GOVERNO DO MARANHÃO
Secretaria de Estado da Fazenda

RELATÓRIO

INFRAESTRUTURA DE ACESSO E INTEGRAÇÃO DE DADOS DA SEFAZ

Título do Projeto:

Uso das Tecnologias da Web Semântica e Big Data para Integrar, Enriquecer e Explorar Dados da SEFAZ

Instituições Parceiras:

- Universidade Federal do Ceará – UFC

  Reitor: José Cândido L. B. de Albuquerque

- Secretaria da Fazenda do Estado do MARANHÃO – SEFAZ/MA

  Secretário: Marcellus Ribeiro Alves

Órgão Financiador: BID PROFISCO2

Sumário

1. Introdução

1.1 Grafo de Conhecimento Semântico

Nos últimos anos, as tecnologias de Grafo de Conhecimento Semânticos (GCS) estabeleceram uma posição sólida no mundo corporativo, servindo como um elemento central na infraestrutura de gerenciamento de dados organizacionais. Os grafos de conhecimento estão se tornando tanto o repositório de dados mestres de toda a organização (esquema ontológico e conhecimento de referência estática) quanto o hub de integração para várias fontes de dados legados.

Grafos de Conhecimento Empresariais (GCE) são uma extensão dos Grafos de Conhecimento Semântico para o âmbito empresarial, possibilitando que empresas consigam obter novos insights sobre seus dados e criar aplicações empresariais inovadoras, mostrando-se um poderoso meio para lidar com o desafio de conectar fontes de dados internas e externas.

A SEFAZ-MA está construindo um grafo de conhecimento inteligente que pode dar suporte a aplicações orientados ao conhecimento em toda a organização. Nesse sentido, o GCE da SEFAZ-MA funcionará como uma fábrica de conhecimento, onde novos conhecimentos serão inferidos à partir do conhecimento existente. Adicionalmente, o grafo de conhecimento faz uso de ontologias, com o objetivo de garantir a fácil integração de múltiplas fontes heterogêneas de dados, como as fontes fornecidas pela RFB, e prover uma representação semântica formal para permitir inferência e processamento de máquina.

1.2. Arquitetura da Plataforma do Grafo de Conhecimento Semântico da SEFAZ-MA

A arquitetura projetada para a SEFAZ-MA fornece um conjunto de componentes para Armazenamento, Acesso, Integração, Consulta e Gerenciamento dos dados da SEFAZ-MA. A **Figura 1** mostra os principais componentes agrupados em cinco camadas: **Camada das Fontes de Dados, Camada Relacional, Camada Semântica, Camada de Acesso e Integração de Dados e Camada de Aplicações.** Cada camada representa um nível de abstração, sendo a camada das fontes de dados com mais baixo nível de abstração e a camada de aplicações a de mais alto nível de abstração.

Na camada das fontes de dados estão armazenados os dados de todas as fontes que terão seus dados integrados. Vale destacar que uma fonte de dados pode ser composta de várias outras fontes, como o caso dos dados da RFB, que estão armazenados em 20 arquivos que devem ser integrados também.

Na camada relacional, os dados integrados de cada fonte são mapeados para vários bancos de dados relacionais. Nesta camada, existem ainda visões relacionais homogêneas que integram as várias fontes de Dados da SEFAZ-MA.

Na camada semântica, um grafo de conhecimento para cada fonte de dados é instanciado a partir de dados da camada relacional e através de ontologias. Nessa camada, são construídas ligações semânticas entre diferentes grafos de conhecimento.

Na camada de acesso e integração é provido o suporte necessário para viabilizar o acesso integrado aos dados. Portanto, os usuários e aplicações podem realizar consultas de forma transparente, sem ter que entender sobre as fontes de Dados e estruturas de dados. Por último, a camada de aplicações irá apresentar as interfaces necessárias para que usuários e aplicações acessem os metadados e dados via GCE da SEFAZ/MA.

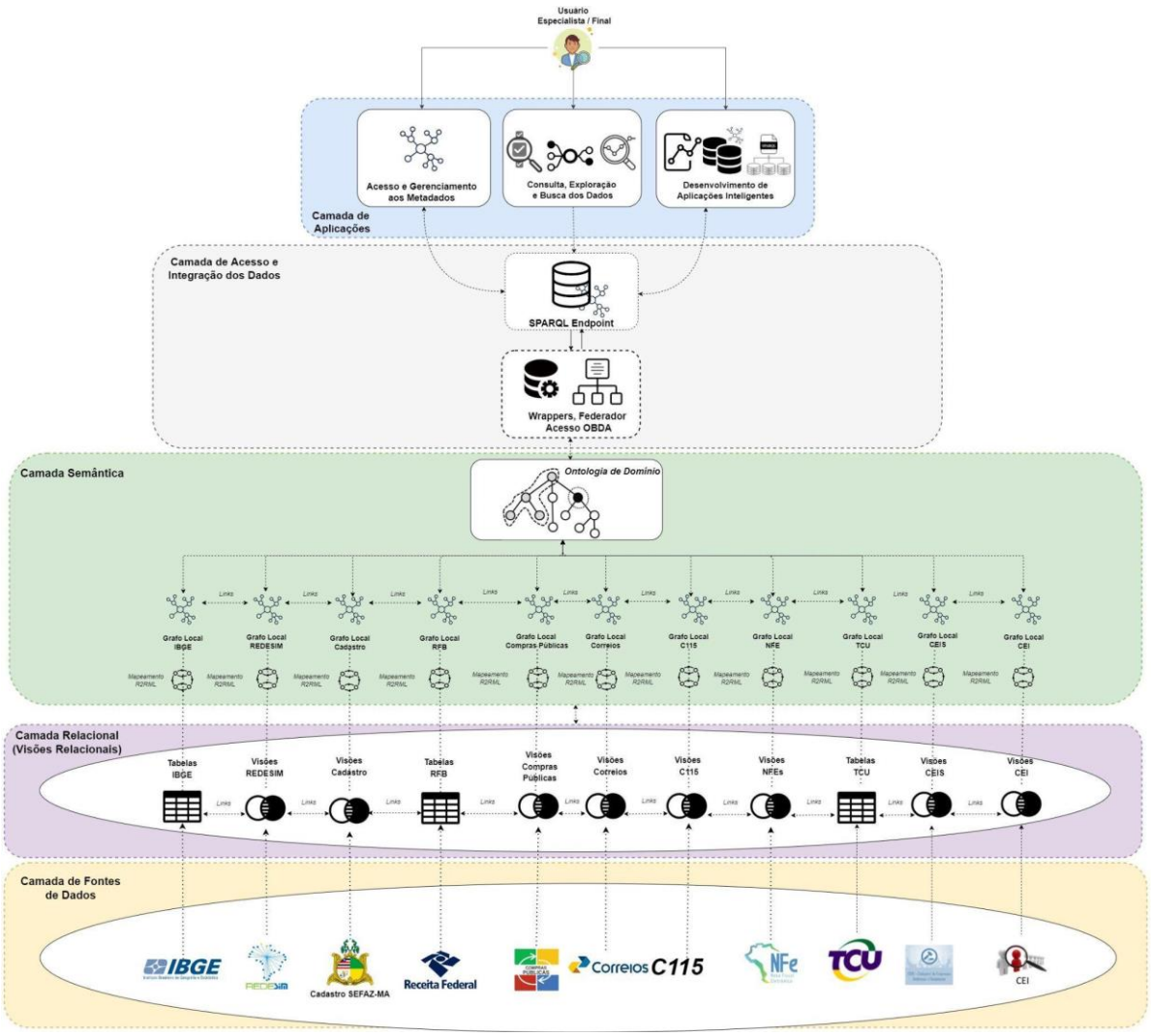


Figura 1. Arquitetura da Plataforma de Conhecimento Semântico da SEFAZ-MA.

1.3 Processo de Construção da Camada Semântica

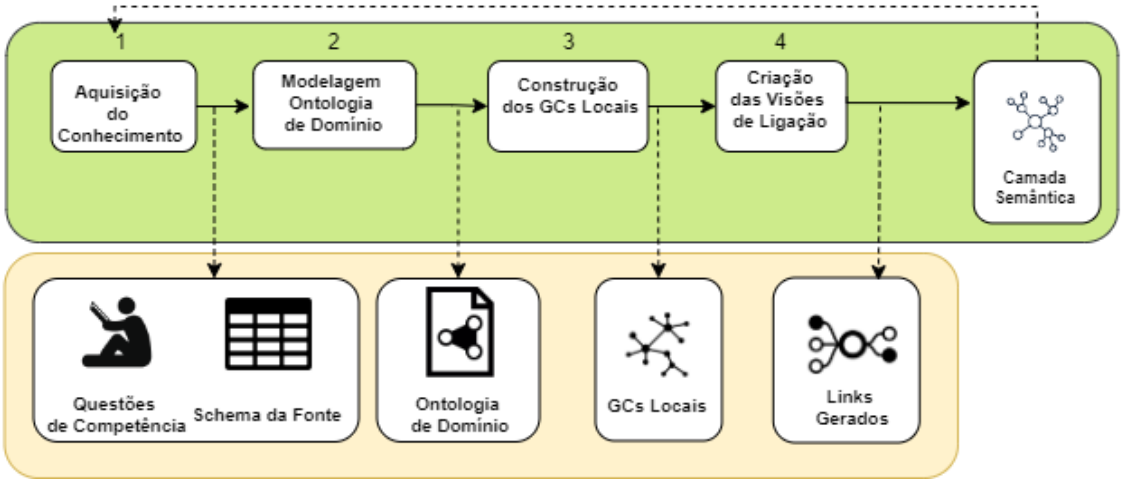


Figura 2. Processo de Construção da Camada Semântica.

O processo de construção da camada semântica (**Figura 2**), é realizado de forma incremental em 4 passos: (i) Aquisição do Conhecimento; ii) Modelagem da Ontologia de Domínio; iii) Construção dos Grafos Locais (GCL); e iv) Criação das Visões de Ligação, conforme descritos a seguir.

Inicialmente, no Passo 1 de Aquisição do Conhecimento, em primeiro lugar deve-se estabelecer a meta por trás da coleta e integração dos dados, assim como definir as questões de competência que o usuário deseja que sejam respondidas. Em seguida, procurar descobrir quais as fontes de dados seriam mais úteis para atingir seu objetivo em termos de domínio, escopo, proveniência, manutenção, etc.

Posteriormente, no Passo 2, o processo de Modelagem da Ontologia de Domínio requer uma análise minuciosa dos diferentes esquemas das fontes de dados, objetivando a criação de um modelo comum para resolver o problema da heterogeneidade das fontes. Nessa atividade, deve-se procurar reusar ontologias já existentes, e formalizar o modelo usando os padrões como *RDF Schema* e OWL.

Já no Passo 3, para cada fonte de dados, é construído um grafo de conhecimento local usando o vocabulário da ontologia de domínio. A construção de um GCL é realizada em 3 passos: I) Primeiro, é gerado o mapeamento entre a fonte de dados e a ontologia de domínio; II) Então, baseado nos mapeamentos gerados, é especificada a ontologia do GC local; III) Por último, o GCL é implementado usando o enfoque virtual ou materializado de forma independente.

Considerando a Abordagem Materializada os dados relevantes são extraídos das fontes de dados originais, transformados em representação RDF de acordo os mapeamentos para a ontologia de domínio, e armazenados em um *triplestore*. As consultas sobre o grafo são processadas diretamente nos dados (visão RDF

materializada). A desvantagem desse enfoque é que o grafo precisa ser atualizado quando atualizações são efetuadas na fonte de dados.

Já na Abordagem Virtual o GCL é uma visão virtual, onde as consultas realizadas sob o *endpoint* do GCL são reescritas em consultas sobre as fontes de dados originais. O uso dessa abordagem garante que os dados disponibilizados estejam sempre atualizados com relação às fontes originais.

Por fim no Passo 4, são criadas visões de ligação relacionando recursos em diferentes GCLs que tenham classes em comum. Para construir uma visão de ligação, o usuário terá que criar primeiro a regra de ligação (*linkage*) que especifica as condições para inferência de links “*same-As*”.

2. Camada das Fontes de Dados

Figura 3. Camada de Fontes de Dados.

Na camada das fontes de dados (**Figura 3**) estão armazenados os dados de todas as fontes que terão seus dados integrados. O formato de armazenamento destas fontes pode estar em diferentes formatos, e.g., bancos de dados relacionais, CSV, *triple stores* RDF, documentos JSON, etc. Vale destacar que uma fonte de dados pode ser composta de várias outras fontes, como o caso dos dados da RFB, que estão armazenados em 20 arquivos.

2.1 RFB: Cadastro da Receita Federal

A Receita Federal, ou Secretaria Especial da Receita Federal do Brasil, é um órgão que tem como responsabilidade a administração dos tributos federais e o controle aduaneiro, além de atuar no combate à evasão fiscal (sonegação), contrabando, descaminho, contrafação (pirataria) e tráfico de drogas, armas e animais. Até 1 de janeiro de 2019, foi subordinada ao Ministério da Fazenda e, a partir daí, passou a integrar a estrutura básica do novo Ministério da Economia do governo Jair Bolsonaro.

A RFB é fonte de dados primária responsável por fornecer a informação fidedigna de CPFs e CNPJs. Se uma pessoa não possui CPF válido e ativo na RFB, para fins de concessão de inscrição, a participação como representante desta pessoa em empresas não será considerada válida junto aos Estados. Se uma Empresa não tem CNPJ então é uma empresa de fato, mas não de Direito, de forma que estará fazendo exercício ilegal das suas atividades. Sendo o segundo ente a dar parecer no REDESIM, para a abertura de empresas.

- **Fonte:** https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/consultas/dados-publicos-cnpj
- **Formato:** CSV
- **Taxa de atualização:** A atualização é mensal.
- **Características especiais:** Os dados são disponibilizados em uma estrutura própria e em grande volume.

2.2. IBGE

O Instituto Brasileiro de Geografia e Estatística (IBGE) é um instituto público da administração federal brasileira criado em 1934 e instalado em 1936 com o nome de Instituto Nacional de Estatística; seu fundador e grande incentivador foi o estatístico Mário Augusto Teixeira de Freitas. O nome atual data de 1938. A sede do IBGE está localizada na cidade do Rio de Janeiro.

O IBGE tem atribuições ligadas às geociências e estatísticas sociais, demográficas e econômicas, o que inclui realizar censos e organizar as informações obtidas nesses censos, para suprir órgãos das esferas governamentais federal, estadual e municipal, e para outras instituições e o público em geral.
Desta fonte de dados são extraídos dois conjuntos de dados que serão integrados ao grafo semântico da SEFAZ: IBGE – CNAE e IBGE – Localização.

2.2.1 IBGE – CNAE

O CNAE tem como objetivo primário identificar e representar os conceitos pertencentes ao domínio de atividades econômicas desempenhadas por organizações / empresas seguindo uma hierarquia de relações.

- **Fonte:** https://concla.ibge.gov.br/classificacoes/download-concla.html
- **Formato:** Excel ou PDF
- **Taxa de atualização:** Indeterminado
- **Características especiais:** Historicamente existiram diversas versões da classificação CNAE, onde cada uma foi vigente por um período de tempo específico. Para este projeto é necessário armazenar todas as versões da classificação já adotadas, sendo necessário indicar o período de validade de cada versão.

2.2.2 IBGE – Localização

Os dados de Localização fornecidos pelo IBGE tem como intuito representar informações de Localização através de relações de parte entre Localização, País, Estado, Município, Mesorregião, Microrregião, Bairro e Logradouro.

- **Fonte:** https://www.ibge.gov.br/explica/codigos-dos-municipios.php
- **Formato:** CSV
- **Taxa de atualização:** Indeterminado.

2.3 Cadastro SEFAZ-MA

O Cadastro de Dados de Contribuintes do Estado do Maranhão é representado por informações de Cadastro de Pessoas Jurídicas (Empresas) e Físicas (Pessoas) bem como dados de Estabelecimentos, Representantes Legais, Sócios, Baixas e Razões de Desabilitação.

No âmbito do Estado do Maranhão é o banco de dados mantido na Secretaria de Estado da Fazenda, que contém informações gerais de todos os contribuintes de impostos. Estes são obrigados a inscrever seus estabelecimentos antes de iniciarem suas atividades e a comunicarem quaisquer alterações dos dados declarados para sua inscrição.

- **Fonte:** Dados do Cadastro de Contribuintes do Maranhão**.**
- **Formato:** Banco de Dados Relacional.
- **Taxa de atualização:** Frequente.
- **Características especiais:** Os dados são armazenados historicamente sempre ao haver ocorrência de alteração nas informações de uma Empresa / Estabelecimento ou na constituição de seu Quadro de Sócios.

2.4 REDESIM

A Rede Nacional para a Simplificação do Registro e da Legalização de Empresas e Negócios, REDESIM, é um sistema integrado que permite a abertura, fechamento, alteração e legalização de empresas em todas as Juntas Comerciais do Brasil, simplificando procedimentos e reduzindo a burocracia ao mínimo necessário.

Esta é uma fonte de amplitude nacional que contém informações sobre a razão social, siglas, nome comercial, data de formação, data de fechamento, tipo de classe e sociedade, seu endereço, etc. Além disso, também é possível recuperar dados sobre seus representantes legais e funcionários. Isto permite o rastreio das atividades de possíveis suspeitos através de diferentes organizações. Tal informação pode ser utilizada para a recomendação da análise aproximada das atividades de uma pessoa ou organização por parte de um auditor. As informações desta fonte podem ser agrupadas e comparadas contra as demais fontes, o que permite a identificação de inconsistências nas informações declaradas.

A SEFAZ/MA tem todos os registros do REDESIM, em função de acordo com a empresa VOX e a Junta Comercial do Maranhão - JUCEMA, mas tal fonte até hoje não foi utilizada, a qual representa uma integração entre diversos órgãos e

instituições, conforme legislação. O REDESIM é um mashup de dados governamentais normatizado, não tratado.

- **Fonte de Acesso:** Dados da REDESIM do Maranhão**.**
- **Formato:** Banco de Dados Relacional.
- **Taxa de atualização:** Frequente.
- **Características especiais:** Os dados são obtidos através de um Web Service em formato JSON mediante parceria com a empresa VOX. Por conseguinte, esses dados são armazenados no Oracle em estrutura relacional seguindo um esquema de tabelas e colunas no ambiente da SEFAZ-MA.

2.5 Notas Fiscais

A Nota Fiscal Eletrônica (NF-e), juntamente com a Nota Fiscal de Consumidor Eletrônica (NFC-e) foram desenvolvidas de forma integrada pelas Secretarias de Fazenda dos Estados e Secretaria da Receita Federal do Brasil, a partir da assinatura do Protocolo ENAT 03/2005, que atribui ao Encontro Nacional de Coordenadores e Administradores Tributários Estaduais (ENCAT) a coordenação e a responsabilidade pelo desenvolvimento e implantação do Projeto NF-e.

O modelo de dados das Notas Fiscais contém informações sobre diversos outros dados que se relacionam com as Notas, como informações sobre produtos, pessoal autorizado, impostos aplicados e tributos devolvidos, entre outros, com ênfase nas informações que definem a nota fiscal, suas características principais e relações com outras notas fiscais, podendo ou não estar presentes no banco de dados (ou arquivo XML).

- **Fonte de Acesso:** Banco de dados de Notas Fiscais do Maranhão.
- **Formato:** Banco de Dados Relacional povoado via arquivo xml.
- **Taxa de atualização:** Frequente.
- **Características especiais:** Os dados são recebidos no formato de um arquivo XML que então é analisado e usado para o povoamento do banco de dados relacional das Notas Fiscais do Maranhão, que busca refletir ao máximo a estrutura dos dados no arquivo original.

2.6. Convênio 115

O convênio ICMS 115/03 da Conselho Nacional de Política Fazendária - CONFAZ, que data de dezembro de 2003, estabelece uma série de exigências que visa orientar o procedimento de emissão em relação a escrituração dos livros fiscais, emissão e manutenção de documentos fiscais correspondentes a serviços prestados por contribuintes do ICMS dos setores de telecomunicações e energia elétrica.

As informações serão mantidas e prestadas através dos seguintes arquivos:

a) MESTRE DE DOCUMENTO FISCAL, com informações básicas dos documentos fiscais;
b) ITEM DE DOCUMENTO FISCAL, com detalhamento das mercadorias ou serviços prestados;
c) DADOS CADASTRAIS DO DESTINATÁRIO DO DOCUMENTO FISCAL, com as informações cadastrais do destinatário do documento fiscal;
d) IDENTIFICAÇÃO E CONTROLE, com a identificação do contribuinte e resumo da quantidade de registros e somatório de valores dos arquivos acima referidos;

- **Fonte de Acesso:** Dados do Convênio.
- **Formato:** Original em TXT e processados em Formato Relacional.
- **Taxa de atualização:** Frequente com intervalos de 15 minutos.
- **Características especiais:** Os dados são armazenados em 4 arquivos de texto (.txt) e inseridos em um esquema relacional através de um script que verifica se existem novos arquivos no diretório e faz a carga no banco. As informações são enviadas e carregadas as informações de todos os usuários, e não somente o que houve alteração.

2.7. Compras Públicas

As Compras Públicas são constituídas a partir da integração de um recorte dos dados de Notas Fiscais do Maranhão e de Compras Governamentais no âmbito do Maranhão. Essa integração fornece dados sobre compras, fornecedores, itens (materiais ou produtos) e uasgs / órgãos.

- **Fonte de Acesso:** Dados de Compras Públicas de Notas Fiscais Eletrônicas (SEFAZ-MA).
- **Formato:** Banco de Dados Relacional
- **Taxa de atualização:** Sincronizado com o ambiente Nacional do SPED.

2.8. TCU

O TCU apresenta a relação atualizada dos responsáveis declarados inidôneos para licitar e aqueles considerados inabilitados para o exercício de cargo ou função pública.

- **Fonte:** https://contas.tcu.gov.br/ords/f?p=704144:1:15032420593870:::::
- **Formato:** Tabelas de Inabilitados e Inidôneos Publicadas em CSV.
- **Taxa de atualização:** Instantâneo.

2.9. CEIS

O Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) apresenta a relação de empresas e pessoas físicas que sofreram sanções que implicaram a restrição de participar de licitações ou de celebrar contratos com a Administração Pública.

- **Fonte**: https://www.portaldatransparencia.gov.br/download-de-dados/ceis
- **Formato:**CSV
- **Taxa de atualização:** Diária**.**

2.10. CEI

O CEI é um banco de dados onde estão registrados os nomes de pessoas físicas e jurídicas em débito para com órgãos e entidades da administração pública estadual, direta e indireta. As informações estão centralizadas no Sistema de Cadastro Estadual de Inadimplentes – SISCEI, cabendo à Secretaria de Estado da Fazenda expedir orientação de natureza normativa.

- **Fonte:** Dados do CEI (providos pela SEFAZ-MA).
- **Formato:** Tabelas do Cadastro de Empresas Inadimplentes
- **Taxa de atualização:** Mensal.

3. Camada Relacional Normalizada

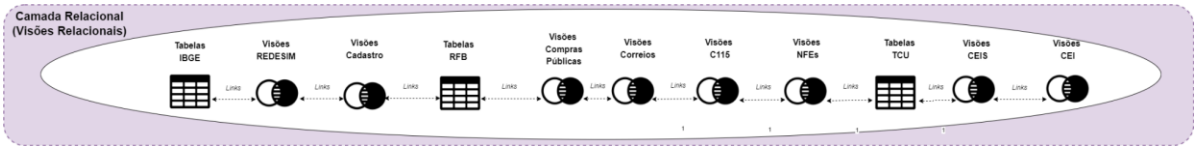


Figura 4. Camada Relacional Normalizada.

O objetivo da Camada Relacional (**Figura 4**) é criar Visões Relacionais homogêneas que integram as várias fontes de Dados da SEFAZ-MA. As visões relacionais usam um vocabulário comum, sendo reusado o vocabulário da Ontologia de Domínio.

As vantagens da criação da camada Relacional, são:

- O mapeamento das visões relacionais para a Ontologia de domínio tem a semântica dos mapeamentos diretos, o que facilita a criação e manutenção desses mapeamentos. Assim como, a diminuição do “gap” semântico entre o modelo relacional e modelo RDF.

- As Visões Relacionais são criadas usando a linguagem SQL, que é a tecnologia usada nos banco de dados da SEFAZ. Portanto, os desenvolvedores da SEFAZ já dominam bem essa tecnologia.

- A Camada Relacional oferece uma visão relacional integrada das fontes de dados, o que permitirá que desenvolvedores, ainda não familiarizados com as tecnologias da web semântica, possam desenvolver aplicações avançadas que requerem a integração de dados de várias fontes de dados.

- A Camada Relacional poderá ser consultada usando a ontologia de domínio, o que permitirá a realização de busca semântica também na camada relacional.

- Na Camada Relacional Normalizada, visões de ligação foram criadas de modo a facilitar a descoberta de relações envolvendo tabelas de fontes distintas, facilitando identificar conceitos e indivíduos idênticos.

As Visões Relacionais Normalizadas podem ser criadas usando o enfoque virtual ou materializado. No caso do enfoque materializado, os dados são extraídos, transformados e armazenados em uma tabela da camada Relacional. Já no enfoque virtual, os dados ficam armazenados nas tabelas originais, sendo que as consultas realizadas nas tabelas virtuais são reescritas em consultas sobre as tabelas das fontes de dados originais.

A seguir apresentamos as Visões Relacionais e Tabelas Normalizadas criadas para cada fonte de dados.

3.1 Fontes Internas

3.1.1 Cadastro SEFAZ

- **Esquema Relacional Normalizado:** Camada Relacional/Fontes Internas/Cadastro/Esquema Normalizado/Diagrama_Relacional_Cadastro.drawio.svg

- **Script de Criação - Visões Cadastro:** Camada Relacional/Fontes Internas/Cadastro/Script SQL/Cadastro.sql

3.1.2 REDESIM

- **Esquema Relacional Normalizado:** Camada Relacional/Fontes Internas/REDESIM/Esquema Normalizado/Diagrama_Relacional_REDESIM.svg
- **Script Criação Visões REDESIM:** Camada Relacional/Fontes Internas/REDESIM/Script SQL/REDESIM.sql

3.1.3 CEI

- **Esquema Relacional Normalizado:** Camada Relacional/Fontes Internas/CEI/Esquema Normalizado/Diagrama_Relacional_CEI.drawio.svg

- **Script de Criação - Visões Relacionais:** Camada Relacional/Fontes Internas/CEI/Script SQL/CEI.sql

3.1.4 Compras Públicas

- **Esquema Relacional Normalizado:** Camada Relacional\Fontes Internas\Compras Publicas\Esquema Normalizado\Diagrama_Relacional_Compras.drawio.svg

- **Script de Criação - Visões Relacionais:** Camada Relacional\Fontes Internas\Compras Publicas\Script SQL\Compras.sql

3.1.5 C115

- **Esquema Relacional Normalizado:** Camada Relacional/Fontes Internas/C115/Esquema Normalizado/Diagrama_Relacional_C115.drawio.svg

- **Script de Criação - Visões Relacionais:** Camada Relacional/Fontes Internas/C115/Script SQL/C115.sql

3.2 Fontes Externas

3.2.1 RFB

Tabelas de Extração:

O processo de extração é iniciado pelo script python, baixando os arquivos do site da RFB, convertendo-os em csv e executando a carga nas tabelas de extração do banco de dados pelo SQL*Loader, ferramenta nativa do Oracle. De acordo com o site da Receita, os dados são gerados mensalmente. O ciclo de carga de novos dados para o banco de dados segue o mesmo período.

Tabelas Relacionais Normalizadas

A Camada Relacional da RFB, quanto aos seus dados de extração, possui dois tipos diferentes de tabelas, denominadas pela própria RFB como tabelas de extração (prefixo EXTR) contendo os dados padrões de extração, e tabelas de domínio (prefixo DOM) contendo dados de outros domínios que a RFB utiliza nas tabelas de extração.

Vale ressaltar a mudança que ocorreu no leiaute dos dados de extração da RFB, modificando alguns dados já existentes, tais modificações incluem:

- Informações de faixa etária dos sócios
- Informações de entes federativos das empresas
- Informações adicionais sobre MEI

• Mudanças de algumas informações sobre países e municípios (no leiaute anterior, tinha-se informações sobre código e nome, agora apenas o código. O nome é obtido nas tabelas de domínio).

- A Camada Relacional da RFB também possui tabelas denominadas tabelas externas, que representam tabelas que serão ligadas a outras fontes de dados, com o IBGE, CORREIOS e outras fontes da RFB (Situação Cadastral e Motivo da Situação Cadastral)

- As tabelas da camada relacional da RFB foram divididas em três tipos diferentes de tabelas, tabelas da RFB são as tabelas que foram obtidas diretamente dos dados de extração, tabelas externas da RFB, que são tabelas referentes à RFB, porém, a obtenção desses dados não vem diretamente dos dados de extração, mas sim de outras fontes da RFB (já que a RFB é uma fonte composta por outras fontes dependendo do escopo dos seus dados), e tabelas externas que não são dados diretamente da RFB, como dados do IBGE e Correios.

Criação dos Tablespaces, das Tabelas de Extração e Tabelas Normalizadas:

- **Esquema das Tabelas de Extração:** Camada Relacional/Fontes Externas/RFB/Diagrama de Extracao RFB.pdf
- **Esquema das Tabelas Normalizadas:** Camada Relacional/Fontes Externas/RFB/Esquema Normalizado/Diagrama_Relacional_RFB.drawio.svg

- **Criação dos Tablespaces das Tabelas de Extração:** Camada Relacional/Fontes Externas/RFB/Script SQL/create_tablespace_rfb_extracao.sql
- **Criação das Tablespaces das Tabelas Normalizadas:** Camada Relacional/Fontes Externas/RFB/Script SQL/create_tablespace_rfb_normalizadas.sql
- **Criação das Tabelas de Extração:** Camada Relacional/Fontes Externas/RFB/Script SQL/cria_tb_extracao_rfb.sql
- **Criação das Tabelas Normalizadas:** Camada Relacional/Fontes Externas/RFB/Script SQL/cria_tb_norm_rfb.sql

Povoamento das Tabelas de Extração e Tabelas Normalizadas:

- **Criação das funções de povoamento das tabelas:** Camada Relacional/Fontes Externas/RFB/Script SQL/funcoes_povoamento_rfb.sql
- **Criação dos índices:** Camada Relacional/Fontes Externas/RFB/Script SQL/indices_constraints.sql
- **Script decodificador:** Camada Relacional/Fontes Externas/RFB/Script Python/decodificador_v2.zip

Tabelas administrativas e adicionais:

Para responder questões administrativas na RFB, como quantas cargas já foram efetuadas, quais as datas das cargas foram realizadas, quantas e quais cargas foram realizadas com um determinado layout, a tabela guarda as datas das cargas com a versão do layout da Receita. Esta tabela é povoada pelo script python após realizar a carga das tabelas de extração. A trigger é disparada após a inserção de uma tupla na tabela e inicia o povoamento das tabelas normalizadas. Como algum erro pode ocorrer durante o povoamento das tabelas normalizadas, a procedure grava o erro e o nome da tabela normalizada que ocorreram na tabela de Log de Erros.

- **Criação das tabelas administrativas e adicionais:**

  Procedure: Camada Relacional/Fontes Externas/RFB/Script SQL/log_povoa_procedure_rfb.sql

  Tabelas Administrativas / Trigger: Camada Relacional/Fontes Externas/RFB/Script SQL/tabelas_administrativas_rfb.sql

- **Processo completo - Manual de Implantação** Camada Relacional\Fontes Externas\RFB\Manual de Implantacao.pdf

3.2.2 IBGE-CNAE

- **Esquema IBGE CNAE:** Camada Relacional/Fontes Externas/IBGE CNAE/Esquema Normalizado/Diagrama_Relacional_IBGE_CNAE.drawio.svg
- **Criação e Povoamento das Tabelas Normalizadas:** Camada Relacional/Fontes Externas/IBGE CNAE/Script SQL/IBGE_CNAE.sql

3.2.3 IBGE-Localização

- **Esquema IBGE-Localização:** Camada Relacional/Fontes Externas/IBGE Localizacao/Esquema Normalizado/Diagrama_Relacional_IBGE_Localizacao.drawio.svg
- **Criação e Povoamento das Tabelas Normalizadas:** Camada Relacional/Fontes Externas/IBGE Localizacao/Script SQL/IBGE_Localizacao.sql

3.2.4 TCU

- **Esquema Relacional Normalizado:** Camada Relacional/Fontes Externas/TCU/Esquema Normalizado/Diagramas_Relacional_TCU.drawio.svg
- **Criação e Povoamento das Tabelas Normalizadas:** Camada Relacional/Fontes Externas/TCU/Script SQL/TCU.sql
- **Script:** Camada Relacional/Fontes Externas/TCU/Script Python/script.7z

3.2.5 CEIS

- **Esquema Relacional Normalizado:** Camada Relacional/Fontes Externas/CEIS/Esquema Normalizado/Diagrama_Relacional_CEIS.drawio.svg
- **Criação e Povoamento das Tabelas Normalizadas:** Camada Relacional/Fontes Externas/CEIS/Script SQL/CEIS.sql
- **Script:** Camada Relacional/Fontes Externas/CEIS/Script Python/script.7z

3.2.6 Correios

- **Esquema Relacional Normalizado:** Camada Relacional/Fontes Externas/Correios/Esquema Normalizado/Diagrama_Relacional_Correios.drawio.svg
- **Script de Criação - Visões Relacionais:** Camada Relacional/Fontes Externas/Correios/Script SQL/CORREIOS.sql

4. Camada Semântica

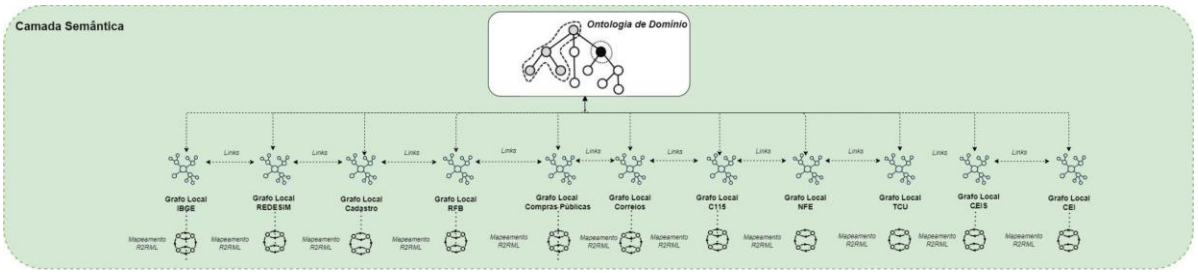


Figura 5. Camada Semântica.

A camada semântica (**Figura 5**) oferece uma visão semântica unificada de várias fontes de dados através de modelos ontológicos e recursos de Linked Data. O principal objetivo da camada semântica é permitir o acesso integrado aos dados.

Os principais componentes da Camada Semântica são:

- Ontologia de Domínio
- Grafo de conhecimento das Fontes Locais
- Ligações semânticas entre diferentes GCS

A Camada Semântica é obtida da integração semântica das fontes de dados. Chamamos de integração semântica o processo que faz uso de uma representação conceitual dos dados e seus relacionamentos para eliminar possíveis heterogeneidades. A construção da camada semântica é uma tarefa complexa que envolve quatro etapas:

(1) Aquisição de Conhecimento;
(2) Modelagem da Ontologia de Domínio;
(3) Especificação dos Grafos de Conhecimentos das Fontes de Dados Locais;
(4) Identificação de links entre recursos em diferentes grafos de Conhecimento;

Na etapa de Aquisição de Conhecimento, utilizou-se a abordagem Criação de Ontologias Orientada a Perguntas (COOP) através de Questões de Competência (QCs) em conjunto com os profissionais e especialistas da SEFAZ-MA afim de maior compreensão e aprofundamento no domínio e nos dados trabalhados.

Com base na compreensão dos conceitos nas QCs, ocorreu a identificação destes nos esquemas das fontes de dados, buscando identificar aspectos conceituais e seus relacionamentos, que serão utilizados para refinar as QCs originais, além de relacioná-las aos dados.

Como resultado desta etapa foram obtidas questões de competência refinadas fornecidas pelos especialistas de domínio da SEFAZ-MA, conforme demonstrado no catálogo abaixo.

4.1 Catálogo de Questões de Competência

Quais empresas de CNAEs de incidência de ICMS existem na RFB e não existem na SEFAZ?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?empresa ?empresa_ ?atividadeICMS WHERE {

   ?empresa a sefazma:Estabelecimento_RFB;
         rdfs:label ?empresa_;

         sefazma:tem_atividade_economica_principal|sefazma:tem_atividade_economica_secund
aria ?atividadeICMS.

   ?atividadeICMS a sefazma:Atividade_ICMS.

  MINUS {
         ?empresa owl:sameAs ?empresaSEFAZ.
   ?empresaSEFAZ a sefazma:Empresa_Cadastro.
  }

}
```

SQL

```
SELECT rfb.CNPJ
FROM RFB_ESTABELECIMENTO rfb
INNER JOIN
DOM_ATIVIDADE_ICMS dom
ON
rfb.ID_ATIVIDADE_ECONOMICA = dom.CNAE
WHERE
rfb.CNPJ NOT IN
(
SELECT cad_est.CNPJ
FROM CAD_ESTABELECIMENTO cad_est
)
```

Quais empresas na RFB ou SEFAZ não tem mais sócio pessoa física?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?o1 ?empresa WHERE {
{
   ?o1 a sefazma:Empresa_RFB;
      sefazma:cnpj ?empresa.

   MINUS {
         ?o1 sefazma:tem_sociedade_fisica ?x.
   }
 }
 UNION
 {
    ?o1 a sefazma:Empresa_Cadastro;
      sefazma:cnpj ?empresa.

          MINUS {
                ?o1 sefazma:tem_sociedade ?x.
                ?x a sefazma:Sociedade_PF.
    }
 }
}
```

SQL

```
SELECT CNPJ_RAIZ FROM RFB_EMPRESA
MINUS
(
   SELECT ID_EMPRESA FROM RFB_TEM_SOCIEDADE_FISICA tem
)
UNION ALL
SELECT CNPJ_RAIZ FROM CAD_EMPRESA
WHERE
CNPJ_RAIZ NOT IN
(
   SELECT CNPJ_RAIZ FROM CAD_TEM_SOCIEDADE
INNER JOIN
CAD_EMPRESA emp
ON
   TIPO = 'FISICA'
AND
   ID_EMPRESA = emp.ID
)
```

Quais empresas não estão ativas na RFB, mas estão na SEFAZ?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?cnpj_estabelecimento_rfb ?nome_fantasia_rfb ?cnpj_estabelecimento_cadastro
?nome_fantasia_cadastro ?tipo_situacao_rfb ?tipo_situacao_cadastro WHERE {

  ?estabelecimento_rfb a sefazma:Estabelecimento_RFB;
     sefazma:cnpj ?cnpj_estabelecimento_rfb;
     sefazma:nome_fantasia ?nome_fantasia_rfb;
     sefazma:tem_situacao_cadastral ?situacao_cadastral_rfb.

  ?situacao_cadastral_rfb rdfs:label ?tipo_situacao.
    FILTER(!CONTAINS(?tipo_situacao, 'ATIVA'))
    BIND(REPLACE(STR(?tipo_situacao),"(\\d|\\-|\\_)","") as ?tipo_situacao_rfb)

 ?estabelecimento_cadastro a sefazma:Estabelecimento_Cadastro;
       sefazma:cnpj ?cnpj_estabelecimento_cadastro;
       sefazma:nome_fantasia ?nome_fantasia_cadastro;
       sefazma:tem_situacao_cadastral ?situacao_cadastral_cadastro.

  ?situacao_cadastral_cadastro sefazma:tipo_situacao ?tipo_situacao_cadastro.

  ?estabelecimento_rfb owl:sameAs  ?estabelecimento_cadastro.
    FILTER((?tipo_situacao_rfb != ?tipo_situacao_cadastro) && ?tipo_situacao_cadastro = 'ATIVA')


}
```

SQL

```
SELECT sit.ID_SITUACAO_CADASTRAL AS SITUACAO_RFB, cadastro.ID AS
SITUACAO_CADASTRO
FROM
RFB_ESTABELECIMENTO est
INNER JOIN
RFB_SITUACAO_CADASTRAL sit
ON
REGEXP_SUBSTR(sit.ID_SITUACAO_CADASTRAL, '\w*', 1) <> 'ATIVA' and
est.ID_SITUACAO = sit.ID_SITUACAO_CADASTRAL
INNER JOIN
CAD_SITUACAO_CADASTRAL cadastro
ON
est.CNPJ = cadastro.CNPJ_CPF
AND cadastro.TIPO_SITUACAO = 'ATIVA'
```

Quais organizações são públicas?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?empresa_publica ?nome WHERE {
   {
   ?empresa_publica a sefazma:Ente_Publico;
                rdfs:label ?nome.
   }
   UNION
   {
      ?empresa_publica a foaf:Organization;
                rdfs:label ?nome;
                sefazma:tem_natureza_legal ?natureza_legal.
                    ?natureza_legal sefazma:codigo_natureza_legal ?codigo;
   FILTER(?codigo = "201-1"^^xsd:string || ?codigo = "2011"^^xsd:decimal)
   }
}
```

SQL

```
SELECT CNPJ, NOME FROM PUB_ENTES_PUBLICOS
UNION
SELECT CNPJ_RAIZ AS CNPJ, RAZAO_SOCIAL AS NOME
FROM RFB_EMPRESA WHERE ID_NATUREZA_LEGAL = 'EMPRESA_PUBLICA'
```

Quais situações cadastrais da RFB são incompatíveis com as existentes na SEFAZ?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?estabelecimento_rfb ?estabelecimento_cadastro ?cnpj_rfb ?cnpj_cadastro
?nome_fantasia_rfb ?nome_fantasia_cadastro ?situacao_rfb ?situacao_cadastro   WHERE {

 ?estabelecimento_rfb a sefazma:Estabelecimento_RFB;
    sefazma:cnpj ?cnpj_rfb;
    sefazma:nome_fantasia ?nome_fantasia_rfb;
    sefazma:tem_situacao_cadastral ?situacao_rfb.

  ?estabelecimento_rfb owl:sameAs ?estabelecimento_cadastro.
   ?estabelecimento_cadastro a sefazma:Estabelecimento_Cadastro;
         sefazma:cnpj ?cnpj_cadastro;
         sefazma:nome_fantasia ?nome_fantasia_cadastro;
         sefazma:tem_situacao_cadastral ?situacao_cadastro.
   OPTIONAL {
```

```
      BIND(?situacao_cadastro AS ?situacao_cadastro2)
         ?situacao_rfb owl:sameAs ?situacao_cadastro2.
   }
   FILTER(!BOUND(?situacao_cadastro2))

}
```

SQL

```
SELECT sit.ID_SITUACAO_CADASTRAL AS SITUACAO_RFB, cadastro.ID AS
SITUACAO_CADASTRO
FROM
RFB_ESTABELECIMENTO est
INNER JOIN
RFB_SITUACAO_CADASTRAL sit
ON
est.ID_SITUACAO = sit.ID_SITUACAO_CADASTRAL
INNER JOIN
CAD_SITUACAO_CADASTRAL cadastro
ON
est.CNPJ = cadastro.CNPJ_CPF
AND
est.ID_SITUACAO <> cadastro.ID
```

Quem são os contribuintes?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?contribuinte ?inscricao_estadual WHERE {
   ?contribuinte a sefazma:Contribuinte_Geral;
        rdfs:label ?inscricao_estadual.
}
```

SQL

```
SELECT * FROM "UFC2"."CAD_CONTRIBUINTE_GERAL"
```

Quem são os contribuintes ST?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
```

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?contribuinte ?inscricao_estadual WHERE {
   ?contribuinte a sefazma:Substituto_Tributario;
        rdfs:label ?inscricao_estadual.
}
```

SQL

```
SELECT * FROM "UFC2"."CAD_SUBSTITUTO_TRIBUTARIO"
```

Quem são os contribuintes SN?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?contribuinte ?inscricao_estadual WHERE {
   ?contribuinte a sefazma:Simples_Nacional;
        rdfs:label ?inscricao_estadual.
}
```

SQL

```
SELECT *  FROM "UFC2"."CAD_SIMPLES_NACIONAL"
```

Quem são os contribuintes Extra-Cadastro?

```
SELECT ?contribuinte ?inscricao_estadual WHERE {
   ?contribuinte a sefazma:Extra_Cadastro;
        rdfs:label ?inscricao_estadual.
}
```

```
SELECT * FROM "UFC2"."CAD_EXTRA_CADASTRO"
```

Quem são os Não Contribuintes?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
SELECT ?contribuinte ?inscricao_estadual WHERE {
   ?contribuinte a sefazma:Nao_Contribuinte;
        rdfs:label ?inscricao_estadual.
}
```

SQL

```
SELECT * FROM "UFC2"."CAD_NAO_CONTRIBUINTE"
```

Quais as empresas na SEFAZ não tem sócios?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?empresa WHERE {
   ?empresa a sefazma:Empresa_Cadastro.
   MINUS {
      ?empresa sefazma:tem_sociedade ?sociedade.
   }
}
```

SQL

```
SELECT TO_CHAR(CNPJ_RAIZ) FROM "UFC2"."CAD_EMPRESA"
MINUS
SELECT TO_CHAR(URI_EMPRESA) FROM "UFC2"."CAD_TEM_SOCIEDADE"
```

Quais são os órgãos públicos por esfera do poder (Executivo, Legislativo e Judiciário) e por nível (Federal, Estadual e Municipal)?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?ente_publico ?ente ?esfera ?nivel WHERE {
   ?ente_publico a sefazma:Ente_Publico;
            rdfs:label ?ente;
                        sefazma:esfera ?esfera;
            sefazma:nivel ?nivel.
}
```

SQL

```
SELECT CNPJ, NOME, ESFERA , NIVEL
```

```
FROM COMPRAS_ENTES_PUBLICOS entes_publicos
ORDER BY ESFERA, NIVEL
```

O que os órgãos públicos estão comprando?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?item_compra WHERE {
   ?compra a sefazma:Compra_Publica;
         sefazma:tem_item ?item.
   ?item rdfs:label ?item_compra.
}
```

SQL

```
SELECT itens.NUMERO_ITEM,
itens.DESCRICAO,
itens.INFORMACOES_ADICIONAIS
FROM COMPRAS_ITEM itens
```

Quem vende para órgãos públicos?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?fornecedor_compra ?cnpj_cpf WHERE {
   ?compra a sefazma:Compra_Publica;
         sefazma:tem_fornecedor ?fornecedor.
   ?fornecedor rdfs:label ?fornecedor_compra;
            sefazma:cnpj_cpf ?cnpj_cpf.
}
```

SQL

```
SELECT DISTINCT fornecedor.CNPJ_CPF, fornecedor.NOME
FROM COMPRAS_FORNECEDOR fornecedor
INNER JOIN
COMPRAS_COMPRA_PUBLICA compras
ON
compras.CPF_CNPJ_FORNECEDOR = fornecedor.CNPJ_CPF
```

Quais fornecedores estão com alguma restrição ou sanção nas fontes de dados integradas: (CEI, CEIS ou TCU)?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?fornecedor_restrito ?cnpj_cpf WHERE {
    ?fornecedor a sefazma:Fornecedor_Restrito;
            foaf:name ?fornecedor_restrito;
            sefazma:cnpj_cpf ?cnpj_cpf.
}
```

SQL

```
SELECT * FROM RESTRICOES_FORNECEDOR_RESTRITO
```

Quais fornecedores estão com alguma restrição em uma determinada data?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?fornecedor ?data_transito_julgado WHERE {
   ?processo a sefazma:Processo;
           sefazma:tem_fornecedor_restrito ?fornecedor;
           sefazma:data_transito_julgado ?data_transito_julgado;
           sefazma:data_final ?data_final_processo.

 FILTER(
       ?data_transito_julgado
       = "18/10/2014"
   )

} LIMIT 100
```

SQL

```
SELECT fornecedores.CNPJ_CPF,
fornecedores.NOME,
processo.DATA_TRANSITO_JULGADO AS DATA_INICIAL,
processo.DATA_FINAL
FROM
RESTRICOES_FORNECEDOR_RESTRITO fornecedores
INNER JOIN
```

```
RESTRICOES_PROCESSO processo
ON
fornecedores.CNPJ_CPF =  REGEXP_REPLACE(processo.CNPJ_CPF, '\D', '')
WHERE
DATA_TRANSITO_JULGADO
LIKE ' 14-10%'
```

Quais os Fornecedores Maranhenses?

SPARQL

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT * WHERE {
        ?fornecedor_restrito a sefazma:Fornecedor_Restrito ;
   sefazma:tem_unidade_federativa ?uf.

   ?uf a sefazma:Unidade_Federativa;
      rdfs:label ?estado.

   FILTER(?estado = "MA")

}
```

SQL

```
SELECT *
FROM RESTRICOES_FORNECEDOR_RESTRITO
WHERE
UF = 'MA'
```

4.2 Ontologia de Domínio

O processo de Modelagem da Ontologia de Domínio requer uma análise minuciosa dos diferentes esquemas das fontes de dados, objetivando a criação de um modelo comum para resolver o problema da heterogeneidade das fontes. Nessa atividade, deve-se procurar reusar ontologias já existentes, e formalizar o modelo usando os padrões como RDF-S e OWL.

A Modelagem e Construção da Ontologia de Domínio foi orientada através dos conceitos identificados nas QCs e na reutilização dos esquemas das fontes de dados. Foram feitos inicialmente modelos ontológicos em alto nível, que foram então agregados em um esquema global único, por conseguinte, esse modelo ontológico foi implementado em OWL através da ferramenta Protégé[1], tendo a definição de regras e axiomas necessários de modo a enriquecer a capacidade de inferência e descoberta sobre os dados. Abaixo são informados os links da ontologia de domínio implementada em OWL, sua especificação e o link para visualização em forma de grafo.

OWL: **Camada Semantica/Ontologia Dominio/OWL/ontologia_dominio.ttl**

Especificação: **Camada Semantica/Ontologia Dominio/Especificacao/index-pt.html**

4.3 Grafos de Conhecimento Locais (GCLs)

Para cada fonte de dados, é construído um grafo de conhecimento, tratando-se os dados já transportados na camada relacional construída, composto por tabelas relacionais que usam o mesmo vocabulário da ontologia de domínio. Assim, os mapeamentos das visões relacionais para a ontologia tem a semântica refletida a partir dos mapeamentos diretos, removendo a complexidade pertencente às visões.

Assim, considerando que cada fonte de dados exporta uma visão RDF definida sobre o esquema relacional da fonte de dados, o mapeamento de cada GCL foi realizado usando a linguagem R2RML sobre as tabelas normalizadas tendo como base o vocabulário da ontologia local de cada fonte de dados. Abaixo são expostos para cada GCL sua ontologia local contendo links para o diagrama de classes, sua implementação em owl e sua especificação em conjunto com os mapeamentos.

[1] https://protege.stanford.edu/

4.3.1 GCL CAD_SEFAZ

Ontologia Local

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL Cadastro/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_Cadastro.drawio.svg
- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL Cadastro/Ontologia Local/OWL/Cadastro.owl
- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL Cadastro/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL Cadastro/Mapeamentos/Cadastro.ttl

4.3.2 GCL RFB

Ontologia Local:

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL RFB/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_RFB.drawio.svg
- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL RFB/Ontologia Local/OWL/RFB.ttl
- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL RFB/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL RFB/Mapeamentos/RFB.ttl

4.3.3 GCL REDESIM

Ontologia Local:

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL REDESIM/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_REDESIM.drawio.svg
- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL REDESIM/Ontologia Local/OWL/REDESIM.owl

- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL REDESIM/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL REDESIM/Mapeamentos/REDESIM.ttl

4.3.4 GCL IBGE CNAE

Ontologia Local:

- **Diagrama de Classe:** Camada Semantica/Grafos de Conhecimento Locais/GCL IBGE CNAE/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_IBGE_CNAE.drawio.svg

- **Arquivo OWL:** Camada Semantica\Grafos de Conhecimento Locais\GCL IBGE Localizacao\Ontologia Local\OWL\IBGE_Localizacao.ttl

- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL IBGE CNAE/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL IBGE CNAE/Mapeamentos/IBGE_CNAE.ttl

4.3.5 GCL IBGE Localização

Ontologia Local:

- **Diagrama de Classe:** Camada Semantica/Grafos de Conhecimento Locais/GCL IBGE Localizacao/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_Localizacao.drawio.svg

- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL IBGE CNAE/Ontologia Local/OWL/IBGE_CNAE.ttl

- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL IBGE Localizacao/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL IBGE Localizacao/Mapeamentos/IBGE_Localizacao.ttl

4.3.6 GCL-Correios

Ontologia Local:

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL Correios/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_Correios.drawio.svg

- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL Correios/Ontologia Local/OWL/Correios.owl

- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL Correios/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL Correios/Mapeamentos/Correios.ttl

4.3.7 GCL C115

Ontologia Local:

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL C115/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_C115.drawio.svg

- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL C115/Ontologia Local/OWL/C115.owl

- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL C115/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica\Grafos de Conhecimento Locais\GCL C115\Mapeamentos\C115.ttl

4.3.8 GCL Compras Públicas

Ontologia Local:

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL Compras Publicas/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_Compras.drawio.svg

- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL Compras Publicas/Ontologia Local/OWL/Compras Municipais.ttl

- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL Compras Publicas/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL Compras Publicas/Mapeamentos/Compras.ttl

4.3.9 GCL TCU

Ontologia Local:

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL TCU/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_TCU.drawio.svg

- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL TCU/Ontologia Local/OWL/TCU.owl

- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL TCU/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL TCU/Mapeamentos/TCU.ttl

4.3.10 GCL CEIS

Ontologia Local

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL CEIS/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_CEIS.drawio.svg

- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL CEIS/Ontologia Local/OWL/CEIS.owl

- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL CEIS/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL CEIS/Mapeamentos/CEIS.ttl

4.3.11 GCL_CEI

Ontologia Local

- **Diagrama de Classes:** Camada Semantica/Grafos de Conhecimento Locais/GCL CEI/Ontologia Local/Diagrama de Classes/Diagrama_Ontologia_CEI.drawio.svg
- **Arquivo OWL:** Camada Semantica/Grafos de Conhecimento Locais/GCL CEI/Ontologia Local/OWL/CEI.owl
- **Especificação:** Camada Semantica/Grafos de Conhecimento Locais/GCL CEI/Ontologia Local/Especificacao/index-pt.html

Mapeamentos: Camada Semantica/Grafos de Conhecimento Locais/GCL CEI/Mapeamentos/CEI.ttl

4.4 Visões de Ligação

Visões de ligação são usadas para relacionar recursos em diferentes GCLs que tenham tenham entidades em comum, portanto os seus vocabulários devem ter classes em comum. Devem ser especificadas ligações "same-As" entre os objetos, de classes em comum, que representem a mesma entidade do mundo real.

Para a criação das visões de ligação, foram estabelecidas regras para descoberta dos links entre os grafos de conhecimento locais que especificam as condições para inferência de links "*same-As*" utilizando o predicado ***owl:sameAs*** através de junção em mapeamentos para estabelecer relações entre recursos dos grafos locais. As regras e as visões de ligação geradas são apresentadas a seguir.

4.4.1 Regra de Linkage Pessoa [Cadastro -> RFB] (vcad:cpf – foaf:name)

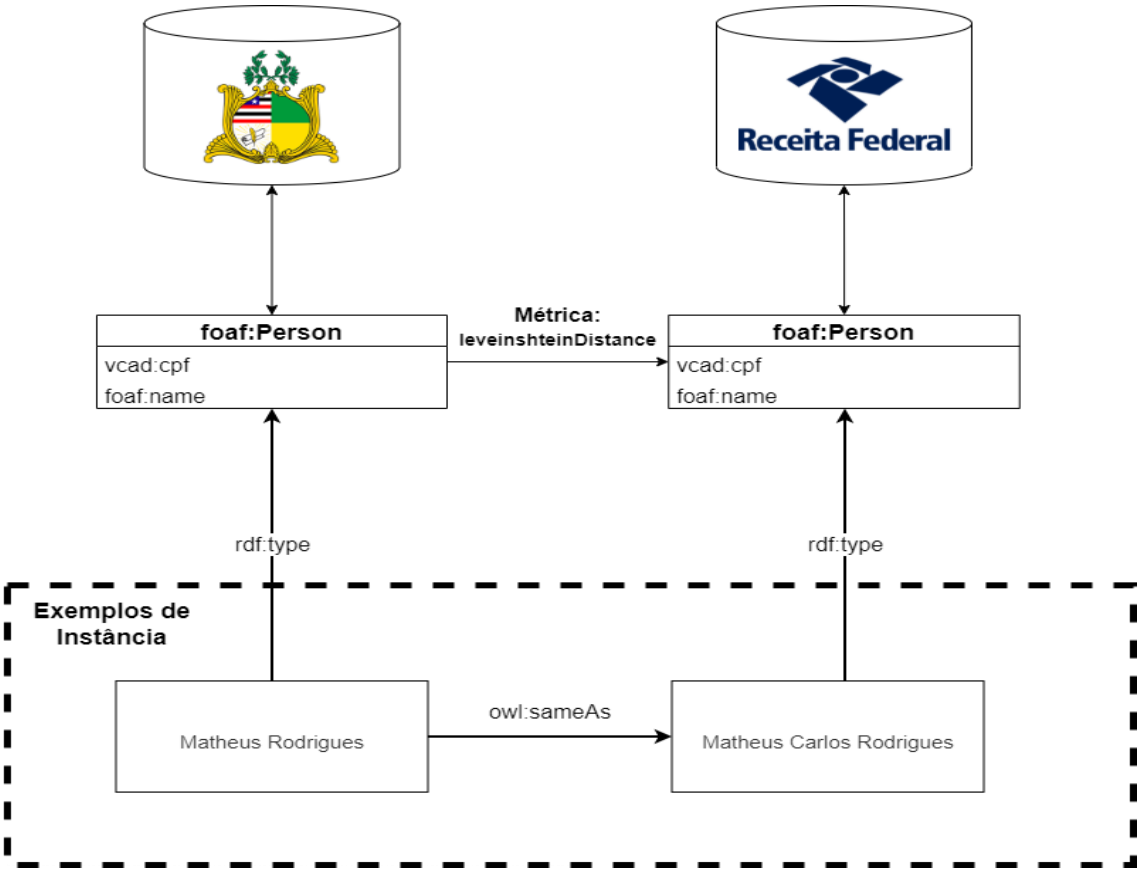


Figura 6. Regra de Linkage Pessoa [Cadastro -> RFB] (vcad:cpf – foaf:name).

Link sameAs RDF

Camada Semantica/Visoes de Ligacao/CADASTRO_RFB/Pessoa/sameAs_RDF.ttl

Link sameAs Relacional

Camada Semantica/Visoes de Ligacao/CADASTRO_RFB/Pessoa/sameAs_Relacional.sql

4.4.2 Regra de Linkage Pessoa [Cadastro -> REDESIM] (vcad:cpf)

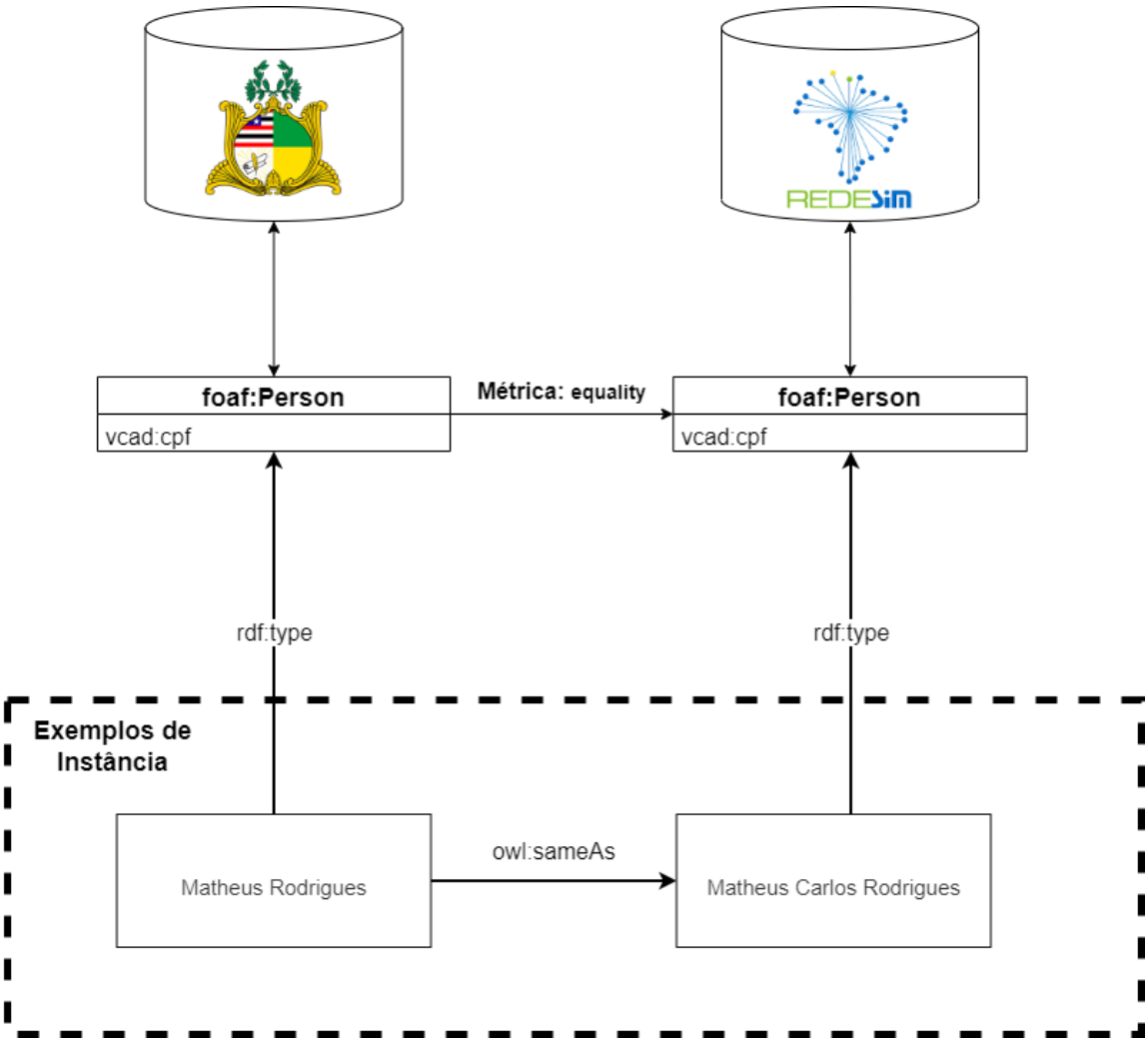


Figura 7. Regra de Linkage Pessoa [Cadastro -> REDESIM] (vcad:cpf).

Link sameAs RDF

Camada Semantica/Visoes de Ligacao/CADASTRO_REDESIM/Pessoa/sameAs_RDF.ttl

Link sameAs Relacional

Camada Semantica/Visoes de Ligacao/CADASTRO_REDESIM/Pessoa/sameAs_Relacional.sql

4.4.3 Regra de Linkage Endereço [Cadastro -> RFB] (vcad:cnpj)

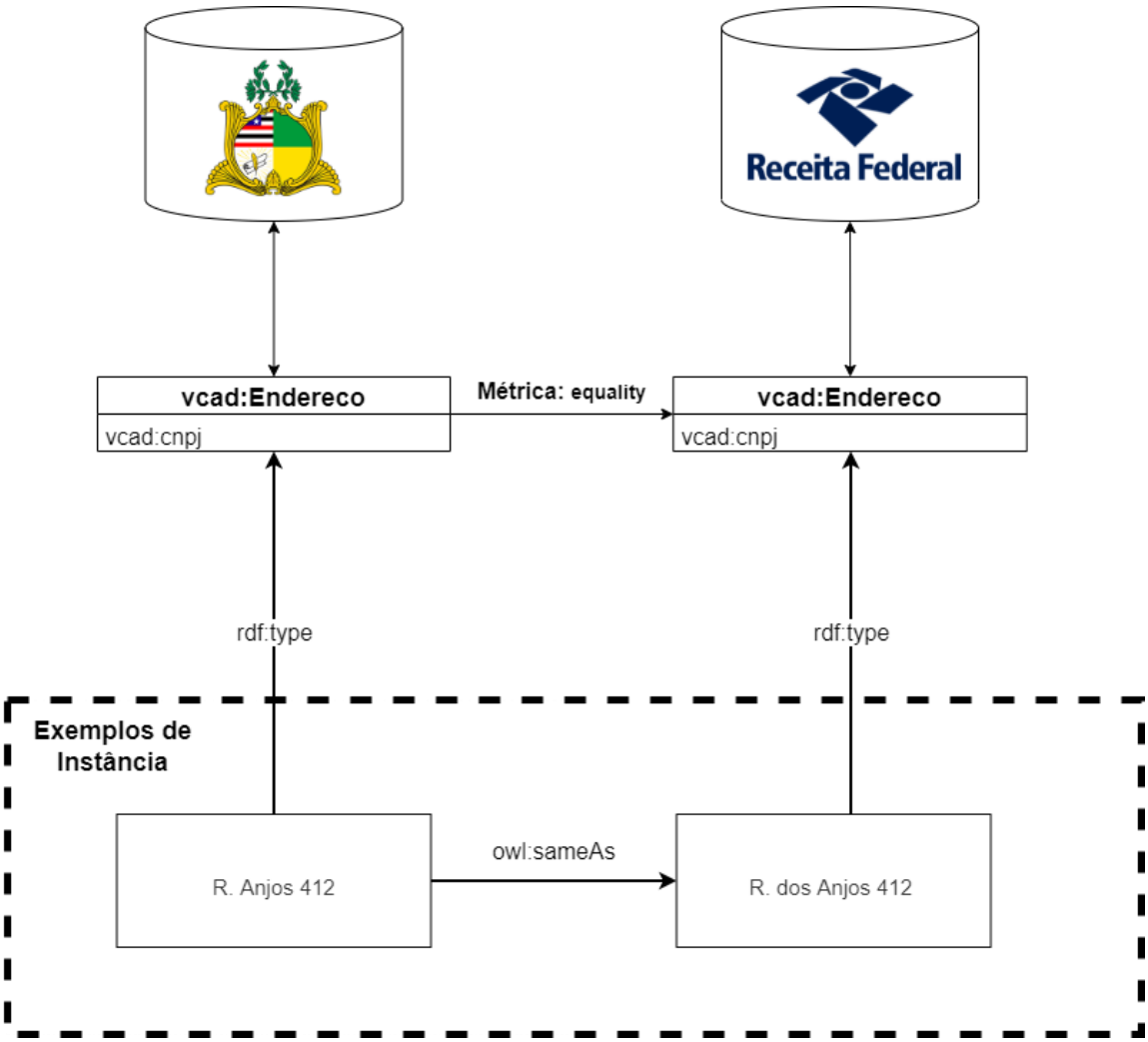


Figura 8. Regra de Linkage Endereço [Cadastro -> RFB] (vcad:cnpj).

Link sameAs RDF

Camada Semantica/Visoes de Ligacao/CADASTRO_RFB/Endereco/sameAs_RDF.ttl

Link sameAs Relacional

Camada Semantica/Visoes de Ligacao/CADASTRO_RFB/Endereco/sameAs_Relacional.sql

4.4.4 Regra de Linkage Endereço [Cadastro -> REDESIM] (vcad:cnpj)

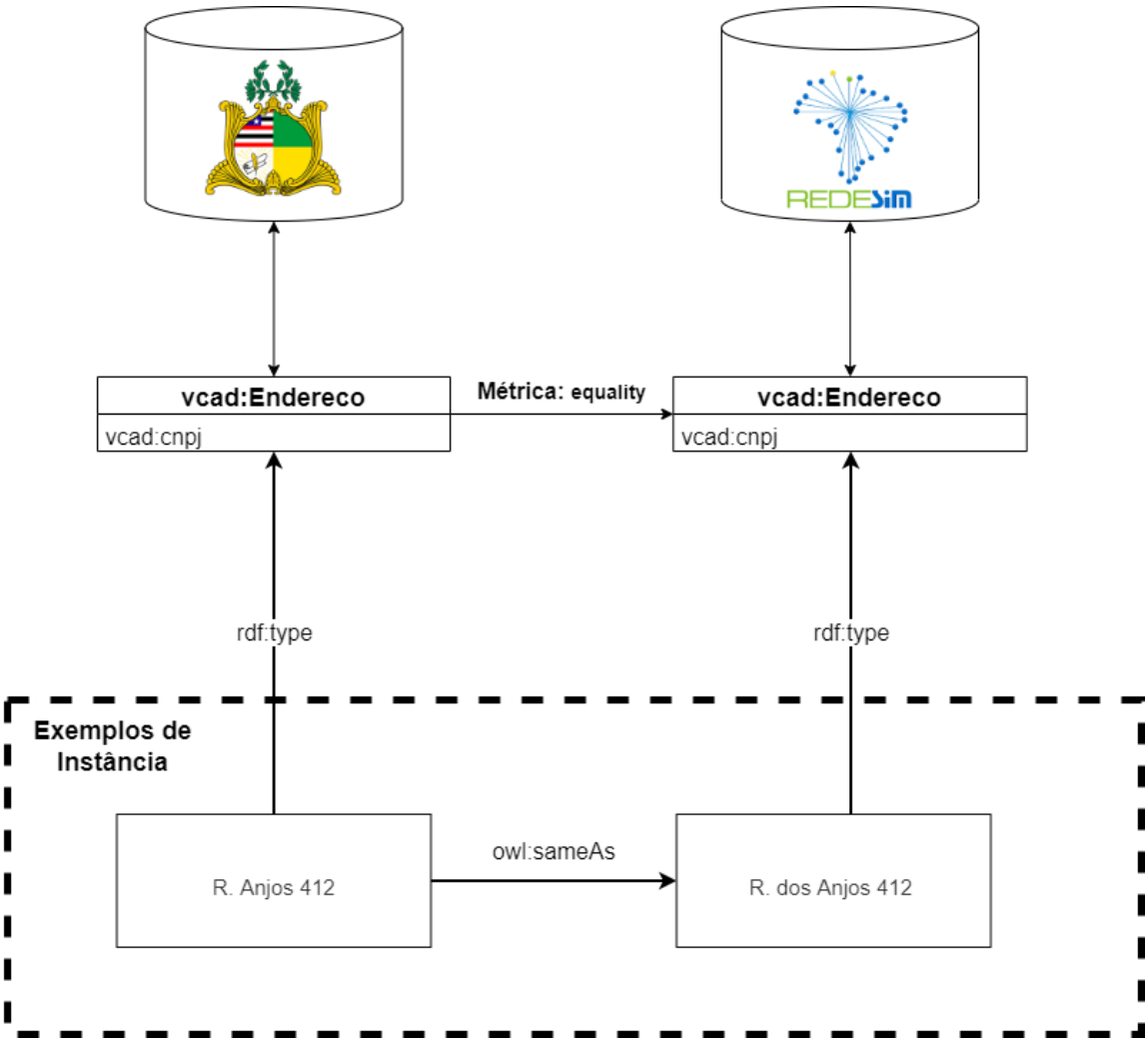


Figura 9. Regra de Linkage Endereço [Cadastro -> REDESIM] (vcad:cnpj).

Link sameAs RDF

Camada Semantica/Visoes de Ligacao/CADASTRO_REDESIM/Endereco/sameAs_RDF.ttl

Link sameAs Relacional

Camada Semantica/Visoes de Ligacao/CADASTRO_REDESIM/Endereço/sameAs_Relacional.sql

4.4.5 Regra de Linkage de Estabelecimento [Cadastro -> RFB] (vcad:cnpj)

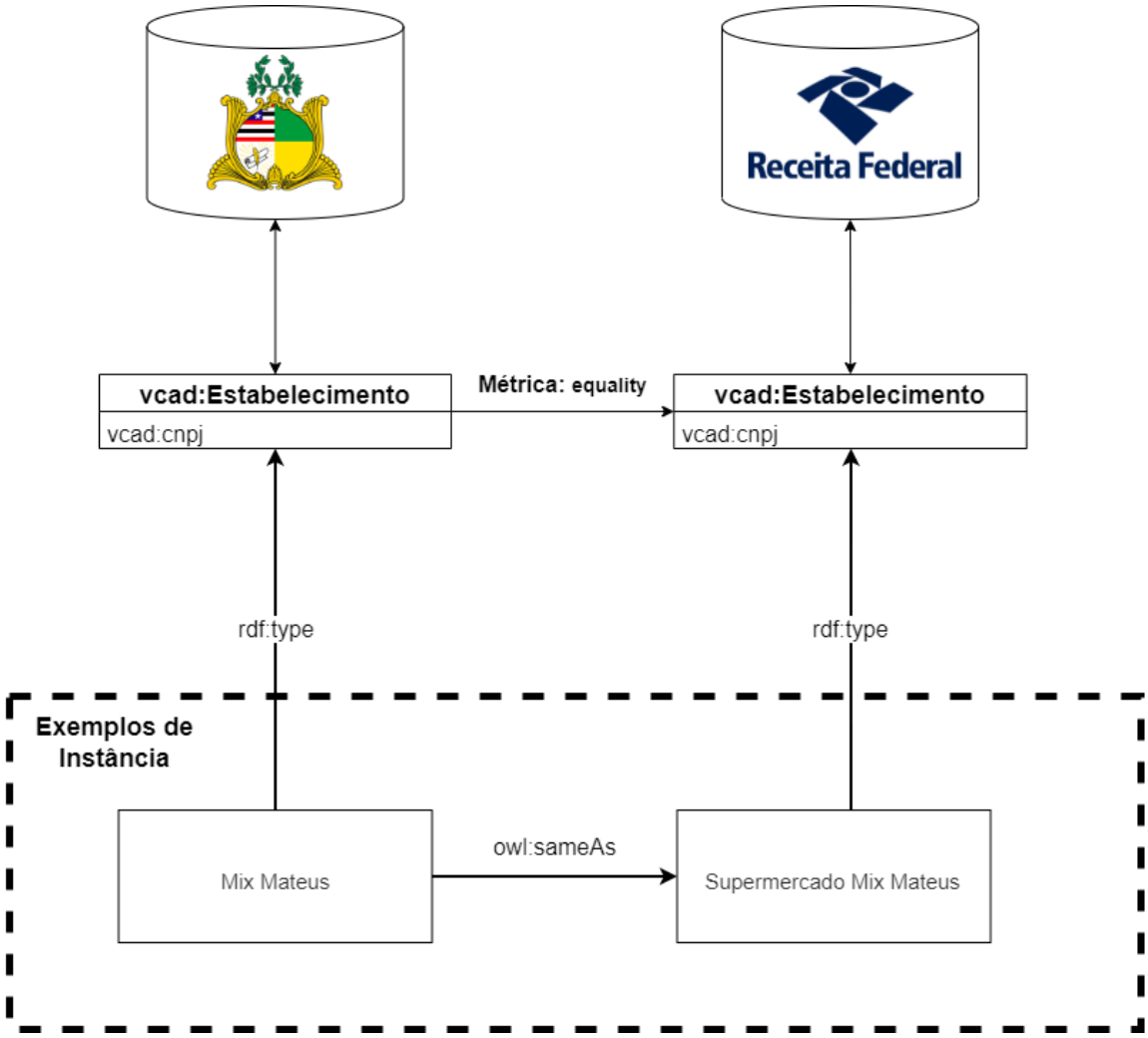


Figura 16. Regra de Linkage de Estabelecimento [Cadastro -> RFB] (vcad:cnpj)

Link sameAs RDF

Camada Semantica/Visoes de Ligacao/CADASTRO_RFB/Estabelecimento/sameAs_RDF.ttl

Link sameAs Relacional

Camada Semantica/Visoes de Ligacao/CADASTRO_RFB/Estabelecimento/sameAs_Relacional.sql

4.4.6 Regra de Linkage de Estabelecimento [Cadastro -> REDESIM] (vcad:cnpj)

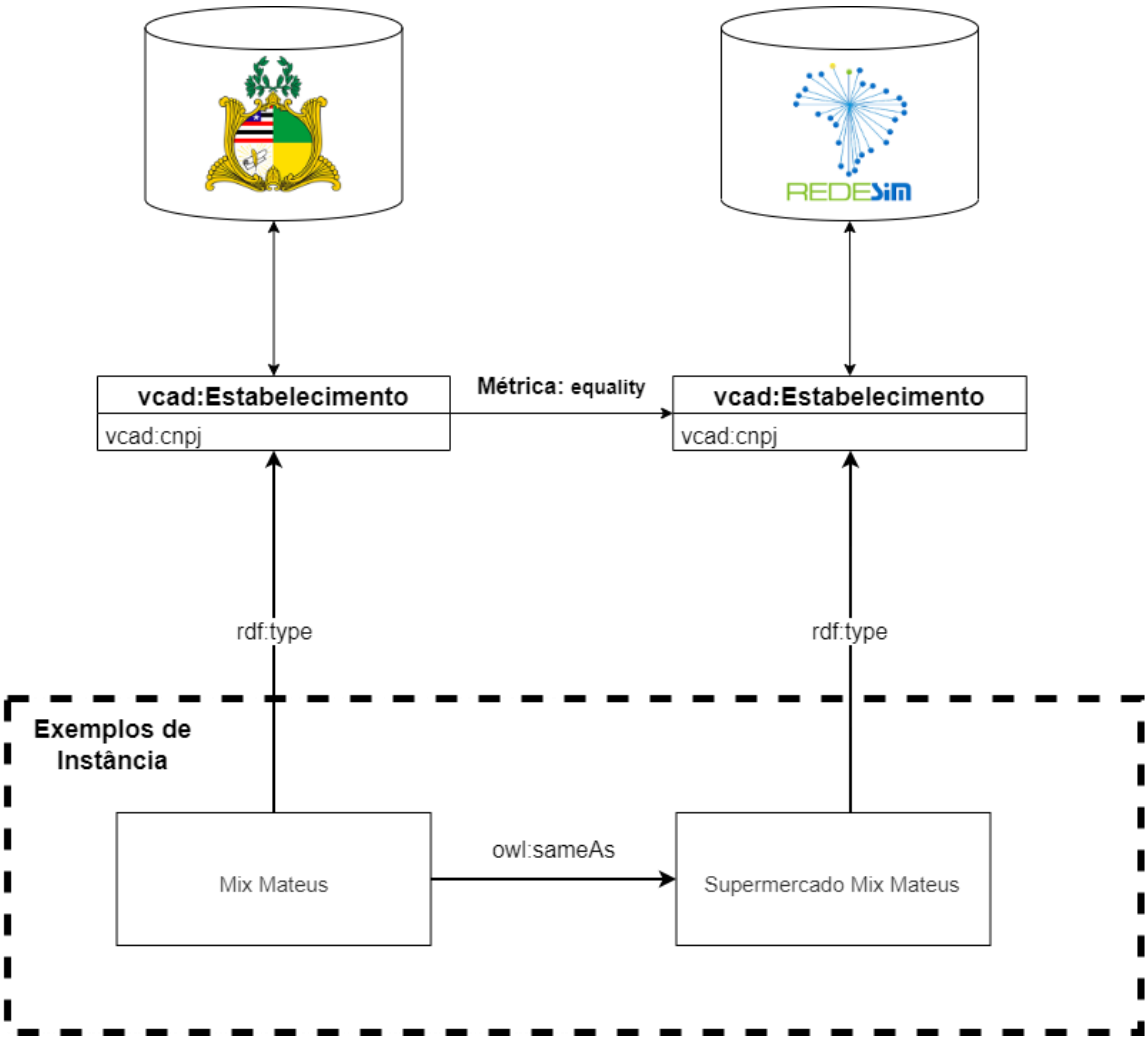


Figura 17. Regra de Linkage de Estabelecimento [Cadastro -> REDESIM] (vcad:cnpj).

Link sameAs RDF

Camada Semantica/Visoes de Ligacao/CADASTRO_REDESIM/Estabelecimento/sameAs_RDF.ttl

Link sameAs Relacional

Camada Semantica/Visoes de Ligacao/CADASTRO_REDESIM/Estabelecimento/sameAs_Relacional.sql

4.4.7 Regra de Linkage de Empresa [Cadastro -> RFB] (vcad:cnpj_raiz)

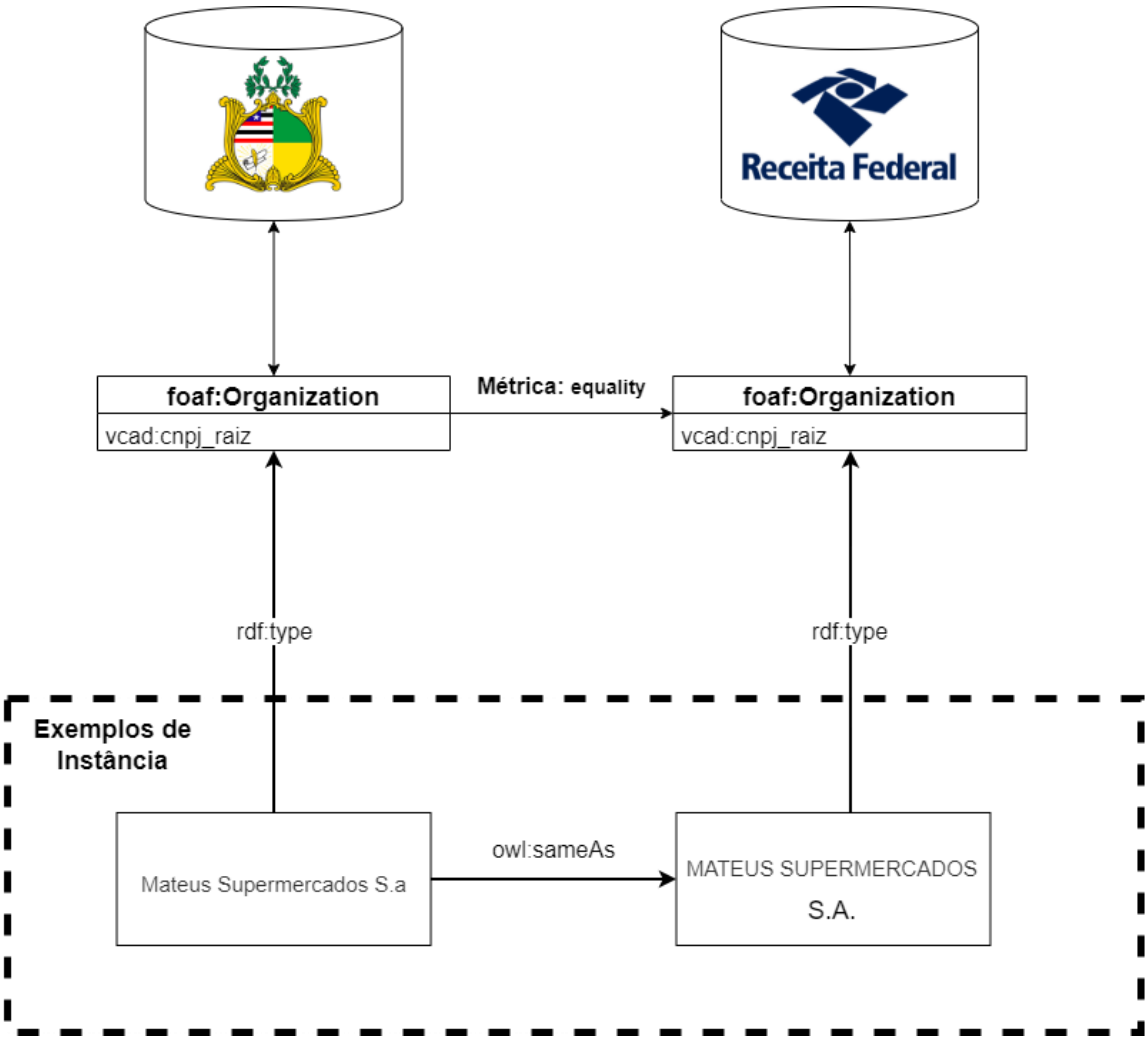


Figura 18. Regra de Linkage de Empresa [Cadastro -> RFB] (vcad:cnpj_raiz).

Link sameAs RDF

Camada Semantica/Visoes de Ligacao/CADASTRO_RFB/Empresa/sameAs_RDF.ttl

Link sameAs Relacional

Camada Semantica/Visoes de Ligacao/CADASTRO_RFB/Empresa/sameAs_Relacional.sql

4.4.8 Regra de Linkage de Empresa [Cadastro -> REDESIM] (vcad:cnpj_raiz)

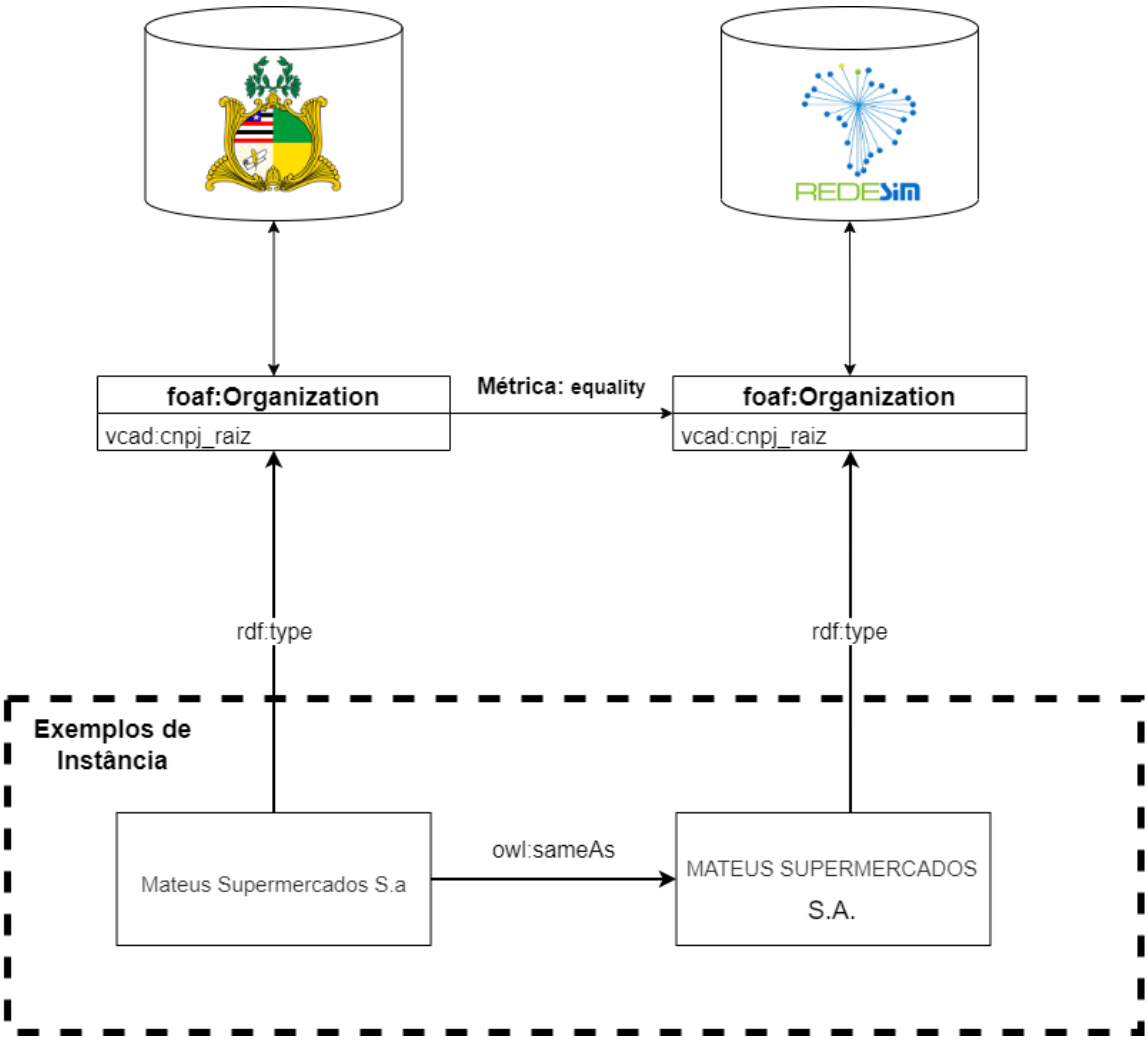


Figura 19. Regra de Linkage de Empresa [Cadastro -> REDESIM] (vcad:cnpj_raiz).

Link sameAs RDF

_Camada Semantica/Visoes de Ligacao/CADASTRO_REDESIM/Empresa/sameAs_RDF.ttl

Link sameAs Relacional

_Camada Semantica/Visoes de Ligacao/CADASTRO_REDESIM/Empresa/sameAs_Relacional.sql

5. Camada de Acesso e Integração de Dados

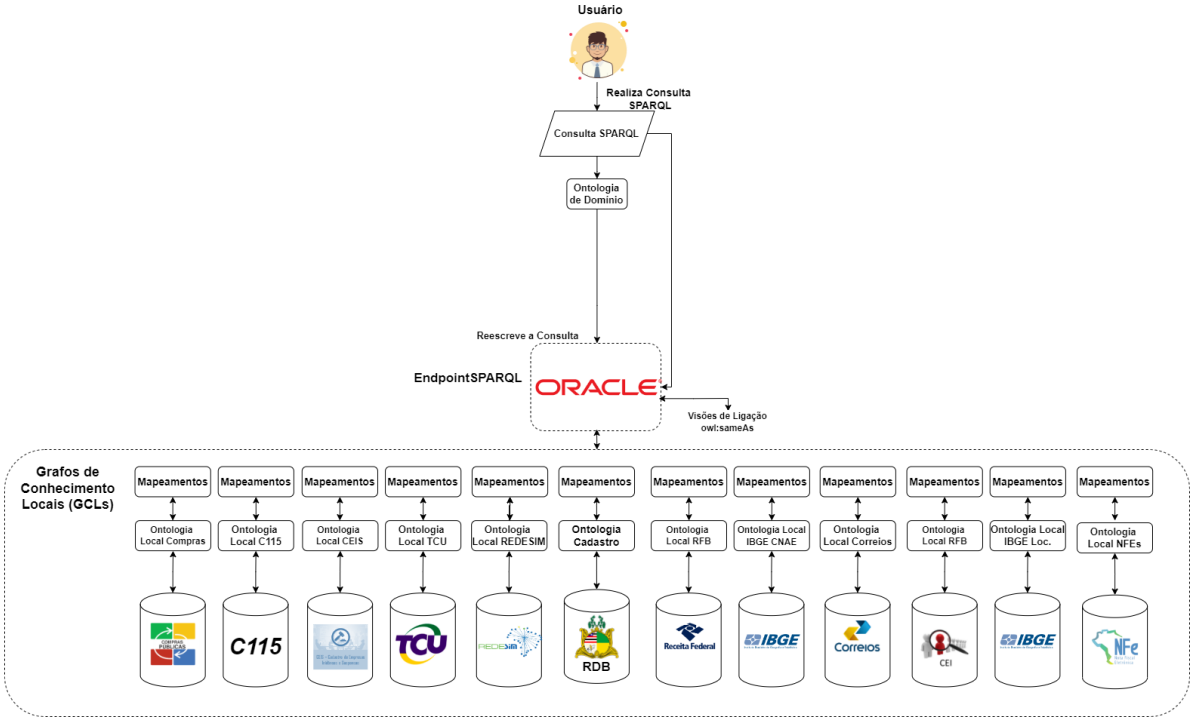
A principal função da camada de acesso é permitir o acesso integrado aos dados usando a Ontologia de Domínio. Portanto, os usuários e aplicações podem realizar consultas de forma transparente, sem ter que entender sobre as fontes, esquema ou estrutura dos dados.

A plataforma fornece as seguintes capacidades:

1) Integração de fontes de dados internas e externas com abordagem virtual;
2) Visão unificada em fontes de dados distribuídas e heterogêneas;
3) Criação e Consulta *on-the-fly* sob demanda dos grafos locais;
4) Possibilidade de Atualização dos Grafos Locais e suas relações bem como identificação de novas visões de ligação (sameAs);
5) Extensibilidade quanto à quantidade de fontes de dados a serem integradas;
6) Capacidade de conexão a vários repositórios ou endpoints através de federação local ou remota de fontes RDF ou não virtualmente integradas e expostas por meio de uma interface SPARQL.

Como forma de prover flexibilidade no acesso dos dados integrados e consulta ao grafo semântico, foram fornecidos 2 endpoints, 1 baseado no Oracle RDF Graph, com vista na capacidade de usufruir da licença Oracle da SEFAZ-MA, e um outro utilizando o GraphDB + Ontop, tendo foco na capacidade de exploração e visualização gráfica do grafo semântico, ambas as propostas trabalhando com uma abordagem virtual.

5.1.1 Endpoint SPARQL Oracle

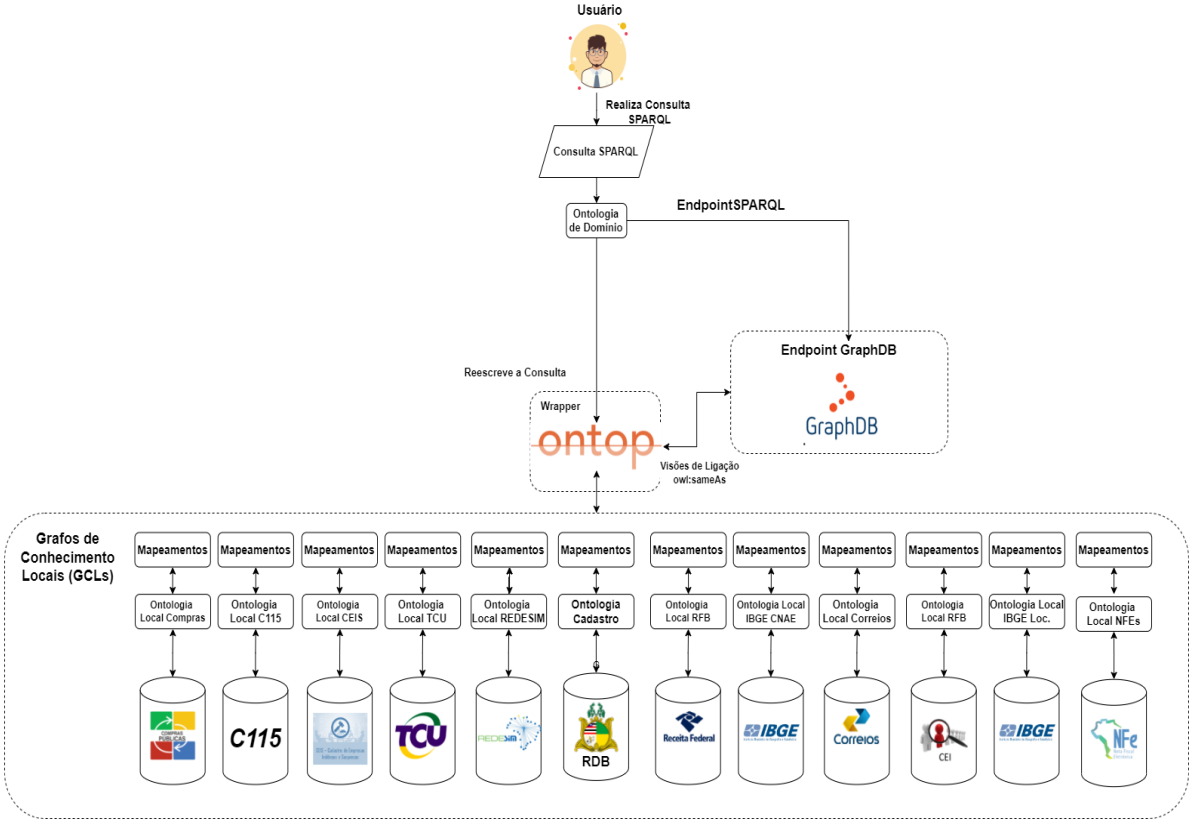


A Figura acima apresenta os componentes da camada de acesso e Integração dos Dados utilizados para construção do GC-SEFAZMA. Para tanto, o Oracle foi utilizado como Endpoint repositório de acesso aos GCLs. O Oracle fornece através de uma RDF View, acesso aos dados integrados na camada relacional com base em mapeamentos R2RML, traduzidos em formato RDF.

A realização de consultas SPARQL por usuários ou aplicações é transparente, sendo tratada, reescrita e otimizada pelo Oracle RDF, possibilitando um acesso contínuo e concorrente.

- **Endpoint SEFAZ Oracle:** http://10.33.96.18:8080/orardf-21.3.1/
- **Tutorial de Acesso:** Camada de Acesso e Integracao\Endpoint Oracle\Tutorial de Acesso ao Oracle RDF Graph.pdf

5.1.2 Endpoint SPARQL GraphDB + Oracle



A Figura acima apresenta os componentes da camada de acesso e Integração dos Dados utilizados para construção do GC-SEFAZMA. Para tanto, o GraphDB foi utilizado como Endpoint repositório de acesso aos GCLs. Já o Ontop foi utilizado como *wrapper* acoplado ao GraphDB para construção dos grafos locais utilizando a abordagem virtual, tendo como parâmetro de entrada a ontologia de domínio e mapeamentos contidos na especificação de cada GCL.

O Ontop adota um workflow para construção do grafo local virtual com base em 2 estágios (offline e online). Onde no estágio offline, Ontop realiza o processo de leitura e processamento da ontologia, mapeamentos e verifica as restrições de integridade do banco de dados, compilando a ontologia nos mapeamentos e gerando os *T-mappings*.

No segundo estágio, O ontop segue a abordagem *Ontology-Based Data Access* (OBDA), onde é feito um processamento online de uma dada consulta SPARQL e sua tradução em SQL usando os mapeamentos. Ainda nesse estágio são feitas otimizações na consulta SQL, sendo executada pelo mecanismo de banco de dados e ao final realizando a tradução do resultado para RDF.

- **Endpoint SEFAZ GraphDB:** http://10.33.96.18:7200/sparql
- **Arquivo de Configuração do Endpoint:** Camada de Acesso e Integracao\Endpoint GraphDB\GRAFO_SEFAZMA_PRODUCAO.zip