UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS QUIXADÁ
CURSO DE SISTEMAS DE INFORMAÇÃO

Salomão da Silva Santos

UM PROCESSO PARA CONVERSÃO E PUBLICAÇÃO DE DADOS PARA MODELO RDF SEGUINDO OS PRINCÍPIOS DE *LINKED DATA*

Quixadá, Ceará
2016

Salomão da Silva Santos

UM PROCESSO PARA CONVERSÃO E PUBLICAÇÃO DE DADOS PARA MODELO RDF SEGUINDO OS PRINCÍPIOS DE *LINKED DATA*

Trabalho de Conclusão de Curso submetido à Coordenação do Curso de Sistemas de Informação do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Sistemas de Informação.

Orientador: Prof Msc. Regis Pires Magalhães

Quixadá, Ceará
2016

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Campus de Quixadá

S233p Santos, Salomão da Silva
Um processo para conversão e publicação de dados para modelo rdf seguindo os princípios de Linked Data / Salomão da Silva Santos. – 2016.
104 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Bacharelado em Sistemas de Informação, Quixadá, 2016.
Orientação: Prof. Msc. Regis Pires Magalhães
Área de concentração: Computação

1. Fluxo de trabalho 2. Web semântica 3. Linked data 4. Metadados 5. Framework (Programa de computador) I. Título.

CDD 005.1

Salomão da Silva Santos

UM PROCESSO PARA CONVERSÃO E PUBLICAÇÃO DE DADOS PARA MODELO RDF SEGUINDO OS PRINCÍPIOS DE *LINKED DATA*

Trabalho de Conclusão de Curso submetido à Coordenação do Curso de Sistemas de Informação do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Sistemas de Informação. Área de concentração: Computação.

Aprovada em: __ /__/ ____

BANCA EXAMINADORA

Prof Msc. Regis Pires Magalhães (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Flávio Rubens de Carvalho Sousa
(Membro)
Universidade Federal do Ceará (UFC)

Profª. Msc. Ticiana Linhares Coelho da Silva
(Membro)
Universidade Federal do Ceará (UFC)

Dedico este trabalho a Deus e a minha família principalmente minha mãe, meu pai e meu irmão.

AGRADECIMENTOS

Agradeço primeiramente a Deus por me dar saúde, força e sabedoria para superar as dificuldades e por ter me proporcionado mais esta conquista.

Sou grato pela educação e todos os esforços de minha família, em especial aos meus pais, Antonia Vieira da Silva e José Soares dos Santos, e ao meu irmão, Elias da Silva Santos.

Aos demais familiares, amigos e amigas da igreja que acompanharam todo o período de estudos da graduação, pela torcida e por compreenderem minha ausência em algumas ocasiões.

Aos professores da Universidade Federal do Ceará campus Quixadá, em especial, o professor Msc. Regis Pires Magalhães, meu orientador, pela paciência, incentivos, ensinos e dedicação, neste trabalho.

Ao Prof. Dr. Davi Romero, meu orientador durante quase três anos no grupo PET, por seu incentivo, apoio, paciência e ensinamentos, que serão levados comigo para toda vida.

Ao colega Roberval G. Mariano e à Profa. Dra. Vânia M. P. Vidal, pela troca de experiências e pelo trabalho em conjunto para integração de nossas abordagens de publicação de dados conectados na Web, seguindo o padrão de *Linked Data*.

Aos professores Msc. Ticiana Linhares e Dr. Flávio Rubens pela gentileza de compor a banca examinadora e pelas ricas contribuições.

Agradeço aos amigos do grupo PET e do ARiDA, aos amigos de graduação que sempre estiveram presentes, em especial, Ana Klyssia, Araújo Filho, Italo Pessoa, Adeilson, Alex, Anderson, Claudio, Cinthia Maria, Gerlan, Guilherme, Mardson, Júnior Leonel, Júnior Holanda, Ricardo Lopes, Tércio Jorge, Sávio de Castro, Warnly, William, Otávio Augusto, Cleiton Brito, Romário Férias, Bruno Gomes, Caio Pessoa, Thiago Vinutto, Ricardo Avila, Narciso Arruda pelas muitas madrugadas de estudos, brincadeiras e companheirismo, ao longo desse curso.

Aos companheiros de estágio, professores e desconhecidos que contribuíram sem perceber para que esse sonho se realizasse. Obrigado a todos.

"Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito.
Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes."
(Marthin Luther King)

RESUMO

Passado o período inicial de grande entusiasmos pela publicação de novas bases de dados em *Linked Data*, a comunidade científica resolve agora passar a avaliar as bases quanto à sua qualidade. Contudo, alguns problemas têm se apresentado frequentemente, são eles: falta de um processo que incentive, guie e contribua com o aprimoramento da publicação e reutilização de dados conectados na Web e não poder identificar e ou verificar o fluxo de trabalho realizado antes da publicação. Portanto, este trabalho tem como objetivo propor um processo que incentive, guie e contribua com o aprimoramento da publicação e reutilização de dados conectados na Web, bem como, simplificar e recomendar a utilização de algumas ferramentas, padrões, princípios e boas práticas para transformação, interligação, exposição e compartilhamento de recursos de dados no modelo RDF, levando em consideração os padrões de Linked Data. Foi realizado um exemplo de aplicação real envolvendo dados abertos e transparência do governo Brasileiro sobre o combate às empresas fraudulentas de licitações de compras públicas para evidenciar as contribuições da abordagem proposta. Além disso, a execução de um *workflow* de ETL (Extração, Transformação e Carga) através da ferramenta Pentaho e do *plugin* ETL4LOD para converter fontes de dados de diversos formatos, para RDF, a fim de demonstrar que esta ferramenta é capaz de automatizar o processo de geração e atualização de RDF, bem como, realizar consultas SPARQL.

Palavras-chaves: Integração de Dados. *Linked Data*. Publicação de dados. Processo de Triplificação.

ABSTRACT

After the initial period of great enthusiasm through the publication of Databases in linked data, the scientific community decides now evaluate the bases as their quality. However, some problems have often presented, such as: lack of a process which encourages, guides and contributes with the improvement of publication and reuses of data connected to the web and can not identify, verify and/or evaluate the source and workflow conducted prior to publication. Therefore, this study aims to propose a process that encourages, guides and contributes with the improvement of publication and reutilization of data connected in the Web, as well as simplifies and recommends the use of some tools, standards, principles and good practices for transformation, interconnection, exposure, and sharing of data resources in RDF model, considering Linked Data patterns. It was performed an example of a real application involving open data and transparency of the Brazilian government related to the combat of fraudulent companies of bidding public commerce to highlight the contribution of the proposed approach. Besides that, the execution of a workflow ETL (Extraction, Transformation and Load) by Pentaho tool and ETL4LOD plugin to convert data sources of various formats, to RDF, in order to demonstrate that this tool is able to automate the generation process and update of RDF, as well as SPARQL queries.

Key-words: Data Integration. Liked Data. Publishing data. Triplify Process.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo, introduz-se o cenário em que este trabalho está inserido e a síntese das contribuições. Portanto, apresentam-se a motivação, caracterização do problema, objetivo a ser alcançado, assim como a organização dos demais capítulos.

1.1 Motivação

A *Web* é atualmente um enorme espaço global de documentos e dados distribuídos em múltiplas fontes heterogêneas. Diante deste cenário, tem-se procurado um modo de organizar este grande volume de dados, de maneira que, seja possível integrá-los e compartilhá-los com facilidade. Além de permitir fácil processamento e interpretação do conteúdo por parte de aplicações que façam uso desses dados. Em 2001, para suprir a necessidade de interligação entre os dados da *Web* de Documentos e também suprir a carência de semântica explícita, a fim de que os dados pudessem ser acessados e processados por agentes computacionais, Berners-Lee, Hendler e Lassila (2011) propuseram uma extensão para a Web existente até então. A visão proposta por eles, denominada *Web Semântica*, consistia na criação de uma *Web* de Dados, onde os dados seriam disponibilizados e interligados na *Web*, de maneira que pudessem ser usados por computadores não somente com propósitos de apresentação, mas para automação, integração e reutilização entre as aplicações.

A *Web Semântica* fornece tecnologias para publicação, recuperação e integração de dados distribuídos na Web de dados. Resumindo, essas tecnologias são: (i) *Resource Description Framework* (RDF), segundo Manola e Miller (2004) é um modelo de dados descentralizado, baseado em grafo e extensível, possuindo um alto nível de expressividade e permitindo a interligação entre dados de diferentes conjuntos de dados; (ii) *Uniform Resource Identifier* (URI), usado como mecanismo que identifica um recurso físico ou abstrato; (iii) linguagem SPARQL para Prud'hommeaux e Seaborne (2008), é uma linguagem de consulta de alto nível, capaz de abstrair detalhes da sequência de passos necessária para a execução de consultas sobre fontes heterogêneas; e possibilidade de uso de mecanismos de inferência sobre os dados.

Ainda no âmbito da *Web Semântica*, Heath e Bizer (2011a) consideram *Linked Data* um conjunto de melhores práticas para publicação e consumo de dados estruturados na Web, que permite estabelecer ligações entre itens de diferentes conjuntos de dados para formar um único espaço de dados global.

Desde a publicação do memorando de Tim Berners Lee no *World Wide Web Consortium* (W3C) em 2007, listando alguns dos princípios básicos de dados referenciados até hoje, aconteceu um número significativamente de iniciativas internacionais em publicação de dados no modelo RDF.

Essas iniciativas são resultantes da combinação das tecnologias da *Web Semântica*, dos princípios de *Linked Data* e o projeto *Linking Open Data* (LOD) que estão contribuíndo para a ampliação da Web de Dados. *Linking Open Data* é um esforço comunitário iniciado em 2007 e suportado pelo W3C para identificar fontes de dados publicadas sob licenças abertas, fomentando

convertê-las para RDF e publicá-las na Web usando os princípios de *Linked Data* (HEATH; BIZER, 2011a). No Cenário brasileiro o interesse pelo tema é crescente com algumas propostas e protótipos emergentes nos ambientes acadêmicos e governamentais.

Motivados pelo grande crescimento da quantidade de fontes de dados disponíveis na web e pelo sucesso da iniciativa *Linking Open Data*, um conjunto de questões têm sido levantadas, como (i) identificar fontes de dados relevantes e confiáveis, (ii) selecionar os dados relevantes em meio a tantas fontes, (iii) consumir os dados, (iv) mapear vocabulários fonte para vocabulários destino, (v) quais ferramentas utilizar nas operações de preparar, modificar, transformar em RDF e integrar com outros repositórios de RDF, (vi) qual decisão tomar sobre erros ou problemas durante a transformação de dados para o modelo RDF, (vii) publicar os dados atendendo os princípios de *Liked Data*, (viii) proceder na atualização dos dados publicados em RDF, deixam lacunas, que ainda não podem ser respondidas. Considerando os estudos realizados durante o desenvolvimento deste trabalho, por não se encontrar na literatura um processo que detalhe e facilite a orientação dos usuários quanto ao mapeamento, integração, conversão e publicação de dados para o modelo RDF, através do uso de boas práticas, seguindo os princípios de *Linked Data*. O caminho natural segue a busca por soluções para estas lacunas.

1.2 Caracterização do Problema

O processo peculiar de publicação de *Linked Data* consiste normalmente da extração de dados de múltiplas fontes heterogêneas, seguida da execução, através do uso de diversas ferramentas, de uma gama de operações de preparação, modificação e integração dos dados antes de carregá-los em um banco de triplas RDF (CORDEIRO et al., 2011).

Desta maneira, o processo de publicação de *Linked Data* e a suas respectivas operações de extração, preparação, modificação, triplificação, e integração, tem carência por um processo que facilite e oriente os usuários quanto a esses passos.

Além disso, com relação à recuperação efetiva dos dados publicados no contexto de *Linked Data*, surge ainda a necessidade de que estes estejam vinculados a outras fontes de dados publicados na Web (HARTIG; ZHAO, 2010), para que, assim, a partir de um determinado dado na Web seja possível recuperar os dados por ele referenciados. Neste sentido, também configura-se como um problema, a dificuldade de interligar os dados com os demais dados publicados na Web, devido, elementos diferentes representarem o mesmo recurso em diferentes conjutos de dados.

Realizada esta exposição introdutória, segue a caracterização do problema analisado por este trabalho: *"A carência de um processo que incentive, guie e contribua com o aprimoramento da publicação e reutilização de dados abertos conectados na Web, seguindo os princípios de Linked Data, bem como, orientações quanto a mapeamento, integração, triplificação, publicação e sugestões de ferramentas."*

1.3 Objetivo

O objetivo principal deste trabalho é propor um processo que facilite a disponibilização de dados na Web, no formato RDF, seguindo os princípios de *Linked Data*. O processo possui o intuito de simplificar e recomendar a utilização de algumas ferramentas, padrões e diretrizes para coletar, triplificar, interligar, expor e compartilhar recursos de dados nos padrões de *Linked Data*.

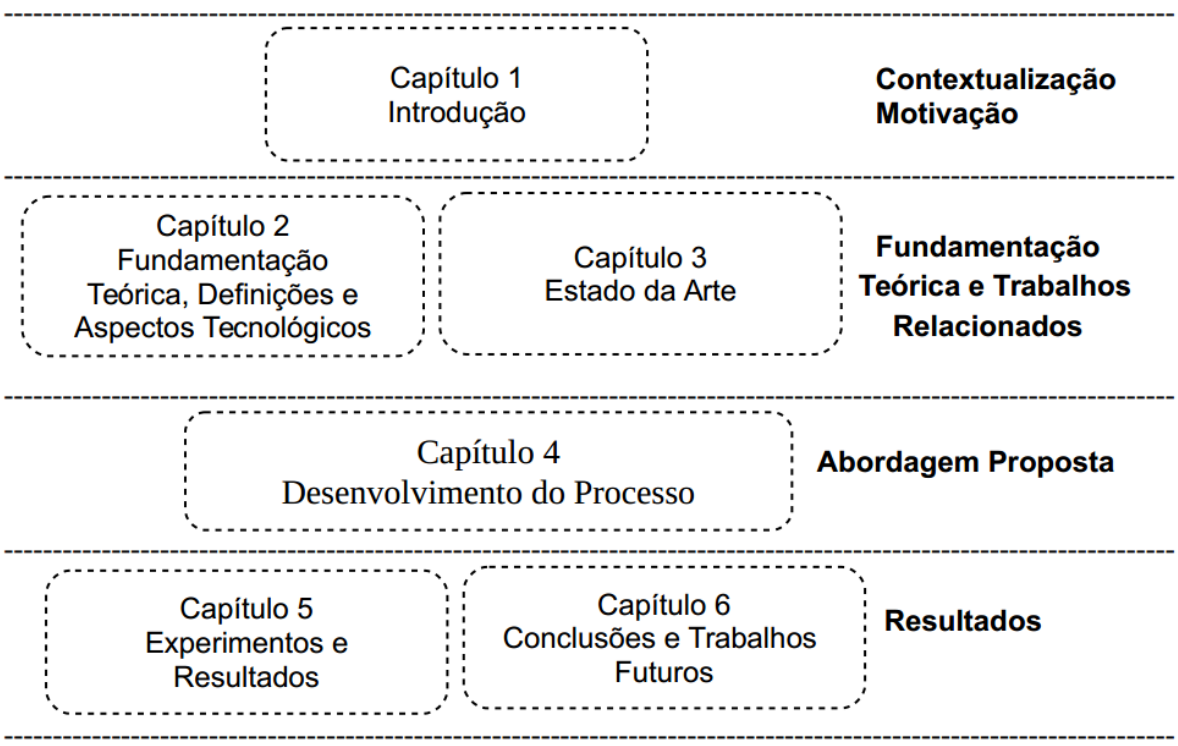
Finalmente, na fase de análise e validação do processo, um estudo de caso, aplicando um cenário com dados reais e abertos a respeito de empresas e pessoas inabilitadas ou inidôneas para desempenharem atividades ou funções públicas.

Os resultados evidenciaram as contribuições do processo quanto a facilitação e simplificação dos passos do processo de publicação de dados na Web Semântica seguindo os princípios e padrões de *Linked Data*. Além disso, o aumento da confiabilidade dos dados por seguir os princípios e padrões de *Liked Data*, bem como, o acompanhamento do processo de publicação dos dados no modelo final.

1.4 Organização do Trabalho

Os estudos e análises realizados no desenvolvimento desta pesquisa estão estruturados em quatro partes organizados em capítulos ilustrados na Figura 1 a seguir:

Figura 1 – Organização deste trabalho.



Fonte: Elaborado pelo autor.

Este trabalho está organizado em 6 capítulos. Não sendo mais necessário tratar deste capítulo introdutório, os demais são delineados a seguir.

No capítulo 2, faz-se uma síntese dos conceitos e definições relevantes para o entendimento dos demais capítulos deste trabalho, bem como os principais trabalhos relacionados. Ele apresenta os conceitos relacionados a integração, mapeamento, conversão, publicação de dados, *Uniform Resource Identifier* (URI), *Resource Description Framework*(RDF), links RDF, *Linked Data*, além de tratar também sobre linguagem *SPARQL.*

O capítulo 3, se destina ao estado da arte, apresentando uma revisão de alguns trabalhos da área. Assim como, análise comparativa considerando as etapas e padrões comuns utilizados no processo de publicação de dados na Web, seguindo os princípios de *Linked Data*.

O capítulo 4, expõe o processo desenvolvido neste trabalho e sua arquitetura. Considerando os principais passos da triplificação e publicação dos dados seguindo os princípios e boas praticas de *Linked Data*.

O capítulo 5, apresenta um estudo de caso, com um exemplo de aplicação envolvendo a publicação e interligação de dados reais e abertos a respeito de empresas e pessoas inabilitadas ou inidôneas para desempenharem atividades ou funções públicas.

Por fim, o capítulo 6 destaca as principais contribuições deste trabalho e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A Web atual deixou de ser apenas um espaço global de documentos interligados e está se tornando também um enorme espaço global de dados vinculados constituído de bilhões de triplas RDF que cobrem os mais variados domínios, denominada Web de Dados (HEATH; BIZER, 2011a). A Web de Dados baseia-se nos princípios *Linked Data* que formam a base para a difusão e uso de dados na Web. Portanto, gerando um volume crescente de dados nos mais diversos domínios e, consequentemente, uma demanda por seu consumo. A Web de Dados fornece um novo cenário à integração de dados, mas também traz novos desafios à pesquisa, como os passos, princípios e padrões essenciais a serem seguidos na publicação, entre outros.

Seguindo esse contexto, apresenta-se as áreas de pesquisas relevantes para este trabalho, que podem ser visualizadas na Figura 2. No lado esquerdo estão os principais tópicos relacionados à integração de dados, destacando-se ontologias e mapeamentos. Quanto ao lado direito da figura, os campos relacionados aos *Linked Data* são descritos. Entre eles, a linguagem SPARQL, RDF/RDFS/OWL, linguagem, protocolo e *endpoint* SPARQL, bem como princípios e boas práticas.

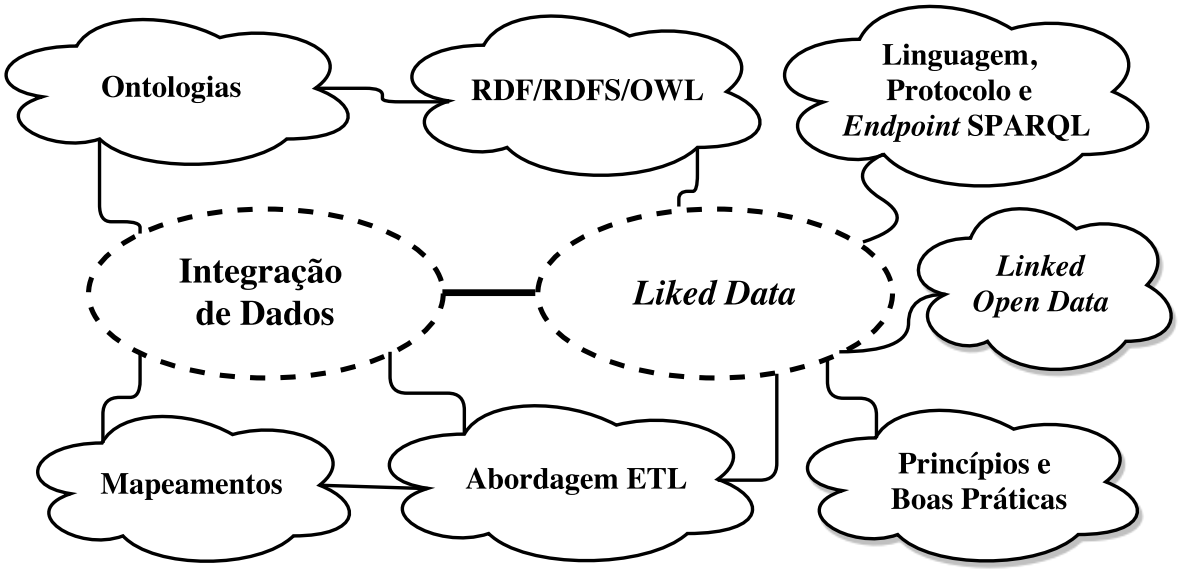


Figura 2 – Áreas e subáreas de pesquisa relacionadas.

A seguir serão apresentados os conceitos essenciais para uma melhor compreensão do contexto em que está inserido este trabalho, bem como da fundamentação necessária ao entendimento dos capítulos seguintes.

2.1 Ontologias

De acordo com Smith e Welty (2001), ontologia é um termo inicialmente ligado à esfera filosófica como sendo a ciência que estuda em quais tipos e estruturas se classificam os objetos, as propriedades, os eventos, os processos e as relações existentes, mas, posteriormente a definição foi ampliada por Fensel (2001): “Ontologia é uma especificação explícita de uma

conceitualização e uma descrição formal dos conceitos e relacionamentos compartilhados de uma área de conhecimento".

Conceitualização refere-se então a um modelo abstrato de um domínio: **explícita** é concernente a definições de nomenclaturas não-ambíguas; **formal** significa passível de ser processada automaticamente; e **compartilhada** representa o conhecimento consensual de um domínio (PINHEIRO, 2011).

Embora existam diferentes definições do termo ontologia, algumas noções básicas quanto à estrutura são compartilhadas pela maioria das abordagens. Considerando-se que uma ontologia pode ser representada por uma hierarquia de conceitos e de relações, a seguir são apresentados alguns elementos usados para representar as ontologias (NOY; MCGUINNESS et al., 2001) (PINHEIRO, 2011):

a) **Classe** (também conhecida como conceito) – descreve conceitos em um domínio. Por exemplo, uma classe de Vinho representa todos os vinhos. Uma classe pode ter subclasses que representam conceitos que são mais específicos do que a superclasse (NOY; MCGUINNESS et al., 2001). Por exemplo, pode-se dividir a classe de todos os vinhos em tinto, branco e rosé. Alternativamente, pode-se dividir uma classe de todos os vinhos em espumantes e não-espumante.

b) **Propriedade** (também chamado de predicado) – vista como relação, uma vez que é usada para estabelecer um relacionamento entre dois termos. O primeiro termo deve ser um conceito que represente o domínio da relação; e o segundo termo deve ser um conceito que represente o contradomínio (range) da relação. Por exemplo, um revisor poderia ser representado como uma relação de tal forma que seu domínio é pessoa e o contradomínio é revisões. O contradomínio de uma propriedade também pode ser um tipo de dado primitivo como string, decimal ou boolean. E uma relação pode ter subrelações.

C) **Instância** (indivíduo) – unidade materializada de uma classe, como Maria é uma instância de pessoa, ou um carro específico que possui uma placa identificando-o unicamente.

Portanto, a motivação principal do paradigma declarativo das ontologias é a modelagem de sistemas em alto nível e mais próximo do conhecimento do domínio a ser modelado. Consiste em uma forma mais flexível de descrever um domínio sem nenhum compromisso com sua implementação (FENSEL, 2001), destacando-se que há a garantia de que, nas ontologias, os dados são estruturados com vocabulário livre de ambiguidades e formalismo passível de processamento automático (NOY; MCGUINNESS et al., 2001), o que torna o uso, em sistemas de integração de dados, muito apropriado (PINHEIRO, 2011).

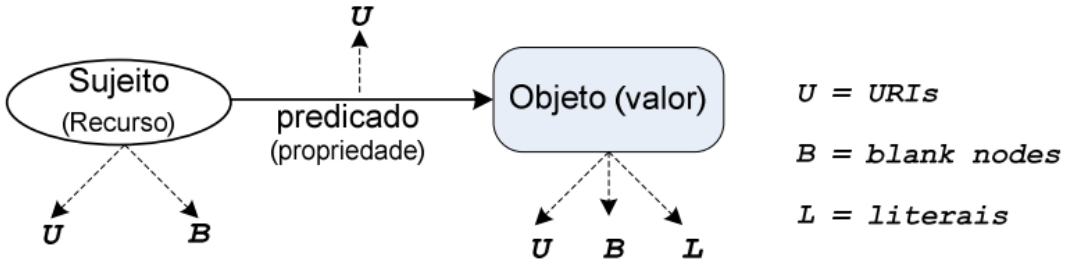
2.2 *Resource Description Framework* (RDF)

Segundo Klyne e Carroll (2004), RDF é o modelo de dados recomendado pelo *World Wide Web Consortium* (W3C) para representar informações sobre os recursos na Web. Manola e Miller (2004) consideram o modelo RDF como um modelo de dados descentralizado, baseado em grafo e extensível, possuindo um alto nível de expressividade e permitindo a interligação entre dados de diferentes conjuntos de dados. Ele foi projetado para a representação integrada de informações originárias de múltiplas fontes.

Logo, os dados são descritos na forma de triplas RDF que é um conjunto de tuplas (s, p, o), em que s, p e o são sujeito, predicado e objeto, respectivamente. O sujeito é um recurso identificado por um espaço de nomes global fornecida pelo uso de URI, o objeto pode ser ser outro recurso relacionado, o valor da propriedade do sujeito uma URI, um literal ou *Blank Node* e o predicado é uma propriedade fornecida pelo uso de URI, que define como sujeito e predicado estão relacionados.

Ainda, sob a representação de grafo, um sujeito (recurso) é representado por uma elipse identificada por um URI, o objeto (valor) por retângulo e o predicado (propriedade) por um arco que conecta o sujeito ao objeto (vide Figura 3).

Figura 3 – Representação gráfica de uma tripla RDF



Fonte: Pinheiro (2011).

O uso de nomes globais é extremamente importante porque as triplas podem sempre ser combinadas (*merged*) sem tradução de nome. Grafos inteiros podem ser transportados e combinados sem tradução, fato que constitui grande vantagem para o intercâmbio e compartilhamento de dados, propiciando integração de dados. Sites como o *eBay* e *Amazon*, por exemplo, fazem uso do RDF na estrutura de pesquisa (PINHEIRO, 2011).

Além disso, o armazenamento de dados no modelo RDF pode ser realizado através de grafo em memória, arquivo texto ou banco de dados específico para armazenamento de triplas RDF, chamado de RDF *Store*, *Triple Store* ou *Quad Store*. Normalmente uma *Triple Store* é de fato uma *Quad Store*, pois suporta Grafos Nomeados (*Named Graphs*). Um grafo nomeado é simplesmente uma coleção de triplas RDF nomeada por uma URI que identifica o grafo, com a finalidade de caracterizar a proveniência dos dados RDF. O armazenamento de triplas em arquivo texto usa algum formato de serialização de RDF, como RDF/XML, Notation3 (N3), Turtle, NTriples ou RDF/JSON (MAGALHÃES, 2012).

2.3 *Resource Description Framework Schema* (RDF-S)

RDF Schema (RDF-S) (BRICKLEY; GUHA, 2004) é o padrão do *World Wide Web Consortium* (W3C) que provê um vocabulário básico para definição de uma taxonomia para as informações descritas em RDF. Além disso, o RDF-S fornece um conjunto de primitivas, das quais se destacam as que permitem a definição de classes (*rdfs:Class*) e suas propriedades, assim como uma hierarquia entre as classes (*rdfs:subClassOf*) e entre as propriedades (*rdfs:subPropertyOf*),

o domínio de classes aceitas pelo sujeito de uma propriedade (*rdfs:domain*) e o domínio de classes ou tipos de dados aceitos pelo objeto de uma propriedade (*rdfs:range*).

2.4 *Web Ontology Language* (OWL)

A linguagem *Web Ontology Language* (OWL) (HARMELEN; MCGUINNESS, 2004) estende o RDF-S, adicionando um vocabulário para permitir a descrição das classes e propriedades com maior enriquecimento semântico. Entre as principais funcionalidades oferecidas, a linguagem OWL permite descrever (i) relações de equivalência e disjunção entre as classes e suas instâncias; (ii) operações de interseção, união e complemento, análogas às da teoria de conjuntos, sobre as classes; (iii) a cardinalidade das propriedades de uma classe e (iv) características de simetria, reflexividade e transitividade entre as propriedades. Assim sendo, a OWL, que oferece três principais sublinguagens (OWL Lite, OWL DL e OWL Full) com expressividade crescente projetadas para uso de comunidades específicas de usuários e tipos de aplicação, tornou-se o padrão do W3C para descrição de ontologias (MENDONÇA, 2013).

2.5 *Linked Data*

Linked Data é um conjunto de melhores práticas para publicação e consumo de dados estruturados na Web, permitindo estabelecer ligações entre itens de diferentes conjuntos de dados para formar um único espaço de dados global (HEATH; BIZER, 2011a). Bizer, Heath e Berners-Lee (2009) resumem *Linked Data* como o uso da Web para criar ligações tipadas entre dados de diferentes conjuntos de dados.

Essa tecnologia permite muitas possibilidades para a integração de dados, pois lida com dados em escala global. Em geral, os dados no padrão de *Linked Data* são (BIZER; HEATH; BERNERS-LEE, 2009):

a) abertos – pode-se acessar *Linked Data* por meio de uma variedade ilimitada de aplicações e aplicativos, pois ela é expressa em formatos abertos, não proprietários;

b) modulares – os *Linked Data* podem ser combinados com quaisquer outros *Linked Data*. Nenhum planejamento prévio é necessário para integrar essas fontes de dados, já que ambas utilizam o mesmo padrão;

c) escaláveis – é fácil adicionar mais *Linked Data* aos já existentes, mesmo quando os termos e definições utilizados mudem ao longo do tempo.

Assim, a adoção dos princípios de *Linked Data* forneceu meios para concretizar a visão da Web Semântica e levou à criação da Web de Dados, um Grafo Gigante Global (GGG- *Giant Global Graph*) segundo Berners-Lee (2007), contendo bilhões de triplas RDF que representam e interligam dados provenientes de diversos domínios.

2.5.1 Princípios e Boas Práticas

Baseado nas tecnologias e boas práticas da Web Semântica, Berners-Lee (2006) estabeleceu quatro regras que ficaram conhecidas como os princípios de *Linked Data* que são listadas a seguir:

1. Usar URIs como nomes para os itens.
2. Usar URIs HTTP para que as pessoas possam consultar esses nomes.
3. Quando alguém consultar uma URI, prover informação RDF útil.
4. Incluir sentenças RDF interligando URIs, a fim de permitir que itens relacionados possam ser descobertos.

Essas regras fornecem a base para a publicação e interligação de dados estruturados na Web, além de serem os princípios fundamentais do termo *Linked Data*. Posteriormente, eles foram estendidos por documentos originados a partir das experiências da comunidade de *Linked Data* (BIZER; CYGANIAK; HEATH, 2007; SAUERMANN; CYGANIAK, 2008), resultando em boas práticas de publicação e consumo de *Linked Data*. Algumas dessas boas práticas relacionadas a *Linked Data* são tratadas na seção 2.7, após a exposição de alguns fundamentos necessários ao seu entendimento.

2.5.2 Padrões

Os padrões abertos adotados em *Linked Data* são amplamente difundidos e amplamente utilizados na Web, são eles: um mecanismo de identificação global e único (URIs - *Uniform Resource Identifiers*), um mecanismo de acesso universal (HTTP - *Hypertext Transfer Protocol*), o modelo de dados *Resource Description Framework* (RDF), e a linguagem de consulta SPARQL para acesso aos dados.

Berners-Lee, Fielding e Masinter (2005) definem *Uniform Resource Identifier* (URI) como uma sequência compacta de caracteres que identifica um recurso físico ou abstrato. Uma URI é uma sequência de caracteres semelhante a um *Uniform Resource Locator* (URL), que é utilizada, por exemplo, para acessar um documento HTML através de um navegador Web. Entretanto, uma URI, não necessariamente precisa especificar a localização do recurso identificado na rede e o mecanismo para recuperá-lo (MEALLING; DENENBERG, 2002).

Assim, uma URI desreferenciada resulta em uma descrição RDF do recurso identificado. Por exemplo, a URI `http://www.w3.org/People/Berners-Lee/card#i` identifica o pesquisador Tim Bernes-Lee (MAGALHÃES et al., 2011).

O protocolo de Transferência de Hipertexto (HTTP) é responsável pelo tratamento de pedidos e respostas entre cliente e servidor na Web. Ele surgiu da necessidade de distribuir informações pela Internet e, para que essa distribuição fosse possível, foi necessário criar uma forma padronizada de comunicação entre os clientes e os servidores da Web. Com isso, o protocolo HTTP passou a ser utilizado para a comunicação entre computadores na Internet e a especificar como seriam realizadas as transações entre clientes e servidores, através do uso de regras básicas.

Assim, o HTTP é um protocolo genérico, sem estado e no nível de aplicação para sistemas distribuídos, colaborativos e hipermídia. Uma característica do HTTP é a tipagem e negociação

de representação de dados, que permitem a construção de sistemas de forma independente dos dados transferidos (FIELDING et al., 1999).

Nesse contexto de *Linked Data*, o modelo de dados RDF traz uma série de benefícios, como a criação de URI-*links* (*links* RDF) entre dados de diferentes fontes de dados, os quais permitem que informações de diferentes fontes se mesclem naturalmente e permitem a representação de informações de diferentes esquemas em um modelo único.

Os *Links* RDF descrevem relacionamentos entre dois recursos (HEATH; BIZER, 2011a). Um *link* RDF consiste em uma tripla com três URIs. As URIs referentes ao sujeito e o objeto identificam os recursos relacionados. A URI referente ao predicado define o tipo de relacionamento entre os recursos. Uma distinção útil que pode ser feita é com relação a links internos e externos. *Links* RDF internos conectam recursos dentro de um único conjunto de dados. *Links* externos conectam recursos servidos por diferentes conjuntos de dados *Linked Data*. No caso de links externos, as URIs referentes ao sujeito e predicado do *link* pertencem a espaço de nomes (*namespaces*) distintos. *Links* externos são cruciais para a Web dos Dados visto que eles permitem interligar as fontes de dados dispersas em um espaço global de dados (MAGALHÃES, 2012).

Contudo, existem diversas maneiras de armazenar dados no padrão de *Linked Data*, por exemplo: usando RDF nativo por meio de APIs (*Application Programming Interface*) como Sesame [1], Jena (CARROLL et al, 2004), ou fornecendo *wrappers* para banco de dados relacionais, como D2R (BIZER e CYGANIAK, 2006), Virtuoso [2], também, pode-se previamente fazer a triplificação de fontes de dados, por exemplo, o Jena SDB [3], Jena TDB [4], entre outros [5]. Além disso, esses dados no padrão de *Linked Data* podem ser acessados através de consultas SPARQL submetidas aos SPARQL *endpoints*, que serão introduzidos na seção seguinte.

2.6 Linguagem, Protocolo e *Endpoint* SPARQL

SPARQL é uma linguagem de consulta declarativa, mas também um protocolo (CLARK; FEIGENBAUM; TORRES, 2008), recomendado pelo *World Wide Web Consortium* (W3C) para recuperar e manipular dados descritos em RDF, incluindo sentenças que envolvem RDF-S e OWL, através de chamadas *Simple Object Access Protocol*(SOAP) ou *Hypertext Transfer Protocol*(HTTP).

Assim como dados armazenados em banco de dados relacionais podem ser consultados e manipulados com a linguagem *Structured Query Language* (SQL), dados armazenados na forma de triplas podem ser consultados e manipulados com a linguagem SPARQL, por meio de SPARQL *Endpoints*, serviços *Representational State Transfer* (REST) que implementam o protocolo SPARQL. Logo, o intuito do protocolo SPARQL é promover a interoperabilidade, em que os clientes podem interagir com os dados SPARQL de forma consistente.

[1] http://www.openrdf.org/
[2] http://virtuoso.openlinksw.com/
[3] http://jena.hpl.hp.com/wiki/SDB
[4] http://www.openjena.org/TDB/
[5] https://www.w3.org/wiki/ConverterToRdf

Fontes *Linked Data* tipicamente fornecem um SPARQL *Endpoint* que é um serviço Web com suporte ao protocolo SPARQL. Ele tem sido grande aliado para o sucesso e disseminação da utilização da linguagem SPARQL, pois possui uma URI específica para receber requisições HTTP com consultas SPARQL, possibilitando por exemplo, a execução de consultas nos dados disponíveis no padrão de *Linked Data* na Web. Alguns exemplos são DBpedia [6], Data.Gov [7], DrugBank[8], dentre outros.

Com o advento da SPARQL 1.1 Prud'hommeaux e Buil-Aranda (2011), a utilização de SPARQL *endpoint* torna-se ainda mais geral, uma vez que é possível enviar instruções SPARQL de atualização (*insert*, *update*, *delete*), usar operação HTTP em uma maneira *RESTful*. Quanto às consultas, os resultados podem ter diferentes formatos. As consultas que usam os comandos SELECT e ASK geralmente são retornadas nos formatos XML, JSON ou texto plano. Já os resultados de consultas através dos comandos DESCRIBE ou CONSTRUCT normalmente usam os formatos RDF/XML, NTriples, Turtle ou N3. A maioria dos *endpoints* SPARQL exibem uma página HTML interativa que permite ao usuário digitar e submeter uma consulta (MAGALHÃES, 2012).

Além disso, prefixos e atalhos são bastante usados em consultas SPARQL e em alguns formatos de serialização de RDF para abreviar URIs. Assim se '`dc:`' é um prefixo para `<http://purl.org/dc/elements/1.1/>`, então '`dc:creator`' é uma notação abreviada da URI `<http://purl.org/dc/elements/1.1/creator>`. Um atalho frequentemente usado é a letra '`a`' ('um' em inglês) que serve para abreviar a URI '`rdf:type`', onde '`rdf:`' é um prefixo para `<http://www.w3.org/1999/02/22-rdf-syntax-ns#>`. A sintaxe '<>' é usada para identificar o próprio documento onde ela está inserida. O símbolo vírgula ('`,`') é usado para delimitar uma tripla e definir que o sujeito da próxima tripla não deverá ser explicitamente definido, pois será o mesmo sujeito da tripla definida antes do símbolo '`;`' (MAGALHÃES, 2012).

2.7 Boas Práticas

A adoção de boas práticas relacionadas a *Linked Data* facilita a descoberta de informações relevantes para a integração de dados entre diferentes fontes.

- **Selecionar URIs adequadas.** Devem-se evitar URIs contendo algum detalhe de implementação ou do ambiente em que estão publicadas. Como exemplo a evitar, consideremos o URI *http://lia.ufc.br:8080/ regispires/cgi-bin/resource.php?id=ufc* que possui detalhes da porta 8080 usada em seu ambiente de publicação e do script implementado em PHP necessário à sua execução.

  É frequente o uso de três URIs relacionadas a cada recurso: (i) um identificador para o recurso; (ii) um identificador para informações sobre o recurso para visualização através de navegadores HTML; (iii) um identificador para informações sobre o recurso em formato

[6] http://dbpedia.org/sparql
[7] http://services.data.gov.uk/finance/sparql
[8] http://wifo5-03.informatik.uni-mannheim.de/drugbank/sparql

RDF/XML. A Figura 4 representa um exemplo de três URIs relacionadas à pesquisadora Vânia Vidal na fonte DBLP. A representação de um recurso através de diferentes URIs permite que a interface *Linked Data* realize o dereferenciamento da URI de acordo com o tipo de conteúdo requisitado no cabeçalho HTTP (i.e. Text/HTML, application/rdf+xml, etc.).

```
http://dblp.l3s.de/d2r/resource/V%C3%A2nia_Maria_Ponte_Vidal
http://dblp.l3s.de/d2r/page/V%C3%A2nia_Maria_Ponte_Vidal
http://dblp.l3s.de/d2r/data/V%C3%A2nia_Maria_Ponte_Vidal
```

Figura 4 – Exemplos de URIs relacionadas a um mesmo recurso

A Figura 5 apresenta dois exemplos de requisições HTTP referente à URI da pesquisadora Vânia Vidal na fonte DBLP. No exemplo referente ao item (a), a requisição define como tipo MIME, dados no modelo RDF e recebe como resposta, através do redirecionamento 303, a URI referente aos dados da pesquisadora. No exemplo referente ao item (b), a requisição solicita os dados no formato HTML e recebe como resposta o redirecionamento para a URI referente à pagina HTML da pesquisadora.

```
(a)
$ curl -H "Accept: application/rdf+xml"
   http://dblp.l3s.de/d2r/resource/V%C3%A2nia_Maria_Ponte_Vidal

303 See Other: For a description of this item,
see http://dblp.l3s.de/d2r/data/V%C3%A2nia_Maria_Ponte_Vidal
```

```
(b)
$ curl -H "Accept: text/html"
   http://dblp.l3s.de/d2r/resource/V%C3%A2nia_Maria_Ponte_Vidal

303 See Other: For a description of this item,
see http://dblp.l3s.de/d2r/page/V%C3%A2nia_Maria_Ponte_Vidal
```

Figura 5 – Exemplos de requisições HTTP com tipos MIME RDF e HMTL

- **Usar URIs dereferenciáveis** para que a descrição do recurso possa ser obtida da Web.
- **Utilizar URIs estáveis.** A alteração de URIs quebra links já estabelecidos, criando um problema para a localização de recursos. Para evitar esse tipo de alteração, recomenda-se um planejamento meticuloso das URIs que serão usadas e também que o responsável pela publicação detenha a propriedade do espaço de nomes.

- **Criar links para outras fontes de dados** de modo a permitir a navegação entre as fontes de dados. Os *links* podem ser criados de forma manual ou automatizada.

- **Publicação de Metadados.** Análise dos metadados facilita a seleção dos dados relevantes. Devem ser fornecidos metadados sobre proveniência e licenciamento dos dados. Também é recomendável a disponibilização de metadados sobre a fonte de dados. O vocabulário mais usado atualmente para publicação de metadados sobre conjunto de dados disponíveis é o *VoiD – Vocabulary of Interlinked Datasets*.

- **Utilizar termos de vocabulários amplamente usados**. Embora não haja restrições para seleção de vocabulários, é considerada uma boa prática o reuso de termos de vocabulários RDF amplamente usados para facilitar o processamento de *Linked Data* pelas aplicações clientes (BIZER; CYGANIAK; HEATH, 2007). Novos termos só devem ser definidos se não forem encontrados em vocabulários já existentes. A seguir apresentamos alguns vocabulários bastante difundidos: *Friend-of-a-Friend* (FOAF), *Semantically-Interlinked Online Communities* (SIOC), *Simple Knowledge Organization System* (SKOS), *Description of a Project* (DOAP), *Creative Commons* (CC) e *Dublin Core* (DC). Uma relação mais extensa desses vocabulários é mantida pelo projeto *Linking Open Data* no *ESW Wiki*[9].

- **Estabelecer relações entre os termos de vocabulários proprietários para termos de outros vocabulários.** Isso pode ser feito através do uso das propriedades *owl:equivalentClass*, *owl:equivalentProperty*, *rdfs:subClassOf*, *rdfs:subPropertyOf*. A Figura 6 mostra que a classe Pessoa de um vocabulário local é equivalente à definição da classe *Person* no vocabulário da *DBpedia*. A definição de relações entre vocabulários facilita a integração de dados que utilizam esses vocabulários.

```
<http://lia.ufc.br/Pessoa> owl:equivalentClass <http://dbpedia.org/ontology/Person> .
```

Figura 6 – Relação de equivalência entre termo proprietário e termo da DBpedia

- **Explicitar formas de acesso adicional aos dados** como *endpoints SPARQL* e *RDF dumps*.

2.8 *Linked Open Data*

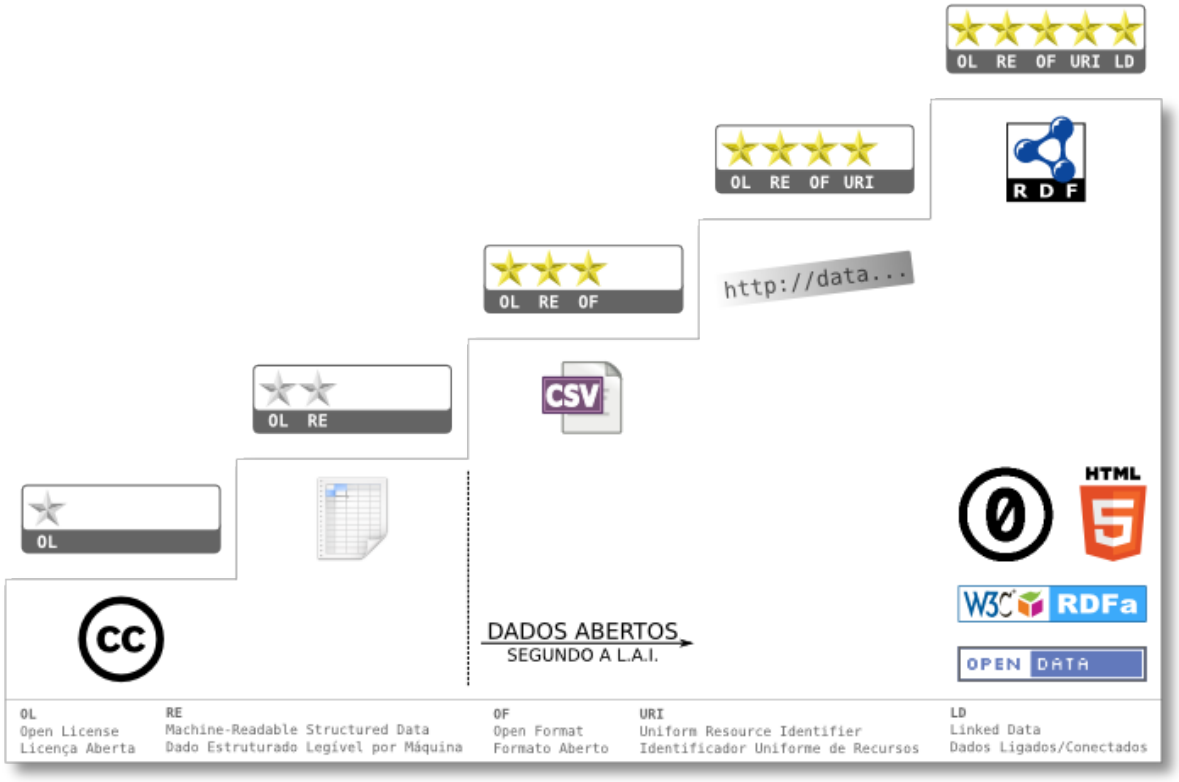
O conceito *Linked Data* pode ser especializado para *Linked Open Data* (LOD) (BAUER; KALTENBÖCK, 2011), quando está vinculado com o conceito de Dados Abertos, ou seja, "dados que podem ser usados, reutilizados e redistribuídos livremente, sujeitos, no máximo, à exigência de atribuição e compartilhamento pela mesma licença".

Nesse contexto, em 2010, na exposição Gov 2.0, em Washington - DC, EUA, Tim Berners-Lee anunciou o modelo de 5 estrelas para para publicação de dados abertos, no qual se

[9] http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/CommonVocabularies

agrega valor, por meio de associação de um contexto, no qual ele é útil, mas inicial, por que a sociedade é livre para montar, remontar, agrupar e dar novos contextos gerando inovação e mais conhecimentos. As 5 estrelas do modelo são visualizadas na Figura 7 e descritas abaixo:

Figura 7 – As 5 Estrelas dos Dados Abertos



Fonte: http://5stardata.info/pt-BR/.

- Torne seus recursos disponíveis na Web (tanto faz o formato) sob uma licença aberta.
- Torne seus recursos disponíveis como dados estruturados (Por exemplo um excel no lugar de imagem escaneada).
- Utilize formatos não-proprietários (ex. CSV e não excel).
- Utilize URIs para identificar recursos. Isso vai ajudar as pessoas a apontarem para eles.
- Conecte seus dados com dados de outras pessoas para prover contexto (dados ligados).

Com isso, surgiram nos últimos anos, inúmeras iniciativas voltadas para fomentar a criação da Web de Dados, como por exemplo, o projeto *Linking Open Data* [10] que é um esforço comunitário iniciado em 2007 e suportado pelo W3C para identificar fontes de dados publicadas sob licenças abertas, convertê-las para RDF e publicá-las na Web usando os princípios de *Linked*

[10] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Data (BIZER; JENTZSCH; CYGANIAK, 2011). A Figura 8 mostra um diagrama de nuvem[11] com as fontes de dados publicadas pelo projeto LOD e as interligações entre elas em até agosto de 2014. O tamanho dos círculos corresponde ao número de triplas de cada fonte de dados. As setas indicam a existência de pelo menos 50 links entre duas fontes. A origem de uma seta indica a fonte que possui o link e a fonte referenciada é a fonte para a qual a seta está apontando. Setas bidirecionais representam fontes que se referenciam mutuamente. A espessura da seta corresponde ao número de links.

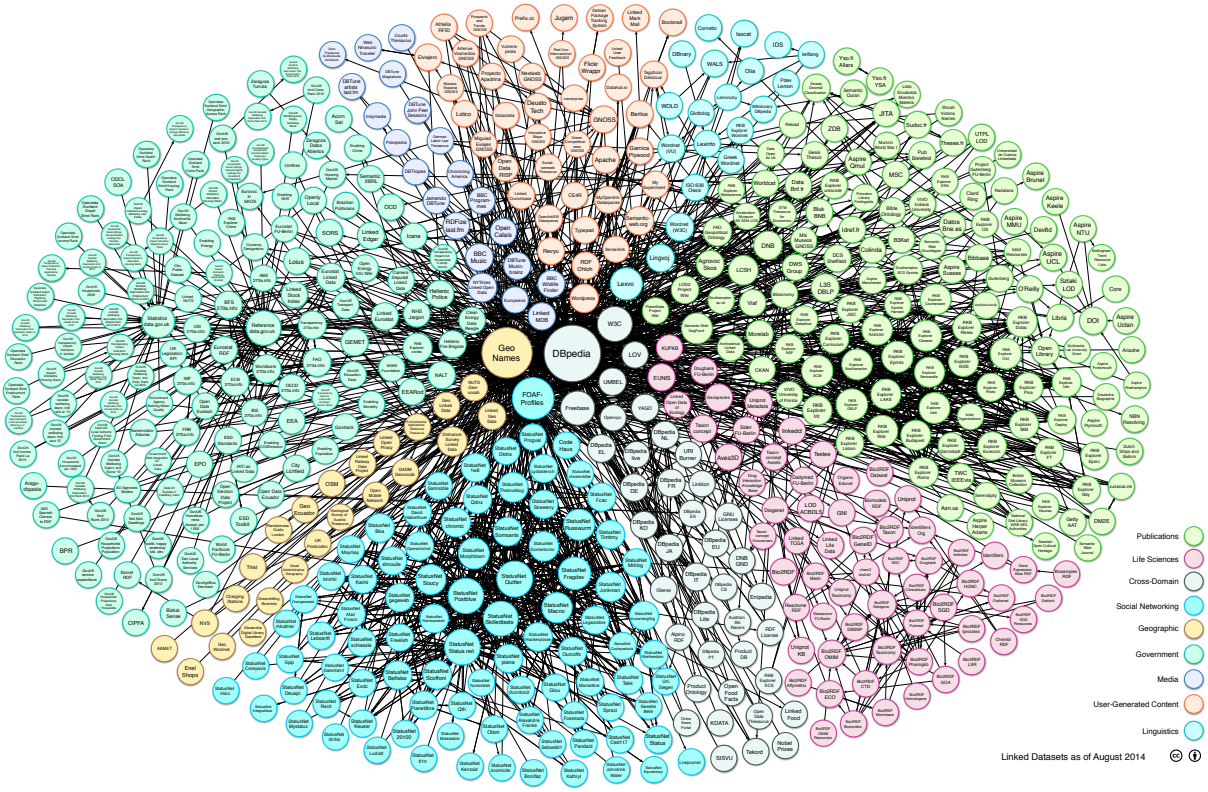


Figura 8 – Diagrama de nuvem *Linking Open Data*, por Richard Cyganiak e Anja Jentzsch. Atualizado em 30/08/2014.

Outra importante iniciativa no âmbito governamental foi a criada em 2011 no Brasil, a Infraestrutura Nacional de Dados Abertos (INDA) [12], para aplicar os princípios de *Linked Data* na publicação de dados governamentais abertos.

2.9 Integração de Dados

O problema de integração de dados tem sido tema de pesquisa na área de banco de dados há bastante tempo (ZIEGLER; DITTRICH, 2004) sendo reconhecido como amplo e de grande relevância.

Para Langegger (2010), integração de dados é o processo de permitir acesso transparente à múltiplos sistemas de informação heterogêneos e distribuídos.

11 http://lod-cloud.net/
12 http://wiki.gtinda.ibge.gov.br/

O principal objetivo dos sistemas de integração de dados é permitir que usuários consultem simultaneamente múltiplas fontes de dados heterogêneas, distribuídas e autônomas por meio de uma única interface de consultas, mantendo transparentes os procedimentos de acesso, extração e integração dos dados. Assim o sistema de integração de dados deve tratar de forma transparente problemas de heterogeneidade (estrutural, conceitual e tecnológica), distribuição e autonomia das fontes durante a execução de consultas.

De acordo com Özsu e Valduriez (1999), os principais desafios para a integração de dados são justamente esses três aspectos ortogonais: heterogeneidade, distribuição e autonomia, descritos a seguir.

- **Distribuição.** As fontes de dados estão dispersas geograficamente, sendo interligadas por meio de uma rede de computadores. Como consequência da distribuição, é necessário lidar com os problemas envolvidos nas redes, como replicação, fragmentação e custo da transmissão dos dados e a capacidade de processamento de cada servidor.

- **Autonomia.** Refere-se ao nível de independência de operação de cada fonte de dados que participe de um sistema de integração, em que as fontes possuem controle total sobre os dados e, geralmente, não podem afetar suas funcionalidades e requerer modificações.

- **Heterogeneidade.** Como os esquemas das fontes são desenvolvidos independentemente, eles possuem estruturas e terminologias diferentes (heterogeneidade estrutural e semântica), o que ocorre tanto com os esquemas que vêm de domínios diferentes, quanto com os modelados no mesmo domínio do mundo real, pelo fato de serem desenvolvidos por pessoas diferentes, em diferentes contextos. Para serem efetivos, os sistemas de integração de dados devem ser capazes de transformar dados de diferentes fontes para responder a consultas feitas sobre esse esquema.

Contudo, o processo de integração, ou seja, interligação dos dados remete a dois problemas principais: o problema de encontrar termos correspondentes nos diferentes vocabulários utilizados e o empecilho de encontrar entidades correspondentes presentes em *datasets* distintos. Esses problemas são resolvidos com atividades de mapeamento descritas na seção subsequente.

2.10 Mapeamentos

Mapeamentos consistem em encontrar termos correspondentes nos diferentes vocabulários utilizados, e atividades de resolução de entidades, que consistem em encontrar e interligar entidades similares presentes em *datasets* distintos. Para isso, técnicas de mapeamento e alinhamento de ontologias (EUZENAT; MOCAN; SCHARFFE, 2008) são fortemente utilizadas, com emprego de um grande conjunto de parâmetros e produção de uma grande quantidade de resultados (MENDONÇA, 2013).

Além disso, seja qual for a abordagem adotada, é fundamental para os sistemas de integração que se estabeleça como os elementos do esquema de mediadores ou final estão relacionados com os elementos dos esquemas das fontes de dados. Portanto, são definidos

os mapeamentos entre o esquema de mediação e os esquemas das fontes de dados, as quais possibilitam tanto materializar o esquema de mediação ou final (abordagem materializada) quanto reescrever consultas sobre o esquema de mediação ou final em subconsultas sobre as fontes locais (abordagem virtual) (PINHEIRO, 2011).

2.11 Abordagem ETL

A natureza e a variedade das fontes de dados são os fatores principais que devem ser considerados para a escolha da estratégia mais apropriada de publicação de *Linked Data*(HEATH; BIZER, 2011a).

No processo de publicar dados no modelo RDF para serem disponibilizadas no contexto de *Linked Data*, as atividades de extração dos dados de múltiplas fontes heterogêneas, seguida das atividades de limpeza, consolidação, agregação e integração dos mesmos, podem ser orquestradas por uma abordagem ETL (CORDEIRO; CAMPOS; BORGES, 2011).

A abordagem que utiliza um *workflow* ETL para publicar *Linked Data* herda o potencial oferecido pelas técnicas e ferramentas de ETL, que foram desenvolvidas e refinadas durante anos, em desafiantes cenários de Inteligência de Negócio (Business Intelligence).

Logo, os benefícios imediatos da abordagem ETL são (i) a sistematização do processo de publicação de *Linked Data*; (ii) o monitoramento e gerenciamento de suas atividades de extração, transformação e carga e (iii) a oportunidade de reutilização do *workflow* para carregar novos dados e atualizar os dados previamente publicados no contexto de *Linked Data*.

Com relação às ferramentas de ETL, existe um conjunto significativo de ferramentas como: Oracle Data Integrator (ODI)[13], Talend Studio for Data Integration[14] e Pentaho Data Integration[15] (também conhecida como Kettle).

O Pentaho Data Integration será utilizado no processo de publicação de dados abordado neste trabalho, devido suas características de ser *open source* e possuir uma ampla comunidade de usuários.

2.11.1 Pentaho Data Integration (Kettle)

A ferramenta Pentaho Data Integration, também chamada de Kettle [16], é a ferramenta de *workflow* Extração, Transformação e Carga (ETL - Extract, Transform and Load)ETL integrante do conjunto de ferramentas da plataforma de BI Pentaho (CASTERS; BOUMAN; DONGEN, 2010). Além da característica comum das ferramentas de ETL, que é apoiar o processo de ETL em projetos de *Data Warehousing* (DW), o Kettle também pode ser utilizado para outros propósitos como, por exemplo, (i) migração de dados entre aplicações ou banco de dados; (ii) exportação e importação de dados entre diferentes tipos de repositórios; (iii) carga massiva de dados em um repositório de dados; (iv) limpeza de dados em um repositório e (v) integração entre aplicações.

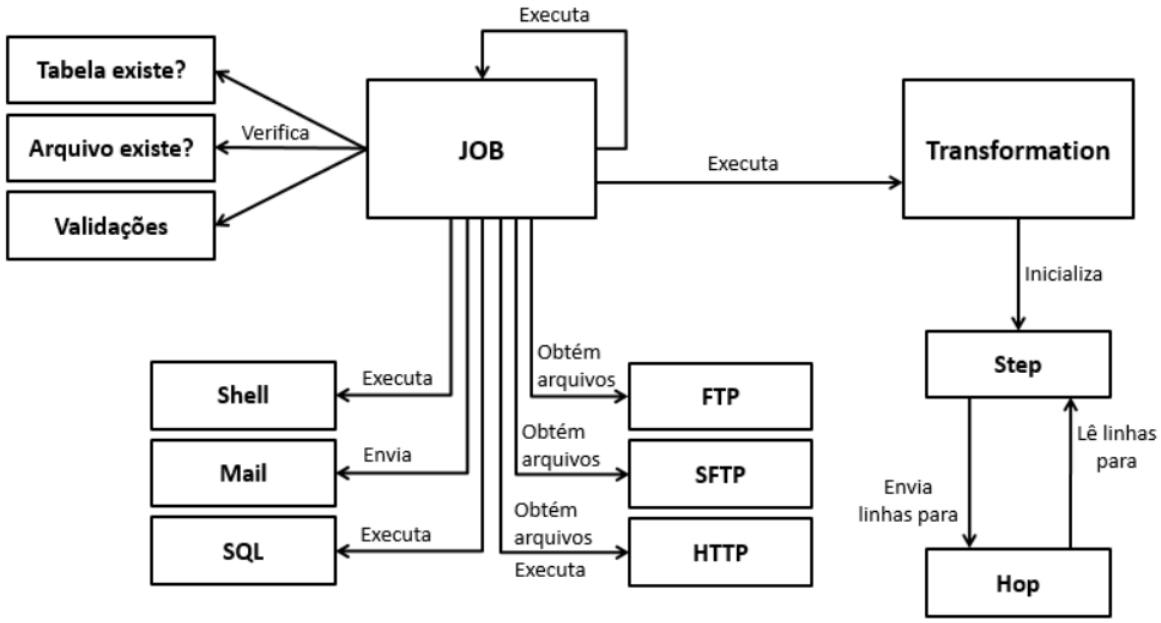
---

13 http://www.oracle.com/technetwork/middleware/data-integrator/overview/index.html
14 http://www.talend.com/products/talend-open-studio
15 http://kettle.pentaho.com
16 http://kettle.pentaho.com

Entre os princípios sobre os quais o Kettle foi desenvolvido, encontram-se (i) a facilidade de instalação; (ii) a facilidade de desenvolvimento do *workflow* ETL; (iii) a disponibilização de interface gráfica intuitiva em detrimento da necessidade de programação por linhas de código; (iv) a extensibilidade das funcionalidades da ferramenta através de uma API e (v) a forte orientação e transparência na disponibilização de metadados sobre o *workflow* ETL. Sendo assim, as composições dos *jobs* e das *transformações* podem ser armazenadas em tabelas de um repositório de um tipo banco de dados ou em arquivos XML.

No Kettle, o *workflow* ETL pode ser especificado através de dois tipos de componentes, denominados *transformations* e *jobs*. Um *transformation* consiste de um conjunto de passos conectados, onde cada passo, denominado *step*, é responsável por uma atividade de extração, transformação ou carga de dados. A conexão entre dois *steps* de um *transformation*, denominada *transformation hop*, permite que os dados fluam em um único sentido e de maneira assíncrona. Um *job* também consiste de um conjunto de passos conectados. No entanto, os passos de um *job*, denominados *job entries*, são responsáveis por executar um *transformation*, outro job ou atividades auxiliares como manipular e transferir arquivos, enviar e receber *emails* e executar uma série de validações. A conexão entre dois *job entries*, denominada *job hop*, determina a ordem de execução dos passos do *job*, que, diferente dos passos do *transformation*, são executados de maneira síncrona. A Figura 9 ilustra o modelo conceitual do Kettle.Tanto um *transformation*, quanto um *job* podem possuir notas para documentar informações sobre o *workflow* especificado por eles.

Figura 9 – Modelo conceitual da ferramenta Pentaho Data Integration (Kettle).



Fonte: Mendonça (2013).

Além disso, o Kettle oferece ainda um conjunto de aplicativos para apoiar o desenvolvi-

mento e a execução dos *workflows* ETL. Este conjunto inclui as seguintes ferramentas:

- *Spoon*: ferramenta de interface gráfica para desenvolvimento e execução do fluxo de dados, desde sua entrada até a saída, através da criação e execução de *jobs* e *tranformations*;
- *Pan*: ferramenta de linha de comando para execução dos *transformations* desenvolvidos no *Spoon*. Normalmente é utilizada para agendamento da execução de *transformations*;
- *Kitchen*: ferramenta de linha de comando para execução dos *jobs* desenvolvidos no Spoon. Normalmente é utilizada para agendamento da execução dos *jobs*;
- *Carte*: servidor web que possibilita o agrupamento em *cluster* do processo de ETL, através da execução remota de *transformations* e *jobs*.

2.11.2 ETL4LOD - Componentes do Kettle relacionados a Linked Data

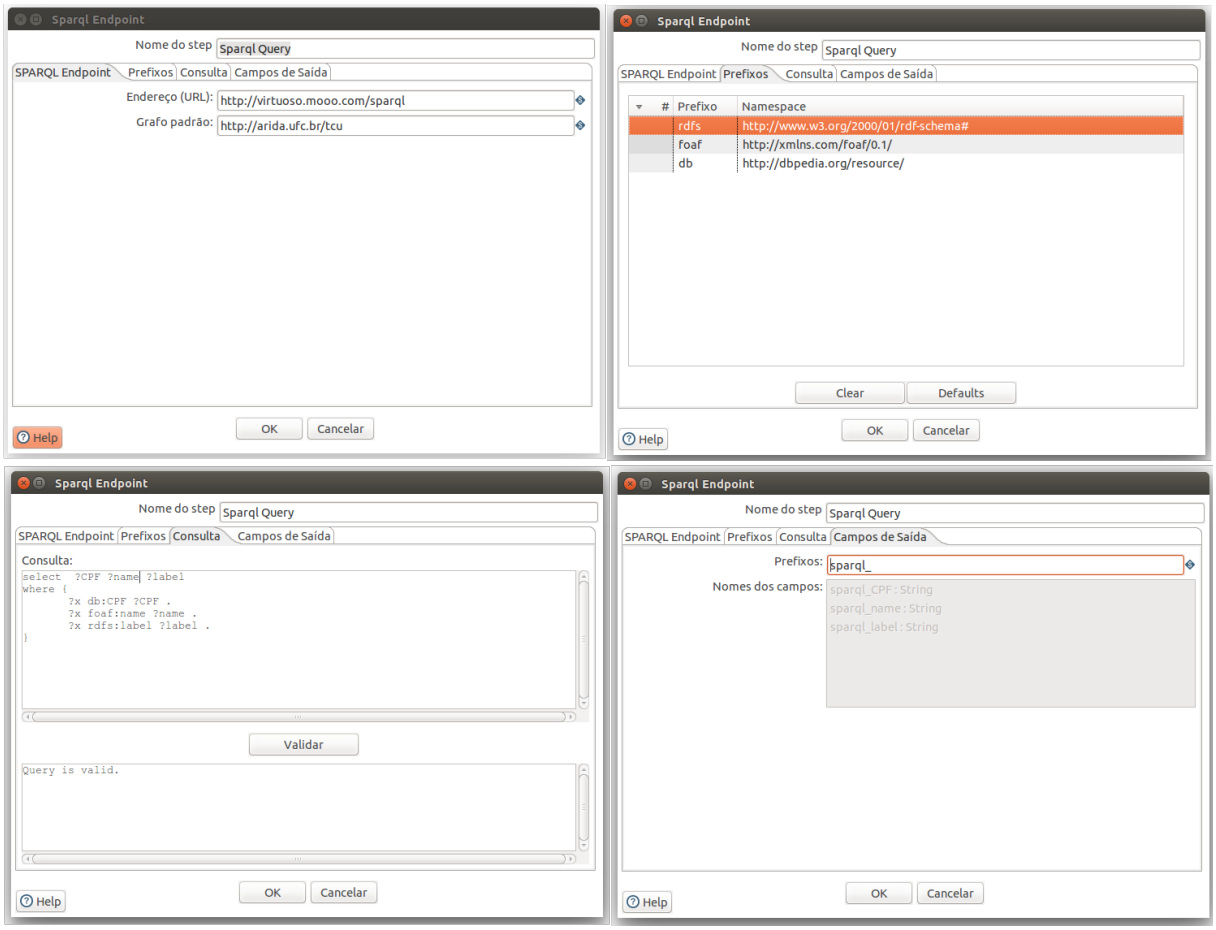
A ferramenta Pentaho Data Integration (Kettle) não oferece componentes específicos para o contexto de *Linked Data*. Para suprir esta carência do Kettle, o projeto LinkedDataBR (CAMPOS; GUIZZARDI, 2010) desenvolveu e disponibilizou 4 novos *steps* utilizando a API Java do Kettle: Sparql *Endpoint*, Sparql *Update Output*, *Data Property Mapping* e *Object Property Mapping*. Este conjunto de *steps*, denominado ETL4LOD, executa atividades relevantes com as tecnologias relacionadas com *Linked Data*, que possibilitam a publicação como triplas RDF tanto dos dados de domínio, quanto dos dados de proveniência (MENDONÇA, 2013).

O *step* Sparql *Endpoint* oferece a capacidade de extrair dados de um SPARQL *Endpoint*, a partir da especificação da URL relacionada ao SPARQL *Endpoint* e da definição de uma consulta SPARQL. Em conjunto com a API do Kettle, a API Jena[17] foi intensamente utilizada para implementação das operações relacionadas com RDF e SPARQL. A Figura 10 ilustra a interface de configuração do *step* Sparql *Endpoint*, com a especificação dos parâmetros necessários para extrair do *dataset* do Grupo de Pesquisa *Advanced Research in Database*[18](ARiDA), as propriedades Cadastro de Pessoas Físicas (CPF), nome e *label* de pessoas inabilitadas para função pública.

O *step* Sparql *Update Output* oferece a capacidade de carregar triplas RDF em um banco de triplas. Para isso, é necessário informar a URL do SPARQL *Endpoint* que provê suporte a operações de atualização, disponibilizadas na versão 1.1 do SPARQL; o usuário e a senha de autenticação no banco de triplas; a URI do grafo onde se deseja carregar as triplas RDF e o campo que contém a tripla RDF no formato NTriple. A Figura 11 ilustra a interface de configuração do *step* Sparql *Update Output*, com a especificação dos parâmetros necessários para inserir triplas RDF relacionadas a organizações inabilitadas para função pública e licitantes inidôneas, disponibilizados pelo Tribunal de Contas da União (TCU), no banco de triplas http://virtuoso.mooo.com, especificamente no grafo http://arida.ufc.br/tcu.

[17] http://jena.apache.org
[18] http://www.arida.ufc.br/site/

Figura 10 – Interface de configuração do *step* Sparql Endpoint.

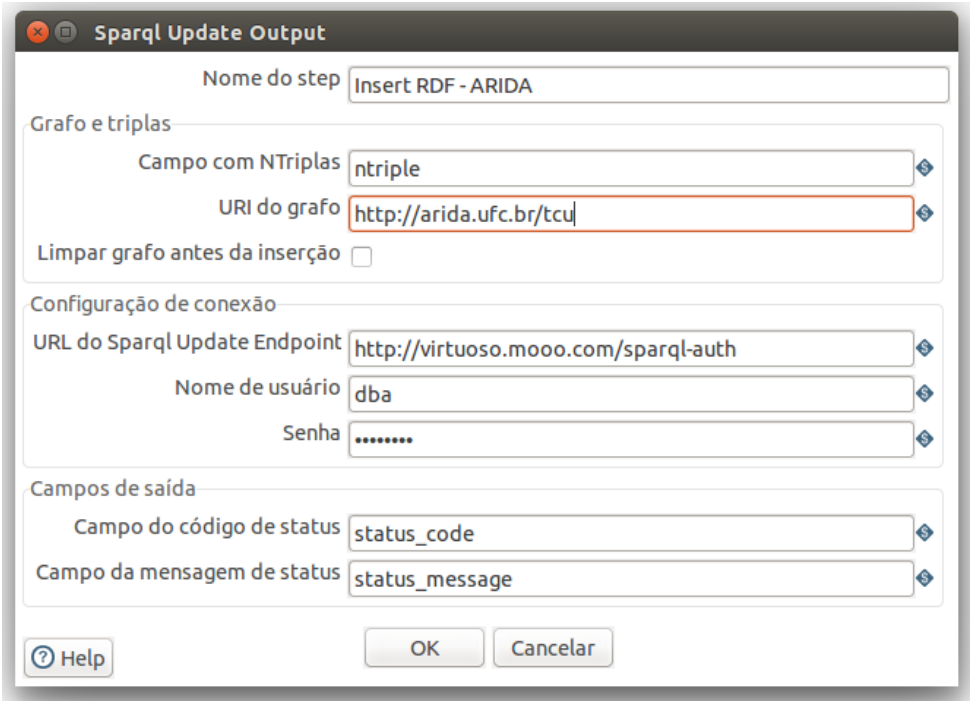


O *step Data Property Mapping* representado pela Figura 12 oferece a capacidade de mapear, a partir das linhas do fluxo de entrada, os componentes de uma tripla RDF (sujeito, predicado e objeto) nas linhas do fluxo de saída, sendo o objeto um valor literal. Cada linha de entrada deve conter um campo com a URI que identifica o recurso do sujeito. Na configuração do *step*, é necessário definir uma ou mais URIs especificando o tipo do recurso e um mapeamento de propriedades do tipo literal, informando a URI da propriedade e o campo da linha de entrada que contém o valor da propriedade. Além disso, é possível definir, no mapeamento, o tipo do literal e a marcação de idioma.

O *step Object Property Mapping* que é representado pela Figura 13, é similar ao *step Data Property Mapping*, com a diferença de que o valor do objeto enviado no fluxo de saída é uma URI de um recurso. A configuração deste step é mais simples e consiste em indicar o campo da linha de entrada que contem a URI do sujeito, o campo da linha de entrada que contem a URI do objeto e definir a URI da propriedade.

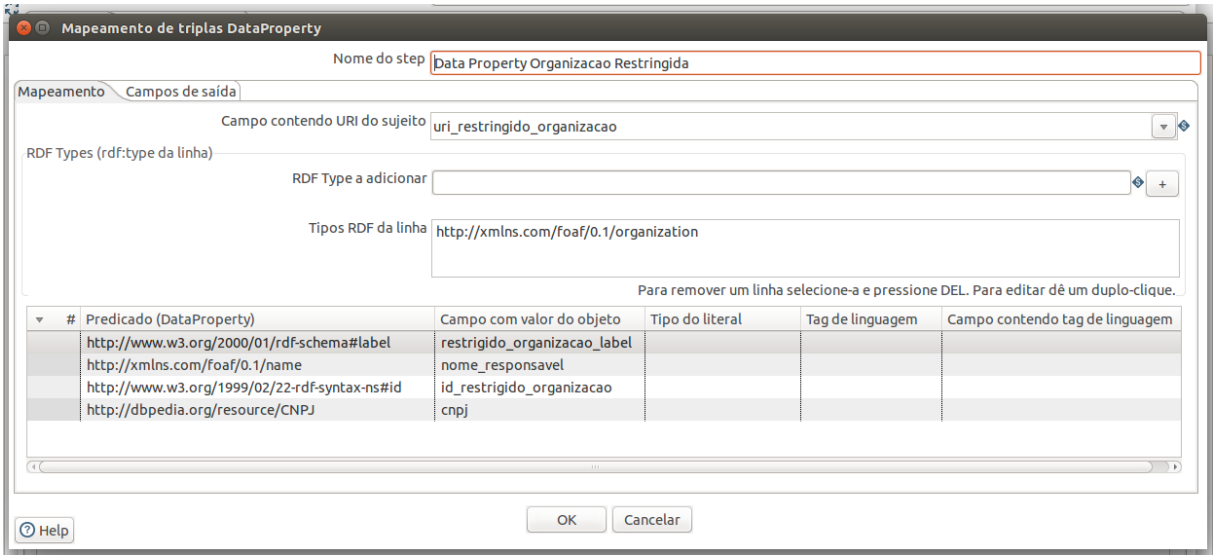
A Figura 12 ilustra a interface de configuração do *step Data Property Mapping*, com a especificação do tipo e do mapeamento das propriedades *label*, *name*, identificador (id) e Cadastro Nacional de Pessoas Jurídicas (CNPJ) do recurso Organização Restringida para Organizações inabilitadas para função pública e licitantes inidôneas e a Figura 13 ilustra a

Figura 11 – Interface de configuração do *step* Sparql *Update Output*.



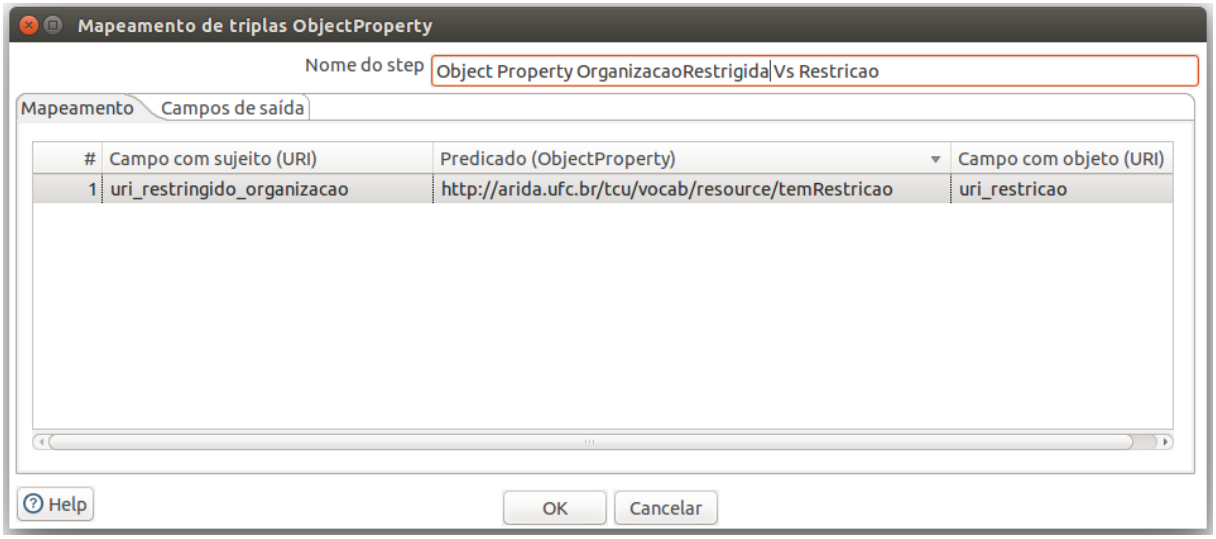
interface de configuração do *step Object Property Mapping*, com a especificação do mapeamento da propriedade Organização restringida e a restrição.

Figura 12 – Interface de configuração do *step Data Property Mapping*.



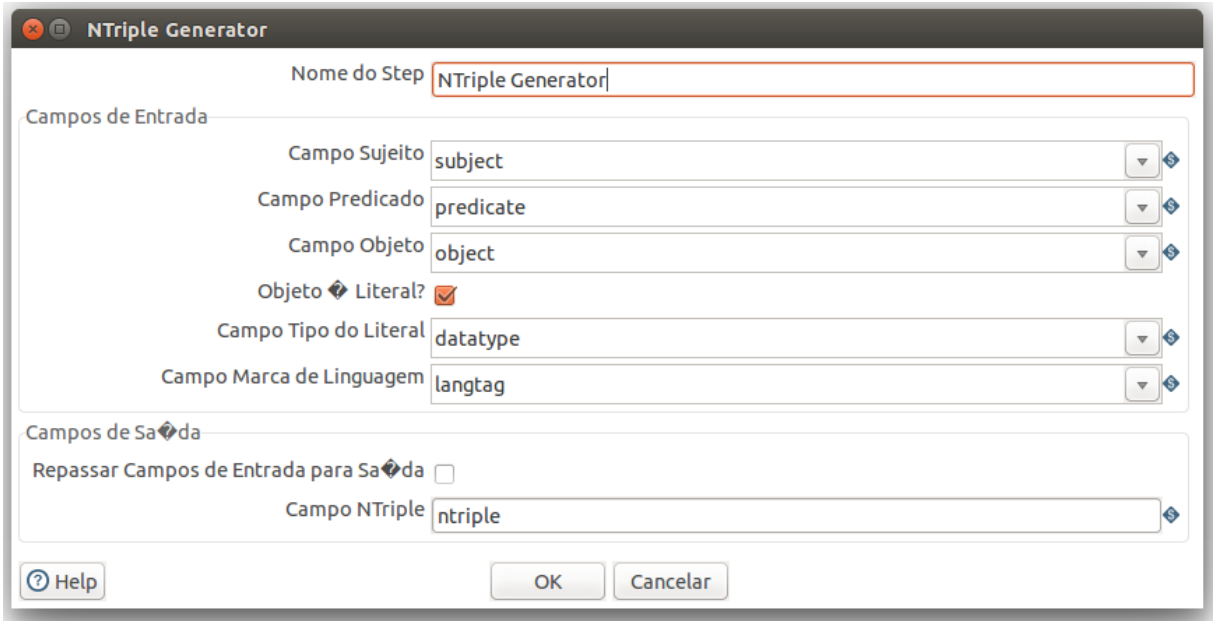
Além disso, Mendonça (2013) estendeu os steps ETL4LOD para possibilitar o armazenamento dos metadados de composição dos mesmos também em um repositório Kettle do tipo banco de dados. Portanto, o trabalho de Mendonça (2013) desenvolveu o *step* NTriple *Generator*, que pode ser utilizado juntamente com os 4 *steps* do ETL4LOD, desenvolvido no projeto LinkedDataBR (CAMPOS; GUIZZARDI, 2010).

Figura 13 – Interface de configuração do *step* Data Object Mapping.



O *step NTriple Generator* oferece a facilidade de geração de sentenças RDF no formato NTriple. A configuração consiste em informar cada campo da linha de entrada relacionada aos respectivos componentes da tripla RDF, ou seja, sujeito, predicado e objeto. Se o objeto for literal, ainda é possível informar os campos relacionados com o tipo do literal e a marca de idioma. Assim sendo, o *step NTriple Generator* torna-se útil, por exemplo, para receber as linhas de dados enviadas por um *step Data Property Mapping* ou *Object Property Mapping* e gerar, no fluxo de saída, as linhas de dados com as triplas RDF a serem inseridas em um banco de triplas, por meio do *step* Sparql *Update Output*. A Figura 14 ilustra a interface de configuração do *step NTriple Generator*.

Figura 14 – Interface de configuração do *step* NTriple Generator.



2.12 Considerações do Capítulo

Este capítulo apresentou um panorama dos principais padrões utilizados na Web Semântica que possibilitam a disseminação dos dados em *Linked Data*. Os padrões incluem RDF/RDFS, OWL, Linguagem, protocolo e *endpoints* SPARQL, cujos conceitos e definições foram introduzidos no decorrer do capítulo, servindo de fundamentação para o entendimento dos demais capítulos. Finalmente, foi abordada a integração de dados e a utilização de uma ferramenta ETL para o processo de publicação de dados na Web com o modelo RDF.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta o estudo de alguns trabalhos da área que foram usados como base para criação deste trabalho e uma breve descrição de cada, acompanhada com pontos comuns e distintos em relação a este. Pretende-se com esse estudo, obter subsídios e contribuições relevantes para o desenvolvimento deste trabalho.

Além disso, será apresentada uma tabela que simplifica a comparação entre os trabalhos, levando em consideração as etapas do processo de publicação de dados na Web usando o padrão *Linked Data*.

3.1 OpenSBBD: Usando Linked Data para Publicação de Dados Abertos sobre o SBBD

Batista e Lóscio (2013) apresentam os passos realizados para divulgar os dados abertos do Simpósio Brasileiro de Banco de Dados (SBBD) seguindo o padrão *Linked Data*. Esse trabalho é essencialmente dividido em três fases que são: (i) A criação de um conjunto de dados seguindo os princípios de *Linked Data* com alto potencial para ser reutilizado por futuros trabalhos, em conjunto com a criação de uma ontologia, denominada SBC*Event*, com novos termos sobre o domínio de conferências e com reuso de vocabulários existentes; (ii) Disponibilização de um SPARQL Endpoint para a realização de consultas SPARQL sobre o conjunto de dados criado e iii) Criação de visualizações dos dados do SBBD em formatos amigáveis, como gráficos e tabelas.

O trabalho de Batista e Lóscio (2013) é análogo a este, pois apresenta um modo utilizado para publicar dados no modelo RDF, seguindo os princípios de *Linked Data*. Entretanto, a forma utilizada para publicar os dados é específica e centrada na ferramenta D2R-Server que solicita como pré-requisito a existência de uma base relacional para utilizar o SPARQL *Endpoint*, possibilitando que os mapeamentos sejam feitos em tempo real.

Diferentemente da proposta deste trabalho, que visa a definição de um processo genérico para publicação dos dados seguindo os padrões de *Linked Data*, de modo que, facilite a integração com diversas fontes de dados, além do uso de uma ferramenta capaz de converter para o modelo RDF fontes de dados em formatos diversos.

3.2 An approach for managing and semantically enriching the publication of Linked Open Governmental Data

Cordeiro et al. (2011) expõe uma plataforma, na qual inclui uma ferramenta de Extração, Transformação e Carga (ETL) para gerenciar o processo de publicação e mapeamento dos dados brutos em modelo RDF, ligado a um repositório para gerenciar e armazenar os dados em formato RDF. Além disso, o trabalho visa apoiar o processo de publicação de dados vinculados de modo que possui como foco o enriquecimento semântico.

O trabalho de Cordeiro et al. (2011) apresentado assemelha-se a com este no sentido de facilitar o mapeamento dos dados brutos em formatos diversos, além de, facilitar e apoiar

publicação no formato RDF.

Apesar de Cordeiro et al. (2011) facilitar o mapeamento e a publicação, não se trata de um processo, mas uma plataforma. Além disso, o trabalho apresentado possui foco restrito na publicação de dados abertos governamentais. Este presente trabalho trata de um processo que servirá para um usuário publicar dados no formato RDF seguindo os padrões de *Linked Data*.

3.3 StdTrip: An a priori design approach and process for publishing Open Government Data

Salas et al. (2010) apresenta uma abordagem de design e processo para para publicação de dados abertos do Governo que visa guiar usuários durante a modelagem conceitual do processo de triplicação, que pode ser definido como mapeamento do modelo relacional para RDF. Além disso, o *StdTrip* promove a reutilização de padrão *World Wide Web Consortium* (W3C), recomendando vocabulários RDF em primeiro lugar e, se não for possível, sugere a reutilização de outros vocabulários já empregados por outros conjuntos de dados RDF na Web, para facilitar a integração com outros conjuntos de dados.

O trabalho de Salas et al. (2010) se assemelha a este, pois é um processo de publicação de dados no modelo RDF, seguindo os padrões de *Linked Data*.

No entanto, o trabalho Salas et al. (2010) apresenta um processo que enfatiza especificamente o mapeamento do modelo relacional para RDF.

Contudo, este trabalho apresenta um processo de triplicação ressaltará o mapeamento e a utilização dos principais modelos dos dados, entre eles: *JavaScript Objetc Notation* (JSON), *Extensible Markup Language* (XML), *Comma-Separated Values* (CSV) e o modelo relacional.

3.4 Uma abordagem para coleta e publicação de dados de proveniência no contexto de Linked Data

Mendonça (2013) propõe a abordagem *ETL4LinkedProv*, uma abordagem de proveniência para o processo de publicação de *Linked Data* realizado dentro dos limites de uma organização. A abordagem utiliza um agente de proveniência que é executado através de um workflow de Extração, Transformação e Carga (ETL). O trabalho ainda detalha a implementação da abordagem *ETL4LinkedProv* e para demonstrar as contribuições foi realizado um exemplo de publicação e integração envolvendo dados reais do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Rede Nacional de Ensino e Pesquisa (RNP), organizações de incentivo à pesquisa no Brasil.

O trabalho de Mendonça (2013) se assemelha a este, no tocante de um modelo real de integração e publicação de dados no modelo RDF.

Todavia, esse trabalho está focado em um contexto genérico como manual de boas práticas para publicação no modelo RDF, enquanto Mendonça (2013) enfatiza a proveniência, além disso, detalha a implementação de uma abordagem e aplicação desta em um exemplo específico.

3.5 Comparação entre trabalhos relacionados

No Quadro 1 apresentamos uma comparação, que visa facilitar e mostrar algumas das principais diferenças entre os trabalhos citados acima.

Quadro 1 – Comparação entre os principais trabalhos citados anteriormente

Abordagem	**Batista e Lóscio(2013)**	**Cordeiro et al. (2011)**	**Salas et al. (2010)**	**Mendonça (2013)**	**Este trabalho**
Modelo de Dados Origem	Relacional	csv, xml, json, relacional	Relacional	csv e xml	csv, xml, json, relacional
Ferramentas	D2RQ, Protégé	ETL	Não apresentou	Protégé, ETL	Protégé, ETL
Integração/ interligação de dados	Não	Não	Não apresentou	Sim	Sim
Dados abertos	Sim	Sim	Não	Sim	Sim
Mapeamento/ Reuso de vocabulário	Sim	Não apresentou	Sim (Foco principal)	Sim	Sim
Esquema e modelo de dados final	Ontologia, RDF	Ontologia, RDF	Não apresentou	Ontologia, RDF	Ontologia, RDF
Padrões e Boas prática de Liked Data(W3C)	Sim	Sim	Sim	Sim	Sim
Mecanismo de acesso dos dados em RDF	Dumps RDF, D2R-Server e SPARQL Endpoint	Não apresentou	Não apresentou	Dumps RDF e SPARQL Endpoint (Virtuoso)	Dumps RDF e SPARQL Endpoint (Virtuoso)

3.6 Considerações do capítulo

Este capítulo representou o estado da arte, através do estudo de alguns trabalhos de publicação de dados na Web que seguem o padrão *Linked Data*, destacando suas vantagens e desvantagens. Esse estudo conclui que trabalhos de publicação de dados na Web, ainda estão em estágio de amadurecimento e há espaço para melhorias.

Logo, o trabalho proposto foi desenvolvido considerando o processo de publicar dados na Web, desde obtenção dos dados até sua publicação de maneira adequada, seguindo o padrão de *Linked Data* . Os capítulos seguintes discutem em detalhes o processo para publicação de dados na Web seguindo os princípios de *Linked Data* e em seguida um estudo de caso demonstrando o uso do processo.

4 PROCESSO

Passado o período inicial de grande entusiamo pela publicação de novas bases de dados em *Linked Data*, a comunidade científica resolve agora passar à avaliação das bases quanto à sua qualidade. E alguns problemas têm se apresentado frequentemente, são eles: falta de um processo que incentive, guie e contribua com o aprimoramento da publicação e reutilização de conectados na Web. Outros problemas visualizados, consequentes do anterior, é que na maiorias dos casos pode-se analisar apenas o resultado da publicação dos dados, ou seja, os dados publicados, mas não pode-se verificar e ou avaliar a o fluxo de trabalho (*workflow*) realizado antes da publicação.

Neste trabalho, considera-se como fundamental para qualidade de dados, o fluxo de trabalho (*workflow*), ou seja, os passos realizados para que aconteça a publicação. Sendo assim, propõe-se desenvolver um processo que incentive, guie e contribua com o aprimoramento da publicação e reutilização de dados abertos ligados na Web, levando em consideração os padrões de *Linked Data*, bem como a descrição detalhada de cada passo levantado como relevante.

Logo, este capítulo tem como objetivo apresentar o processo desenvolvido neste trabalho, que possui o intuito de simplificar e recomendar a utilização de algumas ferramentas, padrões e diretrizes para transformação, interligação, exposição e compartilhamento de recursos de dados no modelo RDF, seguindo os princípios de *Linked Data*.

O título do processo surgiu da seguinte maneira, do verbo da língua Inglesa "triplify", que quer dizer triplificar, ou seja, gerar ou transformar dados em triplas RDF. E a substantivo "*Process*"que também, vem da língua Inglesa e significa processo. Depois uniu-se as duas palavras e surgiu o termo "*Triplify Process*"que significa Processo de Triplificação. A seção seguinte detalha o processo.

4.1 *Triplify Process*

O *Triplify Process* incentiva, guia e contribui com os usuários no aprimoramento do processo de concepção, preparação, transformação, publicação e reutilização de dados abertos e/ou ligados na Web, levando em consideração os padrões de *Linked Data*.

O *Triplify Process* baseia-se no princípio de promover o reuso de padrões através de um processo guiado composto por nove fases. Os passos do *Triplify Process* são representados na Figura 15.

As fases de 1 a 10 foram nomeadas, respectivamente, Concepção do Projeto, Seleção dos dados de origem, Estruturação, Mapeamento, Coleta, Refinamento, Transformação, Armazenamento e Publicação, Enriquecimento e Atualização. Cada passo é descrito detalhadamente como se segue:

1. **Concepção do Projeto**

   Esta é a fase inicial, na qual se tem a finalidade de realizar o planejamento da publicação dos dados, conhecer o cenário, decidir quais dados a serem publicados, bem como o que incluir, identificar equipe(s) e ou órgão(s) que devem participar do processo e as opções

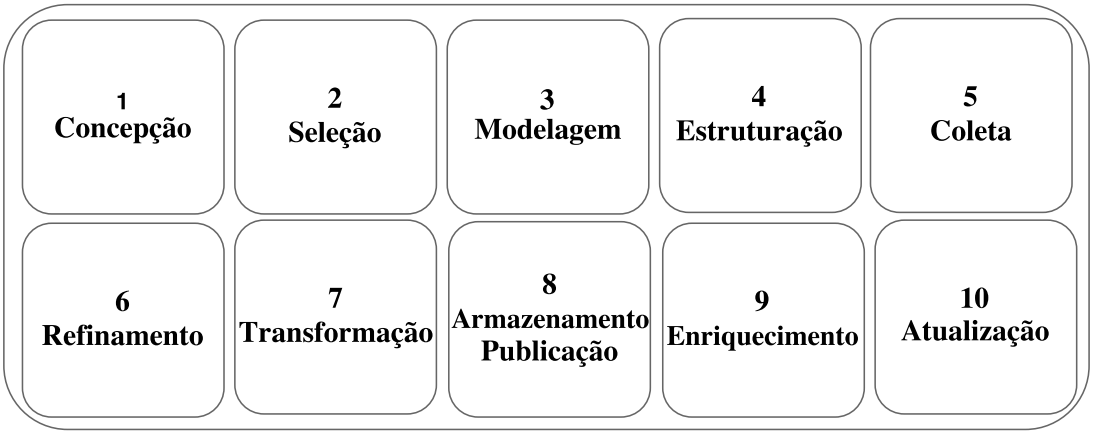


Figura 15 – Passos do *Triplify Process*.

de coleta, infraestrutura utilizada em todo o processo e para publicação. Logo, o objetivo principal dessa fase, é arquitetar e definir o projeto.

Para isso, deve ser realizada uma atividade de Estudo de Domínio, que consiste em captar conhecimento do cenário, como: aplicabilidade, glossário(fundamental definir o que é uma fonte de dados sobre a perspectiva do projeto), requisitos, especificações, restrições, viabilidade, infraestrutura, equipe(s) ou órgão(s) envolvidos e a extensão do projeto. Ou seja, definir o projeto considerando as reais necessidades, limitações, complexidade e importância.

O ideal é que os resultados dos conhecimentos adquiridos no Estudo de Domínio, devam ser descritos em um artefato, que facilite a compreensão e o acompanhamento do projeto pela equipe(s) e/ou órgão(s) envolvidos. propõe-se um modelo de artefato que se chamará de Documento de Visão. Este documento, por sua vez, pode seguir o modelo sugerido no Apêndice A, bem como pode ser modificado ou adaptado de acordo com a real necessidade do projeto.

Portanto, o resultado dessa fase, é uma definição preliminar do projeto, bem como uma área que os dados serão publicados e uma descrição detalhada dessas definições, em um artefato (Documento de Visão). Esclarece-se, que projeto com manipulação de dados são incrementais, ou seja, sempre tem-se a necessidade de realizar mudanças ao decorrer das fases para que se possa chegar a um resultado esperado.

As definições realizadas nesta fase poderão ser modificadas no decorrer do projeto de acordo com as necessidades apresentadas para obter os objetivos do projeto, contudo, precisam ser comprovadas e acordadas entre os envolvidos no projeto. Além disso, recomenda-se a atualização contínua do Documento de Visão, com versões diferentes, para que no final possa evidenciar a escala de mudanças ocorridas ao longo do projeto.

Quanto a decisão de quais dados incluir na publicação, opções de coleta e plataformas de publicação serão tratadas com mais detalhes nas próximas fases.

2. **Seleção dos dados de origem**

Esta fase tem a finalidade de decidir quais fontes de dados serão utilizadas para coleta, bem como, reconhecimento da estrutura e uma análise prévia dos dados. Esta fase é complexa e divide-se em algumas atividades, são elas:

- **Identificar Fontes de Dados Confiáveis e Relevantes**

  Esta atividade tem a finalidade de encontrar fonte(s) de dados relevantes e confiáveis, considerando a definição de fonte de dados (Base de dados) realizada anteriormente no glossário. Logo, esta atividade tem muita importância principalmente para a continuação dos esforços da manipulação dos dados, devido muitos riscos como os de procedência e qualidade dos dados.

  Para que o andamento do projeto não seja comprometido, as tarefas a seguir descritas foram consideradas, como essenciais para esta atividade:

  a) **Considerar dimensões de qualidade e seus respectivos indicadores.** Esta tarefa consiste em verificar se as fontes de dados (Bases de dados) possuem um nível mínimo de qualidade, para isso deve-se considerar dimensões de qualidade e seus respectivos indicadores.

  A literatura de qualidade da dados e informação fornece uma classificação completa das dimensões da qualidade de dados, no entanto, há uma série de discrepâncias na definição de dimensões devido à natureza contextual de qualidade. As dimensões são referências para a qualidade da informação. Em algumas situações, determinado grupo de dimensões podem ser importantes, e este grupo varia conforme a situação (GERMANO; TAKAOKA, 2012).

  Logo, as dimensões de qualidades foram escolhidas com base nos trabalhos de Wang, Ziad e Lee (2006) e com o significado para cada dimensão apresentado a seguir no Quadro 2.

  Segundo Wang, Ziad e Lee (2006), o significado de cada categoria é a seguinte:

  - Intrínseca: características intrínsecas dos dados, independentes da sua aplicação;
  - Acessibilidade: aspectos relativos ao acesso e à segurança dos dados.
  - Contextual: características dependentes do contexto de utilização dos dados;
  - Representacional: características derivadas da forma como a informação é apresentada;

  Assim sendo, recomenda-se utilizar esses indicadores e dimensões como critérios de avaliação para qualidade, compreensão e confiabilidade dos dados de origem, de modo que, garanta um mínimo de credibilidade e não comprometa o processo de publicação dos dados. É importante esclarecer que esses indicadores e dimensões não garantem totalmente a qualidade. Além disso, pode-se adicionar novos indicadores e/ou dimensões de acordo com a necessidade do projeto.

Quadro 2 – Categorias, dimensões e definições da Qualidade da Informação

Categoria	Dimensão	Definição
Intrínseca	Acuracidade	Quanto o dado é correto e confiável
	Objetividade	Quanto o dado é imparcial
	Credibilidade	Quanto o dado é considerado como verdadeiro e verossimil
	Reputação	Quanto o dado considerado em termos de sua fonte ou conteúdo
Acessibilidade	Acessibilidade	Quanto o acesso ao dado esta disponível, ou fácil e rapidamente recuperável
	Segurança no acesso	Quanto o acesso ao dado é, restrito apropriadamente para manter sua segurança
Contextual	Relevância	Quanto o dado é aplicável e útil para a tarefa a ser realizada
	Valor Agregado	Quanto o dado é benéfico e proporciona vantagens para seu uso
	Temporalidade/ Oportunidade	Quanto o dado esta suficientemente atualizado para a tarefa a ser realizada
	Integridade/Perfeição	Quanto o dado não esta extraviado e é suficientemente para a tarefa em amplitude e profundidade
	Quantidade de dados apropriado	Quanto o volume do dado é apropriado para a tarefa ser executada
Representação	Interpretabilidade	Quanto o dado está em linguagem apropriada, símbolos e unidades, e as definições são claras
	Facilidade de entendimento	Quanto o dado é facilmente compreendido
	Representação concisa	Quanto o dado esta compactamente representado
	Representação consistente	Quando o dado é apresentado em um mesmo formato
	Facilidade de manipulação/ Operação	Quando o dado é fácil de ser manipulado e aplicado em diferentes tarefas

Fonte: Adaptado de Wang, Ziad e Lee (2006) e Pipino, Lee e Wang (2002) .

b) **Identificar fontes de dados candidatas.** Essa tarefa consiste em realizar pesquisas na Web, procurando possíveis bases de dados de acordo com o contexto e que atendam as especificações e restrições realizadas na primeira fase. Recomenda-se alguns sites de pesquisas e buscas de dados conhecidos mundialmente, por referenciarem muitos *datasets*, são eles: DATAHUB [1], Europe's Public Data [2], DATA.GOV[3] e OpenSpending[4].

[1] http://datahub.io/dataset
[2] http://publicdata.eu/
[3] http://www.data.gov/
[4] https://openspending.org/

c) **Avaliar fontes(bases) de dados candidatas considerando as dimensões e indicadores de qualidade.** Essa tarefa consiste em avaliar cada fonte de dados candidata, para que se possa escolher a mais adequada ao projeto, bem como, tenha um mínimo de qualidade. Para avaliação, o método adotado será se a fonte de dados candidata possui ou não a dimensão de cada categoria da qualidade de dados. Deve-se utilizar o Quadro 5 que tem as dimensões de cada categoria e marcar com um "X", as dimensões de qualidade que cada fonte de dados atender.

Quadro 3 – Categorias, dimensões e fontes de dados para avaliação

Categoria	Dimensão	Fonte 1	Fonte 2	Fonte (N+1)
Intrínseca	Acuracidade			
	Objetividade			
	Credibilidade			
	Reputação			
Acessibilidade	Acessibilidade			
	Segurança no acesso			
Contextual	Relevância			
	Valor Agregado			
	Temporalidade/ Oportunidade			
	Integridade/Perfeição			
	Quantidade de informação apropriada			
Representação	Interpretabilidade			
	Facilidade de entendimento			
	Representação concisa			
	Representação consistente			
	Facilidade de manipulação/ Operação			

Fonte: Adaptado pelo autor.

d) ***Rankear* as fonte(s) (bases) de dados candidatas.** Essa tarefa consiste em determinar uma ordem crescente para as fontes (bases) de dados candidatas, considerando as que atenderam mais dimensões em cada categoria de qualidade.

e) **Usar as fontes (bases) a serem utilizadas de acordo com a classificação.** Essa tarefa consiste em selecionar ou usar as fontes de dados candidatas com melhor colocação na classificação da tarefa anterior. Logo, as fontes de dados candidatas escolhidas atenderão os principais critérios requeridos e serão consideradas como confiáveis e relevantes para o projeto, por terem atendido melhor as dimensões e indicadores de qualidade, bem como as especificações e restrições apresentadas na primeira fase, devidamente consideradas.

- **Navegar para conhecer a estrutura**

  Esta atividade tem a finalidade de realizar uma visualização prévia dos dados, para que se tenha uma melhor compreensão e entendimento do modelo e da estrutura

que estas fontes (bases) de dados estão disponíveis. Bem como, identificar melhor a acessibilidade dos dados, ou seja, para realizar uma visualização prévia dos dados, deve-se por exemplo: realizar *download* de arquivo (CSV, XML), realizar uma consulta em um banco relacional ou uma chamada a um *endpoint* de um Web Service, entre outros.

- **Analisar**

  Esta atividade tem a finalidade de identificar o modelo de dados, conhecer os principais vocabulários utilizados, bem como, examinar os elementos e atributos fornecidos pelos de dados das fontes escolhidas. Nesse contexto, considera-se elementos e atributos dos dados, como sendo colunas e/ou linhas (tuplas) de tabelas de bancos de dados, planilhas, entre outros.

Portanto, o resultado dessa fase, é conhecimento de fontes que podem ser utilizadas pelo projeto, assim como, o modelo de dados e o vocabulário utilizado. Isso garantirá um mínimo de qualidade, devido atender as principais categorias e dimensões de qualidade, além de facilitar as próximas fases devido o conhecimento do modelo e vocabulário utilizado.

3. **Estruturação**

   Esta fase tem a finalidade de definir a estrutura dos dados a serem distribuídos. Por exemplo, no caso da utilização de Web Semântica e Dados Ligados, que ontologias serão utilizadas para as instâncias dos dados ou se nova ontologia será definida, caso seja necessária. Nesse contexto, define-se ontologia como especificação formal e explícita de uma conceitualização que permita compartilhamento e reutilização de conhecimento, ou seja encontrar uma linguagem legível para máquinas e/ou humanos que explicite de forma compreensível e clara, os conceitos, propriedades, relações, funções, restrições e axiomas.

   A sociedade tem utilizado a anotação dos dados publicados em ontologias para facilitar integração, fusão e posteriores buscas e a etapa de análise de qualidade – explicita que o suporte analítico pode se beneficiar quanto ao contexto e a estrutura. Logo, um dos principais objetivos dessa fase, é determinar uma ontologia de aplicação, que por sua vez pode ser reutilizada, adaptada e/ou criada de acordo com o objetivo a ser atingindo pelo projeto.

   Quanto a isso, o ideal é reutilizar ou adaptar as ontologias existentes, pode-se pesquisar estas nos principais motores de buscas em relação a Web Semântica: ***Falcons*** [5], ***Watson*** [6], ***Swoogle*** [7], ***Schema*** [8] e outros.

[5] http://ws.nju.edu.cn/falcons/objectsearch/index.jsp
[6] http://watson.kmi.open.ac.uk/WatsonWUI/
[7] http://swoogle.umbc.edu/
[8] https://schema.org/

Caso, não encontre-se uma ontologia que esteja no contexto do trabalho proposto, pode ser criada de acordo com o objetivo a ser atingindo pelo projeto, divide-se em algumas atividades, que são:

- **Definir conceitos do domínio**

  Esta atividade consiste em definir os conceitos do domínio selecionados para integrar o esquema da ontologia. Para isso, deve-se identificar e avaliar os principais conceitos envolvidos no cenário do trabalho a ser realizado, bem como, nos dados e seus metadados a serem utilizados. Logo, o intuito é realizar a modelagem da ontologia coerente com domínio do problema a ser tratado.

- **Selecionar Vocabulários**

  Esta atividade tem a finalidade de pesquisar, criar, combinar e misturar os vocabulários que será usado ao criar a ontologia. Embora não haja restrições para seleção de vocabulários, é considerada uma boa prática o reuso de termos de vocabulários RDF amplamente usados para facilitar o processamento de *Linked Data* pelas aplicações clientes (BIZER; CYGANIAK; HEATH, 2007).

  Sendo esta atividade ampla e complexa, para facilitar, foi dividida nas seguintes tarefas:

  – **Reutilizar Vocabulários.** Essa tarefa consiste em realizar uso de termos de vocabulários amplamente usados em outras publicações de dados disponíveis na Web, garantindo que o *dataset* criado estará interconectado com outros *datasets* na *Linked Data Cloud* e facilitar o processamento de *Linked Data* pelas aplicações clientes. Logo, essa tarefa foi dividida em duas, são elas:

    * Conhecer os Principais Vocabulários. O Quadro 4 a seguir apresenta os vocabulários mais comuns.

Quadro 4 – Uso de Vocabulários Comuns

Prefix	Namespace	Used By
dc	http://purl.org/dc/elements/1.1/	66 (31.88 %)
foaf	http://xmlns.com/foaf/0.1/Person	55 (26.57%)
terms	http://purl.org/dc/terms	38 (18.36%)
skos	http://www.w3.org/2004/02/skos/core	29 (14.01%)
akt	http://www.aktors.org/ontology/portal#	17 (8.21%)
geo	http://www.w3.org/2003/01/geo/wgs84_pos#long	14 (6.76%)
mo	http://purl.org/ontology/mo/	13 (6.28%)
bibo	http://purl.org/ontology/bibo/	8 (3.86%)
vcard	http://www.w3.org/2006/vcard/ns#	6 (2.90%)
frbr	http://purl.org/vocab/frbr/core#	5 (2.42%)
sioc	http://rdfs.org/sioc/ns#	4 (1.93%)

Fonte: (4TH. . . ,).

    * Como Adquirir Outros, Bons Vocabulários. Para adquirir outros, bons vocabulários deve seguir os seguintes critérios: Procurar nas aplicações existentes, ativas pela sociedade, apoiado por grandes organizações respeitáveis, simples e com poucas restrições ou com pressupostos ontológicos. Para encontrar vocabulários pode-se utilizar os motores de buscas citados anteriomente (***Falcons***, ***Watson***, ***Swoogle***, ***Schema***, entre outros).

  – **Criar seus Vocabulários.** Novos termos só devem ser definidos se não for encontrado nenhum termo existente refletindo a semântica do conceito que se desejava representar. Logo, o intuito é criar apenas termos essenciais ao esquema a ser projetado e, até então, inexistentes.

  – **Misturar e combinar vocabulários.** Essa tarefa consiste em estabelecer relações entre os termos de vocabulários proprietários para termos de outros vocabulários. Para isso, deve realizar uso das propriedades:**owl:equivalentClass**, **owl:equivalentProperty**, **rdfs:subClassOf** e **rdfs:subPropertyOf**.

- **Selecionar URIs adequadas**

  Essa atividade tem a finalidade de escolher minuciosamente URIs, com nomes que outros publicadores de dados possam usar de forma confiável para criar ligações entre as duas fontes de dados. Além disso, é necessário ter infraestrutura técnica para tornar estas URIs dereferenciáveis, ou seja, prover conteúdo quando as URIs forem acessadas.

  Algumas outras recomendações na escolha de URIs são:

  – Utilizar URIs HTTP, pois o esquema "http://" é o único esquema de URIs que é amplamente suportado pelas ferramentas e infraestrutura dos dias atuais.

  – Definir URIs em um namespace HTTP sob controle, onde se pode implementar o que for necessário para torná-las dereferenciáveis.

  – Tentar manter as URIs estáveis e persistentes. Trocar as URIs em um momento posterior irá quebrar todos os links existentes.

- **Modelar**

  Essa atividade consiste em reunir todas as atividades dessa fase e usar uma ferramata que auxilie a criação ou integração do esquema que servirá como base para a criação do *dataset* RDF. A seguir, algumas das principais ferramentas utilizadas para criar e/ou integrar ontologias: Protégé [9], Vitro [10], Neologism[11], entre outras.

Logo, o resultado dessa fase, é a definição de uma ontologia de aplicação, ou seja, definição de um esquema final para o *dataset* RDF.

[9] http://protege.stanford.edu/
[10] http://vitro.mannlib.cornell.edu/
[11] http://neologism.deri.ie/

4. **Mapeamento de vocabulários fonte (source) para vocabulários destino (target)**

   Esta fase tem a finalidade de especificar como instâncias de dados de um esquema(*source*) correspondem à instância de dados de outro esquema (*target*). Que significa relacionar um vocabulário de um modelo ou fonte de origem para o modelo de dados escolhido como final.

   Esses mapeamentos podem ocorrer de forma manual, semi-automática e até mesmo automática. A forma manual, é quando não se utiliza nenhuma ferramenta e definido de forma mecânica por um participante do projeto. Semi-automático é quando se utiliza uma ferramenta para auxílio e nela define-se uma especificação, de modo que, a ferramenta que reconheça que um elemento do modelo de dados origem, irá representar outro no modelo final, por exemplo, uma tabela pessoa no modelo relacional, representar uma classe no modelo RDF. O mapeamento automático é quando usa-se uma ferramenta que reconhece ou já tem pré-definida, como se representa um termo de um modelo origem para um modelo final.

   Um exemplo de ferramenta é o *Framework* R2R[12] para traduzir os dados da Web que é representado usando termos de vocabulários diferentes em um único vocabulário alvo. Os vocabulários mapeamentos são expressos usando a linguagem de mapeamento R2R[13]. A linguagem prevê transformações simples, bem como para as transformações estruturais mais complexas e de transformações de valor de propriedade, tais como unidades de medida diferentes normalizando ou manipulações de strings complexas. A sintaxe do R2R é muito semelhante à linguagem de consulta SPARQL, o que facilita a curva de aprendizagem.

   Assim sendo, o resultado dessa fase é saber como será realizado o mapeamento de um termo do modelo origem (*source*) para um modelo final(*target*) e qual ferramenta pode ser utilizada para auxiliar.

5. **Coleta dos dados**

   Esta fase tem a finalidade da aquisição dos dados propriamente ditos, que incluem dados a ser criados e dados já existentes, obtidos a partir de *download* de planilhas, consumo de serviços, consultas a bancos de dados ou RDF *Store*, entre outras.

   Em alguns casos, como: dados com estrutura complexa (formatos e padrões a serem corrigidos) e dados que são disponibilizados momentaneamente (como dados de *Global Positioning System* (GPS) de ônibus de uma determinada cidade, por exemplo) devem ser obtidos e armazenados pelo menos temporariamente no formato original para serem analisados, selecionados, modelados, refinados, enriquecidos e transformados para o modelo de dados final, com mais minuciosidade.

[12] http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/
[13] http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/spec/

Em outras situações, os dados tem o formato esperado, como por exemplo o próprio RDF na RDF *Store* e a consulta deste já é o resultado esperado, logo estes são fortes candidatos a serem utilizados, sem necessidade de transformações e/ou limpeza seja para inserção em outra RDF *Store* ou para geração de um *Dump*.

Observação: Em caso de realização de atividades com arquivos, o ideal é utilizar as boas práticas de nomenclatura de pastas e arquivos apresentado no Anexo A, como por exemplo o *download* de um arquivo CSV.

Logo, o resultado dessa fase é obter os dados nas fontes selecionadas anteriormente. Pois os dados serão necessários para as fases seguintes.

6. **Refinamento**

   Esta fase tem a finalidade de realizar atividades de seleção e melhoramento dos dados de origem buscando aumentar sua qualidade. Ainda nesta fase, tem-se a decisão sobre os erros ocorridos durante o refinamento, enriquecimento e transformação dos dados. Esta fase terá subetapas que são as seguintes:

   - **Seleção**

     Essa atividade tem a finalidade de escolher cuidadosamente os atributos e as informações dos dados coletados que são essenciais para o projeto. Com o auxílio da atividade de análise realizada anteriormente na segunda fase, pode-se realizar esta atividade com maior certeza.

     Exemplo: Seleção de determinadas tuplas ou colunas de uma tabela do banco de dados.

   - **Limpeza e formatação**

     Essa atividade está relacionada diretamente ao melhoramento da qualidade e tem a finalidade de eliminar dados redundantes, inconsistentes(brancos, nulos e desnecessários), recuperar dados incompletos e avaliar possíveis dados discrepantes ao conjunto.

     Quanto a formatação, consiste em alterar o formato de origem e deixar em um formato que se considera ideal de acordo com a necessidade final. Um exemplo, é a retirada de caracteres especiais de um determinado campo de texto, ou retirada de pontos e traços de um Cadastro de Pessoa Física(CPF).

     Nessa atividade, pode-se utilizar alguns procedimentos e algumas ferramentas que apoie, como *Pentaho Data Integration* que facilita a seleção das colunas, linhas ou tuplas desejadas e/ou utilização de expressões regulares para identificar dados nulos, incompletos, formatações, entre outras.

   - **Tomada de decisão sobre erros / problemas durante a transformação dos dados.**

     Esta atividade tem a finalidade de realizar decisões sobre erros e problemas no decorrer do refinamento, enriquecimento e/ou até mesmo da transformação(triplificação).

Divide-se essa atividade em três tarefas básicas para facilitar a compreensão e decisão, são elas:

- **Descartar / ignorar os dados problemáticos**
  Esta tarefa tem a finalidade de decidir descartar ou ignorar os dados incompletos, em brancos ou nulos, devido estes não fornecerem resultados ou informações importantes para os usuários finais. O ideal é antes de destacá-los, verificar se a fonte é primária e se esta não fornece os dados de forma mais completa, pois estes podem ter sido adquiridos de terceiros que já processaram deixando os dados voltados para seu objetivo específico e possivelmente incompleto para outros fins.
- **Tentar corrigi-los automaticamente**
  Esta atividade consiste em tentar corrigir alguns dados como em casos incompletos, buscar diretamente na fonte primária, ou realizar uma tarefa a mais, como por exemplo, você tem um arquivo com todas as datas de nascimento dos clientes e quer suas idades atuais, neste caso pode-se usar uma ferramenta que apoie neste cálculo. Em outros casos, tem que pesquisar e encontrar outra(s) fonte(s) que ao integrar, complete os dados que são necessários no projeto atual. Lembrando que é importante ao publicar esses dados informar de forma clara ao usuário final as fontes e integrações realizadas, essa informação pode ser descrita nos metadados.
- **Armazenar dados problemáticos e deixar a correção por conta do usuário que é notificado sobre o problema**
  Esta tarefa consiste em tomar a decisão de simplesmente armazenar os dados problemáticos em outro repositório para não comprometer a triplificação e notificar o problema ao usuário final dos dados que serão publicados, essa notificação pode ser realizada através dos metadados explicando sobre a decisão tomada, que pode ser deixar os dados incompletos ou de tirá-los da publicação devido suas inconsistências. Lembrete, é importante referenciar ou disponibilizar o que foi retirado por ser incompleto ou inconsistente, para que em caso de usuário final precisar, deste dado, ele pesquise mais em outras fontes ou até mesmo obtenha estes e gere-os em casos mais simples.

Assim sendo, o resultado dessa fase será um dado de qualidade com apenas o que é de interesse para o projeto, bem como, com suas devidas formatações. Portanto, o que se tem nessa situação são dados preparados para transformação para o modelo de dados considerado como final pelo projeto.

7. **Transformação**

   Esta é a fase onde os dados de origem são transformados de acordo com o modelo de dados e a plataforma escolhida para publicação.

Logo, nesse processo, como será utilizado o modelo de dados RDF para publicação, converte-se os dados preparados até o presente momento pelas fases anteriores, para triplas RDF, comumente usado em *Linked Data*. Para isso, será utilizada a seguinte abordagem:

- **Materialização dos dados no modelo RDF**

  Consiste em usar um processo de **conversão**, onde os dados não RDF são usados para gerar um arquivo RDF através de uma ferramenta específica [14]. Desse modo, através de conversores específicos é possível converter bancos de dados relacionais, planilhas, arquivos CSV, arquivos XML e outros documentos para o formato RDF. Após a conversão, pode-se com o auxílio da própria ferramenta, inserir esses dados diretamente em uma RDF Store e/ou gerar um arquivo para os dados serem carregados em uma RDF *Store* e publicados.

  Uma vantagem dessa abordagem é a melhoria de desempenho que pode ser obtida ao usar formas de armazenamento especificamente otimizadas para realizar a persistência de triplas RDF. No entanto, o armazenamento das triplas requer espaço extra em relação aos dados originais. Além disso, a conversão demanda certo tempo para ser realizada e os dados em RDF podem ficar desatualizados em relação aos dados originais.

- **Opcional**: adicionar descrição de conjunto de dados usando *Vocabulary of Interlinked Datasets* (VoID)

  Para Alexander et al. (2009). *Vocabulary of Interlinked Datasets* (VoID) é um vocabulário desenvolvido para descrever metadados sobre datasets RDF, que são coleções de dados estruturados como RDF e publicados e mantidos por um mesmo provedor de dados. Os metadados descritos pelo VoID podem ser metadados gerais, descritos em conjunto com Dublin Core; metadados de acesso; metadados estruturais e metadados sobre interligações entre *datasets*. O objetivo do vocabulário VoID é ser uma ponte entre os publicadores e os usuários de um *dataset* RDF, apoiando desde a descoberta dos dados até a catalogação e arquivamento de *datasets*.

Portanto, o resultado dessa fase é ter os dados no modelo RDF, ou seja, triplificados (abordagem materializada).

8. **Armazenamento e Publicação**

   Esta fase consiste em armazenar e publicar os dados usando os princípios e melhores práticas de *Linked Data* para disponibilizar os dados na Web. Para isso, é preciso fornecer uma interface *Linked Data* com URIs dereferenciáveis para cada entidade e links RDF para outras fontes de dados. Esses são os requisitos mínimos, mas além deles é frequente a disponibilização de SPARQL *endpoints* e de *dumps* dos dados. Para simplificar divide-se esta fase em algumas atividades, são elas:

[14] http://www.w3.org/wiki/ConverterToRdf

- **Armazenamento**

  Esta atividade tem a finalidade de agrupar os dados RDF em grafos RDF e carregar-los em memória, arquivo texto ou banco de dados para triplas RDF, chamado RDF *Triple Store*. O armazenamento de triplas como arquivo pode usar algum formato de serialização de RDF, como RDF/XML, Notation3 (N3), Turtle ou NTriples.

- **Disponibilizar um *end-point* SPARQL para consultar e manipular os dados**

  Esta atividade tem a finalidade fornecer um SPARQL *Endpoint* que é um serviço Web com suporte a linguagem SPARQL que possui uma URI específica para receber requisições HTTP com consultas SPARQL, possibilitando por exemplo, a execução de consultas sobre uma fonte de dados disponível no padrão de *Linked Data* na Web. Alguns *endpoints* permitem inclusive atualizações através de SPARQL *Update* (GEARON; PASSANT; POLLERES, 2012).

- **Fornecer uma Interface *Linked Data***

  Esta atividade tem a finalidade disponibilizar uma interface *Linked Data* capaz de tratar requisições de URIs, dereferenciar URIs, tratar dos redirecionamentos 303 requeridos pela arquitetura Web e da negociação de conteúdo entre descrições de um mesmo recurso em diferentes formatos. Através da negociação de conteúdo pode-se retornar a representação mais adequada ao cliente. Assim, um usuário humano pode requisitar uma URI e obter uma representação HTML do recurso. Isso é possível porque os clientes HTTP enviam cabeçalhos indicando que tipo de representação eles preferem obter. Daí o servidor analisa o cabeçalho ao receber uma requisição e seleciona a resposta adequada. Pubby[15] é uma ferramenta para fornecer de forma simples uma interface *Linked Data* para fontes de dados RDF, como SPARQL *endpoints* e arquivos RDF estáticos.

Assim sendo, o resultado dessa fase é o armazenamento e a publicação dos dados no modelo RDF, disponibilizando uma *Interface Linked Data*, SPARQL *endpoints* e de *dumps* dos dados. Ou seja, atender os princípios e boas prática de *Linked Data*.

9. **Enriquecimento**

   Esta fase tem a finalidade de realizar associações de dados, ou seja, a interligação de dados a partir de múltiplas fontes sobre a Web de Dados. Isso significa encontrar termos correspondentes entre os vocabulários utilizados e entidades(recurso) correspondentes entre conjuntos de dados que pretende-se realizar a associação. Esta fase é complexa e para simplificar será divida nas atividades a seguir:

   - **Mapeamentos de esquemas**

     Esta atividade consiste em transformar os dados de acordo com vocabulários correspondentes, entre as fontes ou bases de dados comparadas.

[15] http://www4.wiwiss.fu-berlin.de/pubby/

- **Resolução de identidade**

  Esta atividade consiste em produzir ligações entre entidades com base em comparações definidas de acordo com algum critério utilizado, como alguns algoritmos. Nessa atividade que temos o intuito principal de encontrar diferentes URIs que são usados em diferentes fontes de dados para identificar a mesma entidade no mundo real, para que se possa comparar e criar-se ligações entres esses recursos, pode-se utilizar o auxílio de algumas ferramentas conhecidas pela comunidade e utilizadas pela sociedade. As ferramentas são:

  – Silk[16]. O Silk é um *framework* para encontrar relações entre entidades dentro de dados de diferente fontes. Editores de dados podem usar-lá para definir ligações RDF a partir de sua fontes de dados para outras fontes de dados na web. Silk apresenta uma linguagem declarativa para especificar quais tipos de links RDF devem ser descobertos entre fontes de dados, bem como quais condições entidades terão que cumprir para serem interligadas.
    As condições das ligações podem se basear em várias métricas de similaridade e pode tomar o gráfico à volta entidades em conta, que é dirigida usando uma linguagem seleção baseada em caminho(Heurística). O Silk acessa fontes de dados através do protocolo SPARQL e pode, assim, ser utilizado para replicar conjuntos de dados, tanto na web, como localmente(VOLZ et al., 2009).
    Para realizar as transformações dos valores dos dados, são utilizados um conjunto de operadores, fornecidos pela ferramenta, como: categoria de normalização que tem elementos como retirar caracteres especiais, transformar texto em maiúsculas(Upper Case), categoria de *replace* que aceita expressões regulares, entre outras que podem ser utilizadas de acordo com a necessidade.
    As medidas de distância com seus devidos limiares que servem para comparação de elementos como strings, datas, possuem diversas categorias como: charecter-based que fornece alguns algoritmos como Levenshtein *distance*, Jaro *distance* e outros. A categoria tokenbased, por exemplo, possui alguns algoritmos de comparação bem conhecidos, como: Jaccard, Softjaccard e outros.
    Com o resultados dos algoritmos de distância, realiza-se as agregações, algumas das opções são: *maximum*, *minimum*, *average* e outras.
    Com a heurística especificada no Silk, de acordo, com as transformações, medidas de distâncias e agregadores, ele consegue gerar por exemplo, *links* sameAs, como referência de que em um recurso que esta em um *dataset* é o mesmo recurso no outro *dataset* que esta sendo comparado e carregar esses links diretamente em um *Endpoint* SPARQL ou arquivo.

  – Limes[17]. O Limes é um framework para descoberta de ligação entre entidades

[16] http://silkframework.org/
[17] http://aksw.org/Projects/LIMES.html

de dados em larga escala com base em características de espaços métricos. É facilmente configurável através de uma interface web e pode ser baixado como ferramenta independente para a realização de descoberta de ligações de dados, localmente.

Além disso, o *framework* Limes utiliza diferentes abordagens técnicas de aproximação para calcular as estimativas da semelhança entre os recursos de diferentes fontes de dados. Para isso, ele utiliza algoritmos de aprendizagem de máquina, como: EAGLE [18], COALA[19] e EUCLIDES [20].

O Limes acessa fontes de dados através do protocolo SPARQL e utiliza uma linguagem de especificação, onde se declara as bases de dados a serem comparadas, bem como, as métricas a serem utilizadas na comparação. De acordo, com o algoritmo definido nas métricas, ele consegue gerar *links* sameAs entre os recursos que são semelhantes nos *datasets* comparados.

- **Fusão**

  Esta atividade de fusão está relacionada com enriquecimento dos dados e consiste no processo de integração de múltiplas representações do mesmo objeto do mundo real em uma única, consistente e limpa representação. O principal desafio na fusão de dados é a resolução de conflitos de dados, isto é, a escolha de um valor em situações onde múltiplas fontes fornecem valores diferentes para a mesma propriedade de um objeto (HEATH; BIZER, 2011b). Para maiores detalhes pode-se utilizar dois trabalhos como bases: (SCHULTZ et al.,) e o (VIDAL et al., 2015).

Portanto, o resultado dessa fase pode ser o enriquecimento especifico do *dataset* na etapa de enriquecimento, com um conjunto de ligações **owl: sameAs** (Criado entre dois URIs) que representam os recursos equivalentes a outra base, resultante da etapa de ligação(resolução de identidade), devido a(s) técnica(s) de similaridade utilizada(s) na(s) comparação(ões). Bem como, a fusão que é a criação de um novo *dataset* com o resultado da junção de outros dois *datasets* e suas devidas relações.

Mais detalhes sobre esta fase podem ser obtidos nos seguintes trabalhos: LDIF(SCHULTZ et al.,) e o *Specification and incremental maintenance of linked data mashup views*(VIDAL et al., 2015).

10. **Atualização**

    Esta fase consiste na atualização do *dataset* RDF de acordo com novos dados disponibilizados pelas fontes de origem, no mesmo modelo e estrutura fornecida inicialmente. Não será abordada atualização para os casos que a fonte de dados de origem, mudar o modelo e/ou a estrutura dos dados. Caso, aconteça mudanças na estrutura ou no modelo dos dados

[18] http://svn.aksw.org/papers/2012/ESWC_EAGLE/public.pdf
[19] http://link.springer.com/chapter/10.1007%2F978-3-642-38288-8_30
[20] http://disi.unitn.it/ p2p/OM-2013/om2013_Tpaper3.pdf

fornecidos pela fonte de origem, deve-se reinicializar todo o processo desde a concepção. Essa é uma fase será dividida em atividades, são elas:

- **Atualizar do *dataset* RDF final apenas os dados de origem novos**

  Esta atividade tem a finalidade de atualizar o *dataset* RDF, com os novos dados disponibilizados pelas fontes de dados origem sem modificação no modelo inicial. Para essa atividade, tem que utilizar um mecanismo que identifique os acréscimos de novos dados pela fonte de dados origem obedecendo o modelo fornecido inicialmente

- **Atualizar do *dataset* RDF final por completo**

  Esta atividade tem a finalidade de atualizar o *dataset* RDF, fazendo um *backup* dos dados RDF e regerando a triplificação de todos os dados novamente, incluindo os novos dados disponibilizados pelas fontes de dados origem. Para essa atividade, tem que utilizar um mecanismo que identifique os acréscimos de novos dados pela fonte de dados origem obedecendo o modelo fornecido inicialmente

  Bem como, uma forma de obter os dados triplificados na ocasião anterior, afinal, será triplificado tudo novamente.

4.2 Considerações do Capítulo

Neste capitulo, apresentou-se o processo proposto com as fases essenciais e comuns na publicação de dados ligados na Web. Além disso, foram expostas as principais ferramentas, orientações e dicas para atender os padrões, princípios e boas práticas de *Linked Data*. O capítulo a seguir procura aplicar e avaliar este processo proposto, assim como, as demais recomendações.

5 ESTUDO DE CASO

Neste capítulo será apresentado um estudo de caso mostrando o funcionamento de todos os passos do *Triplify Process*, servindo assim para validá-lo.

O estudo de caso realizado neste trabalho utilizará dados abertos e mostrará que ao se trabalhar com o modelo RDF e com o padrão *Linked Data* se tem maior facilidade para interligar e integrar os dados, evitando problemas de divergências.

Para isso, será utilizado o *Triplify Process* para servir como guia e apoio para publicação de dados ligados na Web, seguindo o padrão *Linked Data*. Além disso, será utilizada a ferramenta Pentaho Data Integration (kettle) e os *plugins* disponibilizados pelo projeto LinkedDataBR, denominados ETL4LOD, para automatizar a conversão dos dados para RDF e o framework Silk para interligação dos dados. Disponibilizará-se à sociedade e aos governos um *endpoint* SPARQL. Além disso, será mostrado a dificuldade de se trabalhar dados abertos que não sejam ligados.

5.1 Aplicação do Triplify Process

Para começar a aplicação do *Triplify Process* na publicação de dados ligados, a primeira atividade realizada foi a compreensão das fases e das recomendações. A seguir é detalhado a aplicação do processo, através da execução de cada fase.

5.1.1 Concepção do Projeto

Nesta primeira fase do processo foi definida uma área para atuar e escolheu-se trabalhar com dados abertos, a respeito do combate às empresas e pessoas inabilitadas ou inidôneas para desempenharem atividades ou funções públicas. Para mais conhecimento deste cenário, realizou-se um estudo de domínio.

No estudo, constatou-se que existem divergências nos dados abertos do Brasil, devido à existência de múltiplas fontes de um mesmo dado, fazendo com que em algum momento ocorra falta de consistência entre as diversas fontes, informações desatualizadas, incompletas, etc. Estes problemas dificultam o uso destes dados pela sociedade e pelo próprio governo.

Com isso, percebe-se que no Brasil o problema do combate às empresas fraudadoras de licitações de compras públicas não é devido à falta de lei e nem de ação dos órgãos competentes, como a CGU, o TCU e muitos outros. Uma das principais razões para que isso ocorra decorre da dificuldade de integração e divulgação de forma ampla dos dados abertos.

Além disso, percebe-se que muito dos dados disponibilizados pelos governos estão em formato CSV, o que corresponde ao nível de 3 *star*, no modelo 5 *star*. Para que se alcance as 5 *starts* é necessário agregar semântica e ser processável por máquinas (agente de software). Para isso é necessário converter os dados para o modelo *Resource Description Framework* (RDF) recomendado pela W3C, o qual representa os dados por meio de triplas formadas de sujeito, predicado e objeto.

Para isso, este trabalho propõe-se a publicação de *datasets* RDF, com dados abertos de

pessoas físicas e Organizações inabilitadas ou inidôneas para realizarem serviços público. Com o intuito de conseguir o nível máximo do modelo 5 *star*.

Com isso, foi definida a equipe que atuará nesse projeto e que tem um conhecimento avançado e bastante experiência em Dados Abertos, RDF, SPARQL, *Ontology*, Pentaho Data Integration (kettle) e *Linked Data*.

A infraestrutura fundamental que será utilizada na implementação desse processo de criação de *datasets* RDF consistiu do sistema operacional Ubuntu versão 64 bits, da plataforma Java Development Kit [1] 64 (JDK) versão 1.8.072 e da edição comunitária (Community Edition - CE) do Pentaho Data Integration (Kettle) versão 6.0.1.0-386. Além dos *steps* e *job entries* padrão do Kettle, os 4 *steps* do ETL4LOD e o *step* NTriple *Generator* para serem utilizados no desenvolvimento dos *workflows*. Quanto ao banco de triplas, será utilizado a edição *open source* do Virtuoso versão 06.01.3127, para armazenar as triplas RDF geradas pelo processo de publicação.

Ainda na primeira fase do *Triplify Process*, foi criado artefato (Documento de Visão), com as informações citadas anteriormente, bem como, com alguns outros detalhes, que se encontram no Apêndice B.

5.1.2 Selecionar dados de origem

Esta fase consiste em escolher fontes, navegar nos dados, para reconhecimento do modelo e estrutura dos dados de origem, assim como, analisar e avaliar outros fatores dos dados que são importantes para uma modelagem e publicação coerente do domínio.

O próximo passo nesta segunda fase, é encontrar fontes candidatas, considerando dimensões de qualidade e seus respectivos indicadores. Para isso, pesquisou-se nos principais motores de busca de dados abertos de acordo com as recomendações do processo, encontrou-se as seguintes fontes:

- Tribunal de Contas da União (TCU)[2], o qual tem em Responsabilização pública, os seguintes itens: Contas julgadas irregulares; Inabilitados para função pública; Licitantes inidôneos; Combate à corrupção e Eleições:
  - Cadastro de Responsáveis com Contas Julgadas Irregulares – CADIRREG. O CADIRREG é um cadastro histórico que reúne o nome de todas as pessoas, físicas ou jurídicas, vivas ou falecidas, detentoras ou não de cargo/função pública, que tiveram suas contas julgadas irregulares pelo TCU. Só é possível o acesso por nome, CPF/CNPJ e pelo número do processo. Isto impede o acesso amplo.
  - Lista de inabilitados: Uma relação que contém os nomes de todos os responsáveis a quem o Tribunal de Contas da União - TCU declarou inabilitados para o exercício de cargo em comissão ou função de confiança no âmbito da Administração Pública Federal, nos termos do art. 60 da Lei nº 8.443/92 (LOTCU). Esta lista está disponível

[1] http://www.oracle.com/technetwork/pt/java/javase/downloads/index.html
[2] http://portal.tcu.gov.br/comunidades/responsabilizacao-publica/

em um endereço variável, o qual dificulta a automatização da captura e carga (<https://contas.tcu.gov.br/pls/apex/f?p=2046:4>). Consulta no site ou através de *download* de arquivo no formato PDF ou CSV.

  - Lista de Inidôneos: A relação contém os nomes de todos os inidôneos para participarem de licitações realizadas pela Administração Pública Federal, nos termos do art. 46 da Lei nº 8.443/92 (LOTCU). Esta lista está disponível em um endereço variável, o qual dificulta a automatização da captura e carga <https://contas.tcu.gov.br/pls/apex/f?p=2046:5>). Consulta no site ou através de *download* de arquivo no formato PDF ou CSV.
  - Eleições: O CADIRREG serve para a elaboração da lista de responsáveis com contas julgadas irregulares a ser encaminhada à Justiça Eleitoral. Essa lista se constitui em um subconjunto do CADIRREG. No entanto, ter o nome identificado no CADIRREG não implica estar na lista encaminhada, pois outros critérios devem ser considerados.

- Dados Abertos[3]:
  - Licitantes Inidôneos segundo TCU / Relação de Licitantes Inidôneos em CSV (http://data.gov.br/dataset/licitantes-inidoneas-segundo-tcu/resource/ea594e4e-1708-4e97-b57c-7c60999dedcc). Através de *download* de arquivo em http://www.tcu.gov.br/inidoneos-csv. Faz referência a http://data.gov.br/dataset/licitantes-inidoneas-segundo-tcu.
  - Inabilitados para função pública segundo TCU (<http://data.gov.br/dataset/search?q=inabilitado>).

- Portal da Transparência [4]
  - Tem Cadastro de Empresas Inidôneas e Suspensas (CEIS) em <http://www.portaltransparencia.gov.br/downloads/snapshot.asp?c=CEIS#get>
  - Não tem *link* para as inabilitações de servidores.

Identificadas as fontes de dados candidatas será realizada a avaliação destas de acordo com os critérios e dimensões de qualidade e a classificação quanto as que atendem melhor o objetivo desse projeto.

Após avaliação, classifica-se as fontes em ordem crescente, ou seja, a que atendeu melhor os categorias e dimensões e a ordem ficou da seguinte maneira:

1. Tribunal de Contas da União (TCU)
2. Portal da Transparência
3. Dados abertos

[3] http://data.gov.br/
[4] http://www.portaltransparencia.gov.br/downloads/

Quadro 5 – Categorias, dimensões e fontes de dados para avaliação

Categoria	Dimensão	TCU	Dados Abertos	Portal da Transparência
Intrínseca	Acuracidade	X	X	X
	Objetividade	X	X	X
	Credibilidade	X	X	X
	Reputação	X	X	X
Acessibilidade	Acessibilidade	X	X	X
	Segurança no acesso	X	X	X
Contextual	Relevância	X		X
	Valor Agregado	X		X
	Temporalidade/ Oportunidade	X		X
	Integridade/Perfeição	X		X
	Quantidade de informação apropriada	X	X	X
Representação	Interpretabilidade	X	X	X
	Facilidade de entendimento	X	X	X
	Representação concisa	X	X	X
	Representação consistente	X	X	X
	Facilidade de manipulação/ Operação	X		X

Fonte: Adaptado pelo autor.

Com esta classificação, utilizou-se dados de duas fontes, TCU e do Portal da Transparência. Além disso, realizou-se uma visualização prévia dos dados para conhecer melhor suas estruturas e detalhes como acessibilidade.

Além disto, realizou-se uma análise minuciosa, identificando e conhecendo o modelo de dados, bem como, detalhes dos elementos e atributos fornecidos pelos de dados das fontes escolhidas.

5.1.3 Estruturação

Esta fase que tem a finalidade de definir ou reutilizar ontologia, o conhecimento do cenário, tal como, dos dados que irão ser utilizados para criação dos *datasets* facilitou a definição dos conceitos deste domínio.

Neste contexto, foi realizada a modelagem de duas ontologias com domínios bem parecidos, buscando utilizar e a semelhar o máximo as duas ontologias, principalmente vocabulários, para que ao realizar a interligação dos dados, não se tenha problemas de identidades dos recursos. Ou seja, desde o início destas modelagens, buscou-se facilitar o máximo a interligação dos dados.

Os conceitos dos domínio selecionado para integrar os esquemas das duas ontologias, a do TCU e a da CGU, foram os seguintes:

- **Pessoa Física**. Pessoa física com nome declarado pelo Tribunal de Contas da União(TCU), como inabilitados para o exercício de cargo em comissão ou função de confiança no âmbito da Administração Pública Federal, nos termos do art. 60 da Lei nº 8.443/92 (LOTCU).

- **Organização**. Organização com nome inidôneo para participar de licitações realizadas pela Administração Pública Federal, nos termos do art. 46 da Lei nº 8.443/92 (LOTCU).

- **Restrição**. Esse termo detalha quais as restrições de Organizações e/ou pessoas, em

relação aos Órgãos públicos. Exemplo: Especificar qual processo e intervalo de tempo que restringe a pessoa ou empresa.

- **Restritor**. Esse termo detalha as informações do Órgão Publico responsável por realizar as restrições. Por exemplo: TCU e/ou CGU.

- **Provenance**. Esse termo detalha informações de metadados dos dados fontes, como por exemplo, o nome do arquivo e a data que foi realizado o *download*.

Após a definição dos conceitos incluídos nos esquemas das ontologias, buscou-se analisar vocabulários conhecidos e estáveis com o objetivo de reaproveitar o maior número de termos possível. Além da procura de classes existentes com definições semânticas equivalentes a estes conceitos, houve também uma pesquisa por propriedades que representem bem as relações entre estas classes.

Consultando tanto vocabulários de caráter geral como vocabulários com um foco específico de Pessoas, Organizações, Restrição, Restritor e Provenance, um total de 2 classes tiveram vocabulário reutilizados e as outras foram criadas. A Quadro 6 mostra quais termos foram reusados (prefixados com o *namespace* do respectivo vocabulário).

Quadro 6 – Termos reusados para classes

Conceito	**Termo reusado**
Pessoa	foaf:Person
Organização	foaf:Organization

Além destas classes, várias propriedades também foram reutilizadas e para facilitar o entendimento da aplicação destas propriedades nos esquemas deste trabalho, lista-se na Quadro 7.

Quadro 7 – Termos reusados para propriedades

Prefix	**Termo reusado**
foaf:name	http://xmlns.com/foaf/0.1/name
dbr:cnpj	http://pt.dbpedia.org/property/cnpj
dbr:cpf	http://pt.dbpedia.org/property/cpf
dbr:Filename	http://pt.dbpedia.org/property/filename
time:date	http://www.w3.org/2006/time#date
time:startDate	http://www.w3.org/2006/time#startDate
time:endDate	http://www.w3.org/2006/time#endDate
time:hour	http://www.w3.org/2006/time#hour
rdf:id	http://www.w3.org/1999/02/22-rdf-syntax-ns#id
rdfs:label	http://www.w3.org/2000/01/rdf-schema#label
dcterms:source	http://purl.org/dc/terms/source

Apesar dos vocabulários existentes cobrirem uma ampla parte dos termos necessários para a criação de das ontologias, ainda foram necessários definir alguns novos termos, já que em alguns casos não foi encontrado nenhum termo existente refletindo a semântica do conceito que se desejava representar. Desta forma, um conjunto de termos foi criado com o objetivo de complementar o esquema da ontologia.

O vocabulário destinado ao Tribunal de Contas da União, tem como *namespace* o termo tcu, que durante este trabalho é será representado por http://arida.ufc.br/tcu/vocab/resource, que possua vez, estão destacados na Quadro 8.

Quanto ao vocabulário destinado a Controladoria Geral da União, tem como *namespace* o termo cgu, que durante este trabalho é será representado por http://arida.ufc.br/cgu/vocab/resource, que possua vez, estão destacados na Quadro 9.

Alguns vocabulários possuem propriedades que ligam uma pessoa e/ou uma organização a uma restrição, e uma restrição a um restritor.

Ao invés de criar termos para todo o domínio em questão, nesta fase o foco foi na criação apenas dos terfig:transformacaoTCUs essenciais ao esquema projetado e, até então, inexistentes.

Quadro 8 – Termos criados para classes relacionados a TCU

Classe	Descricao
tcu:Restricao	Classe responsável por mostrar o que restringe a pessoa ou Organização, como qual restrição e periodo desta
tcu:Provanance	Classe responsável por organizar informações de metadados, como nome do arquivo e outros
tcu:Restritor	Órgão Publico responsável por realizar restrição devido alguma irregularidade.

Quadro 9 – Termos criados para classes relacionados a CGU

Classe	Descricao
cgu:Restricao	Classe responsável por mostra o que restringe a pessoa ou Organização, como qual restrição e periodo desta
cgu:Provanance	Classe responsável por organizar informações de metadados, como nome do arquivo e outros
cgu:Restritor	Órgão Publico responsável por realizar restrição devido alguma irregularidade.

Quadro 10 – Termos criados para *Datatype Properties* do TCU

Propriedade	Domain	Range
tcu:númeroProcesso	tcu:Restricao	String
tcu:tipoRestricao	tcu:Restricao	String

Quadro 11 – Termos criados para *Datatype Properties* da CGU

Propriedade	Domain	Range
cgu:númeroProcesso	cgu:Restricao	String
cgu:tipoRestricao	cgu:Restricao	String

A ferramenta utilizada para a criação dos vocabulários foi o Protégé[5]. E também as ontologias foram criadas em OWL, com esta mesma ferramenta.

[5] http://protege.stanford.edu/

Quadro 12 – Termos criados para *Object Properties* do TCU

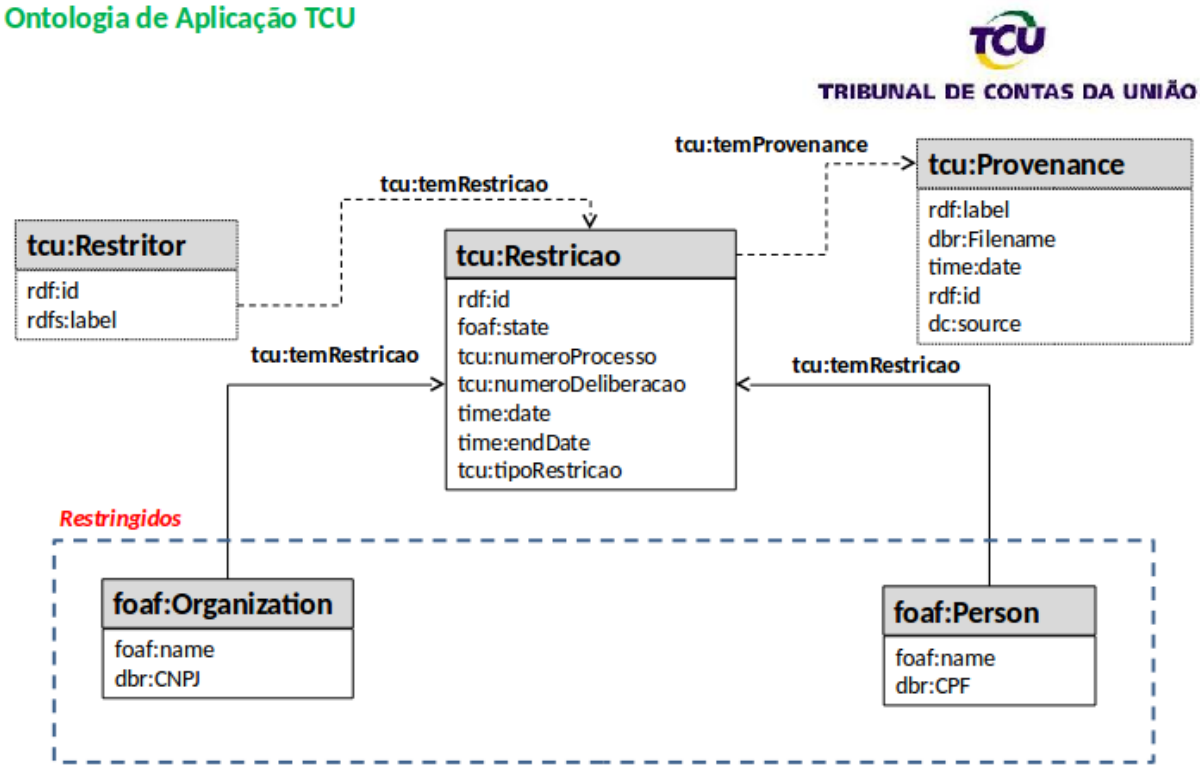
Propriedade	Domain	Range
tcu:temRestricao	foaf:Organization	tcu:Restricao
tcu:temRestricao	foaf:Person	tcu:Restricao
tcu:temProvenance	tcu:Restricao	tcu:Provenance

Quadro 13 – Termos criados para *Object Properties* da CGU

Propriedade	Domain	Range
cgu:temRestricao	foaf:Organization	cgu:Restricao
cgu:temRestricao	foaf:Person	cgu:Restricao
cgu:temProvenance	cgu:Restricao	cgu:Provenance

Com a criação destes termos, o esquema da ontologia do TCU foi finalizado. Este esquema serviu como base para a criação do *dataset* RDF. A Figura 16 mostra o esquema final do TCU.

Figura 16 – Ontologia de aplicação do TCU

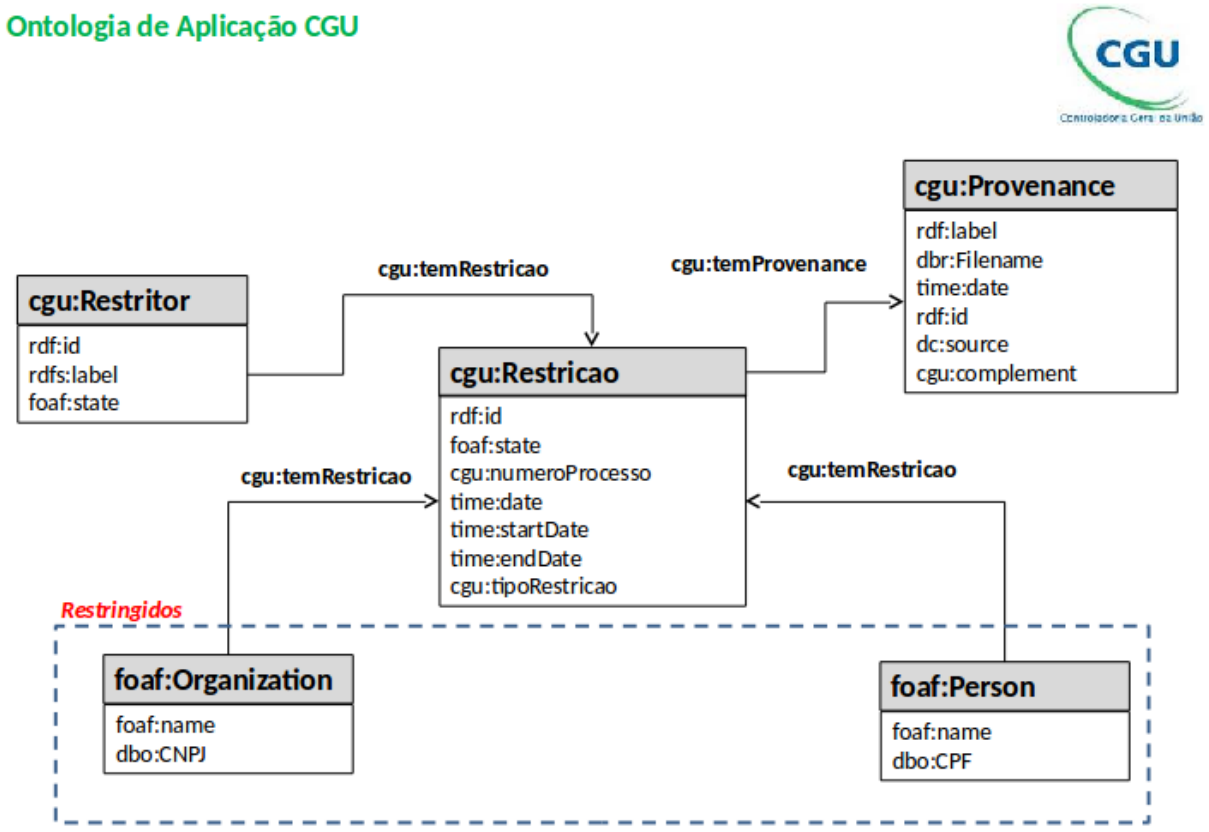


Quanto a criação destes termos, o esquema da ontologia da CGU, também foi finalizado. O esquema serviu como base para a criação do *dataset* RDF. A Figura 17 mostra o esquema final do CGU.

5.1.4 Mapeamento de vocabulários fonte (source) para vocabulários destino (target)

Nesta fase começa-se a direcionar as atividades para a ferramenta que foram utilizadas para a transformação dos dados. Logo, chegando ao Pentaho Data Integration(Kettle) e os

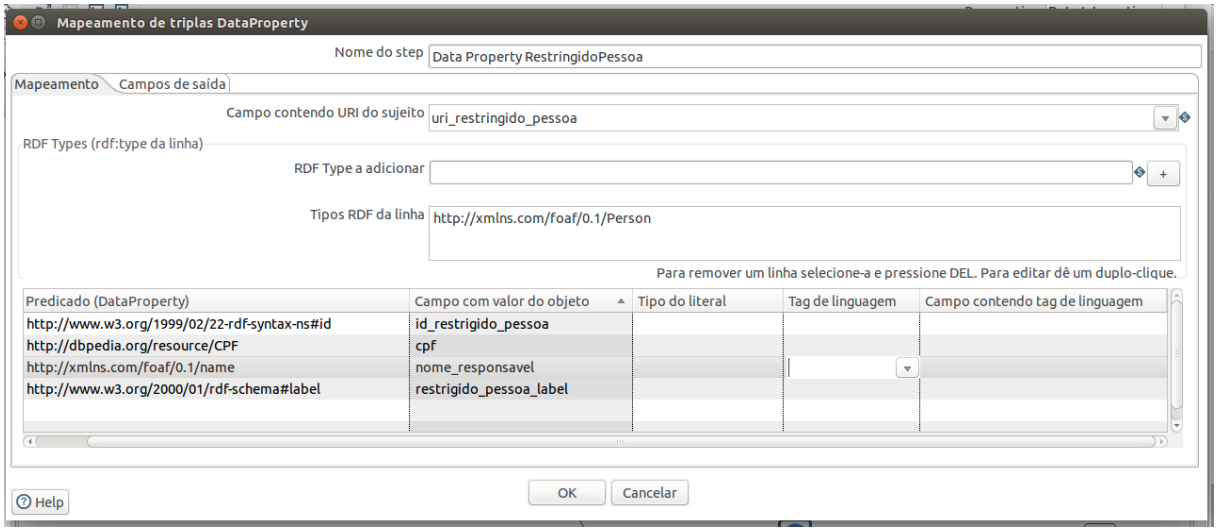
Figura 17 – Ontologia de aplicação da CGU



plugins do projeto LinkedDataBr, o *step Data Property Mapping* oferece a capacidade de mapear, a partir das linhas do fluxo de entrada, os componentes de uma tripla RDF (sujeito, predicado e objeto) nas linhas do fluxo de saída, sendo o objeto um valor literal. Cada linha de entrada deve conter um campo com a URI que identifica o recurso do sujeito. Na configuração do *step*, é necessário definir uma ou mais URIs especificando o tipo do recurso e um mapeamento de propriedades do tipo literal, informando a URI da propriedade e o campo da linha de entrada que contém o valor da propriedade. Além disso, é possível definir, no mapeamento, o tipo do literal e a marcação de idioma.

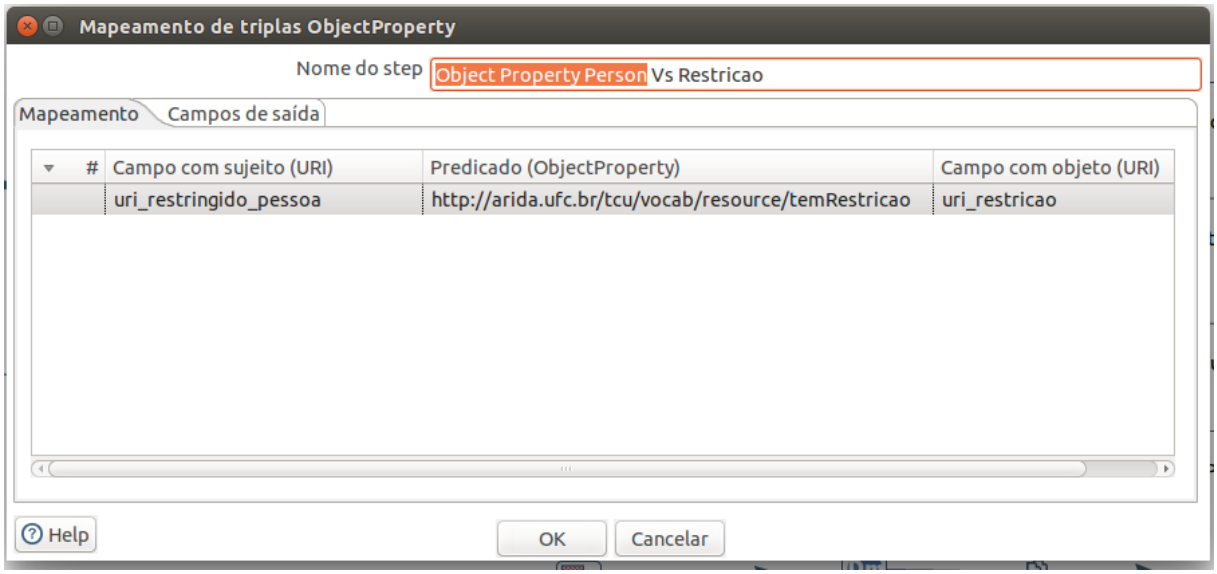
O *step Object Property Mapping* é similar ao *step Data Property Mapping*, com a diferença de que o valor do objeto enviado no fluxo de saída é uma URI de um recurso. A configuração deste step é mais simples e consiste em indicar o campo da linha de entrada que contem a URI do sujeito, o campo da linha de entrada que contem a URI do objeto e definir a URI da propriedade. A Figura 18 ilustra a interface de configuração do *step Data Property Mapping*, com a especificação do tipo e do mapeamento das propriedades id, cpf, nome e label do recurso Person e a Figura 19 ilustra a interface de configuração do *step Object Property Mapping*, com a especificação do mapeamento do relacionamento da propriedade Person e Restrição.

Figura 18 – Data Property Mapping Person



Fonte: Pentaho Data Integration(Kettle).

Figura 19 – Object Property Mapping Person vs Restrição



Fonte: Pentaho Data Integration(Kettle).

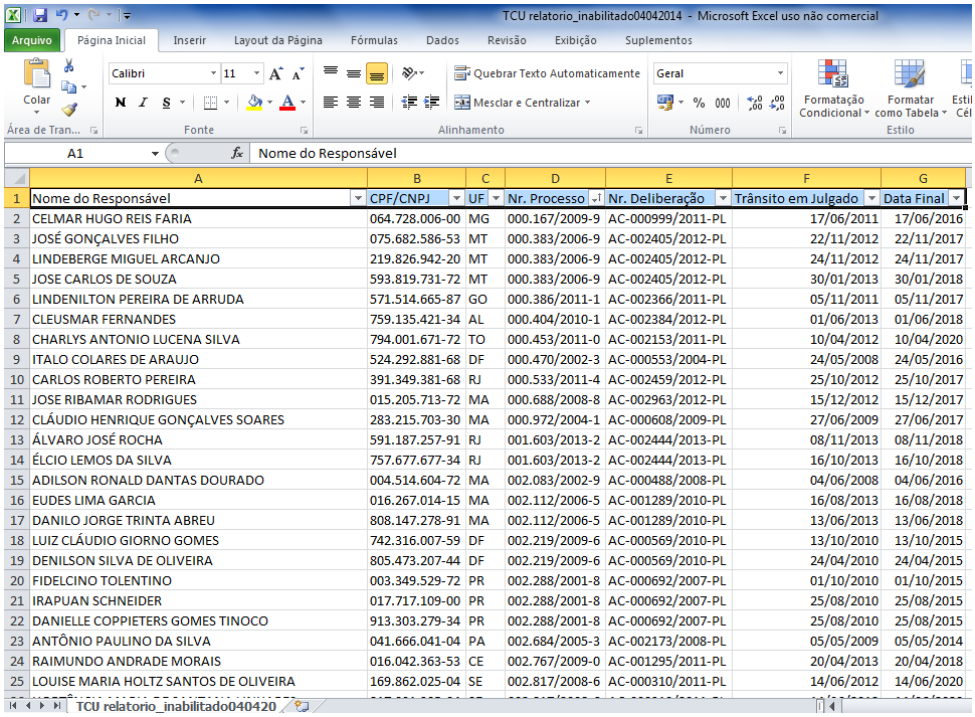
5.1.5 Coleta de dados

Nesta fase, os dados coletados vieram das duas fontes selecionadas anteriormente. A coleta foi realizada diretamente no site do TCU e do Portal da Transparência, através de *downloads* dos arquivos no formato CSV, com endereço especifico para inabilitados[6] e inidôneos[7,8]. Além disso, usou-se a nomenclatura recomendada para arquivos no Apêndice A e armazenou-se

6 https://contas.tcu.gov.br/ords/f?p=2046:4::CSV
7 https://contas.tcu.gov.br/ords/f?p=2046:5::CSV
8 http://www.portaltransparencia.gov.br/downloads/snapshot.asp?c=CEIS#get

temporariamente os dados enquanto eram preparados e realizadas as próximas fases, que são detalhadas a seguir. As Figuras 20 e 21 mostram dois arquivos que exemplificam os dados coletados.

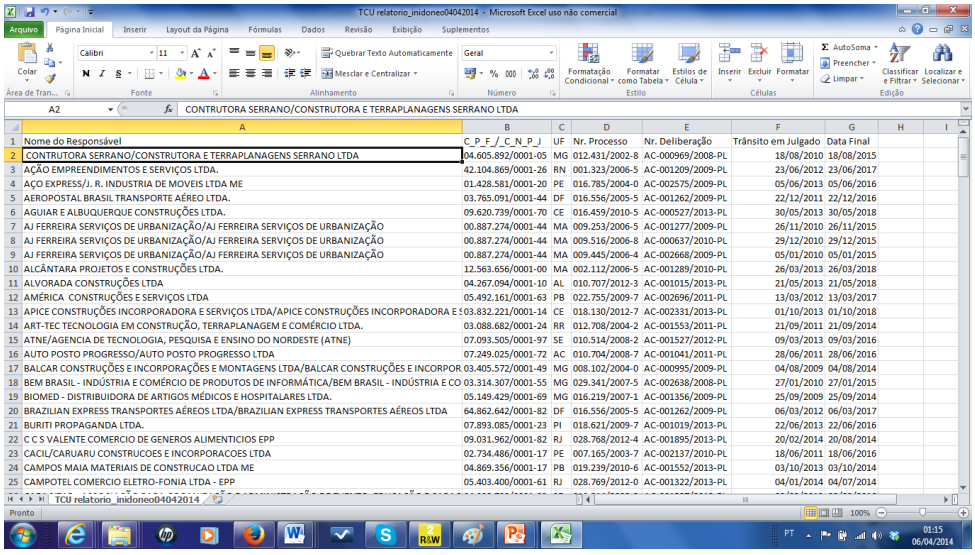
Figura 20 – Relatório de Pessoas Físicas e Jurídicas Inabilitadas



	Nome do Responsável	CPF/CNPJ	UF	Nr. Processo	Nr. Deliberação	Trânsito em Julgado	Data Final
2	CELMAR HUGO REIS FARIA	064.728.006-00	MG	000.167/2009-9	AC-000999/2011-PL	17/06/2011	17/06/2016
3	JOSÉ GONÇALVES FILHO	075.682.586-53	MT	000.383/2006-9	AC-002405/2012-PL	22/11/2012	22/11/2017
4	LINDEBERGE MIGUEL ARCANJO	219.826.942-20	MT	000.383/2006-9	AC-002405/2012-PL	24/11/2012	24/11/2017
5	JOSE CARLOS DE SOUZA	593.819.731-72	MT	000.383/2006-9	AC-002405/2012-PL	30/01/2013	30/01/2018
6	LINDENILTON PEREIRA DE ARRUDA	571.514.665-87	GO	000.386/2011-1	AC-002366/2011-PL	05/11/2011	05/11/2017
7	CLEUSMAR FERNANDES	759.135.421-34	AL	000.404/2010-1	AC-002384/2012-PL	01/06/2013	01/06/2018
8	CHARLYS ANTONIO LUCENA SILVA	794.001.671-72	TO	000.453/2011-0	AC-002153/2011-PL	10/04/2012	10/04/2020
9	ITALO COLARES DE ARAUJO	524.292.881-68	DF	000.470/2002-3	AC-000553/2004-PL	24/05/2008	24/05/2016
10	CARLOS ROBERTO PEREIRA	391.349.381-68	RJ	000.533/2011-4	AC-002459/2012-PL	25/10/2012	25/10/2017
11	JOSE RIBAMAR RODRIGUES	015.205.713-72	MA	000.688/2008-8	AC-002963/2012-PL	15/12/2012	15/12/2017
12	CLÁUDIO HENRIQUE GONÇALVES SOARES	283.215.703-30	MA	000.972/2004-1	AC-000608/2009-PL	27/06/2009	27/06/2017
13	ÁLVARO JOSÉ ROCHA	591.187.257-91	RJ	001.603/2013-2	AC-002444/2013-PL	08/11/2013	08/11/2018
14	ÉLCIO LEMOS DA SILVA	757.677.677-34	RJ	001.603/2013-2	AC-002444/2013-PL	16/10/2013	16/10/2018
15	ADILSON RONALD DANTAS DOURADO	004.514.604-72	MA	002.083/2002-9	AC-000488/2008-PL	04/06/2008	04/06/2016
16	EUDES LIMA GARCIA	016.267.014-15	MA	002.112/2006-5	AC-001289/2010-PL	16/08/2013	16/08/2018
17	DANILO JORGE TRINTA ABREU	808.147.278-91	MA	002.112/2006-5	AC-001289/2010-PL	13/06/2013	13/06/2018
18	LUIZ CLÁUDIO GIORNO GOMES	742.316.007-59	DF	002.219/2009-6	AC-000569/2010-PL	13/10/2010	13/10/2015
19	DENILSON SILVA DE OLIVEIRA	805.473.207-44	DF	002.219/2009-6	AC-000569/2010-PL	24/04/2010	24/04/2015
20	FIDELCINO TOLENTINO	003.349.529-72	PR	002.288/2001-8	AC-000692/2007-PL	01/10/2010	01/10/2015
21	IRAPUAN SCHNEIDER	017.717.109-00	PR	002.288/2001-8	AC-000692/2007-PL	25/08/2010	25/08/2015
22	DANIELLE COPPIETERS GOMES TINOCO	913.303.279-34	PR	002.288/2001-8	AC-000692/2007-PL	25/08/2010	25/08/2015
23	ANTÔNIO PAULINO DA SILVA	041.666.041-04	PA	002.684/2005-3	AC-002173/2008-PL	05/05/2009	05/05/2014
24	RAIMUNDO ANDRADE MORAIS	016.042.363-53	CE	002.767/2009-0	AC-001295/2011-PL	20/04/2013	20/04/2018
25	LOUISE MARIA HOLTZ SANTOS DE OLIVEIRA	169.862.025-04	SE	002.817/2008-6	AC-000310/2011-PL	14/06/2012	14/06/2020

Fonte: Tribunal de Contas da União.

Figura 21 – Relatório de Pessoas Físicas e Jurídicas Inidôneas



	Nome do Responsável	C_P_F_/_C_N_P_J	UF	Nr. Processo	Nr. Deliberação	Trânsito em Julgado	Data Final
2	CONTRUTORA SERRANO/CONSTRUTORA E TERRAPLANAGENS SERRANO LTDA	04.605.892/0001-05	MG	012.431/2002-8	AC-000969/2008-PL	18/08/2010	18/08/2015
3	AÇÃO EMPREENDIMENTOS E SERVIÇOS LTDA.	42.104.869/0001-26	RN	001.323/2006-5	AC-001209/2009-PL	23/06/2012	23/06/2017
4	AÇO EXPRESS/J. R. INDUSTRIA DE MOVEIS LTDA ME	01.428.581/0001-20	PE	016.785/2004-0	AC-002575/2009-PL	05/06/2013	05/06/2016
5	AEROPOSTAL BRASIL TRANSPORTE AÉREO LTDA.	03.765.091/0001-44	DF	016.556/2005-5	AC-001262/2009-PL	22/12/2011	22/12/2016
6	AGUIAR E ALBUQUERQUE CONSTRUÇÕES LTDA.	09.620.739/0001-70	CE	016.459/2010-5	AC-000527/2013-PL	30/05/2013	30/05/2018
7	AJ FERREIRA SERVIÇOS DE URBANIZAÇÃO/AJ FERREIRA SERVIÇOS DE URBANIZAÇÃO	00.887.274/0001-44	MA	009.253/2006-5	AC-001277/2009-PL	26/11/2010	26/11/2015
8	AJ FERREIRA SERVIÇOS DE URBANIZAÇÃO/AJ FERREIRA SERVIÇOS DE URBANIZAÇÃO	00.887.274/0001-44	MA	009.516/2006-8	AC-000637/2010-PL	29/12/2010	29/12/2015
9	AJ FERREIRA SERVIÇOS DE URBANIZAÇÃO/AJ FERREIRA SERVIÇOS DE URBANIZAÇÃO	00.887.274/0001-44	MA	009.445/2006-4	AC-002668/2009-PL	05/01/2010	05/01/2015
10	ALCÂNTARA PROJETOS E CONSTRUÇÕES LTDA.	12.563.656/0001-00	MA	002.112/2006-5	AC-001289/2010-PL	26/03/2013	26/03/2018
11	ALVORADA CONSTRUÇÕES LTDA	04.267.094/0001-10	AL	010.707/2012-3	AC-001015/2013-PL	21/05/2013	21/05/2018
12	AMÉRICA CONSTRUÇÕES E SERVIÇOS LTDA	05.492.161/0001-63	PB	022.755/2009-7	AC-002696/2011-PL	13/03/2012	13/03/2017
13	APICE CONSTRUÇÕES INCORPORADORA E SERVIÇOS LTDA/APICE CONSTRUÇÕES INCORPORADORA E	03.832.221/0001-14	CE	018.130/2012-7	AC-002331/2013-PL	01/10/2013	01/10/2018
14	ART-TEC TECNOLOGIA EM CONSTRUÇÃO, TERRAPLANAGEM E COMÉRCIO LTDA.	03.088.682/0001-24	RR	012.708/2004-2	AC-001553/2011-PL	21/09/2011	21/09/2014
15	ATNE/AGENCIA DE TECNOLOGIA, PESQUISA E ENSINO DO NORDESTE (ATNE)	07.093.505/0001-97	SE	010.514/2008-2	AC-001527/2012-PL	09/03/2013	09/03/2016
16	AUTO POSTO PROGRESSO/AUTO POSTO PROGRESSO LTDA	07.249.025/0001-72	AC	010.704/2008-7	AC-001041/2011-PL	28/06/2011	28/06/2016
17	BALCAR CONSTRUÇÕES E INCORPORAÇÕES E MONTAGENS LTDA/BALCAR CONSTRUÇÕES E INCORPOR	03.405.572/0001-49	MG	008.102/2004-0	AC-000995/2009-PL	04/08/2009	04/08/2014
18	BEM BRASIL - INDÚSTRIA E COMÉRCIO DE PRODUTOS DE INFORMÁTICA/BEM BRASIL - INDÚSTRIA E CO	03.314.307/0001-55	MG	029.341/2007-5	AC-002638/2008-PL	27/01/2010	27/01/2015
19	BIOMED - DISTRIBUIDORA DE ARTIGOS MÉDICOS E HOSPITALARES LTDA.	05.149.429/0001-69	MG	016.219/2007-1	AC-001356/2009-PL	25/09/2009	25/09/2014
20	BRAZILIAN EXPRESS TRANSPORTES AÉREOS LTDA/BRAZILIAN EXPRESS TRANSPORTES AÉREOS LTDA	64.862.642/0001-82	DF	016.556/2005-5	AC-001262/2009-PL	06/03/2012	06/03/2017
21	BURITI PROPAGANDA LTDA.	07.893.085/0001-23	PI	018.621/2009-7	AC-001019/2013-PL	22/06/2013	22/06/2016
22	C C S VALENTE COMERCIO DE GENEROS ALIMENTICIOS EPP	09.031.962/0001-82	RJ	028.768/2012-4	AC-001895/2013-PL	20/02/2014	20/08/2014
23	CACIL/CARUARU CONSTRUCOES E INCORPORACOES LTDA	02.734.486/0001-17	PE	007.165/2003-7	AC-002137/2010-PL	18/06/2011	18/06/2016
24	CAMPOS MAIA MATERIAIS DE CONSTRUCAO LTDA ME	04.869.356/0001-17	PB	019.239/2010-6	AC-001552/2013-PL	03/10/2013	03/10/2014
25	CAMPOTEL COMERCIO ELETRO-FONIA LTDA - EPP	05.403.400/0001-61	RJ	028.769/2012-0	AC-001322/2013-PL	04/01/2014	04/07/2014

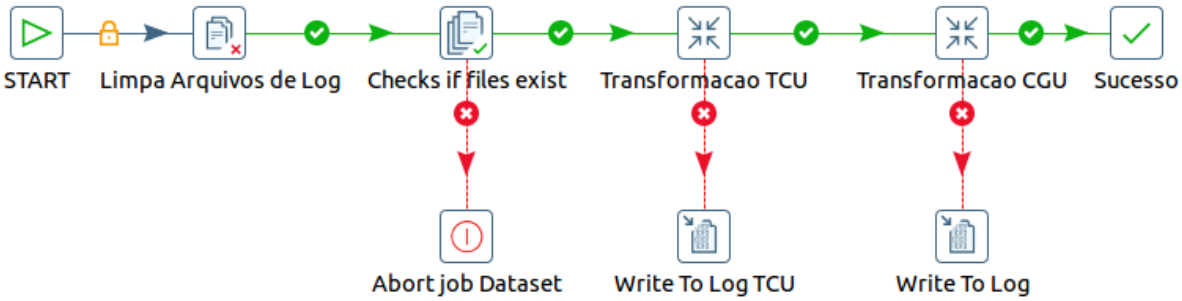
Fonte: Tribunal de Contas da União.

5.1.6 Refinamento, Transformação, Armazenamento e Publicação

Seguindo o fluxo do processo, para o exemplo aplicado, será apresentado o *workflow* ETL que utilizou-se para publicar como *Linked Data* as duas fontes de dados do TCU e uma fonte de dados da CGU. O workflow ETL, apresentado a seguir, será detalhado adiante de acordo com seguintes fases definidas no *Triplify Process*: refinamento, transformação, armazenamento e publicação.

A Figura 22 ilustra o workflow ETL encapsulado, implementado com um job do Kettle, e a Figura 23 e a Figura 24 ilustram as transformações responsáveis por publicar os dados das fontes de dados do Tribunal de Contas da União(TCU) e da Controladoria Geral da União (CGU).

Figura 22 – Job TCU-CGU: Implementação do workflow ETL que publica como *Linked Data* duas fontes de dados do TCU e uma fonte de dados da CGU



Fonte: Pentaho Data Integration(Kettle).

O Job TCU-CGU apresentado na Figura 22 ao ser executado segue o seguinte fluxo:

1. Confirma a existência dos arquivos de origem necessário ao processo de publicação, caso contrário, ele aborta a execução.

2. Apaga os arquivos de logs.

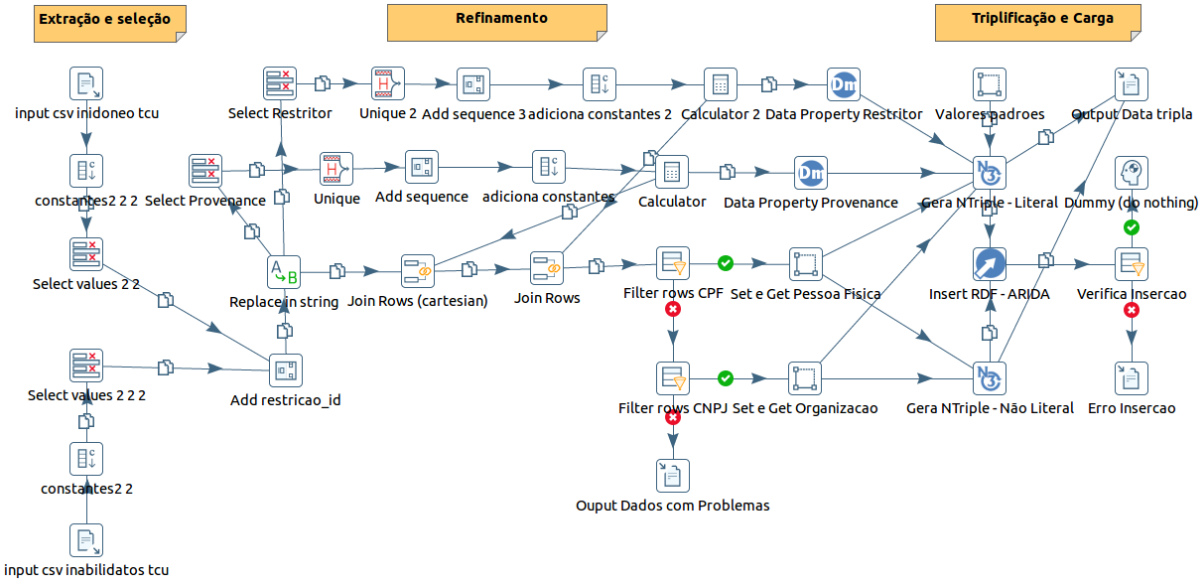
3. Permite a execução integrada das transformações do TCU e da CGU, que constituem o processo de publicação e em caso de erros durante as transformações escrevem os logs em arquivos.

Quanto as transformações do TCU e da CGU apresentadas na Figura 23 e na Figura 24 realizam as seguintes atividades:

1. Extrai os dados dos arquivos fontes.

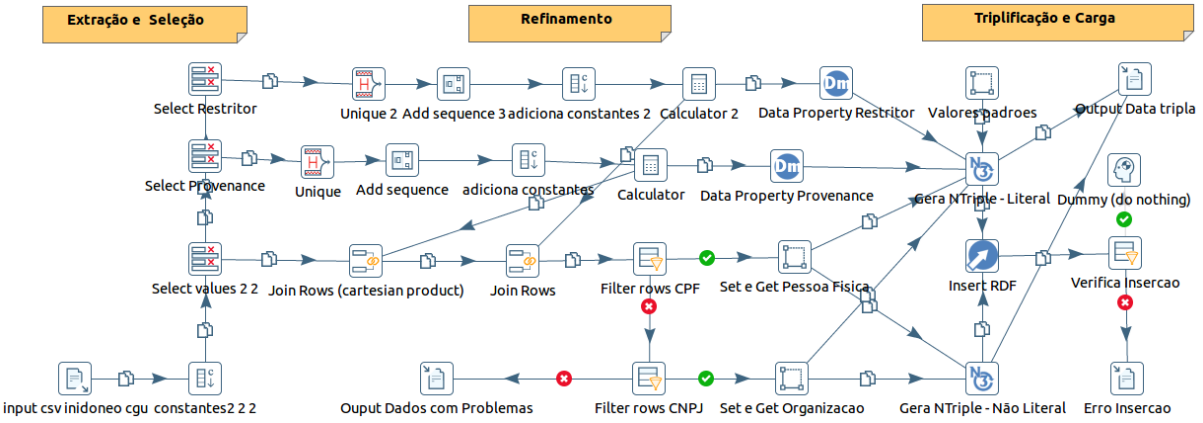
2. Seleciona e renomeia os campos lidos dos arquivos de origem para os que foram descritos na ontologia.

Figura 23 – Transformação transf_cgu: Implementação de uma transformação ELT responsável por publicar como *Linked Data* a fonte de dados do Tribunal de Contas da União(TCU).



Fonte: Pentaho Data Integration.

Figura 24 – Transformação transf_cgu: Implementação de uma transformação ELT responsável por publicar como *Linked Data* a fonte de dados da Controladoria Geral da União(CGU).



Fonte: Pentaho Data Integration.

3. Formata os valores do campo CPF_CNPJ e depois os separam em campos diferentes, através de do filtro *Filter Rows* CPF e o *Filter Rows* CNPJ.

4. Seguindo, realiza-se a chamada da sub-transformação através do *step* Set e Get Pessoa Física, que é representado na Figura 25 que realiza as seguintes atividades:

    a) Anota os dados recuperados com os conceitos presentes no vocabulário da ontologia

de referência.

b) Formata URI a ser utilizada nos valores selecionados.

Ou seja, a sub-transformação realiza o mapeamento dos dados referentes a Pessoa Física, Restritor, Restrição, Provenance, bem como, os relacionamentos destes. Ao realizar os mapeamentos a sub-transformação retorna os dados para a transformação que a chamou, também através do mesmo *step* Set e Get Pessoa Física.

5. Seguindo, realiza-se a chamada a outra sub-transformação através do *step* Set e Get Organização, que é representado nas Figuras 25 e 26 que realizam

   as seguintes atividades:

   a) Anota os dados recuperados com os conceitos presentes no vocabulário da ontologia de referência.

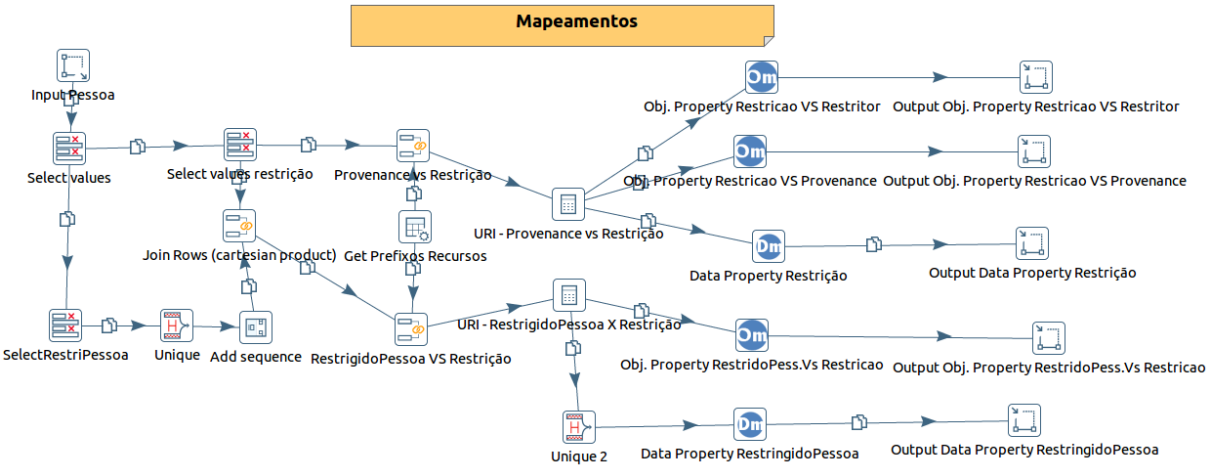
   b) Formata URI a ser utilizada nos valores selecionados.

   Logo, a sub-transformação realiza o mapeamento dos dados referentes a Organização, Restritor, Restrição, Provenance, bem como, os relacionamentos destes. Ao realizar os mapeamentos a sub-transformação retorna os dados para a transformação que a chamou, também através do mesmo *step* Set e Get Organizacao.

6. Tranforma dados recebidos das sub-transformações para o formato de triplas através de dois *steps* o Gera NTriple Literal e o Gera NTriple - Não Literal.

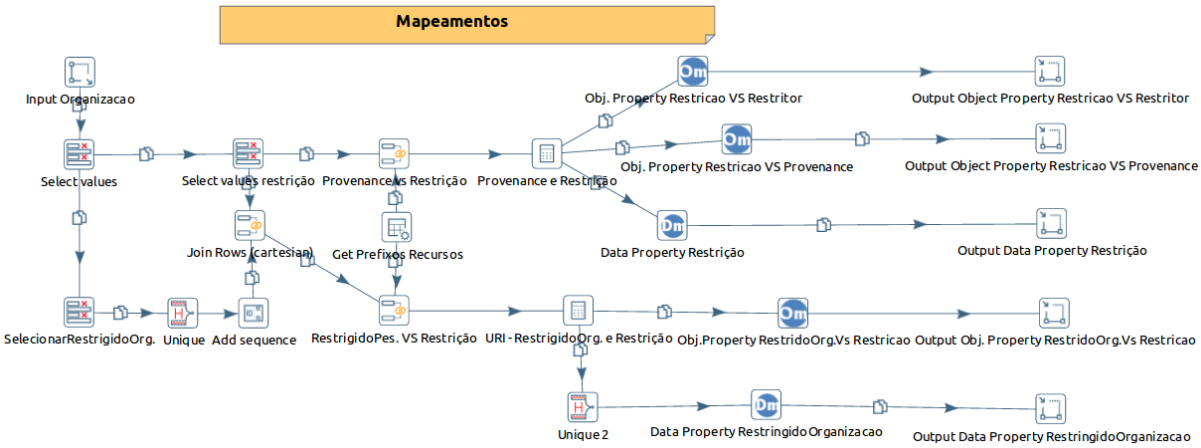
7. Carrega os dados no *Endpoint* da RDF *store* e no arquivo, através dos respectivos *steps* nomeados como *Insert* RDF e o *Output Data* tripla.

Figura 25 – Sub-transformação sub_transformacao_pessoa_fisica: Implementação de uma sub-transformação ELT responsável pelo mapeamentos e URIs de Pessoa Restringida, Restrição, Restritor



Fonte: Pentaho Data Integration.

Figura 26 – Sub-transformação sub_transformacao_organizacao: Implementação de uma sub-transformação ELT responsável pelo mapeamentos e URIs de Organização Restringida, Restrição, Restritor



5.1.6.1 Refinamento

Nesta sexta fase, realizou-se atividades com os dados de origem com o intuito de melhorar a sua qualidade, para isso, usou-se e aproveitou-se os diversos recursos do Pentaho Data Integration (Kettle). Para ser mais específicos, ao importar os dados para Kettle, foram selecionados e renomeados alguns atributos dos dados de origem e começaram as atividades relacionadas a qualidade, entre eles, adicionamos um identificador para cada recurso utilizando o *step* add seguence, formatando o atributo CPF_CNPJ deixando apenas os números, para isso utilizou-se o *step replace* String com uma expressão regular, e por fim, utilizou-se um filtro neste mesmo campo CPF_CNPJ para separar os que tinha onze dígitos no caso de CPF e quartoze para CNPJ e arquivou os que não atendessem essa filtragem. Logo, melhorou-se a qualidade dos dados retirando caracteres indesejáveis no CPF ou CNPJ que dificultaria uma comparação por exemplo, bem como, retirada de elementos que não atendesse a filtragem. Um problema que estava comprometendo a qualidade dos dados era a codificação *Unicode* e alinhou-se todos os arquivos para UTF8.

5.1.6.2 Transformação

Nesta fase, os dados já estão no fluxo de trabalho do kettle, e já tinham passo por todas as outras fases que deixaram os dados preparados e qualificados para transformação. Com o auxílio dos *steps Data Property*, *Object Property* e NTriple, os dados são transformados em triplas RDF, no formato N-TRIPLE e na medida que são gerados podem ser carregados em um determinado arquivo ou armazenados diretamente em um RDF *store*.

5.1.6.3 Armazenamento e Publicação

Nesta fase os dados das duas fontes foram publicados no banco de triplas Virtuoso com o seguinte endereço válido <virtuoso.mooo.com/sparql>, organizados em dois grafos

RDF: o primeiro grafo, nomeado http://arida.ufc.br/tcu, contém as triplas RDF do Tribunal de Contas da União e o segundo grafo, nomeado http://arida.ufc.br/cgu, contém as triplas RDF da Controladoria Geral da União. Na Figura 23 e na Figura 24 os *steps* SPARQL *Update Ouput* nomeados respectivamente, como Insert RDF, recebem as configurações de conexões do banco de triplas e o grafo especificado para realizar a inserção dos dados. Durante a inserção das triplas nos seus respectivos grafos, caso ocorra algum problema, usamos os *steps* do Data Integration, para verificar e guardar as triplas que não foram inseridas, em um arquivo, para serem inseridas em um outro momento. Os *steps* utilizados foram chamados de Verifica Insercao e Erro Insercao. Paralelamente com a inserção dos dados no banco triplas, utilizamos outro *step*, o *Text file output* que nomeamos como *Output Data tripla*, para salvamos todas as triplas em um arquivo especificado e com o formato N-TRIPLE, para eventuais necessidades.

5.1.7 Enriquecimento

A nona fase que consiste no enriquecimento dos dados através da interligação e integração. A primeira atividade realizada foi identificar diferentes URIs que são usadas em diferentes fontes de dados para identificar a mesma entidade no mundo real(interligação).

Para interligação do *dataset* do TCU e da CGU, utilizou-se o *framework* Silk, que fornece os recursos necessários para essa atividade. Para isso, foi preciso selecionar propriedades discriminativas para comparação, como:

- Aplicar transformações nos dados para normalizar valores de propriedade antes de comparação.
- Aplicar medidas de distância combinada com adequada limiares distância.
- Agregar o resultado de múltiplas comparações utilizando linear, bem como funções de agregação não-lineares.

Para realizar as transformações dos valores dos dados, são utilizados um conjunto de operadores, fornecidos pela ferramenta, como: categoria de normalização que tem elementos como retirar caracteres especiais, transformar texto em maiúsculas(Upper Case), categoria de *replace* que aceita expressões regulares, entre outras que podem ser utilizadas de acordo com a necessidade.

As medidas de distância com seus devidos limiares que servem para comparação de elementos como strings, datas, possuem diversas categorias como: charecterbased que fornece alguns algoritmos como Levenshtein *distance*, Jaro *distance* e outros. A categoria tokenbased, por exemplo, possui alguns algoritmos de comparação bem conhecidos, como: Jaccard, Softjaccard e outros.

Com o resultados dos algoritmos de distância, realiza-se as agregações, algumas das opções são: *maximum*, *minimum*, *average* e outras. Portanto, os *links* de referência de que em um *dataset* um recurso é o mesmo recurso em outro *dataset* é determinado pelo conjunto dessas propriedades, transformações, limiares de distância e comparações.

A ferramenta Silk acessa as fontes de dados via protocolo SPARQL, permitindo então o acesso tanto *Endpoints* SPARQL (remoto), como localmente.

Além disso, o Silk tem dois paradigmas:

- Silk Workbench que é uma aplicação web que guia o usuário através do processo de interligar diferentes fontes de dados. As principais características oferecidas são:
  - Ele permite ao usuário gerenciar diferentes conjuntos de fontes de dados, tarefas que ligam e tarefas de transformação.
  - Ele oferece um editor gráfico que permite ao usuário criar e editar ligando tarefas e tarefas de transformação facilmente.
  - Como encontrar uma boa heurística de ligação é geralmente um processo iterativo, além de torna possível o usuário a avaliar rapidamente os links que são gerados pela especificação da interligação atual.
  - Ele permite ao usuário criar e editar um conjunto de *links* de referência utilizados para avaliar a especificação atual do *link*.
- Silk Single Machine é utilizada para gerar ligações RDF em uma única máquina. Os conjuntos de dados que devem ser interligados pode residir na mesma máquina ou em máquinas remotas que são acessados através do protocolo SPARQL. Além disso, fornece multithreading, cache e o desempenho é reforçado usando o algoritmo de bloqueio multi-bloco. Usando uma linguagem de especificação baseada em XML declarativa, o usuário pode especifica os fontes de dados RDF e quais métodos de extração deve ser usadas. A saída resultante é um arquivo RDF contendo os valores das ligações dos recursos semelhates, entre as fontes comparadas.

As especificações das ligações entre as fontes podem ser criadas usando a interface gráfica do Silk Workbench ou manualmente no XML.

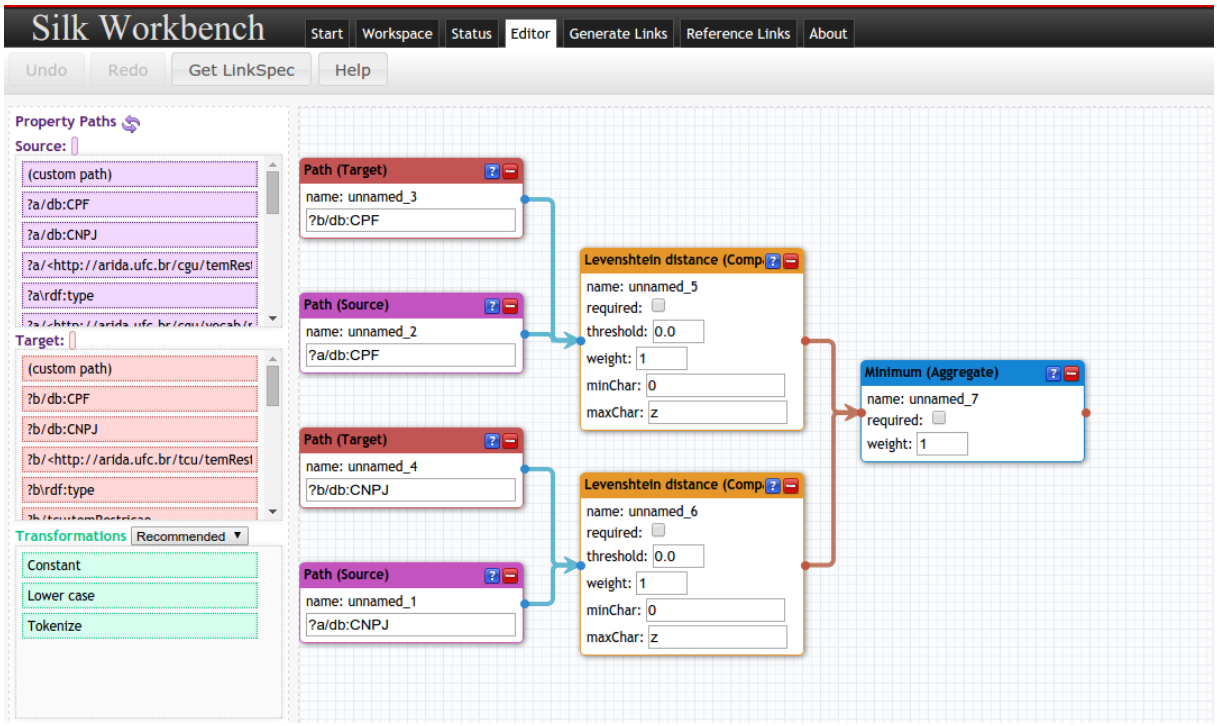
Voltando ao contexto do processo de triplificação, foi utilizado o Silk Workbench, para realizar a ligação dos *datasets* TCU e CGU criados neste estudo de caso. Para isso, identificamos elementos em comum nos dois *datasets* que servirão para a especificação das heurísticas para descobertas de same-as links, serão os seguintes:

- sameAs para Pessoas Físicas com o mesmo CPF completo e Organizações com o mesmo CNPJ completo(Heurística 1).
- sameAs para Pessoas Físicas com o mesmo CPF completo e nome. Organizações com o mesmo CNPJ completo e nome. Observação: o nome pode variar em quantidade de caracteres (Heurística 2).

Inicialmente realizou-se apenas a heurística para links sameAs para Pessoas Físicas com o mesmo CPF completo e Organizações com o mesmo CNPJ completo. A Figura 27

apresenta os elementos que foram utilizados na heurística. Recebeu-se os CPF *target* e *source*, passou-os para comparação de similaridade que foi utilizado o algoritmo Levenshtein *distance* que foi configurado para permitir apenas os valores que fossem iguais e depois para o agregador *minimum agregate*. Para as Organizações realizou-se as mesmas operações. Recebeu-se os CNPJ *target* e *source*, passou-os para comparação de similaridade que foi utilizado o algoritmo Levenshtein *distance* que foi configurado para permitir apenas os valores que fossem iguais e depois para o agregador *minimum agregate*.

Figura 27 – Linkage Rule Editor



Fonte: Silk Workbench.

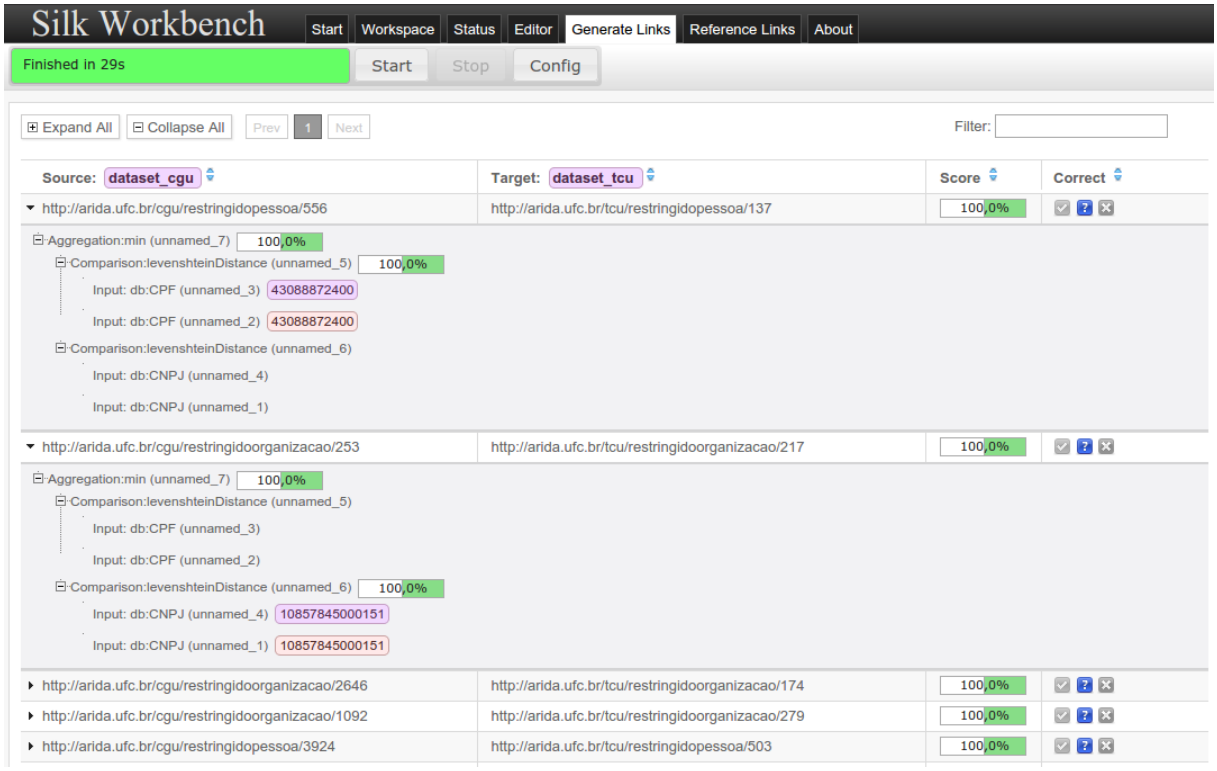
Definida a heurística com auxílio da interface do Silk Workbench, o próximo passo foi gerar os *links* de conexões entre as bases RDF escolhidas. A Figura 28 mostra quais elementos comparados e a porcentagem de similaridade utilizado.

O resultado da criação da heurísta 1, que considera apenas CPF completo e CNPJ completo, segue no Quadro 14 abaixo. Nessa heurística se considera como número de ligações, a soma do número de Pessoas Físicas e o número de Organizações similares entre os dois *datasets*, Tribunal de Contas da União(TCU) e a Controladoria Geral da União(CGU).

Quadro 14 – Estatística de Ligações na Heurística 1

Número de entidades do *dataset source* CGU	22498
Número de entidades do *dataset target* TCU	2690
Número de ligações	277

Figura 28 – Silk generate links



Fonte: Silk Workbench.

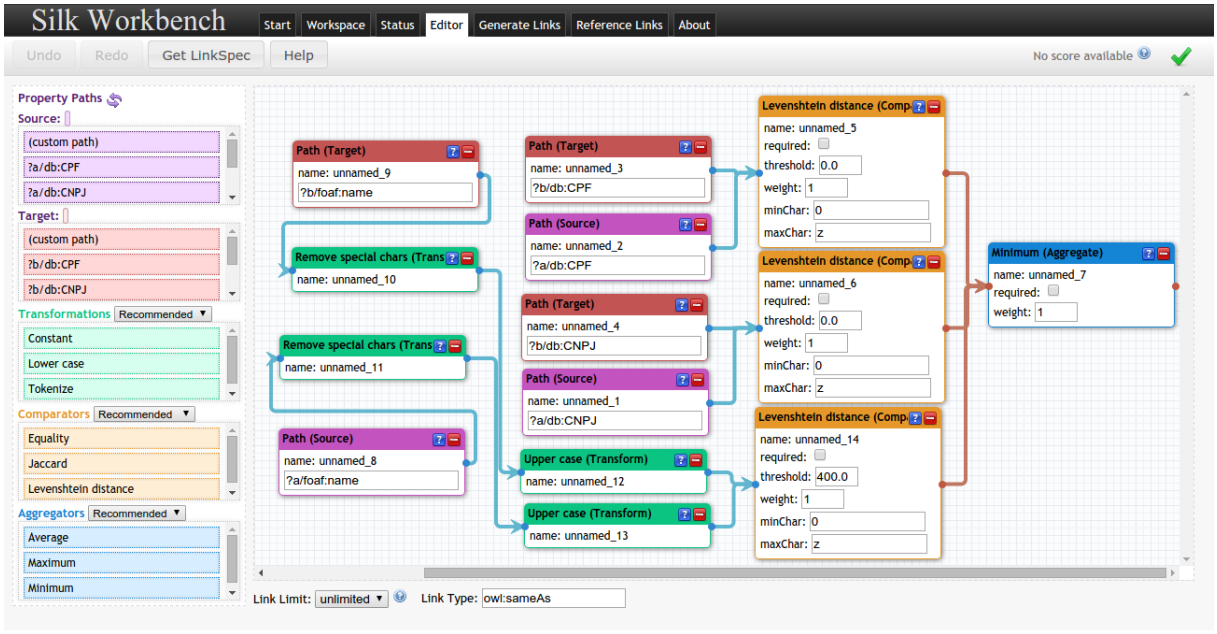
Passado essa primeira verificação, criou-se a heurística definitiva com *links* sameAs considerando o CPF completo e nome da Pessoa Física, CNPJ completo e o nome da Organização. Importante esclarecer que apenas o CPF e CNPJ deve ser completos, ou seja, o nome pode variar.

A Figura 29 apresenta os elementos que foram utilizados na heurística. Recebeu-se os CPF *target* e *source*, passou-os para comparação de similaridade que foi utilizado o algoritmo Levenshtein *distance* que foi configurado para permitir apenas os valores que fossem iguais e depois para o agregador *minimum agregate*. Além disso, recebeu-se o nome *source* e *target*, passou-os em transformações de normalização para facilitar as comparações. As transformações de comparação utilizadas, foram: retirada de caracteres especiais e *upper case*. Após as transformações passou-se para comparação de similaridade, na qual foi utilizado o algoritmo Levenshtein *distance* e definiu-se o a porcentagem limite(threshold) de distância igual a 400.0. Depois se utilizou o agregador *minimum agregate*.

Definida a heurística com auxílio da interface do Silk Workbench, o próximo passo foi gerar os links de conexões entre as bases RDF escolhidas. A Figura 30 mostra quais elementos comparados e a porcentagem de distância determinada pelo algoritmo de similaridade utilizado.

O resultado da heurística que considera o CPF, CNPJ e o nome, segue no Quadro 15 abaixo. Nessa heurística se considera como número de ligações, a soma do número de Pessoas Físicas e o número de Organizações similares entre os dois *datasets*, Tribunal de Contas da União(TCU) e a Controladoria Geral da União(CGU).

Figura 29 – Linkage Rule Editor



Fonte: Silk Workbench.

Quadro 15 – Estatística de Ligações na Heurística 2

Número de entidades do *dataset source* CGU	22498
Número de entidades do *dataset target* TCU	2690
Número de ligações	258

A seguir no Quadro 16 realizou uma comparação entre os resultado total de ligações da primeira e segunda heurísticas, nos *datasets* RDF do TCU e da CGU, criados neste trabalho.

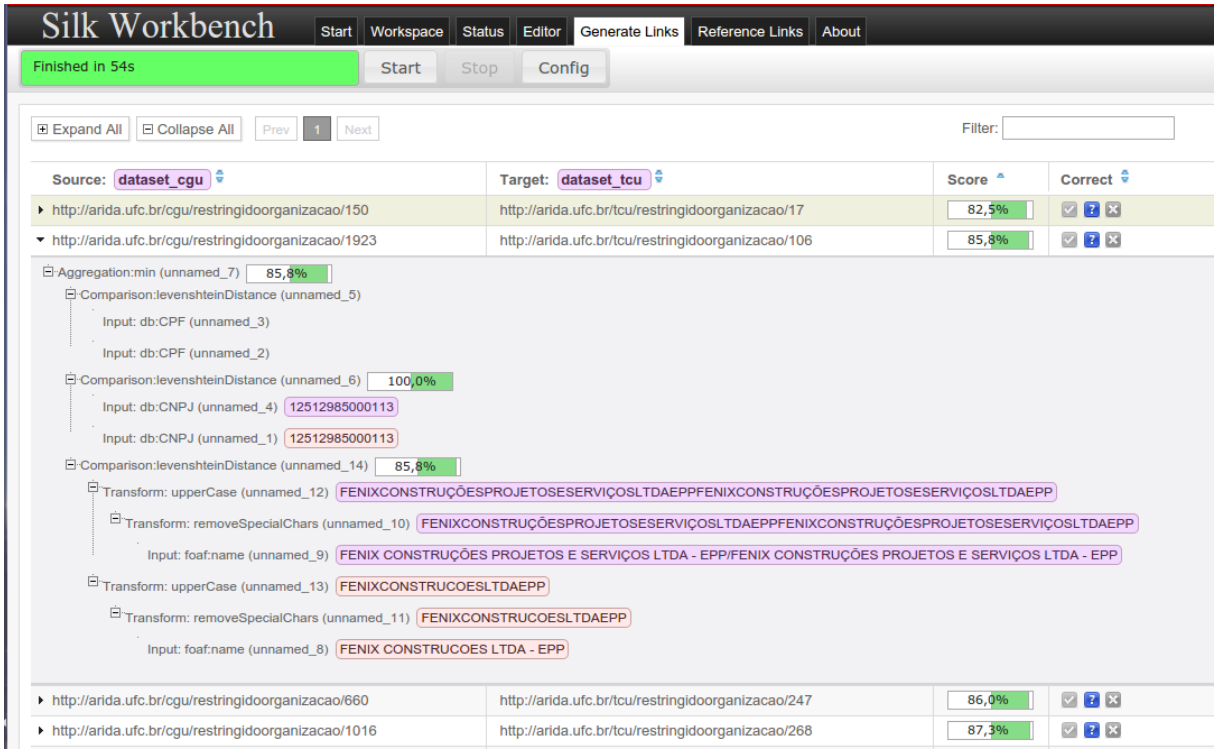
Quadro 16 – Comparação do Total de Recursos Similares considerando as heurísticas um e dois

	Quantidade de Recursos Similares
Heurística 1	277
Heurística 2	258
Diferença	19

Considerando o valor de ligações da heurística 1 que é igual 277, como o todo e o valor da heurística 2 que é 258, a diferença tem o valor igual a 19. Logo, concluímos que a heurística 2 tem um aproveitamento menor em aproximadamente 6.859 por cento. Os *links* gerados forma inseridos nos dois *datasets* servindo como enriquecimento para que as informações se completem.

A partir deste resultado, podemos concluir que o impacto da da heurística esta relacionada diretamente aos atributos que estão sendo comparados, bem como, quais métricas estão sendo utilizadas.

Figura 30 – Silk generate links



Fonte: Silk Workbench.

5.1.8 Atualização

Esta décima fase, atualização, não foi utilizada até o presente momento, mais a abordagem que será utilizada é a triplicação completa de todos os dados novamente. A intensão é evitar dados duplicados.

5.2 Considerações do Capitulo

Este capítulo apresentou a implementação do *Triplify Process* em projeto de publicação de dados conectados na Web. O processo apoiou e facilitou a publicação dos dados devido as fases, dicas e orientações.

Além disso, foi possível automatizar a triplicação dos dados com a ferramenta Penhato Data Integration (kette) e os *plugins* do projeto LinkedDataBr, conseguindo gerar dois *datasets*, assim como, criar *links* entre os *datasets* através do framework Silk. Por último, todo o workflow realizado no Pentaho Data Integration(Kettle), esta disponível no GitHub[9] do Group Advanced Research in Database(ARiDA).

[9] https://github.com/ARiDa

6 CONCLUSÃO

6.1 Considerações Finais

Este trabalho apresentou o cenário da *Web* atual como um enorme espaço global de documentos e dados distribuídos em múltiplas fontes heterogêneas. E o desafio de organizar este grande volume de dados, de maneira que seja possível integrá-los e compartilhá-los com facilidade. Além de permitir fácil processamento e interpretação do conteúdo por parte de aplicações que façam uso desses dados. Nesse contexto, apresentou-se a proposta de extensão da Web existente, por Berners-Lee, Hendler e Lassila (2011). A visão proposta por eles, denominada *Web Semântica*, consiste na criação de uma *Web* de Dados.

Com grande o crescimento da quantidade de fontes de dados disponíveis na web e pelo sucesso da iniciativa *Linking Open Data* e a busca por qualificar a publicação de dados, surgiu a carência de um processo que incentive, guie e contribua com o aprimoramento da publicação e reutilização de dados abertos conectados na Web, seguindo os princípios de *Linked Data*, bem como, orientações quanto a mapeamento, integração, triplificação, publicação e sugestões de ferramentas.

Logo, este trabalho buscou desenvolver um processo para apoiar a publicação de dados ligados na Web, seguindo os princípios de *Linked Data*, com suas devidas fases, orientações e recomendações de ferramentas.

Além disso, realizou-se a implementação do *Triplify Process*, por meio da materialização de dados, utilizando a ferramenta Pentaho Data Integration (Kettle) com os *plugins* do projeto LinkedDataBR implementando um exemplo de aplicação com dados reais relacionados ao Tribunal de Contas da União e a Controladoria Geral da União. Assim como a utilização da ferramenta Silk para realizar as devidas interligações entres os novos *datasets* RDF. Com isso, buscou-se validar o *Triplify Process* e evidenciar a importância de ligar os dados, devido fornecer maiores possibilidades de organizar grandes volumes de dados, de maneira que, facilita integrá-los e compartilhá-los. Além de permitir fácil processamento e interpretação do conteúdo por parte de aplicações que façam uso desses dados.

Conclui-se que o presente trabalho conseguiu atingir seus objetivos, fornecendo uma proposta viável como guia e apoio a publicação e interligação de dados na Web, seguindo os padrões de *Linked Data* recomendados pela W3C, bem como, aplicação de um cenário com dados reais, preenchendo uma importante lacuna em função da carência de trabalhos relacionados.

Portanto, as contribuições deste trabalho que se destacam são:

- Desenvolvimento de um processo que orienta e apoia a publicação de dados ligados na Web.
- Recomendações de padrões e boas práticas de *Linked Data*, bem como, de ferramentas, motores de buscas relacionadas a Web Semântica.
- A criação de um conjunto de dados seguindo os princípios de *Linked Data*, descrito em

RDF, com alto potencial para ser reutilizado por futuros trabalhos e aplicações envolvendo dados a respeito de empresas e pessoas inabilitadas ou inidôneas para desempenharem atividades ou funções públicas.

- Disponibilização de um SPARQL *Endpoint* para a realização de consultas SPARQL sobre os *datasets* criados.
- Apoio na disseminação do movimento *Linked Data* na comunidade acadêmica.

O conjunto de dados interligados sobre o TCU e a CGU apresentam um alto grau de flexibilidade e reusabilidade, podendo serem utilizados por demais aplicações interessadas em extrair dados.

6.2 Trabalhos Futuros

Com a realização deste trabalho, deu-se um passo para melhoria da qualidade da publicação de dados ligados na web, ou seja, com base de conhecimento para atender o padrão *Linked Data* definido pela W3C como ideal para publicação de dados na web. Com isso, alguns dos aspectos de complexidade da publicação de dados ligados na web foram abordados neste trabalho. Em trabalhos futuros, outros aspectos podem ser manipulados e melhorados, considerando os aspectos já tratados. Além disso, o processo, a arquitetura e as publicações de dados realizadas poderão ser evoluídas. Os trabalhos futuros foram agrupados a seguir, de acordo com principais assuntos apresentados neste trabalho.

- Obter um *feedback* do processo pela comunidade, que ficou fora do alcance desse trabalho.
- Definição de uma estratégia de atualização do modelo e publicação dos dados RDF de acordo com as mudanças de estrutura no modelo de dados origem.
- Adaptar o Silk para um *plugin* do Pentaho Data Integration, seria viável pois os dois utilizam API Java.
- Apresentação dos resultados às instituições proprietárias das fontes de dados e solicitação, para que disponibilizem seus dados como *Linked Data*, via *endpoint* SPARQL.

Em suma, este trabalho contribui para a disseminação das práticas *Linked Data* na comunidade acadêmica. As contribuições desde o processo de triplificação até o estudo de caso, realizado com dados reais, demonstram as possibilidades que a Web de Dados está proporcionando com sua expansão.

REFERÊNCIAS

4TH Linked Data on the Web Workshop (LDOW 2011). Disponível em: <http://events.linkeddata.org/ldow2011/slides/ldow2011-slides-intro.pdf>. Acesso em: 11.06.2015.

ALEXANDER, K. et al. Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In: *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*. Madrid, Spain: [s.n.], 2009.

BATISTA, M. G. R.; LÓSCIO, B. F. Opensbbd: Usando linked data para publicação de dados abertos sobre o sbbd. *XXVIII SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD)*, Recife/PE, 2013.

BAUER, F.; KALTENBÖCK, M. Linked open data: The essentials. *Edition mono/monochrom, Vienna*, 2011.

BERNERS-LEE, T. *Linked Data - Design Issues*. 2006. <http://www.w3.org/DesignIssues/LinkedData.html>.

BERNERS-LEE, T. Giant global graph. *Decentralized Information Group*, 2007.

BERNERS-LEE, T.; FIELDING, R.; MASINTER, L. *RFC 3986 – Uniform Resource Identifier (URI): Generic Syntax*. 2005. <http://tools.ietf.org/html/rfc3986>.

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. 2001. *Cited on*, p. 18, 2011.

BIZER, C.; CYGANIAK, R.; HEATH, T. *How to Publish Linked Data on the Web*. [S.l.]: Web-based Systems Group, Freie Universität Berlin, 2007. Http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, v. 5, n. 3, p. 1–22, 2009.

BIZER, C.; JENTZSCH, A.; CYGANIAK, R. *State of the LOD Cloud*. 2011. <http://www4.wiwiss.fu-berlin.de/lodcloud/state/>.

BRICKLEY, D.; GUHA, R. V. {RDF vocabulary description language 1.0: RDF schema}. 2004.

CAMPOS, M.; GUIZZARDI, G. *GT-LinkedDataBR–Exposição, compartilhamento e conexão de recursos de dados abertos na Web (Linked Open Data)*. 2010.

CASTERS, M.; BOUMAN, R.; DONGEN, J. V. *Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration*. [S.l.]: John Wiley & Sons, 2010.

CLARK, K. G.; FEIGENBAUM, L.; TORRES, E. *SPARQL Protocol for RDF*. 2008. <http://www.w3.org/TR/rdf-sparql-protocol/>.

CORDEIRO, K.; CAMPOS, M.; BORGES, M. Empowering citizens and government with collaboration on linked open data. In: *Proc. of the Extended Semantic Web Conference (ESWC)*. [S.l.: s.n.], 2011.

CORDEIRO, K. de F. et al. An approach for managing and semantically enriching the publication of linked open governmental data. *III WORKSHOP DE COMPUTAÇÃO APLICADA EM GOVERNO ELETRÔNICO (WCGE)*, Florianópolis/SC, 2011.

EUZENAT, J.; MOCAN, A.; SCHARFFE, F. Ontology alignments. In: *Ontology Management*. [S.l.]: Springer, 2008. p. 177–206.

FENSEL, D. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce, SpringerVerlag Berlin and Heidelberg GmbH & Co*. [S.l.]: KG, 2001.

FIELDING, R. et al. *RFC 2616 – Hypertext Transfer Protocol – HTTP/1.1*. 1999. <http://tools.ietf.org/html/rfc2616>.

GEARON, P.; PASSANT, A.; POLLERES, A. Sparql 1.1 update. *Working draft WD-sparql11-update-20110512, W3C (May 2011)*, 2012.

GERMANO, E. C.; TAKAOKA, H. Uma análise das dimensões da qualidade de dados em projetos de dados governamentais abertos. 2012.

HARMELEN, F. V.; MCGUINNESS, D. L. Owl web ontology language overview. *World Wide Web Consortium (W3C) Recommendation*, 2004.

HARTIG, O.; ZHAO, J. Publishing and consuming provenance metadata on the web of linked data. In: *Provenance and annotation of data and processes*. [S.l.]: Springer, 2010. p. 78–90.

HEATH, T.; BIZER, C. *Linked Data: Evolving the Web into a Global Data Space*. 1st. ed. [S.l.]: Morgan & Claypool, 2011. 136 p. ISBN 9781608454303.

HEATH, T.; BIZER, C. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, Morgan & Claypool Publishers, v. 1, n. 1, p. 1–136, 2011.

KLYNE, G.; CARROLL, J. J. *Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation, 2004*. 2004.

LANGEGGER, A. *A Flexible Architecture for Virtual Information Integration based on Semantic Web Concepts*. Tese (Doutorado) — J. Kepler University Linz, 2010.

MAGALHÃES, R. P. *Uma ambiente para processamento de consultas federadas em Linked Data*. Dissertação (Mestrado em Ciência da Computação) — Departamento de Ciência da Computação, Mestrado e Doutorado em Ciência da Computação, Universidade Federal do Ceará, Fortaleza/CE, 2012.

MAGALHÃES, R. P. et al. Linked data: construindo um espaço de dados global na web. *Minicurso apresentado no XXVI SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD)*, Florianópolis/SC, 2011.

MANOLA, F.; MILLER, E. *RDF Primer*. 2004. <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>. (W3C Recommendation).

MEALLING, M.; DENENBERG, R. Report from the joint w3c/ietf uri planning interest group: Uniform resource identifiers (uris), urls, and uniform resource names (urns): Clarifications and recommendations. 2002.

MENDONÇA, R. R. de. *Uma abordagem para coleta e publicação de dados de proveniência no contexto de Linked Data*. Dissertação (Mestrado em Informática) — Programa de Pós-Graduação em Informática, Instituto de Matemática, Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Dezembro 2013.

NOY, N. F.; MCGUINNESS, D. L. et al. *Ontology development 101: A guide to creating your first ontology*. [S.l.]: Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, 2001.

ÖZSU, M. T.; VALDURIEZ, P. *Principles of distributed database systems*. [S.l.]: Prentice-Hall, Inc., 1999.

PINHEIRO, J. C. *Processamento de consulta em um framework baseado em um mediador para integração de dados no padrão de Linked Data*. Dissertação (Mestrado em Ciência da Computação) — Departamento de Ciência da Computação, Mestrado e Doutorado em Ciência da Computação, Universidade Federal do Ceará, Fortaleza/CE, Setembro 2011.

PRUD'HOMMEAUX, E.; SEABORNE, A. *SPARQL Query Language for RDF*. 2008. <http://www.w3.org/TR/rdf-sparql-query/>.

PRUD'HOMMEAUX, E.; BUIL-ARANDA, C. *SPARQL 1.1 federated query. W3C*. 2011. <http://www.w3.org/TR/rdf-sparql-protocol/>. Acesso em 20 janeiro 2016.

SALAS, P. E. et al. Stdtrip: An a priori design approach and process for publishing open government data. *XXV SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD)*, Belo Horizonte /MG, 2010.

SCHULTZ, A. et al. LDIF - A Framework for Large-Scale Linked Data Integration. In: *21st International World Wide Web Conference (WWW2012), Developers Track*. [S.l.: s.n.].

SMITH, B.; WELTY, C. Ontology: Towards a new synthesis. In: ACM PRESS, USA, PP. III-X. *Formal Ontology in Information Systems*. [S.l.], 2001. p. 3–9.

VIDAL, V. M. et al. Specification and incremental maintenance of linked data mashup views. In: SPRINGER. *Advanced Information Systems Engineering*. [S.l.], 2015. p. 214–229.

VOLZ, J. et al. Silk-a link discovery framework for the web of data. *LDOW*, v. 538, 2009.

WANG, R. Y.; ZIAD, M.; LEE, Y. W. *Data quality*. [S.l.]: Springer Science & Business Media, 2006.

ZIEGLER, P.; DITTRICH, K. R. Three decades of data integration-all problems solved? In: SPRINGER. *IFIP congress topical sessions*. [S.l.], 2004. p. 3–12.

A DOCUMENTO DE VISÃO

LOGO DO PROJETO
NOME DA INSTUIÇÃO
OU ÓRGÃOS QUE PARTICIPAM DO PROJETO

TÍTULO DO PROJETO

EQUIPE ENVOLVIDA

Cidade – UF
Ano da realização do projeto

SUMÁRIO

1 INTRODUÇÃO

1.1 Objetivos

Use diferentes seções para cada conceito-chave do seu trabalho. Em geral, temos três conceitos chave. Em geral, os três conceitos chaves são identificados em seu título.

1.2 Público Alvo deste documento

[Inserir Público Alvo]

2 CONCEPÇÃO

[Forneça uma descrição resumida da definição do projeto. O que é o projeto? Quais objetivos?]

2.1 Contexto

[Forneça uma descrição resumida do contexto que o projeto esta envolvido.]

2.2 Aplicabilidade

[Forneça uma descrição resumida da aplicabilidae do projeto. Para que serve esse projeto?]

2.3 Glossário

[Forneça uma descrição dos principais termos utilizados no projeto]

Verbete	Definição

2.4 Requisitos

[Forneça uma descrição resumida dos requistos do projeto. O que esse projeto precisa para acontecer?]

2.5 Retrições

[Forneça uma descrição das restrições que o projeto não pode realizar.]

2.6 Viabilidade

[Forneça uma descrição das viabilidade que o projeto aconteça.]

2.7 Equipe

[Forneça informações do membros da equipe]

2.8 Infraestrutura

[Forneça informações da infraestrutura que será utilizada para implementação do projeto. Como, quais ferramentas]

2.9 Extensão do Projeto

[Descreva a extensão do projeto, ou seja, o tamanho deste.]

B DOCUMENTO DE VISÃO DO PROJETO DE CONVERSÃO E PUBLICAÇÃO DOS DADOS DO TCU E CGU PARA O MODELO RDF

UNIVERSIDADE
FEDERAL DO CEARÁ

Conversão e Publicação de Dados Abertos do Tribunal de Contas da União e Controladoria Geral da União para o modelo RDF, seguindo os príncipios de *Linked Data*

Salomão Santos, Regis Magalhães

Quixadá – CE
2016

SUMÁRIO

1 INTRODUÇÃO

1.1 Objetivos

Objetivo desse trabalhado é a conversão e publicação dos dados fornecidos pelo TCU e pela CGU, no formato RDF, com os padrões de *Linked Data,* através de ferramentas open-source, bem como, a disponbilização para a sociedade através de um endpoint SPARQL. Além disso, realizar a integração dessas fontes de dados.

1.2 Público Alvo deste documento

O público-alvo deste trabalho é a sociedade e o próprio Governo Brasileiro.

2 CONCEPÇÃO

2.1 Contexto

Publicação de *datasets* RDF, com dados abertos de pessoas físicas e Organizações inabilitadas ou inidôneas para realizarem serviços público.

2.2 Aplicabilidade

Combate à Organizações e pessoas físicas inabilitadas ou inidôneas para desempenharem atividades ou funções públicas.

2.3 Glossário

Verbete	Definição
Pessoa Física	Pessoa física com nome declarado pelo Tribunal de Contas da União(TCU), como inabilitados para o exercício de cargo em comissão ou função de confiança no âmbito da Administração Pública Federal, nos termos do art. 60 da Lei nº 8.443/92 (LOTCU)
Organização	Organização com nome inidôneo para participar de licitações realizadas pela Administração Pública Federal, nos termos do art. 46 da Lei nº 8.443/92 (LOTCU).
Restrição	Esse termo detalha quais as restrições de Organizações e ou pessoas, em relação aos

	Órgãos públicos. Exemplo: Especificar qual processo e intervalo de tempo que restringe a pessoa ou empresa.
Restritor	Esse termo detalha as informações do Órgão Publico responsável por realizar as restrições. Por exemplo: TCU e ou CGU.
Provenance	Esse termo detalha informações de metadados dos dados fontes, como por exemplo, o nome do arquivo e a data que foi realizado o download.

2.4 Requisitos

Converter e publicar dados do TCU e da CGU para o modelo RDF, seguindo o padrão de Linked Data.

Fornecer um *end-point* para consultas SPARQL, pela sociedade.

Realizar a interligações dos dois *dataset* gerados.

2.5 Retrições

Nenhuma.

2.6 Viabilidade

Esse projeto é viável devido se trabalhar com dados abertos e disponibilizados pelo Governo Brasileiro através do Portal de Dados Abertos e do Site de Transparência.

2.7 Equipe

Salomão Santos e Regis Pires, que acumulam as funções de planejamento e analistas de dados.

2.8 Infraestrutura

A infraestrutura fundamental que será utilizada na implementação desse processo de criação de *datasets* RDF consistiu do sistema operacional Ubuntu versão 64 bits, da plataforma Java Development Kit 64 (JDK) versão 1.8.072 (http://www.oracle.com/technetwork/pt/java/javase/downloads/index.html) e da edição comunitária (Community Edition - CE) do Pentaho Data Integration (Kettle) versão 6.0.1.0-386. Além dos steps e job entries padrão do Kettle, os 4 steps do ETL4LOD e o step NTriple Generator para serem utilizados no desenvolvimento dos workflows. Quanto ao banco de triplas, será utilizado a edição open source do Virtuoso versão 06.01.3127, para armazenar as triplas RDF geradas pelo processo de publicação.

2.9 Extensão do Projeto

A extensão desse projeto é a conversão e publicação dos dados fornecidos pelo TCU e pela CGU, no formato RDF, com os padrões de *Linked Data,* através de ferramentas open-source, bem como, a disponbilização para a sociedade através de um endpoint SPARQL.

A NOMENCLATURA PARA AJUDAR ORGANIZAR PASTA E ARQUIVOS

Quanto a nomeação de arquivos eletrônicos e pastas, não existem normas ou politicas oficiais de nomenclatura e armazenamento. Portanto, a comunidade faz o que acha melhor, mas sem essas políticas adequadas, os resultados podem ser imprevisíveis e extremamente caro. Logo, a gestão desses arquivos são complicadas e comprometidas, principalmente ao tentar integrá-los.

Qualquer atividade de negócio possui uma estrutura de nomenclatura ideal. Entretanto, uma convenção de nomenclatura estruturada que tente ser abrangente pode resultar em exagero e ou afetar o manejo.

Adiante, dez regras básicas que podem servir de oritentação geral para convenções estruturada de nomes de arquivos e pasta:

1. Evitar nomes longos ou e estruturas hierárquicas complexas, mais usar nomenclatura rica em informação:

- **Fazer:**
  Z:\Prod\QA\AssL7_WO_Suzuki_L3688_20090725.xls
  Z:\Pubs\ Article_eXadox_ File-Naming-Conventions_V03.doc

- **Não fazer:**
  Z:\Production\QualityControl\AssemblyLine7\Work Orders\Clients\SuzukiMotors\ LOT3688_July-25-2009.xls
  Z:\Publications\Articles\eXadox\File-Naming-Conventions_V03.doc

  Motivo: Estrutura de pastas hierárquicas complexas requesitam navegação extra no tempo de armazenamento e na recuperação dos arquivos. Tendo, apenas a informação essencial de forma concisa no próprio nome do arquivo, facilita tanto a pesquisa, quanto a identificação do arquivo é mais simples e precisa.

2. Colocar elementos suficientes na estrutura para facilitar a recuperação e identificação, sem exagerar.

- **Fazer:**
  NOVALEC_37507_INVOICE_20090703.pdf
  FUJITSU_S1500_SPEC_Scanner.pdf

Fonte:<http://www.exadox.com/files/pdf/en/Folder-File-Naming-Convention-How-To-Organize.pdf>.

- **Não fazer:**
  NOVALEC_INVOICE.pdf
  FUJITSU_S1500_SPEC_Black_Desktop_Scanner_ModelReplacesS510_.pdf

  Motivo: Precisão na recuperação exigir elementos suficientes para evitar que os resultados da pesquisa sejam ambíguos, mas também, muita informação exige esforço excessivo e muito tempo na nomeação do arquivo, além de pouco ou nenhum retorno nas pesquisas.

3. Usar o sublinhado (_) como elemento delimitador. Não utilizar espaços ou caracteres especiais, tais como: ! # $ % & ‘ “ @ ^ `~+,.;=) (

- **Fazer:**
  SMITH-J_AXA_7654-6_POLICY_20120915.pdf
  FUJITSU_S1500_SPEC_Scanner.pdf

- **Não fazer:**
  FUJITSUSMITH-JAXA7654-6POLICY20120915.pdf
  FUJITSU $S1500$ SPEC$Scanner.pdf

  Motivo: O sublinhado (_) é um padrão mais utilizado para campo de delimitação. Além disso, algumas ferramentas de busca não trabalham com os espaços ou caracteres especiais, logo estes devem ser evitados, especialmente para arquivos de internet. Alguns caracteres especiais pode ser interessante, mas visualmente confuso e estranho.

4. Use hífen (-) para demilitar palavras de elemento ou capitalizar a primeira letra de cada palavra dentro de um elemento.

- **Fazer:**
  Smith-John_AIG_7654-6_POLICY_2009-09-15.pdf
  WhitePaper_StructuredFileNamingStrategy.doc

- **Não fazer:**
  Smith John AIG 7654 6 POLICY 2009 09 15.pdf
  White Paper Structured file naming strategy.doc

  Motivo: Os espaços são pobres delimitadores visuais e algumas ferramentas de busca não trabalhar com espaços. O hífen (-) é um delimitador comum. Alternativamente, capitalizando as palavras dentro de um elemento é um método eficiente de diferenciando palavras, mas é mais difícil de ler.

5. Os elementos devem ser ordenados do geral para o detalhe específico de importância, tanto quanto possível.

- **Fazer:**
  FY2009_Acme-Corp_Q3_TrialBal_20091015_V02.xls Production_Paint-Shop_WorkOrder_775-2.xls

- **Não fazer:**
  TrialBal_Q3_20091015_Acme-Corp_V02_FY2009.xls Paint-Shop_775-2_WorkOrder_Production.xls

  Motivo: Em geral, os elementos devem ser ordenados logicamente, na mesma seqüência que você normalmente procurar um arquivo alvo.

6. A ordem de importância da regra se aplica quando elementos incluem data e hora. As datas devem ser ordenadas: ano, mês, dia. (por exemplo, YYYYMMDD, YYYYMMDD, YYYYMM). O tempo deve ser ordenada: Hora, Minutos, Segundos (HHMMSS).

- **Fazer:**
  RFQ375_Cables-Unlimited _BID_20091015-1655.pdf
  2009-11-20_AMATProj_Phase1_Report.docProduction_Paint-Shop_WorkOrder_775-2.xls

- **Não fazer:**
  RFQ375_Cables-Unlimited_BID_10152009-1655.pdf Nov-20-2009_AMATProj_Phase1_Report.doc

  Motivo: Para garantir que os arquivos serão organizados em ordem cronológica correta, os componentes mais importantes de data e hora devem aparecer em primeiro lugar seguido com os componentes menos significativos.

7. Nomes pessoais dentro de um elemento deve ter em primeiro lugar o nome da família, seguido pelo primeiro nome ou iniciais.

- **Fazer:**
  Tate-Peter_SunLife _1-7566-2_POLICY_10YrTerm.pdf
  SmithJ_ ID3567_ADMIN_WageReview.xls

- **Não fazer:**
  Peter-Tate_SunLife _1-7566-2_POLICY_10YearTerm.pdf
  JSmith_ID3567_ADMIN_WageReview.xls

Motivo: O nome de família é uma referência padrão para recuperar registros. Tendo o nome da família em primeiro lugar garante que os arquivos são classificados em ordem alfabética adequada.

8. Abreviar o conteúdo de elementos sempre que possível.

- **Fazer:**
  RevQC _QST_2009-Q2.xls
  MCIM_27643_POD.doc

- **Não fazer:**
  Minister of Revenue Quebec _Quebec-Sales-Tax_2009-2ndQuarter.xls
  MultiCIM-Technologies-Inc_27643_Proof-Of-Delivery.pdf

  Motivo: Abreviação ajuda a criar nomes de arquivos concisos, mais fáceis de ler e reconhecer.

9. Um elemento para controle de versão, deve começar com V seguido por pelo menos dois dígitos e deve ser colocado como o último elemento a mais. Para distinguir entre rascunhos de trabalho (ou seja, revisões menores) usar Vx-01-> Vx-99 gama e para o projeto final (ou seja, grande lançamento da versão) usar V1-00-> V9-xx. (em que x = 0-9)

- **Fazer:**
  MCIM_Proposal_V09.doc
  eXadox_UserManual_V1-02.doc

- **Não fazer:**
  MCIM_Proposal_9.doc
  eXadox_UserManual_V2FinalDraft.doc

  Motivo: O "V" ajuda a denotar que o elemento refere-se a um número de versão. Um mínimo de dois dígitos com uma zero à esquerda é necessária para garantir que busca resultados são devidamente ordenados. A intenção é evitar a situação onde, por exemplo, um nome de arquivo com um "V1-13" vai erradamente comparecer perante um nome de arquivo idêntico a uma "V1-2" número da versão, quando classificados em ordem crescente alfabética ou numérica. Para distinguir entre o trabalho, revisão e redação final de um único dígito prefixo seguido de hífen "-" é o preferido para facilitar a triagem adequada; utilização palavras no nome do arquivo, tais Final, Projeto ou Comentário no nome do arquivo afetam a ordem e deve ser evitado.

10. Prefixar os nomes das sub-pastas pertinentes ao nome do arquivo de arquivos que estão sendo compartilhados via e-mail ou dispositivos de armazenamento portáteis.

- **Fazer:**
  Prod_PS_AssL7_WO_Suzuki_J3688-20090725.xls
  FY2009_Acme-Corp _Q3_TrialBal_20091015_V02.xls

- **Não fazer:**
  WO_Suzuki_J3688-20090725.xls
  Q3_TrialBal_20091015_V02.xls

  Motivo: Arquivos anexados e arquivos compartilhados através de dispositivos portáteis incluir apenas o nome do arquivo e pode ser totalmente desprovido de contexto, o que é geralmente fornecida pela estrutura da pasta de origem. Para compensar e evitar confusão, às vezes é essencial prefixar o nome da subpasta (s) para tais nomes de arquivos.