorar a integração e interoperabilidade dos recursos armazenados nas Bibliotecas Digitais.

A proposta descrita facilita a existência de uma variedade de métodos e padrões nos processos de cada provedor de recursos digitais. A **Figura 9** mostra o processo de publicação de dados conectados de um conjunto de repositórios selecionados.

Figura 9 - Visão geral da estrutura



Fonte: PIEDRA *et al.* (2015, p. 61)

Piedra *et al.* (2015) apresentam a coleção da biblioteca digital que corresponde às universidades membros da Rede de repositórios do Consórcio Equatoriano para Desenvolvimento Avançado da Internet (CEDIA8). O CEDIA é composto por mais de 30 instituições de ensino superior do Equador e possui várias alianças internacionais com outras redes. Os materiais apresentados nos repositórios selecionados incluem: atlas, CDs, DVDs, *ebooks,* enciclopédias, brochuras, jogos, livros, memórias, revistas, teses; o que representa uma oportunidade imbatível para analisar os dados contidos.

Piedra *et al.* (2015) apresentam as etapas da visão geral do projeto que está apresentado na **Figura 9**. O protocolo OAI-PMH é a base do modelo de colheita de metadados de recursos acadêmicos, definidos de acordo com o esquema *Dublin Core*. O aplicativo *Harvester2* foi usado para coletar metadados através do OAI-PMH através do verbo "*listRecords*". Os metadados extraídos são armazenados em um banco de dados relacional na forma de triplas.

A limpeza dos dados extraídos é feita para detectar e corrigir dados corrompidos ou errados. O processo consiste em analisar padrões inconsistentes nos dados e executar um esquema de limpeza. Entre os casos detectados está a variação no formato de determinados metadados.

Ontologias e vocabulários abertos constituem o esquema baseado que descreve os recursos e entidades da *Web*. Portanto, eles foram examinados para representar recursos digitais, classificação de tópicos, descrição de organizações, autores, catálogos de dados e repositórios.

Para a geração em RDF dos dados coletados, um gerador próprio foi desenvolvido baseado em *Jena*. Um passo importante no processo de geração de dados RDF é atribuir URIs a textos extraídos; nesse sentido, os metadados dos recursos bibliográficos foram mapeados com os URIs dos termos mais apropriados, permitindo que os recursos possam ser interoperáveis e integrados a outros conjuntos de dados.

Os processos apresentados no trabalho contextualizam diretrizes importantes para que se possa mapear e permitir que esses dados sejam disponibilizados em repositórios. Visualiza-se ainda que determinam apenas um domínio específico, mas já apresentam características importantes para o trabalho proposto.

2.1.10 *Framework of Semantic Data Warehouse for heterogeneous and incomplete data*

O *Framework* para armazenagem de dados *(Data Warehouse*) semânticos para dados heterogêneos e incompletos apresentado por Nugraheni, Akbar e Saptawati (2016) tem como principal objetivo a construção do *Semantic Data Warehouse* (SDW), que analisa e extrai a demanda de conhecimentos derivados da colaboração de várias fontes.

Os dados apresentam desafios, pois possuem características estruturais heterogêneas, sintaxe e semântica, além da possibilidade de fontes de dados incompletos fornecerem os dados necessários para a análise.

É possível que os dados necessários não sejam fornecidos pela fonte de dados original, porém existem outros dados que podem ser usados para gerá-los. Nugraheni, Akbar e Saptawati (2016) propõem uma estrutura para o desenvolvimento de *Data Warehouse* semântico que lida com dados incompletos e problemas de dados de heterogeneidade (formato, sintaxe e estrutura) por meio do uso de Ontologias.

A estrutura também fornece ferramentas para lidar com fontes de dados incompletas, ou seja, a função de ferramentas disponíveis transforma as instâncias de objetos relevantes em classes de ontologia externas para gerar os dados necessários. A identificação dos elementos multidimensionais e o *design* conceitual do esquema multidimensional usam uma abordagem híbrida que combina a necessidade e a oferta.

A construção do esquema de cubo usando vocabulários QB4OLAP gera esquemas multidimensionais e cubos de dados RDF que podem executar operações OLAP (*Online Analytical Processing*).

No contexto de negócios dinâmicos e competitivos, os dados armazenados sem disponibilização não fornecem informações suficientes para processos de tomadas de decisão. Verificou-se, nesse trabalho, que as análises de negócios uma vez aprimoradas incluindo *Linked Open Data* (LOD) oferecem várias perspectivas para os tomadores de decisão, e ainda fornecem um novo modelo multidimensional, denominado *Unified Cube*, que oferece uma representação genérica para os dados armazenados e LOD no nível conceitual, de acordo com as necessidades dos tomadores de decisão.

2.1.11 OAM: *An Ontology Application Management Framework for Simplifying Ontology-Based Semantic Web Application Development*

Buranarach *et al.* (2016) apresentam o OAM, um *Framework* de gerenciamento de aplicativos ontológicos para simplificar o desenvolvimento de aplicativos Web Semânticos baseados em ontologia.

Embora os padrões de dados da *Web* Semântica já serem estabelecidos, os aplicativos baseados em ontologia criados com base nesses padrões, ainda são relativamente limitados, pois se devem, em parte, à alta curva de aprendizado e aos esforços exigidos na criação de aplicativos *Web* Semânticos baseados em ontologia.

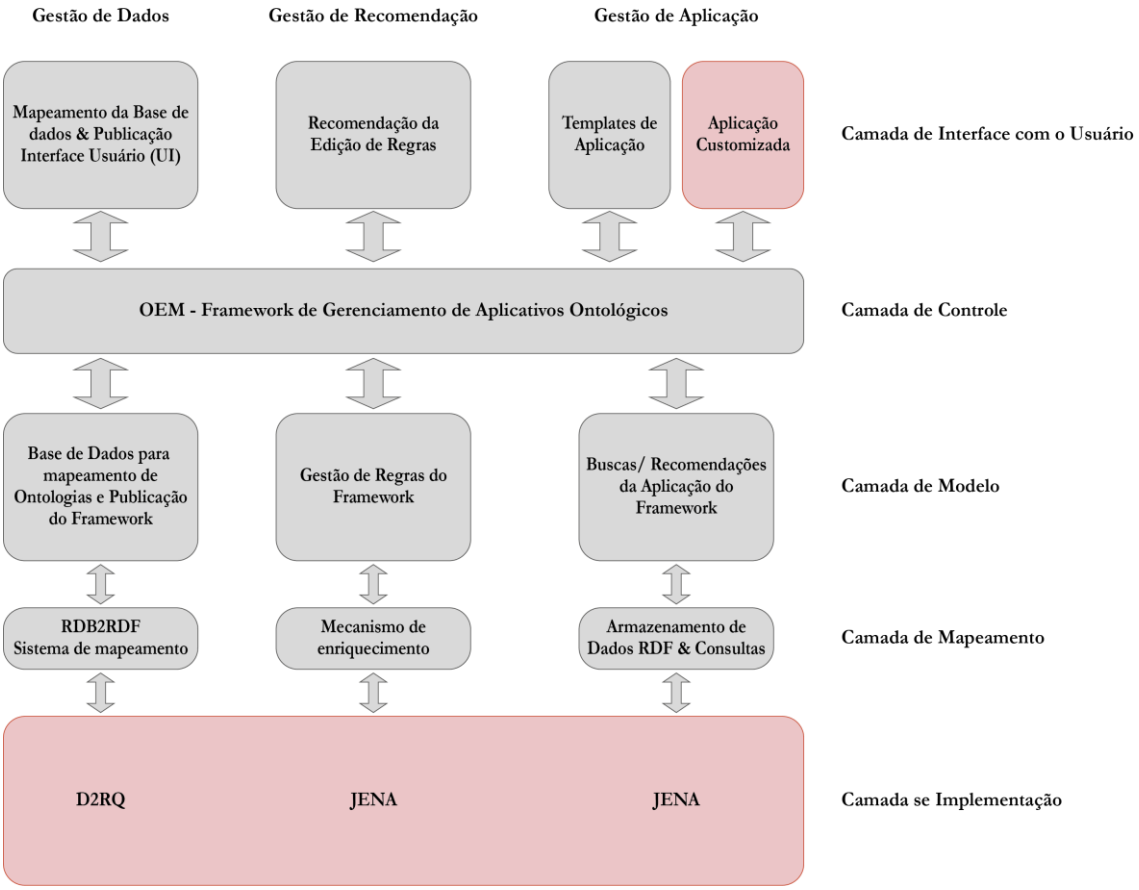
Dentro de tal contexto, Buranarach *et al.* (2016) descrevem uma estrutura de gerenciamento de aplicativos ontológicos (OAM) que visa simplificar a criação e adoção de aplicativos baseados em ontologias na tecnologia da *Web* Semântica.

O OAM introduz uma camada intermediária entre o aplicativo do usuário e o ambiente de programação e desenvolvimento, a fim de oferecer suporte à publicação e acesso de dados baseados em ontologia, abstração e interoperabilidade.

A estrutura se concentra no fornecimento de dados reutilizáveis e configuráveis e modelos de aplicativos, que permitem aos usuários criar os aplicativos sem a necessidade de habilidades de programação.

Três formas de modelos são introduzidas: configuração de mapeamento de banco de dados para ontologia, regra de recomendação e modelos de aplicativo, como mostra a **Figura 10**.

Figura 10 - Arquitetura em camadas da estrutura do OAM



Fonte: Adaptado pela autora de BURANARACH *et al.* (2016, p. 122)

Esse *Framework* mostra que, embora os padrões de dados da Web Semântica sejam estabelecidos, os aplicativos baseados em ontologia desenvolvidos com base nos padrões são relativamente limitados, podendo assim entender que isso se deve à alta curva de aprendizado e aos esforços exigidos na construção de aplicativos da Web Semântica baseados em ontologia.

A estrutura do *Framework* proposto, nessa seção, consiste em gerenciar aplicativos ontológicos (OAM) que visam simplificar a criação e a adoção de aplicativos baseados em ontologia que se baseiam na tecnologia da Web Semântica.

2.1.12 *Enriching Linked Data with Semantics from Domain-Specific Diagrammatic Models*

O trabalho “Enriquecendo dados conectados com semântica a partir de modelos diagramáticos específicos de domínio”, apresentado por Buchmann e Karagiannis (2016), apresenta uma abordagem para o enriquecimento de dados conectados com base em um modelo esquemático que é o RDFizer, que pode ser citado como exemplos os projetos D2RQ (*Accessing Relational Databases as Virtual RDF Graphs*) para bancos de dados relacionais, e XLWrap (para planilhas), onde as abordagens visam remover limites de troca de dados corporativos, abrindo-os para *links* entre organizações dentro de uma Web de dados.

Buchmann e Karagiannis (2016) afirmam que uma fonte insuficiente de semântica legível por máquina é a natureza gráfica subjacente dos modelos conceituais esquemáticos - um tipo de informação mais rica em comparação com o que normalmente é extraído dos esquemas de tabela, especialmente quando uma linguagem de modelagem específica de domínio é empregada. Grafos de dados de sistemas legados, empregando vários adaptadores originalmente desenvolvidos no contexto do projeto de pesquisa *ComVantage FP7*[4].

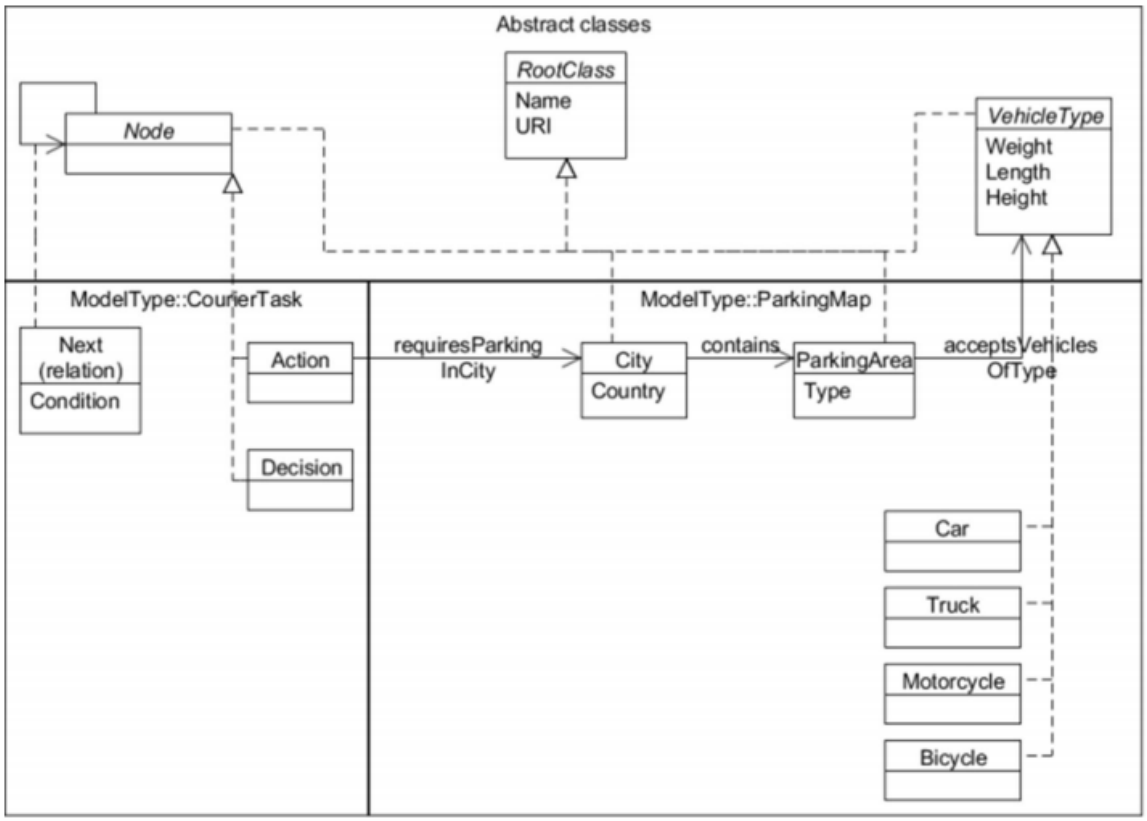
O trabalho fornece um exemplo ilustrativo, a partir do qual os argumentos serão generalizados, levando a uma visão proposta de sistemas de informação com "modelo conceitual desse processo de enriquecimento”.

A **Figura 11** apresenta o processo descrito em um diagrama UML (*Unified Modeling Language*) onde mostra o reaproveitamento da linguagem composta por dois tipos de modelos, segundo Buchmann e Karagiannis (2016).

O primeiro deles é um tipo de modelo *CourierTask rudimentar* fluxo de controle, permitindo a concepção de sequências de ações e decisões que descrevem as tarefas. E o segundo é o tipo de modelo *ParkingMap* que mostra a alocação de *ParkingAreas* (de diferentes tipos) para áreas geográficas (cidades).

[4] *Collaborative Manufacturing Network for Competitive Advantage* (CORDIS, 2017)

Figura 11 - Metamodelo para a linguagem de modelagem exemplar

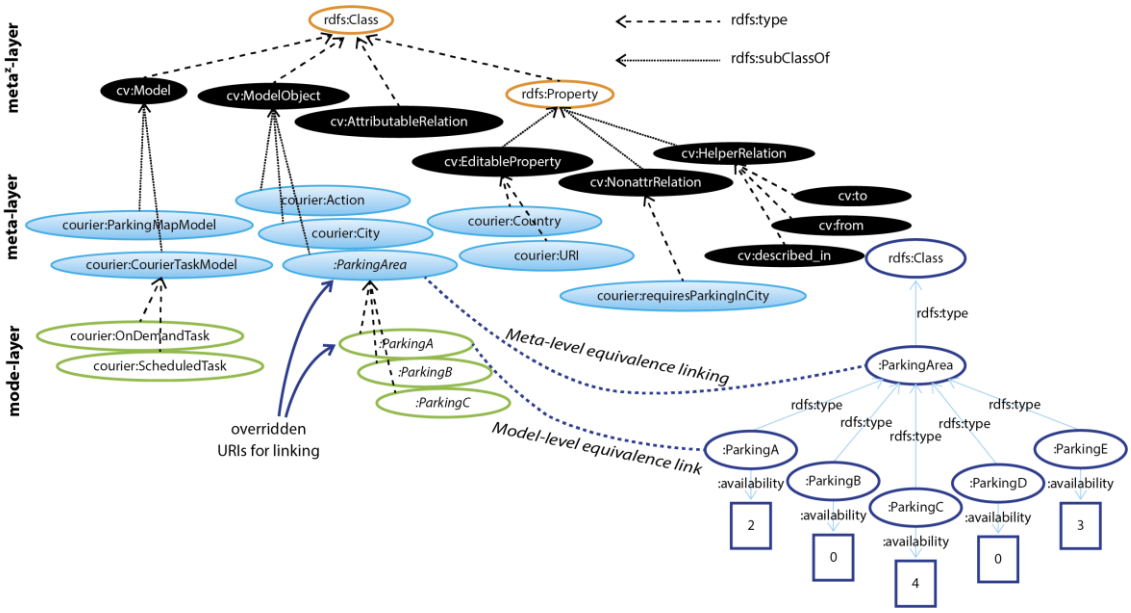


Fonte: BUCHMANN; KARAGIANNIS (2016, p. 344)

A **Figura 12** apresenta a reutilização fora do domínio do projeto, o esquema em cuja serialização foi baseado e modularizado em camadas de abstração com diferentes graus de flexibilidade.

Links podem ser estabelecidos entre elementos do modelo e dados abertos em tempo de execução em diferentes níveis de abstração.

Figura 12 - O vocabulário de serialização do modelo e os níveis de ligação do modelo



Fonte: BUCHMANN; KARAGIANNIS (2016, p. 347)

O artigo defende uma ligação entre os paradigmas de modelagem conceitual e *Linked Data*, para o benefício de permissão de sistemas de informação com reconhecimento de modelo conceitual, cujas características principais são o fato de diagramar os modelos ajuda na camada de abstração que os dados de execução (em relação a algum metamodelo/esquema de abstração superior) e assim os dados são enriquecidos semânticamente e disponíbilizados em forma de diagrama.

2.1.13 *Linked Data Enrichment with Self-Unfolding URIs*

Enriquecimento de dados vinculados com URIs auto desdobráveis é um trabalho desenvolvido por Szász, Fleiner e Micsik (2016), que apresenta um mecanismo que facilita o manuseio de inserções complexas e frequentes e fornece os benefícios da criação simples, juntamente com consultas e raciocínios avançados sobre os dados.

Verificou-se essa necessidade uma vez que os dados conectados são visualizados pelos identificadores de recursos uniformes. É uma etapa importante em qualquer projeto de Dados conectados, para definir as convenções para atribuições de URI.

Em alguns casos, os recursos já possuem seus identificadores naturais ou podem ser herdados de bancos de dados anteriores. No entanto, há casos em que inserções frequentes de conjuntos de triplas ocorrem sem nenhuma maneira conveniente de identificação e agrupamento deles.

O foco deste artigo é a expansão automática de Conjuntos de dados ligados, focando em dados ligados enriquecidos. Também apresentou como eliminar o tempo para consumir as etapas de criação de dados ligados com a busca de URIs que já estavam disponíveis em repositórios de forma automática, gerando uma descrição RDF para as entidades com base em informações codificadas em seu URI. Para a geração dessas novas triplas RDF, os autores propuseram modelos que podem ser implementados pelas consultas de inserção do SPARQL.

Neste artigo, Szász, Fleiner e Micsik (2016) apresentam um método para o tratamento de inserções complexas em um conjunto de dados. O método combina padrões de URI com modelos RDF para garantir que a representação RDF completa a nova entidade inserida e é gerado automaticamente.

2.1.14 *Working Framework of Semantic Interoperability for CRIS with Heterogeneous Data Sources*

O projeto “Estrutura de trabalho de interoperabilidade semântica para fontes de dados heterogêneas – CRIS”[5] apresentado por Leiva-Mederos *et al.* (2017) tem como objetivo fazer com que as informações dos sistemas atuais de informações de pesquisa *Community Research and Development Information Service for Science Research and Development* (CRIS), criado por Ivanović *et al.* (2013) e Joint (2008) armazenam em diferentes formatos e em plataformas incompatíveis ou mesmo em redes independentes.

Utilizaram uma metodologia apresentada na **Figura 13**, por Leiva-Mederos *et al.* (2017) para permitir o processamento de dados heterogêneos

[5] As bases do CRIS são definidas nos trabalhos de Pesquisa Comunitária Serviço de Informação e Desenvolvimento para Pesquisa e Desenvolvimento Científico. Essa mesma organização também criou, alguns anos depois, CERIF (*Common European Information Information Format*), um formato de dados de pesquisa que facilita a interoperabilidade dos sistemas e padroniza os dados necessários para a avaliação de Ciência (IVANOVIć *et al.*, 2013; JOINT, 2008).

de gerenciamento em um único *site*, de modo a aproveitar a capacidade de vincular dados dispersos encontrados em diferentes sistemas, plataformas, fontes e/ou formatos.

Com base nas funcionalidades do projeto VLIR (desenvolvido pelo laboratório de ICT de Cuba), Leiva-Mederos *et al. (*2017) trouxeram, como objetivo, apresentar um modelo que fornecesse a interoperabilidade por meio de técnicas de alinhamento semântico e metadados que facilitassem a fusão de informações armazenadas em diversas fontes.

Os autores trouxeram diversos estudos sobre mecanismos para alcançar a interoperabilidade semântica, e como análise trouxeram os seguintes aspectos: a cobertura específica dos conjuntos de dados (tipos de dados, cobertura temática e cobertura geográfica); as especificações técnicas necessárias para recuperar e analisar uma distribuição do conjunto de dados (formato, protocolo entre outros.); as condições de reutilização (direitos autorais e licenças); e as “dimensões” incluídas no conjunto de dados, bem como a semântica dessas dimensões (a sintaxe e as taxonomias de referência).

A estrutura de interoperabilidade semântica apresentada implementou um alinhamento semântico e uma faixa de metadados para converter informações de três vocabulários diferentes (*ABCD*, *Moodle* e *DSpace*) para integrar todos os bancos de dados em um único arquivo RDF.

Figura 13 - Metodologia desenvolvido para a interoperabilidade do CRIS para redes de usuários (TIC)

Fonte: Adaptado pela autora de LEIVA-MEDEROS *et al.* (2017, p. 486)

O trabalho trouxe como resultados uma avaliação baseada na comparação - por meio de cálculos de *recall* e precisão, ou seja, do modelo proposto e consultas idênticas feitas no *Open Archives Initiative* e SQL, para verificar sua eficiência. Os resultados foram satisfatórios o suficiente, devido ao fato de que a interoperabilidade semântica facilita a recuperação exata das informações.

2.1.15 A *Semantic Web based Framework for the Interoperability and Exploitation of Clinical Models and EHR Data*

Segundo Legaz-García *et al.* (2016), o advento dos sistemas de registros eletrônicos de saúde [*electronic healthcare records* (EHR)], desencadeou a necessidade de sua interoperabilidade semântica, o que é reforçado pelas oportunidades para o uso secundário de dados de EHR.

O uso conjunto de padrões EHR e recursos semânticos foram identificados como chave para a interoperabilidade semântica no projeto. Legaz-García *et al.* (2016) afirmam que, até o momento, as ferramentas existentes focadas nos padrões de EHR permitem criar, pesquisar, explorar modelos e mapear fontes de dados para modelos clínicos, mas não fornecem um suporte e integração adequados de recursos semânticos ou permitem o uso secundário de dados de EHR.

Visando a este conjunto de informações, foi descrita uma estrutura baseada em *Ontology Web Language* (OWL), o projeto denominado como “Uma estrutura semântica baseada na *Web* para a interoperabilidade e exploração de modelos clínicos e dados de RSE” que aproveita as tecnologias EHR e *Web* Semântica para a interoperabilidade e exploração de arquétipos, dados EHR e ontologias. Também permite o uso secundário de dados clínicos.

A estrutura foi implementada no *Archetype Management System* (ArchMS) e foi feita uma documentação que descreve como o ArchMS foi usado em um estudo real no domínio do câncer colorretal.

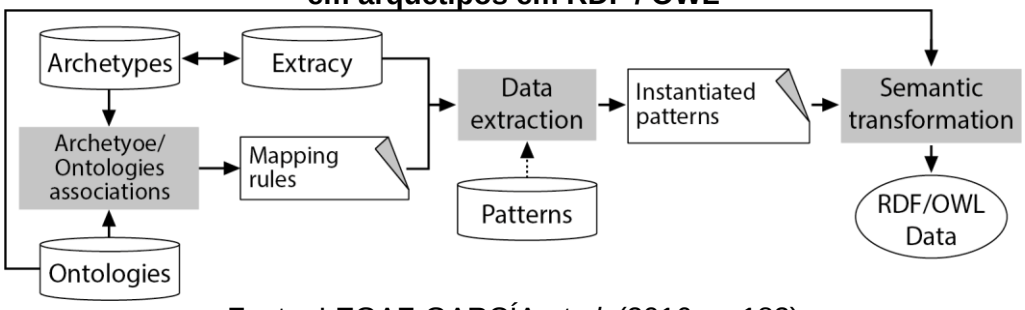
A **Figura 14**, segundo Legaz-García *et al.* (2016) fornece uma explicação mais detalhada deste processo. As regras de mapeamento são definidas entre os processos associados às ontologias de domínio de dados usados.

Verifica-se também que o processo de transformação usa padrões de ontologia, que são padrões que definem os modelos para a criação de indivíduos OWL com tipos específicos de axiomas e permitem a definição de regras de mapeamento complexas.

Um padrão é criado a partir de descrições formais de tipos de entidades e suas relações. Os padrões de ontologia são expressos usando as entidades

da ontologia do domínio de destino e podem conter variáveis. As variáveis são mapeadas nos elementos de dados.

Figura 14 - Visão geral do método para a transformação de dados EHR baseados em arquétipos em RDF / OWL



Fonte: LEGAZ-GARCÍA *et al.* (2016, p. 182)

O trabalho apresentado nessa seção, chamado ArchMS, trouxe a combinação de gestão e de serviços de interoperabilidade desenvolvidos e novas funções, entre as quais se destacam a transformação semântica e exploração de dados.

Os autores Legaz-García *et al.* (2016) ainda afirmam que os resultados mostraram o potencial das tecnologias da Web Semântica para o gerenciamento e exploração de dados para o EHR, e acharam que a abordagem poderia ser aplicada a outros padrões de modelo duplo, permitindo assim a distribuição dos dados de forma aberta e usados na integração de novos padrões e na melhoria dos métodos de transformação e recomendação.

2.1.16 Luzzu - *A Framework for Linked Data Quality Assessment*

A crescente variedade de dados vinculados na Web torna difícil determinar a qualidade desses dados e, posteriormente, tornar essas informações explícitas para os consumidores de dados.

Apesar da disponibilidade de várias ferramentas e estruturas para avaliar a qualidade dos dados vinculados, a saída dessas ferramentas não é adequada ao consumo da máquina e, portanto, os consumidores dificilmente podem comparar e classificar os conjuntos de dados na ordem de adequação ao uso.

Auer, Debattista e Lange (2016) apresentam o *Framework Luzzu*, uma estrutura para a avaliação de qualidade de dados conectados. O *Framework* segue um fluxo de trabalho em três etapas, começando com o processo de

inicialização da métrica (Etapa 1). Nesta etapa, as métricas definidas pelo usuário são compiladas e inicializadas junto com métricas implementadas em *Java*.

A avaliação da qualidade: o processo é iniciado (Etapa 2) por *streaming* sequencial de instruções do conjunto de dados candidato nas métricas de qualidade inicializadas. Uma vez que este processo é concluído, o processo de anotação (Etapa 3) gera metadados de qualidade e compila um relatório abrangente de qualidade.

Esta estrutura permite que os curadores de dados melhorem a qualidade do conjunto de dados usando o relatório para identificar problemas de qualidade no conjunto de dados.

O *Luzzu* é baseado em quatro componentes principais:

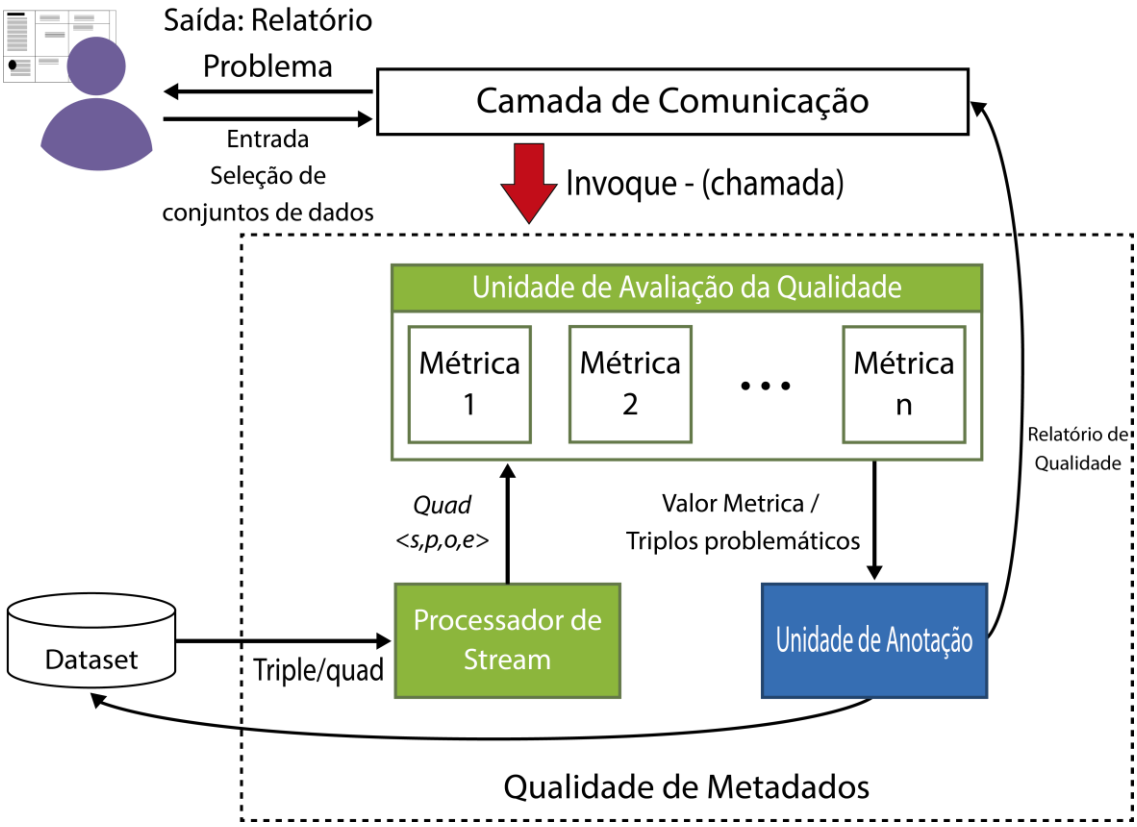
- (1) uma interface extensível para definir novas métricas de qualidade,
- (2) um *back-end* interoperável e orientado por ontologia para representar metadados e problemas de qualidade que podem ser reutilizados em diferentes estruturas semânticas,
- (3) um processador de fluxo escalável para despejos de dados e pontos de extremidade SPARQL; e
- (4) um algoritmo de classificação personalizável, levando em consideração os pesos definidos pelo usuário.

A **Figura 15** apresenta a Unidade de Avaliação de Qualidade que oferece muitas métricas de qualidade comuns para *download* na página inicial do *Luzzu*. Na sua implementação foram seguidos passos de uma pesquisa abrangente de dados conectados (*linked data*) com qualidade por Zaveri *et al.* (2015).

A estrutura foi criada com métricas personalizadas utilizando *simple Java interfaces* ou LQML (AUER; DEBATTISTA; LANGE, 2016), uma nova linguagem métrica de qualidade, desenvolvida pelos próprios autores.

Assim, o *Framework Luzzu* escala linearmente em relação ao número de triplas em um conjunto de dados e sua aplicabilidade, avaliando e analisando vários conjuntos de dados estatísticos em relação às métricas relevantes.

Figura 15 - Uma estrutura para avaliação de qualidade de dados conectados: o fluxo de trabalho da avaliação



Fonte: AUER; DEBATTISTA; LANGE (2016, p. 10, tradução nossa)

Assim, o *Framework Luzzu* apresenta uma escala linear em relação ao número de triplas em um conjunto de dados e sua aplicabilidade, avaliando e analisando vários conjuntos de dados estatísticos em relação às métricas relevantes.

2.1.17 *Linked Data* e Ciência da Informação: Diretrizes para Publicação de *Datasets* Institucionais Abertos

O trabalho apresentado por Nhancuongue, Rozsa e Dutra (2018) segue mais um contexto informacional propondo diretrizes para a publicação de dados ligados, segundo os princípios do *Linked Data*, a partir de um olhar da Ciência da Informação.

Nhancuongue, Rozsa e Dutra (2018) fizeram uma revisão bibliográfica e documental, seguida de um procedimento exploratório e aplicado, para avaliar

os alcances dos princípios do *Linked Data* sobre a disseminação de recursos nas transformações sociais.

Para gerar as diretrizes propostas, os autores supracitados utilizaram como base os trabalhos de Tim Berners-Lee relacionados às práticas de *Linked Data* e analisaram projetos e *datasets* da área educacional, por serem compostos, em sua maioria, por recursos abertos, acessíveis e reutilizáveis.

Os resultados da pesquisa, apresentados no **Quadro 1**, contêm os resultados resumidos nas principais diretrizes para a Publicação de Conjuntos de Dados, reforçando a iniciativa de profissionais, pesquisadores e instituições para o reuso e interoperabilidade.

Quadro 1 - Diretrizes para a Publicação de *Datasets* Abertos

Diretriz	Descrição
Utilização de URIs adequadas	As URIs são os identificadores únicos dos recursos nos *datasets*. Então, além dos princípios básicos do *Linked Data*, considere: ● Utilizar apenas URIs de domínios que você controle; ● Ocultar detalhes de implementação; ● Utilizar chaves naturais nas URIs.
Vínculo com outros datasets	De modo a conectar-se a outros *datasets* e aumentar a visibilidade dos dados (tanto próprios como de terceiros) crie: ● Ligações de relacionamento; ● Ligações de identidade; e ● Ligações de vocabulários. ● Também, quando possível, gere ligações a partir de outros *datasets* para o seu (ex.: a partir do DBPedia).
Reuso de Vocabulários	Buscar reutilizar vocabulários populares para que os dados possam ser compreendidos por aplicações que consomem dados de diferentes repositórios. Manter equilíbrio entre vocabulários populares e específicos de domínio.

Disponibilização de Metadados	Gerar metadados sobre o *dataset* por meio da utilização do vocabulário VoID, para facilitar seu reconhecimento por potenciais usuários.
Publicação do *dataset*	Submeta o *dataset* ou crie uma referência para ele em uma ferramenta centralizadora (ou catálogo) *online*, como o *Datahub*.

Fonte: NHACUONGUE; ROZSA; DUTRA (2018, p. 73)

As principais conclusões estão ligadas à linha que versa sobre as novas perspectivas na Ciência da Informação sobre o *Linked Data,* ou dados conectados, e à outra linha é sobre a necessidade de comprometimento dos profissionais da informação com questões relacionadas ao *Linked Data,* de modo a ampliar o universo informacional e o ambiente de satisfação das necessidades dos usuários.

2.1.18 *La publicación em Linked Data de Registros Bibliográficos: Modelo e Implementación*

O projeto “A publicação em *Linked Data* de registros bibliográficos: modelo e implementação” dos autores Senso e Machado (2018) traz uma visualização de todos os processos necessários para publicação de dados de bibliotecas, incluindo o contexto do mapeamento e enriquecimento Semântico.

Senso e Machado (2018) apresentam que as bibliotecas estão intimamente ligadas ao *Linked Data* devido ao alto nível de estruturação de seus dados, embora os projetos relacionados a ele sejam elaborados principalmente por grandes bibliotecas.

O trabalho proposto por Senso e Machado (2018) analisa alguns projetos, ciclos de vida e ferramentas envolvidos durante o processo da publicação de dados de forma ligada e estabelecem uma metodologia e executam a implementação completa, convertendo registros bibliográficos em dados conectados e enriquecendo-os através de outros conjuntos de dados, disponibilizando na Web.

O trabalho traz um estudo de caso utilizando um conjunto de registros extraídos da Biblioteca da Universidade de Granada, a fim de conhecer alguns

dos problemas que podem ser encontrados por qualquer centro que queira converter seus registros em dados ligados sem precisar alterar o sistema de automação de biblioteca.

Como mostra o **Quadro 2**, Senso e Machado (2018) criaram diretrizes para publicar os dados na *Web*.

Quadro 2 - Proposta de metodologia para publicação de registros bibliográficos como *Linked Data*

Etapa	Descrição	Tarefas
1 – Determinar	Identificação e descrição dos dados	*a. Identificar e analisar os dados e a fonte de dados (software, formato, banco de dados ...)* *b. Identificar sua licença* *c. Determinar uma licença*
2 – Limpar	Armazenamento e correção de dados	*Curadoria de dados*
3 – Modelar	Desenvolvimento de um vocabulário para descrever Dados RDF	*a. Selecionar vocabulários* *b. Criação de Mapa* *c. Atribuir URIs*
4 – Gerar	Geração de recursos RDF	***a.*** *Selecionar as tecnologias para a geração do RDF* ***b.*** *Transformar dados de origem em RDF* ***c.*** *Validar*
5 - Ligação dos Dados	Conecte o conjunto de dados a outras pessoas que Enriquecer	*a. Buscar datasets relevantes* *b. Descobrir relacionamentos* *c. Ligar os dados* *d. Verificar as ligações dos dados (integridade)*

6 – Publicar	Publicação do conjunto de dados	a. Escolher o formato e a plataforma b. Publicar o dataset c. Publicar seus metadados

Fonte: SENSO; MACHADO (2018, p. 6, tradução nossa)

Senso e Machado (2018) enfatizam no texto e mostram sua operação empiricamente, que produziu como resultados esperados a aplicação a conjuntos de dados que geram muitas complicações em seu gerenciamento, mas produzem um conjunto de dados no *Linked Data* com o esquema de implantação de 5 estrelas.

2.1.19 *Expressing the Tacit Knowledge of a Digital Library System as Linked Data*

Di Iorio e Schaerf (2019) apresentam uma abordagem para codificar, como Vocabulário Vinculado (conectado), o conhecimento "tácito" do sistema de informação que gerencia uma fonte de dados. O objetivo do trabalho é o processo de interpretação do significado dos dados vinculados dos conjuntos de dados publicados.

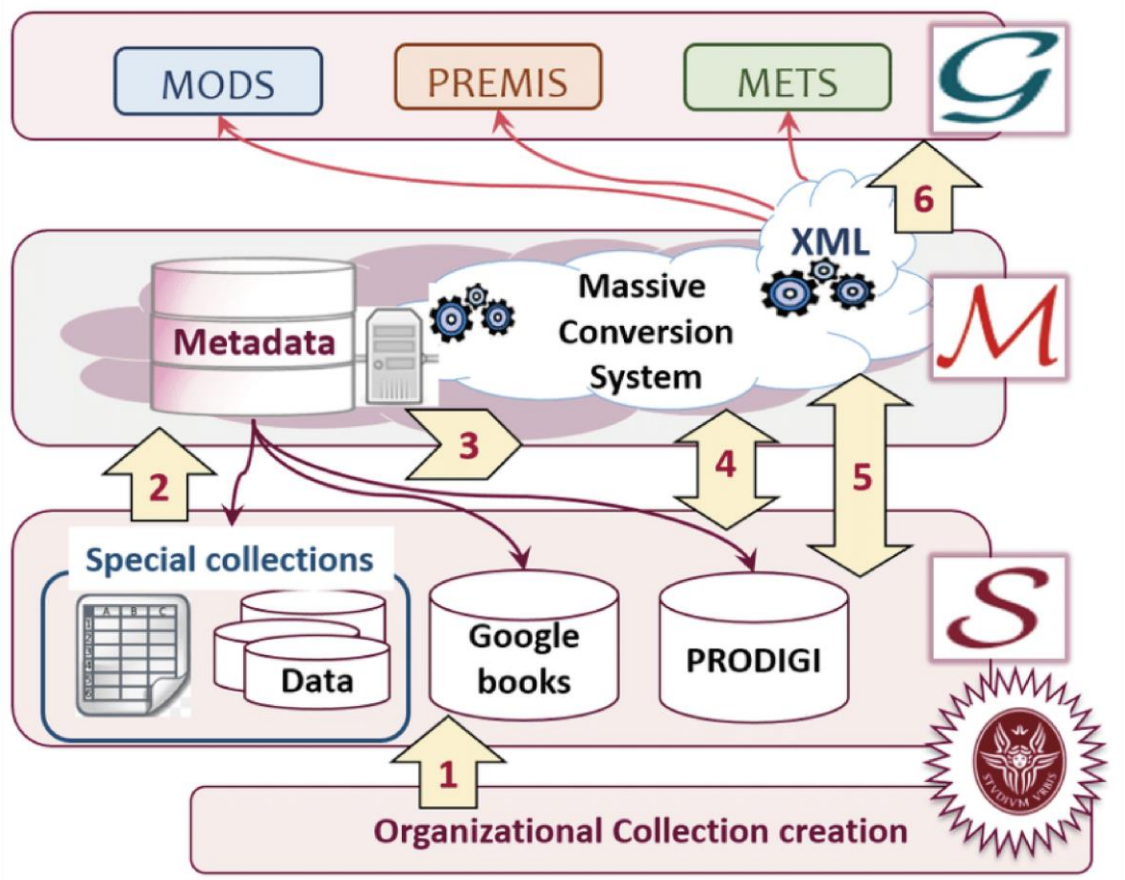
O trabalho traz uma análise de um sistema de biblioteca digital, como um estudo de caso, para prototipar o método de “gerenciamento de dados semânticos”, onde os dados e seu conhecimento são gerenciados de forma nativa, levando em consideração os pilares de dados vinculados.

Nesse contexto, o gerenciamento dos dados semânticos vem para curar a interpretação correta dos dados pelos consumidores e facilitar a reutilização adequada. O protótipo define as entidades ontológicas que representam o conhecimento do sistema de biblioteca digital, que não é armazenado na fonte de dados, nem nas ontologias existentes relacionadas à semântica do sistema.

Assim, a ontologia local e sua correspondência com as ontologias existentes, Estratégias de Implementação de Metadados de Preservação (PREMIS) e Esquema de Descrição de Objetos de Metadados (MODS) e são discutidas triplas de dados vinculados prototipados do banco de dados relacional herdado, usando a ontologia local.

Esta seção descreve o sistema de conversão massiva (*MassConv*), o estudo de caso do sistema DigLib. A **Figura 16** mostra uma visão geral abstrata do gerenciamento de dados realizado pelo sistema *MassConv,* como componente da biblioteca digital *Sapienza* (SDL), um trabalho em nome da Universidade Sapienza.

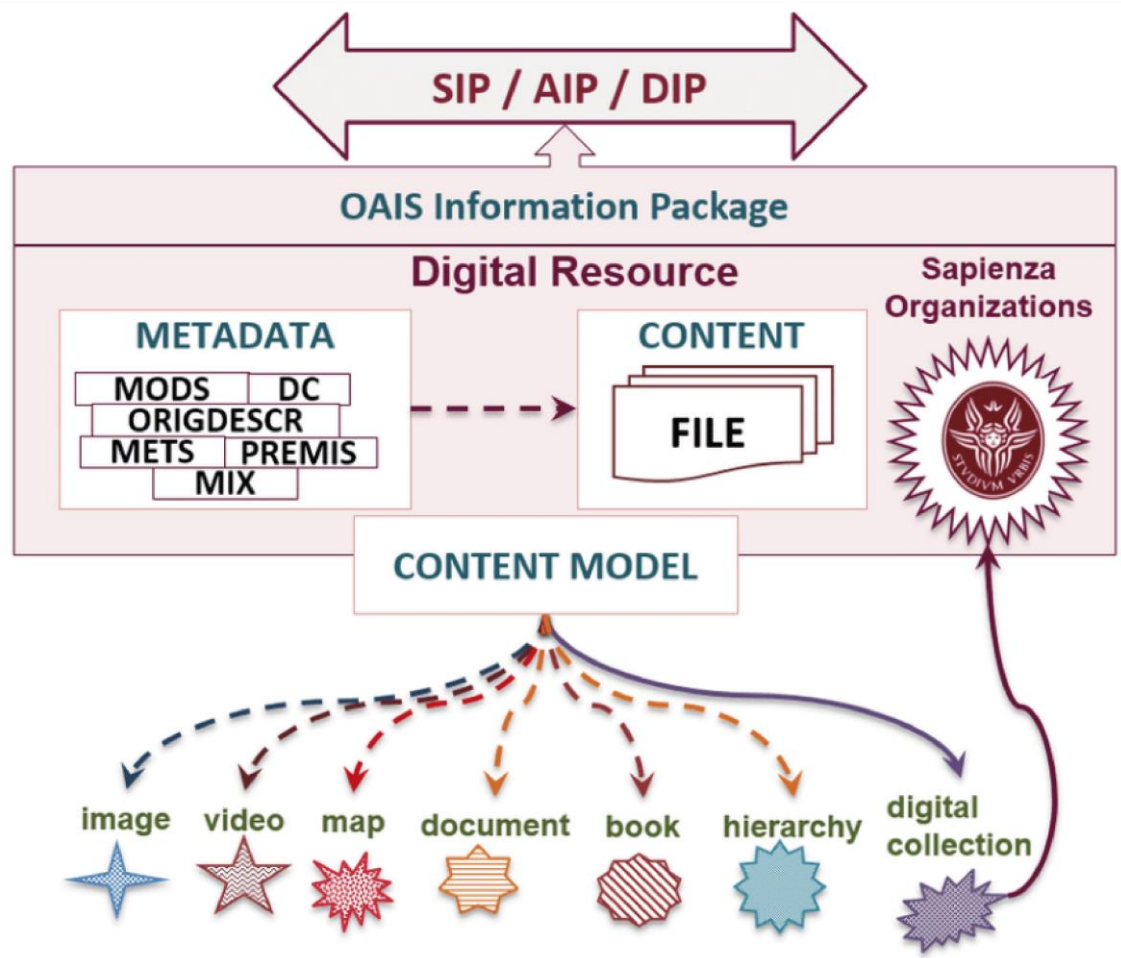
Figura 16 - Representação abstrata de elementos arquitetônicos e fluxo de trabalho de dados do *Sapienza* digital sistema de conversão massiva da biblioteca (SDL)



Fonte: DI IORIO; SCHAERF (2019, p. 6, tradução nossa)

MassConv é um sistema de gerenciamento de dados baseado em RDF, onde os dados gerenciados são coletados e usados para produzir recursos digitais (**Figura 17**) em conformidade com os padrões de metadados estabelecidos para os recursos digitais. Os dados, gerenciados pelo sistema *MassConv*, descrevem o SDL objetos multimídia.

Figura 17 - Modelo estrutural de um recurso digital SDL



Fonte: DI IORIO; SCHAERF (2019, p. 7, tradução nossa)

O protótipo mostra como o gerenciamento de dados semânticos pode lidar com a inconsistência dos dados do sistema e verificou que uma mudança específica na mentalidade do desenvolvedor do sistema é necessária para extrair e “codificar” o conhecimento tácito, necessário para melhorar os dados. processo de interpretação.

2.1.20 *Ontology-driven Open Data Enrichment Framework for Value Optimisation: The Case of National Performance Indicators for Botswana*

O trabalho apresentado por Sebubi; Zlotnikova; Hlomani (2019) apresenta um artefato de *design*, o *Open Data Enrichment Framework,* dirigido por uma Ontologia, como uma entrega preliminar de uma pesquisa que propõe uma estrutura semântica de dados abertos sobre Indicadores Nacionais de

Desempenho (NPI's) dentro do contexto do *National Integrated Indicator Framework* de Botsuana.

A estrutura proposta aborda o problema de menor impacto dos dados abertos resultante de heterogeneidades relacionadas a plataformas de dados, padrões, modos de acesso, modelos, formatos e definições. A estrutura mapeia uma transição de dados abertos estáticos e desintegrados nos NPIs para o ILOD (*Intelligent Linked Open Data*) para otimização do valor.

Os autores supracitados ainda confirmam que esta pesquisa em andamento segue a metodologia de pesquisa em ciência do design, e o trabalho apresenta a saída das três primeiras fases da metodologia. Os resultados completos e totais só serão apresentados assim que as fases de implementação e avaliação estiverem concluídas.

A estrutura finalizada segundo Sebubi; Zlotnikova; Hlomani (2019) servirá como um modelo na conversão de dados abertos estáticos nos NPIs em conteúdo interligado e inteligente, além de se obterem benefícios acadêmicos, sociais e comerciais significativos e ainda desejam mostrar a amplitude da pesquisa, em que hoje a estrutura está apenas alinhada com Botsuana, mas a ideia é permitir que ela possa ser personalizada para atender às necessidades de outros países em desenvolvimento.

2.2 ANÁLISES DOS TRABALHOS

Nesta seção, analisam-se os vinte projetos de mapeamento e enriquecimento semântico dos dados e as diretrizes que levaram a construir esses processos, baseando-se em características. Tal atividade é descrita detalhadamente no **Quadro 3** e no **Quadro 4**.

O **Quadro 3** apresenta características de autor, título e ano de publicação dos artigos.

Em relação às características apresentadas no **Quadro 4**, a primeira característica está associada ao uso ou não de Ontologias e/ou vocabulários, uma vez que a construção do enriquecimento de dados se baseia em padrões já estipulados e determinados para que o mapeamento seja realmente relevante à informação buscada.

Em relação ao nível de automação, segunda característica do **Quadro 4**, está relacionada em como são extraídos e analisados os dados, ou seja, o processo minerador desses dados, e quais ferramentas foram utilizadas ou não no processo de mapeamento.

Na característica determinada como Repositório, verifica-se a importância do enriquecimento dos dados nos trabalhos citados, se utilizaram vocabulários e se eles foram disponibilizados como um repositório aberto para uso futuro.

Outra característica importante no **Quadro 4** é referente a como os trabalhos foram tratados, ou seja, como eles assumiram suas categorias de tratamento. Verifica-se que não existe um padrão adotado, em algumas são usados como método/metodologia, diretriz ou *Framework*. Essa característica é importante para adotar e justificar o uso da palavra *Framework* como padrão para esse trabalho.

Segundo Pereira (2001), um conjunto de métodos ou metodologias pode ser classificado como processos práticos que, através de técnicas e ferramentas voltadas para coleta, processamento e difusão e uso das informações, ou, como no contexto os dados de determinado domínio, são usados com o intuito de transformação desses dados em informações de relevância e um determinado propósito, para que assim sejam usadas e vistas à tomada de decisão que as cabe no contexto semântico.

March e Smith (1995, p. 257) esclarecem que “[...] métodos baseiam-se em um conjunto de constructos subjacentes (linguagem) e uma representação (modelo) em um espaço de solução [...]”. Os autores ainda explicam que os métodos podem ter modelos como entrada que submetidos a uma sequência de passos levam à solução de um problema.

Sendo assim *Framework* pode ser apresentado como um método, ou então como um conjunto de passos para se executar uma tarefa.

O intento de um *Framework* Conceitual é gerar soluções para uma família de problemas de um determinado domínio no âmbito de entendimento que esse problema se encontra. Podendo gerar um modelo de dados que seja compreensível e então aplicado a um modelo de *Framework*, de forma mais conceitual ou computacional, segundo os autores Foote e Jhonson (1988),

Gamma *et al.* (1995), Taligent White Paper (1995) e Shehabuddeen, Probert e Phaal (2000).

Bem e Coelho (2013) apresentam *Frameworks* como essenciais em áreas onde não se têm a total compreensão conceitual de problemas propostos em algumas áreas, ou seja, que ainda alguns pontos estudados não possuem uma harmonia e nem uma clareza conceitual na literatura.

Quadro 3 – Trabahos Correlatos

Número	Trabalho	Autor	Data publicação
1	*PowerMap: Mapping the Real Semantic Web on the Fly*	Lopez, Motta e Sabou	2006
2	*Semantically Enhanced Information Retrieval: An Ontology-Based Approach*	Fernández *et al.*	2011
3	*A Framework for Semantic Enrichment of Sensor Data*	Mararu e Mladenic	2012
4	*OpLiDaF Open Linked Data Framework: a platform for the creation and publication of Linked Data*	Possemato	2013
5	*Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling*	Schaible, Gottron e Scherp	2015
6	*Towards an Intelligent Framework for Multimodal Affective Data Analysis*	Poria *et al.*	2015
7	*Ferramenta para Extração de Dados Semiestruturados para Carga de um Big Data*	Furtado *et al.*	2015
8	*Semantic Representation and Enrichment of Information Retrieval Experimental Data*	Silvello *et al.*	2015
9	*Marco de Trabajo para la Integración de Recursos Digitales Basado en un Enfoque de Web Semántica*	Piedra *et al.*	2015
10	*Framework of Semantic Data Warehouse for heterogeneous and incomplete data*	Nugraheni, Akbar e Saptawati	2016
11	*OAM: An Ontology Application Management Framework for Simplifying Ontology-Based Semantic Web Application Development*	Buranarach *et al.*	2016
12	*Enriching Linked Data with Semantics from Domain-Specific Diagrammatic Models*	Buchmann e Karagiannis	2016
13	*Linked Data Enrichment with Self-Unfolding URIs*	Szász, Fleiner e Micsik	2016
14	*Working Framework of Semantic Interoperability for CRIS with Heterogeneous Data Sources*	Leiva-Mederos *et al.*	2017
15	*A Semantic Web based Framework for the Interoperability and Exploitation of Clinical Models and EHR Data*	Legaz-García et al.	2016

16	*Luzzu - A Framework for Linked Data Quality Assessment*	Auer; Debattista; Lange	2016
17	*Linked Data e Ciência da Informação: Diretrizes para Publicação de Datasets Institucionais Abertos*	Nhancuongue, Rozsa e Dutra	2018
18	*La publicación em Linked Data de Registros Bibliográficos: Modelo e Implementación*	Senso e Machado	2018
19	*Expressing the Tacit Knowledge of a Digital Library System as Linked Data*	Di Iorio e Schaerf	2019
20	*Ontology-driven Open Data Enrichment Framework for Value Optimisation: The Case of National Performance Indicators for Botswana*	Sebubi; Zlotnikova; Hlomani	2019

Fonte: Elaborado pela autora (2021)

Quadro 4 - *Proposta* de metodologia para publicação de registros bibliográficos como *Linked Data*

TRABALHO	USO DE ONTOLOGIA E/OU VOCABULÁRIOS	NÍVEL DE AUTOMAÇÃO	REPOSITÓRIO	CATEGORIZAÇÃO
1	Sim	Automatizado	Não	*Framework* Metodologia
2	Sim	Automatizado	Não	Diretrizes
3	Sim	Automatizado	Sim disponibilizado no LOD	*Framework*
4	Sim	Automatizado	Sim disponibilizado no LOD	*Framework*
5	Sim	Manual	Não	Diretrizes
6	Sim	Automatizado	Sim disponibilizado no LOD	*Framework*
7	Sim	Manual Ontologia Automatizado Mapeamento	Não site de busca	*Framework*
8	Não apresenta	Automatizado	Sim disponibilizado no LOD	Diretriz
9	Sim Vocabulário DBPedia	Automatizado	Sim disponibilizado em 4 repositórios	Diretriz

10	Sim	Automatizado	Sim disponibilizado no LOD	*Framework*
11	Sim	Automatizado	Sim para publicação	*Framework*
12	sim	Automatizado	Sim, em formato de diagramas e repositórios abertos.	Método Metodologia
13	Sim	Automatizado, usando Sparql	Sim, em repositórios abertos.	Método Metodologia
14	DSpace	Análise Teórica	Não	*Framework*
15	Sim	Sim	Repositório Clínico	*Framework*
16	DBPedia	Sim	Sim, em repositórios abertos	*Framework*
17	Não	Manual	Não	*Diretrizes*
18	MARC21	Sim	Sim, em repositórios abertos.	Metodologia
19	Sim	Sim	Sim disponibilizado no LOV	*Diretrizes*
20	Não	Manual	Não	*Framework*

Fonte: Elaborado pela autora (2021)

Justificam-se a busca e apresentação de todos esses trabalhos para que a visibilidade computacional traga benefícios e ajude na visão da CI para o entendimento do processo que está em estudo, que é o processo de extração, limpeza, enriquecimento e mapeamento dos dados de forma semântica. Mostra ainda, que a maioria dos trabalhos apresentados propõe diversas formas de desenvolver esse mapeamento e enriquecimento, mas não todas aplicáveis, nem todas compreendidas a qualquer domínio.

3. METODOLOGIA

3.1 PROCEDIMENTOS METODOLÓGICOS

Nesta seção são apresentados os procedimentos metodológicos aplicados no trabalho, compreendendo a fundamentação, o percurso da pesquisa, sua caracterização e a análise realizada para obtenção dos resultados.

Gil (2010, p.12) afirma que a pesquisa é o “[...] procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos”, isto quer dizer que, por meio da pesquisa, propõe-se buscar um retorno para o assunto em questão, seja uma dúvida ou uma pergunta.

O projeto de pesquisa, segundo Severino (2007), precisa ter vários elementos, e os procedimentos metodológicos são de extrema importância para a construção mais concisa da pesquisa. Com base em tais aspectos, as próximas seções apresentam, no contexto do objetivo do presente trabalho, que é a construção de um ***Framework*** para o processo de mapeamento e enriquecimento semântico de dados, todos os passos para a construção do procedimento metodológico, desde a finalidade da pesquisa, seus meios e qual a abordagem e natureza, os métodos aplicados, a técnica de análise e a apresentação gráfica escolhida.

3.1.1 Abordagem, Natureza, Tipo e Método de Pesquisa

A pesquisa caracteriza-se como sendo de natureza aplicada, do tipo descritiva e de abordagens qualitativa. Flick (2009) afirma que a aplicação de muitos métodos no processo da pesquisa auxilia na construção com um olhar analítico, gerando uma diversidade de particularidades, dependendo onde se quer aplicar. Cria-se, assim, o contexto sobre a triangularização da pesquisa defendida por Denzin (1978) que corroborou a perseguir com interpretações conflitantes, mesmo que contendo axiomas paradigmáticos comuns, o autor supracitado defende que visões diferentes devem ser estudadas a ponto de se chegar a uma “verdade”, para assim serem apresentadas e discutidas.

Com base nos objetivos destacados do presente estudo, uma pesquisa científica pode ter abordagens de tratamentos qualitativos, os quais se complementam, permitindo análises e discussões que fomentam os resultados (MINAYO, 2012).

Tratando-se de pesquisa qualitativa, segundo Chizzotti (2006, p. 42), o termo “[...] implica uma partilha densa com pessoas, fatos e locais que constituem objetos de pesquisa, para extrair desse convívio os significados visíveis e latentes que somente são perceptíveis a uma atenção sensível.”

Sampieri, Collado e Lucio (2013) determinam que uma investigação qualitativa é guiada por áreas ou tópicos de pesquisa significativos. Mas, em vez de clareza sobre questões de pesquisa e hipóteses, precedem a coleta e análise de dados. Estudos qualitativos podem desenvolver perguntas e hipóteses antes, durante ou depois da coleta e análise de dados.

Muitas vezes as atividades servem em um primeiro momento para descobrir quais são as questões de pesquisa mais importantes, e depois refinar e respondê-las. A ação investigativa move-se para dinâmica em ambas as direções: entre os fatos e sua interpretação, e é um processo bastante “circular” e a sequência nem sempre é a mesma, varia de acordo com cada estudo em particular.

A pesquisa qualitativa pode ser considerada como uma pesquisa mais aprofunda, rica de detalhes, e que possui características como: observar, compreender e explicar (GERHARDT *et al.*, 2009, p.31).

Minayo (2012, p. 623) defende que a pesquisa qualitativa “[...] trabalha com o universo de significados, motivos, aspirações, crenças, valores e atitudes, o que corresponde a um espaço mais profundo das relações [...]”, isto é, uma vasta explanação sobre os assuntos.

Observa-se que a pesquisa quantitativa é algo tangível, o que possibilita a contagem dos resultados obtidos. Além disso, permite a representação em gráficos ou quadros, exibindo melhor os dados pesquisados.

Gil (2010) alega que a aproximação de fatores qualitativos com amostragens de dados quantitativos nas pesquisas relacionadas à área da Ciência viabiliza uma amplitude na concepção dos problemas que encontrados na atual realidade.

Grimes e Lewis (2005) partem do contexto definido por Denzin (1978) e

apresentam o propósito da triangularização determinando três fases: Fundamentos, Análise de Dados e Construção de Teorias.

No que diz respeito a Fundamentos apresentam que as diretrizes devem estar relacionadas a prover foco, dar agilidade ao processo interpretativo, atingir um entendimento, sentido ou significado àquilo que se acredita, e reconhecer o paradigma de origem, além de salientar referências empíricas comuns (GRIMES; LEWIS, 2005).

A análise de dados contempla vertentes ligadas ao reconhecimento das influências paradigmáticas enfatizando diferenças, mas mantendo um equilíbrio para semear interpretações distintas dos dados apresentados. É um processo que destaca vários tipos de paradigmas, além de gerar a experiência da sua linguagem e gerenciar os entendimentos cumulativos (GRIMES; LEWIS, 2005).

Quando se diz respeito à “Construção das Teorias”, Grimes e Lewis (2005) afirmam que é necessário conduzir os experimentos cognitivos e aproximar as ideias dos tipos de paradigma, abrangendo as discrepâncias e as correlações; ponto que motiva a relação e indica a qualidade e o processo de construção teórica.

O estudo utilizou de uma abordagem uma abordagem quantitativa, visto que, conforme apresenta Sampieri, Collado e Lucio (2013), a abordagem qualitativa busca "dispersar ou expandir" dados e informações, abordagem a trabalhar no projeto.

Uma vez que o modelo de *Framework* proposto apresenta as diretrizes de Mapeamento e Enriquecimento Semântico de dados, precisa de uma abordagem que traga o estudo quantitativo, de como todos os passos do mapeamento devem ser feitos, e quais passos e ferramentas tecnológicas devem-se usar. Na abordagem qualitativa, leva-se em conta o resultado dos dados enriquecidos, tendo como apoio a aplicação das Boas Práticas para publicação de dados na *Web* (LÓSCIO et al., 2018) utilizando-se de conceitos oriundos da Web Semântica para construir um modelo que apresente as teorias propostas, buscando responder àpergunta de pesquisa explicitada anteriormente.

Quanto à natureza da pesquisa, verifica-se que ela assume dois aspectos: básica e aplicada. Gil (2010, p. 128) esclarece que a pesquisa

básica “[...] objetiva gerar conhecimentos novos para avanço da ciência sem aplicação prática prevista [...]”, enquanto a pesquisa aplicada “[...] objetiva gerar conhecimentos para aplicações práticas [dirigidas] à solução de problemas específicos.” (GIL, 2010, p. 128).

A Pesquisa aplicada visa resolver problemas que já existem na prática, de forma imediata ou não, e no contexto do trabalho. A natureza da pesquisa está relacionada com os procedimentos práticos, ou seja, aplicada, uma vez que determinando a construção do *Framework*, mesmo que conceitual, auxiliará a orientação e possíveis soluções dos passos existentes e necessários ao mapeamento dos dados e sua transformação em dado semântico gerando conteúdos enriquecidos.

Outro recurso para colaborar com o estudo em questão é quanto aos objetivos da pesquisa que, segundo Gil (2010), podem ser classificados como exploratório, descritivo ou explicativo.

Em conjunto com a pesquisa aplicada, verifica-se que o presente trabalho se caracteriza também como sendo do tipo descritivo, uma vez que pretende descrever o modelo do *Framework* conceitual, trazendo diretrizes de enriquecimento semântico de dados e todas as etapas do mapeamento dos dados.

A pesquisa descritiva, de acordo com Gil (2010, p. 42), “[...] tem como objetivo primordial a descrição das características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis”. Significa que descreve os fatos de uma realidade, além de permitir a generalização sobre a realidade pesquisada.

Na pesquisa descritiva leva-se em conta aspectos como formulação de perguntas que direcionam a pesquisa e, nesse contexto, estabelece uma relação entre as dimensões propostas no objeto de estudo em análise (GERHARDT et al., 2009).

Segundo TRIVIÑOS (1987), tal direcionamento estimula o pesquisador a desenvolver toda a análise, registro e esclarecer os fatos, sem que haja interferência ou manipulação da análise, ou seja, é realizado um estudo detalhado, com coleta de dados, análise e interpretação, sem que haja interação ou envolvimento do pesquisador no assunto analisado. Sua responsabilidade é expor a periodicidade de como o fato estudado ocorre ou

como se comporta ou se estrutura dentro de um determinado sistema, método, processo ou realidade operacional.

A pesquisa descritiva aplica técnicas padronizadas de coleta de dados para apresentar as diferentes e possíveis propostas referentes ao objeto de estudo, que podem estar ligadas, desde características socioeconômicas de um grupo, até características que podem ser alteradas durante o processo, sendo apresentadas na forma de: pesquisa documental, estudo de campo e levantamentos de dados (TRIVIÑOS, 1987).

Em relação ao conjunto de técnicas de análise das comunicações, ou seja, que permitiu a análise dos significantes como significados da mensagem, a Análise de Conteúdo foi escolhida para ser utilizada no trabalho e, segundo Gil (2010, p.112), é “[...] uma técnica de investigação que, através de uma descrição objetiva, sistemática e quantitativa do conteúdo manifesto das comunicações, tem por finalidade a interpretação destas mesmas comunicações”.

Para Bardin (1979), a análise de conteúdo se constitui de várias técnicas em que se busca descrever o conteúdo emitido no processo de comunicação, seja ele por meio de falas ou de textos. É composta por procedimentos sistemáticos que proporcionam o levantamento de indicadores, permitindo a realização de inferência de conhecimentos. Sendo um conjunto de técnicas de análise das comunicações, visando a obter, por procedimentos sistemáticos e objetivos de descrição do conteúdo das mensagens, indicadores (quantitativos ou não) que permitam a inferência de conhecimentos relativos às condições de produção/recepção (variáveis inferidas) destas mensagens (BARDIN, 1979).

Além disso, a autora supracitada faz a menção de três fases em que a análise de conteúdo se desenvolve, sendo: pré-análise, que diz respeito à fase de organização, isto é, escolhas dos métodos, formulação das hipóteses e preparação do material para análise; exploração do material, que se refere ao gerenciamento das decisões tomadas na pré-análise, e é uma fase mais longa; e tratamento dos dados, que tem por finalidade tornar os dados válidos e significativos (BARDIN, 1979; GIL, 2010).

Uma vez definida como pesquisa descritiva e técnica de análise dos dados como análise de conteúdo, descrita na Introdução, leva-se em conta a

seguinte questão: **quais os métodos a serem aplicados no desenvolvimento da Pesquisa no contexto da metodologia?**

Creswell (2010) afirma que é vital para o desenvolvimento de uma pesquisa o processo de escolha do método a ser utilizado, o qual deve estar de acordo com as características da pesquisa, e como poderão ser escolhidas diferentes modalidades de pesquisa, sendo possível aliar a sua abordagem.

A pesquisa, ou revisão bibliográfica permitirá, no contexto da pesquisa, conhecer o que já se estudou sobre o mapeamento de dados e sua transformação em dados semânticos, ressaltando a ocorrência de enriquecimento semântico de dados e assim desenvolver o referencial teórico-metodológico.

Tudo o que for explorado e investigado necessita de embasamento, argumentos, e não meramente suposições, para que assim a pesquisa seja fundamentada e tenha qualidade em suas informações.

Um dos métodos escolhidos para a pesquisa é a revisão bibliográfica, considerada indispensável para a elaboração do objeto de estudo. Por meio dessa revisão, é possível explicar e justificar os assuntos pesquisados do mapeamento semântico de dados, até a construção do *Framework* que direciona os passos para que apresente a possibilidade de um catálogo de dados enriquecidos que venham a ser publicados de uma forma para consumo.

Gil (2010) apresenta uma explanação da revisão bibliográfica em que a revisão ou levantamento bibliográfico é um importante estágio na elaboração do quadro inicial da pesquisa. Se o pesquisador utiliza teorias e conceitos para estudar fenômenos, a leitura é um hábito que deve ser cultivado. Pela leitura, o pesquisador fica conhecendo o que outros pesquisadores e autores disseram a respeito do fenômeno que pretende estudar. Além de comentar resumidamente as ideias apresentadas, é possível destacar o que o próprio autor diz sobre a obra ao apresentá-la. Pode, também, escrever e destacar trechos para serem usados em citações como fundamentação teórica, e apresentando de forma.

Gil (2010) ainda aponta que a pesquisa bibliográfica é um trabalho de natureza exploratória, que assegura bases teóricas ao pesquisador, auxilia no processo reflexivo e crítico sobre o objeto de estudo. Com um conjunto de conhecimento, o pesquisador faz o levantamento da base teórica para o

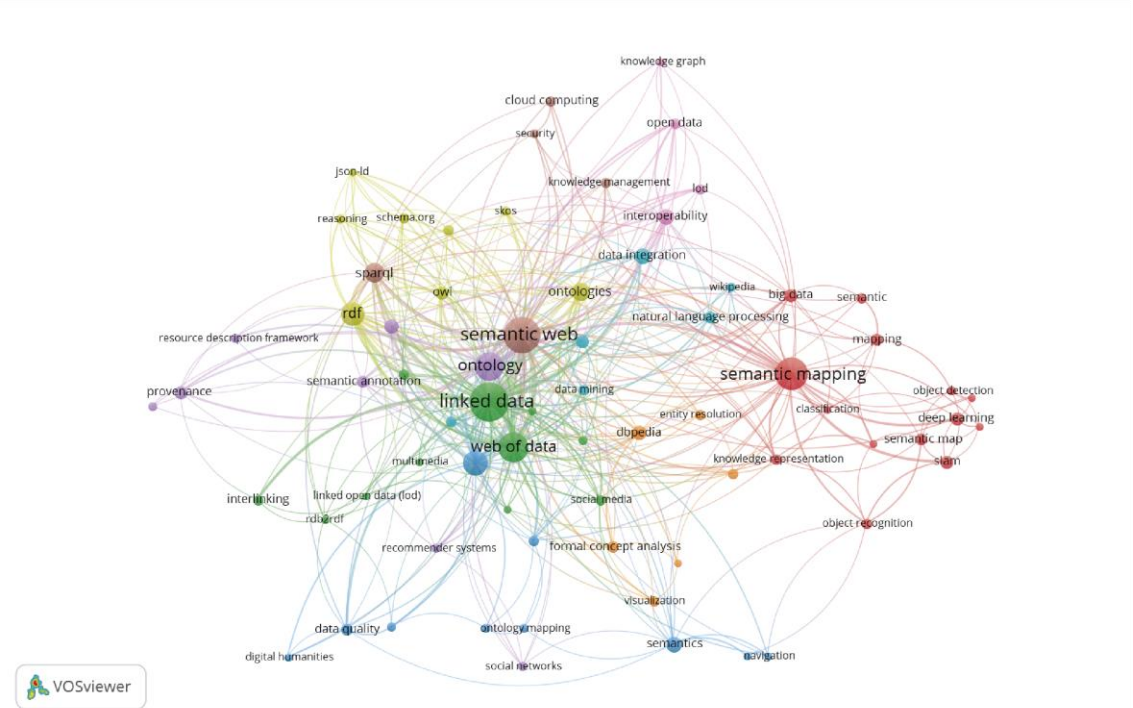
estudo, e se familiariza com o conteúdo, sendo seletivo no processo de leitura, analítico e interpretativo em relação ao conteúdo buscado, como livros, artigos, reportagens, textos da Internet, filmes, imagens e sons, além de ser fiel às fontes utilizadas.

Ainda, a pesquisa adota como métodos a pesquisa bibliográfica e a análise de conteúdo. A pesquisa bibliográfica permite conhecer o que já se estudou sobre o contexto de mapeamento e enriquecimento de dados semânticos e desenvolver o referencial teórico-metodológico.

Além disso, a partir dos resultados da WoS e da Scopus, os metadados das referências encontradas foram exportados em arquivo texto para que fosse realizada uma análise de coocorrência de palavras-chave, por meio do *software* VOSViewer, o que certamente contribuiu para o posicionamento do trabalho na área de CI.

De posse da análise de ocorrência das palavras-chave obtidas por meio da busca realizada na base *Scopus* (**Figura 18**) é possível ilustrar os que apresentam uma maior aderência, considerando somente aqueles termos que ocorreram em, pelo menos, 5 (cinco) trabalhos distintos:

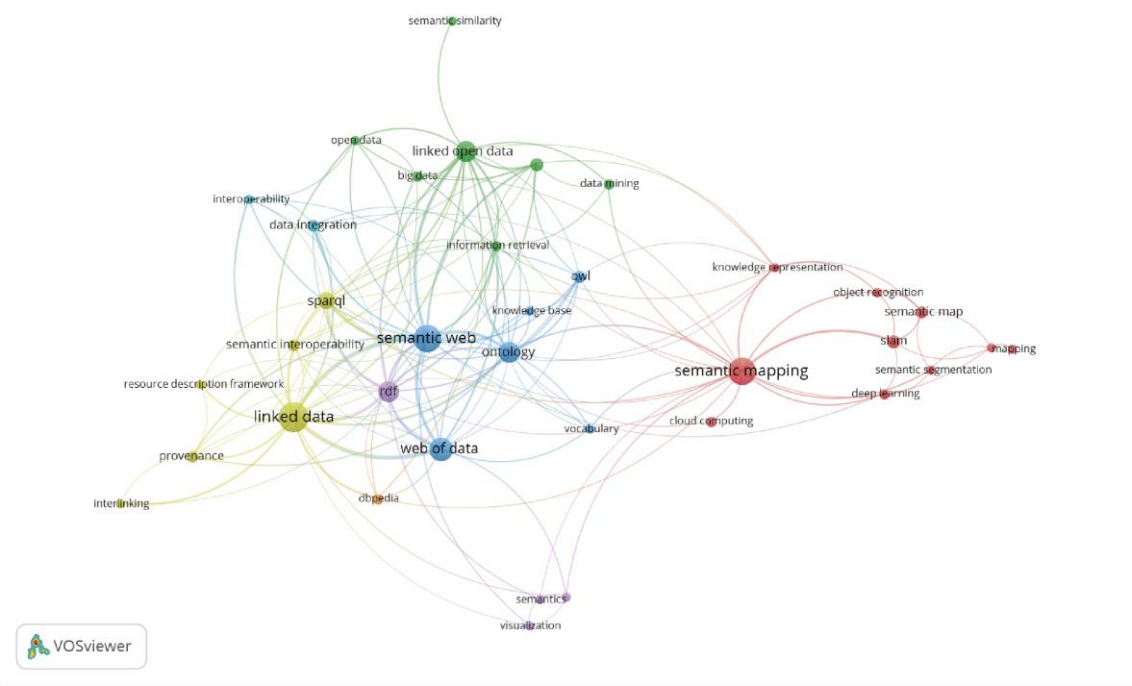
Figura 18- Imagem geradas com as palavras-chave na base da *Scopus*



Fonte: Elaborado pela autora (2019)

Na **Figura 19**, apresenta-se a imagem gráfica gerada na análise de coocorrência de palavras-chave na *Web of Science*, assim como na **Figura 18** consideraram-se somente os termos que ocorreram em, pelo menos, 5 (cinco) trabalhos distintos.

Figura 19 - Imagem geradas com as palavras-chave na base da *Web of Science*



Fonte: Elaborado pela autora (2019)

Por conseguinte, a literatura coletada na pesquisa bibliográfica passou por uma análise de conteúdo (BARDIN, 1979), com algumas categorias analíticas definidas a *priori:* mapeamento semântico, *Web* Semântico, Dados abertos e dados abertos conectados, mapeamento conceitual e *Framework* conceitual.

O Mapa Conceitual tem, por objetivo, ajudar na organização do conhecimento, podendo ser utilizado para várias finalidades em diversas áreas como: ciência, educação, filosofia, entre outros. Para melhor compreensão, há algumas definições que ajudam no entendimento dele.

Souza e Boruchovitch (2010, p. 197) definem mapa conceitual como “[...] representações gráficas de uma estrutura de conhecimento demonstrada hierarquicamente, apresentando forma e representação condizentes com a maneira como os conceitos são relacionados, diferenciados e organizados”.

Em outra definição, Gaines e Shaw (1995) descrevem mapa conceitual da seguinte maneira: uma forma de diagrama especificamente direcionada para fornecer uma linguagem visual parecida com as características da linguagem natural do texto, no sentido de que eles possam estar sujeitos às limitações sintática e semântica, e sua capacidade de representação pode variar de “muito informal” a “extremamente formal”.

Assim, percebe-se que o mapa conceitual possibilita as relações entre conceitos ou entre palavras. Além disso, colabora para a estruturação do trabalho, daquilo que se deseja representar, e permite a compreensão das relações e conceitos.

Os Mapas conceituais são classificados em quatro tipos, sendo eles: Estrutura de Teia, ou seja, o tema principal é colocado no meio do mapa; Estrutura Hierárquica, que se refere ao modo como as informações são dispostas, e nesse caso, a mais importante é colocada no início; Estrutura *Flowchart* que diz respeito ao formato linear em que as informações são organizadas e Estrutura Conceitual, isto é, as informações são classificadas de modo parecido com um fluxograma (LIMA, 2004).

O Mapa conceitual, no contexto da pesquisa apresentada, permitiu uma melhor visibilidade das ligações dos dados e o processo da construção e os passos designados no *Framework* para que esse dado ou esses dados estejam disponíveis de forma enriquecida e pronta para uso, ou seja, para concatenar os conceitos e as técnicas no processo de mapeamento de dados abertos semânticos e enriquecidos, no contexto da publicação de dados abertos na Web, apresentando o modelo do *Framework* em formato de grafo.

3.2 *FRAMEWORK*: DEFINIÇÕES E CONTEXTO DE USO NA TESE

A visão mais apurada referente ao que é um *Framework conceitual* e como ele atua no âmbito da CI é de extrema importância nesse contexto. Suas definições e classificações, baseadas em autores da área da CI e da CC, delimitam o objetivo do estudo e possibilitam uma maior abrangência acadêmica e uma visão ampla de como o processo de mapeamento semântico pode ocorrer.

Para cunhar a pesquisa, o termo *Framework* conceitual estará relacionado com as diretrizes de duas áreas específicas: Ciência da Informação (CI) e Ciência da Computação (CC), uma vez que o intuito da construção do *Framework* remete-se a técnicas dessas áreas e ainda o principal conceito será encaminhado para área da Ciência da Informação.

A definição dos autores Foote e Jhonson (1988), que deriva da combinação das palavras *frame*(estrutura) e *work* (trabalho), determinam que *Framework* é o conjunto de categorias que faz parte de um projeto que não é concreto para solucionar problemas de uma mesma parentela. Algumas das definições são apresentadas no texto de forma cronológica, delimitando as especificações de cada respectiva área.

Gamma *et al.* (1995) utilizam como base a definição que *Framework* é um conjunto de peças, que será tratado no trabalho como objetos, com o objetivo de dar diretrizes e encaminhamento a um conjunto de tarefas que possa trazer resoluções em um âmbito específico ou em um domínio de aplicação, além de ainda afirmar permite o reuso desses objetos em projetos para recorrentes problemas.

Esse conjunto de objetos, que pode ser visto como classes, já que tem uma relação mais computacional, segundo Taligent White Paper (1994) e Taligent White Paper (1995) que definem como o conjunto de classes de *softwares* correlacionadas que possuem a tarefa de gerar soluções de questões que se encontram em um mesmo domínio, permitindo seu uso, sua amplitude e ainda podendo ser personalizada conforme o profissional da área da computação sentir necessidade.

Pinto (2000) afirma que a reutilização de projetos de soluções para problemas recorrentes, denotado por Gamma *et al.* (2000), faz com que o *Framework* acoplado a tecnologias gere família de produtos a partir de uma única estrutura que captura conceitos mais gerais da família de aplicação” (GAMA *et al.,* 2000)

Shehabuddeen, Probert e Phaal (2000) também usam o termo família para denotar definições sobre *Framework*. Os autores supracitados classificam ainda como um conjunto de possíveis hipóteses que formam a família de objetos como base para uma determinada ação, verificando-se, assim, um

processo que auxilia a tomada de decisões das resoluções de problemas, pois a estrutura do *Framework* possibilita a comparação entre diferentes situações.

Dentro das diversas definições apresentadas, o intento de um *Framework* conceitual é gerar soluções para uma família de problemas de um determinado domínio, ou melhor, no âmbito de entendimento que esse problema se encontra. Podendo gerar um modelo de dados que seja compreensível e então aplicado a um modelo de *Framework* de *software* que transformará em codificação, utilizando linguagens de programação específicas.

Neste trabalho definiu-se ***Framework*** como: o levantamento de requisitos necessários para geração de diretrizes, o referencial teórico. São apresentados os objetivos definidos para apresentar o levantamento de requisitos necessários para geração de diretrizes, o referencial teórico considerando os conceitos, processos e técnicas referentes ao mapeamento e enriquecimento semântico de dados conectados baseados nas Boas Práticas de Publicação de dados na *Web*, que se pretende utilizar na proposição de um *Framework*.

Verifica-se que todos os processos apresentados no capítulo 2 foram descritos de forma fundamentada, uma vez que a proposta do trabalho é relatar tudo o que foi proposto na introdução.

Caso a proposta estivesse relacionada ao desenvolvimento computacional de um *Framework*, o foco seria se basear no trabalho que trata *Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling*, apresentados por Schaible, Gottron e Scherp (2014), uma vez que, computacionalmente, a pesquisa dos autores supracitados gera uma visão mais geral e explícita do tópico da modelagem e reuso dos dados.

Porém o foco do trabalho não é desenvolver um *Framework* computacional mais um modelo que apresente o tratamento dos processos que cheguem no ponto de enriquecer semanticamente o dado e permitir sua manipulação no contexto de máquinas e humanos, principalmenteno se diz respeito ao reuso dos dados de vocabulários já existentes.

4. REVISÃO DA LITERATURA

4.1 *WEB* SEMÂNTICA

Ao se falar tanto da *Web* quanto da *Web* Semântica é indispensável se ressaltar a Internet. A propósito, é impressionante notar como ambas se complementam, pois seria difícil compartilhar informações sem uma ferramenta que facilitasse esse processo, do mesmo modo que, se não houvesse a Internet não haveria possibilidade de compartilhamento e outros serviços. Por isso, é importante destacar cada uma delas.

Segundo Dizard Jr. (2000, p. 24), Internet é “um sistema de redes de computadores interconectados de proporções mundiais, atingindo mais de 150 países e reunindo cerca de 300 milhões de computadores”. Isso quer dizer que, observando o próprio nome “Internet”, torna-se autoexplicativo, pois *Inter* de conexão, relacionamento e *net*: rede.

A Internet teve seu surgimento em meados da década de 60 e, ao longo dos anos, foi se desenvolvendo e melhorando para que os objetivos pretendidos se tornassem possíveis, isto é, quando ela surgiu, a finalidade era ter uma rede sem controle algum, a partir da qual seriam transmitidas mensagens divididas em “pacotes”. Depois, em 1969, houve uma junção com ARPAnet (*Advanced Research Projects Agency Network* - Rede de Agências Para Projetos de Pesquisas Avançadas) que no princípio teria quatro computadores interligados e posteriormente mais computadores.

Nos anos 80, houve o desenvolvimento e utilização do TCP/IP (*Transmission Control Protocol/Internet Protocol* - Protocolo de Controle de Transmissão/Protocolo da Internet) e em 1990 a ARPAnet foi transformada em NSFnet (*National Science Foundation’s Network* - Fundação Nacional de Ciência Rede) ligando-se a outras redes existentes (MONTEIRO, 2001, p. 27-28).

A *Web* surge na década de 1990, desenvolvida por Tim Berners-Lee com o intuito de facilitar e colaborar o compartilhamento de dados. Por isso, juntamente com a *Web*, veio uma Linguagem de Marcação de Hipertexto (HTML) também elaborada por Tim Berners-Lee para ajudar o usuário ao acesso de informações (MONTEIRO, 2001, p. 29).

Para esclarecer melhor o conceito *Web*, segundo Laufer (2015, p.10) e desenvolvido por vários parceiros como: nic.br, ceweb.br e cgi.br, possui a seguinte explicação: “A *Web* é um local onde as pessoas consultam informações, fazem compras, trabalham, conversam, estabelecem relacionamentos etc.”. Isto é, a *Web* foi criada com o propósito de ser uma forma de navegação.

Web de Documentos, *Web* Programável e *Web* de Dados são aprimoramentos elaborados desde o surgimento da *Web*, acompanhando a evolução da Internet, ao passo que cada uma delas teve uma utilidade cumprindo suas funções em determinados períodos.

Primeiramente, a *Web* de Documentos foi criada com base no conceito cliente-servidor. É um sistema composto por três elementos:

- URL (*Uniform Resource Locator*) - identificador dos documentos, dos nós da estrutura.
- HTML (*Hypertext Markup Language*) - linguagem de marcação para a descrição dos documentos.
- HTTP (*Hypertext Transfer Protocol*) - protocolo de comunicação para acesso aos documentos.

Outros elementos importantes são os *links* que fazem a interligação entre outros nós. Assim sendo, cada documento da *Web* pode ser acessado por meio de uma URL, podendo estar interligado com diversos documentos por meio de um *link*.

A *Web* Programável tem esse nome, pois evolui de um ambiente com documentos estáticos para diversos tipos de aplicações. Afinal, é possível realizar compras, procedimentos bancários, enviar mensagens por meio da utilização dos navegadores e o espaço programável surgiu para facilitar a execução desses processos como por exemplo: NET, JSPJava (*Java Server Pages Java*), PHP (*Program Hypertext Preprocessor*), Perl, Python, entre outros.

A estrutura dessa *Web* Programável é dividida em três camadas: apresentação, lógica de negócios e armazenamento. A primeira camada diz respeito a um navegador *Web*, a segunda camada refere-se à utilização de

alguma tecnologia (PHP, Python), e a terceira camada é o uso de um banco de dados.

A *Web* de Dados é algo mais amplo e complexo que tem sido estudado, estruturado e aprimorado para tornar cada vez mais fácil a localização de informações e disponibilizar todos os dados que são conectados em um só ambiente, e assim promover uma navegação mais eficaz (CALEGARI, 2016).

Ainda de acordo com Laufer (2015, p. 14), a respeito da *Web* de Dados tem-se a seguinte explicação:

> Páginas da Web, exibidas por um navegador, contém um conjunto de informações que são consumidas por pessoas. São textos, fotos, vídeos etc., dispostos na página, de forma que uma pessoa possa extrair um significado dessas informações. Essas informações agrupam, em geral, um conjunto de dados que têm algum relacionamento entre si e que, por algum motivo, faz sentido apresentá-los em uma única página, um documento único.

Observa-se a magnitude de *Web* de Dados pelo fato de tornar cada nó, ou seja, cada interligação não apenas em um documento, mas um documento de dado específico, um recurso, permitindo dessa maneira acesso mais eficiente e possibilitando que desenvolvedores possam criar aplicações para agrupar esses dados das mais variadas formas.

Uma vez que todos os dados e informações são entendidos por usuários, como fazer uma máquina compreender para realizar os procedimentos necessários? É nesse contexto que aparece a *Web* Semântica. Uma extensão da *Web* atual que visa tornar a informação entendida para computadores ajudando no processamento dos dados (Berners-Lee, Hendler e Lassila, 2001). Ou seja, a *Web* Semântica trouxe consigo a organização e a estruturação de informações para que seres humanos e máquinas se entendam.

De acordo com W3C (2020, *online,* não paginado, tradução nossa), a *Web* Semântica pode ser definida da seguinte maneira:

> O termo "Web Semântica" refere-se à visão do W3C da Web de dados vinculados. As tecnologias da Web semântica permitem que as pessoas criem armazenamentos de dados na Web, construam vocabulários e escrevam regras para lidar

> com dados. Os dados vinculados são capacitados por tecnologias como RDF, SPARQL, OWL e SKOS[6].

Nada mais é do que afirmar que a *Web* Semântica é a *Web* de Dados - de datas, títulos; números, propriedades químicas e qualquer outro dado que se possa conceber -, dando às pessoas a capacidade de criar repositórios de dados na *Web*, construir vocabulários e escrever regras para interoperar com esses dados.

Nota-se que a *Web* Semântica é ampla e subdivide-se em outros aspectos paralelos, como vocabulários, ontologias, dados abertos e dados abertos conectados que serão abordados nas seções a seguir. Embora possuam funções diferentes, objetivam a facilidade e agilidade no processo da informação e, desta forma, possibilitam o suporte de interação na rede.

4.1.1 A Semântica no Contexto da *Web*

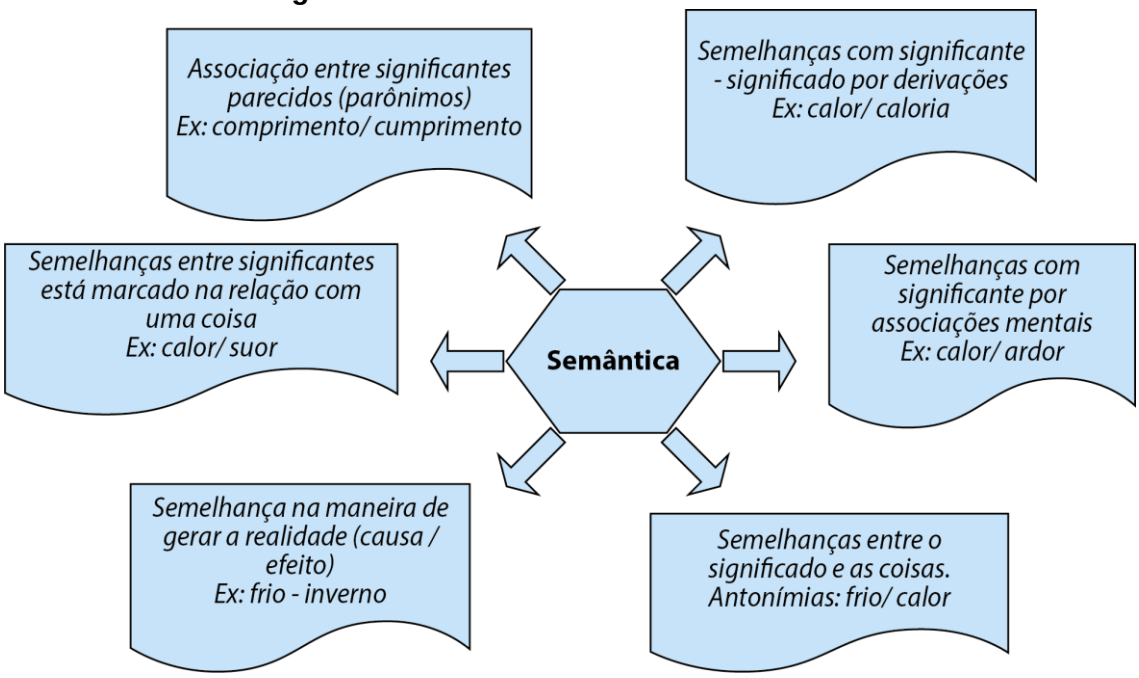
Linguisticamente, o campo semântico pode ser classificado como reconhecimento ou mesmo semelhança de significados que uma palavra possui. Sinonímia (palavras que têm significação semelhante), polissemia (palavras que têm significação aproximada), antonímia (palavras que têm significação oposta), hipônimos e hiperônimos, que podem gerar contextos de sentidos mais especializados ou mais genéricos, sintático, semântico e léxico em conteúdos que geram concepções da realidade e suas diferentes expressões linguísticas de acordo com o contexto abordado, como apresentam Pietroforte e Lopes (2004).

O significado de uma sentença não é dado apenas pela soma do conteúdo de suas partes, mas pela soma do conteúdo de suas partes acrescido à estrutura que a carrega, como apresenta a **Figura 20**:

---

[6] *"In addition to the classic "Web of documents" W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS."* (W3C, 2020, não paginado).

Figura 20 - Inferências do Termo Semântico



Fonte: adaptado pela autora de LEÓN, 2016, p. 81

Neste contexto, para atribuir esses conjuntos de informações na *Web* Semântica, verifica-se a necessidade de se criar mecanismos e ferramentas que possam reconhecer a informação digital existente na *Web*.

Seu registro de forma heterogênea, ou seja, em diferentes formatos, as diferentes variantes e a riqueza da linguagem natural, permitem assim a organização e recuperação de determinada informação, tomando em consideração a comunicação do seu significado e não somente a estrutura sintática dos termos que as identificam.

Na história da evolução da *Web*, verificam-se mudanças consideráveis, principalmente no que tange ao contexto de permitir resolver informações perdidas, ou até mesmo permitir a troca de informação em campos diversos.

Castells (2002) apresenta um conjunto de datas e acontecimentos (**Quadro 5**) relativos à evolução da *Web*:

Quadro 5 - Datas e acontecimentos relacionados a *Web*

Data	Acontecimento
1989	Tim Berners Lee apresenta o projeto *World Wide Web* para o CERN[7]
1993	Criação dos primeiros servidores *Web* e o navegador Mosaic

[7] CERN - *Organisation Européenne pour la Recherche Nucléaire* (CERN, 2020).

1994	Criação da W3C (*World Wide Web Consortion*)
1997	Criação da SHOE (Simple HTML Ontology Extensions)

Fonte: Elaborado pela autora com base em CASTELLS (2002)

A criação da SHOE gerou os primeiros passos da *Web* Semântica baseando-se em linguagem de marcação, HTML (*Hypertext Markup Language*), gerando as primeiras páginas editadas manualmente, porém já ligadas umas às outras, determinando o contexto das primeiras versões da *Web*, classificadas como *Web* 1.0, criados por Berners Lee.

Fato este que permitiu a possibilidade de conectividade entre as máquinas e o crescimento exponencial da quantidade de informação oferecida, mesmo as páginas sendo ainda estáticas, permitindo uma recuperação da informação apenas para leitura informativa e toda desenvolvida em HTML.

Spivack (2007) afirma que na *Web* 1.0 (1990 a 2000) foram apresentadas interfaces de navegadores mais agradáveis como Internet Explorer (IE), Netscape entre outros e a exploração de base de dados e a criação de linguagens como SQL (*Structured Query Language*), para construção de softwares colaborativos que viessem a auxiliar pessoas em tarefas comuns, além de prover interfaces mais amigáveis e compartilhadas para este uso, determinados como "*groupware".* Porém, nota-se que a criação da *World Wide Web* foi o fato mais importante, gerando assim a transição para *Web* 2.0.

Evans (2011) apresenta que o termo *Web* 2.0 trouxe avanços significativos na tecnologia *Web*. Termo criado por O'Reilly em 2004 permitiu uma mudança radical nos processos de produção e troca de informação e colaboração entre usuários, uma vez que deixou de ser um conteúdo estático e deu lugar a uma *web* dinâmica, onde os usuários, que eram limitados à observação passiva das informações geradas por eles, conseguiram, a partir daquele momento, fornecer seu próprio conteúdo de maneira simples.

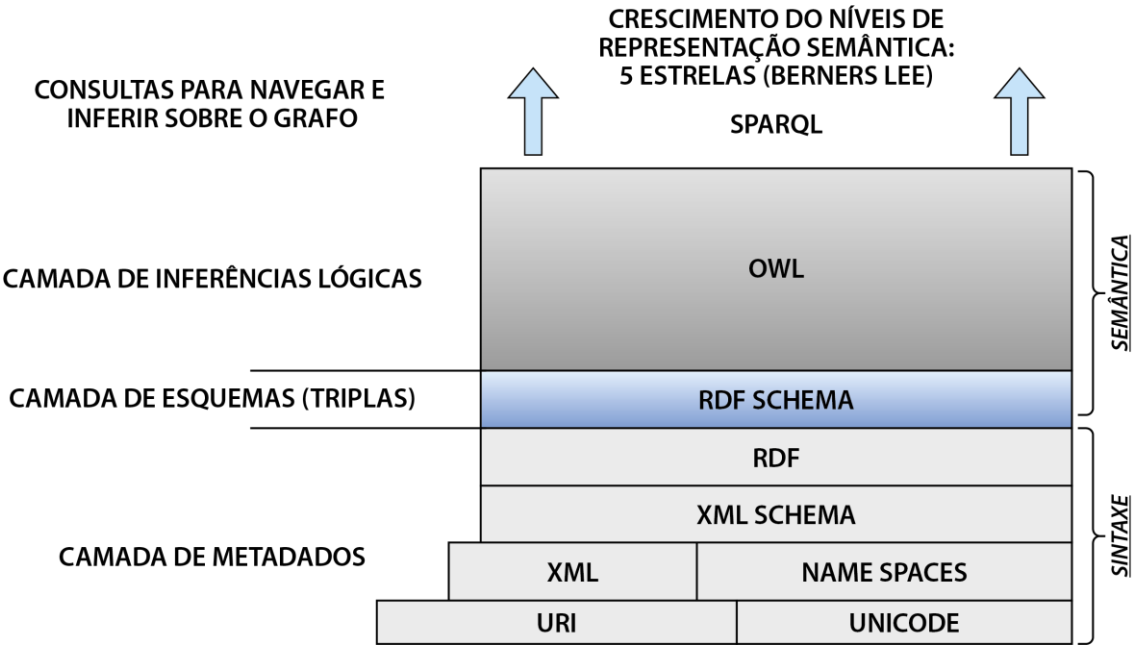
Aparecem, aqui, o surgimento de redes sociais (Facebook e Twitter), serviços *Web*, serviços de rede social, as Wikis, blogs, e a indexação de informação, cunhada como folksonomia, gerando assim comunidade virtuais com usuários que apresentam conteúdo publicado, transformando-os em contribuidores de informações publicadas e que realizam trocas de dados.

Cesca (2018) mostra que, com esse processo de troca de informações, número de usuários, valor do *site* e seus conteúdos, ecossistemas informativos começaram a surgir. Google, Wikipedia, Ebay, Youtube, Skype, Writely, Blogger são alguns exemplos desses ecossistemas, que marcaram a entrada de recursos como palavras chaves para a recuperação da informação e motores de busca utilizando algoritmos mais complexos, o que muda assim a estrutura dos códigos até aqui existentes.

Nesse processo de uma nova arquitetura criou-se a necessidade da semantização da *Web*, que é classificada como a *Web* 3.0, sob o conceito de *Web* semântica entendida e assumida, e apresentada em camadas.

A **Figura 21** apresenta a visão da *Web* Semântica em camadas e o que elas representam.

Figura 21 - Estrutura da *Web* Semântica



Fonte: Elaborado pela autora com base em BERNERS-LEE; HENDLER; LASSILA (2001) e BREITMAN (2005)

A arquitetura da *Web* Semântica é definida como ilustrada na **Figura 21** (BERNERS-LEE; HENDLER; LASSILLA 2001; BREITMAN, 2005). Diversos autores como Berners-Lee, Hendler e Lassila (2001), Pastor-Sánchez (2012), Breitman (2005) apresentam a mesma contextualização. Visualiza-se que,

nesta arquitetura, cada camada estende a funcionalidade e expressividade de suas camadas inferiores.

A camada base refere-se à representação padrão para uso de identificadores únicos para nomeação dos recursos informacionais na *Web* de forma única na *Web*, como os URIs que se caracterizam na codificação de caracteres em nível internacional.

As camadas referentes às linguagens XML (W3C, 2015) e RDF (MILLER, 2004) desempenham papéis fundamentais relacionados à sintaxe (DECKER *et al.*, 2000). Sua estruturação de documentos, como os *namespaces,* para associar com propriedade o esquema que define cada propriedade, além do XML Schema que define elementos que devem conter em um documento XML, como serão organizados, seus atributos e de que tipo cada elemento pode ser composto.

Um modelo básico para estabelecer propriedades sobre os recursos é o RDF, pois ele definirá as relações entre os recursos pelo modelo de seus dados, ou seja, definirá por meio de classes e objetos que se expressam mediante o esquema RDF (RDF Schema). Essas camadas são elementares, porém são necessárias outras camadas de suporte às ontologias e à lógica para fornecer a semântica esperada para a evolução da *Web* atual.

A camada de ontologia define, mediante a descrição detalhada das camadas anteriores, as propriedades e as classes, destacando suas relações, cardinalidades, igualdades, tipologias e propriedades mais complexas, definindo e caracterizando-se como semântica.

A linguagem SPARQL consulta esses conjuntos de dados RDF, incluindo o formato XML que detalha todo o modo em que se estruturaram os resultados obtidos.

Os mecanismos dos *sites* anteriores são recompilados e otimizados por meio de um acesso rápido e fácil à troca de informações e participação em redes sociais, facilitando as atividades dos usuários. Para esse fim e como característica única são trabalhados projetos para atribuir contextos semânticos e ontológicos a WWW, enriquecendo com informações adicionais ao texto primário, para descrever o conteúdo, o significado e assim estabelecer uma relação entre os dados, para que seja possível seu entendimento e

avaliação por máquinas de processamento, através da inteligência Artificial (CESCA, 2018).

Neste contexto, verifica-se que a *Web* 1.0 e a *Web* 2.0 fazem parte de um contexto de uma *Web* sintática, que geraram coleções de dados desestruturados e que se ligavam através de documentos, porém na *Web* 3.0 trouxe a necessidade de codificação das páginas para uma interpretação semântica para analisadores automáticos e não mais somente humanos.

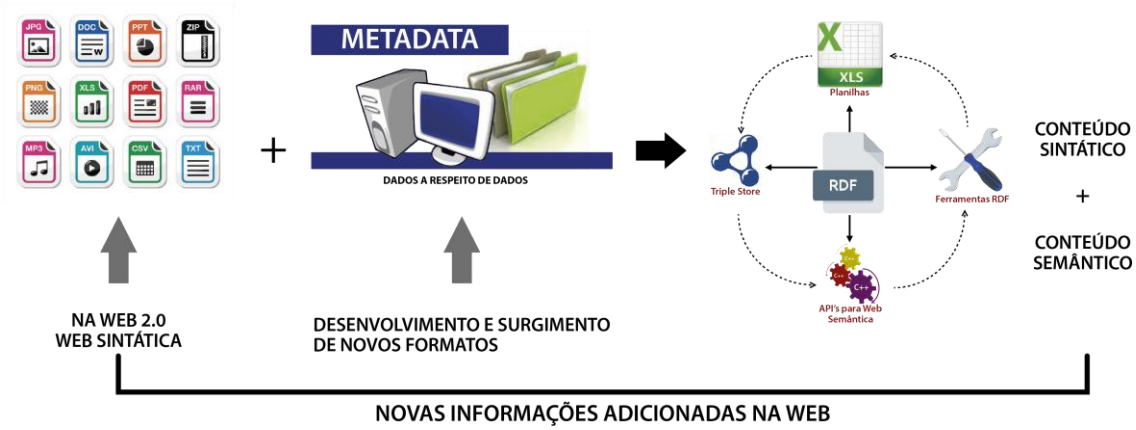
Pastor-Sánchez *et al.* (2013) afirmam que a *Web* atual gera relação entre dados, de forma semântica e por atributos que as determinam. Mas ainda afirma que, para alcançar uma *Web* Semântica na sua totalidade, os dados e as páginas devem ser estruturados e escritos com um formato que permite que computadores e máquinas selecionem melhor as informações existentes na Internet.

A *Web* Semântica pode ser vista como ideologia criada para a troca de dados, cujas informações são lidas e interpretadas entre máquinas, facilitando a comunicação entre o ser humano e um sistema de informação.

Verifica-se que essa estrutura da *Web* Semântica conta com um conjunto de dados, que é a visualização dos documentos, porém sem a identificação do seu conteúdo.

Com a necessidade de *browsers* localizarem os documentos de forma mais próxima à necessidade do usuário, entra o contexto da *Web* Semântica que, para esses dados, são necessários um trabalho de vinculação a outros dados e novos formatos de representação básica como RDF, que atualmente é o que consiste na base da *Web* Semântica. Nesse contexto, a **Figura 22** apresenta a representação dessas informações em formato de imagem.

Figura 22 - Processo de Mudança da *Web* Atual para *Web* Semântica



Fonte: Elaborado pela autora (2019)

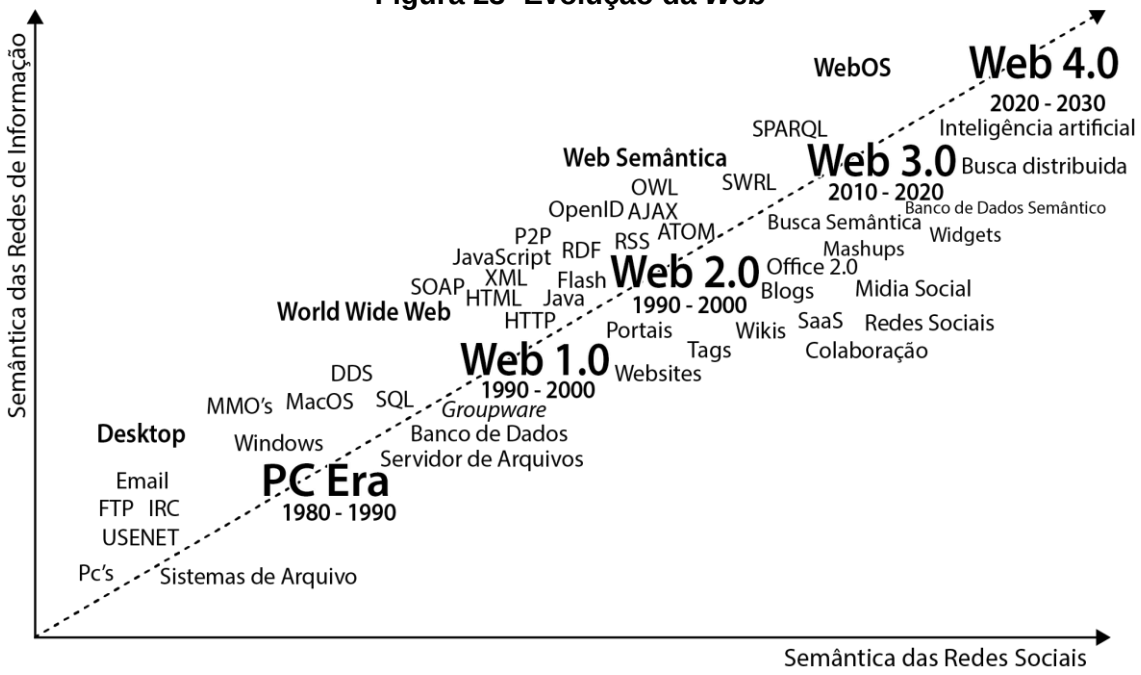
Assim, a *Web* 3.0 segue com todas as diretrizes, administrando um conjunto de dados e metadados e todas as suas partes de maneira holística e semântica que permitem a operação na *Web* automaticamente, podendo descrever sintática e semanticamente os documentos.

O que permite seu processo diferenciador é que, pela necessidade de estruturação dos dados, utilizam-se novas linguagens e padrões como RDF, OWL, SPARQL, RSS (*Rich Site Summary*), sistemas *Open ID*, base de dados e buscas semânticas em pleno processo de desenvolvimento.

A interoperabilidade é um fator preponderante para a melhora da internet, uma vez que os sistemas geram informação na *Web* com agentes inteligentes que se utilizam de dados semânticos e suas descrições lógicas, empregando regras que expressam relações entre conceitos e dados na rede, assim como o aumento da integração de fontes de dados heterogêneos.

A **Figura 23** mostra a evolução em anos dessas *Web*, sob uma linha do tempo. Evolução da *Web* - *Web* 1.0 à *Web* 4.0 (SPIVACK, 2007). Verifica-se que as transformações tecnológicas, o aumento de largura de banda e a grande expansão na quantidade de acessos por parte dos usuários são fatores importantes que contribuem para o crescimento e aplicabilidade dessa quarta geração da *Web* (QUINTÃO *et al.,* 2010).

Figura 23- Evolução da *Web*



Fonte: SPIVACK (2007) *apud* QUINTÃO *et al.* (2010, p. 232)

Nesse cenário, o uso de plataformas que utilizam conceitos da *Web* Semântica com o objetivo de organizar diferentes domínios - apresentando e explicando os relacionamentos existentes - trouxe um grande auxílio para que a comunidade em geral acesse as informações de forma mais relevante.

Para exemplificar, Santarem Segundo, Coneglian e Lucas (2017) apresentam um estudo da plataforma VIVO, apresentando conceitos e tecnologias da *Web* Semântica dentro do contexto acadêmico científico e o quanto à disponibilização das informações são fundamentais para que os dados possam ser posteriormente recuperados.

Foram desenvolvidos testes e análises, utilizando-se um conjunto de publicações científicas de periódicos brasileiros da área da Ciência da Informação e trouxe, como resultado, o uso da *Web* Semântica para um público que visa buscar dados relacionados a níveis de ambientes que constam temas ligados à colaboração científica, estudos métricos e organização das informações.

4.1.2 Triplificação de Dados

O processo de triplificação de dados pode ser apresentado como um processo que gera uma base de dados, seguindo o modelo RDF. Porém não deve atentar apenas para o contexto computacional envolvido, uma vez que, para que a *Web* semântica se torne realidade, Berners-Lee e Fischetti (1999) estabeleceu algumas condições informacionais importantes.

A representação do conhecimento é uma das condições para que a *Web* Semântica possa funcionar. Fazer com que máquinas tenham o acesso a uma determinada coleção ou coleções de informações estruturadas e a conjuntos de regras que inferem a tal ponto que permitam o poder de raciocínio da máquina automaticamente.

Neste ponto, verifica-se a dificuldade de transformar para a máquina esse processo, pois como essa tecnologia é frequente, suas aplicações possuem um potencial, porém é preciso se associar a um sistema global simples para entender todos esses passos (BREITMAN, 2005).

Tem muito mais relação à disponibilidade de dados sobre um tema, ou domínio escolhido. Em que muitas vezes não possui tamanha proporção, sendo que grande parte das bases de dados disponibilizadas estão em formatos que não são de fácil interligação com outras bases de dados existentes.

Neste contexto, existe uma tendência de representar dados de uma forma rica e mais flexível, visando facilitar a interligação de dados, de modo que o dado é representado em uma unidade menor e anotado com metadados para reduzir a sua ambiguidade na *Web*. Esta forma de representar dados é feita a partir da construção de triplas, ou como se trata no texto, triplificação de dados.

Sistemas tradicionais de representação do conhecimento possuem suas próprias regras para fazer inferências a partir dos dados, por isso é classificado como um processo complexo, pois uma vez determinadas as regras, os dados até podem ser transferidos de um sistema para outro, porém as regras cabem apenas para aquele domínio atuante.

Por isso a questão da *Web* Semântica dentro do processo de triplicação cabe muito mais como um contexto de agregação lógica à *Web*, ou seja,

prover uma linguagem que represente tanto os dados como as regras de raciocínio sobre os dados, permitindo assim que as regras criadas para um determinado sistema de representação do conhecimento possam ser exportadas para a *Web* (BREITMAN, 2005).

Os artigos *A Linguagem de Marcas eXtensível (eXtensible Markup Language – XML)* e *Marco para Descrição de Recursos (Resource Description Framework – RDF)* apresentam características que a linguagem XML permite a criação de etiquetas (tags) próprias, ou seja, etiquetas usadas para anotar as páginas *Web* ou partes de texto na página, mas elas não indicam nada sobre seu significado.

Alguns programas que contenham conjunto de instruções para que uma função seja executada em determinado aplicativo são usados de formas mais refinadas, mas, quem os escreve precisa conhecer qual a finalidade de uso das *tags*. A linguagem XML (QUIN, 2015) permite que os usuários acrescentem livremente estruturas aos seus documentos, mas, como já dito anteriormente, não inferem significado.

Essa tarefa se concretiza quando se usa o RDF (W3C, 2014), uma vez que sua função é codificar em grupos de triplas, ou denominadas como ´tríades´ que são formadas por sujeito, verbo e objeto. As tríades podem ser escritas usando tags XML, porém somente em RDF, o documento trará o processo de significado mais assertivo sobre a estrutura de cada um dos três componentes envolvidos, tornando-se um processo mais natural em descrever a grande maioria dos dados processados pelas máquinas.

Cada sujeito e objeto determinados pela tarefa do RDF (BRICKLEY; GUHA, 2004) são identificados por um URI - Identificador Uniforme do Recurso (*Uniform Resource Identifier*) que é usado como um ligador na Página *Web*, denominado como URL - Localizador Uniforme do Recurso (*Uniform Resource Locator*).

Os URLs são classificados como o tipo mais comum de URI, que determinam dentro do contexto das triplas que os verbos são identificados por URIs, permitindo a qualquer um definir um novo conceito, um novo verbo, ou simplesmente definindo sua URI em algum lugar da *Web*.

A linguagem humana se desenvolve usando o mesmo termo com significados diferentes, mas a automatização não (BECKETT, 2004). As tríades

de RDF formam redes de informação sobre coisas relacionadas, e como usam URIs para codificar essa informação em um documento, as URI's garantem que os conceitos não sejam somente palavras no documento, mas que sejam amarradas a uma única definição onde se pode encontrar na *Web* (BECKETT et al., 2014, não paginado).

Verifica-se, então, que a triplificação de dados não é uma tarefa trivial como a maioria das conversões de dados existentes. Ela difere de dados estruturados como base de dados que se encontram em formatos de dados estruturados SQL ou mesmo quando estamos tratando de dados não estruturados como os de formato de extensão JSON (*JavaScript Object Notation)*, pois a função das triplas RDF serve para agregar conhecimento aos dados e, por isso, classifica-se como um processo mais delicado de construí-las.

As próximas seções tratarão de explicações mais técnicas relacionadas a essas linguagens e sua importância no contexto da *Web* Semântica, baseadas no contexto de conhecimento na sua utilização.

4.1.3 Formatos, Padrões e Linguagens

Toda informação apresentada hoje na Internet se encontra em três categorias de apresentação: estruturada, semiestruturada e não estruturada. Isso implica que a *Web* de Documentos evoluiu para uma *Web* Programável e a oferta de aplicações tem sido necessária e crescente, pois executa tarefas que manipulam dados publicados das mais diversas formas nesse ecossistema de dados na *Web*.

Porém, a *Web* ainda se apresenta de forma desordenada e não estruturada, dificultando assim que máquinas e algoritmos consigam entender e aproveitar das informações de forma automática.

Elmasri e Navathe (2005) definem dado como uma ocorrência separada, permitindo sua gravação e gerando um significado implícito, ou seja, que permita um conceito incluído. Elmasri e Navathe Hurwitz *et al.* (2015) trazem como classificação três diferentes tipos de Dados: estruturado, semiestruturado e desestruturado.

Dados estruturados, segundo Dong e Halevy (2007) podem ser classificados como dados organizados de acordo com critérios definidos, respeitando vários atributos, delimitando escopo, tipos, entre outros. Em outras palavras, podem conter uma organização para serem recuperados que possuem relações, classificados em grupos possuindo as mesmas descrições e que são mantidos em um Sistema Gerenciador de Banco de Dados (SGBD) e são analisados através de linguagem própria que é SQL.

Dados semiestruturados, segundo Manuja e Garg (2011), classificam-se como dados que nem sempre são possíveis de prever todos os aspectos, ou seja, alguns atributos gerais podem ser conhecidos, enquanto outros poderão ser adicionados posteriormente.

Segundo Taylor (2018), dados desestruturados classificam-se como aqueles que não possuem estruturas definidas. Normalmente são textos, documentos, gravações de serviços ao consumidor, fotos, imagens e vídeos, e não é possível armazená-los em uma base de dados tradicional, devido as inúmeras informações aleatórias.

Todos os conteúdos hoje são visualizados de forma importante e valiosa, uma vez que, segundo Caldón *et al.* (2010) afirmam que é necessário definir mecanismos para dar estrutura e enriquecer semanticamente a informação existente, para que possa ser aproveitada por outras aplicações de software.

Tanto para a realização de técnicas de processamento de linguagem natural, de anotação semântica e suas variações, ou de qualquer tecnologia semântica requerem padrões, linguagens especializadas para conseguir representar, identificar as estruturas dos documentos, agrupar e relacionar os conceitos de um domínio de conhecimento, facilitando a recuperação da Informação de forma mais simples possível em ambientes tão heterogêneos.

Berners-Lee, Handler e Lassila (2001) já apresentam que a ideia de agregar semântica a esses dados só facilita o entendimento e a interoperabilidade dos dados nesse universo de informações heterogêneas, uma vez que sua publicação pode ocorrer nos mais variados formatos e protocolos de acesso.

Laufer (2015) apresenta no guia da *Web* Semântica os blocos básicos que definem a *Web* Semântica que são: um modelo de dados padrão; um conjunto de vocabulários de referência; um protocolo padrão de consulta.

A seguir serão apresentadas as diferentes tecnologias utilizadas no ambiente da *Web* Semântica, em específico XML, RDF e RDF Schema, Ontologias, como o foco na linguagem OWL e a linguagem de consulta SPARQL, de forma objetiva e suas características em termos de conceitos e o papel de cada uma delas no ecossistema de dados publicados na *Web*.

4.1.3.1 XML

Extensible Markup Language é uma linguagem de marcação de Texto e sua especificação define uma forma padrão de adicionar marcas, ou classificadas como *tags* aos documentos que permitem identificar as estruturas desse documento (QUIN, 2015).

Classifica-se como sendo uma linguagem de formato simples, baseado em texto para representar informação estruturada como: dados, documentos diversos, configuração, transações, faturas entre outros (W3C, 2015). Foi derivado da linguagem SGML (*Standard Generalized Markup Language*) que é um padrão ISO (ISO 8879), que especifica as regras para a criação de linguagens de marcação que independe da plataforma para manter repositórios de documentação estruturada há mais de uma década, mas não é muito adequado para atender estruturas de documentos da *Web*.

Quin (2015) define XML como um perfil de aplicação de SGML, ou seja, significando que qualquer sistema SGML é totalmente hábil e capaz de ler documentos XML, podendo ser definido na visão do autor supracitado, como sendo uma forma restringida de SGML.

Como a *Web* necessita de instruções e expressões semânticas, a necessidade de uma estrutura para proporcionar métodos que não trouxesse nenhum equívoco foi definida.

No seu contexto de criação, baseado nos documentos disponibilizados pela W3C, Quin (2015) afirma que o XML tem características muito parecidas com o HTML (*HyperText Markup Language*), porém sua principal característica é a capacidade de definir marcas, etiquetas ou como trabalhado no texto, *tags*

próprias. Esta característica tem como qualidade a produção de documentos personalizados que, segundo Quin (2015), é de extrema valia quando se tem a necessidade de realizar consultas ou qualquer tarefa em alguns tipos de dados, principalmente no que se diz respeito quando esse conteúdo está na *Web*, gerando assim um *dataset* de cada documento, não criando sobrecarga e muito menos gastos computacionais em outros *datasets* que carregam um projeto *Web*.

Todos os documentos XML podem ser processados de forma confiável por *softwares* de computadores, uma vez que têm uma sintaxe rigorosa. A W3C reconhece que XML é um dos formatos que mais se utiliza para troca de informação estruturada e é considerado como a camada sintática da *Web* Semântica, sendo a base para todas as tecnologias que consistem em todo processo semântico, segundo Daconta, Obrst e Smith (2003).

Sobre a base do XML, definem-se linguagens distintas de marcação para diferentes tipos de documentos, o **Quadro 6** apresenta essas linguagens e suas funcionalidades segundo Solís (2015).

Quadro 6 - Linguagens que usam a base XML e suas características predominantes

Linguagens	Características
SHOE (*Simple HTML Extensions*) LUKE; RAGER; SPECTOR (1996)	Linguagem de representação de conhecimento baseado em HTML (extensão das tags XML e as que se adicionam significado)
OWL (*Ontology Web Language*) SMITH *et al.,* 2004	Habilita o processamento semântico da informação mediante as máquinas
OIL (*Ontology Interchange Language*) HORROCKS *et al.,* 2000	Linguagem de troca de ontologias. A união de DAML (*DARPA Agent Markup Language*) +OIL pode derivar a forma mais prática para que os agentes inteligentes compreendam o conteúdo das páginas e, para isso, desenvolvem a *Web* Semântica
XOL (*Ontology eXchange Language*) CHAUDRI; KARP; THOMERE (1999)	é uma linguagem de trocas de ontologias. usa-se como linguagem intermediária. Manipula dados, define classes e objetos, base de fatos. Possui uma sintaxe baseada em XML e uma semântica que simula o modelo de conhecimento.
OKBC (*Open Knowledge Base Connectivity*) CHAUDRI *et al.* (1998)	caracterizado como base expressiva, legível, compreensível e processável por agentes humanos e máquinas
IFF (*Information Flow Framework*)	é uma linguagem de tags associadas à informação. Possui ideias do XML agregando RDF, OIL, OML/CKML

KENT (2000), KENT (2011), KENT (2009)	(*Conceptual Knowledge Framework*)
RDF (*Resource Description Framework*) FENSEL (2001)	Recomendação da W3C, baseado em XML, que proporciona a tecnololgia para descrever metadados que descrevem recursos na *Web*.
WSML (*Web Service Modeling Language*) BRUJIN; LAUSEN (2005)	é uma linguagem para a modelagem de serviços *Web* Semânticos baseados em WSMO (*Web Service Modeling Language*). A linguagem conta de uma série de linguagens e acessórios que definem determinados aspectos referentes à semântica de serviço *Web*.

Fonte: Elaborado pela autora (2020)

Quin (2015) apresenta algumas vantagens do XML que são redundância, autodescrição e efeito de rede e a promessa XML.

A redundância se insere no contexto em que a marcação XML é muito detalhada. Quin (2015) exemplifica que todas as *tags* finais devem ser fornecidas, permitindo que o computador encontre erros comuns, como alinhamento incorreto.

Em relação à autodescrição, Quin (2015) afirma que é relacionado com a legibilidade do XML (é um formato baseado em texto) e a presença de nomes de elementos e atributos no XML significa que as pessoas que olham para um documento XML geralmente conseguem avançar no entendimento do formato (e ajudam as pessoas a encontrar erros).

Quanto ao efeito rede e à promessa XML, Quin (2015) apresenta que qualquer documento XML pode ser lido e processado por qualquer ferramenta XML. Obviamente, algumas ferramentas XML podem querer uma marcação XML específica, mas o próprio formato XML pode ser lido por qualquer analisador XML

Quin (2015) afirma enfaticamente que o XML é o formato mais usado em todo o mundo, significando que todo novo documento XML aumenta o valor de qualquer outro documento XML e de toda ferramenta XML; que por consequência toda nova ferramenta XML aumenta o valor de todo documento XML e, portanto, de qualquer outra ferramenta.

4.1.3.2 RDF *Resource Description Framework* e RDF *Schema*

A crescente disponibilização de dados abertos tem se tornado, nos últimos tempos, uma tarefa de tendência global e ainda em crescimento exponencial. Neste contexto, o presente trabalho tem como principal objetivo apresentar o mapeamento e enriquecimento semântico de dados baseados nos padrões da W3C (W3C, 2014), utilizando as linguagens *Resource Description Framework* (RDF).

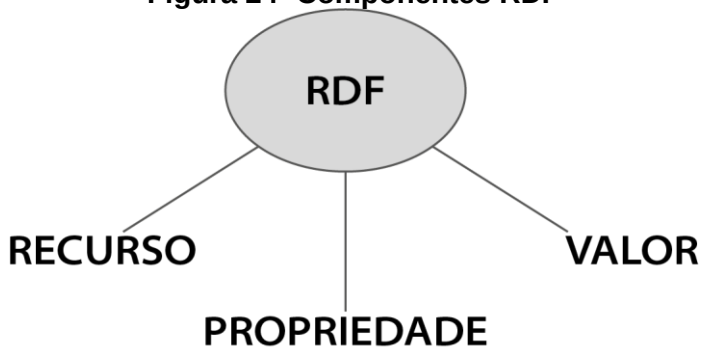
RDF, linguagem que surgiu em agosto de 1997, determinada pela W3C (*World Consortium Web*) e se tornou uma recomendação deste consórcio em 2004, é um modelo de metadados para publicação, ligação, troca de dados e descrição de recursos na *Web* (W3C, 2014).

É uma linguagem de marcação e metadados, que possui uma arquitetura genérica de metainformação baseada em XML (*eXtensible Markup Language*) e está presente em todos os componentes que se adequem à estrutura *Web* (navegadores, servidores, entre outros) a fim de proporcionar um marco padrão para a construção de interoperabilidade na descrição de conteúdo da *Web* sem precisar da intervenção humana (SENSO, 2003).

O RDF possui uma estrutura de triplas determinada por sujeito, predicado e objeto e foi desenvolvido para ser lido e entendido por máquinas.

A **Figura 24** apresenta a estrutura da informação em forma de triplas (YU, 2011).

Figura 24- Componentes RDF



Fonte: Elaborado pela autora com base em W3C (2014).

Como se apresenta na **Figura 24**, a W3C (2014) e os componentes do RDF possuem as seguintes características:

- Recurso - Todas as coisas estão descritas por expressões RDF e se denominam recursos, quer dizer, qualquer recurso entendido como objeto de informação na *Web*, identificado exclusivamente por um identificador uniforme de recursos (URI - *Uniform Resource Identifier*), que podem ser documentos, parte de XML dentro de um documento fonte, *Web site* completo, coleção de páginas, entre outros;
- Propriedade: são aspectos específicos, características, propriedades ou relacionamentos utilizados para descrever recursos. Cada tipo de propriedade tem seus valores específicos, define os valores que lhe são permitidos, os tipos de recursos que podem descrever e as relações que existem entre as distintas propriedades. Determina que tipo de relação se mantém entre o sujeito e objeto;
- Valor: são os dados em que a propriedade de um recurso específico é especificada, alternativamente, pode ser um valor literal como uma sequência ou um número.

Esses componentes, descritos, constrõem declarações, expressões e descrições que nada mais são do que um conjunto de um recurso, um nome de propriedade e o valor dessa propriedade - assunto, predicado e objeto, respectivamente.

Verifica-se que a estrutura RDF se difere do HTML, pois sua estrutura, por ser criada e compreendida por máquina, permite a troca de dados, podendo utilizar uma variedade de domínios para gerar informação, pois é uma infraestrutura que possibilita a restrição, a reutilização e a troca de metadados estruturados.

Possui características que facilitam a fusão de dados, mesmo que todos os esquemas que o compõem sejam diferentes e se apoiem especificamente na evolução dos esquemas ao longo do tempo, sem a necessidade de todos os consumidores de dados mudarem (W3C, 2014).

Velegrakis (2010) e Heflin (2007) afirmam que, embora o RDF seja apenas um modelo de dados e por si só não tem implicações semânticas, é compreendido, por sua solidez e grau de implementação, como um instrumento para representar muitas quantidades de dados. Dados estes que

podem ser usados diretamente em aplicações, sem que tenham que realizar processos de tradução custosos, desde que mecanismos de representação compartilhada sejam usados.

Pastor-Sánchez (2012) apresenta em seus estudos outras declarações complementares que o RDF oferece, que estão no **Quadro 7**.

Quadro 7 - Estruturas e Declarações do RDF

Estruturas	Declarações
Contentores	Permitem a descrição de grupo de elementos. Podem descrever grupos onde a ordem dos elementos não é importante; ou podem descrever onde a ordem é relevante e descrever grupos com elementos alternativos os quais podem ser selecionados para atribuir uma propriedade.
Coleções	Permitem criar grupos fechados de elementos. Esta estrutura tem a funcionalidade oposta dos contenedores, que é aberta, permite enumerar de um modo mais preciso os elementos que os compõem.
Modelagem	Permite descrever com declarações RDF de alto nível com o intuito de expressar conhecimento sobre outras declarações. Verifica que casos como estes, a declaração se contempla como um recurso, representando explicitamente ao sujeito, predicado, objeto e tipo da afirmação.

Fonte: Elaborado pela autora com base em PASTOR-SÁNCHEZ (2012)

Com todo o contexto apresentado, verifica-se que o RDF somente permite a estruturação dos dados, a começar com o uso de algum vocabulário de um domínio específico, permitindo assim inferências e relações semânticas entre os conteúdos.

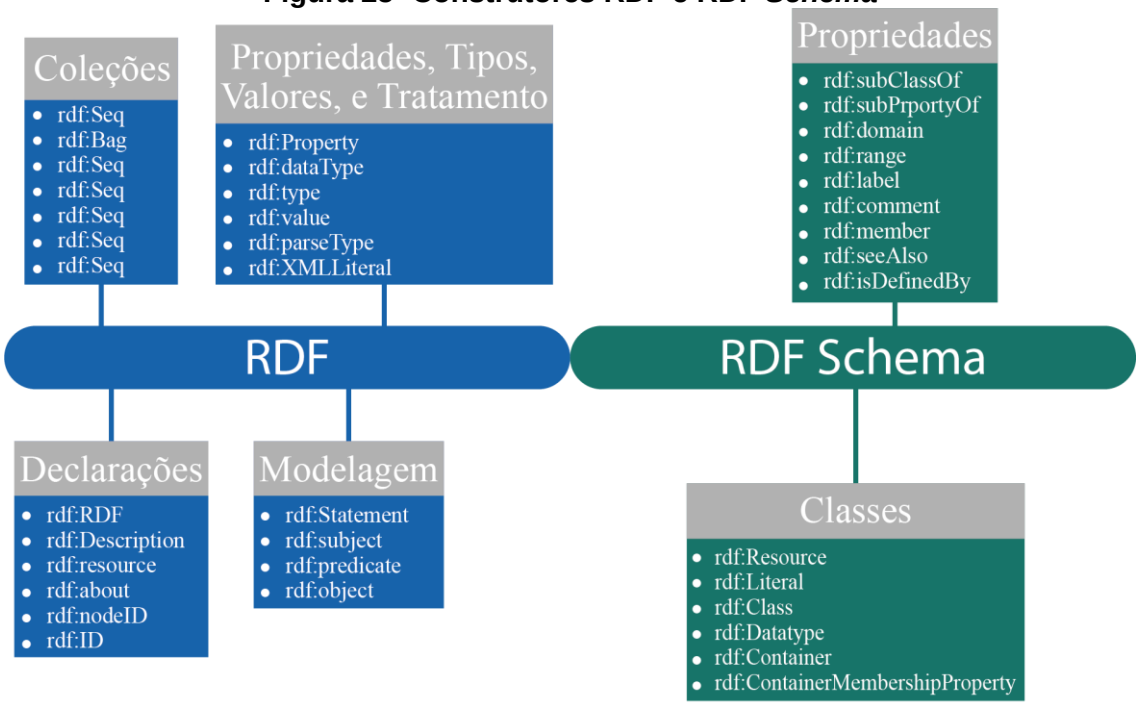
Segundo Antoniou e Harmelen (2008) é o RDF *Schema* (RDFs) que define esses vocabulários, pois permite a criação de vocabulários para definição dos metadados específicos para um conjunto de dados de um determinado domínio, e Brickley; Guha, (2014) afirmam que, devido a estas características descritas, os RDFs podem ser classificados como uma extensão semântica do RDF.

Para fazer a representação de domínios específicos na *Web*, como todo processo, primeiramente é necessário saber o que se quer falar sobre ele, sendo assim, Antoniou; Harmelen (2008) afirmam que o RDFS permite a definição de Classes, que podem ser classificadas como os tipos de entidades que serão descritos, e as Propriedades, que podem ser tipos de

relacionamento entre entidades, como atributos de uma entidade, de acordo com Brickle e Guha (2004) e Manola e Miller (2004).

A partir dos elementos vistos no **Quadro 6**, apresentadas por Pastor-Sánchez, Martínez-Méndez e Rodríguez-Muñoz (2012), a **Figura 25** apresenta os construtores RDF e RDFs que podem se agrupar em categorias funcionais.

Figura 25- Construtores RDF e RDF *Schema*



Fonte: PASTOR-SÁNCHEZ; MARTINÉZ-MÉNDEZ; RODRÍGUEZ-MUÑOZ (2012, não paginado)

A forma de se definir as classes e propriedades no RDFs é determinada a partir dos atributos que seus objetos irão possuir. Suas propriedades são definidas levando-se em conta a quais classes de recursos elas podem ser aplicadas, afirmam Brickley e Guha (2004); e muitas vezes as classes são utilizadas para gerar restrições ao domínio e alcance das propriedades (ANTONIOU; HARMELEN, 2008).

A W3C publicou uma relação de documentos normativos para trabalhar com RDF como apresenta o **Quadro 8**:

Quadro 8 - Relação de normas para trabalhar com RDF

Tipos	Normas W3C
Plataforma de dados conectados 1.0 (SPEICHER, 2012)	Uso de HTTP para: ● acessar, criar e eliminar recursos dos servidores que expõe seus recursos como *Linked Data*. ● Explicações e extensões de regras do *Linked Data*. ● Descrição do seu status utilizando RDF
Esquema RDF 1.1 (BRICKLEY; GUHA, 2014)	Documento que descreve como utilizar RDF para descrever vocabulários RDF: ● Define um vocabulário para este propósito; ● define outros elementos de construção de vocabulário RDF; ○ especificando: modelo RDF e a sintaxe.
RDF 1.1 XML Syntax (GANDON; SCHREIBER, 2014)	Definição de sintaxe XML para RDF ● RDF/XML - em termos de namespaces em XML ● Gramática Formal da sintaxe: ○ geração de triplas RDF no formato de serialização N-Triples RDF (HETTNE et al., 2014)
RDF 1.1 Conceitos e Sintaxe Abstrata (HETTNE *et al.,* 2014)	Definição de uma Sintaxe abstrata (modelo de dados) ● conecta todos os tipos e especificações RDF
RDF 1.1 Turtle (BECKETT *et al.,* 2014)	Descrição de recursos (RDF) ● Linguagem de propósito geral para representar informação na *Web*
RDF 1.1 N-Quads (CAROTHERS, 2014)	Formato de Texto plano para codificar um conjunto de dados RDF em múltiplos grafos.
RDF 1.1 N- Triples (BECKETT *et al.*, 2014)	Formato de Texto plano para codificar um grafo RDF. ● Sintaxe utilizada para caso de Testes em RDF
RDF 1.1 Semantics (PATEL-SCHNEIDER, 2015)	Documento que descreve uma semântica precisa para descrição do Resource *Framework* 1.1 e para o RDF Scheme ● regras de vinculação distintas e que correspondem ao padrão de vinculação de dados
RDF TriG (BIZER; CYGANIAK, 2014)	Documento que define uma sintaxe textual para RDF chamado TriG: ● permite que um conjunto de dados RDF seja escrito de forma compacta de texto e natural (com abreviaturas) ● extensão do formato RDF 1.1 Tutle (BECKETT *et al.* 2014)
JSON-LD 1.0 Algoritmos de Processamento e API (LONGLEY *et al.*, 2020)	Interface de aplicação e programação e um conjunto de algoritmos: ● Transformação para formatos mais fáceis de trabalhar com a máquina: ○ JavaScript, Python e Ruby
JSON-LD 1.0 (SPORNY; KELLOGG; LANTHALER, 2014)	Formato usado para representar grafos dirigidos: ● Combinação do *Linked Data* e *non Linked Data* (em um só documento)

rdf:PlainLiteral (BAO et al., 2012)	Documento de com especificação de dados primitivos e literais simples de RDF
RDF Semantics (HAYES, 2004)	Especificação de uma semântica precisa ● Sistemas completos de regra de inferência ○ RDF e RDF Schema correspondentes
Casos de teste RDF (BECKETT, 2004)	Descrição de casos de teste RDF
Resource Description *Framework* (RDF) (KLYNE; CARROLL, 2004)	Define uma sintaxe abstrata que se baseia em RDF e serve para amarrar sua sintaxe concreta com a semântica formal. ● objetivos de design ● conceitos chave ● datatyping ● normalização de caracteres ● referências URI
RDF Primer (MANOLA; MILLER, 2004)	Manual para os usuários utilizarem os conceitos básicos do RDF

Fonte: Elaborada pela autora com base padrões W3C (2014)

No entanto, RDF e RDF Schema não são suficientes, na sua totalidade, para lidar com a representação de princípios semânticos na *Web*, verificando assim que constitui a causa fundamental para o desenvolvimento de outros padrões, como, por exemplo OWL ou SKOS, baseados em RDF e que fornecem linguagem para definir ontologias estruturadas, que permitem a integração e interoperabilidade de dados entre comunidades descritivas.

4.1.3.3 Ontologias

O termo ontologia é proveniente da filosofia, porém tem verificado de forma marcante e interdisciplinar na linguística e ciências cognitivas em geral o que tem trazido um aporte de diferentes definições para cada área de conhecimento aplicada.

No contexto da *Web* Semântica, a ontologia pode ser determinada como um documento ou arquivo que define formalmente relações entre termos, em que o tipo clássico é uma sistemática, ou denominada como taxonomia e um conjunto de regras inferenciais, que contemplam um conjunto de vocabulários de referência, como maneira de facilitar a comunicação dos metadados, definidos por classes de objetos e relações entre elas, onde uma vez criadas, as ontologias fornecerão os subsídios para que agentes computacionais possam buscar informações que possuam sentido semântico dentro da *Web*.

Gruber (1993), pesquisador que traz um aporte na área computacional, afirma que ontologia é uma especificação explícita e formal sobre um conceitualização compartilhada.

Neches *et al.* (1991) definem ontologia como uma visão procedimental desde a conceitualização como um vocabulário de uma área, mediante a um conjunto de terminologias básicas e relações entre esses termos, relações essas que implicam nas definições determinadas naquele vocabulário.

García-Marco (2007) afirma que ontologia é o ramo relacionado com a representação do conhecimento que se dedica à construção de sistemas inteligentes, ou seja, sistemas que podem descrever e representar uma área de conhecimento que expressa as relações entre eles por meio de uma linguagem formal que pode ser entendida por uma máquina.

Para Guarino (1998, p. 7, tradução nossa), ontologia é “uma maneira de se conceitualizar de forma explícita e formal os conceitos e restrições relacionados a um domínio de interesse”.

O termo ontologia tem tido notoriedade nas áreas da Ciência da Informação e da Ciência da Computação desde a década de 90, pelo fator preponderante que, nestas áreas, ela pode ser classificada como um modelo de dados que representa um conjunto de conceitos dentro de um domínio e os relacionamentos entre estes, pois, dentro deste contexto, uma ontologia é utilizada para realizar inferência sobre os objetos do domínio, segundo Breitman (2005), que afirma que é a definição suficiente para compreender o que ontologias são e suas contribuições.

Noy e McGuinness (2001) apontam que as ontologias têm por objetivo, uma vez que foi conceituado o termo, intervir no entendimento da estrutura das informações entre pessoas ou agentes de *software*, o que deve ser necessário na extração e recuperação de informações, em páginas da *Web*, de conteúdo temáticos; permitir a reutilização de conhecimentos pertencentes a um domínio; explicar as suposições de um domínio para a representação de conhecimento e informação além de considerações técnicas, operacionais e informativas, separar o conhecimento de domínio do conhecimento operacional, pois se refere ao fato de que, eventualmente, o conhecimento que está sendo representado pode estar relacionado a diferentes áreas,ou seja, mais ao conhecimento relacionado a processos; e, para finalizar, analisar o

conhecimento de um campo, no que se refere aos estudos das terminologias e relações.

Todos esses objetivos apresentados por Noy e McGuinness (2001) são pontuados por Gruber (1993) para que sejam alcançados. As ontologias requerem diversos componentes que permitem representar o conhecimento de algum domínio que são: conceitos, relacionamentos, funções, instâncias e axiomas.

O **Quadro 9** apresenta suas características.

Quadro 9 - Componentes de uma ontologia

Componentes	Descrição
Conceitos	Ideias básicas para a formalização: classes de objetos, métodos, planos, estratégias, processos de raciocínio entre outros.
Relacionamentos	Representam a interação e enlace entre os conceitos do domínio, a taxonomia: subclasses
Funções	tipo concreto de relação em que se identifica um elemento mediante a cáculo de uma função que considera vários elementos da ontologia
Instâncias	se utilizam para representar objetos determinados de um conceito
Axiomas	são teoremas sobre os relacionamentos. Que devem cumprir os elementos da ontologia

Fonte: Elaborado pela autora com base em GRUBBER (1993)

As ontologias trabalham com conceitos relacionados, como redes semânticas e tesauros, e permitem definir relações semânticas completas, regras e axiomas, que nem sempre estão presentes em todo o contexto do domínio, isso traz possibilidades de comunicação entre pessoas, agentes e sistemas, pois uma vez que ela permite o reuso, mapeamento de formalismos, compartilhamento de conhecimentos (NOY; McGUINESS, 2001).

Para se tornarem computacionais, além de toda sua estruturação informacional, como decidir seu tipo, critérios de construção, uso de metodologias adequadas, Santarem e Coneglian (2015, p. 227) afirmam que as ontologias “precisam passar de uma estrutura conceitual para implementação através de uma linguagem”.

São inúmeras as linguagens existentes para implementação de ontologias: CycL (Lenat *et al.* 1990), Loom (MACGREGOR, 1991), Ontolingua/KIF (GRUBER, 1993), Flogic (KIEFER; BERNSTEIN; STOCKER,

1995), RDF(S) (KLYNE; CARROLL, 2004), SHOE (HEFLIN; HENDLER, 2000), XOI (FENSEL; BUSSLER, 2002), DAML+OIL (HORROCKS, 2002) e OWL (OWL WORKING GROUP, 2012).

OWL é a sigla usada para referir-se a Linguagem *Web* para Ontologias, ou *Ontology Web Language*, recomendada pela W3C e publicada em 10 de fevereiro de 2004, que é um padrão internacional para codificação e trocas de ontologias baseado em lógicas descritivas que favorecem o desenvolvimento de aplicações semânticas (OWL WORKING GROUP, 2012).

Sintaticamente, uma linguagem de ontologia OWL é um documento RDF e, consequentemente com ele, um documento XML bem estruturado, permitindo que OWL seja processado por diversas ferramentas XML e RDF disponíveis.

Heflin (2007) define que cada um dos termos importantes do RDF Schema está incluso na OWL ou é substituído por novos termos. Já Dean e Ghemawat (2004) afirmam que OWL atribuiu um significado adicional a certas triplas do RDF, aumentando assim a expressividade deste padrão.

4.1.3.3.1 Linguagem OWL

A OWL (*Web Ontology Language*) é uma linguagem que usa o padrão internacional, desenvolvido pela W3C desde 2004, para a codificação e intercâmbio de ontologias, baseado em lógicas descritivas que favorecem o desenvolvimento de aplicações semânticas.

Linguagem que vem para apoiar no desenvolvimento do projeto da *Web* semântica, e foi criada para ter compatibilidade com XML. Sintaticamente, uma ontologia OWL é um documento RDF e, consequentemente com ele, um documento XML bem estruturado.

Isto permite que a linguagem OWL seja processada por diversas ferramentas XML e RDF disponíveis.

Segundo Heflin (2007), verifica-se que cada um dos termos RDF Shchema importantes está incluído na OWL ou está substituído por novos termos; enquanto Dean e Ghemawat (2004) afirmam que a linguagem atribui um significado adicional a certas triplas RDF, o que aumenta a expressividade deste padrão.

Neste contexto, Dean e Ghemawat (2004) ainda apresentam que a linguagem OWL está desenhada para ser utilizada por aplicações que necessitam processar o conteúdo da informação, e não somente apresentar a informação às pessoas. Essa verificação técnica facilita a interpretação do conteúdo *Web* para a máquina, proporcionando o vocabulário adicional junto com uma semântica formal.

Verifica-se que a linguagem OWL tem três divisões expressivas:

- OWL Lite: é uma extensão de RDFs e incorpora os elementos mais comuns da OWL, destinada a usuários que somente necessitam criar taxonomias de classes e restrições simples.
- OWL DL: inclui completamente o vocabulário OWL, fornece 40 elementos primitivos (16 classes e 24 propriedades). OWL DL impõe uma série de restrições sobre grafos RDF e assegura que um documento RDF/XML cumpre com todas as restrições (KUTER *et al.*, 2005).
- OWL Full: fornece maior flexibilidade para representar ontologias que OWL DL. É essencialmente uma ampliação de RDF com vocabulário adicional, sem restrições na forma, e que se utiliza vocabulário (BECHHOFER et al., 2004).

Os diferentes construtores de OWL Lite, OWL DL e OWL Full podem agrupar-se em uma série de categorias funcionais, segundo Pastor-Sánchez; Martínez-Méndez e Rodríguez-Muñoz (2012), e ainda afirma que OWL Full se contempla como uma extensão de RDF, enquanto OWL Lite e OWL DL podem se classificar como extensões de uma visão restringida do RDF.

Todo documento OWL é um documento RDF e cada documento RDF é um documento OWL Full, porém só alguns documentos RDF podem considerar-se documentos OWL Lite ou DL (PASTOR-SÁNCHEZ; MARTÍNEZ-MÉNDEZ; RODRÍGUEZ-MUÑOZ, 2012).

Senso (2003) afirma que OWL fornece axiomas que dotam de restrições as classes definidas e as relações permitidas entre estes elementos. Como linguagem para a construção de ontologias, permite realizar especificações a

níveis abstratos (conceitual) do conhecimento de um domínio, além disso, utiliza a lógica descritiva. O **Quadro 10** apresenta os principais benefícios.

Quadro 10 - Principais benefícios da Linguagem OWL

Características
Restrições de alcance/faixas
Classes disjuntas
Combinações booleanas (união, intersecção etc.)
Restrições de Cardinalidade
Características de propriedades (transitividade, unicidade, inversa)
Existem técnicas de prova para subconjuntos de lógica de primeira ordem
Propriedades: consistência, completude, tratabilidade
RDF(s) e OWL: são subconjuntos de lógicas descritivas
Outros formalismos: cláusulas
Linguagem de regras em desenvolvimento

Fonte: Elaborado pela autora com base em SENSO (2003)

Por essas razões, OWL é a linguagem de maior significado semântico, que possui uma liberdade para implementar uma ontologia (FERNÁNDEZ, 2008) e é descrita em vários documentos publicados na página de W3C que segue:

- OWL 2 *Web Ontology Language Overview* (MCGUINNESS; VAN HARMELEN, 2004).
- OWL 2 *Web Ontology Language: Mapping to RDF Graphs* W3C *Working* (GRAU *et al.*, 2008).
- OWL 2 Web Ontology Language Profiles (MOTIK *et al.*, 2009).
- OWL 2 *Web Ontology Language XML Serialization* (MOTIK; PARSIA; PATEL-SCHNEIDER, 2009).
- OWL 2 *Web Ontology Language Conformance* (SMITH *et al.*, 2012).
- OWL 2 *Web Ontology Language Direct Semantics* (MOTIK; PATEL-SCHNEIDER; GRAU, 2009).
- OWL *Web Ontology Language Quick Reference Guide* (BAO *et al.*, 2012).

- OWL 2 *Web Ontology Language Structural Specification and Functional-Style Syntax* (MOTIK; PATEL-SCHNEIDER; PARSIA; 2012).
- OWL 2 *Web Ontology Language New Features and Rationale* (Segunda Edição) (GOLBREICH; WALLACE, 2012).
- OWL 2 *Web Ontology Language Document Overview* (W3C, 2012).
- RIF RDF and OWL Compatibility (BRUIJIN; WELTY, 2013).

A capacidade da *Web* Semântica, que vincula informação de múltiplas fontes, é uma característica importante de extrema necessidade que se pode utilizar em muitas aplicações. No entanto, a capacidade para combinar dados de múltiplas fontes, combinadas com o poder de inferência de linguagem OWL, tem algumas dificuldades no percurso, devido principalmente à sua heterogeneidade como apresenta Dean; Ghemawat 2004.

4.1.3.3.2 Ontologia e *Web* Semântica

As ontologias transformaram-se em um tema de investigação popular, desde o princípio da década de 90, em que pesquisadores da área de inteligência artificial começaram a desenvolver ontologias e, hoje, várias áreas se incluem e não só a da Ciência da Computação, mas sim área que inclui a engenharia do conhecimento, a aquisição de conhecimentos, linguagem natural, processamento e representação do conhecimento, entre outros, trabalham sobre elas.

Nesse sentido, diferentes áreas de estudo a tem abordado: a engenharia de software, as matemáticas, a informática e, mais recentemente no campo das Ciências de Informação, como ferramentas na representação de informação (como esquema conceitual), em busca e recuperação (como ferramenta), como sistemas de informação cooperativos, e sua aplicação a bibliotecas digitais, comércio eletrônico e gestão do conhecimento.

Projetos desenvolvidos na Internet com linguagens de codificação de ontologias apresentados por Rodriguez Perojo e Ronda León, (2005) se encontram no **Quadro 11,** a seguir:

Quadro 11 - Projetos e Características de Ontologias usando *Web* Semântica

Projeto	Características
Servidor Ontolingua	● resultado do KSE (*Knowledge Sharing Effort*), ○ Ferramentas que criam ontologias, integrando-as com outras existentes e incorporando-as a novos produtos de software.
SHOE (*Simple HTML Ontology Extensions*)	● Complemento semântico de HTML, ○ conteúdo da página *web* e que pode utilizar-se por agentes de software para o descobrimento de informação. ● SHOE evolui até RDF e é OWL (*Web Ontology Language*), ○ fornecer de uma linguagem que possa utilizar-se para descrever classes e relações entre elas, inerentes a documentos e aplicações *Web*.
FERMI (*Formalization and Experimentation on the Retrieval Multimedia Information*)	● Ferramentas de planejamento, descobrimento e seleção de recursos de informação multimídia.
IMP (*Information Manifold Project*)	● Uso das ontologias para identificar as fontes de informação pertinentes a uma busca/pesquisa, acessá-las, obter documentos relevantes, compará-los, selecionar os mais adequados e oferecer um resumo prévio ao usuário.
UMLS (*Unified Medical Lenguage System*)	● No âmbito da medicina, desenvolvido pela *National Library of Medicine* dos Estados Unidos, que utiliza as ontologias como uma ferramenta mais para o acesso, integração e recuperação de informação biomédica.

Fonte: Elaborado pela autora com base em RODRÍGUEZ PEROJO; RONDA LEÓN (2005)

As ontologias têm sido consideradas uma importante contextualização dentro da *Web* semântica, pois elas permitem a descrição e identificam o que os conceitos humanos significam para que possam ser utilizados os metadados para criar serviços de integração, compartilhamento e processar dados. A ontologia dá um entendimento compartilhado e que permite a

reutilização de conhecimento do domínio, útil para visualizar e comunicar (ANTONIOU E HARMELEN, 2008).

4.1.3.4 Consultas SPARQL

SPARQL (*Protocol and RDF Query Language*) é uma linguagem de consulta de conjunto de dados RDF que, dentro de um contexto da Recuperação da informação, se inclui no formato XML, que detalha o modo como se estruturam os resultados obtidos dessa recuperação, utilizando essa ferramenta. Pode ser denominada também, como um protocolo para acesso à RDF elaborado pelo W3C *RDF Data Access Working Group* (HARRIS e SEABORNE, 2013).

É uma linguagem de consultas sobre grafos RDF padronizado pela W3C, onde esse padrão de consultas surge não apenas com o objetivo de realizar consultas mas também facilitar a recuperação da informação em dados provenientes de base de dados relacionais, estruturados em XML e JSON (*JavaScript Object Notation* - Notação de Objetos JavaScript), com duas versões implementadas que são: versão 1.0, que é a recomendação inicial da QUIN (2015) e a versão 1.1 é a mais recente determinada pela KELLOGG; CHAMPIN; LONGLEY (2020).

As consultas, nesta linguagem, estão baseadas em comparações padrões gráficos, e são descritas em uma estrutura clássica RDF (sujeito, predicado e objeto), mas com a opção da possibilidade de diversas consultas ao invés do uso do termo RDF, segundo Castillo (2010).

Consultas estas, que se denominam semânticas e são a base de sistemas de buscas da *Web* Semântica, adaptando-se em sistema de busca pelas interfaces *Web* para que assim usuários possam usar de forma eficiente e recuperar as informações necessárias, cuja eficácia depende da construção de sentenças dispostas em um grafo e dos operadores booleanos.

A estrutura da linguagem SPARQL se compõe em elementos que serão apresentados e descritos no **Quadro 12**.

Quadro 12 - Elementos que compõem a Linguagem SPARQL

Elemento	Funcionalidade
Select	Cláusula que permite retornar os valores de um dado específico e construir consultas usando condições, identificando as variáveis que aparecem nos resultados de consultas
Prefix	Identificação do recurso mediante a uma URI, que possui a denominação exata de um tipo de dado ou vocabulário de onde se vão encriptar os dados
where	Filtro que declara onde selecionar os dados que serão apresentados na consulta. Proporciona o padrão de grafo básico para a concordância com os grafos de dados

Fonte: Elaborado pela autora com base no W3C (2013)

Por ser uma linguagem de consulta, não há interferência no SPARQL e é uma linguagem orientada a dados de forma que só permite consultar as informações presentes nos modelos. O que se consegue verificar é que os modelos de algumas APIs (*Application Programming Interface*) são vistos como “inteligentes”, uma vez que dá a impressão de que certas triplas são criadas sob demanda, incluindo raciocínio OWL. SPARQL nada mais faz do que pegar a descrição do que a aplicação quer, na forma de uma consulta, e retornar à informação, na forma de um conjunto de ligações ou grafo RDF (APACHE, 2021).

O SPARQL disponibiliza 4 funções de objetivo de busca e da precisão do que se requer (HARRIS; SEAORNE, 2013) que são: *select, construct, ask e describe,* formas de consulta que usam soluções que concordem com o padrão para formar conjuntos de resultados de grafos RDF.

4.1.4 Vocabulários Semânticos para *Web*

As tecnologias da *Web* Semântica que são relacionadas à forma como sintaticamente os metadados podem ser agregados às informações foram apresentadas nas seções anteriores, descrevendo os formatos, padrões e linguagens utilizadas em todo esse processo.

Em um contexto de publicação de dados na *Web* é de fundamental importância se alcançar o objetivo de que o significado pretendido dos dados seja o mesmo que o significado entendido pelo consumidor dos dados, isso implica o uso de vocabulários que possuam uma semântica bem definida.

Um fator que facilita esse objetivo é a utilização de vocabulários de referência ou vocabulários de uso mais comum, alguns documentos ou literaturas acabam trabalhando com os termos vocabulário e ontologia, utilizando-os de forma que, às vezes, possuam as mesmas características. Não é possível verificar na literatura uma separação específica dos dois conceitos, em muitos dos casos, o termo vocabulário é utilizado para o caso de ontologias mais simples (GONZÁLEZ; ANTONIO, 2011).

Pessoas utilizam-se de vocabulários para se criar uma comunicação todos os dias, que seja de um grupo particular de pessoas ou que seja para determinações próprias. Neste contexto, este ajuntamento se define por diversos motivos, tais como: localização geográfica, relação familiar, profissional, social, entre outras.

Inúmeras outras situações e características podem ser citadas, verificando que diversos vocabulários podem ser criados em todos os tipos de conversas. Para que o cenário da *Web* Semântica entre neste contexto, é preciso que se estabeleça um conjunto de vocabulários de referência de maneira que facilite a comunicação dos metadados, além de permitir que o leitor deva estar ciente de que para cada publicação específica, é necessário ser feita uma busca de vocabulários existentes que possam ser utilizados.

Hawke (2013) afirma que os vocabulários definem os conceitos e os relacionamentos e os chamam de "termos", que podem ser usados para descrever e representar uma área de preocupação, mas caracterizados com descrever e representar um domínio específico, ou uma área específica. Os vocabulários são usados para classificar os termos e podem ser usados em um aplicativo específico, caracterizar possíveis relacionamentos e definir possíveis restrições ao uso desses termos.

Na prática, os vocabulários podem ser muito complexos, contendo milhares de termos, ou podem se caracterizar como muito simples, com um ou dois conceitos. Hawke (2013) ainda afirma que não existe uma diferença clara entre o que é chamado de "vocabulários" e "ontologias".

No contexto da *Web* Semântica, usar a palavra “ontologia” é mais tendencioso, pois para uma coleção de termos mais complexa é possivelmente bastante formal. Já chamar de “vocabulário” pode ser usado quando esse

formalismo não é necessário para ser usado, ou seja, apenas em um sentido quando quer dar amplitude.

Porém um fator é certo e Hawke (2013) afirma: os vocabulários são os blocos de construção básicos das técnicas de inferência na *Web* Semântica, o que se torna de extrema importância no contexto do estudo da tese, uma vez que essa inferência gera comunicação entre os metadados. Existem repositórios *online* que armazenam conjuntos de vocabulários de diversos domínios, como Schema.org[8], LOV[9] e Dbpedia[10], Dcterms[11], BioPortal[12], sem contar com repositórios relacionados a Dados governamentais, como e-VoG, entre outros. Laufer (2015) afirma que casos em que nenhum vocabulário satisfaça às necessidades de expressividade dos metadados, ou seja, devem ser adequadamente descritos em metadados para permitir que usuários e ferramentas encontrem, entendam e (re)usem, aí sim existe a necessidade de construção, e uma nova ontologia pode ser criada, com o cuidado em reutilizar o maior número possível de elementos de ontologias já existentes, evitando assim a duplicação de referências diferentes aos mesmos conceitos.

Cada vocabulário é descrito por um documento apontado por uma URI e, além de usar vocabulários de domínio apropriados para enriquecer semanticamente os dados de uma determinada plataforma, podem ser empregados usando metadados contextuais, que podem apresentar uma definição mais detalhada dos conjuntos de dados dessa plataforma e permitir que os usuários obtenham mais informações sobre eles.

Os metadados são informações sobre cada conjunto de dados, como título, provedor, data de publicação e muito mais, que será mais bem detalhado no capítulo 4. O **Quadro 13** apresenta uma lista de vocabulários que podem ser usados por qualquer plataforma para descrever conjuntos de dados individuais e sua qualidade com base em vários atributos.

[8] https://schema.org/
[9] https://lov.linkeddata.es/dataset/lov
[10] http://dbpedia.org/ontology/
[11] (http://dublincore.org/documents/dcmi-terms/)
[12] http://bioportal.bioontology.org/

Quadro 13 - Lista de Vocabulários e Descrições

Nome	Descrição
Data Catalog Vocabulary (DCAT)[13]	Vocabulário RDF projetado para facilitar a interoperabilidade entre catálogos de dados publicados na *Web*.
Vocabulary Of A Friend (VOAF)[14]	Especificação de vocabulário que fornece elementos que permitem a descrição de vocabulários (vocabulários RDFS ou ontologias OWL) usados na Nuvem de Dados Vinculada. Em particular, ele fornece propriedades que expressam as diferentes maneiras como esses vocabulários podem confiar, estender, especificar, anotar ou vincular-se. Baseia-se no Dublin Core e no voiD.
Vocabulary for Annotations (VANN)[15]	Vocabulário para anotar descrições de vocabulário.
Vocabulary of Interlinked Datasets (VoID)[16]	Vocabulário do Esquema RDF para expressar metadados sobre os conjuntos de dados RDF. Ele é planejado como uma ponte entre os editores e os usuários de dados RDF, com aplicativos que vão da descoberta de dados à catalogação e arquivamento de conjuntos de dados.
Dublin Core (*Dublin Core Metadata Initiative*)[17]	Vocabulário para a descrição de metadados, tendo como premissas que as descrições tenham independência em relação à sintaxe e tenham uma semântica bem definida
FOAF *(Friend of a Friend Project)*[18]	Um sistema experimental de informações vinculadas é uma linguagem de computador que define um dicionário de termos relacionados a pessoas que podem ser usados em dados estruturados (por exemplo, RDFa, JSON-LD, Linked Data).
SKOS (*Simple Knowledge Organization System*)[19]	É uma área de trabalho que desenvolve especificações e padrões para apoiar o uso de sistemas de organização do conhecimento, como dicionário de sinônimos, esquemas de classificação, sistemas de cabeçalho e taxonomias dentro da estrutura da *Web* Semântica. O uso do RDF também permite que os sistemas de organização do conhecimento sejam usados em aplicativos de metadados distribuídos e descentralizados. Os metadados descentralizados estão se tornando um cenário típico, em que os provedores de serviços desejam agregar valor aos metadados coletados de várias fontes.
Schema.org[20]	Fornece uma coleção de vocabulários que podem ser utilizados para embutir metadados em páginas *Web*, e são entendidas pelas principais máquinas de busca: Google, Microsoft, Yandex e Yahoo!. Os metadados podem ser embutidos utilizando microdados, RDFa ou JSON-LD.
VIVO[21]	Representa pesquisadores no contexto de suas experiências, resultados, interesses, realizações e instituições associadas.

Fonte: Elaborado pela autora (2020)

[13] Disponível em: http://www.w3.org/TR/vocab-dcat/
[14] Disponível em: http://purl.org/vocommons/voaf
[15] Disponível em: http://purl.org/vocab/vann/
[16] Disponível em: http://rdfs.org/ns/void#
[17] Disponível em: https://dublincore.org/
[18] Disponível em: http://www.foaf-project.org/
[19] Disponível em: https://www.w3.org/2004/02/skos/
[20] Disponível em: https://schema.org/
[21] Disponível em: https://bioportal.bioontology.org/ontologies/VIVO

4.1.5 Agentes Inteligentes

Berners-Lee; Hendler e Lassila (2001), no artigo “*The Semantic Web,*” já propunham um mundo em que programas e dispositivos especializados e personalizados, denominados como agentes, iriam interagir por meio da infraestrutura de dados da Internet trocando informações entre si, de forma a automatizar tarefas rotineiras dos usuários. Essa teoria só cresce, pois com as tecnologias que já se tem os agentes podem executar diferentes funções e modificar seu comportamento dinamicamente.

Sistemas Multiagentes, ou como próprio nome já diz, composto por diversos agentes interagindo, já são capazes de resolver problemas complexos, principalmente aqueles que os sistemas tradicionais não conseguem resolver, em que o foco é resolver problemas em ambientes dinâmicos como na *Web*.

> Sistemas Multiagentes são um meio comum de explorar a capacidade potencial de agentes, combinando muitos agentes em um sistema. Cada agente em um Sistema Multiagente tem informações incompletas e os agentes formam um sistema que tem informações suficientes e capacidade para resolver o problema. O sistema não tem um mecanismo de controle centralizado para resolver o problema. (COPPIN, 2017, p. 243).

Petrie (1996) tem como definição que agentes inteligentes são programas que foram definidos por vários autores e que, em muitos casos, não é preciso muito para discutir se é sistema ou não um agente. As classificações que podem ser feitas referentes aos agentes podem ser variadas e depende das diferentes características que são levadas em consideração a cada situação do sistema, pois Aguillo (1997) e Cornella (1999) afirmam as diversas classificações que os agentes podem ter, desde agentes de reações, intencionais, sociais, móveis, híbridos, de informação, de interface e colaborativos, além dos diversos outros tipos existentes.

E ainda Aguillo (1997) e Cornella (1999) afirmam que se deve ter em mente o contexto da Internet, que, mesmo naquela época referente ao estudo sobre agentes inteligentes, já era uma preocupação se entender que o conjunto de recursos acessíveis da *Web* era somente por meio de entradas,

formulários ou conteúdos de criação dinâmica e, na maioria das vezes, os agentes inteligentes tentam oferecer uma solução para esse tipo de situação, pois, em muitos casos, eles são de vital importância na recuperação de informações para a qualidade do conteúdo buscado.

Os agentes inteligentes têm uma série de características que os definem, mas que nem sempre todas essas características estão presentes na sua totalidade e, por outro ponto de vista, boa parte dessas características é suficientemente clara para admitir diferentes interpretações sobre seu conteúdo real e concreto, segundo Aguillo (2000).

Em termos gerais, as características típicas de um agente inteligente são:

- Segundo Wooldridge e Jennings (1995), a **Autonomia.**
    - Os agentes devem trabalhar sem supervisão humana, ao contrário de programas que operam com base em interfaces de manipulação direta do usuário. Assim, uma vez que as condições e restrições exigidas pela parte do usuário, espera-se que o agente tente cobrir ou alcançar seus objetivos, deixando detalhes ocultos para o referido usuário.
- Segundo Roseler e Hawkins (1994), a **Inteligência.**
    - Característica que é descrita em diferentes cenários e formas de especificar essa inteligência e como são mencionados. A inteligência pode ser entendida de maneiras muito diferentes.
- Segundo Finin (1992) e Labrou e Finin (1998) a **Cooperação.**
    - Um agente deve colaborar com outros agentes, trocando informações e resultados de ações próprias. A negociação pode ser feita com agentes que possuem os mesmos objetivos (para que, por exemplo, objetivos já alcançados por um agente possam ser transferidos para outro agente que o persiga) ou com agentes de objetivos diferentes, mas necessários ou pelo menos úteis para atingir os primeiros objetivos. A capacidade de cooperação requer

um mecanismo que permita a negociação entre agentes, mesmo quando estes são heterogêneos.

- Segundo Hendler (2001), a **Comunicação.**
    - Isso não implica apenas a simples capacidade de se comunicar com o usuário, mas também a necessidade de ter conhecimento sobre o mundo ou domínio sobre o qual o agente opera. Esse conhecimento geralmente é implementado por meio de ontologias e regras de inferência.
- Fernández-Berrocal *et al.* (2012), **Reatividade e Adaptabilidade.**
    - Um agente deve ser capaz de responder aos eventos, tomar suas próprias decisões, até mesmo modificando o modo de operar, sempre com o objetivo de realização de seus objetivos e deve ser capaz de aprender com experiências passadas (e com a experiência de outros agentes), bem como as reações do usuário a resultados anteriores. Isso está diretamente relacionado ao aprendizado de máquina.

E com a proliferação de dados e informações, é óbvio que agentes inteligentes geram o processo de exploração da *Web* de forma automática, a fim de recuperar informações relevantes para determinadas necessidades específicas.

A formulação de tais necessidades faz parte das especificações iniciais fornecidas ao agente, onde o processo de explorar a rede, permite a escolha das URIs mais promissoras, acessando a novas páginas e os seus dados, coletando aqueles que podem atender às especificações iniciais.

Bertocchi (2009) afirma que a própria exploração da *Web*, manual ou automática, requer tempo, uma vez que essas abordagens com os agentes inteligentes têm limitações. Não se espera uma resposta imediata, nem mesmo com uma agilidade suficiente para propor um dinamismo, especialmente quando o processo está relacionado com a interatividade com o usuário.

Deve ser muito bem alinhada com os agentes inteligentes, uma vez que é o usuário quem delega as tarefas ao agente, depois de fornecer algumas instruções, ou seja, indicando que tipo de informação é desejada.

É permitido ao agente fazer seu trabalho de forma autônoma e levar tempo, aguardando um período razoável para entregar o resultado de seu trabalho, ou seja, as páginas e seus conteúdos vindos da *Web* que são considerados úteis para satisfazer às necessidades do usuário.

A outra limitação importante dessa abordagem é a renúncia implícita à completude. Dado o tamanho da *Web*, parece claro que a verificação completa, ou mesmo uma parte significativa é implantável; pelo contrário, agentes desse tipo de trabalho explorariam apenas uma parte da *Web*.

Por outro lado, espera-se que os resultados obtidos atinjam uma precisão notável, pois os agentes evitam o efeito de sobrecarga de informação, e aceitando essas limitações, os agentes inteligentes devem resolver uma série de problemas, para os quais várias soluções foram propostas, a maioria delas não excluindo um ao outro.

Para a *Web* Semântica é estrutural entender a importância das ontologias no contexto digital. A troca e aquisição de ontologias da qual fala Bertocchi (2009) é justamente a capacidade gerada por programação para que agentes inteligentes, segundo SOUSA, ALVARENGA (2014), saibam o conteúdo do qual os arquivos são feitos, aprendam sobre eles e troquem essas informações com outros agentes inteligentes.

Muitos serviços na *Web* existem sem semântica, e os agentes não têm como localizar entre eles algum que seja capaz de realizar uma função específica.

Esse processo, denominado descoberta de serviços (*service discovery*), só pode acontecer quando existe uma linguagem comum para descrever o serviço, permitindo a outros agentes entender a função oferecida e como se servir dela.

Agentes produtores e consumidores podem chegar a um entendimento compartilhado, intercambiando ontologias que fornecem o vocabulário necessário para uma negociação. E é aí que novamente entra o conceito determinado por Bertocchi (2009, p. 14), de que a *Web* Semântica possuirá vários agentes inteligentes interagindo entre si, compreendendo, trocando

ontologias, adquirindo novas capacidades racionais adquirindo novas ontologias e formando cadeias que facilitam a comunicação e a ação humana.

Os agentes podem dar o impulso inicial a novas capacidades de raciocínio para a descoberta de novas ontologias, e a semântica pode, também, facilitar o aproveitamento de um serviço que somente satisfaz a demanda parcialmente.

Um processo típico implica a criação de uma cadeia de valor, na qual subconjuntos de informação passam de um agente para outro, cada um agregando valor, para construir o produto requerido pelo usuário final.

Normas e recomendações já foram estipuladas para descrever a capacidade funcional dos instrumentos. As linguagens tornarem-se mais versáteis para lidar com ontologias e lógica. A *World Wide Web Consortium* (W3C, 2012), consórcio que centraliza o desenvolvimento de linguagens e ferramentas para um melhor funcionamento da *Web*, criou a *Web Ontology Language* (OWL), já descrita anteriormente.

A *Web* Semântica não é uma simples ferramenta para realizar tarefas individuais. Ela foi concebida especificamente para que, de forma adequada, sustentar o desenvolvimento do conhecimento humano.

4.2 *WEB* DE DADOS

Baseado em princípios e estudos de Banco de Dados, como já apresentado e estudado por diversos autores, Silberschatz, Sundarshan e Korth (2006) apresentam que dados viviam em formatos de arquivos e, devido a isso, eram isolados com suas informações, gerando processos de redundância e inconsistências. As resoluções desses problemas vieram através de estudos em que os dados poderiam se organizar em uma estrutura denominada como Banco de Dados. Estrutura essa que possui toda uma ordem de uso e estruturação. Como houve essa organização e estruturação, os bancos de dados começaram a virar sistemas, de forma que seus dados puderam ser distribuídos, comportando-se como federações de banco de dados, ou seja, uma abordagem de arquitetura corporativa que permite a interoperabilidade e o compartilhamento de informações entre objetos descentralizados e semiautônomos.

Devido à proliferação da Informação, o surgimento da *World Wide Web* (WWW), apresentando por Tim Berners-Lee em 1990, o objetivo era que máquinas pudessem interpretar os dados da *Web*, criando um silogismo sobre eles, permitindo a inferência de novos conhecimentos, gerando informações relevantes e serviços que ajudassem usuários (BERNERS-LEE; FIELDING; FRYSTYK, 1996).

Cunha, Lóscio e Souza (2011) afirmam que as perspectivas para a produção de dados são animadoras e desafiadoras, devido ao crescimento e ao imenso volume de dados digitais, a *Web* já é um espaço classificado como global de informações. Se por um lado a oferta de dados digitais cresce exponencialmente, por outro lado, um percentual significativo destes dados pode não ser útil para coisa alguma, uma vez que o computador caminha para interpretação dessas informações, baseada mais uma vez em Berners-Lee, Handler e Lassila (2001), que apresentam o conceito de *Web* semântica, afirmando que é uma extensão da *World Wide Web,* a qual permite aos computadores e humanos trabalharem em cooperação e, nela, a informação é dada com um significado bem definido, permitindo melhor interação entre os computadores e as pessoas e determinando o conceito de *Web* de Dados, como uma rede de dados ou vista apenas como uma mudança natural de paradigma em nosso uso diário da *Web*.

O imenso emaranhado de documentos acessíveis na *Web* é composto de dados e informações de praticamente todas as áreas do conhecimento humano. Contudo, ainda é árdua a tarefa de prover meios eficientes que permitam aproveitar todo esse conteúdo, que pode ser composto tanto por dados estruturados, como de dados provenientes de bancos de dados relacionais, quanto por dados não estruturados, como textos e dados multimídia, e o semiestruturados. No cenário da *Web* atual, o grande volume de dados e a falta de metadados dificultam o acesso à informação útil, específica e relevante, conforme propõem Cunha, Lóscio e Souza (2011).

Buscando alcançar esta visão, a W3C vem trabalhando fortemente na construção de uma nova *Web*, desde novas linguagens, ontologias, boas práticas de publicação de dados, tudo para que atenda aos princípios e padrões abertos e que vá muito além da *Web* que se conhece composta

prioritariamente por arquivos e páginas HTML, e nessa nova *Web*, mais conectada e aberta, denominada a “*Web* dos Dados”.

Na *Web* dos Dados, estipula-se que os dados passem a ser facilmente localizáveis bem como sejam associados a elementos semânticos, como exemplo, os vocabulários. E os dados passam a ser visualizados como recursos de dados, precisando assim de identificadores exclusivos que viabilizem o acesso específico para cada recurso.

Na forma como os dados passam a se relacionar entre si, muda dos tradicionais esquemas de tabelas e bancos de dados para um esquema de sujeito-objeto-predicado, conhecido como tripla, dentre outros avanços.

Contextualizando, a *Web* de Dados usa de Navegadores RDF (*Resource Description Framework*), usa de *links* RDF, além de recursos URIs, HTTP, linguagem OWL e linguagens de acesso SPARQL, que é o intuito deste capítulo mostrar todos esses contextos.

4.2.1 Dados Abertos

Segundo o *Open Knowledge Foundation* (2012, sem paginação, tradução nossa), dados abertos “são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa, sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras”.

Verifica-se que esta definição de dados abertos completa que a *Open Knowledge Foundation* criou em 2005, com o intuito de fornecer uma explicação sucinta e uma definição detalhada sobre o assunto.

Neste contexto, James (2013) afirma que existem dois elementos classificatórios para se entender o dado como aberto, que podem ser a abertura Legal e a abertura Técnica. Abertura legal, segundo o autor supracitado, define que deve ter permissão para obter os dados legalmente, desenvolvê-los e compartilhá-los, pois geralmente é fornecida pela aplicação de uma licença (aberta) apropriada e que permite acesso e reutilização gratuitos dos dados, colocando-os em domínio público.

Já a abertura Técnica apresenta características que não devem gerar barreiras técnicas ao uso desses dados, ou seja, que ao ato de fornecer dados torne as informações extremamente difíceis de trabalhar. Para valer a definição

aberta, a abertura técnica possui vários requisitos, como exigir que os dados sejam legíveis por máquinas e estejam disponíveis em massa (JAMES, 2013).

À medida que o movimento de dados abertos cresce, as comunidades se inscrevem para abrir dados, tornando-se cada vez mais importante que existam definições objetivas e de fácil compreensão do que significa "dados abertos" para que se possa aproveitar todos os benefícios da abertura, e evitar os riscos de criar incompatibilidade entre projetos e fragmentar a comunidade.

A *Open Definition* explica em detalhes que Dados abertos são utilizáveis por qualquer pessoa, independentemente de quem eles sejam, onde estão ou o que desejam fazer com os dados e não deve haver nenhuma restrição sobre quem pode usá-lo e sobre o seu uso comercial. Além disso, sobre a disponibilidade em massa, ou seja, a facilidade de se trabalhar com eles, e a sua gratuidade, ou pelo menos a um custo razoável de reprodução.

Salienta ainda que todas as informações devem ser digitais, para a disponibilidade de se fazer *download* na Internet, e serem facilmente processadas por um computador, explorando toda a possibilidade de seu uso. Isso implica, então, que um dado aberto permite que as pessoas o usem, o utilizem e o redistribuam, incluindo a mistura com outros conjuntos de dados e a distribuição dos resultados.

Baseado em três princípios, James (2013) apresenta quando um dado aberto tem sua potencialidade que são: Disponibilidade e acesso, Reutilização e redistribuição e Participação Universal.

Esse conceito não apresenta a simples abertura como essência, na verdade é preciso assegurar a interoperabilidade dos dados para que se possam integrar as diversas bases e assim permitir o uso combinado de múltiplos sistemas (OKF, 2015).

Existem diversos formatos proprietários para o uso de dados abertos que serão explicados posteriormente em cada seção, que são: RDF, XML, CSV, JSON, ODS, SVG e GML, como cita Rautenberg *et al.* (2018).

4.2.2 Dados Conectados

Como já apresentado anteriormente, o conceito da *Web* Semântica é uma *Web* de Dados - de datas e títulos, números de peças e propriedades

químicas e quaisquer outros dados que se possa conceber. A coleção de tecnologias da *Web* Semântica (RDF, OWL, SKOS, SPARQL entre outros) fornece um ambiente em que o aplicativo pode consultar esses dados, fazer inferências usando vocabulários entre outros.

No entanto, para tornar a *Web* de Dados uma realidade, é importante ter uma enorme quantidade de dados na *Web* disponível em um formato padrão, acessível e gerenciável por ferramentas da *Web* Semântica. Além disso, não apenas a *Web* Semântica precisa de acesso aos dados, mas também devem ser disponibilizadas relações entre os dados para criar uma *Web* de Dados (em oposição a uma simples coleção de conjuntos de dados). Essa coleção de conjuntos de dados inter-relacionados na *Web* também pode ser chamada de Dados Conectados (WOOD *et al.,* 2013).

Para obter e criar dados conectados, as tecnologias devem estar disponíveis para um formato comum (RDF), para fazer conversão ou acesso rápido aos bancos de dados existentes (relacional, XML, HTML, entre outros). Também é importante poder configurar pontos de extremidade de consulta para acessar a esses dados de forma mais conveniente.

O W3C fornece uma paleta de tecnologias (RDF, RDFa, SPARQL) para obter acesso aos dados, além de afirmar que fornece um ambiente em que qualquer aplicativo pode consultar esses dados, fazer inferências e usando vocabulários.

Segundo Wood *et al.* (2013), os dados conectados são caracterizados como principais objetos no contexto da *Web* Semântica: integração em larga escala e raciocínio sobre dados na *Web*.

Quase todos os aplicativos listados em estudos de caso na *Web* e casos de uso semânticos baseiam-se, essencialmente, na acessibilidade e integração de dados conectados em vários níveis de complexidade. Para que a *Web* de Dados se torne uma realidade o fator de impacto está relacionado com a quantidade de dados na *Web* disponível em um formato padrão, acessível e gerenciável por ferramentas da *Web* Semântica.

Vale lembrar que não é apenas a *Web* Semântica que precisa ter acesso aos dados, mas devem ser disponibilizadas também as relações entre os dados para criar uma *Web* de Dados, ou seja, não apenas uma simples coleção de conjuntos de dados.

Essa coleção de conjuntos de dados inter-relacionados na *Web* também pode ser chamada de Dados Conectados, ou o termo determinando pela W3C, em inglês que é *Linked Data* (W3C, 2015a).

4.2.2.1 Anotação Semântica de Dados

Garcia-Garcia (2015) afirma, em um dos seus textos, que os usuários irão se acostumar a colocar metadados de todos os tipos em documentos, com o intuito de permitir uma maior recuperação e o acesso à informação de seu interesse. Além disso, esses metadados contribuirão na organização dos conteúdos da *Web*, uma vez que atendam aos requisitos do meio digital.

Segundo Calegari (2016), a anotação semântica de documentos é o primeiro passo para permitir o processamento automático da informação na *Web* e, por conseguinte, para a criação da *Web* Semântica. As anotações semânticas implicam em um processo de análise, extração e marcação da informação, por meio dos metadados, para enriquecê-las semanticamente.

Neste contexto, somente em ambientes controlados verifica-se uma estruturação do uso de esquemas de classificação, podendo citar como exemplos, de acordo com Garcia-Garcia (2015), as bibliotecas e repositórios digitais ou os ambientes cooperativos.

Pastor Sanchéz (2012) aponta que existem diferentes tipos de anotações e apresentam como as identificar. A tipologia mais apresentada e estudada no contexto semântico é estabelecida a partir do agente que realiza a anotação, que pode ser classificada como automática ou assistida.

A anotação automática é realizada utilizando softwares, que parte de um processo que pode ser puramente automático ou semiautomático. Neste contexto, Pastor Sánchez (2012) afirma que o objetivo principal é atribuir a cada elemento do texto uma *tag* específica e única, eliminando assim qualquer evidência que traga equívoco e que precise de participação humana no processo. Caso sejam identificadas as entidades semânticas, e suas relações, no conteúdo fonte, automaticamente se realizam as correspondências de mapeamentos com as entidades e relações com ontologias.

A anotação assistida tem relação com a participação humana, o usuário é responsável por analisar os documentos, identificar as entidades semânticas

e os relacionamentos entre eles. O revisor se encarrega de determinar a *tag* final, ou seja, o metadado que mais corresponde ao contexto daqueles elementos que, uma vez informatizado, não teria sido capaz de definir unicamente, trazendo ambiguidade, mesmo depois de executar rotinas adequadas.

Outra tipologia a ser validada e estudada nas anotações semânticas tem relação com o lugar onde estão armazenadas. Galindo *et al.* (2014) afirmam que podem ser classificadas como internas ou externas, e não se excluem, podendo se fazer o armazenamento com anotações duplas, dependendo do sistema que queira se desenvolver.

As anotações internas são criadas com linguagens de marcação e armazenadas dentro do mesmo documento que se realiza a anotação. Já as anotações externas são armazenadas em arquivos ou serviços diferentes do documento que é anotado, podendo ser acessadas através de metalinguagens e padrões como XML, OWL e RDF.

4.2.3 Dados Abertos Conectados

Verifica-se que a quantidade de portais que publicam seus dados está crescendo a cada dia, isso se deve ao contexto explorado nesta seção, que é a combinação de dados abertos com dados conectados, pois combinam características que geram a base informacional da *Web* de Dados. Essa publicação de dados faz com que aplicações sejam produzidas em menor tempo e ainda determinadas no formato já conectado, não criando problemas em questão de processamento e aumentando a qualidade do uso de usuários (ISOTANI; BITENCOURT, 2015).

Esse poder de processamento dos dados abertos conectados está relacionado com o uso da arquitetura da *Web* de Dados, pois colocam os dados de forma estruturada.

O conceito, definido como Dados Abertos Conectados (*Linked Open Data*) surgiu para trazer uma nova geração, a de novos conhecimentos, que comandam e ainda permitem a capacidade de gerir múltiplos recursos, desde sistemas, dispositivos, pessoas e organizações, criando e produzindo em

conjunto, trabalho ou informação de uma forma eficiente, favorecendo sempre o processo inovador (RAUTENBERG *et al.*, 2018).

4.2.3.1 As Cinco Estrelas da Abertura de Dados

Tim Berners-Lee (5 [ESTRELAS] DOS DADOS ABERTOS, 2012) sugeriu um esquema de implantação de 5 (cinco) estrelas para o Dados Abertos, classificando em nível de discernimento e qualidade baseados no contexto da Publicação desses dados abertos na *Web*. É classificado em um ciclo de cinco etapas a serem consideradas que são: Especificar; Modelar; Gerar; Publicar; Explorar, onde cada uma delas e combinadas recebem a quantidade de estrelas.

A primeira estrela corresponde à disponibilidade dos dados na *Web* em qualquer formato com licença aberta. A segunda refere-se aos dados disponíveis na Internet de modo estruturado (arquivos Excel com extensão XLS). A estrela concerne a disponibilidade na *Web* de maneira estruturada e em formato não proprietário (CSV).

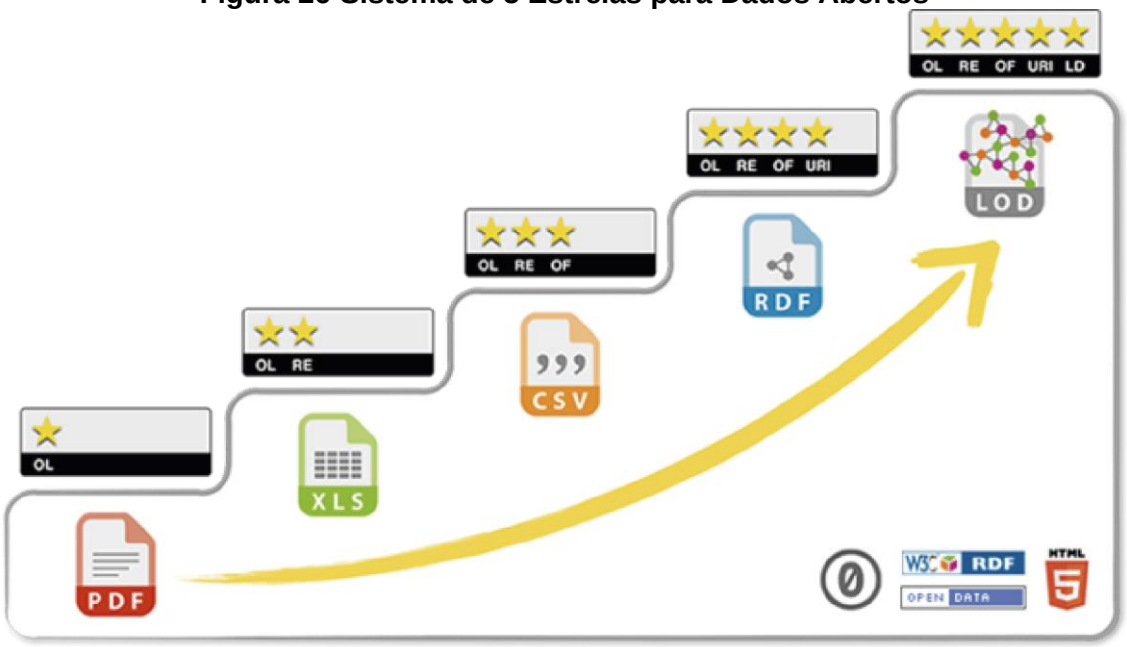
A quarta estrela relaciona-se à utilização das especificações supracitadas, usando URL para identificações, assim os usuários podem discernir. A quinta estrela também emprega todas as outras regras referidas, além da conexão de dados com outros dados para fornecer um contexto.

Portanto, seguindo cada etapa corretamente, é possível alcançar o objetivo pretendido de disponibilizar dados abertos, com qualidade, de modo que outras pessoas possam utilizá-los, segundo (ISOTANI; BITTENCOURT, 2015).

Considerando cada passo, ou estrela, do processo de publicação de dados abertos na *Web*, fica subentendido o uso de ferramentas e processos computacionais para transformação dos dados em padrões aceitos e largamente adotados. Um processo que se aproxima muito dessa evolução das estrelas é o processo de conversão ferramental ou mapeamento semântico, que trabalha com etapas, desde a obtenção dos dados, em qualquer formato (primeira estrela), e vai transformando esses dados através de ferramentas até chegar ao nível mais alto de dados abertos conectados (quinta estrela).

A **Figura 26** exibe o “Sistema de 5 Estrelas” e a explicação de cada uma delas, conforme segue:

Figura 26-Sistema de 5 Estrelas para Dados Abertos



Fonte: 5 [ESTRELAS] DOS DADOS ABERTOS (2012, não paginado)

Shadbolt & O’Hara (2013) salientam que a categorização das 5 estrelas proposta por Tim Berners-Lee é muito usada para classificação, mas também é vista como um esquema de transição de Dados Abertos para Dados Conectados.

Isotani & Bittencourt (2015) afirmam que o dado só pode ser considerado aberto, se ele estiver na classificação de no mínimo três estrelas da escala e, já os dados classificados como Dados Abertos Conectados, deverá receber no mínimo quatro estrelas.

4.2.3.2 Nuvem *Linked Open Data*

Bizer *et al.* (2008) apresentam que o surgimento de uma *Web* de Dados vem do projeto de Dados abertos conectados, conhecido na literatura como *Linked Open data* (LOD), fundado em 2007 e com suporte da *Web* Semântica e da W3C.

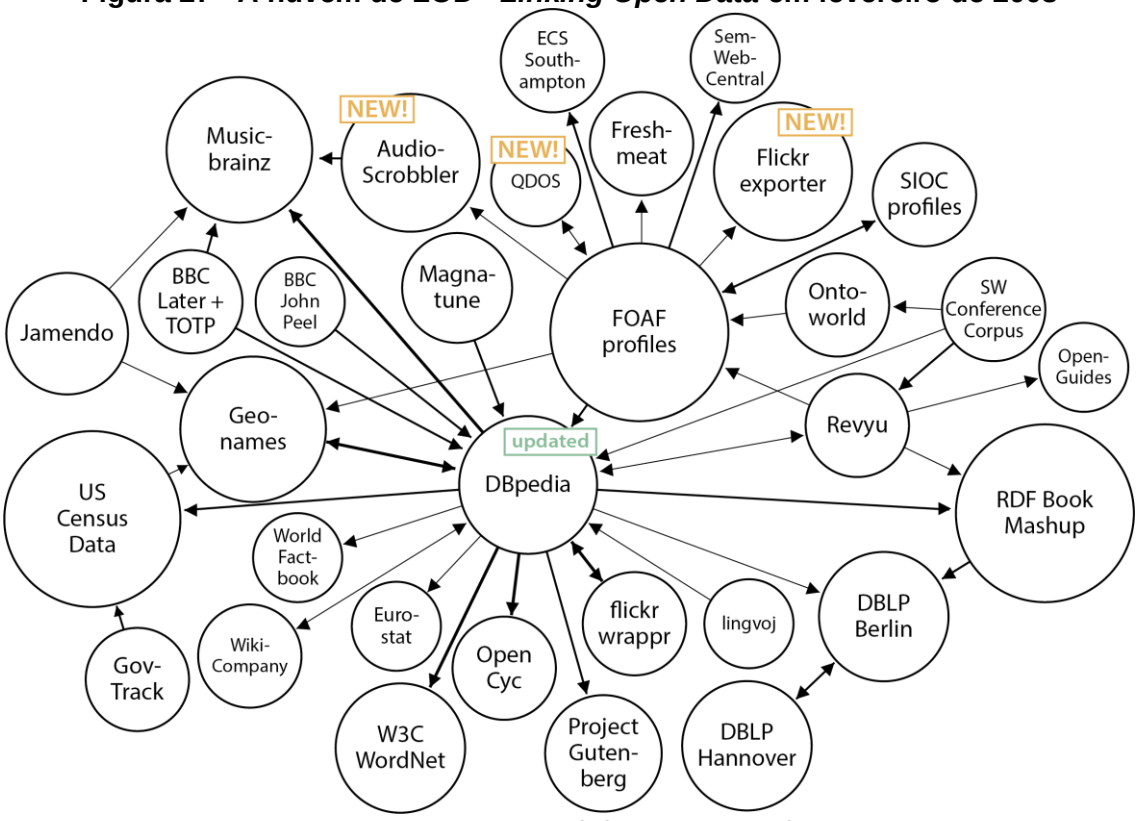
O objetivo do projeto veio para identificar conjuntos de dados disponíveis sob licenças abertas, e o reuso na *Web* de forma interconectada,

ou seja, para identificar conjuntos de dados disponíveis sob licenças abertas, publicando-os novamente em RDF na *Web* e os interconectam.

Em estudos mostram que em 2007 o tamanho da *Web* aumentou para mais de dois bilhões de triplas RDF, verificando todo um esforço e gerenciamento da comunidade para que esses dados viessem a ser uma realidade. Seus resultados foram provenientes de conjuntos de dados em diversos domínios, como regiões geográficas, informações, informações do censo, pessoas, empresas, comunidades online, línguas humanas, publicações científicas, filmes, música, livros e críticas, dados esse que são mostrados por Cyganiak[22] no *site* oficial sobre a evolução do projeto *Linked Open Data Cloud.*

A **Figura 27** apresenta, em formato de grafo, um diagrama com as principais conexões em fevereiro de 2008 da nuvem de dados interligados.

Figura 27 - A nuvem de LOD - *Linking Open Data* em fevereiro de 2008



Fonte: BIZER *et al.* (2008, p. 1266)

Esse diagrama, conforme apresenta Bizer *et al.* (2008) mostra as principais interligações que, na época, eram entre os *sites* como DBpedia e

[22] https://lod-cloud.net/

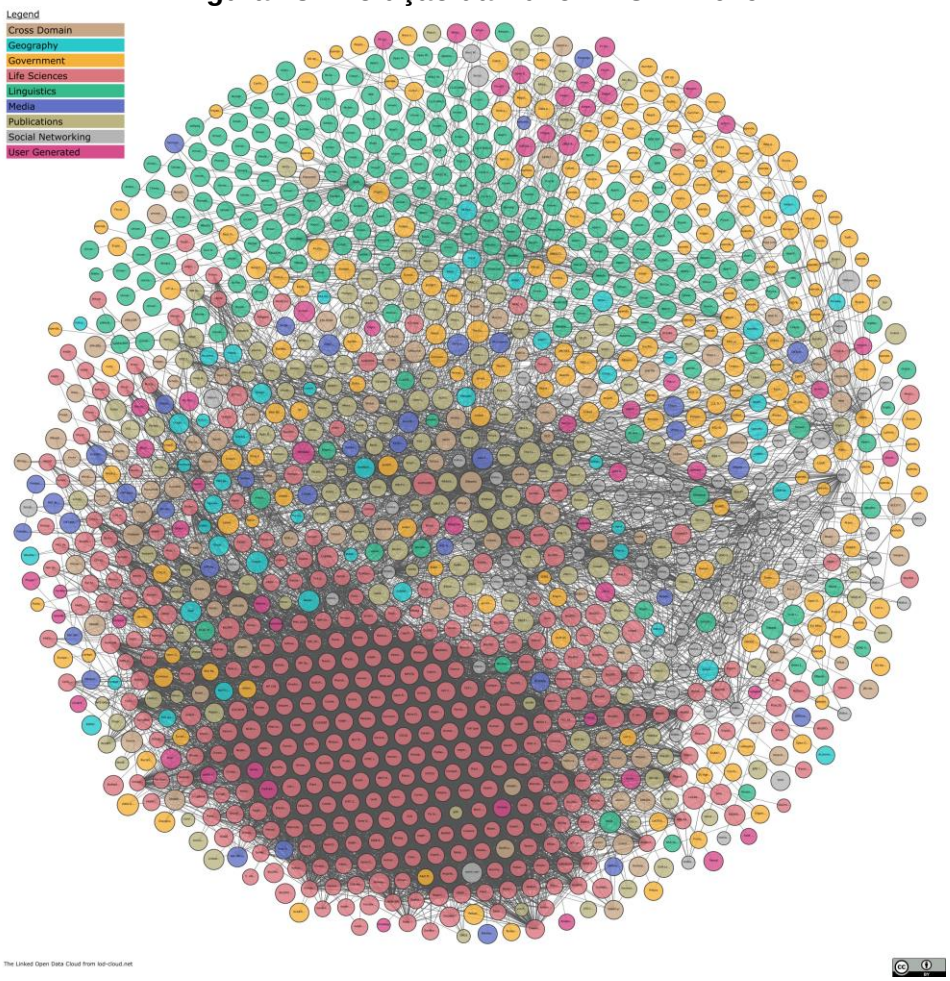
Geonames, de onde DBpedia extraiu triplas RDF de artigos da Wikipedia, tornando-os disponíveis na *Web*, gerando consultas através do SPARQL.

O diagrama de nuvem é mantido por Richard Cyganiak, do *Digital Enterprise Research Institute* (DERI[23]), da Universidade Nacional da Irlanda/Galway, e Jentzsch Anja, da Universidade Freie, Berlin.

De 2007 a 2019, houve novas “nuvens” publicadas. Apresenta a quantidade de *datasets* interligados, apresentando 1239 datasets interligados (CYGANIAK E JENTZSCH (2020).

A **Figura 28** apresenta a imagem da nuvem mostra os conjuntos de dados que são publicados e como estão interligados com outros conjuntos de dados. Outra característica importante é a funcionalidade de oferecer cliques nos *datasets*, possibilitando a navegação. Ao ser clicado, o *dataset* abre todas as suas características e as triplas RDF que possui.

Figura 28-Evolução da Nuvem LOD - 2019



Fonte: CYGANIAK; JENTZSCH (2021, não paginado)

[23] http://www.deri.ie/

4.2.3.3 Conexão entre *datasets* na nuvem LOD

Os vocabulários são indicados para serem usados em aplicações que seguem os princípios de dados conectados, segundo Isotani e Bittencourt (2015), facilitando o processo de integração dos dados de diferentes *datasets* e o processamento desses dados.

Além disso, Heath e Bizer (2011) afirmam que, para a integração desses recursos na *Web* seja possível, é necessário que um dado conectado gere um modelo de dados comum, quanto permita a sua descrição de conceitos e a forma como estão relacionados, pois como já apresentado anteriormente, os vocabulários são utilizados para descrever termos e conceitos específicos no domínio e seus diferentes tipos de relacionamento.

Rozsa, Dutra, Nhacuongue (2017, p. 38) afirmam, baseados nesses contextos apresentados, que “um modelo de dados genérico viabiliza a interoperabilidade sintática, permitindo a integração de dados de fontes diversas, enquanto o compartilhamento de vocabulários viabiliza a interoperabilidade semântica”, o que permite que as aplicações possam conceitualizar os dados formalmente e ainda interpretá-las.

Verifica-se que, dentro de uma mesma conjuntura, os dados de um *dataset* podem ser derivados, devido a diversas situações apresentando uma grande quantidade de termos específicos, e ainda tendo a necessidade da inserção de diferentes vocabulários na descrição destes termos.

Como apresentado por Rozsa, Dutra, Nhacuongue (2017), posto que os vocabulários permitam uma semântica dos dados, significa que *Web* de Dados já possui o alicerce concreto para sua construção e ainda a magnitude do seu papel atrelado à disparidade de situações e circunstâncias que exigem representações específicas, tem como resultado a multiplicação dos vocabulários disponíveis na *Web*, segundo D'Aquin e Noy (2012).

Noy, Fergerson e Musen (2000), como já apresentado anteriormente, afirmam que, caso nenhum vocabulário possua termos que consigam descrever esses dados, aí sim existe a necessidade da criação de novos termos que venham corresponder ao que os dados expressam naquela situação.

Todos esses pontos sugerem uma nova forma de visualização de reuso, enfatizados por Heath e Bizer (2011), pois permitem que aumente a probabilidade de os dados serem consumidos por aplicações que usam os vocabulários já conhecidos, e visto que é uma etapa significativa quando se constrói um *dataset,* pois os dados ficam muito mais fáceis de serem descritos, pois serão buscados de fontes diversas.

Um caso típico, de um grande conjunto de dados conectados, é o DBPedia[24] que disponibiliza o conteúdo da *Wikipedia* em RDF e agrega *links* para outros conjuntos de dados na *Web*. Esses *links*, em formatos de triplas RDF, quando fornecidos permitem que aplicações explorem o conhecimento mais preciso de outros *datasets* (conjuntos de dados), fornecendo uma experiência ao usuário muito melhor.

O *Linked Open Vocabulary* (LOV) significa Vocabulário Aberto Conectado. Esse nome é derivado de LOD, sigla para *Linked Open Data*. Como os dados na *Web* usam propriedades e classes para descrever pessoas, lugares, produtos, eventos e qualquer tipo de coisa um vocabulário no LOV,[25] reúne definições de um conjunto de classes e propriedades, que são denominadas como termos do vocabulário, úteis para descrever tipos específicos de coisas, ou coisas em um determinado domínio ou setor, ou coisas em geral, mas para um uso específico (LINKED..., 2020).

Os termos de vocabulário também fornecem as ligações de dados conectados e as definições de termos fornecidas pelos vocabulários trazem semânticas claras às descrições e *links*, graças à linguagem formal que eles usam (triplas RDF, como RDFS ou OWL).

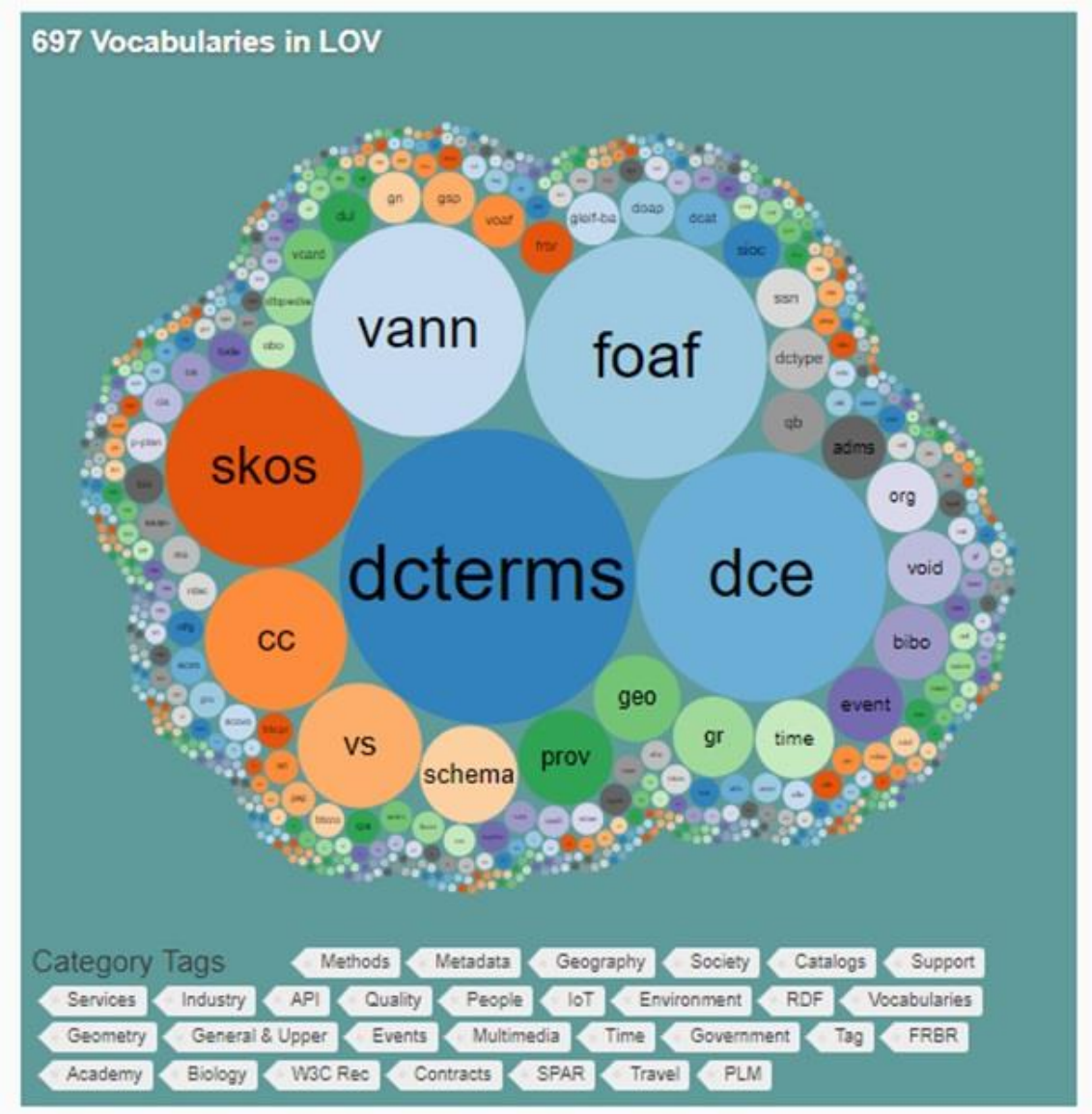
Resumindo, os vocabulários fornecem a semântica, permitindo que os Dados se tornem Dados significativos.

O LOV, segundo a própria documentação no *site* (LINKED..., 2020), oferece uma escolha de várias centenas desses vocabulários, com base em requisitos de qualidade, incluindo estabilidade e disponibilidade de URI na *Web*, uso de formatos padrão e práticas recomendadas de publicação, metadados e documentação de qualidade, organismo de publicação identificável e confiável, política de versão adequada, com o objetivo de auxiliar

[24] https://wiki.dbpedia.org/
[25] https://lov.linkeddata.es/dataset/lov

publicadores e usuários de vocabulários no *Linked Data* a descobrir opções de vocabulários, reutilizá-los e se possível publicar novos vocabulários quando necessário. A **Figura 29** apresenta a nuvem e a representação de 697 vocabulários atrelados aesse conceito.

Figura 29- Vocabulários mais utilizados apresentados na Nuvem LOV



Fonte: LINKED... (2020, não paginado)

4.2.4 Ecossistemas de dados na *Web*

Para entender o contexto de Ecossistema de dados na *Web*, é necessário entender o que é realmente um ecossistema. Da definição do dicionário Aurélio (DICIO, 2020, não paginado) “Ecologia. Sistema (ecológico) que inclui o conjunto das relações dos seres vivos entre si e/ou destes com o ambiente; biogeocenose, biossistema”.

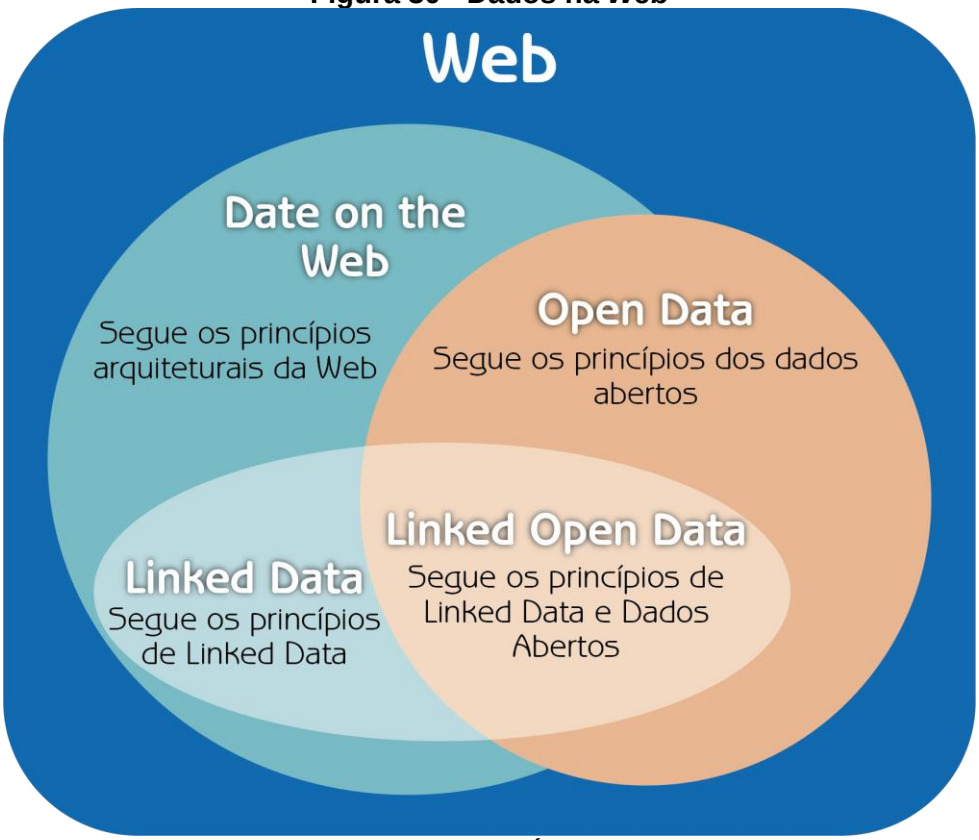
Porém o termo Ecossistema vem sendo usado para definir conjuntos de relações, mas não só que envolvam seres-vivos, mas sim, consiste em um contexto em que uma comunidade de organismos esteja em conjunto com seu ambiente físico.

Isso inclui diversos tipos de Ecossistemas, desde negócios, digitais, software, e os dados na *Web*, entre outros. Segundo Laufer (2015), Tim Berners-Lee criou o primeiro *browser* (navegador) com acesso a documentos por meio de especificação URL e, por consequência, esse acesso permitiu que usuários pudessem editar a informação e a gravasse, nos moldes da Wikipédia.

Dessa forma, o princípio de Ecossistema de Dados começou a ter seu formato, pois a pessoa, classificada como um ator desse ecossistema, exerceria dois papéis: consumidor e publicador de dados.

A **Figura 30**, feita por Oliveira e Lóscio (SBBD, 2019), apresenta como os dados estão dispostos na *Web*, baseados na Arquitetura da *Web* recomendada pela W3C (2015b).

Figura 30 - Dados na *Web*



Fonte: OLIVEIRA; LÓSCIO (2019)

Pollock (2011) fornece a definição mais antiga para um ecossistema de dados: Um ecossistema tem ciclos de dados, nos quais consumidores e intermediários de dados, que podem ser, por exemplo, desenvolvedores de aplicativos, conseguem compartilhar seus dados limpos, integrados e empacotados no ecossistema de forma reutilizável. Geralmente, esses dados limpos e integrados são mais valiosos do que a fonte original.

Zuiderwijk *et al.* (2012) trazem sete propriedades elementares e necessárias para um ecossistema:

- liberação e publicação de dados abertos na internet;
- criação de mecanismos de procura, descoberta, avaliação e visualização de dados e suas licenças relacionadas;
- limpeza, análise, melhoria, combinação e interligação de dados;
- interpretação, discussão e fornecimento de feedback para o provedor dos dados e outras partes interessadas.
- formas de sinalizar ao usuário como os dados abertos podem ser usados;
- sistema de gestão da qualidade e
- diferentes tipos de metadados capazes de conectar os elementos que compõe os dados;

Lee (2014) apresentam princípios e desafios para a construção de um ecossistema de dados abertos. Como princípios, têm-se: suporte do governo, padronização dos elementos, orientação pela demanda, transparência no processo e relacionamento com outras iniciativas internacionais. Como desafios, destacam-se a privacidade de informações, a governança da estrutura decisória, mudanças nos processos internos do governo e utilização dos dados.

Dawes, Vidiasova, Parkhimovich (2016) compilam diversos trabalhos sobre modelos de dados abertos e apresentam, inicialmente, cinco abordagens sobre o tema: foco em dados abertos; orientação a programas de governo; foco no usuário; foco em ecossistemas e foco em impactos e agregação de valor.

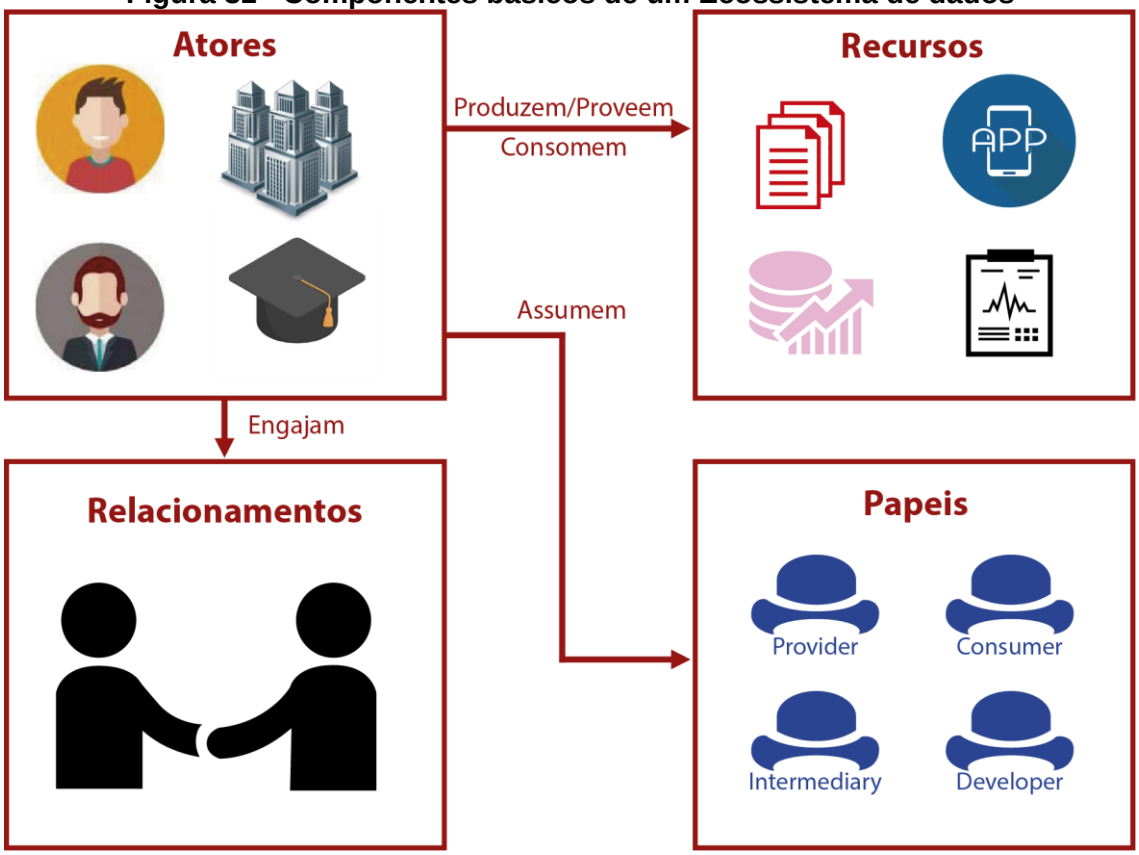
Zubcoff *et al.* (2016) afirmam que um Ecossistema de Dados é composto de muitos atores e pequenas estruturas organizacionais que devem

reconhecer dados como a matéria-prima. Os atores devem formar um ciclo a fim de “alimentar" o ecossistema, proporcionando benefícios a todas as partes.

Zuiderwijk *et al.* (2012) apresentam o ecossistema como todas as atividades para liberar e publicar dados na Internet, para quais usuários de dados podem conduzir atividades como pesquisar, localizar, analisar, enriquecer, integrar e visualizar dados, além de fornecer de *feedback* ao produtor de dados e outras partes interessadas.

Oliveira e Lóscio (2019) apresentaram a **Figura 31** que denota que, a partir das definições usadas pelos autores supracitados com as definições de Ecossistema de Dados na *Web*, eles definem que um ecossistema de dados é um conjunto de redes compostas por atores autônomos que, direta ou indiretamente, consomem, produzem ou fornecem dados e outros recursos relacionados a dados, onde cada ator desempenha um ou mais papéis e está conectado a outros atores por meio de relacionamentos, de forma que a colaboração e a competição dos atores promovem a autorregulação do Ecossistema de Dados.

Figura 31 - Componentes básicos de um Ecossistema de dados



Fonte: OLIVEIRA; LÓSCIO (2019, p. 74)

Oliveira *et al.* (2016) publicaram um artigo que contêm um meta-modelo para ecossistemas de dados, descrevendo todos os pontos importantes e suas características específicas desde os Mapeamento Sistemático, o processo de Meta-Modelo de Descrição de Ecossistemas de Dados na *Web* até a proposta de um *Framework*.

Dentro do contexto de Ecossistema de Dados na *Web*, cada qual possui seu papel e organização, e existe a necessidade de atores que desempenham papéis específicos. Oliveira e Lóscio (2019) afirmam que existem na literatura mais de 35 papéis porém, apresentam 3 papéis de extrema importância: Usuário de dados, que se classifica como responsável pelo consumo direto ou indireto de dados, mas sem a necessidade de consumir dados diretamente dos produtores de dados; Provedor de dados, que se classifica como responsável pela publicação ou fornecimento de dados e o Produtor ou Criador de Dados, que se classifica como responsável pela captura ou geração de dados, podendo também compilar, agregar e empacotar dados.

Na questão da estrutura organizacional, Oliveira e Lóscio (2019) citam diversos exemplos e modelos, desde Organização Centrada em Ator-Chave, como, por exemplo, dados referentes ao governo brasileiro e como as organizações usam esses dados baseados no governo. Organizações baseadas em plataformas, por exemplo, CKAN, Organização baseada em *marketplace* como, por exemplo, Azure da Microsoft, Organização baseada em Intermediários, como por exemplo DATA HUB[26].

Baseado no contexto de atores e organizações, verifica-se uma geração de valor que requer habilidade e capacidades, focando, assim, principalmente no contexto do gerenciamento para Publicação e consumo dos dados na *Web*.

Verifica-se a importância e o cuidado referente a esse gerenciamento, que Kucera (2015) apresenta uma lista de orientações, ou chamados de *guidelines*, para essa publicação e consumo de dados na *Web*. Alguns dos *guidelines:*

- *Best Practices for Publishing Linked Data* (HYLAND; ATEMEZING; VILLAZÓN-TERRAZAS, 2014).

[26] é um modelo de gerenciamento de Big Data que usa uma plataforma Hadoop como o repositório de dados central.

- *Czech Open Government Data Publication Methodology* (CHLAPEK et al., 2014).
- *Government Data Openness and ReUse* (ÁLVAREZ ESPINAR, 2014).
- *Norway Guide for disclosure of public data* (EUROPEAN COMMISSION, 2019).
- *Guidelines on Open Government Data for Citizen Engagement (2ndedition) (*OPEN DATA INSTITUTE, 2020).
- *Open Data Certificate* (OPEN DATA INSTITUTE, 2020).
- *Open Data Field Guide* (SOCRATA, 2020).
- *Open Data Handbook* (OPEN KNOWLEDGE FOUNDATION, 2012*).*
- *Open Data Handbook (Flanders) (*FLEMISH GOVERNMENT, 2014*).*
- *Open Data Institute Guides* (OPEN DATA INSTITUTE, 2020).

O Ecossistema de Dados possui um ciclo de vida dos dados na *Web*, e verificam-se os diferentes papéis e competências que são necessários para a execução das diversas tarefas relacionadas à publicação dos dados e que, segundo Laufer (2015), é o processo de gerenciamento de informações desde a sua fase de seleção até a fase de arquivamento.

O ciclo de vida de dados do ecossistema de dados na Web é composto por 10 fases apresentadas por Laufer (2015), como apresentadas no **Quadro 14** e suas respectivas características.

Quadro 14 - Fases do Ciclo de Vida dos Dados referente ao Ecossistema de Dados na *Web*

Fase	Características
Planejamento	● Fase inicial responsável pelo planejamento da publicação dos dados; ● Seleção dos dados a ser publicados, ● Identificação de equipes e órgãos que devem participar do processo e ● Opções de coleta e plataformas de publicação.
Estruturação	● Definição da estrutura dos dados a serem distribuídos.
Criação e Coleta	● Aquisição dos dados: dados a serem criados e dados já existentes.
Refinamento e Enriquecimento	● Qualidade dos dados: filtro de possíveis incorreções, agregando ou separando informações e fazendo eventuais conexões com dados

	relativos a outras bases; ● Transformação e Integração
Formatação	● Dados que serão publicados passarão por uma formatação de acordo com a plataforma escolhida para publicação
Descrição	● Definição, criação e coleta dos metadados e os documentos que devem ser agregados aos dados: Documentação
Publicação	● Dados são efetivamente colocados na *Web*, na plataforma escolhida para publicação
Consumo	● Dados são consumidos por usuários finais ou por desenvolvedores de aplicações
Realimentação	● Recolhimento de informações relacionadas à utilização dos dados: ○ consumidores dos dados, ○ plataformas de distribuição.
Arquivamento	● Dados são retirados da Web

Fonte: LAUFER (2015, não paginado)

Sob a perspectiva das máquinas, pode-se desenhar a *Web* de Dados como uma camada de dados consumidos por uma camada de aplicações. Tem-se, assim, uma arquitetura onde existem dados publicados de diferentes formas e um conjunto de aplicações que oferecem um conjunto de tarefas e manipulam esse conjunto de dados para gerar seus resultados.

A distribuição desses dados é heterogênea e se comporta ao acesso a dados contidos em páginas *Web*, *downloads* de arquivos armazenados em estruturas de diretórios nos servidores, dados retornados por *Web Services* e *Web APIs*.

Como não existe uma padronização em relação à estrutura e a semântica dos dados distribuídos, são resolvidas pelas próprias aplicações que as usam, uma vez que o entendimento da estrutura e da semântica dos dados fica embutido nas aplicações. Mais uma representação para o ciclo de vida de Dados na Web apresentado por Corrales *et al.* (2018).

4.3 PUBLICAÇÃO DE DADOS NA *WEB*

Obter e/ou disponibilizar informações na Internet é uma prática comum que vem crescendo cada vez mais, pois se percebe que, ao realizar uma busca, desde algo simples até o mais complexo, sempre possui algum subsídio relacionado ao tema.

Assim sendo, o volume de dados tem aumentado consideravelmente ao longo dos anos e a tendência é ampliar cada vez mais esse fator (ISOTANI; BITENCOURT, 2015). Por isso, são necessários um cuidado e um procedimento para que os dados sejam consistentes e disponibilizados corretamente.

Nesse sentido, é necessária a utilização da *Web* Semântica para lidar de maneira adequada com os dados, colaborando assim com a experiência do usuário. Visando a esses fatores, é essencial a junção de dados abertos e *Web* Semântica, pois com o grande volume de dados existentes, torna-se primordial sua disposição de maneira que seja legível para máquina, objetivando uma experiência mais consistente ao usuário.

Para lidar com as limitações das abordagens para publicação de dados, o W3C (*World Wide Web Consortium*) vem recomendando um conjunto de boas práticas para publicar dados de forma estruturada e conectada (BIZER; HEATH; BERNERS-LEE, 2011). Esta abordagem preconiza o uso de vocabulários compartilhados, que são representados através de ontologias, sendo estas formalizadas em linguagens de representação do conhecimento concebidas para a *Web* e objetivam o melhor compartilhamento, consumo e integração dos dados.

Para assegurar que os aspectos supracitados sejam realizados, há um processo e práticas que devem ser seguidos para atingir os resultados desejados na publicação de dados abertos.

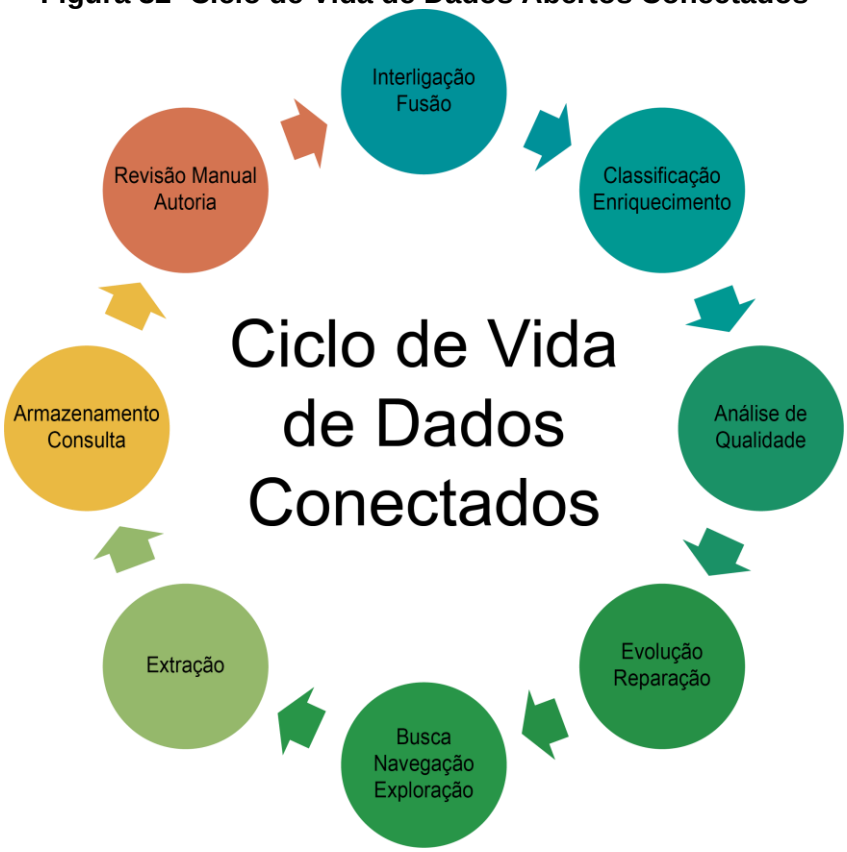
4.3.1 Ciclo de Vida dos Dados Conectados

O Ciclo de Vida de Dados conectados é uma metodologia usada para publicar dados na *Web*, mas de forma conectada, o que permitiu que diversas ferramentas computacionais fossem desenvolvidas, dando o suporte à *Web* nesse quesito. Todas as fases desse processo foram desenvolvidas pelo Instituto de Pesquisa *Agile Knowlegde and Semantic Web* (ARNDT, 2017).

Como já mostrado na seção referente ao Ecossistema de Dados na *Web*, o ciclo de vida de dados conectados também possui 8 etapas, porém utilizando os princípios da *Web* Semântica e criadas de acordo com as

necessidades e características de cada aplicação, como apresentada pela **Figura 32**, determinada por Auer (2014).

Figura 32- Ciclo de Vida de Dados Abertos Conectados



Fonte: AUER (2014, p. 4)

As etapas do ciclo de vida para Publicação de dados abertos conectados, ou do inglês, *Linked Data Lifecycle* segundo Auer (2014) apresenta 8 atividades apresentadas no **Quadro 15**.

Quadro 15 - Atividades referentes ao Ciclo de Vida de Dados Abertos Conectados e suas características

Atividade	Característica
Extração	• Dados não estruturados ou estruturados (diferentes formatos ou provenientes de sistemas legados) ○ Mapeamento para o modelo de dados RDF
Armazenamento/Consulta	• Gerenciamento de dados RDF ○ *Triple Stores* - potencializam as tarefas de publicação e consumo de dados;
Revisão Manual/Autoria	• Tarefas de editoração de dados
Interligação/Fusão	• Dados de uma base são interligados a outros dados de outros conjuntos,

	○ ampliação de contextos informacionais
Classificação/Enriquecimento	● Aumento da expressividade e riqueza semântica dos dados em relação a um domínio, ○ Uso de recursos como ontologias e/ou vocabulários
Análise de Qualidade	● Realização do Tratamento dos aspectos de integridade, precisão, consistência e validade de dados. ○ verificação dos quesitos de consistência, concisão, compreensão, disponibilidade e proveniência dos dados
Evolução/Reparação	● Ações para correção das não conformidades; ○ Uma vez encontradas inconsistências nos dados ou no modelo de representação perante requisitos previamente estabelecidos.
Busca/Navegação/Exploração	● Técnicas de exploração ou visualização ○ Utilizadas para manipular os Dados Conectados em diferentes aplicações.

Fonte: RAUTENBERG *et al.* (2018)

Verifica-se que todas essas características dependem das aplicações a serem desenvolvidas, envolvendo o profissional da informação e o engenheiro de dados para que, assim, sejam disponibilizadas e publicadas na *Web*.

Rautenberg *et al.* (2018) ainda afirmam que deve ser visto todo o processo do mapeamento dos dados ligados, utilizando vocabulários para cada domínio para a *Web* de Dados, garantir a qualidade dos dados e ainda a interligação dos dados que estão em servidores locais com dados já disponibilizados na *Web*, contexto este que motiva para a construção das diretrizes de mapeamento e enriquecimento desses dados na presente tese.

4.3.2 Recomendações e Boas Práticas para a Publicação de Dados na *Web*

As Boas Práticas para Publicação de Dados na *Web* (DWBP - *Data on the Web Best Practices*) são recomendações descritas pela W3C, por Lóscio, Burle e Calegari (2017), que trazem, como contexto principal, técnicas e diretrizes para que os dados que tramitam na *Web* estejam de forma que permita, além da sua facilidade de reuso, que o crescimento seja mútuo e tanto publicadores como consumidores consigam compreender e utilizar esses dados.

As Boas Práticas vêm muito mais com o intuito de compatibilidade entre publicadores e consumidores e, devido a isso, o Grupo de Trabalho de Boas Práticas para Publicação de Dados na *Web*, possui uma missão baseada no contexto da *Data Activity* (W3C, 2020) que apresenta os dados como cada vez mais importantes para a sociedade e o W3C possui um conjunto de padrões da *Web*, com planos para mais trabalho, facilitando o trabalho de desenvolvedores comuns com dados de gráficos e gráficos de conhecimento.

Data Activity ainda afirma que os Dados conectados se referem ao uso de URIs como nomes para coisas, a capacidade de desreferenciar esses URIs para obter mais informações e incluir *links* para outros dados, permitindo fontes cada vez maiores de dados abertos conectados na *Web*, bem como serviços de dados restritos aos fornecedores e consumidores desses serviços.

Neste contexto, Lóscio, Burle e Calegari (2017) trazem como missão três tópicos importantes sobre as Boas Práticas de Publicação de Dados Abertos na *Web*:

- Desenvolver o ecossistema de dados abertos, o que facilita a comunicação entre desenvolvedores e publicadores de dados;
- Providenciar diretrizes para os publicadores de dados, que promoverão consistência na forma como os dados são gerenciados, provendo assim o reuso de dados;
- Fomentar a confiança dos desenvolvedores nos dados, independente da tecnologia utilizada, aumentando o potencial de inovação.

Conforme descrito na **Figura 33**, o contexto dos dados na *Web*, cada dado que possui uma forma física específica de um conjunto de dados passa por diferentes distribuições e é publicado. A facilidade do compartilhamento de dados em larga escala advém baseada nessas distribuições, permitindo assim que conjuntos de dados possam ser utilizados por vários grupos de consumidores de dados, como afirmam Strong, Lee e Wang (1997), não levando em consideração a finalidade, o público, interesse ou licença.

Figura 33- Publicação de Dados na *Web*

Fonte: adaptado pela autora a partir de LÓSCIO; BURLE; CALEGARI (2017, p. 33)

Visto que existe uma heterogeneidade e levando-se em conta que os provedores de dados e os consumidores de dados podem não acessar do mesmo modo ou até conhecê-lo, verifica-se a necessidade de fornecer algumas características sobre os conjuntos de dados e suas distribuições, gerando processo de contribuição para a confiabilidade e reutilização dos dados. Essas características são: metadados estruturais, metadados descritivos, acesso à informação, informação sobre a qualidade de dados, informações sobre a procedência, informações sobre licença e informações sobre uso.

Lóscio, Burle e Calegari (2017) propuseram o conjunto de Boas Práticas, uma vez que elas oferecem diretrizes técnicas para a publicação de dados na *Web*, abrangendo diferentes desafios uma vez que estarão relacionadas com publicação e o consumo desses dados, como formatos de dados, acesso a dados identificadores de dados, vocabulários e metadados.

No documento de casos de uso de dados na *Web* definido por Lee (2014), a relevância da boa prática é evidenciada por requisitos, que aborda por pelo menos uma boa prática e gerando assim as diretrizes do documento composto por Lóscio, Burle e Calegari (2017).

Então o quesito de publicar os dados na *Web* é muito além do que apenas publicar ou permitir o seu reuso, tem relação com questões:

- Como disponibilizar os dados?
- Como deixar os dados interoperáveis?
- Como identificar os recursos dos dados?
- Como receber um retorno?

- Quais formatos usar?
- Quais são as fontes dos dados?
- Quais dados publicar?

Devido a isso, Lóscio, Burle e Calegari (2017) apresentam 12 desafios para publicar os dados na *Web* e ainda contempla com diretrizes e restrições importantes dentro do documento das Boas Práticas.

Os desafios para publicar os dados na *Web* estão relacionados com Metadados e como isso pode ser visto e interpretado por humanos e máquinas, como já apresentado.

Em relação à Licença de Dados, alguns pontos são importantes e necessários:

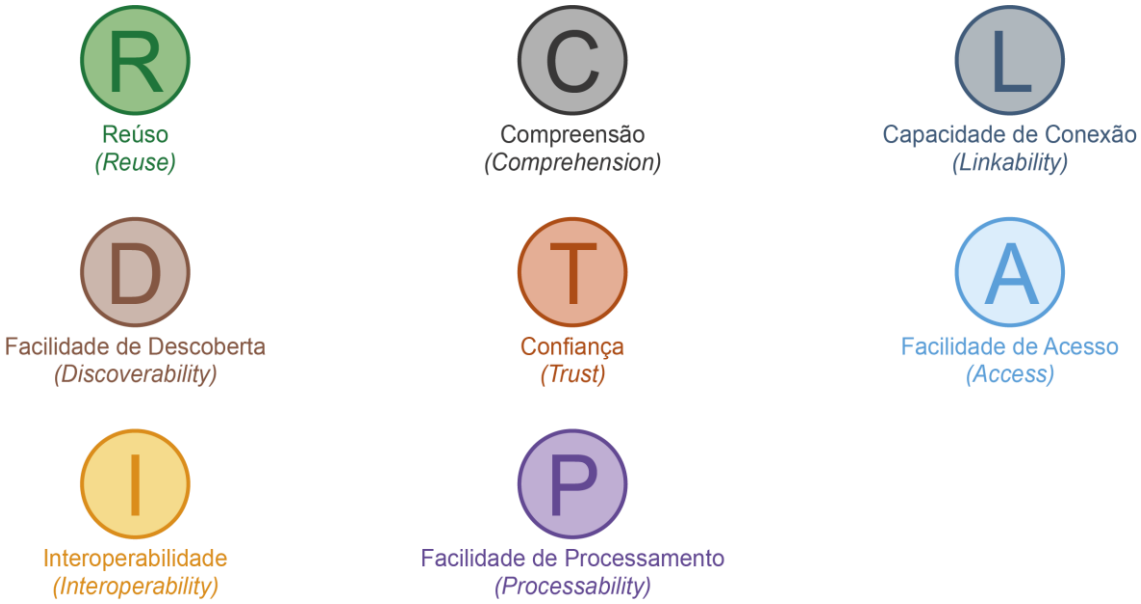
- gerando conteúdo que aponta como permitir e restringir o acesso a esses dados;
- Proveniência e Qualidade dos Dados, apresentando características relacionadas à confiança;
- Versionamento dos Dados que permite o rastreio das versões dos conjuntos dos dados; a Identificação dos dados, que está relacionada com identificação dos conjuntos de dados e suas distribuições;
- Formatos dos dados e quais formatos usar;
- Vocabulários dos Dados e como promover a interoperabilidade entre eles;
- Acesso aos Dados e suas opções de acesso;
- Preservação dos Dados;
- *Feedback*,
    - trazendo o retorno e como: engajar usuários; o processo do Enriquecimento dos Dados, que adiciona valor aos dados e a Republicação dos Dados, que consiste na responsabilidade de reusar os dados.

No documento “*Data on the Web Best Pratices*” ou pela versão em português publicada em outubro de 2019, determinado pela W3C, escrito por Lóscio, Burle e Calegari (2017) descreve as 35 melhores práticas, trazendo os

Benefícios das Boas Práticas que gera uma compreensão sobre os aspectos em que humanos terão um melhor entendimento sobre a estrutura dos dados, seu significado, os metadados e a natureza do conjunto de dados. O processamento dos dados tem relação com as máquinas e podem automatizar o processo e manipular os dados do conjunto de dados. Além de apresentar o descobrimento, ou seja, como as máquinas poderão descobrir automaticamente os dados ou um conjunto de dados para gerar um reuso que permite a possibilidade de aumentar o reuso de conjuntos de dados por distintos grupos de consumidores.

Além de tudo, o documento disponibiliza os benefícios das Boas Práticas para Publicação de Dados na *Web*, onde cada benefício representa uma melhoria na maneira como os conjuntos de dados são disponibilizados na *Web*, como apresentado na **Figura 34.**

Figura 34 - Benefícios na forma como conjuntos de dados são disponibilizados na *Web*



Fonte: adaptado pela autora a partir de LÓSCIO; BURLE; CALEGARI (2017, p. 34-43)

Cada benefício possui suas funcionalidades que serão apresentadas no **Quadro 16**.

Quadro 16 - Benefícios que os publicadores de dados obterão ao optar pela adoção das Boas Práticas

Benefícios ao uso das Boas Práticas	
Reuso	• Todas as Boas Práticas
Facilidade de Acesso	• Fornecer download em massa (*bulk download*) • Fornecer subconjuntos para conjuntos de dados extensos • Usar negociação de conteúdo para disponibilizar dados em formatos múltiplos • Fornecer acesso em tempo real • Fornecer dados atualizados • Disponibilizar dados por meio de uma API • Usar padrões *Web* como base para construção de APIs • Fornecer visualizações complementares
Facilidade de Descoberta	• Fornecer metadados • Fornecer metadados descritivos • Usar URIs persistentes como identificadores de conjuntos de dados • Usar URIs persistentes como identificadores dentro de conjuntos de dados • Atribuir URIs para as versões dos conjuntos de dados e séries • Usar padrões *Web* como base para construção de APIs • Citar a publicação original do conjunto de dados
Confiança	• Fornecer informações sobre a licença de dados • Fornecer informações de procedência dos dados • Fornecer informações de qualidade de dados • Fornecer indicador de versão • Fornecer o histórico de versão • Atribuir URIs para as versões dos conjuntos de dados e séries • Reutilizar vocabulários, dando preferência aos padronizados • Fornecer uma explicação para os dados que não estão disponíveis • Fornecer documentação completa para as APIs • Evitar alterações que afetem o funcionamento de sua API • Preservar identificadores • Avaliar a cobertura do conjunto de dados • Coletar feedback de consumidores de dados • Compartilhar o feedback disponível • Enriquecer dados por meio da geração de novos dados • Fornecer visualizações complementares • Fornecer feedback para o publicador original • Obedecer aos termos de licença • Citar a publicação original do conjunto de dados
Facilidade de Processamento	• Fornecer metadados • Fornecer metadados estruturais • Usar formatos de dados padronizados legíveis por máquinas • Fornecer dados em formatos múltiplos • Reutilizar vocabulários, dando preferência aos padronizados • Fornecer subconjuntos para conjuntos de dados extensos • Disponibilizar os dados por meio de uma API

	• Usar padrões *Web* como base para construção de APIs • Enriquecer dados por meio da geração de novos dados
Interoperabilidade	• Usar URIs persistentes como identificadores de conjuntos de dados • Usar URIs persistentes como identificadores dentro de conjuntos de dados • Reutilizar vocabulários, dando preferência aos padronizados • Escolher o nível de formalização adequado • Disponibilizar dados por meio de uma API • Usar padrões *Web* como base para construção de APIs • Evitar alterações que afetem o funcionamento de sua API • Fornecer feedback para o publicador original
Facilidade de Conexão	• Usar URIs persistentes como identificadores de conjuntos de dados • Usar URIs persistentes como identificadores dentro de conjuntos de dados • Fornecer subconjuntos para conjuntos de dados extensos • Usar padrões *Web* como base para construção de APIs
Compreensão	• Fornecer metadados • Fornecer metadados descritivos • Fornecer metadados estruturais • Fornecer informações de procedência dos dados • Usar representações de dados que sejam independentes de localidade (*locale neutral*) • Reutilizar vocabulários, dando preferência aos padronizados • Escolher o nível de formalização adequado • Coletar *feedback* de consumidores de dados • Enriquecer dados por meio da geração de novos dados • Fornecer visualizações complementares

Fonte: desenvolvido pela autora a partir de LÓSCIO; BURLE; CALEGARI (2017, p. 34-43)

4.3.3 Processos e Técnicas para a Publicação de Dados na *Web*

Em se tratando de divulgação e Publicação de dados na *Web*, há uma complexidade envolvida na sua elaboração. Empiricamente falando, realizar uma determinada tarefa para que posteriormente seja publicada, não se resume somente na maneira pela qual será exposta, porém existem métodos que orientam na estruturação que vão além de uma simples veiculação de informações.

Nesse sentido, as boas práticas de publicação de dados na *Web*, são diretrizes que auxiliam na elaboração de subsídios do que se pretende publicar para o seu posterior reuso. E quando algo é disponibilizado com qualidade, desperta curiosidade para ser explorado ou até utilizado.

A W3C, em conjunto com Lóscio, Burle e Calegari (2017), como já apresentado na seção anterior, dispõe de 35 (trinta e cinco) recomendações denominadas como melhores práticas (*best practice*) divididas em 12 (doze) categorias, como já apresentadas anteriormente, que norteiam a organização de dados coerente para obter-se o melhor resultado em uma futura publicação.

A seguir, será abordado sucintamente, baseado em Lóscio, Burle e Calegari (2017), acerca de cada boa prática para compreender-se melhor o que significa a sua elaboração e como isso reflete sobre os dados, quais são suas técnicas e sua respectiva publicação.

1. **Metadados / 2. Fornecer metadados descritivos/ 3. Fornecer metadados estruturais**

A primeira boa prática, nomeada como Metadados, refere-se ao fornecimento de informações que colaboram para o entendimento do significado dos dados, auxiliando na realização de tarefas. Esta categoria é fragmentada em 3 (três) partes, designadas como, melhor prática 1: orienta o fornecimento de metadados que servirá tanto para uma pessoa, ou para uma máquina para que entendam a informação que foi disposta na *Web*.

A melhor prática 1 aconselha o provimento de metadados descritivos, como o próprio nome diz, serão informações que, através das quais, os usuários conseguirão compreender as características do conjunto de dados.

Referente à primeira categoria, a melhor prática 3 recomenda-se o fornecimento de metadados estruturais, isto é, subsídios organizados em partes que compõem um conjunto de dados para que ele seja explorado ou consultado, facilitando o entendimento.

Os benefícios ligados a estas 3 melhores práticas são: Reuso, Compreensão, Facilidade de Descoberta e Facilidade de Processamento.

4. Fornecer informações sobre a Licença de Dados

A boa prática, denominada como Licença de Dados, tem como finalidade fornecer informações descritivas de conjuntos de dados de forma explícita, e possibilita que agentes de software descubram automaticamente conjuntos de dados disponíveis na *Web*, permitindo que pessoas possam compreender a natureza do conjunto de dados e suas distribuições.

Os benefícios ligados a essa melhor prática são: Reuso, Confiança.

5. Fornecer informações de procedência dos dados

Dando sequência aessas recomendações, a quinta boa prática denomina-se como Procedência de Dados e orienta justamente esse aspecto, o fato de saber a origem da informação para poder conjecturar a qualidade dele, se é confiável ou não.

Os benefícios ligados a essa melhor prática são: Reuso, Compreensão e Confiança.

6. Fornecer informações de qualidade de dados

A sexta boa prática titulada como Qualidade de Dados implica no modo de como o subsídio será utilizado. Dessa forma, a melhor prática concerne no aspecto do fornecimento de tais informações que facilitem o processo de seleção do conjunto de dados, pois documentar com qualidade é primordial.

Os benefícios ligados a essa melhor prática são: Reuso e Confiança.

7. Fornecer indicador de versão

A sétima boa prática designada como Versão de Dados, ou indicador de Versão, está correlacionada com a necessidade de obterem-se novas versões, pois ao longo do tempo podem mudar, por isso é preciso criar e disponibilizar essas novas versões. A melhor prática 7 para esse quesito, orienta que haja um indicador da versão em cada conjunto de dados para que sejafacilmente verificada, ou seja, se a versão mais recente está disponível.

Os benefícios ligados a essa melhor prática são: Reuso e Confiança.

8. Fornecer o histórico de versão

A melhor prática 8 diz respeito ao histórico da versão que informa as mudanças realizadas em cada uma das versões de dados descritas na prática 7.

Os benefícios ligados a essa melhor prática são: Reuso e Confiança.

9. Identificadores de Dados -Usar URIs persistentes como identificadores de conjuntos de dados / 10. Usar URIs persistentes

como identificadores dentro de conjuntos de dados/ 11. Atribuir URIs para as versões dos conjuntos de dados e séries

A melhor prática 9 é denominada como Identificadores de Dados que ajudam na localização de informação. Está segmentada em 3 (três) partes, e recomenda que o uso de URI (Identificador Uniforme de Recursos) seja constante, possibilitando a comparação e o gerenciamento de informação com confiabilidade.

A melhor prática 10 é semelhante à 9; contudo, diferencia-se na maneira de usá-la. Neste caso, o identificador é dentro do conjunto de dados para possam ser mencionados por outros conjuntos. Por isso, é preponderante manter as URIs constantes.

E a melhor prática 11 orienta que seja atribuída URI para versão e séries também. Assim, uma pessoa ou um software poderão citar especificamente em algum momento que se deseja utilizar.

Os benefícios ligados a essa melhor prática são: Reuso e Capacidade de Conexão, Facilidade de Descoberta, Interoperabilidade e Confiança.

12. Usar formatos de dados padronizados legíveis por máquinas / 13. Usar representações de dados que sejam independentes de localidade (*locale neutral*) / 14. Fornecer dados em formatos múltiplos

A décima segunda boa prática é denominada como Formatos de Dados e é indispensável para que os dados possam ser usados. Referente a este quesito, há três (3) três partes como a melhor prática 12 aconselha o uso padronizado de informação para que seja legível por uma máquina, pois caso não haja padronização torna-se ineficiente.

A melhor prática 13 recomenda o uso de representações locais neutras para facilitar a manipulação de subsídios e a melhor prática 14 instrui o fornecimento de dados em vários formatos, isto é, se algum dado for alterado por mais de uma pessoa, esses vários formatos ajudam a evitar erros, além de economizar tempo e dinheiro.

Os benefícios ligados a essa melhor prática são: Reuso, Facilidade de Processamento, Compreensão.

15. Vocabulário de Dados / Reutilizar vocabulários, dando preferência aos padronizados

A boa prática 15, denominada como Vocabulário de Dados, estabelece termos que podem ser empregados em algum aplicativo específico e norteia a reutilização do vocabulário e dá preferência aos dados padronizados.

Os benefícios ligados a essa melhor prática são: Reuso, Facilidade de Processamento, Compreensão, Confiança e Interoperabilidade

16. Escolher o nível de formalização adequado

A melhor prática 16 sugere a escolha do nível de formalização correta, ou seja, simples, mas de qualidade, implicando um nível de semântica formal que se ajuste tanto aos dados quanto às aplicações mais prováveis de serem utilizadas.

Os benefícios ligados a essa melhor prática são: Reuso, Compreensão e Interoperabilidade.

17. Fornecer download em massa (bulk download)/ 18. Fornecer subconjuntos para conjuntos de dados extensos/ 19. Usar negociação de conteúdo para disponibilizar dados em formatos múltiplos/ 20. Fornecer acesso em tempo real/ 21. Fornecer dados atualizados/ 22. Fornecer uma explicação para os dados que não estão disponíveis

A categoria designada como Acesso de Dados concerne com a facilidade das pessoas e máquinas terem acesso e aproveitarem os benefícios, garantindo o fácil acesso aos dados na *Web*, beneficiando-se do compartilhamento de dados e utilizando a infraestrutura *Web*. Esta se divide em seis (6) partes, conforme segue:

- A melhor prática 17 aconselha ao fornecimento de download em grande proporção.
- A melhor prática 18 orienta provimento de subconjuntos e conjuntos de dados grandes, isto é, devido à dificuldade de armazenar e analisar um volume considerável de dados, o usuário pode ter a opção de baixar apenas o subconjunto em vez do conjunto todo.

- Melhor prática 19 norteia usar negociação de conteúdo fornecer dados disponíveis em vários formatos.
- A melhor prática 20 recomenda fornecer acesso em tempo real.
- A melhor prática 21 que provê dados atualizados e, por fim,
- A melhor prática 22 aconselha fornecer uma explicação para informações indisponíveis.

Os benefícios ligados a essa melhor prática são: Reuso, Facilidade de Acesso, Capacidade da Conexão, Facilidade de Processamento e Confiança.

23. Disponibilizar dados por meio de uma API/ 24 Usar padrões *Web* como base para construção de APIs/ 25. Fornecer documentação completa para as APIs/ 26. Evitar alterações que afetem o funcionamento de sua API

Neste mesmo segmento, ainda há algumas sugestões que auxiliam na elaboração de determinada tarefa, que é a categoria intitulada como APIs de Acesso de dados. Seguem as recomendações:

Melhor prática 23 orienta disponibilizar informações por meio de uma API (Interface de Programação de Aplicações); Melhor prática 24 refere-se ao uso de padrões *Web* como base da API; Melhor prática 25 trata-se do fornecimento da documentação completa para API e melhor prática 26, fala sobre evitar quebrar alterações na API, ou seja, qualquer alteração que seja feita, avisar o usuário.

Os benefícios ligados a essa melhor prática são: Reuso, Facilidade de Processamento, Interoperabilidade, Facilidade de Acesso, Capacidade de Conexão, Facilidade de Descoberta, Confiança.

27. Preservar identificadores/ 28. Avaliar a cobertura do conjunto de dados

A categoria nomeada como Preservação de Dados possui as seguintes recomendações de boas práticas referentes a esse aspecto.

Melhor prática 27 instrui preservação dos identificadores, isto é, caso seja removida alguma informação deve haver uma explicação, uma mensagem ao usuário para que ele identifique se a remoção é algo temporário ou

permanente. E a melhor prática 28 avalia a cobertura do conjunto de dados, seu alcance e sua influência.

Os benefícios ligados a essa melhor prática são: Reuso e Confiança.

29. Coletar *feedback* de consumidores de dados / 30. Compartilhar o *feedback* disponível

A categoria denominada como *Feedback,* ou denominada como Comentários apresenta a questão de o usuário ter voz, poder expressar sua opinião, experiência, assim ter o conhecimento de outros feedbacks. Seguem as recomendações relacionadas:

Melhor prática 29 é a agrupação dos comentários de consumidores de dados, ou seja, proporciona algum meio de fácil acesso para sejam apresentadas opiniões, e a Melhor prática 30 aconselha tornar esses comentários disponíveis para que outros usuários tenham acesso.

Os benefícios ligados a essa melhor prática são: Reuso, Compreensão e Confiança.

31. Enriquecer dados por meio da geração de novos dados/ 32. Fornecer visualizações complementares

Enriquecimento de dados refere-se a um conjunto de processos que pode ser utilizado para aperfeiçoar, aprimorar ou outra forma de melhorar dados brutos ou dados previamente processados.

É uma categoria que apresenta o enriquecimento de dados que auxiliam, podendo melhorar informações, ou seja, conceitos semelhantes contribuem para transformar os dados de tornar-se um subsídio ativo valioso.

Melhor prática 31 orienta o desenvolvimento de dados, gerando novos dados, podendo aumentar o valor do mesmo e facilitar o processamento, e melhor prática 32 aconselha fornecer apresentações complementares.

Os benefícios ligados a essa melhor prática são: Reuso, Compreensão, Confiança, Facilidade de Processamento e Facilidade de Acesso.

33. Fornecer *feedback* para o publicador original/ 34. Obedecer aos termos de licença / 35. Citar a publicação original do conjunto de dados

A categoria designada como Republicação, como o nome já diz, refere-se ao fato de publicar novamente. E, na sequência, as três últimas recomendações que dizem respeito a esta categoria:

Melhor prática 33 aconselha fornecer comentários ao editor original para que ele fique informado quando seus dados estão sendo reutilizados; Melhor prática 34 orienta a seguir os termos de licenciamento, isto é, há requisitos que devem ser seguidos para que não haja problema posteriormente e a melhor prática 35 instrui citar obviamente a publicação original e referenciá-la.

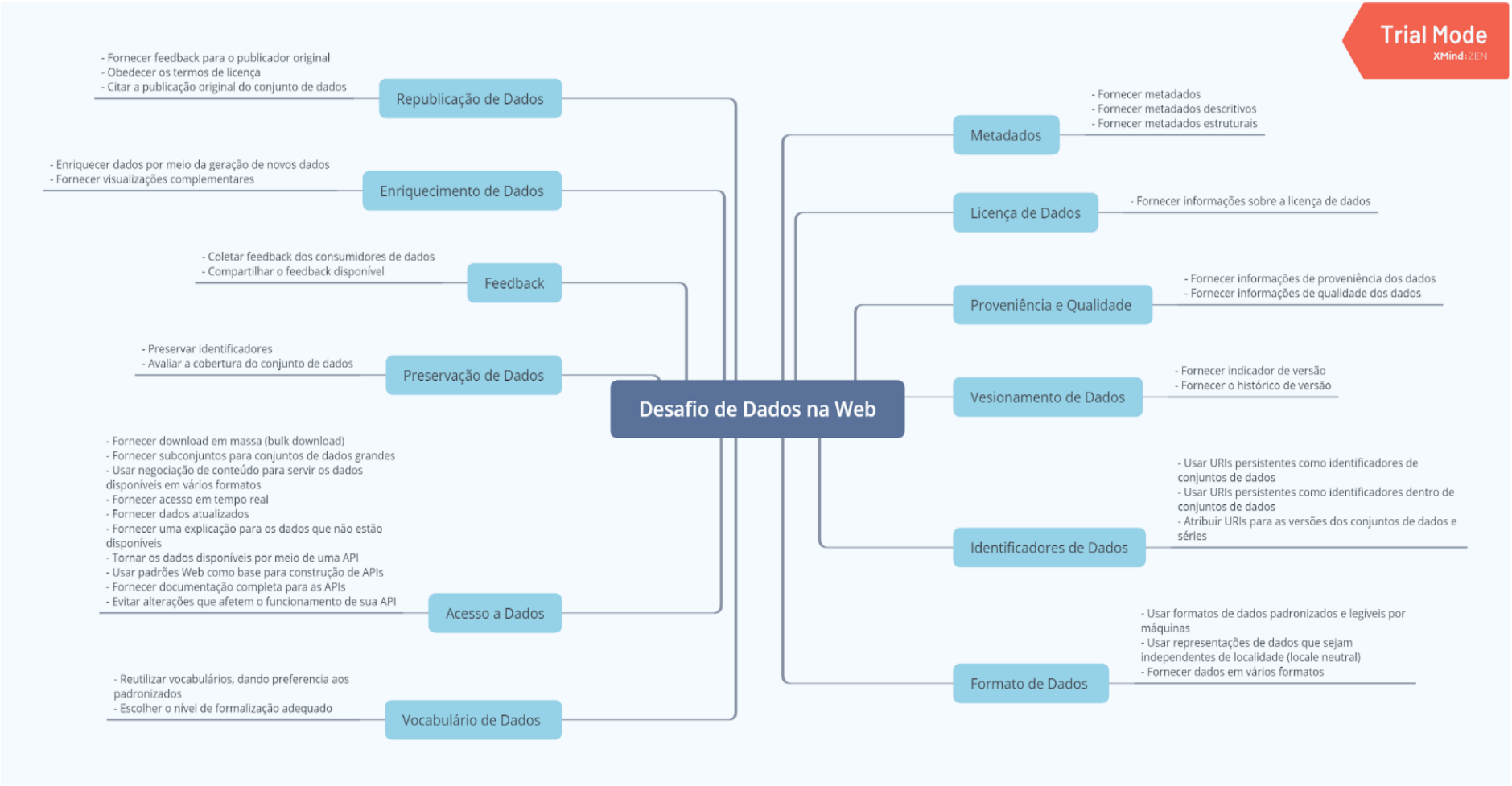
Os benefícios ligados a essa melhor prática são: Reuso, Interoperabilidade, Confiança, Facilidade de Descoberta.

A **Figura 35** e a **Figura 36** apresentam, em forma de grafo, todo o contexto apresentado sobre as categorias e as Boas práticas indicadas nas duas últimas seções.

Figura 35 - As 12 categorias referentes às Boas Práticas de Publicação de Dados na *Web*

Fonte: CONEGLIAN; LUZ; SANTAREM SEGUNDO (2017, p. 6)

Figura 36 - Principais desafios enfrentados ao publicar ou consumir dados na *Web*



Fonte: adaptado pela autora a partir de LÓSCIO; BURLE; CALEGARI (2019, não paginado)

4.3.4 Ferramentas e *Framework*s de Suporte para a publicação de dados na *Web*

Esta Seção apresenta e analisa algumas ferramentas e *Frameworks* que dão suporte à publicação de dados na Web, apresentando algumas aplicações de suporte à *Web* Semântica.

Eller (2014) afirma que existem várias características desejáveis e que devem ser levadas em consideração para se fazer a seleção entre as ferramentas de anotação semântica existentes. As características estão relacionadas com seis propriedades principais: Tipo de Ferramenta, quanto às ontologias, quanto às anotações, documentos, arquitetura, Interface, Interoperabilidade e documentação.

Na categoria Anotação Semântica, como tratado anteriormente, algumas aplicações existem e que permitem gerar esse contexto. Holmes e Kogut (2001) afirmam que existem pelo menos três tipos de ferramentas que podem ser utilizadas para anotação semântica: semiautomáticas, automáticas e híbridas. Semiautomáticas se associam a palavras do texto a classes, instâncias e propriedades da ontologia, utilizando-se da percepção humana; automáticas são técnicas de processamento de linguagem natural (PLN), aprendizado de máquina e extração de informação, que tem como objetivo associar palavras à ontologia que podem ser padrões ou ontologias de domínios específicos e, por último, as híbridas, ou seja, que usa das duas vertentes de automática e semiautomáticas combinadas em uma só ferramenta (HOLMES; KOGUT, 2001).

Reeve e Han (2005) afirmam que anotações semânticas manuais são vistas como processos cansativos, pois devem considerar o volume de documentos existentes na *Web* que precisam ser anotados para se adequarem à *web* semântica. Popov et al. (2003) afirmam que a identificação e classificação totalmente automática das entidades de um documento, trazem mais vantagens e já possuem sistemas, sendo desenvolvidos com um bom desempenho.

Para esse contexto, a maioria das ferramentas adota o contexto apresentado por Reeve e Han (2005), onde as plataformas são baseadas em padrões e podem fazer a descoberta de novos padrões ou executar com

padrões definidos manualmente. A maior parte dos métodos padrão descoberta utiliza um conjunto inicial de entidades definido e o documento é varrido na busca dos padrões em que as entidades aparecem. Já nas Plataformas baseadas em aprendizado de máquina utilizam dois métodos, o probabilístico e de indução, que podem utilizar modelos estatísticos para predizer os locais das entidades no texto.

Para uma melhor compreensão foram divididas em categorias como: Anotação, Armazenamento, *Application Programming Interface* (API), Consulta, Edição, Inferência, Integração, Validação e Visualização. São categorias que apresentam relevância na sua categorização, não criando uma obrigatoriedade que somente uma ferramenta ou aplicação possa pertencer a um único contexto, na sua maioria elas podem pertencer a mais de uma.

A escolha do tipo de ferramenta que faz a anotação semântica está ligada com as necessidades dos usuários, do seu conhecimento em relação ao domínio da aplicação e quais as ontologias existentes para suprir o domínio, esse mapeamento das ferramentas descritas a seguir, estão relacionados com o levantamento dos trabalhos correlatos apresentados no capítulo 2, todas apresentadas nesses trabalhos.

DOSE - *Distributed Open Semantic Platform-* é um projeto que começou em 2003. Desde 2005, a *Intellisemantic* s.r.l. (INTELLISEMANTIC, 2020) está colaborando ativamente com este projeto no que diz respeito à otimização de desempenho e desenvolvimentos futuros. DOSE é uma plataforma distribuída para elaboração semântica que fornece serviços semânticos, como anotação automática de recursos da *Web* no nível da subestrutura do documento, recursos de pesquisa semântica, armazenamento e recuperação de anotação semântica. Seus dados, licenças e versões se encontram disponíveis no site referente ao projeto (DOSE, 2020).

MnM é uma ferramenta que dá suporte semiautomático e automático dirigido por Ontologia para *Web* Semântica (MNM, 2004). É uma ferramenta de anotação que fornece o suporte automatizado e semiautomatizado para anotação de páginas *Web* com conteúdo semântico. A última versão é de 2004, o site disponibiliza todos os dados, licenças e versões referentes ao projeto.

OntoMat Annotizer, segundo (PETRIDIS *et al.,* 2006), é uma ferramenta que permite a anotação de uma página *Web* em OWL. Ela auxilia as tarefas de criação e manutenção de ontologias OWLs, seus atributos e relações. Inclui um navegador para explorar a ontologia e expor partes anotadas do texto.

Petridis *et al.* (2006) afirmam que O M-*OntoMat-Annotizer* permite a exclusão de descrições visuais MPEG-7 de baixo nível para ontologias e anotações convencionais da *Web* Semântica e utilizam o M-*On-Mat-Annotizer* na ordem para construir ontologias que incluem instâncias prototípicas de conceitos de domínio de alto nível, juntamente com uma especificação formal dos descritores visuais correspondentes, formalizando a inter-relação das descrições dos conceitos multimídia de alto e baixo nível, permitindo novos tipos de análise, raciocínio e recuperação de conteúdo multimídia.

Handschuh, Staab e Ciravegna (2002) mostram a possibilidade de acessar ontologias especificadas em mais de um tipo de marcação, como RDF e DAML + OIL. Mas cada ontologia somente pode ser acessada individualmente, além de poder armazenar páginas anotadas em DAML + OIL usando OntoBroker (ontobroker.semanticweb.org/), que é um servidor de anotações e ainda oferece indexadores que podem pesquisar na Internet páginas anotadas para adicioná-las à sua base de conhecimento.

PhotoStuff (HALASCHEK-WIENER *et al.,* 2005) é ferramenta para anotação de imagens, permitindo ao usuário anotar partes da imagem de acordo com conceitos definidos em ontologias. Além disso, é possível publicar o metadado gerado automaticamente.

Halaschek-Wiener *et al.,* (2005) ainda afirmam que fornecem funcionalidade para anotar imagens manualmente usando ontologias da *Web*, além de explorar metadados de imagens incorporadas pré-existentes para anotação automática. Por fim, o PhotoStuff é fracamente acoplado a um portal da *Web* Semântica que fornece gerenciamento de metadados de imagem e funcionalidade de interação.

SemanticWord, segundo Tallis (2003), é uma ferramenta que permite efetuar associações entre marcações OWL e regiões específicas em documento do editor de texto Word.

Tallis (2003) ainda afirma que *SemanticWord* é um ambiente baseado no MS Word, que integra a autoria de conteúdo e marcação, fornecendo ferramentas personalizáveis que permitem a geração simultânea de conteúdo e anotações semânticas, um esquema de anotação que permite que as anotações sejam reutilizadas quando o conteúdo é reutilizado. Biblioteca de modelos personalizáveis contendo texto parcialmente anotado e um sistema automático de extração de informações com as ferramentas para refinar e aumentar sua saída.

A plataforma **KIM**, desenvolvida pela DERI (*Digital Enterprise Research Institute*) na Universidade Nacional da Irlanda em *Galway*, envolvendo também o GATE research team, bem como a OntoText laboratório de Sirma AI Ltd. e apresentada por POPOV *et al.* (2003) é uma parte do projeto SWAN (*Semantic Web ANnotator*), e consiste em uma ontologia KIM, que gera uma base do conhecimento KIM, denominado como KIM Server, ou como uma API para acesso remoto ou incorporação).

KIM emprega uma técnica NLP para fazer a extração da informação, que se baseia no GATE (*General Architecture of Text Engineering*) para extrair, indexar e anotar dados instanciados. Na base de conhecimento, são armazenados recursos com uso do Sesame[27], uma arquitetura para armazenamento de grandes quantidades de metadados, descritos em RDF e RDF Schema (BROEKSTRA; HARMELEN; KAMPMAN, 2002).

Também é vista como parte de suporte, ou parte de suporte de termos generalizados que se referem às etapas inicial e final do processo, classificados como *front-ends*, e fornecem total acesso às funcionalidades do KIM Server ao usuário.

Ela define as classes, entidades e as relações de interesse referente a essa ontologia e foi dividida em duas partes: PROTON[28] (JAIN *et al.,* 2011), uma ontologia de alto nível genérica, e os módulos específicos da plataforma KIM, KIMSO e KIMLO.

---

[27] http://www.openrdf.org/

[28] A PROTON faz algumas distinções filosóficas, e descreve mais a fundo algumas entidades de importância geral (encontros, conflitos militares, governos e organizações etc.). A intenção é que a ontologia seja utilizável para anotações de propósito geral, e seja fácil de estender para domínios específicos.

O **SMORE**, desenvolvida pela *The Semantic Web Research Group na University of Maryland Institute for Advanced Computer Studies*; que, segundo Kalyanpur *et al.* (2006), tem como objetivo gerar para o usuário um ambiente flexível, que permita que usuários consigam anotar os documentos em RDF usando ontologias, e fazer associações com termos específicos do documento com os elementos da ontologia, ou seja, que possa ser criado uma página na *Web* que envolva a anotação dos documentos, sem muitos problemas a acessar, sendo capaz de classificar semanticamente os dados do documento, fazendo a correta associação sujeito-predicado-objeto entre os elementos do documento e as suas referências nas ontologias existentes na Internet.

Kalyanpur *et al.* (2006) ainda afirmam que a ontologia pode ser criada desde o início, aproveitando de conteúdos de ontologias já existentes, se baseando pelas normas da W3C e utilizando a linguagem Owl.

Annotea, desenvolvida pela LEAD (*Live early Adoption and Demonstration*) e faz parte da W3C para gerar a *Web* semântica, classificada como código aberto e contribuindo para o avanço das normas W3C.

Segundo Kahan e Koivunen (2001), a ferramenta permite a criação de anotações com metadados, gerando assim um ambiente compartilhado de anotações, e as armazena em uma base de dados RDF e a comunicação entre a cliente e servidor utiliza métodos HTTP, além de poderem ser associados a qualquer documento da *Web* ou apenas a uma parte selecionada dele, sem precisar mudar a sua estrutura.

Laufer (2015) apresenta no Guia de *Web* semântica três plataformas que são muito utilizadas para Publicação de Dados na *Web*, que são CKAN, Socrata e Junar.

CKAN (*Comprehensive Knowledge Archive Network*) (CKAN, 2020) que é um software livre para catalogação de dados abertos e desenvolvido pela *Open Knowledge*[29] e está disponibilizado e é *Open-source*[30], está em desenvolvimento desde 2007 e já tem disponibilidade em mais de 50 idiomas.

É uma plataforma de código aberto que serve tanto para publicação, como compartilhamento, pesquisa e uso de dados. Permitem a publicação utilizando-se de interface *Web* ou de uma API, com dados de características

[29]http://okfn.org
[30] https://github.com/ckan/ckan

privadas ou públicas (formatos: CSV, JSON, XML, RDF) gerando informações para que possam em um momento serem recuperadas a qualquer momento, permitindo guardar dados e metadados e poder linkar dados diretamente permitindo uma busca fácil de dados.

Suas principais funcionalidades, segundo (CKAN, 2020), estão relacionadas à catalogação pela interface *Web*, pela API ou por ferramentas de importação; fazer pesquisas em todos os metadados, gerando buscas avançadas e organizando as etiquetas (*tags*), formato, licença. Sua organização se baseia em *datasets* (conjuntos de dados) e recursos que sejam relacionados ou similares, mesmos dados em diferentes formatos, períodos distintos, entre outras características.

Socrata (SOCRATA, 2019) é uma plataforma de dados que permite que os governos usem os dados como um ativo estratégico no design, gerenciamento e entrega de programas. Os dados fluem facilmente entre funcionários e departamentos, levando a programas mais eficientes e tomada de decisão com maior assertividade. Serve também para publicação, compartilhamento, pesquisa e consumo de dados. Diferente do CKAN, ela tem como base tecnologias mais recentes, e utiliza MongoDB[31] e *Elasticsearch*[32] e disponibiliza uma API para a publicação e consumo de dados. O Socrata disponibiliza ainda um conjunto de bibliotecas e SDK para o desenvolvimento de clientes em várias linguagens, tais como: R, PHP, Java, Objective-C e outros, além de existirem outras duas versões para o Socrata, uma *Open Source* e outra para uso privado.

Junar (JUNAR, 2020) é uma plataforma de dados que permite transformar seus ativos de dados difíceis de encontrar e inúteis em tabelas dinâmicas, visualizações, mapas, painéis e APIs - para que cidadãos, desenvolvedores e empresas possam reutilizá-los para seus interesses de uma maneira simples, porém é uma plataforma privada, de código fechado e pago e que, segundo Laufer (2015), afirma que cobre todas as etapas do processo de

[31] O MongoDB é um banco de dados distribuído, baseado em documentos e de propósito geral, desenvolvido para desenvolvedores de aplicativos modernos e para a era da nuvem (MONGODb, 2020).

[32] Elasticsearch é um servidor de buscas distribuído baseado no Apache Lucene. Foi desenvolvido por Shay Banon e disponibilizado sobre os termos Apache License. ElasticSearch foi desenvolvido em Java e possui código aberto liberado sob os termos da Licença Apache (ELASTIC, 2020).

publicação de dados. Além de todas as plataformas descritas, os *Frameworks* também se encaixam para dar suporte na no ciclo de publicação de dados na *Web*.

Para a atividade de extração Rautenberg *et al.* (2018) elencam as ferramentas: OpenLink Virtuoso Sponger, DBpedia Spotlight, PoolParty Thesaurus Server (PPT), DR2, R2R, Apache Stanbol e CSVImport, relacionados à atividade de extração do ciclo de vida de dados conectados de publicação de Dados.

O **Quadro 17** apresenta o *Framework* e suas características principais:

Quadro 17 - *Frameworks* para atividades de extração para Publicação de dados na *Web*

Framework	**Características**
OpenLink Virtuoso Sponger (VIRTUOSO SPONGER, 2020)	● versão licença de código aberto ● middleware da ferramenta Virtuoso ○ transforma dados em RDF (formato semântico) ● Nome: RDFizers
DBpedia Spotlight (DBPEDIA SPOTLIGHT, 2020)	● Licença de código aberto ● extrai informações do Wikipedia ● Converte dados em RDF ● solução para ligar informação não estruturada em dados abertos conectados ● geração de anotações em: ○ HTML ○ RDFa ○ XML ○ JSON ● Anotações podem ser filtradas para diversos tipos: ○ ontologia DBpedia ○ entidades ○ Consultas *Sparql*
PoolParty Thesaurus Server (PPT) (POOLPARTY, 2020)	● Alinhamento de Tesauros ○ criar mapeamento entre conceitos ● Listas Dinâmicas ○ listas Sparql pré-defnidas ● Coleção SKOS
DR2	● Componente da plataforma D2RQ ● Publicação de Dados ○ vindo de banco de dados de Grafo ● Consultas: ○ SPARQL ○ RDF Dumps ○ BD conectados na *Web*
R2R	● Linguagem de Mapeamento (R2R) ○ correspondência de vocabulários RDF diferentes ○ Transforma dados da *Web* para um determinado

	vocabulário.
Apache Stanbol (APACHE STANBOL, 2010)	● Conjuntos de componentes reusáveis para manutenção de conteúdo semântico. ● Resultados retornam em: ○ RDF ○ JSON ● Serviços: ○ enriquecimentos de conteúdo ○ Inferências para recuperação da Informação semânticas adicionais ○ Modelos de conhecimento ○ Persistência
CSVImport (AKSW, 2020)	● Plug-in da ferramenta Ontowiki ● Mapeamento de arquivos CSV para RDF

Fonte: RAUTENBERG *et al.* (2018) adaptada pela autora

Em relação à Atividade e Armazenamento/Consulta RDF que são responsáveis pelo gerenciamento dos dados na *Web*. Os *Frameworks* disponíveis estão disponibilizados no **Quadro 18**.

Quadro 18 - *Frameworks* para Atividade e Armazenamento/Consulta RDF

Framework	**Características**
Open Link Virtuoso RDF Store (VOS.VOSRDFFAQ, 2020)	● Armazenar e realizar consultas RDF ● Armazenamento e Acesso persistente em Grafos RDF ● Armazenar dados em XML
SparQLed[33]	● Projeto de código aberto ○ empresa: SindiceTech ● Escrita simplificada de consultas SPARQL
Sparqlify[34]	● Reescrita de consultas SPARQL-SQL ● define visões RDF em BD Relacionais e gera consultas SPARQL ○ SGDB PostGree que suporta essas funções
SIREn - Semantic Information Retrieval Engine (DELBRU, 2009)	● recuperação de Entidades para *Web* ○ indexar e pesquisar dados semiestruturados na *Web* de Dados ● Desempenha técnicas de Recuperação da Informação: ○ escalabilidade ○ atualizações incrementais ○ pesquisas relacionadas à consulta Espaço-Textual ○ Desempenho eficiente

Fonte: adaptada pela autora a partir de RAUTENBERG *et al.* (2018)

[33] https://sparqled.com/?v=19d3326f3137

34 https://aksw.org/Projects/Sparqlify.html

Segundo Rautenberg *et al.* (2018), as Atividades de revisão Manual e autorias de recursos na *Web* são ferramentas que facilitam a autoria de bases de conhecimento semânticas enriquecidas por meio de *Frameworks*. Os *Frameworks* disponíveis estão disponibilizados no **Quadro 19**.

Quadro 19 - *Frameworks* para Atividade de Revisão Manual e Autorias na *Web*

Framework	**Características**
OntoWiki (ONTOWIKI, 2019)	● *sites Web* para compilação e compartilhamento de informação de conhecimento ● Construção de Ontologias ○ OWL ● Informação estruturada em formato de mapa ○ visões diferenciadas sobre os dados ● Controle de acessos e conteúdos
RDFAuthor (HOFFMANN, 2020)	● Solução de Edição de conteúdo distribuído na *Web* ○ de forma estruturada ○ escondendo detalhes do usuário, sobre transformação RDF ● Componentes de editores inteligentes ○ Usa uma estrutura Semântica para recuperar as informações adicionais sobre elementos a partir do esquema utilizado ● Fontes de Dados suportadas: ○ RDF in attributes ■ RDFa ■ SPARQL
PoolParty Thesaurus Server (POOLPARTY, 2020)	● gerenciar metadados e dados conectados ● modelos semânticos (taxonomia, ontologias e entre outros) ● APIs - Integração de Tecnologias semânticas com outros sistemas ○ motores de busca ○ *Content Managament System*

Fonte: Adaptado pela autora a partir de RAUTENBERG *et al.* (2018)

Em relação à Atividade de Interligação/ Fusão, diz respeito ao processo semiautomatizado referente à manutenção e criação das ligações para permitir coerência e facilidades da integração dos dados, segundo Rautenberg *et al.* (2018). O **Quadro 20** apresenta os *Frameworks* responsáveis por essas tarefas.

Quadro 20 - *Frameworks* para Atividade de Interligação/ Fusão

Framework	**Características**
LIMES (*LInk discovery Framework for MEtric Spaces*) (GEORGARLA, 2019)	● Descobre links entre recursos na *Web*
Silk (ISELE *et al.*, 2020)	● Código aberto ● Integração de fontes de dados heterogêneos ○ paradigma de dados conectados ■ RDF modelo de dados estruturado ■ ligações RDF: entidades e fontes de dados diferentes
SIEVE (SZYMANSKI; KRAWCZYK; DEPTUŁA, 2013)	● Ferramenta de Fusão ● Avaliação da Qualidade de dados conectados

Fonte: Adaptado pela autora a partir de RAUTENBERG *et al.* (2018)

A Atividade de Classificação/ Enriquecimento diz respeito a que as instâncias de dados brutos necessitam ser conectados e integrados com ontologias de nível superior, segundo Rautenberg *et al.* (2018).

O processo de Enriquecimento tem como objetivo recuperar essas informações, ou seja, dados RDF que contém referências implícitas para outros tipos de dados, tornando-as mais óbvias.

O *Framework* DL- Learner (DL-LEARNER, 2015) contém vários algoritmos para aprendizado de máquina supervisionado, usando dados estruturados como conhecimento de segundo plano, e pode usar vários formatos RDF e OWL como entrada, podendo se conectar a ontologias OWL mais populares e é fácil e flexível de configurar. Estende as idéias de Programação em Lógica Indutiva e Aprendizagem Relacional à *Web* Semântica, a fim de permitir analisar dados em bases de conhecimento ou ajudar na construção de bases de conhecimento a partir de dados.

A Atividade de Análise de Qualidade que está relacionada a verificar a qualidade das questões relacionadas aos dados, segundo Rautenberg *et al.* (2018).

A ferramenta ORE (*Ontology Repair and Enrichment*[35]) permite que os engenheiros do conhecimento promovem uma ontologia de OWL, corrigindo inconsistências e fazendo sugestões para adicionar outros axiomas a ela.

A Atividade de Evolução e Reparação geram mudanças, que nas bases do conhecimento, vocabulários e ontologias devem ser explícitas e reconhecidas, conforme apresenta Rautenberg *et al.* (2018).

A ferramenta OpenRefine (OPENREFINE, 2020) é uma ferramenta poderosa para trabalhar com dados: limpando-os; transformando-o de um formato para outro (dados primários em RDF); e estendendo-o com serviços da *web* e dados externos. O *OpenRefine* sempre mantém seus dados privados em seu próprio computador até existam necessidades de compartilhar ou colaborar. Seus dados privados nunca saem do computador, funciona executando um pequeno servidor e usa o navegador da *Web* para interagir com ele.

A Atividade de Busca e Navegação e Exploração, como o próprio nome já diz, oferecem interfaces para visualização e utilização de dados RDF, conforme apresenta Rautenberg *et al.* (2018). Algumas das Ferramentas utilizadas nesse processo estão apresentadas no **Quadro 21**.

Quadro 21 - *Frameworks* para Atividade de Busca e Navegação e Exploração

Framework	**Características**
Sigma (ALRABAEE *et al.* 2015)	● Biblioteca JavaScript para gerar grafos ● Facilidade de Publicar redes em páginas *Web* ● Conjunto de diferentes configurações para o desenho das redes
Spatial Semantic Browser (CHEN, 2000)	● Navegador JavaScript para visualizar bases de conhecimento com dimensão espacial ○ Interação com Triple Store via SPARQL.
CubeViz (ABICHT, 2020)	● O CubeViz é um componente do OntoWiki ● fornece navegação facetada em dados estatísticos na *Web* de dados conectados. ● Utiliza o vocabulário RDF DataCube, que é o estado da arte na representação de dados estatísticos no RDF ● Gera um widget de navegação facetada que pode ser usado para filtrar observações interativas a serem visualizadas em gráficos. ○ Com base na estrutura selecionada, o CubeViz oferece tipos e opções de gráficos beneficiários que podem ser

[35] https://aksw.org/Projects/ORE.html

	selecionados pelos usuários.
Facete (*JavaScript SPARQL-based Faceted Search Library and Browsing Widgets*) (HÖFFNER, 2016)	● Facete é composto por uma biblioteca JavaScript para navegação facetada de dados RDF. ● Aplicativo para navegar por dados RDF relacionados a áreas geográficas. ○ oferece filtragem por facetas, rotação e exibição dos dados em um mapa.

Fonte: adaptado pela autora a partir de RAUTENBERG *et al.* (2018)

5. O MAPEAMENTO E ENRIQUECIMENTO SEMÂNTICO DE DADOS PARA PUBLICAÇÃO NA *WEB*

O presente capítulo tem como objetivo elucidar e ligar todo o contexto que envolve critérios ontológicos e semânticos com especificação no processo de Enriquecimento e Mapeamento Semântico na *Web* de Dados.

A atual *Web* de Documentos tem, como pilar fundamental, a possibilidade de seu conteúdo ser compreensível tanto pelos seres humanos quanto por agentes computacionais, como já dito inúmeras vezes no texto, no entanto esse contexto é importante, uma vez que se usa a conceituação da *Web* Semântica para interligação de palavras, o que provê significado (sentido) a todo conteúdo que é publicado na Internet.

Por sua vez, a *Web* Semântica, segundo Berners-Lee, Hendler e Lassila (2001), traz conceitos e determinações que caminham para um uso contínuo de processos inteligentes e de aprendizado de máquina para que informações se tornem cada vez mais acessíveis.

Porém, para que isso se torne possível, é necessário que todos os dados passem por um processo de enriquecimento, ou seja, por um processo em que exista inicialmente uma camada de abstração, a fim de que as diferentes origens não sejam impeditivas para sua publicação. Isso se dá basicamente por meio das anotações dos dados, conhecidas também como Metadados (dados sobre os dados), os quais fornecem uma estrutura comum que permita que os dados sejam compartilhados e reutilizados através dos limites de aplicativos, empresas e comunidades, como apresentado pelo *World Wide Web Consortium* (W3C, 2002).

Gardenfors (2014) afirma ser de extrema importância adicionar uma camada semântica às informações, permitindo, desse modo, uma conceitualização dos dados do mundo real e, por sua vez, uma capacidade de executar a busca e o raciocínio no nível do conhecimento, além de ligá-los a infinitas outras possibilidades de raciocínios gerenciados na *Web*, de forma que a semântica apareça em diferentes níveis, diferentes domínios e em um espaço livre.

Kuipers (2000) afirma que este é o momento que está se tornando realmente funcional em nível de conhecimento, pois se tornam uma fonte

extremamente confiável em muitas aplicações devido ao processo de inferência de informação que se gera.

Esse processo funcional é derivado de classificações e diretrizes importantes para que os dados e informações sejam padronizados de forma estruturada e conectada, com base no “Sistema de 5 Estrelas” criado por Berners-Lee (5 [ESTRELAS], 2012), apresentando conhecimentos técnicos, que envolvem a aplicação semântica na *Web* utilizando modelos e padrões que possibilitam o uso dos dados estruturados disponíveis na *Web*, além de permitir busca e navegação desses dados/informações de forma que softwares possam dar o apoio para essa contextualização (ISOTANI, BITTENCOURT, 2015).

Com a imensa quantidade de tipos de dados disponibilizados na *Web*, a publicação desses dados, por meio de aplicativos desenvolvidos com diferentes linguagens de programação, arquiteturas, sistemas de banco de dados e representação de dados na *Web*, leva-nos a alguns problemas como, por exemplo, o problema da diversidade de dados, que gera um desafio a ser resolvido quando este é referente ao aspecto semântico, e como já foi apresentado no **Capítulo 4**.

5.1 ENRIQUECIMENTO SEMÂNTICO DE DADOS

Enriquecimento semântico, que pode ser visto como o processo de atribuir maior significado aos metadados e dados através da aplicação de recursos auxiliares, com o objetivo de facilitar a compreensão, a integração e o processamento dos dados por pessoas e máquinas. Ou seja, o enriquecimento semântico torna os dados e metadados mais qualificados e descritivos, através do uso da semântica atribuída por recursos semanticamente bem-descritos, como recursos de bases de conhecimento, ontologias, vocabulários existentes, entre outros (LIRA, 2014).

Segundo Costa (2018), enriquecimento semântico diz respeito ao processo de enriquecer conteúdo com contexto de dados por marcação, além de categorizar e/ou classificar os dados uns em relação aos outros e a dicionários ou outras fontes base de referência. Isso significa acrescentar informações contextuais a algum conjunto de dados existente.

Os métodos de enriquecimento geralmente podem ser aplicados em uma abordagem de base para a criação da base de conhecimento. Nessa abordagem, toda a estrutura ontológica não é criada antecipadamente, mas evolui com os dados em uma base de conhecimento. Idealmente, isso permite um desenvolvimento mais ágil de bases de conhecimento. Em particular, no contexto da *Web* de dados vinculados, essa abordagem parece ser uma alternativa interessante aos métodos de engenharia de ontologia mais tradicionais.

5.1.1 Definição

Enriquecer dados conectados é o processo que alimenta a semântica implícita ou oculta incorporada a um conjunto de dados compreensível pelas ferramentas, que é o RDF.

Verificou-se, na seção de Trabalhos Relacionados, como esse campo de estudo está focado principalmente na geração de *links* para recurso semanticamente relacionados, o que na maioria dos casos significa a inserção de instruções de uma determinada ontologia, com o recurso OWL, entre instâncias de recursos diferentes (vocabulário e conjuntos de dados (*datasets*)) ou entre os conceitos correspondentes, como classes ou propriedades de ontologia.

Este último é o principal objetivo do alinhamento ontológico e alinhamento de esquema, pois responder a consultas complexas geralmente requer acesso, e combinando informações de vários conjuntos de dados, que podem ser alcançados pelo processamento de consulta.

Diferentes abordagens para o processamento de consultas através do *Linked Data* são analisadas e autores diversos estudam quão diferentes alternativas de *design* afetam o desempenho e a praticidade de processamento de consultas e definem uma referência para processamento dessas consultas, compreendendo uma seleção de fontes de dados em vários domínios e consultas representativas a partir da perspectiva adotada por Lee (2014).

Criar dados conectados pode ser descrito, de maneira geral, por uma sequência de ações constituída por 5 etapas, conforme descrito em (GUANDALINI; SANTOS, 2018):

(i) identificação e seleção de dados,
(ii) limpeza, anotação e transformação,
(iii) mapeamento,
(iv) interligação e
(v) armazenamento e publicação.

Embora essas etapas integrem o fluxo de triplificação de dados, algumas partes não são obrigatórias e cada etapa pode ser realizada de maneira bem particular e distinta, dependendo do conjunto de dados a ser tipificado e da natureza das informações coletadas.

Segundo Chris Clarke (2009), o enriquecimento semântico pode ser entendido como um recurso projetado para aumentar a riqueza dos dados. O enriquecimento semântico também pode ser visto como o processo de atribuir maior significado aos metadados e dados por intermédio da aplicação de recursos auxiliares, objetivando facilitar a compreensão, a integração e o processamento dos dados por pessoas e máquinas. Ou seja, o enriquecimento semântico torna os dados e metadados mais qualificados, através do uso da semântica atribuída por vocabulários pré-existentes, sinônimos e informações de proveniência.

Para realizar o Enriquecimento Semântico e obter conceitos adicionais, alguns recursos e técnicas são usados, dentre eles podemos citar: Anotação Semântica, Vinculação e Mapeamento de Recursos, além da conversão para modelos de dados semânticos.

Anotação Semântica: Segundo Uren *et al.* (2005), anotação semântica consiste na atribuição de semântica (significado) aos elementos de um esquema de origem de forma manual ou automatizada por meio da adição de informação semântica.

Vinculação e Mapeamento de Recursos: consiste em descobrir links entre as combinações semânticas dos dados e metadados com outros recursos na *Web* de dados. Segundo Sorrentino *et al.* (2013), é muito utilizado para interligar recursos na nuvem LOD.

Conversão para modelos de dados semânticos: Consiste em modelar os dados num formato semântico e estruturado, como RDF/XML, beneficiando sua manipulação por aplicações que consomem esses modelos de dados.

5.1.2 Problemas Encontrados

Isotani e Bittecourt (2015) já apresentam, em seus estudos, os principais problemas encontrados em publicar o dado aberto na *Web*, devido às suas diversas interpretações e à dificuldade do seu reuso.

Quando existe a necessidade de representação de conhecimento, o processo apresentado por Gil (2003), que está relacionado com o processo de descrição, representação e interpretação dos dados, gera muitas discussões. Uma delas está atrelada em descrever os dados e qual vocabulário deve ser usado para aquele determinado domínio.

Então, se um dado aberto for publicado na *Web,* ele deve passar pela diretriz de escolha de um determinado conjunto de termos, de forma que esse conjunto possa descrevê-lo acertada ou apropriadamente. Por consequência, esse conjunto de termos já deve estar contido em algum vocabulário, ou caso não esteja, a determinação de criá-lo.

Isotani e Bittecourt (2015) ainda afirmam que, além dessa escolha, que precisa estar muito bem atrelada e que identifique um conjunto de dados, outra diretriz a ser vista é como serão representados esses dados, de forma que venha aumentar a amplitude do dado sem gerar imprecisão na sua posterior interpretação.

O processo de enriquecimento é outra diretriz, pois assim permite que os novos questionamentos possam ser respondidos de forma mais fácil. Para todos esses processos, apresentando-se a questão de representação onde é possível colocar restrições nos atributos para se possa assumir um determinado papel, de forma clara e sem ambiguidade.

Mecanismos que permitem essas representações são feitos através das representações dos dados, onde devem ser utilizados mecanismos por linguagens que tragam processos visuais ou lógico, ou através de ontologias, Gruber (1993), em que pode ser utilizada a linguagem OWL que, segundo Noy e McGuinness (2001), explicitam as relações entre conceitos, permitindo que tanto pessoas quanto máquinas possam compreender os conceitos que representam os dados abertos disponibilizados e, no contexto da *Web*, é a representação mais adequada.

5.2 Metadados

Visto que, na seção anterior, foram apresentados os problemas encontrados referentes à semântica ou à forma como é feita a interpretação da *Web*, percebe-se que esses problemas vão muito além do que o conceito da *Web* Semântica pode resolver.

Os metadados são usados para descrever um recurso e torná-lo mais acessível para usuários de aplicações, porém a necessidade de ser compreensível está ligada também com as máquinas e, nesse caso, os metadados devem estar estruturados para que essa compreensão chegue de forma correta ao usuário.

Metadados são descritivos, incluindo elementos que atribuem valor agregado ao dado e que podem ser considerados básicos, desde dados que sejam estruturados até dados desestruturados que precisam ser anotados como descrições textuais do conteúdo ou palavras-chave para ajudar a procurar.

Day (2004) afirma que, em geral, os metadados são importantes para uma pesquisa eficaz, permitindo que sejam descobertos baseados em critérios e são úteis no processo de pesquisa, adicionando informações a essa estrutura e organizando-os, criando uma orientação de arquivamento e identificação de informações. Nesse sentido, o uso mais importante para metadados, então, é promover a interoperabilidade, permitindo a combinação de recursos heterogêneos entre plataformas sem perda de conteúdo.

Alguns levantamentos devem ser relevantes quando estão relacionados ao entendimento dos metadados, por exemplo, deve-se saber quais metadados precisam ser aplicados ao conteúdo, como a descrição dessas informações afeta a inferência, ou ainda levantar questionamentos se esses processos apresentados irão melhorar ou dificultar ainda mais a informação ou verificar o que pode ser feito sobre a anotação de conteúdo herdado, entre outros questionamentos apresentados em W3C 2002.

Para cada um dos levantamentos apresentados em W3C (2002), tudo depende dos propósitos para os quais os recursos serão empregados. Para muitos propósitos, os metadados podem se responsabilizar sobre as suas informações, porém em alguns momentos, no corpo do recurso, que não se

pode assumir que esteja estruturado, pode precisar de informações a mais, para que os usuários os encontrem manualmente e ainda se leva em conta que nem toda a informação seja importada de uma única fonte, podendo vir de *datasets* estruturados, como bancos de dados, Knolmayer e Myrach (2001).

As informações de proveniência são extremamente importantes para determinar o valor e integridade de um recurso. Muitos padrões de arquivamento digital descobrem quais informações de proveniência são necessárias, outros apresentam fator-chave na avaliação da confiabilidade de um documento, se é ou não confiável; metadados sobre proveniência, sem dúvida, ajudarão nesses julgamentos, mas não precisam ser necessariamente resolvidos. Representar sua confiança e confiabilidade é uma tarefa muito complexa e difícil na lógica epistêmica, como afirma Jøsang (2001).

No contexto do conhecimento, abordagens de representação incluem: lógica subjetiva, que representa uma opinião como uma tripla com valor real (crença, descrença, incerteza), onde os três itens somam uma resposta e, ainda na classificação com base em julgamentos qualitativos, embora possam receber interpretações numéricas e depois raciocinar matematicamente, aplicar Inteligência artificial como lógica fuzzy e probabilidade (GIL, 2003).

Os metadados relacionados às restrições de licenciamento estão atrelados à lei de direitos autorais, sob as licenças *Creative Commons* que permitem que os autores ajustem o exercício de suas atividades, renunciando a alguns deles para facilitar o uso de seu trabalho em vários contextos especificados (LESSIG, 2001).

As perguntas sobre as dificuldades de raciocinar com metadados baseiam-se no contexto do estudo da anotação semântica, que exige métodos automáticos em larga escala, e exigem compromissos de modelagem de conhecimento (KIRYAKOV *et al.,* 2005), como apresentado na seção 4.2.2.1 Anotação Semântica de Dados.

5.3 Limpeza de dados

O Processo de Limpeza de Dados, apresentado por Rautenberg *et al.* (2018), é uma etapa que depende muito de como os dados foram coletados de um determinado domínio e independe se esse dado se constitui como

estruturado, semiestruturado ou desestruturado. Esse processo parte do princípio do ciclo de vida de Dados apresentado por Sant'Anna (2016) sobre a coleta dos dados que não é o foco de explicação do presente trabalho.

Esse processo de limpeza se procede com a importanção do arquivo criado no processo anterior, como já dito, no processo de busca no ciclo de vida de dados, e seu resultado sempre é composto pela obteção de uma linha para cada registro dos dados determinando em um domínio e uma coluna para cada campo e subcampo deste contexto.

O objetivo principal que se segue nesta fase de limpeza, determinado por Montalvillo Mendizábal (2012), é corrigir ao máximo os possíveis erros que tiveram esses registros advindos de base de dados (que podem se classificar em três categorias), transformando-os em "dados limpos e alta qualidade".

Sem dúvida alguma, esta é a fase em que se encontra a parte mais específica de um trabalho que envolve o contexto de enriquecimento e mapeamento, pois requer um maior tempo de apreciação e apuração desses dados. Porém, ao mesmo tempo, é o processo que determinará se o resto do processo de extração e limpeza gerará em arquivo de conversão de registros que resulte corretamentemente ou não o processo, além de escolher o modelo que a categorize no modelo de 5 estrelas de Berners-Lee (5 [ESTRELAS], 2012), onde sua denominação é o arquivo estruturado, atingindo 3 estrelas.

O êxito desta etapa se encontra no poder de desenvolver todo o restante do processo de enriquecimento e mapeamento, em especial do processo de ligação (*Linked Data*), onde o cruzamento dos dados com os outros *datasets* é crucial, já que devem se encontrar no melhor estado possível, podendo tanto estabelecer relações como garantir que outros usuários e até mesmo outros *datasets* já denominados neste contexto, possam consumi-los da mesma maneira (SENSO, MACHADO, 2018).

5.4 Modelagem de dados

A Modelagem de Dados está ligada ao êxito de poder desenvolver todo o processo de Enriquecimento e Mapeamento, uma vez que que está ligada com o processo de evolução dos dados, segundo Rautenberg *et al.* (2018) e Senso, Machado (2018).

Em especial, o processo de ligação dos dados (*Linked Data*), onde a junção dos dados com os outros *datasets* caracteriza-se de forma crucial, pois devem se encontrar no melhor estado possível, podendo tanto estabelecer relações como garantir que outros usuários e até mesmo outros *datasets,* já denominados neste contexto, possam consumi-los da mesma maneira (SENSO, MACHADO, 2018).

Entende-se como processo de modelagem, como uma seleção de conjuntos de Ferramentas conceituais, que serão descritas com maior finalidade na seção (mapeamento), que descrevem e mapeiam os dados, suas relações, significados e todos os tipos de restrições que as compõem.

A partir do conteúdo ou contexto utilizado pelo usário, segue-se um determinando modelo que é proposto, onde se podem encontrar e localizar ontologias e vocabulários que permitiram, com a maior precisão, realizar o processo descrito, como o *Linked Open Vocabularies*[36].

Uma vez escolhido o vocabulário e o gerenciamento d Ontologia, ou mesmo sua criação como já descrito anteriormente no **Capítulo 4**, entende-se que todas as vezes que se criar uma modelo específico não é uma ideia muito boa, pois não entra no contexto das Boas Práticas, descritas por Lóscio, Burle e Calegari (2017), uma vez que o reuso de datasets ligados é um dos maiores fatores para que a *Web* Semântica crie um processo de interoperabilidade entre outros contextos já apresentados, referente às boas práticas de Publicação na *Web*.

Dimou *et al.* (2016) afirmam que, uma vez escolhidos os vocabulários que serão usados, desenvolvem-se uma sequência mapeada das diferentes classes e subclasses, propriedades e relações que serão estabelecidas com os dados obtidos.

Segundo Santarem Segundo e Coneglian (2015, p. 227), é importante falar sobre a construção e adaptação de ontologias que têm uma importância para a Web Semântica, uma vez que seu uso pode explicar uma situação das informações em si.

---

[36] https://lov.linkeddata.es/dataset/lov/about

Todo o processo de Ontologias, linguagem OWL e suas construções e diversidades já estão explicadas e apresentadas no **Capítulo 4**.

Ainda no contínuo do racíciocinio dos autores, Santarem Segundo e Coneglian (2015, p. 227), a *Web Ontology Language* (OWL) é uma linguagem indicada no que se propõe à construção de ontologias e sua viabilidade à implementação em uma linguagem computacional, e ainda afirmam que:

> Para o uso como tecnologia da Web Semântica, entendem-se as ontologias como: artefatos computacionais que descrevem um domínio do conhecimento de forma estruturada, através de: classes, propriedades, relações, restrições, axiomas e instâncias.

5.5 Mapeamento semântico de dados

Yunianta *et al.* (2013) afirmam que o mapeamento de dados semânticos é a relação entre quatro partes importantes para o mapeamento e integração de dados semânticos de dados.

A parte principal é o mapeamento de dados semânticos, que lida com comunicação e integração com todas as outras partes. A segunda parte é a fonte de dados que será mapeada no processo de mapeamento semântico de dados. A terceira parte é uma aplicação local que usa mapeamento de dados semânticos. E a quarta parte é um outro sistema, que utiliza o protocolo HTTP, que irá se comunicar, integrado de fora do ambiente.

Van Harmellen (2008) afirma que, para se alcançar a interoperabilidade de dados semânticos, os sistemas não precisam apenas trocar seus dados, mas também trocar ou concordar com modelos explícitos desses dados. Interoperabilidade de dados pode ser contextualizada a partir da capacidade fornecida aos sistemas para interpretar, de maneira automática e precisa, o significado dos dados trocados.

O levantamento das ferramentas de mapeamento de dados semânticos existentes traz à pesquisa um processo avaliativo das melhores ferramentas de mapeamento de dados semânticos nos últimos anos, baseados nos estudos referentes aos trabalhos correlatos, apresentados no **Capítulo 2** e segundo a perspectiva de Güell (2017). Essa análise das catorze ferramentas de mapeamento de dados semânticos traz informações para auxiliar no estudo da construção do *Framework* conceitual.

Alguns critérios são de extrema relevância a serem verificados. O primeiro critério é baseado em fontes de dados relacionais, ou seja, em dados estruturados; o segundo é em termos de facilidade de uso; o terceiro é em termos de processo de mapeamento; o quarto é sobre ferramentas pagas ou gratuitas e o quinto é em termos de suporte a várias fontes, que são apresentados na seção 5.3.1.2

5.5.1 Definição

O mapeamento de dados pode ser gerenciado de diversas formas: automática, semiautomática e manual.

Em muitos dos casos, dependendo do domínio estabelecido, ainda se escolhe o processo manual de mapear esses dados, mas devido à evolução de tecnologia e, ainda, à diversidade de tipos de dados encontrados, como já apresentados (estruturados, semiestruturados e desestruturados), existe uma diversidade de ferramentas que, esses dados em conjunto com uma ontologia que possua domínios correlacionados e ainda vocabulários abertos, fica muito mais fácil, utilizar ferramentas para contextualização desse processo, como já dito por Yunianta *et al.* (2017).

Elmasri e Navathe (2003) definem “Dado” como um fato isolado que pode ser gravado e possui um significado implícito e trazem a seguinte classificação para os diferentes tipos de Dados:

- Dados estruturados: São dados organizados de acordo com critérios definidos, respeitando vários atributos, delimitando escopo, tipos, entre outros.
- Dados semiestruturados: São dados que nem sempre é possível prever todos os aspectos. Alguns atributos gerais podem ser conhecidos, enquanto outros poderão ser adicionados posteriormente.

Dados Estruturados, segundo Isotani e Bittencourt (2015), podem ser classificados como dados organizados e representados através de uma estrutura rígida, previamente projetada, se enquadrando no que foi

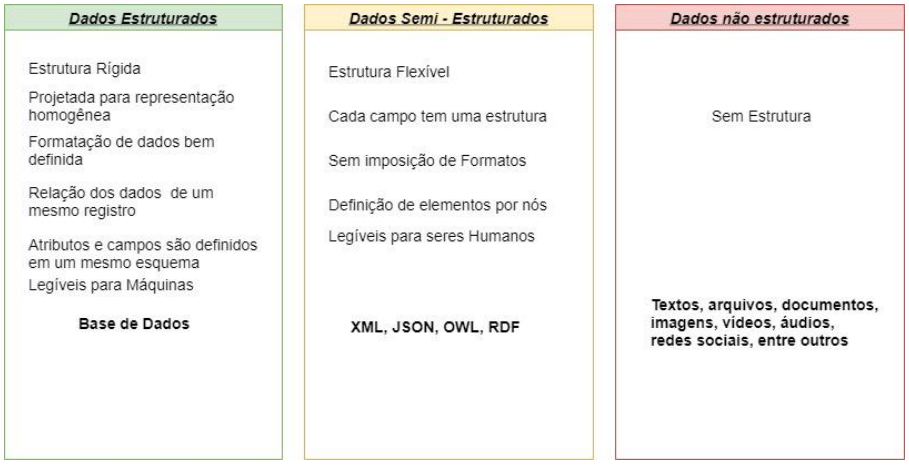
previamente planejado e têm as funcionalidades de armazenamento e a possibilidade de recuperação desse dado.

Um dado pode ser classificado também como não estruturado, pois possui exatamente o oposto da definição determinada quando é estruturado.

Quando se tem textos, vídeos, imagens e entre outros tipos de documentos, é necessário que se guarde toda a informação que se refere a este dado, permitindo futuramente uma recuperação de forma automatizada. Porém isso não acontece, pois esses tipos de documentos não são tratados nessa escala dentro de banco de dados, uma vez pela inviabilidade de classificação por cada palavra, *frame* ou *streaming,* entre outros, como redes sociais relacionadas ao seu conteúdo, aumentando assim o grau de complexidade, inviabilizando essa organização e determinando como uma identificação não clara, ou seja, desorganizada. De Diana e Gerosa (2010) afirmam que os dados na *Web* não são estruturados e isso é um dos fatores que enriqueceram o aparecimento de tecnologias que gerenciam dados diferentes das tecnologias tradicionais.

Os dados semiestruturados são determinados como uma organização de dados bem heterogênea e que possui um armazenamento estratégico dos dados, porém suas estruturas ainda são bem difíceis de entender, o que gera a dificuldade em consultar e classificar o dado (MELLO *et al.,* 2000). A **Figura 37** apresenta de forma visual a classificação dos dados apresentados.

Figura 37 - Classificação dos dados quanto sua Estrutura



Fonte: Elaborado pela autora (2019)

As ferramentas sempre tentam trazer uma forma de padronizar o mapeamento dos dados para gerar consistência entre eles a partir de várias representações, formato de dados heterogêneos de diferentes aspectos semânticos entre aplicativos nas diferentes fontes de dados (YUNIANTA *et al.*, 2014).

Quando se fala em mapear dados, já se visualiza um processo que traga uma busca futura e que seja facilitadora desses dados, e ainda mais, no contexto de significado desses dados no domínio estabelecido pelo usuário, uma vez a usabilidade e interoperabilidade desses devem ter o objetivo de disponibilizar o mapeamento através de bases estruturadas em formato de Dado Aberto Conectado, uma vez que se utilizam de representações padrões preestabelecidas, como Bernees-Lee já apresentou com com o contexto do *Linked Data* e ainda a W3C, através do texto dos autores de Lóscio, Burle e Calegari (2017), apresentam as Boas práticas para se chegar à publicação futura desses dados mapeados semanticamente.

Muitos autores, como visto no **Capítulo 2**, constroem projetos de mecanismos para a disponibilização desses dados. Dessa forma, falar de mapeamento semântico é permitir que um usuário acesse os dados dessas bases estruturadas e conectadas mesmo sem saber detalhes referentes onde se encontram, apenas que sua semântica traga, através da intenção da busca da máquina ou mesmo do usuário, o conteúdo requisitado, aplicando um filtro correspondente sobre essa base de dados, de modo a requisitar informações através da ontologia para o qual se deseja mapear a informação.

5.5.2 Geração do Mapeamento

A crescente disponibilização de dados abertos tem se tornado, nos últimos tempos, uma tarefa de tendência global e ainda em crescimento exponencial, contudo, neste contexto, o presente trabalho tem como principal objetivo apresentar o mapeamento semântico de dados baseados nos padrões da W3C (MANUAL, 2011), utilizando as linguagens *Resource Description Framework* (RDF) e *SPARQL Protocol and RDF Query Language* (SPARQL) que estão diretamente ligadas ao contexto estudado da *Web* Semântica, *Linked Data* e Ontologias.

Berners-Lee (2006b) definiu as regras propostas para publicação de dados na *Web*, determinado como *Linked Data* ou Dados Abertos.

Berners-Lee, Hendler e Lassila (2001) afirmam que, com o cumprimento dessas regras, os dados publicados sejam únicos e unificados em um espaço, além de fornecer a conexão dos dados na *Web* e permitir distinção semântica entre os dados que forem publicados, determinando que essa integração dos dados só possa ocorrer com o uso de uma ontologia.

Verifica-se a necessidade de regras de inferência nesse processo, e como as ontologias oferecidas na *Web* possuem essa característica e ainda uma taxonomia de definição de classes de objetos e as relações entre elas, Berners-Lee, Hendler e Lassila (2001) afirmam que o uso das regras de inferência melhora os aspectos da qualidade da análise dos dados, além de detectar possíveis inconsistências neles.

Esta seção aborda o mapeamento de dados semânticos, verificando o desenvolvimento de tecnologias que, a partir de técnicas e implementação de ferramentas, permitem a automatização do processo. Processo esse que realiza a conversão ou mapeamento de dados em RDF, sejam eles estruturados, semiestruturados ou não estruturados.

5.4.2.1 Principais Ferramentas para Mapeamento

Existem diversas técnicas e ferramentas para guiar os passos do mapeamento de dados semânticos nos últimos anos, afirmam Sahoo et al. (2009) e Yunianta *et al.* (2017). No entanto, esta seção destina-se a apresentar as ferramentas e tecnologias de mapeamento de dados semânticos existentes atualmente e suas características relevantes como: quais fontes de dados as ferramentas suportam, se o mapeamento semântico com o uso da tecnologia é fácil ou não, se o processo é todo automático, ou se tem processos manuais integrados e ainda se são ferramentas gratuitas ou pagas e multiplataformas, ou seja, que atendem a diversos sistemas operacionais de máquinas.

a) *Asio Semantic Bridge for Relational Databases (SBRD) and Automapper*

O Asio *Tool Suite*[37] é um produto comercial desenvolvido pela *BBN Technologies*, que fornece vários componentes integrados: *Asio Semantic Query Decomposition* (SQD), Asio *Semantic Bridge for Relational Databases*[38] (SBRD), Asio *Semantic Bridge for Web Services* e duas ferramentas de código aberto chamadas *Parliament* e *Snoggle*.

Asio SBRD foi projetado para encarar o desafio de volume, variedade e velocidade de dados que, segundo a W3C (2009), o Asio *Tool Suite* oferece suporte à descoberta de informações por meio dos padrões da *Web* Semântica e fornece acessibilidade aos dados por meio de consultas feitas na ontologia do usuário. Além disso, o Asio *Tool Suite* permite que sistemas e usuários alcancem um entendimento compartilhado sobre o significado, estrutura e contexto dos diferentes dados e garante que o usuário possa tomar uma decisão válida sobre a confiabilidade dos dados.

Seu principal processo é transformar as bases de dados relacionais para RDF, através do *automapper*, processo onde se aplica regras de mapeamento para criar a ontologia local que descreve o esquema relacional. O *automapper* usa, como entrada para o mapeamento, a ontologia (OWL) para produzir dois resultados: (1) ontologia de fonte de dados (OWL) e (2) regras e dados de instância de mapeamento, conforme relatam Michel, Montagnat, Zucker (2014).

A ferramenta *Asio* SBRD é a ponte semântica para se comunicar com o banco de dados relacional e mapear dados e regras da instância. Para obter resultados específicos, esta ferramenta usa consulta SPARQL localizada dentro do componente SQD (*semantic query decomposition* - decomposição de consulta semântica).

O mapeamento direto possui restrições de propriedade OWL para modelar tipos de dados e propriedades anuláveis ou não anuláveis. Possui regras de mapeamento adicionais entre o domínio da ontologia e ontologia local. São descritas como conjunto de regras SWRL (*Semantic Web Rule Language* - Linguagem de regras da *Web* semântica) (Horrocks *et al.,* 2000). O *Asio* SBRD implementa o modo de consulta sob demanda, ou seja, ele

[37] http://bbn.com/technology/knowledge/asio_tool_suite
[38] http://bbn.com/technology/knowledge/asio_sbrd

reescreve as consultas SPARQL em consultas SQL e aplica técnicas de planejamento e otimização de consultas.

O módulo *Asio* SQD aborda a associação de várias fontes de dados, divide as consultas SPARQL, baseando-se em uma ontologia de domínio em subconsultas e usando a ontologia local de cada fonte de dados e distribui essas subconsultas otimizadas para as fontes de dados aplicáveis: bancos de dados relacionais, serviços da *Web* ou triplas (HORROCKS *et al.,* 2000).

b) Dartgrid

O Dartgrid é o sistema de consulta semântica para dar suporte ao processo de construção na organização virtual de banco de dados (DB-VO - *database virtual Organization*) baseado em ontologia de larga escala, usando *grids* como plataforma.

Existem três características técnicas principais no Dartgrid como a implementação referencial no modelo OntoVO (*Ontology Virtual Organization - Organização Virtual de Ontologia*) (CHEN; WU, 2005, p. 35).

A primeira característica é o processo de desenvolvimento no Globus 3.0 para construir o VO (*Virtual Organization* - Organização Virtual) na área de pesquisa em computação em grade (*grid*). A segunda característica é sobre o RDF, o modelo de dados padrão para definir protocolos na *Web* Semântica. A terceira característica são as ontologias em si, o *Dartgrid* usou ontologias para cumprir a semântica e a sintaxe da OWL, a linguagem padrão de descrição de ontologias produzida pelo W3C (OWL WORKING GROUP, 2012). O *Dartgrid* ajuda o provedor de dados a conduzir o processo de mapeamento de dados semânticos do esquema relacional na fonte de dados para a ontologia compartilhada.

c) DB2OWL

O DB2OWL é uma ferramenta para mapeamento automático de banco de dados para ontologia que deve gerar dados do banco de dados relacional para a linguagem de ontologia OWL-DL[39]. Existem duas etapas principais no processo de mapeamento de dados semânticos dentro do DB2OWL, segundo Cullot, Ghawi e Yetongnon (2007).

[39] Proveniente das derivações de OWL, que pode ser visto como uma notação alternativa para o idioma lógico da descrição SHOIN (D) (BRUJIN.,2005).

O primeiro passo é ler e extrair todos os esquemas de banco de dados dentro de fontes de dados; todos os esquemas estão incluídos no nome da tabela, estruturas de coluna e todas as restrições dentro do banco de dados. A segunda etapa é converter, diretamente para a linguagem ontológica, o conteúdo; dentro da linguagem ontológica, inclui nome da classe, propriedade dos dados, propriedade do objeto e relacionamento semântico na linguagem ontológica (Bruijn, 2005).

d) Linguagem de Mapeamento D2RQ

A linguagem de mapeamento D2RQ é a ferramenta de mapeamento de dados semânticos classificada como a favorita a ser utilizada, pois o D2RQ é uma ferramenta de mapeamento de dados semânticos gratuita (CYGANIAK *et al.* 2012).

O mapeamento semântico define um grafo virtual RDF que contém informações do banco de dados. Isso é semelhante ao conceito de visualizações no *Structure Query Language* (SQL), exceto que a estrutura de dados virtuais é um grafo RDF em vez de uma tabela relacional virtual. O grafo RDF virtual pode ser acessado de várias maneiras, dependendo do que é oferecido pela implementação.

O D2RQ possui várias habilidades, como extrair e integrar dados de mais de uma fonte de dados, dar suporte ao *dump*[40] Jena e RDF, pode fornecer arquivos de mapeamento de dados semânticos no formato *turtle*[41] e trabalha no protocolo HTTP (*HyperText Transfer Protocol*) usando o servidor D2R (CYGANIAK *et al.*, 2012; BIZER; SEABORNE, 2005).

Os mapeamentos D2RQ podem ser escritos manualmente em um editor de texto, mas geralmente é muito mais rápido começar com a ferramenta de mapeamento que gera um esqueleto ou “mapeamento padrão", a partir do esquema do banco de dados, ou seja, fornece mapeamento automático e manual, para que os usuários possam personalizar os arquivos de mapeamento de dados semânticos para ajustar com os outros arquivos de mapeamento de dados semânticos (CYGANIAK *et al.*, 2012).

---

[40] Um registro da estrutura de tabelas e/ou os dados de um banco de dados (DATE, 2009)
[41] *Terse RDF Triple Language* é um formato de arquivo e sintaxe para expressar dados no modelo de dados do *Resource Description Framework* (RDF) (ISOTANI; BITTECOURT, 2015).

e) Iconomy

A ferramenta *Iconomy* fornece uma interface de usuário multifuncional para facilitar usuários em suas consultas e experiências semânticas.

A Iconomy extrai os dados do banco de dados relacional e sincroniza com o esquema de ontologia para criar uma linguagem de ontologia Vavliakis *et al.* (2011).

Através da ferramenta, é capaz de carregar qualquer ontologia no formato de armazenamento *.owl* ou persistente; criptografar e proteger ontologias geradas por meio de mecanismos de transformação; descriptografar arquivos *.owl* codificados para visualização e edição; navegar, percorrer e editar a classe e o objeto/tipo de dados de propriedade da árvores ontologia, bem como o foco em instâncias; navegar, adicionar e editar restrições nas classes da ontologia através de uma interface amigável; verificar a consistência da ontologia e relatar os termos inconsistentes, criar consultas SPARQL simples ou complexas através de um *menu* interface lista/suspensa ; verificar a explicação das consultas SPARQL em linguagem natural e editar o conteúdo no código SPARQL; habilitar e desabilitar o *Reasoner* embutido (*Pellet*); gerenciar contas de usuário e seus direitos nas ontologias, entre outras diversas funcionalidades descritas por Vavliakis *et al.* (2011).

Iconomy suporta a transformação direta de qualquer relacionamento de banco de dados à sua respectiva ontologia através de uma simples, mas poderosa ferramenta gráfica. Dessa maneira, os usuários podem facilmente mapear um esquema de banco de dados para uma ontologia e instanciá-la, juntamente com permissões de 'inserção' e 'atualização'.

O núcleo do sistema é o mecanismo de transformação semântica, que fornece todos os métodos necessários para a interface gráfica e coordena os módulos de cooperação. Para a manipulação da ontologia, o sistema utiliza estruturas de APIs de *Frameworks* como Jena ou Sesame, mediante solicitação do usuário (VAVLIAKIS *et al.,* 2011).

f) METAmorphosis

Existem muitas abordagens para lidar com banco de dados relacional a fim de apresentar dados na linguagem RDF. Uma delas é a abordagem de dados serem armazenados no esquema clássico relacional e existe algum

mapeamento para RDF, e a ferramenta METAmorphosis utiliza-se dessa abordagem.

O mapeamento é criado de acordo com estrutura e ontologia SQL. Quando existe um pedido de RDF, é traduzido para Consulta SQL. Os resultados da consulta são traduzidos para RDF e isso é enviado como resposta.

Como resposta a este desafio, foi proposta por Svihla e Jelinek (2004) a ferramenta METAmorphoses, que mapeia banco de dados relacional para RDF de fácil utilização.

METAmorphosis possui duas camadas, onde a primeira camada é mapeada e a segunda camada é o modelo (Svihla e Jelinek, 2004).

Existem dois tipos de problemas, que podem ser tratados pela METAmorphosis, relacionados aos dados dentro das bases de dados relacionais. O primeiro está relacionado ao armazenamento dos dados como triplas RDF, que podem consultar os dados diretamente usando o modo RDF. O segundo é a implementação básica da base de dados com o esquema relacional clássico e há algum mapeamento para o formato RDF (SVIHLA; JELINEK, 2004).

g) ODEMapster

O processador ODEMapster é capaz de mapear todas as instâncias dentro do banco de dados relacional para produzir instâncias semânticas baseadas em todas as descrições no documento R2O[42] (*Extensible and Semantically Based Database-to-ontology Mapping Language*) (BARRASA RODRÍGUEZ; CORCHO; GÓMEZ-PÉREZ, 2004).

Existem dois modelos principais de execução dentro do processador ODEMapster: o primeiro é uma atualização orientada a consultas (tradução de consultas *on-the-fly*) e o segundo é o processo em lote de atualização maciça para gerar todos os indivíduos da *Web* Semântica a partir das fontes de dados (GÓMEZ-PERÉZ; ORTIZ-RODRÍGUEZ; VILLAZÓN-TERRAZAS, 2005).

A grande diferença desses dois modelos está no repositório semântico, pois, no primeiro modelo, os clientes podem ser diretamente acessíveis à fonte

[42] Banco de dados extensível e de base semântica para a linguagem de mapeamento de ontologias (BARRASA RODRÍGUEZ; CORCHO; GÓMEZ-PÉREZ, 2004)

de dados usando o processador de consultas e, no segundo, os clientes acessam a fonte de dados por meio da geração do repositório para gerar a partir do repositório semântico.

h) Owlifier

Owlifier é uma ferramenta que descreve uma abordagem para a criação de ontologias que permite que os ecologistas usem ferramentas comuns de planilhas para descrever diferentes aspectos de uma ontologia.

É uma ferramenta que permite criar, relacionar e restringir conceitos por meio de planilhas e fornece ferramentas de software para converter essas ontologias em representações OWL-DL equivalentes. Também são consideradas traduções inversas, ou seja, para converter ontologias representadas usando OWL-DL em formato de planilha. A abordagem apresentada por Bowers; Madin; Schildhaer (2010) permite que grandes listas de termos sejam facilmente relacionadas e organizadas em hierarquias de conceitos, fornecendo uma interface mais intuitiva e natural para o desenvolvimento de ontologias para ecologistas.

Existem quatro componentes importantes contidos no Owlifier para converter dados da planilha no conhecimento ontológico. O primeiro componente é o arquivo de texto das definições de ontologia (blocos) que obtém os dados da planilha. O segundo é o provedor de ontologia OQL para converter os dados da planilha em linguagem de ontologia. O terceiro é a ontologia fundamentada como uma facilitação para medir o conhecimento da ontologia. E o quarto componente é a importação de OWL que tem a função de gerar e importar outra linguagem de ontologia para integrar com os dados da planilha Bowers; Madin; Schildhaer (2010).

i) RDB2Onto

A ferramenta RDB2Onto é uma abordagem para a criação de metadados semânticos a partir de dados de bancos de dados relacionais. Converte os dados selecionados de um banco de dados relacional a um documento de ontologia RDF/OWL com base em um modelo definido. Esses modelos preenchidos podem ser armazenados na memória de conhecimento baseada em ontologia.

A ferramenta traz, como objetivo principal, resolver os problemas da conversão dos dados do RDB para os dados da ontologia quando eles desejam criar conteúdo da *Web* com base nas tecnologias da *Web* Semântica, como arquivos OWL. A ferramenta trouxe a solução de simplificar a extração de dados/informações do banco de dados relacional e produzir o modelo XML RDF/OWL usando a consulta de sintaxe SQL (LACLAVIK et al., 2010).

RDB2Onto obtém duas entradas do banco de dados relacional e de consulta SQL RDF/OWL. Essas entradas são processadas em três etapas usando a execução da consulta SQL, preenchendo o modelo com dados do RDB e armazena os dados RDF/OWL, tendo como saída desse processo os dados da ontologia (LACLAVIK et al., 2010).

j) RDOTE

RDOTE é a ferramenta de mapeamento de dados semânticos para converter dados de vários bancos de dados relacionais em diferentes conhecimentos de ontologia e integrá-los em conhecimento único de ontologia. O RDOTE está disponível sob licença GNU/GPL e fornece interfaces gráficas mais amigáveis, além de expressividade suficiente para criar *dumps* RDF automáticos e personalizados de dados relacionais. O RDOTE também é compatível com as definições de mapeamento D2RQ e R2RML (VAVLIAKIS; MITKAS; KARAGIANNIS, 2012).

Existem dois objetivos principais do RDOTE: o primeiro é a capacidade de instanciar rapidamente o conhecimento da ontologia com os dados reais que esse processo pode implementar com um grande conjunto de dados da ontologia. O segundo é a capacidade de transformar conjuntos de dados, que atualmente residem no banco de dados relacional, em dados semânticos por meio de uma interface gráfica para usuário. Em 2013, o RDOTE se tornou mais complexo e completo com a adição de várias habilidades e processos, como leitor de ontologia, leitor de RDB, processo de mapeamento e escritor de ontologia (VAVLIAKIS; SYMEONIDIS; MITKAS, 2013).

k) R2O

O R2O é uma linguagem extensível e declarativa para descrever mapeamentos entre esquemas de banco de dados relacional e ontologias

implementados no RDF ou OWL. O R2O fornece um grande conjunto de primitivas com semânticas bem definidas. Essa linguagem foi concebida de maneira suficientemente expressiva para lidar com casos complexos de mapeamento decorrentes de situações de baixa similaridade entre a ontologia e os modelos de banco de dados, que visam extrair as informações dentro do banco de dados para serem arquivos no formato RDF ou linguagem de ontologia usando o formato OWL (BARRASA RODRÍGUEZ; CORCHO; GÓMEZ-PÉREZ, 2004).

O processo de extração é produzir um formato padrão a partir de diferentes formatos de representação de banco de dados a serem usados na linguagem ontológica. Existem três camadas principais na arquitetura de mapeamento do R2O: camada de implementação, formalismo e modelagem.

A primeira camada é a camada de implementação relacionada à SQL, sistema de banco de dados e implementação de ontologia. A segunda camada é a camada de formalismo; o objetivo dessa camada é lidar com o modelo relacional e o modelo formal de ontologia. E a terceira é a camada de modelagem relacionada ao modelo de entidade-relacionamento e ao modelo conceitual de ontologia (BARRASA RODRÍGUEZ; CORCHO; GÓMEZ-PÉREZ, 2004).

l) Triplify

O *Triplify* é uma ferramenta de mapeamento de dados semânticos para extrair dados vinculados do banco de dados relacional com base no mapeamento de solicitações HTTP-URI para consultas em bancos de dados relacionais (AUER et al., 2009).

O *Triplify* trabalha entre a base de dados e o servidor da *Web*. O principal objetivo do *Triplify* é recuperar informações valiosas da fonte de dados e converter os resultados da consulta no formato RDF, vinculando os dados no formato JSON[43] (*JavaScript Object Notation*). O *Triplify* pode ser implementado no banco de dados relacional e PHP (*Hypertext Preprocessor*) como uma linguagem de programação nos aplicativos da *web*.

[43] Notação de Objetos JavaScript

m) Ultrawrap

Ultrawrap, criado desde 2009, tem como objetivo específico sintetizar a linguagem de ontologia do esquema SQL dentro do sistema de banco de dados e fornecer consultas SPARQL (W3C. 2012a). O Ultrawrap extrai dados de bases de dados com formato de esquema de tripla ou SQL. Em 2013, o Ultrawrap foi aprimorado e avaliado usando dois conjuntos de *benchmarks* existentes (SEQUEDA, MIRANKER, 2013).

Existem quatro componentes principais na ferramenta de mapeamento de dados semânticos Ultrawrap. O primeiro componente é a tradução do esquema SQL, o segundo é a criação de uma tabela tripla intencional, o terceiro é a tradução de consultas SPARQL e o quarto componente é o otimizador de consulta SQL nativo.

n) Virtuoso RDF *Views*

O Virtuoso pode ser classificado como uma plataforma cruzada para gerenciamento de dados SQL, XML e RDF, e suas funcionalidades estão relacionadas em reduzir o custo de reunir dados de diferentes fontes de dados, também fornece o acesso transparente às suas fontes de dados existentes mesmo de base de dados de diferentes fornecedores. Todas as suas bases de dados são tratadas como uma única unidade lógica.

Oferece mecanismos de banco de dados virtual como plataforma de implantação de serviços *web*, servidor de aplicativos da *Web* e suporte SPARQL e armazenamento de dados RDF totalmente integrado ao seu armazenamento relacional.

O Virtuoso RDF View (ERLING, MIKHAILOV, 2010) são os dados de extração de um ODBC (*Open Database Connectivity*)/ JDBC (Java™ EE Database Connectivity) que fornecem resultados como repositório DAV, pontos finais de protocolo SOAP (*Simple Object Access Protocol*, em português Protocolo Simples de Acesso a Objetos) e *Web* Semântica. O Virtuoso também fornece SPARQL dentro da consulta SQL, isso pode ser implementado na função RDF_MATCH do sistema gerenciador de banco de dados (SGBD) Oracle. Esta ferramenta é lançada com uma versão de código aberto gratuita e a versão comercial. Na versão da licença gratuita é possível combinar um mecanismo

de banco de dados híbrido com o armazenamento triplo RDF. Essa ferramenta possui uma interface gráfica com o usuário para declarar o processo de mapeamento. No entanto, esse recurso é apenas para o banco de dados virtuoso e não pode ser usado em outros SGBDs relacionais.

A seguir, apresenta-se o capítulo que discute a apresentação do ***Framework*** proposto, com diretrizes aplicadas ao todo conteúdo estudado até o momento, com ênfase em Enriquecimento e Mapeamento semântico de Dados.

6. *FRAMEWORK* PARA PUBLICAÇÃO DE DADOS COM ÊNFASE NO ENRIQUECIMENTO E MAPEAMENTO SEMÂNTICO

O debate acerca da questão que envolve o processo de construção de um ***Framework*** com ênfase ao enriquecimento e mapeamento de dados semânticos não se faz tão recente, ao observar o desenvolvimento das pesquisas em torno da temática.

Cood (1970), em seu trabalho *Relational model of data for large shared data banks* (Modelo relacional de dados para grandes bancos de dados, tradução nossa) discute, no contexto da Álgebra Relacional, uma forma para que dados pudessem ser estruturados e trabalhados em possíveis plataformas ou aplicações de usuários, resultando assim em eficiência e qualidade no uso.

A característica principal da Álgebra Relacional, proposta por Codd (1970), usava o sistema de *Closure property*, ou seja, podendo ser classificada como uma álgebra fechada onde cada operação recebe relações e retorna à mesma relação, permitindo uma base formal e otimização de consultas futuras de dados (RAMAKRISHNAN; GEHRKE, 2003).

Comenta-se que as ideias de Cood (1970) não foram completamente implementadas, porém serviram como base para uma estrutura matemática, o *System* R, o qual foi entregue à empresa IBM e, a partir da proposta, tem-se o modelo “relacional” e cria-se a linguagem SEQUEL.

Cood (1970) abandonou o projeto por motivos pessoais, e a empresa chegou ao produto da linguagem proposta por ele em SQL (*Struture Query Language*), que não pode ser classificada como relacional, embora siga todos os pontos e modelos da Álgebra de Cood, contudo alcançou sucesso a partir das vantagens sobre sistemas que já eram pré-relacionais.

A exposição acima se aproxima da realidade atual, na qual devido à introdução da *Web*, apresentada por Tim Berners-Lee e pelo W3C, muitos fatores começaram a mudar com o uso, acesso, reuso, compartilhamento e a qualidade das informações, entre outros fatores evidenciados nas pesquisas em torno do ciclo de vida dos dados (SANTANNA, 2016).

As formas relacionadas ao uso desses dados na *Web* podem enfrentar alguns obstáculos no que tange ao seu acesso, uso, reuso e disponibilidade, devido à sua diversidade de tipagens, e às específicas aplicações para cada

resolução que devem ser elaboradas para cada acesso (SANTAREM SEGUNDO, 2018). Infere-se que a publicação de dados na *Web*, a partir da adoção das Boas Práticas de Publicação de Dados na *Web* (LÓSCIO; BURLE; CALEGARI, 2017) e da implementação de repositórios de ontologias, são necessárias diretrizes para que cada domínio de dados a siga e disponibilize de forma prática e gerenciável, tanto para cientistas da informação como para ser disponibilizado para os setores de desenvolvimento tais aplicações.

Atualmente a busca no cenário da *Web* acontece a partir de palavras-chaves e buscadores indexados a bases que trazem diversas informações e conteúdos, os quais podem não estar direcionados ao objetivo da busca ideal do usuário. Ao inserir a Semântica neste cenário, através da introdução feita por Berners-Lee (2006b), é possível vislumbrar uma *Web* de dados apresentada de forma ligada, com repositórios de dados abertos disponibilizados, ou seja, todo contexto começa a ser visto de outra forma e com novas regras.

É necessária a discussão e visualização de novos contextos e procedimentos para que a pesquisa por palavras-chaves não seja a única forma de busca de informação, mas sim uma busca diferenciada em um aplicativo ou página *Web* que qualquer usuário, ou mesmo máquina, possa fazer através de uma interface.

Sabe-se que, em algumas vezes, a verificação das informações em um buscador serve. Nesse sentido, o buscador serve como intermediário das informações, recebendo os resultados dessa pesquisa. Quando se aplica o contexto da semântica, todo esse processo muda e o que difere é a consequência do enriquecimento dos dados na *Web*.

Muitas vezes o buscador pode trazer milhões de informações que não condizem com a necessidade inicial do usuário, gerando uma busca baseada em palavras chaves, pois os algoritmos que fazem a busca não usam somente palavras-chaves e, na maioria das vezes, possuem comportamentos diferenciados. Tal exposição, apesar de não ser o foco de discussão da tese, demostra que nem sempre a busca apresentará os resultados que o usuário esperava obter, independente de palavra-chave, algoritmo ou até mesmo da tecnologia adotada.

Uma vez determinado o processo de enriquecimento dos dados e feito seu mapeamento semântico na *Web*, baseado em uma determinada ontologia e vocabulário específico, busca-se a informação de forma semântica, ou seja, diretamente no conteúdo e conforme a necessidade do domínio buscado do usuário.

Além da informação, a semântica é tão importante que a possibilidade de retorno em diversos modelos, como, por exemplo, o RDF, que possibilita a interligação de informações de maneira específica e estruturada, e de forma limpa e clara, permitindo o desenvolvimento de aplicativos, páginas *web* e conteúdos para usuários finais.

Na construção do capítulo 2 verificaram-se diversos trabalhos que trouxeram diretrizes, técnicas e implementações voltadas ao enriquecimento semântico; porém, observou-se apenas a ordem de desenvolvimento de aplicações específicas e/ou formas de buscar e enriquecer os dados, para posterior mapeamento no formato de triplas e assim disponibilizá-los, ou não, em repositórios abertos para um reuso. Desta maneira, na análise junto aos trabalhos, verificou-se que não constam especificamente diretrizes que possam guiar o processo de enriquecer e mapear semanticamente os dados para publicação na *Web*, a partir de diretrizes e recomendações de boas práticas para o profissional da Informação.

Observou-se que implementações computacionais e autores da CI apresentam diretrizes específicas para a Publicação de Dados na *Web*, em cada um com formas e possibilidades de solução, com intuito de trazer uma melhor qualidade para o usuário. No entanto, o contexto se deve a como o processo deve ser criado, interpretado e entendido, em um formato informacional, para que, independente de tecnologia aplicada, vocabulário ou ontologia, se obtenha um bom resultado de dados que serão disponibilizados na *Web* de forma aberta e ligada, contextualizando as diretrizes de forma genérica, e principalmente partindo do princípio de dados ligados e da questão das Boas Práticas de Publicação de dados na *Web*, sem que sua explicação seja baseada em um estudo de caso único.

Nesse contexto, justifica-se a ideia de apresentação do estudo da construção de um *Framework* que possa gerar as diretrizes. O intento de um *Framework* Conceitual é propor soluções para uma família de problemas de

um determinado domínio no âmbito do entendimento em que este se encontra. Podendo, desse modo, gerar um modelo de dados que seja compreensível e aplicado a um modelo de *Framework*, de forma conceitual e depois a efetivação computacional, segundo os autores Foote e Jhonson (1988), Gamma *et al.* (1995), Taligent (1995), Shehabuddeen, Probert e Phaal (2000)

Bem e Coelho (2013) afirmam que *Frameworks* são essenciais em áreas onde não se têm a total compreensão conceitual de problemas propostos em algumas áreas, ou seja, que ainda alguns pontos estudados não possuem uma harmonia e nem uma clareza conceitual na literatura. O que vai de encontro com a ideia de gerar as diretrizes no processo de enriquecimento e mapeamento de dados semânticos.

O *Framework* proposto relaciona-se aos estudos de Santarem Segundo (2018), utilizando do contexto de gerar soluções para uma família de problema como já descrito pelos autores aqui supracitados que, diante do propósito de um fluxo segmentado em 6 (seis) fases, descreve as atividades desenvolvidas no macroprocesso de publicação de dados na *Web* em formato semântico, com a finalidade de seguir as melhores práticas de dados conectados.

6.1 DIRETRIZES DO ENRIQUECIMENTO E MAPEAMENTO DE DADOS SEMÂNTICOS

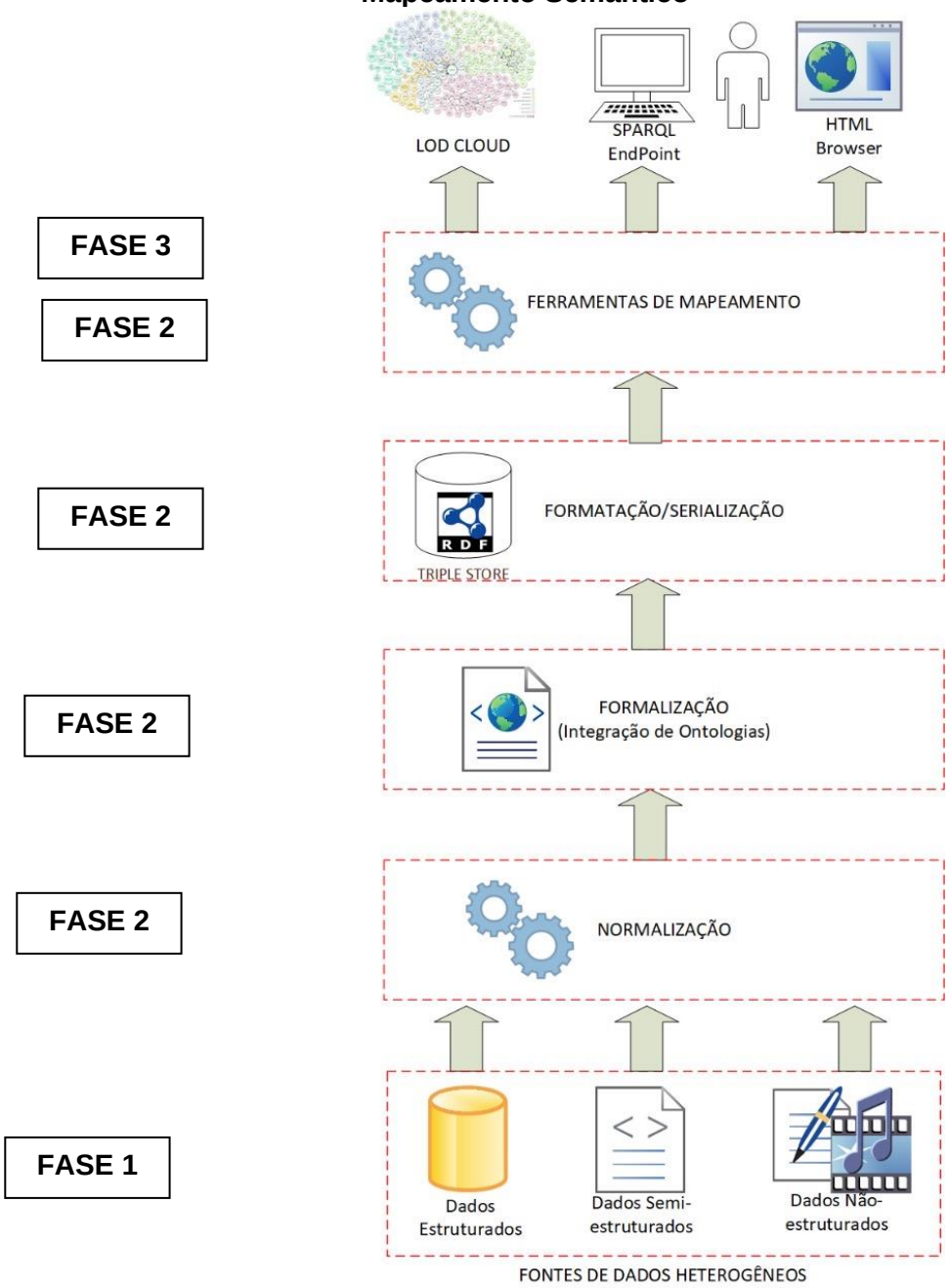
O trabalho apresenta, a partir de um ***Framework,*** diretrizes que enfatizam os possíveis questionamentos sobre Enriquecimento e Mapeamento Semântico de dados, apresentando, desse modo, as diretrizes necessárias no processo para publicar dados heterogêneos de acordo com os princípios do *Linked Data* e baseados nas Boas Práticas de Publicação de Dados propostos pela W3C, independente do domínio dos dados.

O ***Framework*** traz a concepção de apresentar diretrizes necessárias sobre o processo informacional para enriquecer e mapear os dados abertos semânticos, onde a solução não é apenas contextualizar um resultado com uso de um determinado vocabulário, ontologia e ferramenta adequada, mas sim abstrair e contextualizar o processo informacional desse enriquecimento e mapeamento para que se chegue ao contexto da publicação na *Web*, de forma

sistemática, em que os dados se apresentem como um grafo RDF em consonância com os princípios *Linked Data*.

A **Figura 38** explica a ideia inicial do Modelo Proposto do ***Framework*** com ênfase ao enriquecimento e Mapeamento Semântico, definindo como cada fase deve ser visualizada e organizada em passos para explicar nas **Figuras 39 a 46** como são essas diretrizes e suas principais especificações.

Figura 38 - Modelo de *Framework* com ênfase ao Enriquecimento e Mapeamento Semântico



Fonte: Elaborada pela autora (2019)

Compreendendo a noção temática em que tais **Diretrizes** tratam da extração, limpeza de dados para o contexto de enriquecimento e mapeamento dos dados, pressupõe que a **Fase 1**, da **Figura 38,** apresenta dados que já foram extraídos de um conjunto de dados já selecionados, obedecendo às boas práticas adotadas em relação ao publicador original, uma vez que se deve fornecer ao publicador original um *feedback dos* dados, e que serão disponibilizados de forma fidedigna ao mesmos termos originais de licença de dados abertos, como consta no tutorial de Boas Práticas de Publicação de Dados Abertos na *Web* (LÓSCIO; BURLE; CALEGARI, 2017).

A **Fase 2** abarca a normalização, formalização, formatação/serialização e as ferramentas de mapeamento são os pontos relevantes referentes à proposta das diretrizes desse *Framework*. A partir daí, cada uma das fases apresentada na **Figura 38**, gera blocos de transformação desses dados em um subconjunto de dados abertos conectados na *Web*, partindo do princípio determinado por Berners-Lee na categorização das “5 estrelas” (5 [ESTRELAS], 2012), e apresentando em cada etapa as melhores práticas para publicação e consumo na *Web* e seus benefícios (AUER, 2014).

Essas 2 (duas) **Fases** serão divididas em seções para melhor explicação e abordagem, apresentando todos os esforços necessários para que essas diretrizes estejam em concordância com a proposta do trabalho, principalmente na conformidade ao formato RDF sendo enriquecidos e passando por ferramentas existentes para conseguir um mapeamento correto.

A **Fase 3** tem relação em como os dados podem ser difundidos e usados, porém tais passos não entram no contexto da tese, apenas mostram como o resultado do mapeamento gera uma *Tripe Store* que pode ser disponibilizada ao usuário final.

Todas as fases envolvidas com os contextos apresentados na **Figura 38** são descritas de forma mais informacional no processo de Extração e limpeza até o Enriquecimento e Mapeamento Semântico dos dados, apresentando figuras que contribuem com o processo da escrita das diretrizes e todos os seus recursos necessários.

Neste contexto, as Diretrizes são apresentadas como:

- **Diretriz 1:** Extração de Dados;
- **Diretriz 2:** Limpeza e Reparação dos dados;

- **Diretriz 3:** Criação de Colunas URIs e Representação dos dados na *Web* de Dados;
- **Diretriz 4:** Processo de Análise de Similaridade/ Enriquecimento e Mapeamento Semântico; e
- **Diretriz 5:** Fusão de Dados (Formas de Disseminação dos Dados na *Web*).

6.1.1 *Diretriz 1: Extração de Dados*

Após a coleta dados, e estes passados por todas as etapas necessárias de um Ciclo de Vida dos Dados, descritas no fluxo organizacional (SANTAREM SEGUNDO, 2018), a ideia inicial do ***Framework*** vem do contexto do processo de extração dos dados, onde fontes de dados heterogêneas precisam de uma classificação e que atendam às perspectivas específicadas pelo formato “5 estrelas” determinado por Berners-Lee (5 [ESTRELAS], 2012). A extração gera a potencialidade no uso do formato de dados ligados.

Esse formato usado para explicar a **Diretriz 1** deve estar em um formato de dado estruturado, que pode ser, por exemplo, o CSV, que por definição é um formato de arquivo que significa “*comma-separated-values*” (valores separados por vírgulas, tradução nossa), e sempre utilizado para que a gestão de dados seja complexa ou simples.

Qualquer outro formato também pode ser usado aqui nesta diretriz, desde um arquivo com extensão SQL (que vem de SGBDs relacionais), com a extensão xls (arquivo relacionado ao aplicativo excel da empresa Microsoft) ou com a extensão pdf (aplicativo da empresa Acrobat Reader), arquivos do tipo *stream* (como por exemplo música, vídeos) entre outros, também podem ser formatos de extração, mas como são todos heterogêneos, sua transformação para o formato estruturado, acaba sendo um processo facilitador para a continuação da **Diretriz 1** proposta.

Isso significa que os campos de dados indicados no formato estruturado escolhido podem utilizar um *software* que faça essa transformação de formato e importe os arquivos. Utilizando, assim, as melhores ferramentas baseadas em um determinado projeto de um determinado domínio, ou em alguns casos a

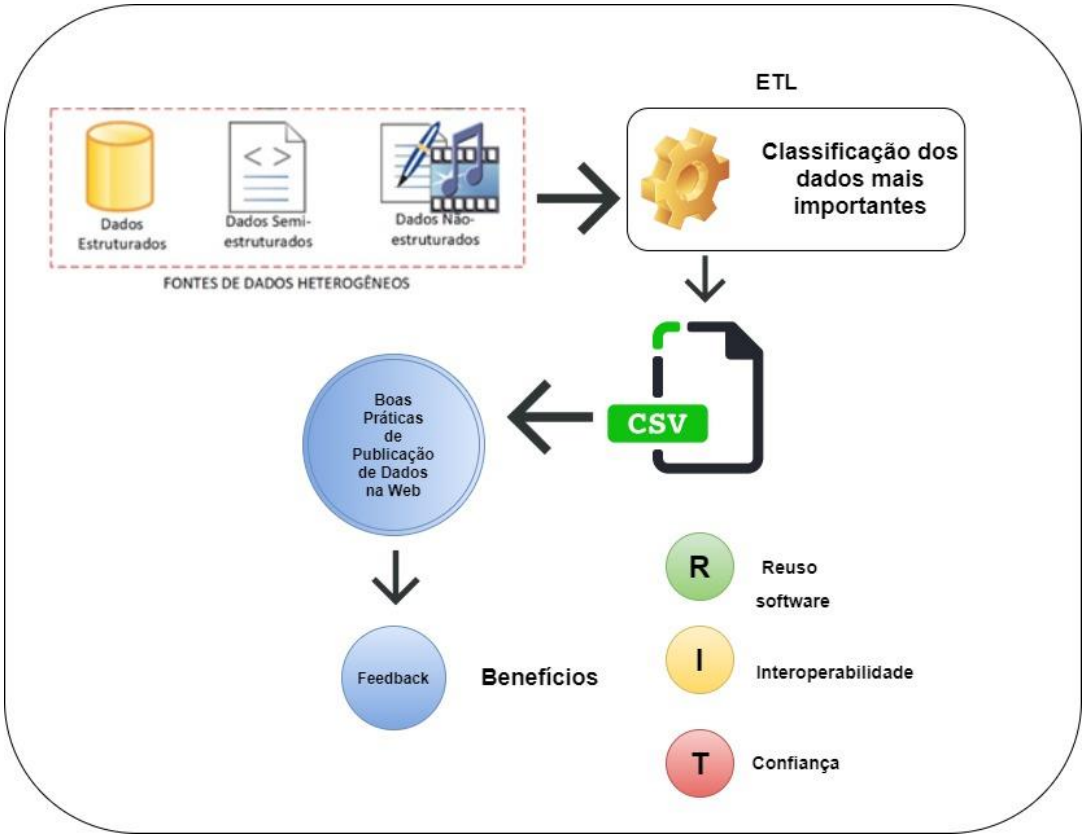
necessidade da criação de uma ferramenta nova que atenda à adequação da transformação.

Classificada como **Diretriz 1**, a extração tem como finalidade principal a própria extração dos dados, mas dos dados primários da fonte utilizada. Como já dito, por Senso e Machado (2018), a etapa está ligada à identificação e descrição dos dados que visa:

1. Identificar e analisar os dados e as fontes de dados (como *softwares*, formatos, banco de dados entre outros);
2. Identificar sua licença;
3. Determinar uma licença.

Verifica-se a entrada de fonte de dados heterogêneos, onde o *software* escolhido fará a classificação dos dados mais relevantes, por meio da *Extract, Transform e Load* (ETL), que permite uma integração de dados, determinados em 3 (três) etapas (extração, transformação, carregamento), usados para combinar dados de diversas fontes, como no caso dos dados heterogêneos apresentados, comumente utilizados para construir uma base de dados, um *data warehouse* entre outros, como apresentado na **Figura 39**.

Figura 39 - Diretriz 1: Extração dos Dados



Fonte: Elaborada pela autora (2020)

No processo dessa **Diretriz 1** visualiza-se a aplicação das Boas práticas para publicação de Dados na *Web* na categoria de coletar *feedback* de consumidores de dados e compartilhar o *feedback* disponível, permitindo comentários, opinião e experiência ao publicador original, além de fornecer a mesma licença, sendo os benefícios do uso dessa prática: reuso, interoperabilidade e confiança dos dados.

A partir da transformação e determinada a tipagem do dado estruturado que, na imagem, utiliza o formato CSV, a **Diretriz 2** permitirá a Limpeza e Reparação dos dados, e tem como objetivo transformar os dados primários encontrados na **Diretriz 1**, utilizam de outras Boas Práticas a partir de ferramentas já existentes e citadas, e transformam os dados em modelo semântico.

6.1.2 *Diretriz 2: Limpeza e Reparação dos dados*

A **Diretriz 1**, até aqui apontada, tem o principal objetivo de subir na categoria da classificação das “5 estrelas”, quando o foco é o dado ligado, determinadas por Berners-Lee (5 [ESTRELAS], 2012), onde independente da tipagem, ou seja, da fonte de dados heterogênea que pode ser proposta, transformando no formato estruturado já permite alcançar a categoria de 3 estrelas.

Neste contexto, uma das especificações é a necessidade de os dados mais importantes determinados na etapa anterior serem transformados em um formato que adere ao contexto semântico, onde os dados serão trabalhados para o formato Semântico.

Neste caso, o formato que adere esse formato semântico pode entrar na seguinte categoria: XML, RDF, RDFa, TURTLE, N-TRIPLES e JSON-LD, porém no contexto da classificação, para se atingir o formato de “5 estrelas”, o mais usado é o formato RDF. O RDF é tido como uma fonte comum dentro da comunidade da *Web* Semântica porque forma um dos blocos de construção básicos à formação da rede de dados semânticos, uma vez que a descrição dos dados e dos metadados são feitos por triplas (recurso-propriedade-valor), e principalmente pela característica de gerar uma forma coesa para catalogação e acesso aos padrões de metadados.

Diversas ferramentas são utilizadas para o processo de transformação dos dados em semântico, mas para que essa transformação ocorra, como será apresentado nas **Diretrizes 3** e **4**, a **Diretriz 2** denota a limpeza e a reparação desses dados. Algumas ferramentas são apresentadas para essa reparação e limpeza e são utilizadas baseadas nas pesquisas dos trabalhos correlatos apresentados no capítulo 2, não é um fator obrigatório, mas para uma melhor qualidade do dado futuro, é de extrema importância sua ocorrência.

Suas principais funcionalidades apontam para a **Diretriz 2**, que executa o processo de reparar e limpar os dados, transformando-os de um formato em outro e estendendo-os com serviços da *Web* e dados externos; dada a sua transformação, eles serão mantidos como dados privados até que se queira compartilhar ou colaborar.

No contexto da *Web* de Dados, verificamos diversos fatores até chegar à publicação final, uma delas é o conceito de Dados Conectados. Devido a tal informação que se deve cuidar e tratar da **Diretriz 2**, classificada como reparação e limpeza de dados com cuidado, pois no final do processo da publicação se exigirá a qualidade desses dados.

Visualiza-se, primeiramente, a necessidade de criação de um projeto, apresentado na **Figura 40**, que importe todos os dados com a extensão escolhida, apresenta o formato CSV, determinados na fase da **Diretriz 1** relacionada à extração.

A **Diretriz 2** gera um agrupamento dos dados conforme o domínio especificado, selecionando apenas os que serão utilizados e deletar aqueles que não se deseja publicar abertamente na *Web*, gerenciando, assim, sempre a manutenção deles, uma vez que podem ser alterados ou removidos da fonte principal e isso deve ser replicado de forma fidedigna.

A limpeza, apresentada na **Figura 40,** e a reparação estão relacionadas com o armazenamento e correção de dados, ou seja, sua principal tarefa é realizar a curadoria de dados. Observa-se que a diretriz está relacionada ao contexto apenas da limpeza e da reparação dos dados, ou seja, as **diretrizes** de enriquecimentos e mapeamento dos metadados são explicados e apresetados nas **Diretrizes 3** e **4**.

Já a reparação e limpeza dos dados necessitam estar em concordância com o funcionamento e qualidade dos dados e onde elas irão se apresentar, ou seja, trazendo um processo de integridade dos dados. Aqui ocorre a transformação dos dados por meio de um *software* escolhido pelo usuário que pode acontecer para possíveis formatos, como RDF, XML, TURTLE, N-TRIPLES e JSON-LD.

A Exportação dos dados transformados que são salvos em um repositório ainda local, que podem ser ferramentas como cita Senso e Machado (2018), por exemplo: *Spoon - Pentaho´s Data Integration, Virtuoso Sponger, D2QR, OpenRefine, GraphDB Free Edition*, entre outros, todos já apresentados na seção de ferramentas, no capítulo 5.

A **Figura 40** mostra todas as etapas que conduzem a **Diretriz 2** para tal transformação, e pode-se visualizar quais são as boas práticas de publicação de dados aplicadas, que estão relacionadas e as seguintes categorias:

- Formatos e Padrões: uso de formatos de dados padrão e que sejam legíveis por máquinas;
- Reuso de Vocabulários: no processo de mapeamento e transformação para os formatos, usam-se vocabulários que, de preferência, já são padronizados;
- Escolha do nível de Formalização certo;
- Seguir todos os termos de licença.

Os benefícios interligados no processo constituem em: Compreensão, Processabilidade, Capacidade de Conexão, Descoberta, Reuso, Interoperabilidade e Confiança.

Todos os benefícios descritos estão ligados com quase todas as características descritas no guia de Boas Práticas (LÓSCIO; BURLE; CALEGARI, 2017).

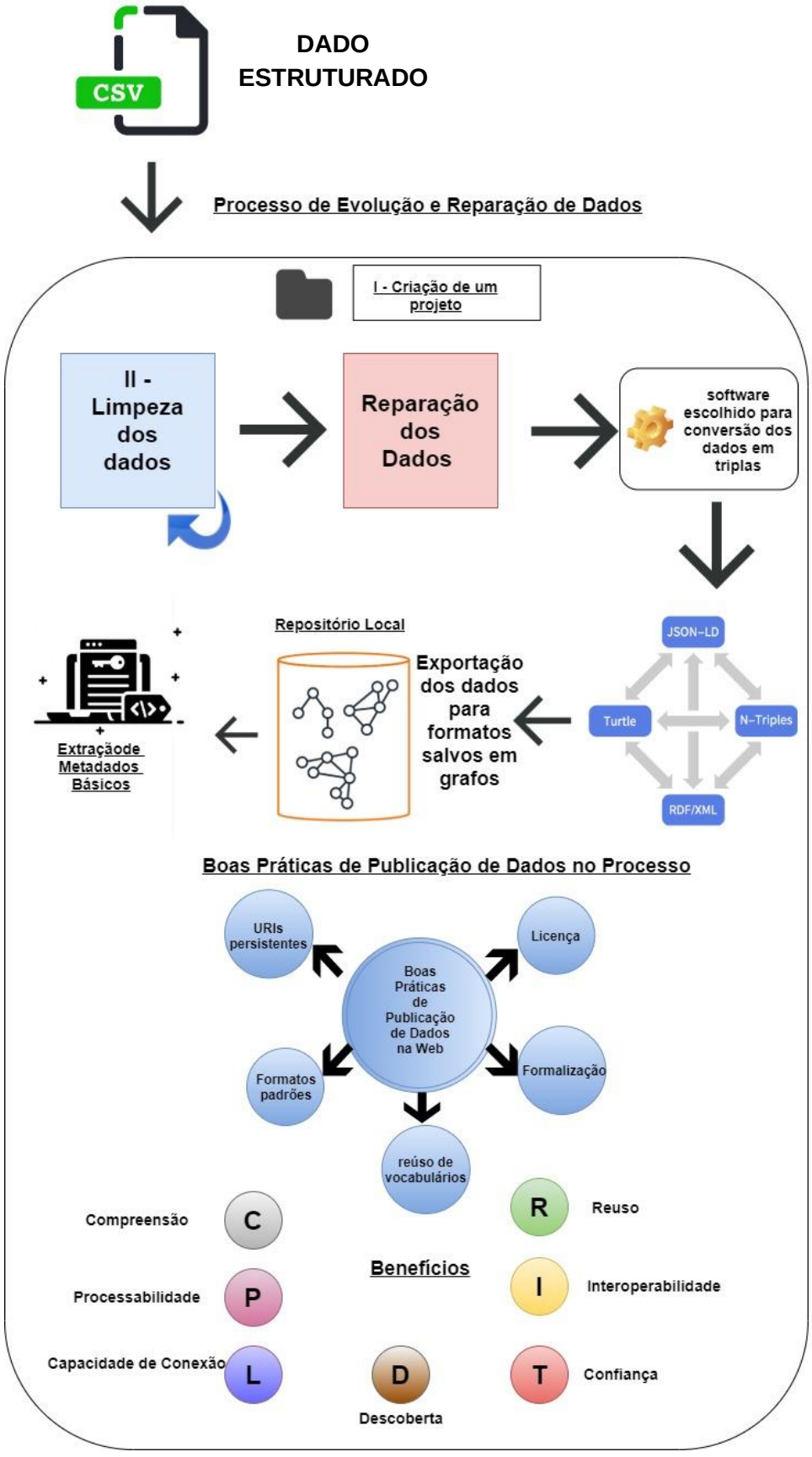
A diretriz de limpeza e reparação dos dados abarca todos os benefícios citados, como, desde o fornecimento de metadados de forma descritiva e estruturada; gerar informações sobre a licença e procedências dos dados; fornecimento de informações de qualidade dos dados e qual indicador e histórico da versão.

Usar formatos de dados padronizados legíveis por máquinas é de extrema importância neste contexto, uma vez que as representações de dados já terão características independentes de localidade, e fornecerão assim os dados em formatos múltiplos, permitindo a reutilização de vocabulários, dando preferência aos padronizados.

A escolha referente ao nível de formalização adequado também é um dos benefícios determinados no mapeamento, pois permite fornecer uma explicação para os dados que não estão disponíveis, criar uma documentação completa e evitar alterações que afetem o funcionamento das APIs.

Preservar os identificadores permite que se avalie a cobertura do conjunto de dados, onde se coleta um *feedback* de consumidores desses dados e o *feedback* referente ao processo de compartilhamento disponível, gerando as **Diretrizes** da próxima seção que é o enriquecimento dos dados por meio da geração de novos dados.

Figura 40 - Diretriz 2 - Limpeza e Reparação de Dados



Fonte: Elaborda pela autora (2020)

Sempre deve fornecer o *feedback* para o publicador original, obedecendo aos termos de licença e à citação da publicação original do conjunto de dados.

6.1.3 *Diretriz 3: Criação de Colunas URIs até a Representação dos dados na Web de Dados*

Uma vez que as **Diretrizes 1 e 2** foram aplicadas e os dados já estão limpos e reparados, por consequência, a boa prática relacionada a este momento identifica-se com a persistência dos dados.

É nesse momento que se criam as colunas das URIs persistentes como identificadores dentro do conjunto de dados que já estão estruturados, necessários para publicá-los na *Web*. Verificam-se etapas importantes:

- Modelagem, que está relacionada ao uso/desenvolvimento de um vocabulário e uso de ontologias para descrever os metadados para que o dado se torne semântico, ou seja, uma tripla RDF, definidas as ações como:
    a. Selecionar vocabulários/ ontologias;
    b. Criação de Mapa; e
    c. Atribuir URIs.
- Geração de recursos RDF, que estão relacionados a:
    a. Selecionar as tecnologias para a geração do RDF;
    b. Transformar dados de origem em RDF; e
    c. Validar.

A representação dos dados está diretamente ligada ao encontrar um vocabulário/ontologia que esteja relacionado ao domínio da aplicação em si para que se possa estar adequado aos dados ali representados.

Em alguns casos, algumas aplicações preferem criar o próprio vocabulário e a própria ontologia, mas como determinado por Nhacuongue, Rozsa e Dutra (2018) são casos de extrema necessidade referente a um domínio em que não se encontra nenhum metadado que seja equivalente ao processo que se está empregando, pois através de um repositório, ou mesmo através do LOV, existe uma vasta nuvem com diferentes vocabulários e que

são extremamente recomendados pelo W3C para representar as informações requeridas, porém o processo da ontologia pode ser mais específico e assim sua criação necessária.

Se o vocabulário/ontologia estiver disponível e o termo corretamente selecionado, optamos sempre por uma ferramenta para definir esse contexto. Para melhorar a organização das colunas de dados, a ferramenta escolhida no processo anterior permitirá reordenar ou criar critérios de uso desses dados.

A **Diretriz 3** deve se repetir a toda URI persistente como identificadores de conjunto de dados e estar associada a um vocabulário, gerando sempre um termo correspondente.

Todas as diretrizes apresentadas até o momento (**Figuras 38**, **39** e **40**), apresentam a evidência de uma tripla RDF resultante para uma determinada classe do domínio escolhido.

Neste momento, as URIs e propriedades estão representadas de acordo com o vocabulário escolhido e a ontologia aplicada.

O último passo relacionado à **Diretriz 3** está relacionado à configuração de propriedades dos dados referente ao tipo especificado da URI informada a uma URI do tipo decimal a partir de um endereço *Web*.

Todo o processo deve-se repetir para todas as URIs geradas, selecionando as colunas para a transformação em RDF, e assim gerando a organização e a representação dos dados em dados conectados.

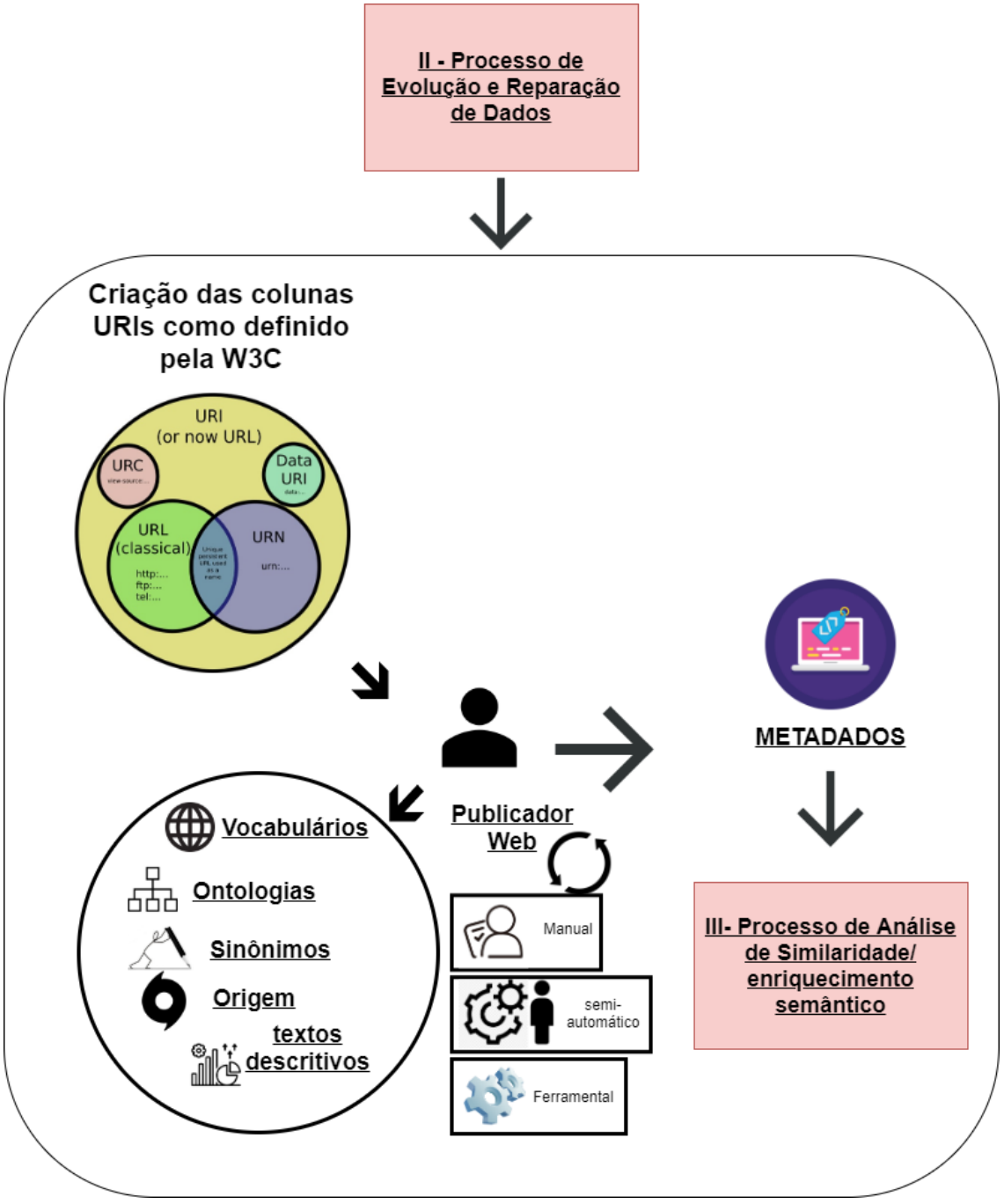
Observamos, logo após essa organização, que o processo está relacionado à representação dos dados, pois uma vez que os dados já estão limpos, reparados e organizados, é necessário designar sua representação de forma com as boas práticas de Dados abertos e Conectados.

Finalizando, o modelo gerado deve ser guardado, ou salvo, pois é aqui que ocorre a verificação de uma representação em tripla RDF para que seja aplicado em uma futura interface que será usada e aplicada, como será discutido na **Diretriz 5**.

Outro ponto a ser discorrido é a necessidade de uma revisão das URIs para a futura Publicação de dados na *Web*, uma vez que são necessárias manutenções das informações escolhidas, diretamente da fonte de onde os dados foram buscados.

Para escrever cada passo da **Diretriz 3** a **Figura 41** ilustra o esquema RDF para ser aplicado a qualquer domínio de estudo de caso, apresentando a organização e a representação dos dados conectados como finalizadas.

Figura 41 - Diretriz 3 - Criação de Colunas URIs até a Representação dos dados Abertos na *Web* de Dados



Fonte: Elaborada pela autora (2020)

Ainda na **Figura 41,** o contexto do processo de Análise e Similaridade está relacionado à exportação dos dados em formato RDF e ao seu processo de enriquecimento, que será apresentado na seção da **Diretriz 4.**

É importante ressaltar que ainda existem alguns casos e preferências de visualização do domínio, e que em todos os passos apresentados podem ser usadas técnicas que envolvem pessoas no levantamento dos dados de forma manual, semiautomática, gerenciando um processo mais longo ou demorado na maioria dos casos, mas isso não significa que não seja eficaz.

Porém, com a quantidade enorme de dados e a facilidade com que as ferramentas estão sendo apresentadas e de licenças abertas (*open source*), verifica-se um aumento na utilização de ferramentas.

Independente da ferramenta utilizada, gera-se armazenamento de arquivos considerados como *backup*, ou cópia do projeto e ainda um documento com todos os passos realizados sobre a massa de dados. Rautenberg *et al.* (2018) afirmam que o arquivo de passos pode ser reutilizado numa próxima transformação de dados, uma vez que toda limpeza, reparo e organização e transformação dos dados em RDF podem ser reutilizados.

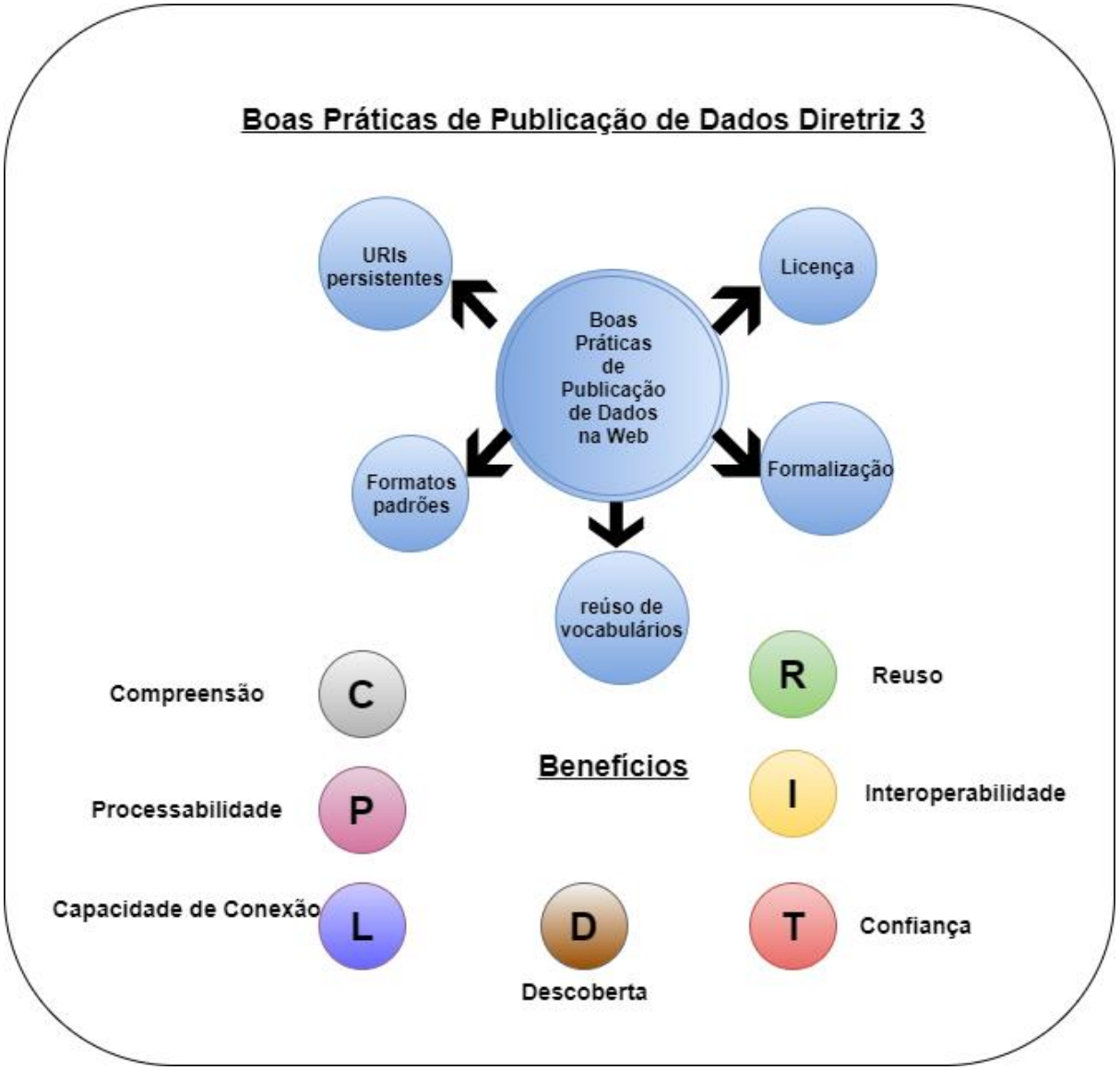
As boas práticas envolvidas na **Diretriz 3** estão associadas ao uso de URI persistentes como identificadores dentro de um determinado conjunto de dados; o uso de formato de dados padrão e que sejam legíveis por máquinas; reuso de vocabulários; escolha de nível de formalização e termos de licença.

Em relação às URIs persistentes, ao evoluir os dados utilizaram-se URIs persistentes como identificadoras dentro do conjunto de dados do domínio proposto, e em relação ao formato de dados, sabe-se que foi padronizado para que máquinas pudessem entender o conteúdo.

O reuso empregado está relacionado ao reuso de vocabulários padronizados e disponibilizados na LOV, e o nível de formalização relaciona-se ao nível de formalismo escolhido, que deve estar ligado ao uso dos vocabulários comuns na *Web* de Dados.

A licença está relacionada ao conjunto de dados gerados, seguindo os termos de licença original dos dados. Como ilustrado na **Figura 42**, referente às Boas Práticas e aos benefícios dos dados abertos no processo de **Diretriz 3.**

Figura 42 - Diretriz 3 - Boas Práticas e Benefícios Relacionados



Fonte: Elaborada pela autora (2020)

Seus benefícios com adoção das boas práticas citadas são:

- Reuso: permite o uso de todas as boas práticas descritas pela W3C;
- Descoberta: uma vez que o uso das URIs seja persistente, os elementos dos conjuntos de dados geram informações provenientes de dados;
- Ligação: também está relacionada ao uso de URIs persistentes como descrito no benefício do reuso;
- interoperabilidade dos dados: esse benefício contempla todas as possibilidades para o processo e a continuidade das diretrizes, permite URIs persistentes como identificadoras de conjunto de dados, o reuso dos vocabulários padronizados e disponibilizados na LOV, o nível de

formalização dos dados de forma certa e o *feedback* para o publicador original;

- processabilidade: está relacionado também à formalização dos dados legíveis por máquina, o que consiste em fornecer dados em múltiplos formatos. Permite também o reuso dos vocabulários padronizados e gera um benefício que impacta na próxima diretriz que é o enriquecimento de dados gerando novos dados;
- compreensão: reuso de vocabulários padronizados, formalização dos dados e ao processo de enriquecimento de dados, que será apresentado na **Diretriz 4**, pois esse enriquecimento permitirá a geração de novos dados;
- confiança nos dados: reuso de vocabulários padronizados, enriquecimento dos dados em novos dados (**Diretriz 4**), feedback para o publicador original, seguindo os termos de licença e citando o publicador original nesse processo.

6.1.4 *Diretriz 4: Processo de Análise e Similaridade/Enriquecimento e Mapeamento Semântico*

A **Diretriz 4** está relacionada ao processo de enriquecimento semântico dos dados. Como já apresentado anteriormente, esse processo é composto pelo processo de anotação semântica que descreve, segundo Eller (2014), todo o seu conteúdo pela associação de palavras relevantes do texto em si e seus conceitos presentes na Ontologia. No caso do enriquecimento de dados abertos semânticos, pode ser definido como uma especificação de uma conceituação.

Observamos, na **Figura 41** que na **Diretriz 3** o resultado do processo gerou um conjunto de metadados básicos que foram extraídos. Básicos, pois compõem informações mínimas para reproduzir a referência do dado.

Esses metadados ainda não estão em processo de consumo, pois só trazem o nome do metadado, sua tipagem e seu tamanho. Para que possam chegar à **Diretriz 5**, que representa a fusão de dados a serem consumidos de forma enriquecida, precisam passar pelo conceito de Anotação Semântica, que é composta por três etapas:

- Análise de Similaridade Léxica;
- Análise de Similaridade Semântica; e
- Enriquecimento Semântico dos dados em si.

Para explicar a Anotação Semântica, Smith et al. (2012 apud Falbo, 1998) discutem que a conceituação pode ser classificada como uma abstração, uma visão simplificada do domínio que se quer conceber.

Esse propósito se dispõe com o intuito de executar os seguintes pontos: a permissão de que múltiplos agentes possam compartilhar seus conhecimentos com o intuito de que pessoas possam compreender melhor um domínio específico e, ainda mais, permitirem o repasse do entendimento desse domínio de forma que uma conceituação identifica o objeto e relações que existem no universo lógico.

Como apresentado na **Diretriz 3** na **Figura 41**, visualizam-se esforços manuais, uma vez que estão relacionados à ação humana no processo dessas anotações semânticas.

Lira (2014, p. 29) afirma que nessa fase de busca em algum repositório de metadados, consegue-se decidir se é possível reutilizar de algum metadado já existente ou se, a partir dessa intervenção humana, é necessária uma nova criação.

Permitir que múltiplos agentes compartilhem seus conhecimentos e possam ajudar as pessoas a compreender melhor um determinado domínio de conhecimento, permitir que pessoas possam atingir um consenso no seu entendimento sobre esse domínio, baseado neste contexto que Ontologias podem ser reutilizadas de diversas formas. Uma delas pode resultar na criação de uma ontologia própria que se baseia nos conceitos de outras, ou pode manter as ontologias originais, que podem ser classificadas como alinhamento de Ontologias.

Campos (2010) afirma que o alinhamento das ontologias permite que sejam reutilizadas de forma inalterada e em seus locais de origem. Entretanto é gerado um conjunto de vínculos (*links*) entre essas ontologias, contendo um conjunto de informações que permitem que se compatibilizem as ontologias reutilizadas apresentando um modelo persistente.

O mapeamento, denominado na literatura como *mapping,* gera um conjunto de *links* expressos em um modelo persistente produzido pelo processo de alinhamento entre as ontologias, a partir das boas práticas determinadas pela W3C.

Esse mapeamento depende dos metadados relacionados ao tipo de vínculo semântico identificado nesses metadados e no tipo de formalismo que foi utilizado na ontologia escolhida para representar a sua semântica no domínio ou contexto gerado pelo usuário.

As semelhanças podem ser classificadas em diferentes graus, ou ainda podem fazer parte de outros contextos, ou ainda podem ter algum outro tipo de relacionamento que pode ser identificado com o auxílio de um entendido na área. É nesse contexto que entra o mapeamento de semelhanças que podem expressar diferentes graus de similaridade, apresentado por Felicíssimo (2004), Breitman (2005), Kalfoglou e Schorlemmer (2003), Aleksovski *et al.* (2006).

Para se determinar o grau de similaridade, geralmente diversos fatores são levados em conta, tais como: similaridade linguística entre os termos, compatibilidade dos seus atributos, posicionamento do termo na estrutura hierárquica da ontologia, dentre outros.

Para a apresentação da **Diretriz 4** do trabalho proposto, esse mapeamento está relacionado à Análise de similaridade léxica, a similaridade semântica e o enriquecimento do conjunto de metadados para disponibilização em uma *Triple Store*, que será apresentado na última **Diretriz 5.**

Um dos aspectos do mapeamento é a questão de como se apresenta um levantamento consistente sobre os tipos de conflitos ao se mapear ontologias, segundo Bruijin *et al.* (2005), no que se diz respeito ao tipo de técnicas que serão utilizadas. Alguns pontos importantes podem basear esse mapeamento:

- na semelhança dos nomes dos termos;
- na estrutura da ontologia, levando em conta o posicionamento dos termos na estrutura hierárquica das ontologias sendo comparadas;
- tipos de relações que sejam utilizadas de forma semelhante nas ontologias comparadas, segundo Euzenat e Shvaiko (2013);

- na inclusão de conhecimento adicional, ontologias e vocabulários que possuam uma hierarquia de conceitos, segundo Miller (2004), que possam ser utilizadas, para procura de sinônimos ou para confronto da distância do posicionamento dos termos das ontologias para serem mapeadas, segundo Hamdi, Reynaud e Safar (2010).

Com todos esses processos descritos e já estudados no capítulo 5, verifica-se que a **Diretriz 4** usa desse mapeamento para que o processo de anotação semântica venha enriquecer os metadados de forma a transformá-los em semânticos, usando ou não a intervenção manual, semiautomática ou automática, como já apresentada na Figura 43, estão associadas ao uso de URI persistentes como identificadores dentro de um determinado conjunto de dados; o uso de formato de dados padrão e que sejam legíveis por máquinas; reuso de vocabulários; escolha de nível de formalização e termos de licença.

Em relação às URIs persistentes, ao evoluir os dados utilizaram-se URIs persistentes como identificadoras dentro do conjunto de dados do domínio proposto, e em relação ao formato de dados, sabe-se que foi padronizado para que máquinas pudessem entender o conteúdo.

O reuso empregado está relacionado ao reuso de vocabulários padronizados e disponibilizados na LOV, e o nível de formalização relaciona-se ao nível de formalismo escolhido, que deve estar ligado ao uso dos vocabulários comuns na *Web* dos Dados.

A licença está relacionada ao conjunto de dados gerados seguindo os termos de licença original dos dados. Como ilustrado na **Figura 42**, referente às boas práticas e aos benefícios dos dados abertos no processo da **Diretriz 3**, a partir do processo de Similaridade.

Esse processo de Similaridade é composto por três etapas determinadas por Madhavan, Bernstein e Rahm (2001): a Normalização, a categorização e comparação. Aponta-se visualizar a necessidade que cada domínio proposto deve usar para que possa ser enriquecido com o uso dessas etapas.

No caso que domínio escolhido tenha a necessidade de mapear termos equivalentes e seus significados para um processo de normalização, e para isso é necessário usar Tesauros que permitam o relacionamento dos termos comuns, uma vez que possam ter nomes diferentes em outros esquemas.

Já relacionado à categorização dos termos, são arranjados em classes, o que restringe o número de comparações entre termos diferentes. E em relação ao contexto de comparação, define um ponto de similaridade entre os termos e seus grupos.

Importante ressaltar que, caso for escolhido o contexto de comparação, onde segundo os trabalhos correlatos apresentados no capítulo 2 apresentam na sua maioria, devem-se analisar dois aspectos importantes, a Análise de Similaridade Léxica e a Análise de Similaridade Semântica.

Como apresentado na **Figura 43**, referente à **Diretriz 4**, a Análise Léxica tem, por objetivo, avaliar duas sequências de caracteres pelo número mínimo de operações necessárias para transformar uma cadeia em outra cadeia de valor, segundo a distância de Levenshtein, que é classificada como uma métrica de *string* para medir a diferença entre duas sequências, segundo Ruberto e Antoniazzi (2017), e é classificado com o nome de *Edit Distance*. Também possui a segunda forma, classificada como *Stemmer* que consiste na avaliação de sequência de caracteres pela redução de uma palavra na sua base.

Já em relação à Análise de Similaridade Semântica está relacionada à expectativa da avaliação semântica entre os termos. Nesse ponto, é imprescindível o uso de um tesauro que avalie as terminologias e conceitos de suas relações, segundo Noll, Saccol e Edelweiss (2007, p. 16).

O resultado desse processo consiste em ter um conjunto de metadados enriquecidos, utilizando do mapeamento e da aplicação da anotação semântica que teve como resultado ou da inclusão das anotações semânticas ou da reutilização de anotações no conjunto de metadados básicos, segundo afirma Lira (2014).

Ao final do processo de anotação semântica da transformação dos metadados em metadados enriquecidos e semânticos, eles são armazenados em repositórios de metadados e repositórios de vocabulários, para que possam ser reutilizados posteriormente, visualizando que sua exportação pode ser feita em formatos como XML ou CSV.

O Repositório de Metadados armazena o conjunto de anotações semânticas associadas aos metadados enriquecidos, como apresentado na forma de uma base de dados na **Figura 43**, pois assim podem ser buscados,

uma vez que está ligado ao contexto de reuso, caso o processo necessite fazer uma atribuição nas descrições aos metadados básicos.

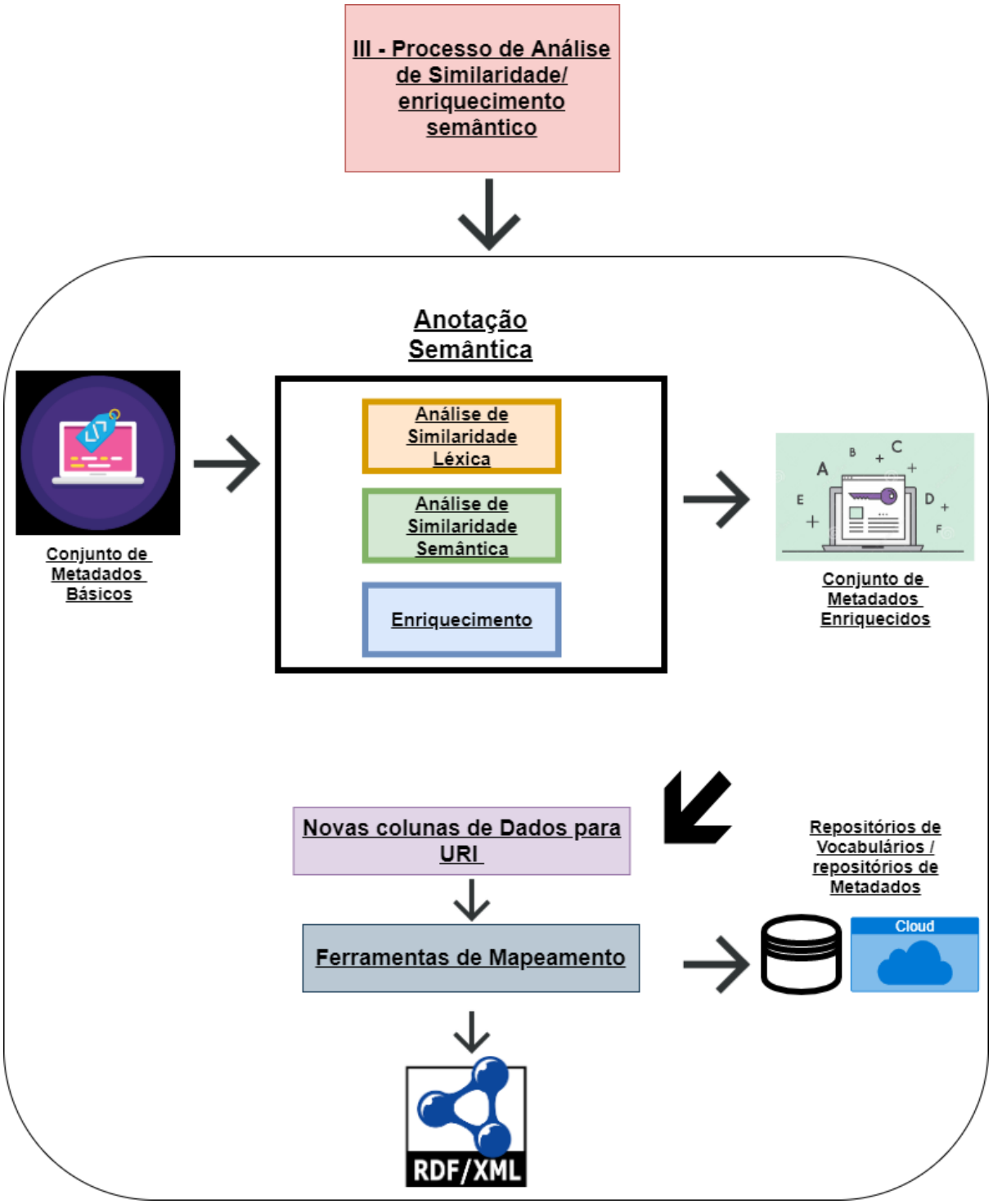
Já o componente do Repositório de Vocabulários está ligado a base específica indicada para armazenar os vocabulários padrões, que servirão de suporte para alimentação do atributo “sinônimos” usados para agregação de informação por anotação semântica, segundo Lira (2014).

Serão criadas colunas de dados para as URIs, sendo elas persistentes e geradas como identificadores dentro do conjunto de coluna de dados, representando-as de acordo com o vocabulário escolhido, gerando sua permissão e alterações no esquema RDF e as configurações de suas propriedades. Neste ponto, a organização e a representação dos dados como dados conectados estão finalizadas.

Na **Figura 43**, referente à **Diretriz 4**, também são apresentadas ferramentas que mapeiam todo esse conteúdo. Diversas ferramentas, já apresentadas no trabalho, possuem essa função, e são usadas conforme a necessidade do domínio e sua devida especificação.

Em muitos dos estudos apresentados no decorrer do trabalho, algumas ferramentas de mapeamento que aparecem com muita frequência são o *Triple Store Virtuoso* e o *Triple Store Jena Fuseki*, apenas como citação de exemplo, gerando o armazenamento dos recursos em RDF e podendo ser representados em formato de grafos de propriedades baseadas em RDF ou em Tabelas SQL.

Figura 43 - Diretriz 4 - Enriquecimento e Mapeamento de Dados Abertos Semântico



Fonte: Elaborada pela autora (2020)

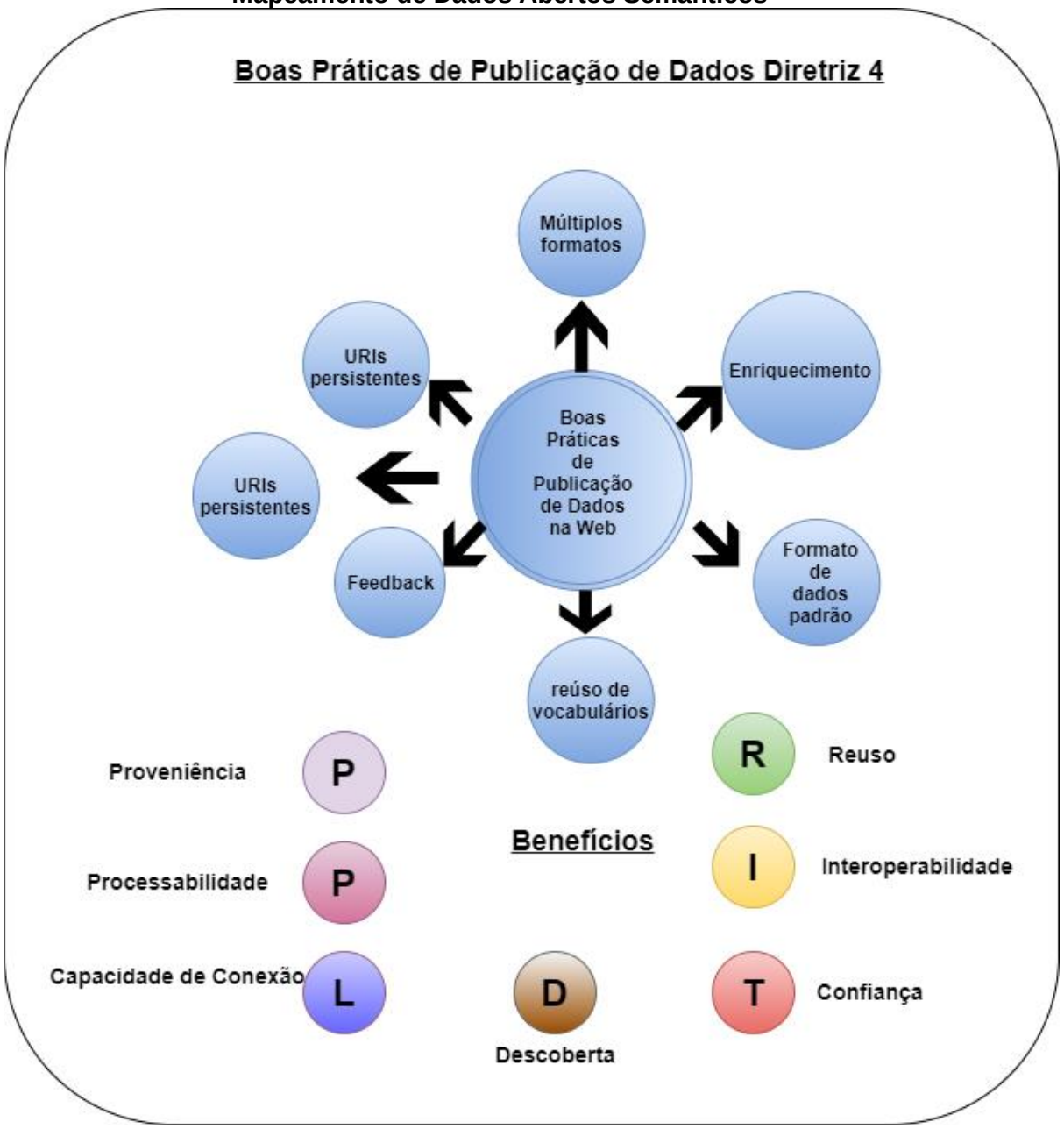
As boas práticas envolvidas nesta **Diretriz 4**, **Figura 44**, estão associadas ao Reuso dos metadados enriquecidos e suas anotações, pois uma vez os dados estando abertos e semânticos, qualquer entidade, instituições ou setores, usando de outra boa prática, que é enriquecimento dos

dados, podem gerar novos metadados enriquecidos. Seu armazenamento e recuperação, estando em um repositório, sugere que o conteúdo seja relevante e ainda utiliza uma padronização e formalização, onde diversas pessoas poderão compartilhar do mesmo conceito e com padrões legíveis por máquina, além de apresentar um feedback para o publicador original.

Benefícios que a **Diretriz 4** oferece:

- Reuso: permite que publicadores de dados abertos faça o reuso do processo da anotação semântica dos metadados enriquecidos;
- Confiança: *Feedback* para o publicador original - geração de credibilidade aos metadados, pois o publicador original recebe um *feedback* desse contexto;
- Processabilidade: possibilita a usabilidade dos dados e metadados, uma vez que são enriquecidos utilizando-se de formato de dados padrões legíveis por máquinas e fornecidos em multiplataformas, e fazem padrões de vocabulários já existentes e ontologias específicas. Isso quer dizer que a abordagem permite uma inserção de muitos vocabulários, o que significa sem limite máximo conhecido, gerando uma qualidade;
- Proveniência: está atrelada à abordagem de disposição de atributos para sua descrição, pois gera reuso dos metadados, uma vez que sua descrição gera informações de proveniência, permitindo a validação de um histórico e ainda uma confiabilidade aos dados e metadados;
- Interoperabilidade, Descoberta e Ligação: uso de URIs persistentes dentro do elemento de conjuntos de dados e fornecimento de informação dos dados.

Figura 44 - Diretriz 4 - Boas Práticas e Benefícios do Enriquecimento e Mapeamento de Dados Abertos Semânticos



Fonte: Elaborada pela autora (2020)

6.1.5 *Fusão de Dados*

A Diretriz classificada como Fusão de Dados é determinada por Auer (2014) no ciclo de vida de dados abertos e conectados como processo de armazenamento de dados (*Triple Store*). Como apresentado na **Diretriz 4,** o mapeamento dos dados e metadados enriquecidos gera uma *Triple Store*.

Esse processo de mapeamento, utilizando um determinado *software* permite o acesso, a integração e gerenciamento de dados relacionais,

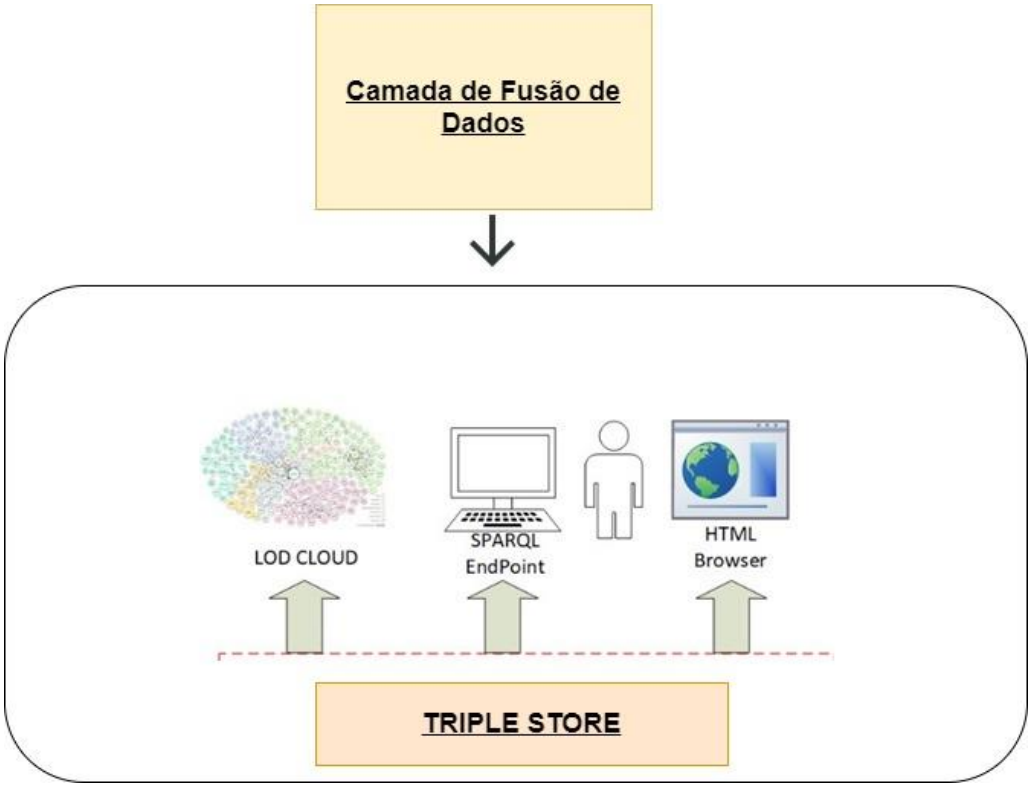
incluindo tabelas SQL ou através da apresentação de grafos de propriedades baseados em RDF, permitindo consultas semânticas.

Esses *softwares* utilizados para mapeamento devem conter um servidor universal de multi modelos, uma unidade central de processamento, classificado como *multithreading*, que possa fornecer vários *threads* de execução simultaneamente, suportados pelo sistema operacional, utilizando múltiplos protocolos que são relacionados com Sistemas Gerenciadores de Banco de Dados, como o padrão ACID (*Atomicity, Consistency, Isolation, Durability*), como descrito por Elmasri e Navathe (2005), além de suportar vocabulários, disponibilizados na *Web* de Dados, como por exemplo o conjunto LOV.

Alguns dos *softwares* que podem ser utilizados para esse mapeamento suportam diversos modelos de dados, como documento, chave-valor, grafo, XML, relacional entre outros. Os suportes relacionados à arquitetura devem ser uma plataforma inteligente, que seja orientada aos metadados, gerando um controle centralizado dos dados e aplicativos que podem ser integrados no contexto do domínio especificado pelo usuário.

Segundo González e Antonio (2011) e Auer (2014), existem alguns pontos importantes a serem suportados por esses *softwares*: o gerenciamento de dados de tabelas relacionais; o gerenciamento de dados de grafos de propriedade; o gerenciamento de conteúdo; a *web* e outros serviços de arquivos de documentos; dados abertos conectados 5 estrelas; servidor de aplicação *Web*, como apresentados na **Figura 45**.

Figura 45 - Diretriz 5 - Fusão de Dados



Fonte: Elaborada pela autora (2020)

Alguns exemplos relacionados aos pontos a esses *softwares*: consultas semânticas utilizando dados de grafos com SPARQL, RDF baseado em *quadstore*. Em relação ao aumento do conteúdo, alguns formatos que podem estar relacionados: HTML, TEXT, RDF/XML, JSON, JSON-LD e XML, em relação à *Web* ou os serviços de documentos de arquivos, pode citar o servidor de arquivos, além do que o servidor dos dados conectados deve ser baseado em RDF, que está relacionado com a categorização 5 estrelas de Berners-Lee (5 [ESTRELAS], 2012), além de poder criar modos de interação como APIs (*Application Programming Interface*), que são aplicações e sistemas inteligentes utilizados, que permitem a navegação *Web*.

Grande parte das APIs são criadas para integração entre um *software* de um determinado domínio e informações associadas, que podem ser utilizados por uma rede para tráfego de informações de uma forma padronizada, que seja por uma rede local ou pela Internet, classificando-se em três categorias: local, baseada em Programa e baseada em *Web*.

Como exemplo, pode-se citar a API RESTful, que é uma API em *web*, fundamentada no protocolo HTTP, um formato padrão de uso que se baseia

em soluções digitais hospedadas em nuvem. Outros padrões que podem ser citados e abordagens que estão sendo utilizados com mais frequência são gRPC[44] e GraphQL[45].

Outro exemplo baseado em *Web* é a SOAP (*Simple Object Access Protocol*), que é um protocolo para troca de informações estruturadas em uma plataforma descentralizada e distribuída. Segundo a W3C, *Soap* (2007), sua especificação define um arcabouço que provê maneiras para se construir mensagens que podem trafegar através de diversos protocolos e que foi especificado de forma a ser independente de qualquer modelo de programação ou outra implementação específica, além de serem implementados através de servidores HTTP e as mensagens SOAP são documentos XML que aderem a uma especificação W3C.

Berners-Lee (2006a, não paginado, tradução nossa) afirma que “o *linked data* é essencial para, de fato, conectar à *Web* Semântica”, o que pode ser traduzido como o fator motivador de ligação de dados abertos e conectados.

Segundo Santarem Segundo (2015, p. 225) afirma que LOD - *LInked Open Data*, ou traduzindo “dados abertos conectados” e pode ser determinado “como a melhor forma de materialização dos conceitos e tecnologias da *Web* Semântica”.

Classifica-se como um projeto, que usa conceitos de ligação semântica da *Web* de Dados, que segue um conjunto de normas, mas como algumas especificações indicando um grau de exigência maior na constituição de sua rede de interligações. Porém, ainda se visualiza que nem todos os dados conectados (*Linked data*), ou conjunto de dados ligados podem ser classificados como dados abertos e, por consequência, não fazem parte do LOD.

O projeto LOD tem seu potencial quando os dados são publicados como *Linked Open Data*, tendo uma maior visibilidade tecnológica. A publicação dos dados na *Web* de forma conectada e aberta, fundamentada em todas as regras do modelo atendido e ainda conectadas a outros conjuntos de LOD, considera-

[44] gRPC é uma estrutura de RPC (chamada de procedimento remoto) de alto desempenho independente de linguagem (LUO E NEWTON, 2021).

[45] GraphQL é uma linguagem de consulta e ambiente de execução voltado a servidores para as interfaces de programação de aplicações (APIs) cuja prioridade é fornecer exatamente os dados que os clientes solicitam e nada mais (REDHAT, 2021).

se que esses dados possuem padrão 5 estrelas de dados abertos, como determinado por Berners-Lee (5 [ESTRELAS], 2012).

O projeto *Linking Open Data*, da W3C (W3C, 2015a), chamado como *LOD-Cloud*[46] tem, como objetivo, publicar várias bases de dados abertas em RDF e estabelecer links RDF entre itens de dados de diferentes fontes de dados e traz, como resultado, a visibilidade dos conjuntos de dados estruturados em RDF no padrão LOD, por serem disponibilizados para reuso de forma livre e gratuita. Os últimos dados publicados na página do projeto *The Linked Open Data Cloud* se encontram com 1269 conjuntos de dados abertos e conectados.

As Boas Práticas adotadas neste processo da **Diretriz 5, Figura 46,** estão relacionadas ao uso de URIs persistentes como identificadores dentro do conjunto de dados, uso de formatos de dados padrão e legíveis por máquina, reuso de vocabulários, dando preferência aos padronizados e que se encontram na LOV, a escolha do nível de formalização correto e seguir todos os termos de licença. Todas elas relacionadas para armazenamentos dos dados relacionados à sua evolução, como já descrito na **Diretriz 3**.

Seus benefícios com adoção das boas práticas citadas são:

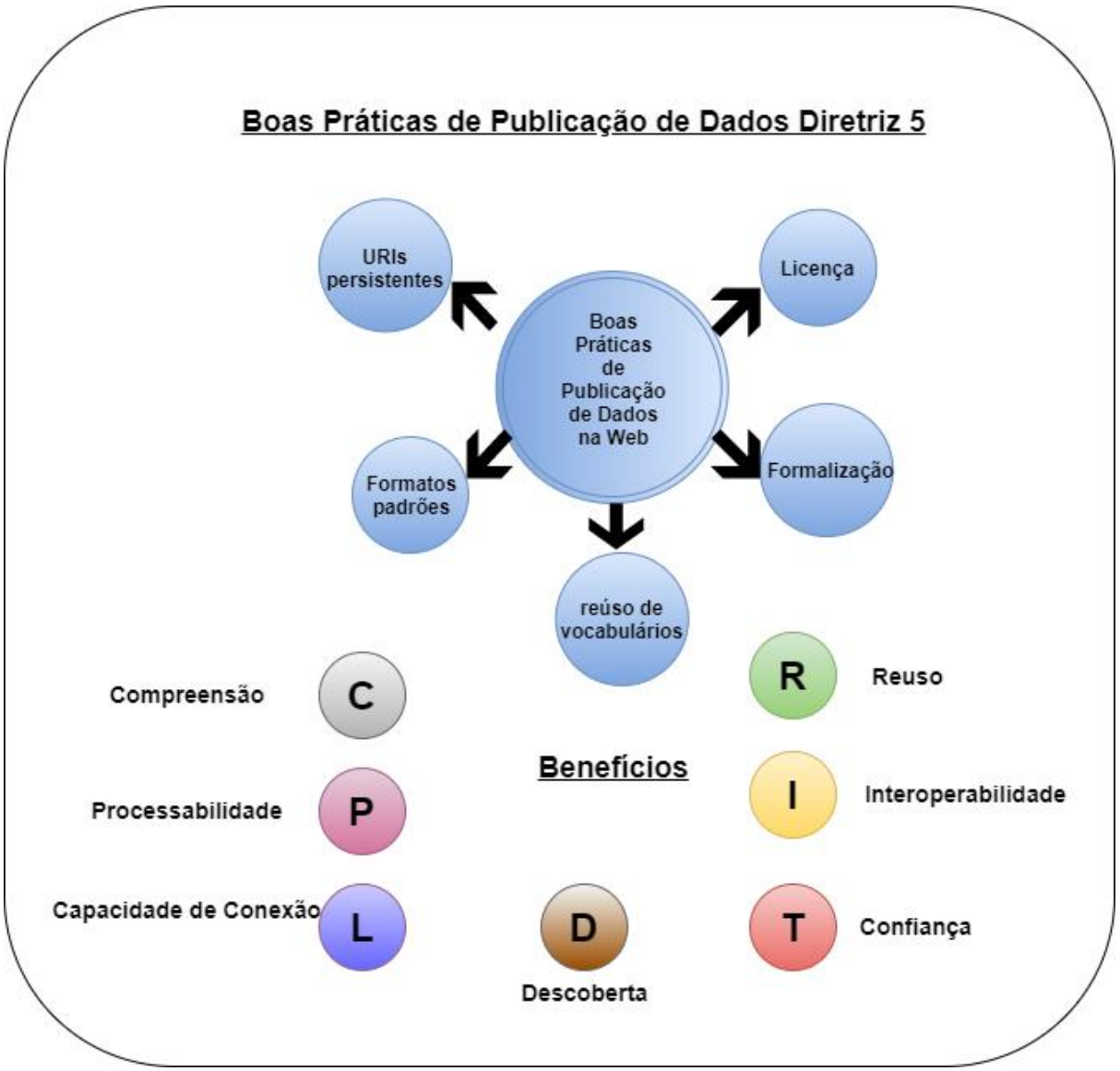
- reuso: permite o uso de todas as boas práticas descritas pela W3C;
- descoberta: uma vez que o uso das URIs seja persistente, os elementos dos conjuntos de dados geram informações provenientes de dados;
- ligação: também está relacionada ao uso de URIs persistentes como descrito no benefício do reuso;
- interoperabilidade dos dados: esse benefício contempla todas as possibilidades para o processo e a continuidade das diretrizes, permite URIs persistentes como identificadoras de conjunto de dados, o reuso dos vocabulários padronizados e disponibilizados na LOV, o nível de formalização dos dados de forma certa e o feedback para o publicador original;
- processabilidade: está relacionada também à formalização dos dados legíveis por máquina, o que consiste em fornecer dados em múltiplos formatos. Permite também o reuso dos vocabulários padronizados e gera um

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

benefício que impacta na próxima diretriz que é o enriquecimento de dados gerando novos dados;

- compreensão: reuso de vocabulários padronizados, formalização dos dados e ao processo de enriquecimento de dados, que será apresentado na **Diretriz 4**, pois esse enriquecimento permitirá a geração de novos dados;
- confiança nos dados: reuso de vocabulários padronizados, enriquecimento dos dados em novos dados (**Diretriz 4**), *feedback* para o publicador original, seguindo os termos de licença e citando o publicador original nesse processo.

Figura 46 - Diretriz 5 - Boas Práticas e Benefícios da Fusão de Dados



Fonte: Elaborada pela autora (2020)

6.2 PROVA DE CONCEITO: ENRIQUECIMENTO E MAPEAMENTO SEMÂNTICO DE DADOS ABERTOS E CONECTADOS DO METRÔ DE SÃO PAULO

Para a presente pesquisa tem os seguintes pressupostos iniciais oriundos da problemática proposta. Nessa perspectiva, apresenta-se um ***Framework*** que oriente e que apresente diretrizes com ênfase ao enriquecimento e mapeamento semântico de dados, baseadas nas boas práticas de Publicação de Dados na *Web*, com os dados do Metrô de São Paulo.

Apresentando cinco diretrizes específicas de forma conceitual, a proposta da tese vem com o intuito de compreender esse processo para que independente de domínio aplicado, são processos informações e contextualizar um exemplo aplicado.

A solução proposta não é apenas contextualizar um resultado com uso de um determinado vocabulário, ontologia e ferramenta adequada, mas sim abstrair e descrever o contexto do processo informacional do enriquecimento e mapeamento semântico para que se chegue à condição da publicação dos dados na *Web*, de forma sistemática, em que uma das possibilidades de apresentação desses dados esteja em formato de um grafo *Resource Description Framework* (RDF) em consonância com os princípios Linked Data.

A partir da **Figura 47**, que se verifica o *Log* da aplicação quando requisitada uma informação, que já foi avaliada e definida como a correta, utilizou os diversos elementos que foram obtidos a partir das diretrizes informacionais conceituais apresentadas.

A aplicação de conceito realizada passou por todas as diretrizes propostas pelo ***Framework***, demonstrando o seu funcionamento, atendendo o enriquecimento e mapeamento de dados em semânticos, de forma que se torne em uma *Tripe Store*, ou melhor, em um repositório, para um uso futuro.

Para contextualizar e trazer alguns resultados referentes ao estudo apresentado, nessa seção é apresentada uma proposta de um aplicativo que traga dados enriquecidos semanticamente, a partir do uso dos dados abertos na *Web*, tendo como estudo de caso os dados do Metrô da cidade de São Paulo. Esta aplicação de conceito foi um estudo que já está publicado na

Revista e-F@tec e intitulado como “Aplicativo que gera enriquecimento semântico de dados abertos” (SILVA; LUZ, 2020).

O Metrô é um meio de transporte extremamente importante principalmente em cidades grandes no Brasil, como as cidades do Rio de Janeiro e São Paulo, que são movidas por esse meio de transporte.

Verifica-se, então, que ainda hoje, além da importância da locomoção em si, existe a importância do usuário poder saber quais as linhas, horários e todas as possibilidades de informações de seu funcionamento.

Fundamentada nesse contexto e na linha de pesquisa que envolve *Web* Semântica e todo o processo de dados abertos e informações em relação ao transporte público, realizou-se uma pesquisa exploratória por meio de estudo bibliográfico que permitiu a criação de um aplicativo que traga informações de busca semântica, preservando-o como aplicado.

Desta forma, a abordagem do trabalho publicado foi qualitativa, pois foi feito um levantamento e análise desses dados abertos, trazendo como resultados parciais o desenvolvimento de um aplicativo que enriquecesse esses dados abertos do Metrô da cidade de São Paulo, com a base no enriquecimento semântico dos dados.

O aplicativo gera enriquecimento semântico a partir de dados abertos, possibilitando um fácil acesso e reuso, tendo como estudo de caso os dados do Metrô de São Paulo.

O trabalho apresentou o processo de extração de dados abertos (usando os conceitos delimitados nas **Diretrizes 1** e **2**) dessa base e o processo de anotação semântica (apresentado na **Diretriz 3**); a realização do mapeamento dos dados abertos (apresentado na **Diretriz 4**); a realização do enriquecimento semântico através de um aplicativo (descrito na **Diretriz 5**).

Para maior detalhamento das **Diretrizes 1, 2** e **3,** os dados extraídos no exemplo do Metrô, foram através do Portal de Transparência do Estado de São Paulo[47], onde ficam todos dados abertos sobre infraestrutura, horários e demais informações do Metrô de SP, providas em diversos formatos como csv, texto, pdf e imagens.

[47] http://www.transparencia.sp.gov.br/

O processo do enriquecimento semântico, como já apresentado através das Diretrizes anteriores denotam todos os pontos importantes para realizar o enriquecimento semântico dos dados do Metrô da cidade de São Paulo.

Para que fique mais claro o processo, para cada uma das diretrizes propostas (desde a limpeza até o processo de fusão dos dados) foi gerenciado e descrito cada etapa, com o intuito de mostrar como esse processo genérico ocorre quando escolhido um domínio específico.

Na coleta dos dados, que engloba as **Diretrizes 1, 2** e **3**, foi preciso selecionar as maiores bases de dados abertos referente ao Metrô, de acordo com a ontologia e recurso necessário tornando um processo automático juntamente com o seu manual correspondente.

A ontologia criada nesse caso, usou Metrô, sendo a palavra-chave para as maiores ligações de dados referente ao mesmo, permitindo que o recurso fosse dinâmico imputado pelo usuário que está a fim de consumir as informações geradas.

A maior base de dados encontrada foi o Portal da Transparência do Metrô do São Paulo[48] contendo a maior gama de dados como horários, a rede de funcionamento, pontos e estações, sendo eles dados em qualquer uma da fase de cinco estrelas proposto por Tim Berners-Lee (5 [ESTRELAS], 2012), utilizando-se de arquivos com formatos PDF, XLS ou CSV, e também foi utilizado outros domínios públicos como o *Google images* e *Google maps* para permitir a ligação das imagens e localização desses pontos, incluindo a latitude e longitude com os dados coletados na transparência, uma vez a aplicação da **Diretriz 3** já gerou os dados de forma enriquecida.

No processo de coleta de dados escolheceu-se o formato automática e inteligente os separando por etapas, cuja primeira etapa foi a extração dos metadados contidos na própria *Web* utilizando a biblioteca *Beautifulsoup 4* da linguagem *Python* que é utilizada para analisar documentos HTML e XML retirando e coletando juntamente as URLs dos arquivos PDF, XLS e CSV.

A segunda parte do processo é transformar os dados de um formato para outro, a fim de atingir a quarta estrela citada anteriormente e assim enriquecendo os dados com apresentado na **Diretriz 3**.

[48] https://transparencia.metrosp.com.br/

Para transformar documentos em arquivos PDF para XLS, que é a aplicação das **Diretrizes 1, 2** e **3** que permitem o mapeamento, foi utilizado a ferramenta *PDFTables*, transformado da primeira para a segunda estrela (processo de enriquecimento de dados – **Diretriz 3**) e logo em seguida são convertidos os arquivos XLS para CSV para que nesse processo fosse utilizado a biblioteca *Pandas* responsável por não só converter, (**Diretriz 4**) mas também manipular todos os arquivos CSV, tendo assim todos os dados coletados para o formato CSV a fim de facilitar a limpeza e anotação desses mesmos dados.

Assim, tendo os novos dados já para fácil manejo é feito a limpeza utilizando a biblioteca *OpenRefine-client* e a *Pandas* para separar e remover dados e metadados não utilizados e o próprio poder computacional como aprendizado de máquina, para classificar os dados e gerar uma informação mais sólida.

Tendo os dados limpos foi feito o mapeamento (**Diretriz 4**) dos dados ligando todos os dados sendo informações textuais ou novas URIs e URLs de maneira a ser convertida para o modelo de grafo RDF ou JSON.

Para gerar grafos RDFs foi utilizado a biblioteca *RDFLib* adicionando aos literais de textos principais e os demais nós.

Tomando-os como *URIREF* e literal ou outros demais tipos existentes de dados, no contexto RDF, e a conversão para o formato JSON foi utilizando a biblioteca JSON da própria linguagem sendo mais fácil, pois não foi preciso realizar a tipagem quando se trata de um documento nesta tipagem específica JSON.

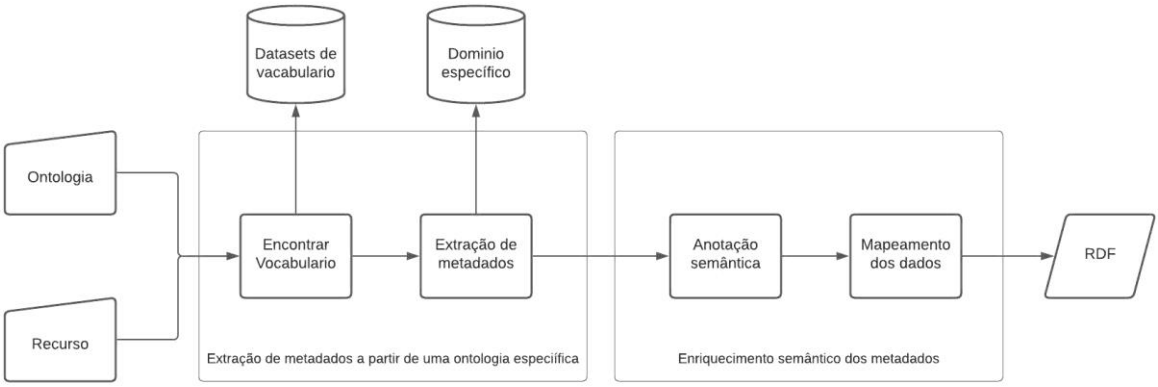
Com o RDF e JSON gerados, foi preciso determinar a persistência das URIs, a fim de facilitar o consumo e localização das informações enriquecidas, e para esse processo foi utilizado a biblioteca *Bottle* capaz de criar um *Web Service* que disponibiliza o consumo de quaisquer aplicações e usuários.

Para realização da aplicação utilizando os dados do Metrô ultilizou-se, em sua maior parte, o tipo de dado em texto, ultilizando a ferramenta *Beautiful Soup 4 python* para lidar com a extração de dados HTML e XML, e outra parte também foi pelos arquivos em formato CSV, passando por um processo mais complexo, pois foi necessário separar os dados de maneira com que poderia ficar mais fácil a localização de um dado específico ultilizando a API *Pandas*

python para manipulá-los e seguir com o processo do enriquecimento semântico.

O processo de enriquecimento semântico, como já apresentado e explicado na **Diretriz 4**, utilizou uma abordagem para ontologias dinâmicas introduzidas pelo usuário a consumir as informações, como apresenta a **Figura 47**.

Figura 47 - Abordagem para enriquecimento semântico



Fonte: SILVA; LUZ, 2020

De acordo com a **Figura 47** averiguamos que o dado de entrada foi combinado com a ontologia desejada e com o recurso a ser acessado, realizou-se assim a classificação da ontologia para a extração dos metadados dos dados abertos, buscando no domínio dos dados do Metrô de São Paulo.

A classificação consistiu em juntar a ontologia com o recurso para poder obter um significado melhor, já com os dados selecionados passaram para fase de anotação semântica onde foi atribuído o significado aos dados e metadados encontrados, facilitando a realização do mapeamento e, por fim, realizou-se o processo de enriquecimento, onde já tendo os dados selecionados, semanticamente anotados e mapeados, foi possível gerar triplas RDF, tornando o dado mais rico e pronto para ser implementado para o aplicativo proposto.

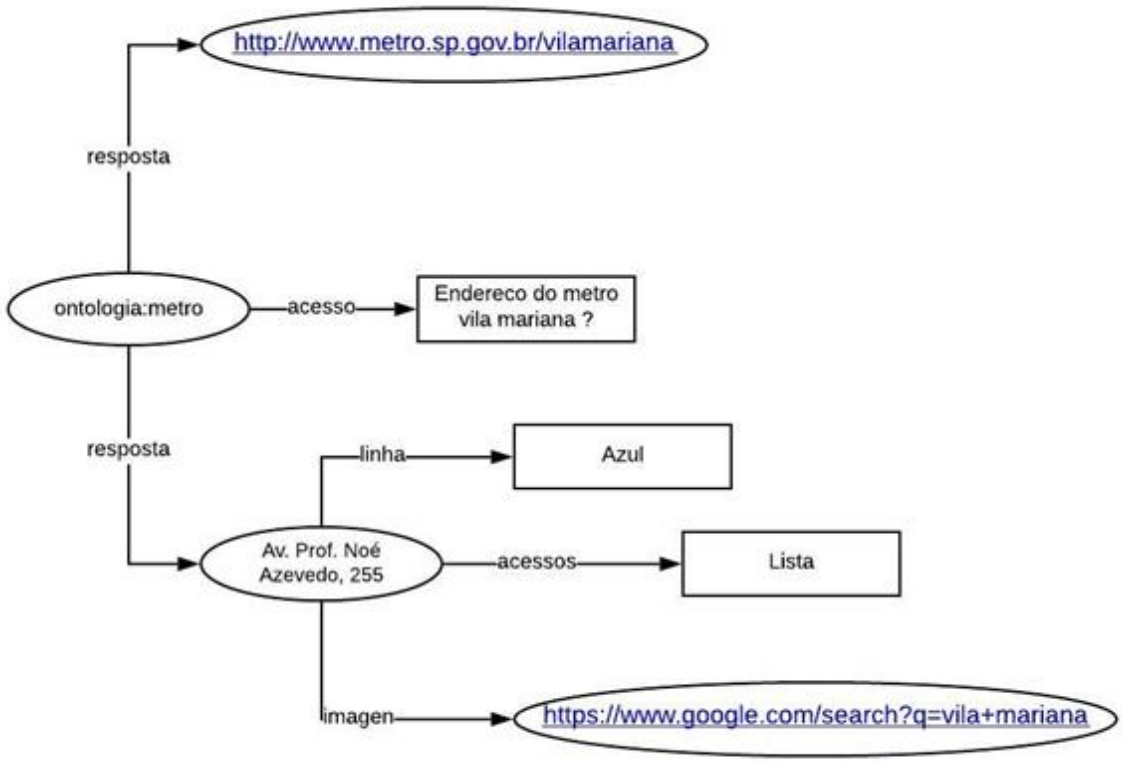
Como resultado, na **Figura 48** consta o dado enriquecido e mapeado com base do recurso escolhido e na ontologia selecionada, podendo visualizar um exemplo de um grafo RDF gerado a partir da ontologia Metrô e do recurso

“Endereço do Metrô Vila Mariana”, e conduzindo a duas respostas de possíveis buscas:

1- Qual é o endereço em si, já ligado com a linha que o metrô pertence e aos seus acessos que gerando assim uma lista e uma imagem;

2 - Que seria um *link* para página onde tem informações do metrô da Vila Mariana.

Figura 48 - Exemplo do grafo RDF enriquecido e mapeado



Fonte: SILVA; LUZ, 2020

Silva e Luz (2020) desenvolveram uma API (*Application Programing Interface*) para prover os dados de maneira sutil, para deixar o projeto mais dinâmico e aberto.

O uso das possibilidades da URI para API foi elaborado com um padrão a fim de facilitar a busca e os resultados providos pelo mesmo, seguindo o padrão /wiki/<ontology>/<type> onde <ontology> seria qual vocabulário se deseja acessar e <type> o formato do arquivo que se quer gerar: que podem ser em XML ou JSON, e como resultado esperado da URI, por exemplo, como “/wiki/metro/json” ou “/wiki/metro/xml”.

Após entrar com a URI desejada, o aplicativo gera uma tripla RDF contendo o máximo de informações procuradas nos domínios específicos. Um exemplo usando a ontologia Metrô e recurso Vila Mariana, e utilizando um cliente HTTP para executar a requisição, teve-se como resultado um arquivo JSON a fim de prover os dados ligados e enriquecidos.

Como resultado, criamos um modelo de aplicação com grande potencial de se tornar um ***Framework*** para ajudar e contribuir com o avanço e crescimento da *Web* Semântica, e apresentar um fluxo onde se difere de um buscador normal ou um simples processo de busca de informação, que apenas retorna possibilidades de resultados, o que se torna difícil de utilizar ou de mesmo reutilizar algum dado e informação disponível como artigos, documentos, pesquisas e até do próprio Metrô.

A mostra um pouco de como o processo final ocorre quando solicitada uma informação para API.

O *log* da aplicação consiste em todo o processo do enriquecimento semântico de maneira simplificada mostrando apenas algumas informações do processo e o seu resultado. O fluxo do *log* se inicia quando o cliente solicita uma requisição na API utilizando a URL específica, como no exemplo que está “http://dominio/wiki/metro/json” trazendo consigo a data, horário e protocolo.

Após esse processo, é separado e mostrado o recurso que foi solicitado, em seguida é mostrado o vocabulário que será utilizado para que possa se assimilar com o recurso e assim ter um meio de busca desses dados que também é representado no *log*.

Assim o algoritmo faz uma busca de dados abertos na *Web* trazendo uma gama de informação para ser anotada, no log; esse processo é repetido para cada domínio em que foram encontradas as informações, finaliza com o mapeamento e enriquecimento, obtendo como resultado uma tripla RDF representada no *log* com sua propriedade e URI ou, se caso for, um valor literal também será exibido. A **Figura 49** apresenta um log da aplicação quando é requisitada a informação.

Figura 49 - Log da aplicação quando requisitado uma informação

Fonte: SILVA; LUZ, 2020

O *log* da aplicação apresentada na **Figura 49** do trabalho de Silva e Luz (2020), é dividido de uma maneira fácil de se entender.

Como ocorre o fluxo, mostrado o recurso solicitado que, no exemplo é “vila mariana” e o vocabulário onde a ontologia cujo é “metro”, a junção das entradas, que é representada pela busca, ficou “vila mariana metro”.

O próximo passo está relacionado aos processos de anotação semântica e mapeamento semântico representando os dados que são ligados como “maps”, “imagens”, “metro” e, por fim, a tripla RDF gerada já pronta para ser consultada ou utilizada.

7. CONSIDERAÇÕES FINAIS

Considerando os conceitos e as análises dos resultados apresentados neste trabalho, concluímos que a *Web* Semântica permite a expansão e a interpretação de um determinado conteúdo que pode ser executado por uma determinada máquina, aplicações gerais e humanos, de modo a proporcionar também, como resultado dos capítulos anteriores, ampliação do processo de assimilação da estrutura do documento produzindo o aperfeiçoamento e a recuperação da Informação.

Este trabalho propôs um modelo de ***Framework*** que apresentasse diretrizes para Publicação de Dadas Semânticos na *Web* com o intuito de mostrar os passos necessários, que inclui desde a extração e limpeza dos dados, para gerar o processo da ênfase ao Enriquecimento e Mapeamento em relação à perspectiva da Ciência da Informação.

Dessa forma, este ***Framework*** apresenta ao usuário quais as diretrizes importantes e necessárias nesse processo, de maneira que ele possa saber a construção de repositórios enriquecidos semanticamente e ainda permitir que o acesso a esses dados semânticos possam ser recuperados ou usados através de aplicativos facilitando o uso informação, além também das próprias máquinas conseguirem essa informação, baseado no processo de recuperação da Informação, pois ela considera a semântica dos termos e da busca do usuário.

Este trabalho teve como base teórica o estudo da Publicação de Dados na *Web*, o uso das Boas Práticas da Publicação, além de todo o contexto da *Web* Semântica.

O enriquecimento e o mapeamento dos dados neste contexto da semântica é uma forma pela qual se pode entender os passos de como esse contexto é criado semanticamente e podendo ainda estabelecer diretrizes compreensíveis.

Sendo assim, a revisão teórica desenvolvida demonstra quatro conceitos-chave, publicação de dados na *Web*, boas práticas de publicação de dados, enriquecimento e mapeamento de dados semânticos são apresentados. Com isso, quer destacar a sua evolução nas últimas décadas, mas, em

especial, como se encontram no presente momento e a possibilidade de um relacionamento efetivo entre essas quatro áreas.

Essa primeira etapa demonstrou uma série de levantamento de estudos correlacionados entre esses campos, em especial no que tange à necessidade do enriquecimento de dados semânticos e como são mapeados, além da possibilidade de disponibilização ao usuário máquina ou usuário humano, através de aplicações e da recuperação da informação através desses repositórios.

A partir desse cenário, discutiu-se a forma como são enriquecidos e mapeados os dados até que se tornem semânticos, visualizando modelos proprietários sem que haja diretrizes informacionais e claras de processos a qualquer usuário. Verificou-se que, independentemente da quantidade de vocabulários ou ontologias já prontas e ainda ferramentas disponíveis, em muitos dos casos, ainda se constroem novos modelos para um uso específico, uma vez que, tendo a ideia das diretrizes, pode ser capaz de criar um processo genérico e assim delimitar o seu escopo de uso, ou melhor, domínio.

O objetivo geral deste trabalho fora apresentar um ***Framework*** capaz de realizar o tratamento dos dados e que oriente, através de diretrizes, o processo de enriquecimento e mapeamento semântico, baseadas nas Boas Práticas de Publicação de Dados na *Web*, utilizando os princípios da representação da informação para que o significado e o contexto dos dados estejam explícitos para o processo de uso. Aponta-se que tal objetivo foi alcançado a partir do modelo proposto e da discussão realizada para a construção e definição desse modelo, além de atender ao conjunto dos objetivos específicos relatados. Além disso, ao apresentar a prova de conceito, demonstra-se como as diretrizes se comportam, destacando que o modelo propicia esse novo processo de enriquecimento e Mapeamento semântico de dados.

O projeto apresentou uma abordagem para realizar enriquecimento e mapeamento semântico de dados que visa utilizar vocabulários e ontologias genéricas tornando-se útil em qualquer área a ser aplicado, simplesmente para ter um dado ligado e rico de informações que possa trazer uma ampla utilidade e contribuição a fim de ajudar em trabalhos futuros, tend como grande diferencial para um buscador comum a sua alta ligação, separação e a

atribuição de significado nos dados devido ao grande aumento de informações providas na *Web*.

Pode-se explicar o processo através das três fases citadas anteriormente, salientando que os dados que alimentam o sistema podem vir de qualquer formato, ou seja, de forma heterogênea.

Neste contexto, o ***Framework*** gera Processo de Extração e Limpeza de Dados que tem a função de extrair e limpar os termos do domínio do qual se utilizará, usando o componente externo. Possui uma abordagem de domínio geral, interligada com repositórios. O resultado do processo dessa etapa é enviado para o Processo de Enriquecimento desses metadados.

Mapeamento é realizado de maneira que possibilita identificar entidades de um domínio de interesse. O componente de mapeamento semântico foi desenvolvido utilizando os dados gerados pelo módulo anterior, Processo de Extração, a fim de gerar um documento em uma estrutura RDF. Essas estruturas permitem que o documento seja entendido tanto pelas máquinas quanto pelas pessoas e, ainda, realização de consultas SPARQL.

A geração do documento RDF, com as triplas de anotações identificadas, pode seguir o padrão de um vocabulário específico.

No componente Fusão de Dados, é realizado o armazenamento do documento RDF de triplas em um banco de dados (*tripleStore*) que fornece um modo padrão para compartilhamento de dados, intercâmbio, que permite consultar e manipular os dados usando a linguagem de consulta SPARQL.

Portanto, com este trabalho, foi possível identificar as diretrizes necessárias para o processo de enriquecimento e mapeamento de dados semânticos apresentando possibilidades de usos de vocabulários, ontologias e ferramentas, que parte dos conceitos e dos elementos da Ciência da Informação, apoiado pelo processo de Boas Práticas de Publicação de Dados permitindo assim um avanço e uma evolução nesses campos de estudos.

Vale destacar que, a partir dessa nova visão do ***Framework*** posposto com as diretrizes, gera uma nova visão de contribuição no modo como a Ciência da Informação se relaciona com o aspecto do enriquecimento e do mapeamento dos dados, em especial tornando-os semânticos e visualizando tanto o processo informacional como computacional.

Aponta-se, ainda, que a relação entre Ciência da Informação e Ciência da Computação guiou o desenvolvimento desta pesquisa, se complementando e utilizando elementos que permitiram a prova de conceito determinada nas diretrizes do ***Framework*** de forma Informacional, expandindo assim quando unido as teorias e práticas da *Web* Semântica, possibilita a interdisciplinaridade do trabalho.

Como fim do processo, elencou os dados semânticos, legíveis para as pessoas e máquinas, abrindo possibilidades para os motores de busca inteligentes e a realização de consultas sofisticadas.

A principal ideia por trás do enriquecimento semântico é fazer o reuso de metadados para facilitar a atividade do publicador em gerar os metadados e publicar estes metadados nos Portais juntamente com os *datasets*. Além disso, minimiza o problema da ausência de metadados ou metadados com pouca descrição semântica, sabendo que é a partir dos metadados que se pode entender os dados.

Apontam-se algumas limitações do presente trabalho, como Nhacuongue e Dutra (2018) apresentam a necessidade de utilização de algumas ontologias e vocabulários que ainda não estão disponíveis e, sendo assim, a necessidade de sua criação. A limitação se aponta devido a um levantamento e processo demorado para que os dados possam ser enriquecidos. Contudo uma das soluções poderia ser o processo automático de se criarem ontologias que determinam o domínio a ser usado ou não, podendo ser uma das etapas da **Diretriz 3** proposta no trabalho.

Na intenção de dar continuidade e desdobramento continuado dos estudos e pesquisas sobre a temática, esta tese visa estudar a qualidade dos dados como apresentado por Melo, Botega e Santarem Segundo (2017) em relação à avaliação de Qualidade para dados conectados, com foco no processo de extração e limpeza apresentado no ***Framework,*** e a contribuição da interdiciplinaridade entre a Ciência da Informação e Ciência da Computação, tornando o modelo do ***Framework*** proposto implementado.

A prova de Conceito utilizando os dados do Metrô de São Paulo já trouxe uma ideia de cenário aplicado, mas seria interessante que o ***Framework*** pudesse expandir para a escolha do domínio do usuário a aplicabilidade de todas as diretrizes propostas, baseados em ontologias e

vocabulários próprios que estejam realmente relacionados com o domínio enriquecendo assim os metadados, propostos pelo usuário, e as ferramentas que sejam competentes a esse processo, mapeando de forma integra e gerando repositórios ligados.

REFERÊNCIAS

5 [ESTRELAS] dos dados abertos. 2012. Disponível em: https://5stardata.info/pt-BR/. Acesso em 25 mar. 2020.

ABICHT, K. **CubeViz: statistical data discovery and visualization**. 2020. Disponível em: http://cubeviz.aksw.org/. Acesso em: 25 mar. 2020.

ABITEBOUL, S.; BUNEMAN, P.; SUCIU, D. **Data on the web:** from relations to semistructure data and XML. San Francisco, CA: Morgan Kaufmamm, 2000.

AGUILLO, I. F**.** Infranet: buscar en los escondrijos de Internet. **Net Conexion**, v. 15, p. 55-56, 1997.

AKSW. **CSV import:** representing multi-dimensional statistical data as RDF using the RDF data cube vocabulary. 2020. Disponível em: http://aksw.org/Projects/CSVImport.html. Acesso em: 25 mar. 2020.

ALEKSOVSKI, Z. et al. Matching unstructured vocabularies using a background ontology. In: STAAB, S.; SVÁTEK, V. (eds). Managing knowledge in a world of networks. EKAW 2006. **Lecture notes in computer science**, v. 4248. Berlin: Springer, Berlin, 2006. Disponível em: https://doi.org/10.1007/11891451_18. Acesso em: 25 mar. 2020.

ALRABAEE, S. et al. SIGMA, a Semantic Integrated Graph Matching Approach for identifying reused functions in binary code. *In*: **DFRWS EU 2015**. 2015. Disponível em: https://dfrws.org/presentation/sigma-a-semantic-integrated-graph-matching-approach-for-identifying-reused-functions-in-binary-code/ Acesso em: 25 mar. 2020.

ÁLVAREZ ESPINAR, M. **Government data openness and re-use**. 2014. Disponível em: http://governobert.gencat.cat/web/.content/01_Que_es/04_Publicacions/col·leccio_govern_obert/GovernObert_2/governobert_2_enG.pdf. Acesso em: 25 mar. 2020.

ANTONIOU, G.; HARMELEN, F. V. **A semantic web primer**. Massachusetts Institute of Technology, 2. ed. 2008. Disponível em: http://people.mpi-inf.mpg.de/~dstepano/KRSW/literature/SWPrimer.pdf. Acesso em: 25 mar. 2020.

APACHE Jena Fuseki. 2021. Disponível em: https://jena.apache.org/documentation/fuseki2/. Acesso em: 25 mar. 2020.

APACHE STANBOL. 2010. Disponível em: https://stanbol.apache.org/. Acesso em: 5 jan. 2020.

ARNDT, N. **Agile knowledge engineering and semantic web**. 2017. Disponível em: http://aksw.org/About.html. Acesso em: 25 mar. 2020.

AUER, S. Introduction to LOD2. *In:* AUER, S.; BRYL, V.; TRAMP, S. (eds.). **Linked Open Data**: creating knowledge out of interlinked data: results of the LOD2 project. Cham: Springer, 2014. p. 1-17, 2014. (Lecture notes in computer science, v. 8661). Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-09846-3_1. Acesso em: 25 mar. 2020.

AUER, S.; DEBATTISTA, J.; LANGE, C. Luzzu: a framework for linked data quality assessment. **IEEE Tenth International Conference on Semantic Computing** (ICSC). 2016. Disponível em: https://ieeexplore.ieee.org/abstract/document/7439317. Acesso em: 9 jan. 2020.

AUER, S. et al. Triplify: light-weight linked data publication from relational databases. **International World Wide Web Conference Committee** (IW3C2), Apr. 2009. Disponível em: https://jens-lehmann.org/files/2009/triplify.pdf. Acessp em: 26 jun. 2019.

BAO, J. et al. **rdf:PlainLiteral:** a datatype for RDF plain literals**.** 2. ed. W3C, 2012. Disponível em: https://www.w3.org/TR/owl2-quick-reference/. Acesso em: 15 jan. 2020.

BARDIN, L. **Análise de conteúdo.** Lisboa: Edições 70, 1979.

BARRASA RODRÍGUEZ, J.; CORCHO, Ó.; GÓMEZ-PÉREZ, A. R_2O, an extensible and semantically based database-to-ontology mapping language. 2004.

BARRETO, A. de A. Perspectivas da ciência da informação. **Revista de Biblioteconomia de Brasília**, v. 21, n. 2, 156-166 p. 1997. Disponível em: https://brapci.inf.br/_repositorio/2010/03/pdf_43caaf49d9_0008818.pdf. Acesso em: 15 jan. 2021.

BECHHOFER, S. et al. **OWL Web Ontology Language reference**. 2004. Disponível em: https://www.researchgate.net/publication/200043063_OWL_Web_Ontology_Language_Reference. Acesso em: 12 nov. 2020.

BECKETT, D. New syntaxes for RDF. **WWW2004**, May 2004, p. 17-22. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.474&rep=rep1&type=pdf. Acesso em: 10 dez. 2020.

BECKETT, D. et al. **RDF 1.1 Turtle**: terse RDF triple language. 2014. Disponível em: https://www.w3.org/TR/turtle/. Acesso em: 15 jan. 2020.

BEM, R. M. de; COELHO, C. C. de S. R. Aplicações da gestão do conhecimento na área de biblioteconomia e ciência da informação: uma revisão sistemática. **Brazilian Journal of Information Science**, v. 7, n. 1, 69-97 p. 2013. Disponível em:

http://www2.marilia.unesp.br/revistas/index.php/bjis/article/view/2987/2395. Acesso em: 23 jan. 2020.

BERNERS-LEE, T. **Linked data: design issues**. W3C, 2006a. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html. Acesso em: jan. 2020.

BERNERS-LEE, T. **Linked data principles**. W3C. 2006b. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: jan. 2020.

BERNERS-LEE, T.; CONNOLLY, D.; SWICK, R. R. **Web architecture:** describing and exchanging data. 1999. Disponível em: https://www.w3.org/1999/04/WebData. Acesso em: 15 jan. 2020.

BERNERS-LEE, T.; FIELDING, R.; FRYSTYK, H. **Hypertext transfer protocol, HTTP/1.0**. May 1996. Disponível em: https://www.hjp.at/doc/rfc/rfc1945.html. Acesso em: 25 abr. 2020.

BERNERS-LEE, T.; FISCHETTI, M. **Weaving the Web:** the original design and ultimate destiny of the world wide web by its inventor. San Francisco, CA: Harper, 1999. Disponível em: http://www.abebooks.com/servlet/SearchResults?an=Berners-Lee&tn=Weaving+Web&paratrk=16303. Acesso em: 15 jan. 2020. ISBN 0062515861.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. **The semantic web:** a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. v. 284, n. 5, p. 28-37, 2001. Disponível em: https://www.researchgate.net/publication/225070375_The_Semantic_Web_A_New_Form_of_Web_Content_That_is_Meaningful_to_Computers_Will_Unleash_a_Revolution_of_New_Possibilities. Acesso em: 15 jan. 2020.

BERTOCCHI, D. Ciberjornalismo e Web Semântica: Considerações sobre o uso de tags em narrativas jornalísticas digitais. *In*: Encontro Nacional de Pesquisadores em Jornalismo (7.: São Paulo, 2009). **Anais do**.... São Paulo, 2009.

BIZER, C.; CYGANIAK, R. **RDF 1.1 TriG**. 2014. Disponível em: https://www.w3.org/TR/trig/. Acesso em: 11 jan. 2021.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: the story so far. International **Journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1-22, 2011. DOI: 10.4018/978-1-60960-593-3.ch008. Disponível em: http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf. Acesso em: 17 out. 2018.

BIZER, C.; SEABORNE, A. **D2RQ**: treating non-RDF databases as virtual RDF graphs. 2005. Disponível em: https://www.researchgate.net/publication/239490591_D2RQ-Treating_non-RDF_databases_as_virtual_RDF_graphs. Acesso em: 14 set. 2019.

BIZER, C. et al. **Linked data on the web** (LDOW2008). WWW '08: Proceedings of the 17th international conference on World Wide Web, April

2008, p. 1265–1266. Disponível em: https://doi.org/10.1145/1367497.1367760. Acesso em: 20 set. 2019.

BORKO, H. Information science: what is it? **American Documentation**, v. 19, n. 1, p. 3-5, 1968.

BOWERS, S.; MADIN, J. S.; SCHILDHAUER, M. P. Owlifier: creating OWL-DL ontologies from simple spreadsheet-based knowledge descriptions. **Ecologial Informatics**, v. 5, n. 1, p. 19-25, Jan. 2010. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1574954109000727. Acesso em: 10 jan. 2019.

BREITMAN, K. **Web semântica:** a Internet do futuro. São Paulo: LTC, 2005. ISBN 8521614667

BRICKLEY, D.; GUHA R. V. **RDF Schema 1.1**. 2014. Disponível em: https://www.w3.org/TR/rdf-schema/. Acesso em: 15 jan. 2020.

BRICKLEY, D.; GUHA, R. V. **RDF vocabulary description language 1.0:** RDF schema. W3C, 2004. Disponível em: http://www.w3.org/TR/rdf-schema/. Acesso em: 16 jan. 2020.

BROEKSTRA, J.; HARMELEN, F.; KAMPMAN, A. Sesame: a generic architecture for storing and querying RDF and RDF schema. **The Semantic WebISWC,** v. 2342, p. 54-68, Springer, 2002. Disponível em: https://link.springer.com/chapter/10.1007/3-540-48005-6_7. Acesso em: 15 jan. 2020.

BRUIJN, J. de et al. OWL DL vs. OWL flight: conceptual modeling and reasoning for the semantic web. **Proceedings of the** 14th International Conference on World Wide Web, p. 623-632, May 2005. Disponível em: https://dl.acm.org/doi/10.1145/1060745.1060836. Acesso em: 28 set. 2019.

BRUIJN, J. de; WELTY, C. **RIF RDF and OWL compatibility**. W3C, 2013. Disponível em: https://www.w3.org/TR/rif-rdf-owl/. Acesso em: 15 jan. 2020.

BRUIJN, J. de; LAUSEN, H. (eds.). **Web service modeling language (WSML)**. W3C, 2005. Disponível em: https://www.w3.org/Submission/WSML/. Acesso em: 15 jan. 2020.

BUCHMANN, R.A; KARAGIANNIS, D. Enriching linked data with semantics from domain-specific diagrammatic models. **Bus Inf Syst Eng** v. 58, p. 341-353, 2016. Disponível em: https://doi.org/10.1007/s12599-016-0445-1. Acesso em: 15 jan. 2020.

BURANARACH, M. OAM: an Ontology Application Management framework for simplifying ontology-based semantic web application development. **International Journal of Software Engineering and Knowledge Engineering**, v. 26, n. 1, p. 115-145, 2016. Disponível em: https://www.worldscientific.com/doi/abs/10.1142/S0218194016500066. Acesso em: 15 jan. 2020.

CALDÓN, E. F. et al. Mecanismos de anotación semántica de contenidos en plataformas de redes sociales. **Cadernos de Informática,** v. 5, n. 1, 2010. Disponível em: https://seer.ufrgs.br/cadernosdeinformatica/article/view/v5n1p89-99/10077. Acesso em: 15 jan. 2020.

CALEGARI, N. **Proposta de uma ferramenta de anotação semântica para publicação de dados estruturados na Web**. Dissertação (mestrado). Pontifícia Universidade Católica. São Paulo, 2016. Disponível em: https://newtoncalegari.com.br/CALEGARI-2016-mestrado.pdf. Acesso: 15 jan. 2020.

CAMPOS, M. L. de A. O papel das definições na pesquisa em ontologia. **Perspectivas em Ciência da Informação**, Belo Horizonte, v;15, n. 1, 2010. Disponível em: https://www.scielo.br/scielo.php?pid=S1413-99362010000100013&script=sci_arttext. Acesso em: 27 nov. 2019.

CAREY, M. J.; ONOSE, N.; PETROPOULOS, M. Data services. **Communications of the ACM**, v. 55, n. 6, p. 86-97, 2012. Disponível em: https://dl.acm.org/doi/pdf/10.1145/2184319.2184340?download=true. Acesso em: 15 jan. 2020.

CAROTHERS, G. (ed). **RDF 1.1 N-Quads**. 2014. Disponível em: https://www.w3.org/TR/n-quads/. Acesso em: 10 jan. 2021.

CASTELLS, P. Aplicación de técnicas de la web semántica. *In*: **Workshop de Investigaciónen Entornos de Interacción Coletiva** (COLINE'02). Granada, Espanha, 2002. Disponível em: https://moodle2.unid.edu.mx/dts_cursos_mdl/pos/TI/BE/AM/12/tecnicas_de_la_web_semantica.pdf. Acesso em: 15 jan. 2020.

CASTILLO, R. Selecting materialized views for RDF data. *In*: International Conference on Web Engineering. **Current Trends in Web Engineering**, 2010. DOI: 10.1007/978-3-642-16985-4_12.

CERN. **Accelerating science**. Geneva, Switzerland, 2020. Disponível em: https://home.cern/. Acesso em: 15 jan. 2020.

CESCA, R. **Modelo de repositório institucional com suporte a web semântica e rede social**. Dissertação (Mestrado)- Universidade Federal de Santa Catarina, Florianópolis. 2018. Disponível em: http://btd.egc.ufsc.br/wp-content/uploads/2018/05/Renato-Cesca.pdf. Acesso em: 15 jan. 2020.

CHAUDRI, V. et al. OKBC: a programmatic foundation for knowledge base interoperability. **Proceedings of AAAI-98**, Madison, WI, EUA, 1998.

CHAUDRI, V.; KARP, P.; THOMERE, J. **XOL:** an XML-based ontology exchange language. 1999. Disponível em: www.ai.sri.com/~pkarp/xol. Acesso em: 15 jan. 2020.

CHEN, C. **Individual differences in a spatial-semantic virtual environment**. 2000. Disponível em:

https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/%28SICI%291097-4571%282000%2951%3A6%3C529%3A%3AAID-ASI5%3E3.0.CO%3B2-F. Acesso em: 15 jan. 2020.

CHEN, H.; WU, Z. **DartGrid III**: a semantic grid toolkit for data integration. International Conference on Semantics, Knowledge and Grid (SKG 2005), Beijing, 2005. Disponível em: 10.1109/SKG.2005.60. Acesso em: 25 set. 2019.

CHIZZOTTI, A. **Pesquisa qualitativa em ciências humanas e sociais**. Petrópolis: Vozes, 2006.

CHLAPEK, D. et al. Otevřená a propojitelná data: metodiky, postupy, nástroje a praxe [= Linked Open Data: methodologies, approaches, tools and the current practices]. In: ŠALOUN, P.; CHLAPEK, D. (eds.). **DATAKON 2014**. Ostrava: Vysoká škola báňskáTechnická univerzita Ostrava, 2014, p. 17–37.

CKAN. **The world's leading open-source data portal platform**. 2020. Disponível em: https://ckan.org/. Acesso em: 15 jan. 2020.

CLARKE, C. A resource list management tool for undergraduate students based on linked open data principles. **ESWC 2009**: The semantic web: research and applications, p. 697-707, 2009. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-02121-3_51. Acesso em: 23 abr. 2019.

COMVANTAGE. Public deliverables. 2015. Disponível em: http://www.comvantage.eu/results-publications/public-deriverables/. Acesso em: 22 set. 2019.

CONEGLIAN, C. S.; LUZ, L. P.; SANTAREM SEGUNDO, J. E. Boas práticas para publicação de dados na web: aplicação nos dados referentes aos resultados de pesquisa científica. **Encontro Nacional de Pesquisa em Ciência da Informação** (18., ENANCIB, Marília, 2017), 2017. Disponível em: http://www.brapci.inf.br/index.php/res/v/104935. Acesso em: 15 jan. 2020.

COOD, E. F. A relational model of data for large shared data banks. **Communications of the ACM**, v. 13, n. 6, p. 377-387, June 1970. Disponível em: https://www.seas.upenn.edu/~zives/03f/cis550/codd.pdf. Acesso em: 15 jul. 2019.

COPPIN, B. **Inteligência artificial**. Rio de Janeiro : LTc, 2017.

CORNELLA, A. La infranet: ¿dónde está el valor? **El profesional de la información**, v. 8, n. 5, p.3, 1999.

CORRALES, B. R. et al. A importância das boas práticas para dados na Web e o caso do Cetic.br. **Congreso Internacional en Gobierno. Administración y Políticas Públicas** GIGAPP (9., Madrid, ES, 2018. Disponível em: https://ceweb.br/media/docs/publicacoes/19/A%20importa%CC%82ncia%20das%20Boas%20Pra%CC%81ticas%20para%20Dados%20na%20Web%20e%20o%20caso%20do%20Cetic.br%20%20-%20FINAL.pdf. Acesso em: 15 jan. 2020.

COSTA, L. S. da. **Enriquecimento semântico de informação geográfica voluntária usando linked data e tesauro**. 2018. Dissertação (Mestrado)- Universidade Federal de Viçosa, 2018. Disponível em: https://www.locus.ufv.br/bitstream/123456789/17963/1/texto%20completo.pdf. Acesso em: 20 mar. 2019.

CRESWELL John W. **Projeto de pesquisa**: métodos qualitativo, quantitativo e misto. 3. ed. São Paulo: Bookman, 2010.

CULLOT, N.; GHAWI, R.; YETONGNON, K. DB2OWL: a tool for automatic database-to-ontology mapping. **Proceedings of the** Fifteenth Italian Symposium on Advanced Database Systems, SEBD, 2007. Disponível em: https://www.researchgate.net/publication/220973908_DB2OWL_A_tool_for_automatic_database-to-ontology_mapping. Acesso em: 16 set. 2019.

CUNHA, D. R. B.; LÓSCIO, B. F.; SOUZA, D. **Linked data**: da web de documentos para a web de dados. 2011. Disponível em: https://www.cin.ufpe.br/~daise/arquivos/publications/2011/2011_Cap%204%20Linked%20Data%20-%20MC2.pdf. Acesso em: 25 ago. 2020.

CYGANIAK, R. et al. **The D2RQ mapping language**. 2012. Disponível em: http://d2rq.org/d2rq-language. Acesso em: 22 abr. 2019.

CYGANIAK, R.; JENTZSCH, A. **The linked open data cloud**. 2020. Disponível em: http://lod-cloud.net/. Acesso em: 15 jan. 2020.

D'AQUIN, M.; ADAMOU, A.; DIETZE, S. Assessing the educational linked data landscape. **WebSci'13**: proceedings of the 5th annual ACM Web Scoence Conference, p. 43-46, May 2013. Disponível em: https://dl.acm.org/doi/abs/10.1145/2464464.2464487. Acesso em: 25 jun. 2020.

DACONTA, M. C., OBRST, L. J; SMITH, K. T. **The semantic web:** a guide to the future of XML, web services, and knowledge management. Wiley, 2003. ISBN-13: 978-0471432579.

DATE, C. J. **Introdução a sistemas de banco de dados**. 7. ed. Rio de Janeiro: Campus, 2000.

DAWES, S. S.; VIDIASOVA L.; PARKHIMOVICH, O. Planning and designing open government data programs: an ecosystem approach. **Government Information Quarterly**, v. 33, n. 1, p. 15-27, 2016. Disponível em: https://www.sciencedirect.com/science/article/pii/S0740624X1630003X. Acesso em: 15 jan. 2020.

DAY, M. **Cedars guide to preservation metadata**. 2004. Disponível em: https://www.researchgate.net/publication/277177828_Cedars_Guide_to_Preservation_Metadata. Acesso em: 15 jan. 2020.

DBPEDIA Spotlight. **Shedding light on the web of documents**. 2020. Disponível em: https://www.dbpedia-spotlight.org/. Acesso em: 15 jan. 2020.

DE DIANA, M; GEROSA, M. A. NoSQL na Web 2.0: um estudo comparativo de bancos não-relacionais para armazenamento de dados na Web 2.0. *In*: WORKSHOP DE TESES E DISSERTAÇÕES EM BANCO DE DADOS, 9.,2010, Belo Horizonte. **Anais...** Belo Horizonte, 2010. Disponível em: https://www.ime.usp.br/~mdediana/nosql_wtdbd10.pdf. Acesso em: 10 jun. 2019.

DEAN, J.; GHEMAWAT, S. MapReduce: simplified data processing on large clusters. **Symposium on Operating Systems Design and Implementation** (6., 2004). 2004. Disponível em: https://static.usenix.org/publications/library/proceedings/osdi04/tech/full_papers/dean/dean.pdf. Acesso em: 11 jan. 2021.

DECKER, S. et al. The semantic web: the roles of XML and RDF. **IEEE Internet Computing**, v. 15, n. 3, 2000. Disponível em: https://ieeexplore.ieee.org/abstract/document/877487. Acesso em: 15 jan. 2020.

DELBRU, R. **SIREn:** entity retrieval system for the web of data. 2009. Disponível em: http://www.renaud.delbru.fr/doc/pub/fdia2009-siren.pdf. Acesso em: 15 jan. 2020.

DENZIN, N. K. **The research act:** a theoretical introduction to sociological methods. 2. ed. New York: Mc Graw-Hill, 1978.

DI IORIO, A.; SCHAERF, M. **Expressing the tacit knowledge of a digital library system as linked data**. DIAG, Department of Computer, Control, and Management Engineering Antonio Ruberti Sapienza University of Rome. Roma, IT, 2019. Disponível em: https://www.mdpi.com/2073-431X/8/2/49/xml. Acesso em: 15 jan. 2020.

DICIO: dicionário online de português. 2020. Disponível em: https://www.dicio.com.br/aurelio/. Acesso em: 15 jan. 2020.

DIMOU, A. et al. Modeling, generating, and publishing knowledge as linked data. **Knowledge Engineering and Knowledge Management**. Bologna, IT: Springer, 2016. p. 3-14.

DIZARD JR., W. **A nova mídia:** a comunicação de massa na era da informação. Rio de Janeiro: Jorge Zahar, 2000.

DL-LEARNER. **Is a framework for supervised machine learning in OWL, RDF and description logics**. 2015. Disponível em: https://dl-learner.org/. Acesso em: 15 jan. 2020.

DONG, X. L.; HALEVY, A.; YU, C. Data integration with uncertainty. **The VLDB Journal**, v. 18, n. 2, 2009. p. 469-500. Disponível em: https://link.springer.com/article/10.1007%2Fs00778-008-0119-9. Acesso em: 14 nov. 2020.

DOSE. 2020. Disponível em: https://sourceforge.net/projects/dose/. Acesso em: 15 jan. 2020.

ELASTIC. 2020. Disponível em: https://www.elastic.co/pt/. Acesso em: 15 jan. 2020.

ELLER, M. P. **Anotação semânticas de fontes de dados heterogêneas:** um estudo de caso com a ferramenta Smore. Dissertação (Mestrado em Computação) - Universidade Federal de Santa Catarina, Departamento de Informática e Estatística, 2014.

ELMASRI, R.; NAVATHE, S. B. **Fundamentals of database systems**. 4. ed. [S.l.]: Addison-Wesley, 2003.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 4. ed. [S.l.]: Pearson: Addison Wesley, 2005.

ERLING, O.; MIKHAILOV, I. **Virtuoso**: RDF support in a native RDBMS. Semantic Web Information Management, p. 501-519, 2010. Disponível em: http://dx.doi.org/10.1007/978-3-642-04329-1_21. Acesso em: 9 set. 2019.

EUROPEAN COMMISSION. **Digital government factsheet 2019**: Norway. 2019. Disponível em: https://joinup.ec.europa.eu/sites/default/files/inline-files/Digital_Government_Factsheets_Norway_2019.pdf. Acesso em: 27 dez. 2019.

EUZENAT, J.; SHVAIKO, P. **Ontology matching**. [S. l.] : Springer, 2013. DOI 10.1007/978-3-642-38721-0.

EVANS, D. **Internet de las cosas:** cómo la próxima evolución de Internet lo cambia todo. Cisco Internet Business Solutions Group (IBSG), 2011. Disponível em: https://www.cisco.com/c/dam/global/es_mx/solutions/executive/assets/pdf/internet-of-things-iot-ibsg.pdf. Acesso em: 15 jan. 2020.

FELICÍSSIMO, C. H. 2004 **Interoperabilidade semântica na Web:** uma estratégia para o alinhamento taxonômico de ontologias. Dissertação (mestrado)- PUCRio, 2004. Disponível: http://www-di.inf.puc-rio.br/~julio/CAROL.pdf. Acesso em: 15 jan.2021.

FENSEL, D. **Ontologies:** a silver bullet for knowledge management and electronic commerce. Berlim: Springer, 2001.

FENSEL, D.; BUSSLER, C. The web service modeling framework WSMF. **Electronic Commerce Research and Applications**, v. 1, n. 2, p. 113–137, 2002. DOI:10.1016/s1567-4223(02)00015-7.

FERNÁNDEZ, M. et al. Semantically enhanced information retrieval: an ontology-based approach. **Journal of Web Semantics,** v. 9, n. 4, p. 434-452, 2011. Disponível em: https://doi.org/10.1016/j.websem.2010.11.003. Acesso em: 15 jan. 2020.

FERNÁNDEZ, R. C. **Representación del conocimiento**. Web Semántica. Madrid: Universidad Carlos III, 2008. Disponível em: http://www.it.uc3m.es/jvillena/irc/practicas/08-09/05.pdf. Acesso em: 15 jan. 2020.

FERNÁNDEZ-BERROCAL, P. et al. Gender differences in emotional intelligence: the mediating effect of age. **Behavioral Psychology**, v. 20, n. 1, p. 77-89, 2012. Disponível em: http://jornadasaludemociongenero.uji.es/wp-content/uploads/2014/11/Fernandez-Berrocal.pdf. Acesso em> 29 jan. 2020.

FERREIRA, A. B. de H. **Dicionário Aurélio da língua portuguesa**. 5. ed. Curitiba: Positivo, 2010.

FININ, T. W. et al. (eds.). **Information and knowledge management: expanding the definition of "database"**. Lecture notes in computer science. Baltimore: CIKM, 1992. Disponível em: https://www.springer.com/gp/book/9783540574194. Acesso em: 10 jan. 2020.

FLEMISH government: open data handleiding [open data handbook] 2014.Disponível em: http://www.opendataforum.info/images/Open_Data_Handboek_2014. Acesso em: 14 jan. 2020.

FOOTE, B; JOHNSON, R. E. Designing reusable classes. **Journal of Object-Oriented Programming**, v. 1, p. 22-35, 1988.

FURTADO, J. et al. Ferramenta para extração de dados semi-estruturados para carga de um big data. **Revista Brasileira de Computação Aplicada**, v. 7, n. 3, p. 43-52, 2015. Disponível em: https://doi.org/10.5335/rbca.2015.3842. Acesso em: Jan. 2020.

GAINES, B.R.; SHAW, M. Concept maps as hypermedia components. **International Journal of Human Computer Studies**., v. 43, n. 3, p. 323-361, 1995. Disponível em: https://www.semanticscholar.org/paper/Concept-maps-as-hypermedia-components-Gaines-Shaw/17ac6d10a1af5b9e17892e1fca89a71e8fe08826. Acesso em: 15 jan. 2020.

GALINDO, M. et al. A rede memorial e sua missão informacional: sistemas memoriais e redes de colaboratividade. *In:* **ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO**, Belo Horizonte. ECI/UFMG, 2014, p. 4759-4775. Disponível em: http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt10. Acesso em: 15 jan. 2020.

GAMMA, E. et al. **Padrões de projeto:** soluções reutilizáveis de software orientado a objetos. Porto Alegre: Bookman, 1995.

GANDON, F.; SCHREIBER, G. **RDF 1.1 XML Syntax**. 2014. Disponível em: https://www.w3.org/TR/rdf-syntax-grammar/. Acesso em: 15 jan. 2020.

GARCÍA-GARCÍA, A. **Integración de contenidos semánticos en un portal web de científicos y humanistas valencianos**: vestigium. 2015. Disponível em: https://www.semanticscholar.org/paper/Integraci%C3%B3n-de-contenidos-sem%C3%A1nticos-en-un-portal-y-Garc%C3%ADa-Garc%C3%ADa/787b7233a9af12c4700813eb688f4eea82fdbd95. Acesso em: 15 jan. 2020.

GARCIA-MARCO, F. J. Ontologies and knowledge organization: challenges and opportunities for information professionals. **El Profesional de la Información**, v. 16, p. 541-550, 2007.

GARDENFORS, P. What is semantics? *In*: GARDENFORS, P. **The geometry of meaning**: semantics based on conceptual spaces. Boston, MIT Press, 2014. p. 3-20.

GEORGALA, K. **LIMES:** LInk discovery framework for MEtric Spaces. 2019. Disponível em: http://aksw.org/Projects/LIMES.html. Acesso em: 15 jan. 2020.

GERHARDT, T. E. et al. Estrutura do projeto de pesquisa. *In:* GERHARDT, T. E.; SILVEIRA, D. T. (orgs.). **Métodos de pesquisa.** Porto Alegre: UFRGS, 2009. p. 65-88. Disponível em: http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf. Acesso em: 15 jan. 2020.

GIL, A. C. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010.

GIL, Y. Knowledge mobility: semantics for the Web as a white knight for knowledge-based systems. *In:* FENSEL, D., **Developing the Semantic Web**, 2003, p. 1-36.

GOLBREICH, C.; WALLACE, E. K. **OWL 2 Web Ontology Language new features and rationale**. 2. ed. W3C, 2012. Disponível em: https://www.w3.org/TR/owl2-new-features/. Acesso em: 15 jan. 2020.

GÓMEZ-PÉREZ, A.; ORTIZ-RODRÍGUEZ, F.; VILLAZÓN-TERRAZAS, B. Legal ontologies for the Spanish e-government. **Current Topics in Artifitial Intelligence**, p. 301-310, 2005. Disponível em: https://link.springer.com/chapter/10.1007/11881216_32. Acesso em: 14 maio 2019.

GONZÁLEZ, M.; ANTONIO, J. **Linguagens documentárias e vocabulários semânticos para a web**: elementos conceituais. Salvador: EDUFBA, 2011.

GRIMES, A. J.; LEWIS, M. W. Metatriangulação: a construção de teorias a partir de múltiplos paradigmas. **RAE**, v.45, n. 1, p. 72-91, 2005. Disponível em: http://bibliotecadigital.fgv.br/ojs/index.php/rae/article/viewFile/37104/35874. Acesso em: Jan. 2020.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, n. 2, p. 199-200, 1993. Disponível em: https://doi.org/10.1006/knac.1993.1008. Acesso em: 15 jan. 2020.

GUANDALINI, C. A.; SANTOS, A. A. dos. Linked open data: conceito, relações e importância na era da informação. **Encontro Regional dos Estudantes de Biblioteconomia, Documentação, Gestão e Ciência da Informação**, 5., Belo Horizonte, 2018. Disponível em: https://brapci.inf.br/index.php/res/download/138877. Acesso em: 17 mar. 2020.

GUARINO, N. Formal ontology and information systems. In N. Guarino (Ed.), **Proceedings of the Conference on Formal Ontology in Information Systems**, FOIS‟98, 6–8 June 1998, Trento, Italy, p. 3–15 (Amsterdam: IOS Press). 1998.

GÜELL, F. **Results of the European innovation scoreboard (EIS).** 2017. Disponível em: https://www.fguell.com/en/results-of-the-european-innovation-scoreboard-eis-2017/. Acesso em: 25 abr. 2019.

HALASCHEK-WIENER, Christian et al. **A flexible approach for managing digital images on the semantic web**. Washington, 2005. Disponível em: https://www.researchgate.net/publication/228585349_A_flexible_approach_for_managing_digital_images_on_the_semantic_web/link/0fcfd5089aafd4a458000000/download. Acesso em: 13 fev. 2020.

HAMDI, F.; REYNAUD, C.; SAFAR, B. Pattern-based mapping refinement. **EKAW 2010, Knowledge Engineering and Management by the Masses**, 2010. p. 1-15. (Lecture notes in computer science, v. 6317).

HANDSCHUH, S.; STAAB, S.; CIRAVEGNA, F. S-cream: semi-automatic CREAtion of metadata; **EKAW 2002**: Knowledge engineering and knowledge management: ontologies and the semantic web, p. 358-372, 2002. Disponível em: https://link.springer.com/chapter/10.1007/3-540-45810-7_32. Acesso em: 16 jan. 2020.

HARRIS, S.; SEABORNE, A. (eds.). **SPARQL 1.1 query language**. W3C, 2013. Disponível em: https://www.w3.org/TR/sparql11-query/. Acesso em: 15 jan. 2020.

HAYES, P. (ed). **RDF semantics**. W3C, 2004. Disponível em: https://www.w3.org/TR/rdf-mt/. Acesso em: 31 jan. 2021.

HAWKE, S. **W3C vocabulary services**. 2013. Disponível em: https://www.w3.org/2013/04/vocabs/. Acesso em: 11 ago. 2020.

HEATH, T.; BIZER, C. **Linked data**: evolving the Web into a global data space. Chicago: Morgan & Claypool, 2011. (Synthesis lectures on the semantic web: theory and technology).

HEFLIN, J. **An introduction to the OWL Web Ontology Language**. 2007. Disponível em: http://www.cse.lehigh.edu/~heflin/IntroToOWL.pdf. Acesso em: 15 jan. 2020.

HEFLIN, J.; HENDLER, J. Searching the web with SHOE. **AAAI Technical Report WS-00-01**. 2000. Disponível em:

https://www.aaai.org/Papers/Workshops/2000/WS-00-01/WS00-01-007.pdf. Acesso: 20 jan. 2021.

HENDLER, J. Agents and the semantic web. **IEEE Intelligent Systems**, v. 6, p. 30-37, 2001. Disponível em: https://ieeexplore.ieee.org/abstract/document/920597. Acesso em: 15 jan. 2020. DOI: 10.1109/5254.920597.

HETTNE, K. M. et al. Structuring research methods and data with the research object model: genomics workflows as a case study. **Journal of Biomedical Semantics**, v. 5, n. 41, p. 1-16, 2014. Disponível em: https://link.springer.com/content/pdf/10.1186/2041-1480-5-41.pdf

HOFFMANN, C. **RDFauthor**. 2020. Disponível em: http://aksw.org/Projects/RDFauthor.html. Acesso em: 16 jan. 2020.

HÖFFNER, K. **Facete**. 2016. Disponível em: http://aksw.org/Projects/Facete.html. Acesso em: 15 jan. 2020.

HOLMES, W; KOGUT, P. **AeroDAML:** applying information extraction to generate DAML annotagions from web pages. Philadelphia, 2001. Disponível em: http://semannot2001.aifb.uni-karlsruhe.de/positionpapers/AeroDAML3.pdf. Acesso em: 15 jan. 2020.

HORROOCKS, I. DAML+OIL: a descriptive logic for the semantic Web. **Bulletin of the IEEE Computer Society Technical Committee on Data Engineering**, 2002. Disponível em: https://www.cs.ox.ac.uk/people/ian.horrocks/Publications/download/2002/ieeede2002.pdf. Acesso em: 12 abr. 2020.

HORROCKS, I. et al. **The ontology inference layer OIL**.2000. Disponível em: http://www.cs.ox.ac.uk/ian.horrocks/Publications/download/2000/oil.pdf. Acesso em: 13 jan. 2020.

HYLAND, B.; ATEMEZING, G.; VILLAZÓN-TERRAZAS, B. **Best practices for publishing linked data**. 2014. Disponível em: https://www.w3.org/TR/ld-bp/. Acesso em: 12 jan. 2020.

INTELLISEMANTIC. **Intelligent solution provider**. 2020. Disponível em: https://www.intellisemantic.com/. Acesso em: 11 jan. 2020.

ISELE, R. et al. **Silk:** the linked data integration framework. 2020. Disponível em: http://silkframework.org/. Acesso em: 11 jan. 2020.

ISOTANI, S.; BITTENCOURT, I. **Dados abertos conectado**. Ibert Bittencourt. São Paulo: Novatec, 2015. Disponível em: http://www.pgcl.uenf.br/arquivos/dadosabertosconectados_011120181613.pdf. Acesso em: 11 jan. 2020.

IVANOVIĆ, L. et al. User interface of web application for searching PhD dissertations of the University of Novi Sad. **International Symposium on**

Intelligent Systems and Informatics (SISY), IEEE. Subotica, p. 117-120, 2013.

JAIN, P. et al. Contextual ontology alignment of LOD with an upper ontology: a case study with Proton. *In*: ANTONIOU, G. et al. (eds). The semantic web: research and applications. ESWC 2011. **Lecture Notes in Computer Science**, v. 6643. Berlin: Springer, 2011. Disponível em: https://doi.org/10.1007/978-3-642-21034-1_6. Acesso em: 11 jan. 2020.

JAMES, L. **Defining open data**. 2013. Disponível em: https://blog.okfn.org/2013/10/03/defining-open-data/. Acesso em: 11 jan. 2020.

JOINT, N. Current research information systems, open access repositories and libraries: ANTAEUS. **Library Review**, v. 57, n. 8, p. 570-575, 2008.

JØSANG, A. A logic for uncertain probabilities. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 9, n. 3, p. 279-311, 2001. Disponível em: https://www.worldscientific.com/doi/abs/10.1142/S0218488501000831. Acesso em: 11 jan. 2020.

JUNAR. **Open data that is easy to re-use and interpret by many different audiences**. 2020. Disponível em: https://www.junar.com/. Acesso em: 11 jan. 2020.

KAHAN, J; KOIVUNEN, M. R. Annotea: an open RDF infrastructure for shared Web annotations. *In*: WWW '01: **Proceedings of the** 10th International Conference on World Wide Web, 2001. p. 6230632. Disponível em: https://dl.acm.org/doi/10.1145/371920.372166. Acesso em: 10 jan. 2020.

KALFOGLOU, Y.; SCHORLEMMER, M. IF-Map: an ontology-mapping method based on information-flow theory. **Journal on Data Semantics**, 2003. DOI: 1. 0.1007/978-3-540-39733-5_5.

KALYANPUR, A. et al. Smore: semantic markup, ontology, and RDF editor. **Proceedings of Internacional Semantic Web Conference** (3., ISWC-2004), Japan, 2006. Disponível em: https://apps.dtic.mil/dtic/tr/fulltext/u2/a447989.pdf. Acesso em: 10 jan. 2020.

KARGER, D. The semantic web and end users: whats´s wrong and how to fix it. **Internet Computing**, IEEE, v. 18, n. 6, p. 64-70, nov. 2014.

KELLOGG, G.; CHAMPIN, P.-A.; LONGLEY, D. (eds.). **JSON-LD 1.1**. 2020. Disponível em: https://www.w3.org/TR/json-ld11/. Acesso em: 16 set. 2020.

KENT, R. The information flow foundation for conceptual knowledge organization. *In:* INTERNATIONAL CONFERENCE OF THE INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION, (6., Toronto, Canada), 2000.

KENT, R. **Conceptual knowledge markup language:** the central core. 2011. Disponível em:

https://pdfs.semanticscholar.org/103e/8ea32bfc207f8ce699c4e7a84b8e0b44b674.pdf. Acesso em: 11 jan. 2020.

KENT, R. The IFF foundation for ontological knowledge organization. **Cataloging & Classification Quarterly**, v. 37, n. 1-2, p. 187-203, 2009. DOI: https://doi.org/10.1300/J104v37n01_13. Acesso em: 18 jan. 2020.

KIEFER, C.; BERNSTEIN, A.; STOCKER, M. **The fundamentals of iSPARQL**: a virtual triple approach for similarity-based semantic web tasks. 2007. DOI: 295-309. 10.1007/978-3-540-76298-0_22.

KIRYAKOV, A. et al. Semantic annotation, indexing and retrieval. **Journal of Web Semantics**, v. 2, n. 1, p. 49-79, 2005. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1570826804000162. Acesso em: 8 jan. 2020.

KUCERA, J. Open government data publication methodology. **Journal of Systems Integration**, v. 6, n. 2, p. 52-62, 2015. Disponível em: http://si-journal.org/index.php/JSI/article/viewFile/231/232. Acesso em: 12 jan. 2020.

KUIPERS, B. The spatial semantic hierarchy. **Artificial Intelligence**, v. 119, n. 1-2, p. 191-233, May 2000. Disponível em: https://www.sciencedirect.com/science/article/pii/S0004370200000175. Acesso em: 17 out. 2020.

KUTER, U. et al. Information gathering during planning for Web service composition. **SSRN Electronic Journal**, v. 3, p. 183-2005, Jan. 2005. Disponível em: https://www.researchgate.net/publication/220291124_Information_Gathering_During_Planning_for_Web_Service_Composition. Acesso em: 13 nov. 2020.

KLYNE, G,; CARROLL, J. **Resource Description Framework (RDF)**: concepts and abstract syntax: W3C recommendation. W3C, 2004. Disponível em: https://www.researchgate.net/publication/272829811_Resource_Description_Framework_RDF_Concepts_and_Abstract_Syntax_W3C_Recommendation_10_February_2004. Acesso em: 21 jan. 2021.

KNOLMAYER, G. F; MYRACH, T. Concepts of bitemporal database theory and the evolution of Web documents. *In*: **Proceedings of** the 34th Annual Hawaii International Conference on System Sciences, 2001, p. 1-10. Disponível em: https://ieeexplore.ieee.org/document/927091. Acesso em: 6 jan. 2020.

LABROU, Y.; FININ, T. **A proposal for a new KQML specification**. Mar., 1998. Disponível em: https://www.researchgate.net/publication/2818605_A_proposal_for_a_new_KQML_specification. Acesso em: 10 jan. 2020.

LACLAVIK, M. et al. Tools for email based recommendation in enterprise. **ENTERprise Information Systems**, p. 209-218, 2010. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-16402-6_23. Acesso em: 23 set. 2019.

LAUFER, C. **Guia web semântica**. 2015. Disponível em: https://nic.br/media/docs/publicacoes/13/Guia_Web_Semantica.pdf. Acesso em: 21 jan. 2020.

LEE, D. Building an open data ecosystem: an Irish experience. *In:* ICEGOV '14: **Proceedings of the** 8th International Conference on Theory and Practice of Electronic Governance. 2014, p. 351-360. Disponível em: http://dl.acm.org/citation.cfm?id=2691258. Acesso em: 11 jan. 2020.

LEGAZ-GARCÍA, M. del C. et al. A semantic web-based framework for the interoperability and exploitation of clinical models and EHR data. **Knowledge-Based Systems**, v. 105, p. 175-189, Aug. 2016 Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0950705116301058. Acesso em: 11 jan. 2020.

LEIVA-MEDEROS, A. et al. Working framework of semantic interoperability for CRIS with heterogeneous data sources. **Journal of Documentation**, v. 73, n. 3, p. 481-499, 2017. Disponível em: https://doi.org/10.1108/JD-07-2016-0091. Acesso em: 11 jan. 2020.

LENAT, D. B. et al. **Cyc**: toward programs with common sense. Communications of the ACM, v. 33, n. 8, Aug., 1990.

LESSIG, L. **The future of ideas**: the fate of the commons in a connected world. New York: Random House, 2001.

LIMA, G. A. B. Mapa conceitual como ferramenta para organização do conhecimento em sistema de hipertextos e seus aspectos cognitivos. **Perspectivas em ciência da informação**, Belo Horizonte, v. 9, n. 2, p. 134-145, 2004. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/355. Acesso em: 11 jan. 2020.

LINKED open vocabularies (LOV). 2020. Disponível em: https://lov.linkeddata.es/dataset/lov/. Acesso em: 17 ago. 2020.

LIRA, M. A. B. de. **Uma abordagem para enriquecimento semântico de metadados para publicação de dados abertos**. 2014. 95 f. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal de Pernambuco, Recife, 2014. Disponível em: https://repositorio.ufpe.br/handle/123456789/11570. Acesso em: 25 jan. 2021.

LONGLEY, D. et al. **JSON-LD 1.1 processing algorithms and API**. 2020. Disponível em: https://www.w3.org/TR/json-ld-api/. Acesso em: 31 jan. 2021.

LOPEZ, V.; MOTTA, E; SABOU M. PowerMap: mapping the real semantic web on the fly. *In:* INTERNATIONAL SEMANTIC WEB CONFERENTE. The Semantic Web, ISWC. **Lecture notes in computer science**, v. 4273, p. 414-427, 2006. Disponível em: https://doi.org/10.1007/11926078_30. Acesso em: 11 jan. 2020.

LÓSCIO, B. F. F.; BURLE, C.; CALEGARI, N. (eds). **Data on the Web: best practices**: W3C recommendation. 31 jan. 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 12 dez. 2018

LÓSCIO, B. F. et al. Fundamentos para publicação de dados na Web. São Paulo: Comitê Gestor da Internet no Brasil, 2018. Disponível em: https://ceweb.br/publicacao/livro-fundamentos-dados-web/. Acesso em: 11 jan. 2020.

LUO, J.; NEWTON, J. **Introdução do gRPC no .NET**. 2021. Disponível em: https://docs.microsoft.com/pt-br/aspnet/core/grpc/?view=aspnetcore-5.0. Acesso em: 30 mar. 2021.

LUKE, S.; RAGER, D.; SPECTOR, L. Ontology-Based Knowledge Discovery on the World-Wide Web**.** *In:* **Proceedings of the Workshop on Internet-based Information Systems**, AAAI-96, Portland, Oregon, USA, 1996.

MACGREGOR, R. M. Inside the LOOM description classifier. **ACM Sigart Bulletin**, v. 2, n. 3, June, 1991. Disponível em: https://doi.org/10.1145/122296.122309. Acesso em: 20 jan. 2021.

MADHAVAN, J.; BERNSTEIN, P. A.; RAHM, E. **Generic schema matching with Cupid**. August, 2001. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.2432&rep=rep1&type=pdf. Acesso em: 14 jul. 2019.

MANOLA, F.; MILLER, E. **RDF primer W3C recommendation**. W3C, 2004. Disponível em: http://www.uazuay.edu.ec/bibliotecas/mbaTI/pdf/RDF%20Primer.pdf. Acesso em: 11 jan. 2020.

MANUAL dos dados abertos: governo: traduzido e adaptado de opendatamanual.org. 2011. Disponível em: https://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf. Acessp em: 19 set; 2019.

MANUJA, M.; GARG, D. Semantic web mining of un-structured data: challenges and opportunities. **International Journal of Engineering**, v. 5, n. 3, p. 268-276, 2011. Disponível em: https://www.researchgate.net/profile/Arunesh-Chandra-2/publication/271520706_Analysis_of_Hand_Anthropometric_Dimensions_of_Male_Industrial_Workers_of_Haryana_State/links/54cb31b00cf22f98631e3dc2/Analysis-of-Hand-Anthropometric-Dimensions-of-Male-Industrial-Workers-of-Haryana-State.pdf#page=34. Acesso em: 11 jan. 2021.

MARCH, S. T.; SMITH, G. F. Design and natural science research in information technology. **Decision Support Systems**, v. 15, p. 251-266, 1995.

MCGUINNESS, D. L.; VAN HARMELEN, F. **OWL Web Ontology Language overview**. W3C. 2004. Disponível em: https://www.w3.org/TR/owl-features/. Acesso em: 11 jan. 2020.

MELO, J. O. S.; BOTEGA, L. C.; SANTAREM SEGUNDO, J. E. Metodologia de avaliação de qualidade para dados conectados. V. 4 N. 2 (2017): **Informação & Tecnologia**, especial ENANCIB 2017, v. 4, n. 2, 2017. Disponível em: https://periodicos.ufpb.br/index.php/itec/article/view/40539 Acesso em: 12 jan.2021.

MELLO, R. dos S. et al. Dados semi-estruturados. 2000. Disponível em: https://www.ime.usp.br/~jef/semi-estruturado.pdf. Acesso em: 15 set. 2019.

MICHEL, F.; MONTAGNAT, J.; ZUCKER, C. F. **A survey of RDB to RDF translation approaches and tools**. Rapport de recherche, 2014. Disponível em: https://hal.inria.fr/hal-00903568. Acesso em: 25 maio 2019.

MILLER, E. **The semantic web**. University of Michigan. Jan. 2004. Disponível em: https://www.w3.org/2004/Talks/0120-semweb-umich/. Acesso em: 25 set. 2019.

MINAYO, M. C. de S. **Pesquisa social:** teoria, método e criatividade. Petrópolis: Vozes, 2012.

MNM. **Welcome to MnM homepage**. 2004. Disponível em: http://projects.kmi.open.ac.uk/akt/MnM/index.html. Acesso em: 11 jan. 2020.

MONGODB. **The database for modern applications.** 2020. Disponível eletronicamente: https://www.mongodb.com/. Acesso em: 11 jan. 2020.

MONTALVILLO MENDIZABAL, L. **Definicion y desarrollo de herramienta web de gestión de metadatos Business Intelligente.** Dissertação (mestrado)- Universidad Politecnica de Catalunia, Barcelona, 2012. Disponível em: https://core.ac.uk/download/pdf/41807126.pdf. Acesso em: 11 jan.2021.

MONTEIRO, L. **A Internet como meio de comunicação:** possibilidades e limitações. Dissertação (Mestrado)- Pontifícia Universidade Católica do Rio de Janeiro, 2001. Disponível em: http://www.portcom.intercom.org.br/pdfs/62100555399949223325534481085941280573.pdf. Acesso em: 11 jan. 2020.

MORARU, A.; MLADENIĆ, D. A framework for semantic enrichment of sensor data. **Journal of Computing and Information Technology**, v. 20, n. 3, p. 167–173, 2012. Disponível em: https://hrcak.srce.hr/index.php?id_clanak_jezik=132071&show=clanak. Acesso em: 11 jan. 2020.

MOREIRA, M. A.; BUCHWEITZ, B. **Novas estratégias de ensino e aprendizagem:** os mapas conceituais e o vê epistemológico. Lisboa: Plátano Edições Técnicas, 1993.

MOTIK, B. et al. (eds.) **OWL2 web ontology language profiles**. W3C, 2009. Disponível em: https://www.w3.org/2007/OWL/draft/ED-owl2-profiles-20090420/all.pdf. Acesso em: Jan. 2020.

MOTIK, B.; PARSIA, B; PATEL-SCHNEIDER, P. **OWL2 web ontology language**: XML serialization. W3C, 2009. Disponível em:

https://www.w3.org/2007/OWL/draft/ED-owl2-xml-serialization-20090420/all.pdf. Acesso em: 11 jan. 2020.

MOTIK, B.; PATEL-SCHNEIDER, P. F.; GRAU, B. C. **OWL2 web ontology language direct semantics**. W3C, 2009. Disponível em: https://www.w3.org/2007/OWL/draft/ED-owl2-direct-semantics-20090914/all.pdf. Acesso em: 11 jan. 2020.

MOTIK, B.; PATEL-SCHNEIDER, P. F.; PARSIA, B. **OWL2 web ontology language structural specification and functional style syntax.** 2. ed. W3C, 2012. Disponível em: https://www.w3.org/TR/owl2-syntax/. Acesso em: 11 jan. 2020.

NECHES, R. et al. **Enabling technology for knowledge sharing. AI Magazine**, v. 12, n. 3. 1991. Disponível em: https://doi.org/10.1609/aimag.v12i3.902. Acesso em: 11 jan. 2020.

NHACUONGUE, J. A.; DUTRA, M. L. De Paul Otlet à web semântica: aportes teóricos sobre a organização do conhecimento. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 13, n. 2. 2018.

NHACUONGUE, J. A.; ROZSA, V.; DUTRA, M. D. Linked data e ciência da informação: diretrizes para a publicação de datasets institucionais abertos. **Biblios**, Pittsburgh, n. 73, p. 20-34, 2018. Disponível em: http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1562-47302018000400002&lng=es&nrm=iso. Acesso em: 11 jan. 2020.

NOLL, R. P.; SACCOL, D. B.; EDELWEISS, N. Uma proposta para análise de similaridade entre documentos XML e ontologias definidas em OWL. *In*: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 2007. **Proceedings** … [S.l.: s.n.], 2007.

NOY, N. F., FERGERSON, R. W., MUSEN, M. A. The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility. *In*: DIENG, R., CORBY, O. (eds.). **Knowledge engineering and knowledge management**: methods, models, and tools. EKAW 2000. Lecture Notes in Computer Science, v. 1937, 2000. Disponível em: https://doi.org/10.1007/3-540-39967-4_2. Acesso em: 16 ago. 2020.

NOY, N. F.; MCGUINNESS, D. L. Ontology development 101: a guide to creating your first ontology. **Tech. Rep**. 2001. Disponível em: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf. Acesso em: 11 jan. 2020.

NUGRAHENI, E.; AKBAR, S.; SAPTAWATI, G. A. P. Framework of semantic data warehouse for heterogeneous and incomplete data. *In:* IEEE. **Conference Paper**. 2016. Disponível em: https://ieeexplore.ieee.org/abstract/document/7519397/citations. Acesso em: 11 jan. 2020.

OLIVEIRA, B. C. N. et al. A platform to enrich, expand and publish linked data of police reports. *In:* **Proceedings of the IADIS International Conference on WWW/Internet** (ICWI). Manheim, Germany: [s.n.], 111–118 p. 2016. Disponível em: http://iadisportal.org/digital-library/a-platform-to-enrich-expand-and-publish-linked-data-of-police-reports. Acesso em: 11 jan. 2020.

OLIVEIRA, M. I. S.; LÓSCIO, B. F. **Ecossistemas de dados na Web**: teoria e desafios. SBBD, 2019. Disponível em: https://sbbd.org.br/2019/wp-content/uploads/sites/6/2019/10/Apresentacao_Minicurso_3_EcossistemasdeDados.pdf. Acesso em: 05 jan. 2020.

ONTOWIKI. 2019.Disponível em: http://ontowiki.net/ Acesso em: 11 jan.2020.

OPEN DATA INSTITUTE. **How to open up data**. 2015. Disponível em: http://opendatahandbook.org/guide/en/how-to-open-up-data/. Acesso em: 13 out. 2019.

OPEN DATA INSTITUTE. **Open data certificate**: the mark of quality and trust for open data. 2020. Disponível em: https://certificates.theodi.org/. Acesso em: 11 jan. 2020.

OPEN DATA INSTITUTE. **Guides**. 2020.Disponível em: http://theodi.org/guides. Acesso em: 11 jan. 2020.

OPEN KNOWLEDGE FOUNDATION. **The open data handbook**. 2012. Disponível em: http://opendatahandbook.org/guide/en/. Acesso em: 11 jan. 2020.

OPENREFINE. 2020. Disponível em: https://openrefine.org/. Acesso em: 11 jan. 2020.

OWL WORKING GROUP. **Web Ontology Language (OWL)**. 2012. Disponível em: https://www.w3.org/OWL/. Acesso em: 13 fev. 2020.

PASTOR SÁNCHEZ, J. A. Tecnologías de la web semántica. **Revista Española de Documentación Científica**. Barcelona: Editorial UOC, 2012.

PASTOR-SÁNCHEZ, J.A.; MARTÍNEZ-MÉNDEZ, F.; RODRÍGUEZ-MUÑOZ, J. V. SKOS application for interoperability of controlled vocabularies in the field of linked open data. **El Profesional de la Informacion**, v. 21 n. 3, p. 245-253, 2012.

PATEL, A.; JAIN, S. Present and future of semantic web technologies: a research statement. **International Journal of Computers and Applications**, p. 1-10, 2019. Disponível em: https://www.tandfonline.com/loi/tjca20. Acesso em: 11 jan. 2020. DOI:10.1080/1206212x.2019.1570666.

PATEL-SCHNEIDER, P. Using description logics for RDF constraint checking and closed-world recognition. **Proceedings of AAAI Conference on Artificial Intelligence**, v. 29, n. 1, p. 247-253, 2015. Disponível em: https://ojs.aaai.org/index.php/AAAI/article/view/9177. Acesso em: 9 jan. 2021.

PATEL-SCHNEIDER, P.; MOTIK, B. (eds.). **OWL 2 web ontology language**: mapping to RDF graphs. 2. ed. W3C, 2012. Disponível em: https://www.w3.org/TR/owl2-mapping-to-rdf/. Acesso em: 11 jan. 2020.

PEREIRA, C. M.; LUZ, L. P. da; SANTAREM SEGUNDO, J. E. 2019. 1-21 p. Papéis do profissional da informação na evolução da web semântica. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, ENANCIB (19., 2019, Florianópolis), 2019. Disponível em: https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/619/908. Acesso em: 11 jan.2020.

PEREIRA, E.C. **Gestão da Informação no agribusiness paranaense: estudo exploratório do Programa Paraná Industrial.** Dissertação (Mestrado)-Pontíficia Universidade Católica de Campinas, Campinas, 2001.

PETRIDIS, K. et al. M-OntoMat-annotizer: image annotation linking ontologies and multimedia low-level features. *In:* INTERNATIONAL CONFERENCE ON KNOWLDEGE-BASED AND INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, 2006. **Lecture notes in computer science**, p. 633-640, 2006. Disponível em: https://link.springer.com/chapter/10.1007/11893011_80. Acesso em: 11 jan. 2020.

PETRIE, C. J. Agent-based engineering, the web, and intelligence. **IEEE Expert**, v. 11, n. 6, p. 24-29, 1996.

PIEDRA, N. et al. Marco de trabajo para la integración de recursos digitales basado en un enfoque de web semántica. **Revista Iberoamericana de Sistemas e Tecnologias de Informação**, v. 3, p. 55-70, 2015. Disponível em: https://www.researchgate.net/publication/279634259_Marco_de_Trabajo_para_la_Integracion_de_Recursos_Digitales_Basado_en_un_Enfoque_de_Web_Semantica. Acesso em: 11 jan. 2021.

PIETROFORTE, A. V. S.; LOPES, I. C. Semântica lexical. *In:* **Introdução à lingüística II**: princípios de análise**.** São Paulo: Contexto. 2014.

PINTO, S. C. C. da S. **Composição em WebFrameworks**.Tese (doutorado)-Pontifícia Universidade Católica do Rio de Janeiro (PUC-RJ), 2000.

POLLOCK, R. **Building the (open) data ecosystem**. 2011. Disponível em: https://blog.okfn.org/2011/03/31/building--the-open-data-ecosystem/. Acesso em: 11 jan. 2020.

POOLPARTY thesaurus server. 2020. Disponível em: https://www.poolparty.biz/wp-content/uploads/2013/04/pp-thesaurus-server.pdf. Acesso em: 11 jan. 2020

POPOV, B. et al. KIM: semantic annotation platform. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, ISWC 2003. **Lecture notes in computer science**, v. 2870. Berlin: Springer, 2003. Disponível em: https://doi.org/10.1007/978-3-540-39718-2_53. Acesso em: 11 jan. 2020.

PORIA; S. et al. Towards an intelligent framework for multimodal affective data analysis. **Neural Networks**, v. 63, p. 104-116, 2015. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0893608014002342. Acesso em: 11 jan. 2020.

POSSEMATO, T. OpLiDaf: Open Linked Data Framework: a platform for the creation and publication of linked data. **JLIS.it,** v. 4, n. 1, p. 271, 2013.

QUIN, L. **Extensible Markup Language (XML)**. W3C, 2015. Disponível em: https://www.w3.org/XML/. Acesso em: 11 jan. 2020.

QUIN, L. R. E. **XML essentials**. 2015. Disponível em: https://www.w3.org/standards/xml/core. Acesso em: 10 ago. 2020.

QUINTÃO, P. *et al.* Análise dos desafios e melhores práticas para resguardar a segurança e privacidade dos dispositivos móveis no uso das redes sociais. **Conferência IADIS**. Ibero-Americana WWW/Internet, p.297-236, dez. 2010.

RAMAKRISHAN, R.; GEHRKE, J. **Database management systems**. 3. ed. Rio de Janeiro: McGraw-Hill, 2003.

RAUTENBERG, S. et al. **Guia prático para publicação de dados abertos conectados na web**. Curitiba, PR: Appris, 2018.

REDHAT. **GraphQL**: o que é e para que serve?. 2021. Disponível em: https://www.redhat.com/pt-br/topics/api/what-is-graphql. Acesso em: 30 mar. 2021.

REEVE, L.; HAN, H. Survey of semantic annotation platforms. *In*: **Proceedings of the 20th Annual ACM Symposium on Applied Computing**, Santa Fe, New Mexico, USA, 2005, p. 1634–1638. Disponível em: https://dl.acm.org/doi/10.1145/1066677.1067049. Acesso em: 25 mar. 2020

RODRÍGUEZ PEROJO, K.; RONDA LEÓN, R. Web semántica: un nuevo enfoque para la organización y recuperación de información en el web. **ACIMED**, Habana, vol.13, n.6, nov.-dic., 2005.

ROESLER, M.; HAWKINS, D. T. Intelligent agents: software servants for an electronic information world (and more!). **Online**, v. 18, n. 4, July, 1994. Disponível em: https://dl.acm.org/doi/abs/10.5555/189322.189326. Acesso em: 20 jan. 2020.

ROZSA, V.; DUTRA, M. L.; NHACUONGUE, J. A. Linked open data no contexto acadêmico: identificação e análise de vocabulários utilizados na academia e na pesquisa científica. **Brazilian Journal of Information Studies**: research trends, v. 11, n. 3, p. 34-52, 2017.

RUBERTO, D. L. V. G.; ANTONIAZZI, R. L. Análise e comparação de algoritmos de similaridade e distância entre strings adaptados ao português brasileiro. *In*: ESCOLA REGIONAL DE BANCO DE DADOS (ERBD), 13., 2017, Passo Fundo. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação,

2017. p. 27-36. Disponível em: https://sol.sbc.org.br/index.php/erbd/article/view/3035/2997. Acesso em: 18 jan. 2019.

SAHOO, S. S. et al. **A survey of current approaches for mapping of relational databases to RDF**. 2009. Disponível em: https://www.w3.org/wiki/images/a/ac/Rdb2RdfXG$$StateOfTheArt$RDB2RDF_SurveyReport.pdf. Acesso em: 15 nov. 2019.

SALVADORI, I. et al. Improving entity linking with ontology alignment for semantic microservices composition. **International Journal of Web Information Systems**., v. 13, n. 3, p. 302-323, 2017. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/IJWIS-04-2017-0029/full/html. Acesso em: 11 jan. 2020.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. P. B. **Metodologia de pesquisa**. 5 ed. Porto Alegre: Penso. 2013.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, Londrina, v. 21, n. 2, p. 116-142, 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940. Acesso em: 11 jan. 2020.

SANTAREM SEGUNDO, J. E. Web Semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente as iniciativas internacionais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v.8, p.219–239, 2015. Disponível em: http://inseer.ibict.br/ancib/index.php/tpbci/article/view/207. Acesso em: 12 set. 2019.

SANTAREM SEGUNDO, J. E. Web semântica: fluxo para publicação de dados abertos e ligados. **Informação em Pauta**, Fortaleza, v. 3, p. 117-140, 2018. Disponível em: http://www.periodicos.ufc.br/informacaoempauta/article/view/39721. Acesso em: 11 jan. 2020.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Tecnologias da web semântica aplicadas a organização do conhecimento: padrão SKOS para construção e uso de vocabulários controlados descentralizados. In: GUIMARÃES, J. A. C.; DODEBEI, V. (orgs.). **Organização do conhecimento e diversidade cultural**. Marília: Fundepe, v. 3, p. 224-233, 2015.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S.; LUCAS, E. R. O. Conceitos e tecnologias da Web semântica no contexto da colaboração acadêmico-científica: um estudo da plataforma Vivo. **Transinformação**, Campinas, v. 29, n. 3, p. 297-309, 2017. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862017000300297&lng=en&nrm=iso. Acesso em: 11 jan. 2020.

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

SCHAIBLE J., GOTTRON T., SCHERP A. Survey on common strategies of vocabulary reuse in linked open data modeling. European Semantic Web Conference. The Semantic Web: trends and challenges. **Lecture Notes in Computer Science**, v. 8465, p. 457-472, 2014. Disponível em: https://doi.org/10.1007/978-3-319-07443-6_31. Acesso em: 11 jan. 2020.

SEBUBI, O.; ZLOTNIKOVA, I.; HLOMANI, H. Ontology-driven open data enrichment framework for value optimisation: the case of national performance indicators for Botswana. **IST-Africa Week Conference** (IST-Africa). 2019. Disponível em: https://ieeexplore.ieee.org/document/8764881. Acesso em: 11 jan. 2020.

SENSO; J. A. Herramientas para trabajar con rdf. **El profesional de la información**, v. 12, n. 2, p. 132-139, 2003. DOI:10.1076/epri.12.2.132.15476

SENSO, J. A.; MACHADO, W. A. La publicación em linked data de registros bibliográficos: modelo e implementación. **Revista Española de Documentación Científica**. Universidad de Granada, 2018. Disponível em: http://www.digibis.com/images/articulos/redc-2018-lod-digibis.pdf. Acesso em: 11 jan. 2020.

SEQUEDA, J.; MIRANKER, D. P. Ultrawrap: SPARQL execution on relational data. **JWS, Databases & the Semantic Grid**, First Look, 2013. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3199073. Acesso em: 13 set. 2019.

SEVERINO, A. J. **Metodologia do trabalho científico**. 23. ed. rev., atual. São Paulo: Cortez. 2007.

SHADBOLT, N., O'HARA, K. Linked data in government. **IEEE Internet Computing**, v. 17, n. 4, p. 72-77, 2013. Disponível em: https://ieeexplore.ieee.org/abstract/document/6547636. Acesso em: 11 jan. 2020.

SHEHABUDDEEN, N.; PROBERT, D.; PHAAL, R**. Representing and approaching complex management issues:** part 1: role and definition. Working Paper UC, Cambridge, 2000.

SILBERSCHATZ, A.; SUDARSHAN, S. KORTH, H. F. **Sistema de banco de dados**. 5. ed. Rio de Janeiro: Elsevier. 2006.

SILVA, G, S.; LUZ, L, P. Aplicativo que gera enriquecimento semântico de dados abertos: estudo de caso do metrô de São Paulo. **Revista e-F@tec**, Garça, v.10, n.1, out.2020. Disponível em: https://fatecgarca.edu.br/ojs/index.php/efatec/article/view/206/163. Acesso em: 11 jan. 2021.

SILVELLO, G. et al. Semantic representation and enrichment of information retrieval experimental data. **International Journal on Digital Libraries**, n. 18, p. 145-172, 2017. Disponível em: https://link.springer.com/article/10.1007/s00799-016-0172-8. Acesso em: 22. jan. 2021.

SILVELLO, G. et al. Semantic representation and enrichment of information retrieval experimental data. **International Journal on Digital Libraries**, v. 18, p. 145–172, 2017. Disponível em: https://link.springer.com/article/10.1007/s00799-016-0172-8. Acesso em: 11 jan. 2020

SMITH, M.; WELTY, C., MCGUINNESS, D. **OWL web ontology language guide**. 2004. Disponível em: http://www.w3.org/TR/2003/WD-owl-guide20030210/. Acesso em: 11 jan. 2020.

SMITH, M. et al. **OWL2 web ontology language conformance**. 2. ed. W3C, 2012. Diwsponível em: https://www.w3.org/TR/owl2-conformance/. Acesso em: 11 jan. 2020.

SOCRATA. 2020. Disponível em: https://www.tylertech.com/products/socrata. Acesso em: 11 jan. 2020.

SOLÍS, S. M. **La web semántica**. [S. l.]: Santiago Márquez Sol, 2015.

SORRENTINO, S. et al. Semantic annotation and publication of linked open data. *In:* **13th International Conference**, ICCSA 2013, Ho Chi Minh City, Vietnam, 2013. p. 462-474.

SOUSA, P. T.C., ALVARENGA E. F. R. Agentes Inteligentes: agentes que aprendem e redes neurais – Dissertação (mestrado)- Universidade Católica de Brasília, UCB –Disciplina: Inteligência Artificial e Agentes Inteligentes. Disponível em: https://github.com/sindice/sparqled. Acesso em: jan 2020.

SOUZA, N. A. de; BORUCHOVITCH, E. Mapas conceituais: estratégia de ensino/aprendizagem e ferramenta avaliativa. **Educação em Revista**, v. 26, n. 3, dez. 2010. p. 195-217. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982010000300010&lng=en&nrm=iso. Acesso em: 14 abr. 2020.

SPEICHER, S. **ACTION-10**: convert member submission to editor's draft, with support of JohnArwe ans mhausenblas. 2012. Disponível em: https://www.w3.org/2012/ldp/track/actions/10. Acesso em: 12 jan. 2021.

SPIVACK, N. **How the WebOS evolves?:** minding the planet. Fev., 2007. Disponível em: http://novaspivack.com/nova_spivacks_weblog/2007/02/steps_towards_a.html. Acesso em: 11 jan. 2020.

SPORNY, M.; KELLOGG, G.; LANTHALER, M. **JSON-LD 1.0**: a JSON-based serialization for linked data. 2014. Disponível em:

https://www.w3.org/TR/2014/REC-json-ld-20140116/diff-20131105.html. Acesso em: 12 jan. 2021.

STRONG, D.M.; LEE, Y.W.; WANG, R.Y. 10 Potholes in the road to information quality. **IEEE Computer**, v. 18, n. 162, p. 38-46, 1997. Disponível eletronicamente: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.1299&rep=rep1&type=pdf. Acesso em: 11 jan. 2020.

SVIHLA, M.; JELINEK, I. **Two layer mapping from database to RDF**. 2004. Disponível em: https://www.researchgate.net/publication/228893102_Two_layer_mapping_from_database_to_RDF. Acesso em: 18 mar. 2019.

SZÁSZ, B.; FLEINER, R.; MICSIK, A. Linked data enrichment with self-unfolding URIs. IEEE 14th **International Symposium on Applied Machine Intelligence and Informatics** (SAMI), 2016.

SZYMANSKI J., KRAWCZYK H., DEPTUŁA M. (2013) Retrieval with Semantic Sieve. *In*: SELAMAT, A.; NGUYEN, N. T., HARON, H. (eds). Intelligent Information and Database Systems. ACIIDS 2013. **Lecture Notes in Computer Science**, v. 7802, Berlin: Springer,2013. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-36546-1_25. Acesso em: 11 jan 2020.

TALIGENT WHITE PAPER. **Building object-oriented frameworks**. Taligent, 1994. Disponível em: https://lhcb-comp.web.cern.ch/lhcb-comp/Components/postscript/buildingoo.pdf. Acesso em: 11 jan. 2020.

TALIGENT WHITE PAPER. **Leveraging object-oriented frameworks**. Taligent, 1995. Disponível em: http://lhcb.web.cern.ch/lhcb/computing/Components/postscript/leveragingoo.pdf. Acesso em: 11 jan. 2020.

TALLIS, M. **Semantic word processing for content authors**. 2003. Disponível em: https://www.researchgate.net/publication/228852352_Semantic_word_processing_for_content_authors. Acesso em: 18 jan. 2020.

TAYLOR, C. **Structured vs. unstructured data.** 2018. Disponível em: https://www.datamation.com/big-data/structured-vs-unstructured-data.html. Acesso em: 15 jan. 2020.

THAKKER, D. et al. Taming digital traces for informal learning: a semantic-driven approach. *In* **Proceedings of the 7th European Conference on Technology Enhanced Learning**, Berlin. Heidelberg: Springer, 2012. p. 348-362.

TRIVIÑOS, A. **Introdução à pesquisa em ciências sociais:** a pesquisa qualitativa em educação. São Paulo: Atlas, 1987.

UREN, V. *et al.* Semantic annotation for knowledge management. Journal of Web Semantics, v. 4, p. 14-28, 2005.

VAN HARMELEN, F. **Semantic web technologies as the foundation for the information infrastructure**. 2008. Disponível em: 10.1201/9781420070729.ch3. Acesso em: 12 maio 2019.

VAVLIAKIS, K. N. et al. An integrated framework for enhancing the semantic transformation, editing and querying of relational databases. **Expert Systems with Applications**, v. 38, n. 4, p. 3844-3856, April, 2011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S095741741001002X. Acesso em: 19 out. 2019.

VAVLIAKIS, K. N.; MITKAS, P. A.; KARAGIANNIS, G. **Semantic web in cultural heritage after 2020**: 'what will the semantic web look like 10 years from now?' 2012. Disponível em: https://www.researchgate.net/profile/Konstantinos-Vavliakis/publication/276294705_Semantic_Web_in_Cultural_Heritage_After_2020_What_will_the_Semantic_Web_look_like_10_years_from_now/links/59b18dda0f7e9b37434abafb/Semantic-Web-in-Cultural-Heritage-After-2020-What-will-the-Semantic-Web-look-like-10-years-from-now.pdf. Acesso em: 11 jun. 2019.

VAVLIAKIS, K. N.; SYMEONIDIS, A. L.; MITKAS, P. A. Event identification in web social media through named entity recognition and topic modeling. **Data & Knowledge Engineering**, v. 88, p. 1-24, Nov. 2013. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169023X13000827. Acesso em: 19 nov. 2019.

VELEGRAKIS, Y. Relational Technologies, Metadata and RDF. *In:* VIRGILIO, R. de; GIUNCHIGLIA, F.; TANCA, L. (eds). **Semantic web information management: a model-based perspective**. Heidelberg; New York: Springer, 2010. p. 41-66. Disponível em: http://disi.unitn.it/~velgias/docs/Velegrakis10.pdf. Acesso em: 10 jan. 2020.

VIRTUOSO SPONGER. 2020. Disponível em: http://vos.openlinksw.com/owiki/wiki/VOS/VirtSponger. Acesso em: 10 jan. 2020

VOS.VOSRDFFAQ. **RDF triple store FAQ**. 2020. Disponível em: http://vos.openlinksw.com/owiki/wiki/VOS/VOSRDFFAQ. Acesso em 4 dez. 2020.

W3C; **Asio**. 2009. Disponível em: https://www.w3.org/2001/sw/wiki/Asio. Acesso em: 18 abr. 2019.

W3C. **Building the web of data**. 2020. Disponível em: https://www.w3.org/2013/data/. Acesso em: 8 jan. 2020.

W3C. **Linked data**. 2015a. Disponível em: https://www.w3.org/standards/semanticweb/data. Acesso em: 20 jan. 2020.

W3C. **Metadata activity statement**. 2002. Disponível em: https://www.w3.org/Metadata/Activity.html. Acesso em: 14 jun. 2020.

W3C. **OWL 2 web ontology language document overview.** 2. ed. 2012. Disponível em: https://www.w3.org/TR/owl2-overview/. Acesso em: 11 jan. 2020.

W3C. **Resource description framework (RDF)**. 2014. Disponível em: https://www.w3.org/RDF/. Acesso em: 23 jan. 2020.

W3C. **Semantic web**. 2020. Disponível em: https://www.w3.org/standards/semanticweb/. Acesso em: 13 jan. 2020.

W3C. **SOAP version 1.2 part 1**: messaging framework. 2. ed. 2007. Disponível em: https://www.w3.org/TR/soap12/. Acesso em: 19 nov. 2019.

W3C. **SPARQL 1.1 overview**. 2013. Disponível em: https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/. Acesso em: 9 jan. 2020.

W3C. **Ultrawrap**. 2012a. Disponível em: https://www.w3.org/2001/sw/wiki/Ultrawrap. Acesso em: 23 ago. 2019.

W3C. **Xml technology**. 2015. Disponível em: https://www.w3.org/standards/xml/. Acesso em: 15 jan. 2020.

W3C. **Web architecture**. 2015b. Disponível em: https://www.w3.org/standards/webarch/. Acesso em: 15 jan. 2020.

WOOD, D.et al. **Linked data: structured data on the Web.** [S. l.]: Manning, 2013. ISBN 9781617290398.

WOOLDRIDGE, N.R.; JENNINGS, M. Intelligent agents: theory and practice. **Knowledge Engineering Review**, v. 10, n. 2, p. 115-152, 1995.

YU, L. **A developer´s guide to the Semantic Web.** New York: Springer. 2011.

YUNIANTA, A. et al. Ontology development to handle semantic relationship between Moodle e-learning and question bank system. **Proceedings of** The First International Conference on Soft Computing and Data Mining (SCDM-2014), p. 691-701, 2014. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-07692-8_65. Acesso em: 25 set. 2019.

YUNIANTA, A. et al. Semantic data mapping on e-learning usage index tool to handle heterogeneity of data representation. **Technology, education and science of international conference** (TESIC), v. 69, n. 5, 2013. Disponível em: https://journals.utm.my/jurnalteknologi/article/view/3193. Acesso em: 30 mar. 2019.

YUNIANTA, A. et al. Semantic data mapping technology to solve semantic data problem on heterogeneity aspect. **International Journal of Advances in**

Intelligent Informatics, v. 3, n. 3, p. 161-172, 2017. Disponível em: https://repository.unmul.ac.id/bitstream/handle/123456789/3254/10.%20131-509-5-PB.pdf?sequence=1&isAllowed=y. Acesso em: 30 mar. 2019.

ZAVERI, A. et al. Quality assessment for linked data. **Semantic Web Journal**, Wright State University, USA, 2015. Disponível em: http://www.semantic-web-journal.net/system/files/swj773.pdf. Acesso em: 11 jan. 2020.

ZUBCOFF, J. J.et al. The university as an open data ecosystem. **International Journal of Design & Nature and Ecodynamics**., v. 11, n. 3, p. 250-257, 2016. Disponível em: https://www.researchgate.net/publication/305640251_The_University_As_An_Open_Data_Ecosyste. Acesso em: 17 jan. 2020.

ZUIDERWIJK A. et al. **Socio-technical impediments of open data**. Electronic Journal of e-Government, v. 10, n. 2, p. 156-172, 2012.