Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Proposta de Arquitetura de Publicação Automatizada de Dados Abertos Conectados Utilizando Meta-Dados e Ontologias

Luiz Carlos Barbosa Martins

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Márcio de Carvalho Victorino

Coorientadora
Prof.a Dr.a Maristela Terto de Holanda

Brasília
2018.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Proposta de Arquitetura de Publicação Automatizada de Dados Abertos Conectados Utilizando Meta-Dados e Ontologias

Luiz Carlos Barbosa Martins

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Márcio de Carvalho Victorino (Orientador)
FCI/UnB

Prof. Dr. Marcelo Ladeira
CIC/UnB

Prof.a Dr.a Ana Paula Bernardi da Silva
MGCTI/UCB

Prof.a Dr.a Aletéia Patrícia Favacho de Araujo Von Paumgartten
Coordenadora do Programa de Pós-graduação em Computação Aplicada

Brasília, 4 de dezembro de 2018.

Dedicatória

Dedico este trabalho à minha mãe, Glória, e ao meu pai, Jorge(*in memorian*) que me ensinaram os valores que norteiam minha vida.

Aos meus irmãos, meus melhores amigos, e minhas sobrinhas e afilhada, luzes de alegria na minha vida.

Á minha querida avó Conceição e demais familiares.

Aos grande educadores do Brasil nas pessoas de Darcy Ribeiro, Anísio Teixeira e Paulo Freire, que acreditavam neste povo no qual faço parte.

Agradecimentos

Agradeço à Deus por me conceder o dom da vida e a possibilidade de lutar pelos meus ideais.

À Juliana, esposa, companheira e amiga que esteve ao meu lado durante toda esta caminhada, pela compreensão nos momentos difíceis e apoio.

Á UnB que é minha segunda casa, onde contribuo na construção de uma sociedade mais justa como servidor no Centro de Informática e que me abriu as portas acadêmicas para desenvolvimento deste trabalho.

À todos os meus professores do Programa de Pós-Graduação em Computação Aplicada que pelos conhecimentos que me passaram e pelos laços de amizade criados especialmente na pessoa do Prof. Marcelo Ladeira.

Ao meu orientador Prof. Márcio Victorino, que esteve presente no devolvimento deste trabalho sempre sendo solícito e compartilhando seus conhecimentos. À minha coorientadora Profª. Maristela Terto pelo auxílio e incentivo.

Aos meus amigos e colegas do CPD que estiveram presentes durante esse período, dando suporte nos momentos mais preciosos e viabilizando este trabalho especialmente nas pessoas do Jacir e Consuelo.

Resumo

O governo brasileiro tem investido no aumento da transparência de suas ações visando incentivar a participação ativa da sociedade na gestão. Neste sentido, uma das principais ações é a abertura dos dados de órgãos federais para a comunidade. Hoje existe uma quantidade considerável de dados abertos nos poderes Executivo, Legislativo e Judiciário, além das esferas da União, Estados ou Municípios, mas não existe uma maneira clara de realizar a conexão entre estes dados e a sua publicação. Esta pesquisa visa propor uma arquitetura que auxilie as instituições a abrir seus dados de maneira mais eficiente e agregue o máximo possível de qualidade a eles. A qualidade dos dados está relacionada a dois fatores: o dado ter possibilidade de ser ligado a outros dados e ser o mais atual possível. Assim, propomos um modelo que busca agregar diversas tecnologias que possibilitem que os dados possam ser descritos semanticamente, tornando assim dados conectados e viabilizando as ligações, além de propiciar que os conjuntos de dados possam ser atualizados sem a intervenção humana, garantindo intervalos reduzidos entre publicações. A arquitetura foi dividida em três camadas desacopladas, onde a origem do dado deve ficar a critério da entidade publicadora e o local da publicação final aos usuário final deve ser uma instância da plataforma CKAN. A camada intermediária entre extração dos dados e publicação é realizada pela solução desenvolvida pela pesquisa UnBGOLD que, através da definição de parâmetros específicos, realiza a indexação semântica do dados utilizando um vocabulário controlado, preferencialmente ontologias, e também publica automaticamente os dados no CKAN. Além disso, foi criado de um catálogo de conjuntos de dados também descritos de modo conectado e uma interface para realização de pesquisa pelos conjuntos de dados abertos em que a resposta é enriquecida semanticamente.

Palavras-chave: Dados Abertos, Dados Conectados, Web Semântica, Metadados, Ontologia, UnBGOLD

Abstract

The Brazilian government has been investing in increasing the transparency of its actions aiming to encourage society to active participate in the country's management. In this sense, one of the main actions is the opening of data from federal agencies to the community. Today there is a considerable amount of open data in the Executive, Legislative and Judicial branches, beyond the spheres of the Union, States or Municipalities, but there is no clear way to link this data with its publication. This research aims to propose an architecture that assists institutions to open their data more efficiently and to add as much quality as possible to them. The quality of the data is related to two factors: the data has the possibility of being linked to other data and be as current as possible. Thus we propose a model that seeks to aggregate several technologies that allow the data to be described semantically, thus making data linked and making connections possible, in addition to providing datasets that can be updated without human intervention, guaranteeing reduced intervals between publications. The architecture was divided into three decoupled layers, where the data origin should be at the discretion of the publisher and the final publication site to the final users must be an instance of the CKAN platform. The intermediate layer between data extraction and publication is performed by the solution developed by the UnBGOLD that, through the definition of specific parameters, performs the semantic indexing of the data using a controlled vocabulary, preferably ontologies, and also published automatically the data in the CKAN. Besides that, it was created from a catalog of datasets also described in connected mode and an interface for performing search by the open data sets in which the response is enriched semantically.

Keywords: Open Data, Linked Data, Semantic Web, Metadata, Ontology, UnBGOLD

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AIISO *Academic Institution Internal Structure Ontology.*

CKAN *Comprehensive Knowledge Archive Network.*

CPD Centro de Informática.

CSV *Comma-separated values.*

DAG Dados Abertos Governamentais.

DBgoldbr Data Base Government Open Linked Data.

DC *Dublin Core Metadata Element.*

DCMI *Dublin Core Metadata Iniciative.*

DW Data Warehouse.

ESB *Enterprise Service Bus.*

ETL *Extract Transform Load.*

FOAF *Friend of a Friend.*

GCIEO *Global City Indicators Education Ontology.*

GUI *Graphical User Interface.*

HTML *Hyper Text Markup Language.*

INDA Infraestrutura Nacional de Dados Abertos.

IRI *Internationalized Resource Identifier.*

JSON *JavaScript Object Notation.*

LAI Lei de Acesso à Informação.

LOD *Linked Open Data Project.*

LUBM *Lehigh University Benchmark.*

OGP *Open Government Partnership.*

OKF *Open Knowledge Foundation.*

OWL *Ontology Web Language.*

PBDA Portal Brasileiro dos Dados Abertos.

PDA Plano de Dados Abertos.

RDF *Resource Description Framework.*

SPARQL *Protocol and RDF Query Language.*

SWEO *Semantic Web Education and Outreach.*

UnB Universidade de Brasília.

UnB-LOD UnB *Load Open Data.*

UnBGOLD UnB - *Governamental Linked Open Data.*

URL *Uniform Resource Locator*.

UVOC UnB *Vocabulary.*

VCGE Vocabulário Controlado do Governo Eletrônico.

W3C *Word Wide Web Consortium.*

XML *eXtensible Markup Language.*

Capítulo 1

Introdução

É apresentado neste contexto no qual a pesquisa esta inserida e os objetivos delimitados para alcançar a solução. A Sessão 1.1 aborda o contexto do tema da pesquisa. Na Sessão 1.2 apresenta o os motivos que justificam o trabalho. A Sessão 1.3 apresenta ao objetivo geral e a Sessão 1.4 detalha os objetivos. Já a Sessão 1.5 detalha a estrutura utilizada no documento.

1.1 Definição do Problema

O governo brasileiro tem investido, nos últimos anos, em meios para atrair a sociedade para a participação ativa da gestão. Por meio da transparência, a população pode acompanhar e fiscalizar as ações governamentais e, assim, auxiliar na promoção da eficiência da gestão pública. Esta política foi consolidada através da Lei de Acesso à Informação (LAI), que regulamentou o Art. 5º, inciso XXXII da Constituição Federal que garante o acesso a qualquer informação que possua interesse público, desde que esta informação não seja imprescindível à segurança da sociedade e do Estado [11].

Segundo Santarem Segundo [32], a LAI foi criada a partir de um movimento chamado Dados Abertos (*Open Data*) em que países como Inglaterra e Estados Unidos avançam desde 2009 em um modelo de gestão que propõe aumentar a visibilidade das informações governamentais com o principal objetivo de induzir à sociedade a participação ativa na gestão, contribuindo para que ela tenha eficiência e transparência.

Deste modo, a LAI [12] estabelece em seu Art 8º que os órgãos e entidades públicas devem promover o acesso fácil às informações de interesse coletivo, independente de solicitação. Dentre estes dados estão telefones para contatos, estruturas organizacionais, repasses e transferências de recursos, processos licitatórios, entre outros. O Brasil, em 2011, aderiu à Parceria para Governo Aberto - *Open Government Partnership (OGP)*[1],

[1] `http://www.opengovpartnership.org/`

uma iniciativa internacional criada para assegurar que os governos promovam a transparência, a participação civil, o combate à corrupção e o uso de novas tecnologias para tornar a administração mais eficaz e aberta. Desde 2012 o governo brasileiro promove ações visando incentivar a abertura de dados para a sociedade.

O Decreto Nº 8.777, de 2 de maio de 2016 [13], instituiu a Política de Dados Abertos no Poder Executivo Federal e estabeleceu que qualquer cidadão pode solicitar a abertura de bases de dados nos termos da LAI, desde que o dado não esteja regido pelas regras de sigilo da informação e, além disso, define quais as autoridades serão responsáveis por publicar e atualizar o Plano de Dados Abertos (PDA) de cada órgão. O PDA é o documento que orienta a implementação e a promoção de abertura de dados em cada órgão ou entidade da administração pública federal, estabelecendo padrões mínimos de qualidade que visem facilitar a manipulação e o reuso dos dados. Nos últimos anos, o Brasil tem evoluído na questão da abertura dos dados alcançando a sétima posição no *ranking* mundial de dados abertos [2] organizado pela *Open Knowledge International*[3]. Na Figura 1.1 é vista a evolução do Brasil no ranking desde 2013 até 2016[8].

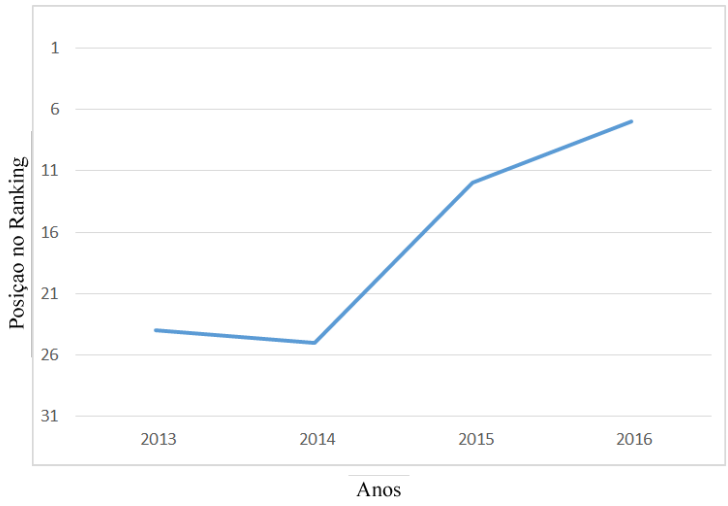


Figura 1.1: Evolução do Brasil no *Ranking* Mundial de Dados Abertos.
Fonte: Elaboração Própria.

Para auxiliar na publicidade dos dados, o Governo Federal criou o Portal Brasileiro dos Dados Abertos (PBDA), ferramenta que disponibiliza o livre acesso aos dados governamentais em formato bruto para que qualquer cidadão possa utilizá-los da maneira que lhe convier[30].

Neste sentido, o Governo Federal disponibiliza um massivo volume de dados públicos estruturados, semiestruturados e não estruturados de interesse coletivo. Assim, torna-se

[2] `https://index.okfn.org/place/`
[3] `https://okfn.org/`

um grande desafio a criação de aplicações capazes de gerar *insights* em uma velocidade apropriada a partir do enorme volume de dados nos mais variados formatos.

Este trabalho tem por objetivo propor uma arquitetura para dar suporte a publicação automatizada de dados abertos com incremento de qualidade desde a extração até o consumo pelos usuários finais.

1.2 Justificativa

Existe uma quantidade considerável de dados disponibilizados no PBDA de diversos órgãos, entretanto, os dados publicados estão relacionados através de marcadores (*tags*) o que é bastante superficial. Assim, há grande dificuldade de realizar integração, comparações ou combinações de conjunto de dados de origens diferentes. Um exemplo seria utilizar dados abertos sobre a ocupação profissional do Ministério do Trabalho, com dados sobre programas sociais do Ministério de Desenvolvimento Social, que poderiam gerar um novo conhecimento sobre como as ações sociais do Estado podem ter influência no mercado de trabalho. Diante das diversas atividades dos órgãos governamentais, nem sempre a atualização dos conjuntos de dados abertos recebe a devida dedicação, justamente porque muitos órgãos não possuem a automatização desta atividade, fazendo com que os dados fiquem defasados com frequência.

Faz-se necessária, então, a criação de uma arquitetura que auxilie na publicação dos dados para promover a melhor integração com as diversas bases, e dessa forma, enriquecer a publicação e facilitar a análise dos dados, além de automatizar a atividade de disponibilização para que os dados possam ser os mais atuais possível. A arquitetura proposta neste trabalho foi desenvolvida por meio do uso de tecnologias existentes para conexão de dados e de ferramentas desenvolvidas especificamente para este fim. Foram utilizados metadados e ontologias para representação semântica dos dados. Esta arquitetura foi validada através de um estudo de caso utilizando dados abertos provindos dos sistemas de informação da Universidade de Brasília (UnB) que tratam de dados acadêmicos.

1.3 Objetivo Geral

Este trabalho tem por objetivo propor uma arquitetura para dar suporte a publicação automatizada de dados abertos com o incremento de qualidade desde a extração até o consumo pelos usuários finais. Esta arquitetura prevê a criação de um catálogo de dados abertos conectados utilizando metadados para descrição das características e o uso de ontologias para a indexação semântica dos conjuntos dados e catálogo.

1.4 Objetivos Específicos

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

- Propor um processo de publicação de dados abertos desde a obtenção nos sistemas operativos até o seu consumo pelo usuário final;
- Estabelecer um conjunto mínimo de metadados descritivos para dados abertos;
- Desenvolver uma ferramenta de indexação de dados abertos utilizando triplas RDF e ontologias;
- Validar a arquitetura utilizando os dados abertos acadêmicos da UnB;
- Publicar um catálogo de dados abertos conectados indexados semanticamente e desenvolver uma interface de busca para os conjuntos de dados abertos com resultados enriquecidos semanticamente.

1.5 Estrutura do Trabalho

Este trabalho será estruturados nos seguintes capítulos:

- Capítulo 2 mostra o referencial teórico dos principais conceitos e tecnologias utilizados para desenvolvimento desta pesquisa, Abordando como as informações estão sendo disponibilizadas e organizadas na internet e como os dados abertos conectados se encontram neste contexto.
- Capítulo 3 aborda trabalhos que possuem relacionamento com esta pesquisa, quais os seus benefícios, limitações e objetivos.
- Capítulo 4 apresenta a arquitetura proposta nesta pesquisa e aborda seus diversos componentes e como eles interagem possibilitando a automação da publicação e a indexação semântica dos dados.
- Capítulo 5 detalha um estado de caso de abertura de alguns conjuntos dados da UnB a partir da abordagem proposta pela arquitetura.
- Capítulo 6 apresenta a conclusão do trabalho e possíveis trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo é apresentada a fundamentação teórica dos principais conceitos e tecnologias consideradas importantes para desenvolvimento desta pesquisa. Na Seção 2.1 é apresentada a origem da Web Semântica e sua evolução. A Seção 2.2 aborda o conceito de dados conectados, as boas práticas para abertura de dados e a classificação de qualidade de dados conectados. Na Seção 2.3 fala sobre ontologia e sua definição. Na Seção 2.4 explica o que vem a ser metadados. Na Seção 2.5 é abordada a tecnologia RDF e suas principais características. Na Seção 2.6 é apresentado a definição de dados abertos, suas características e a sua extensão para dados abertos governamentais. Por fim na Seção 2.7 abordamos o barramento de serviço ErlangMS.

2.1 Web Semântica

A Web foi criada para compartilhamento de documentos que pudessem ser interligados uns aos outros sendo que seu conteúdo foi estruturado para ser facilmente interpretado por seres humanos, sem a preocupação com análises computacionais mais sofisticadas. Em 2001, Berners-Lee publicou o artigo que apresentou a Web Semântica[5], na qual seria possível conectar os dados automaticamente por computador gerando um contexto de informações a partir da conexão de dados.

De acordo com Berners-Lee et al. [5], a Web Semântica é uma extensão da Web em que as informações têm seu significado bem definido, possibilitando que máquinas e pessoas possam colaborar no dia a dia. Em vista disso, os computadores não seriam utilizados apenas para processar dados existentes, mas também para interpretá-los dentro da realidade em que são criados. Neste contexto, o desafio da Web Semântica é criar uma linguagem que seja capaz de representar e raciocinar através do conteúdo existente na Web, adicionando lógica aos dados publicados.

Após este esboço inicial, o estudo da Web Semântica evoluiu rapidamente, já que a perspectiva das camadas da Web Semântica foi atualizada diversas vezes, acrescentando ou retirando tecnologias de acordo com a abordagem. A Figura 2.1 apresenta a comparação entre a arquitetura proposta por Berners-Lee et al. [5] em 2001 e versão atualizada em 2007[19]. Na primeira versão da arquitetura proposta para a Web Semântica teve o objetivo de propor principalmente os conceitos e abstrações, já a a segunda versão se concentrava em especificar as tecnologias para implementação.

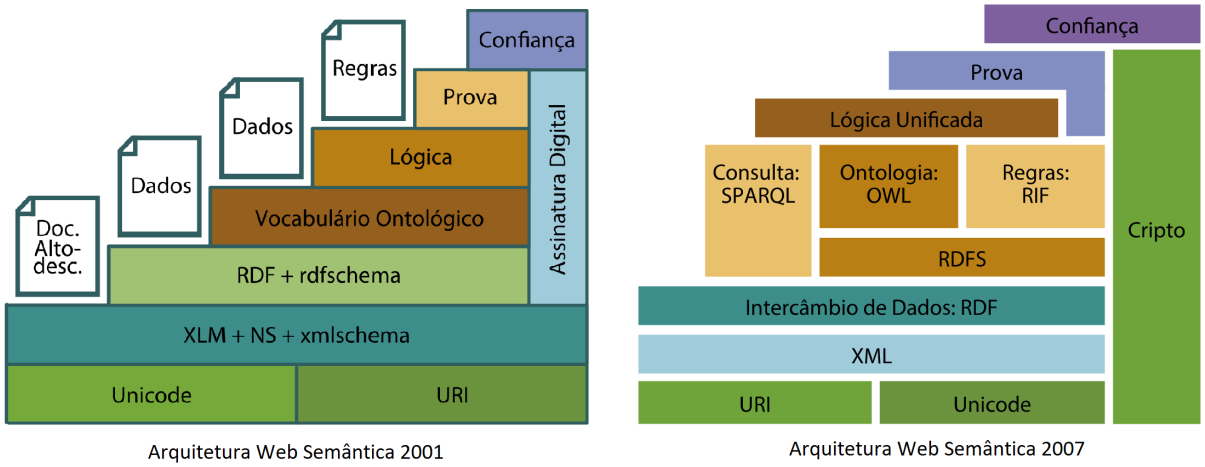


Figura 2.1: Arquiteturas Propostas Para a Web Semântica.
Fonte: Adaptado de [25].

A evolução da Web Semântica é um processo ainda contínuo, Isotani e Bittencourt [25] apresentam o que seria uma visão mais atualizada das tecnologias que envolvem a Web Semântica. Como pode ser visualizado na Figura 2.2, é uma visão que possui duas dimensões: Conceitos e Abstrações e Especificações e Soluções. Dentre as diversas especificações, essa nova arquitetura da Web Semântica define os formatos XML, TURTLE, RDFa e uFORMATS para os dados, que o intercâmbio de informações é realizado através de RDF e que as consultas aos dados é realizada utilizando linguagem SPARQL.

Também é possível identificar através das cores que especificam cada tecnologia, como o conceito de Dados Conectados está no contexto da Web Semântica utilizando alguns de seus conceitos e especificações.

Visto a complexidade de implementar a Web Semântica em todos os conteúdos da Web, Berners-Lee [4] propôs que fosse utilizado um meio mais simples de conectar dados diferentes através do mapeamento semântico de conjuntos específicos de dados, de modo a gerar um contexto conforme previsto na Web. Esse conceito é conhecido como Dados Conectados (*Linked Data*), que é detalhado na próxima seção.

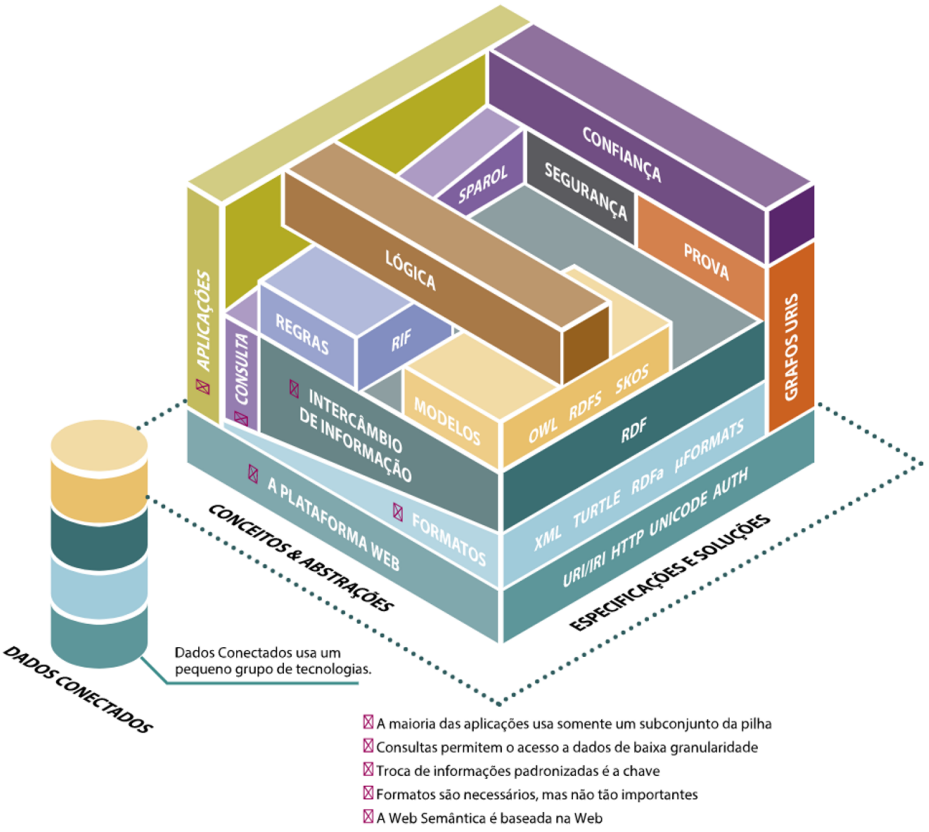


Figura 2.2: Estado Atual da Web Semântica.
Fonte: [25].

2.2 Dados Conectados

O conceito de Dados Conectados é definido por Bizer et al. [8] como o uso da Web para conectar dados digitalmente de fontes diversas, independente do local ou da origem, no qual seu significado pode ser lido e interpretado por máquinas, gerando assim a chamada Web de Dados. Enquanto os dados comuns da Web usam hipertexto semiestruturado em *Hyper Text Markup Language (HTML)* que são ligados através de *hiperlinks*, os Dados Conectados têm por base o uso do formato *Resource Description Framework* (RDF) que possibilita a ligação dos dados a coisas arbitrárias no mundo.

Para auxiliar na publicação de Dados Conectados o *Word Wide Web Consortium* (W3C) criou um grupo de trabalho sobre dados abertos governamentais que definiu um conjunto de 10 itens, chamados de boas práticas, para serem adotados aos interessados em abrir seus dados [39], esses itens são:

1. **Preparar os *Stakeholders***: formar os usuários para criar e manter os dados conectados;

2. **Selecionar o conjunto de dados**: definir quais dados serão publicados e conectá-los para reuso;
3. **Modelar Dados**: Os *Stakeholders* definem como irão representar os dados e como eles se relacionam com os demais dados.
4. **Especificar a licença**;
5. **Nomear bons URIs**;
6. **Usar vocabulários-padrão**: definir se irá construir um vocabulário para publicação do dado ou, preferencialmente, utilizar um vocabulário já existente, facilitando conexão com outros dados;
7. **Converter os dados**: Converter os dados oriundos de uma fonte original para Dados Conectados;
8. **Prover acesso aos dados**: definir como humanos e máquina terão acesso aos dados;
9. **Anunciar novo conjunto de Dados Conectados**;
10. **Reconhecer a função social**: para que o publicador dos dados se comprometa em fomentar a publicação ao longo do tempo.

Buscando uma forma de melhorar a qualidade da publicação dos dados, Tim Berners-Lee propôs um princípio para categorizar o nível de abertura de um dado [4]. Este índice conhecido como "5 Estrelas dos Dados Conectados" (*5 Stars Linked Data*), no qual quanto mais estrelas um dado tiver, mais conectados e, consequentemente, mais qualidade um dado aberto possui. Além disso, estabelece que um dado só consegue atingir o máximo da qualidade quando ele consegue se conectar a outros dados a partir das tecnologias propostas pelo W3C. A seguir, as características sobre cada classificação são apresentadas:

1. Se um dado está disponível na Internet com licença aberta independente do formato, ele é um dado com uma estrela;
2. Se o dado disponível está em algum formato estruturado (como XLS), é um dado com duas estrelas.
3. Se o formato do dado disponível está estruturado em um formato não proprietário como *Comma-separated values* (CSV) ou *eXtensible Markup Language* (XML), ele possui três estrelas.
4. Se utilizar *Uniform Resource Locator* (URL) para identificação e está em conformidade com os padrões estabelecidos pelo W3C, RDF e SPARQL, de modo que se possa direcionar publicações, é considerado um dado quatro estrelas.

5. Se, após cumprir as regras anteriores, o dado for conectado a outros dados de modo a criar uma lógica de ligação, este dado é um dado de 5 estrela, ou seja, de alta qualidade.

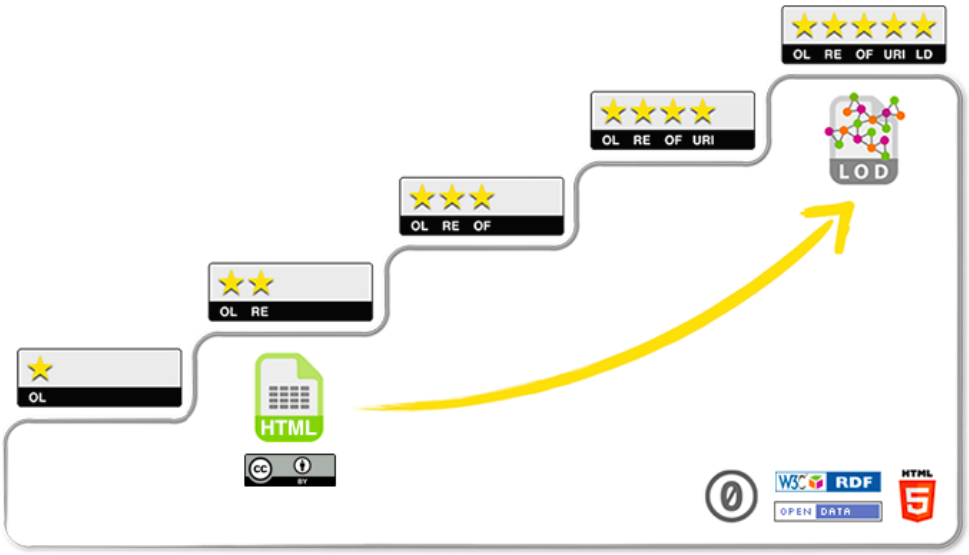


Figura 2.3: Sistema 5 Estrelas.
Fonte: online `http://5stardata.info/pt-BR/`.

A Figura 2.3 mostra visualmente as cinco camadas do Sistema 5 Estrelas, na qual também são apresentadas as tecnologias utilizadas para tornar o dado mais conectado.

Esta pesquisa visa a possibilidade de publicação de dados com mais qualidade e um dos conceitos que iremos adotar como parâmetro de qualidade é o sistema 5 estrelas para dados conectados.

2.3 Ontologia

Ontologia é um termo oriundo da filosofia da ciência que descreve tipos de entidades e a forma na qual estão relacionadas, sendo o estado a natureza do "ser" e a "existência"[1]. Quando a ontologia é aplicada a Ciência da Computação tem o significado de um conjunto de conceitos fundamentais e suas relações. Permite representar o entendimento de maneira formal, entendível por humanos e computadores [28].

Dentre os estudo das ontologias nos últimos anos muitas definições surgiram, mas no contexto da Ciência da Computação e Ciência da Informação, a mais utilizada é a proposta por Gruber que se desdobra nas seguintes definições:

[1] Aristóteles 384-322 a.C.

- Definição 1 — Gruber [21] propôs que a ontologia é uma especificação de uma conceituação. A conceitualização é o significado dos conceitos e suas relações de acordo com o contexto e "especificação" e a representação formal, declarativa e explícita dos conceitos e relações;
- Definição 2 — Borst [10] complementou afirmando que a ontologia é uma especificação de uma conceituação compartilhada; e,
- Definição 3 — Studer et al. [36] combinaram as definições supramencionadas ao estabelecer que a ontologia é uma especificação explícita e formal de uma conceituação compartilhada.

As ontologias nesta pesquisa consistem em utilizar as especificações já existentes que visem descrever semanticamente dados que desejamos publicar através da conceituação formal.

2.4 Metadados

A definição mais comum sobre Metadado é um dado ou a informação sobre o dado, assim, nos Metadados são guardadas informações que servem de sumário dos dados, além da maneira de recuperá-lo ou acessá-lo. O termo surgiu em 1995, durante um simpósio realizado em Dublin, Ohio, do qual originou a *Dublin Core Metadata Iniciative* (DCMI)[18].

Taylor [37] propõe a categorização de metadados em três tipos: administrativos, estruturais e descritivos. Metadados administrativos têm por objetivo o gerenciamento, o suporte à tomada de decisão e a manutenção do registro das informações. Um exemplo seria as informações sobre requisitos de armazenamento e o processo de migração de informações digitais. Já metadados estruturais referem-se à estrutura do suporte físico da informação que está sendo descrita, como um arquivo digital, um livro, uma fotografia ou outro suporte. E por fim, metadados descritivos são aqueles com a função de descrever as características intelectuais do conteúdo de um documento.

Em sistemas de apoio à decisão, como ambientes de *Data Warehouse (DW)*, de acordo com Kimball [26], os metadados são divididos em três tipos: técnicos, de negócio e de processos. Os metadados técnicos são utilizados para descrever as estruturas de dados (como por exemplo, tabelas, campos e tipos de dados), enquanto os metadados de negócio descrevem o conteúdo do DW de modo que sejam compreendidos pelo usuário final e, por último, os metadados de processos buscam descrever os resultados das operações que são executados em um DW, tais como extração, transformação e carga.

Na atual bibliografia não existe um consenso de um padrão único de metadados que consiga abordar todas as áreas do conhecimento humano, contudo, existem padrões di-

ferentes de metadados para finalidades específicas. De acordo com Souza et al. [35], os padrões de metadados têm como objetivo subsidiar as definições e formar uma rede automatizada para registros de propriedades de dados cadastrais de forma padronizada e consistente.

2.5 RDF

RDF é um modelo com objetivo de descrever recursos na Web de modo que estes recursos possuem uma identificação única por meio de uma *Internationalized Resource Identifier* (IRI)[6]. Os recursos podem ser descritos por uma série de propriedades, com tipo (*type*) e valor (*value*) chamadas descrição RDF (descrição sintática). Além da descrição, o RDF permite representar a relação e o significado com outros recursos (descrição semântica) através de chamas triplas RDF. Esta representação é realizada através de uma tripla de informação que é composta inicialmente pelo "Sujeito" que é uma entidade indexada que tem relação com uma outra entidade que representa uma informação, que é o "Objeto", e a relação entre eles chamamos de "Predicado". Na Figura 2.4 é apresentada a representação gráfica de um grafo da tripla RDF.

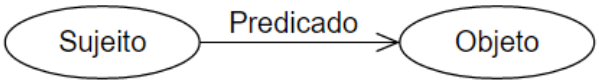


Figura 2.4: Tripla RDF
Fonte: Adaptado de [6]

É responsabilidade do IRI identificar de maneira única um recurso na Web, assim numa tripla RDF o sujeito, predicado ou objeto podem ser identificados pelo RDF do recurso referenciado pelo IRI. Quando os valores no grafo não possuírem um IRI ele é chamado de "Literal", sendo que este pode assumir diversos tipos de dados, como números, cadeias de caracteres (*strings*) ou datas.

O RDF-S (ou *RDF Schema*) é uma extensão do RDF que permite a descrição semântica e a descrição de grupos de recursos e seus relacionamentos através de um vocabulário, ampliando a expressividade do RDF [25]. Existem dois conceitos básicos do RDF-S que são importantes descrever: o primeiro é a Classe que pode descrever o recurso em RDF (*rdfs:resource*), mas também pode descrever um tipo primitivo de dado (*rdfs:literal*) ou derivativo (*rdfs:dateType*); o segundo conceito é a Propriedade que tem a função de estabelecer a relação entre sujeito e objeto. Alguns tipos de propriedade são o *rdfs:type*, para definir que uma classe é uma instância da outra, o *rdfs:domain*, para indicar o sujeito do relacionamento, e o *rdfs:range*, para indicar o objeto do relacionamento [14].

Nesta pesquisa buscaremos transformar os dados abertos estruturados em triplas RDF, nas quais utilizaremos metadados e ontologias para dar significado semântico aos dados através do predicado.

2.6 Dados Abertos

Segundo a International [24] dados abertos são *"...blocos de construção do conhecimento aberto. O conhecimento aberto é o que os dados abertos se tornam quando é útil, utilizável e usado"*. Sendo que, de acordo com OKF [29], o termo "aberto" neste contexto significa que o dado é livre para que qualquer pessoa possa reutilizá-lo e distribuí-lo, independente se o objetivo seja comercial, informativo ou acadêmico, estando sujeito, no máximo, a necessidade de informar a autoria dos dados e compartilhar utilizando a mesma licença.

Além disso, [24] define as três principais características da abertura de dados:

- **Disponibilidade e acesso**: estabelece que os dados devem estar disponíveis por completo com mínimo custo para reproduzi-lo, sendo preferencialmente na Internet para *download* em formato conveniente e modificável.
- **Reutilização e redistribuição**: os termos legais que regem os dados devem permitir a sua reutilização, redistribuição e o intercâmbio com outros dados. Além disso, é importante que os dados estejam em formato que possam ser lidos por computadores.
- **Participação universal**: o uso de dados deve estar universalizado, pois todos devem ter a capacidade de usar, reutilizar e distribuir, não podendo haver nenhum tipo de restrição sobre o destino que possa ocorrer, seja ele utilizado para pesquisa ou exploração econômica.

Segundo Malamud et al. [27], a Internet é o espaço público do mundo moderno que possibilita que os governos tenham a oportunidade de compreender melhor as necessidades dos cidadãos e que os cidadãos possam estar mais envolvidos na gestão governamental. Assim, existe uma agregação de valor às informações publicadas que podem promover a melhoria do bem-estar público e a maior eficiência na aplicação de recursos públicos.

2.6.1 Dados Abertos Governamentais

Nos dias 7 e 8 de dezembro de 2007, um grupo de 30 pessoas reuniu-se para desenvolver um conjunto de princípios que norteasse a publicação de Dados Abertos Governamentais (DAG), assim Malamud et al. [27], foram estabelecidos 8 princípios:

1. **Completos**: Os dados publicados não estão sujeitos a limitações de privacidade, segurança ou privilégios válidas.
2. **Primário**: Os dados são extraídos diretamente da fonte e devem ter o maior nível de granularidade possível sem tratamentos, como agregação ou modificação.
3. **Atuais**: Os dados devem ser atualizados o mais rápido possível para garantir maior valor.
4. **Acessível**: Os dados devem ser disponibilizados para o maior número de usuários possíveis para atender o maior número de finalidades.
5. **Processável por Máquina**: Os dados devem estar o minimamente estruturado para possibilitar o processamento automatizado por máquina.
6. **Não Discriminatório**: Não deve conter nenhum tipo de exigência de registro para obter acesso aos dados.
7. **Não Proprietário**: Os dados devem ser disponibilizados em um formato não proprietário que possibilite a livre leitura.
8. **Livre de licença**: Os dados não devem estar sujeitos a nenhum tipo de regulamentação de direitos autorais, patentes, marcas registradas ou segredos comerciais.

O ativista e pesquisador David Eaves [20] propôs 3 leis para caracterizar DAG que foram adotadas pelo W3C:

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

A publicação de dados governamentais tem por objetivo aumentar a transparência das informações da administração pública por meio de políticas que visam garantir que todo cidadão tenha acesso aos dados gratuitamente em um ambiente livre (a Internet). Além disso, é dever do gestor dar sustentabilidade na continuidade da publicação dos dados garantindo que sejam atualizados o mais breve possível.

2.7 ErlangMS

A partir de estudos realizados pela equipe de analista do CPD/UnB visando elaborar uma estratégia para modernização dos sistemas legados da universidade, optou-se por

experimentar uma arquitetura orientada ao serviço, seguindo o estilo arquitetural REST. Para tanto foi implantada uma arquitetura SOA em que sistemas novos e legados co-existam e acessem os mesmo serviços. Foi utilizado a linguagem funcional Erlang no desenvolvimento de um *Enterprise Service Bus* (ESB) batizado de ErlangMS[17].

Na Figura 2.5 podemos verificar que a arquitetura propicia a ligação dos clientes aos serviços na qual é implementada a regra de negócio da organização. A adoção do barramento permitiu diversas melhorias, como gerenciamento das requisições dos clientes por meio da estruturação de um catálogo de serviços. Uma vez que a linguagem Erlang já disponibiliza bibliotecas que tratam de requisições HTTP/REST, a adoção da solução se tornou mais simples.

O ErlangMS implementa o estilo arquitetural RESTful e utiliza, primariamente, o formato *JavaScript Object Notation* (JSON) para a troca de mensagens, sendo que a comunicação do cliente com o barramento ocorre por meio de uma API REST padronizada desenvolvida pelo CPD/UnB. Essa API possibilita o uso de diversos tipos de operadores para facilitar a extração dos dados, tais como *filter*, *sort*, *limit* e *offset*.

Um componente chave desta arquitetura é o conceito de catálogo de serviços que, em linhas gerais, dá visibilidade aos serviços disponibilizados para extração dos dados. O catálogo de serviços contém as definições da API para acesso aos dados e aos metadados para o barramento buscar os dados solicitados na base de dados.

Desse modo, para realizar o acesso a fonte de dados é necessário primeiramente definir a API REST do serviço no catálogos de serviços do barramento. Cada serviço é definido por um contrato em que se define qual é o escopo da consulta, a *url* do serviço, o tipo de autenticação, os parâmetros de entrada que podem ser fixos (caracteres, números) ou temporais (dias, semestres, ano), entre outros atributos. Uma grande vantagem do uso do barramento é a sua flexibilidade para buscar dados de diversas fontes, sejam eles de um banco de dados *SQL-Server* ou *PostgreSQL* ou de arquivos CSV através do atributo *datasource*. No exemplo demonstrado no Código 2.1, foi utilizado um arquivo csv como fonte de dados para os cursos ofertados na UnB.

Código 2.1: Especificação de uma API para extração dos cursos ofertados na UnB.

```
{
  "name": "/hackathon/cursos",
  "comment": "Lista os cursos ofertados na UnB",
  "owner": "samples",
  "version": "1",
  "service" : "ems_api_query_service:find",
  "url": "/hackathon/cursos",
  "type": "GET",
  "datasource" : {
```

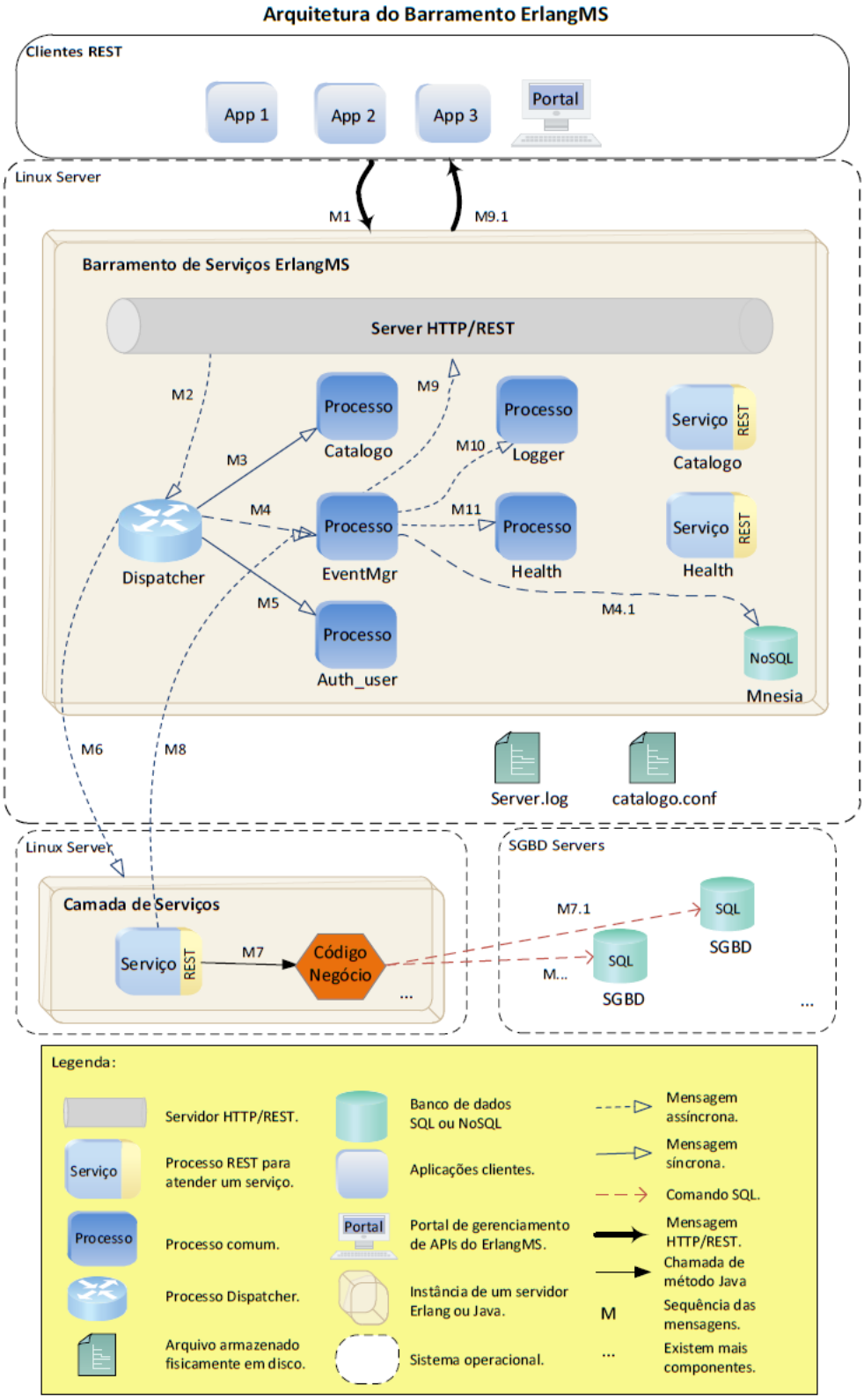


Figura 2.5: Arquitetura ErlangMS
Fonte: [17]

```
    "type" : "csvfile",
    "driver" : "sqlite3",
    "connection" : "cursos.csv",
    "table_name" : "Tbl_Cursos"
  },
  "lang" : "erlang",
  "authorization" : "oauth2",
  "querystring": [
    {
     "name": "filter",
     "type": "string",
     "default" : "",
     "comment": "Filtro principal da pesquisa"
    },
    .
    .
    .
  ]
},
```

Com a definição do serviço especificado no catálogo de serviços conforme o Código 2.1, a extração dos dados pode ser realizada por meio de uma chamada REST ao barramento. Por exemplo, para invocar o serviço /hackathon/cursos filtrando somente os cursos do segundo semestre de 2016, seria preciso fazer uma requisição HTTP/REST no seguinte formato:

```
/hackathon/cursos?filter={ "semestre" : "20162" }
```

O ErlangMS oferece uma interface para acesso aos dados de modo seguro e tem a flexibilidade necessária para implementação de aplicações desacopladas, por esta razão tornou-se o padrão para o desenvolvimento de softwares pelo CPD/UnB.

Capítulo 3

Trabalhos Relacionados

Neste capítulo são abordados alguns trabalhos existentes que possuem alguma relação com nossa pesquisa, eles possuem características em comum com o objetivo de dar significado semântico a dados publicados na Web. Na Seção 3.1 é apresentado o software UnB-LOD para transformação de dados estruturados em dados conectados. Na Seção 3.2 é descrito o Projeto Dados Abertos Conectados que versa sobre a iniciativa de dar semântica a conjuntos de dados já conectados na Web. A ferramenta DBGoldBR é apresentada na Seção 3.3 que tem por objetivo indexar as fontes de dados utilizando metadados e ontologias. Por fim é apresentada na Seção 3.4 o projeto DBPedia, que é um banco de dados que indexa as páginas da Wikipédia através de um vocabulário controlado de forma a melhorar as buscas por informações e identificar assuntos relacionados.

3.1 UnB-LOD

Em 2014 foi apresentada por Silva et al. [33] o UnB *Load Open Data* (UnB-LOD), protótipo de uma ferramenta de código aberto e gratuita com objetivo de criar a integração entre conjuntos de dados abertos, principalmente governamentais, adotando padrões de publicações definidos pelo W3C, utilizando RDF e URI. Traz uma Interface Gráfica com o Usuário, ou GUI, desenvolvida em linguagem JAVA, além de agregar diversas outras tecnologias para auxiliar a aplicação, conforme é apresentada nas Figuras 3.1 e 3.2 em que podemos verificar a arquitetura utilizada e a tela de interface com o usuário, respectivamente.

A integração dos dados no UnB-LOD é realizada a partir de um fluxo de trabalho composto pelas fases de *Preparation*, *Data Collection* e *Data Integration*.

Na fase de *Preparation* é modelado um esquema dos dados para ser associado aos conjuntos de dados. Esse esquema possui os componentes *Class*, *Property* e *Reference-Property*. A Figura 3.3 apresenta a interface GUI desta fase, na qual foi modelado um

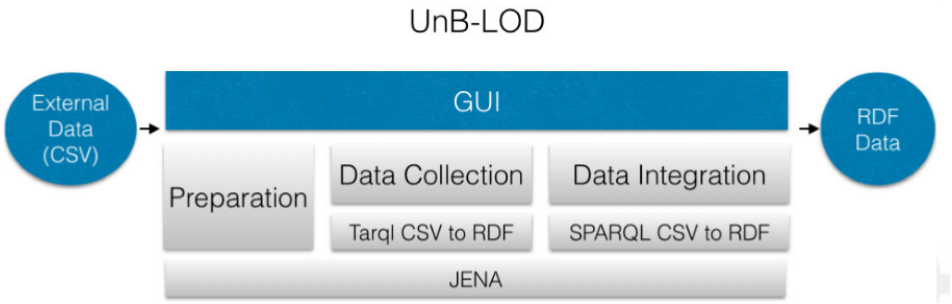


Figura 3.1: Arquitetura UnB-LOD (2014)
Fonte: [33]

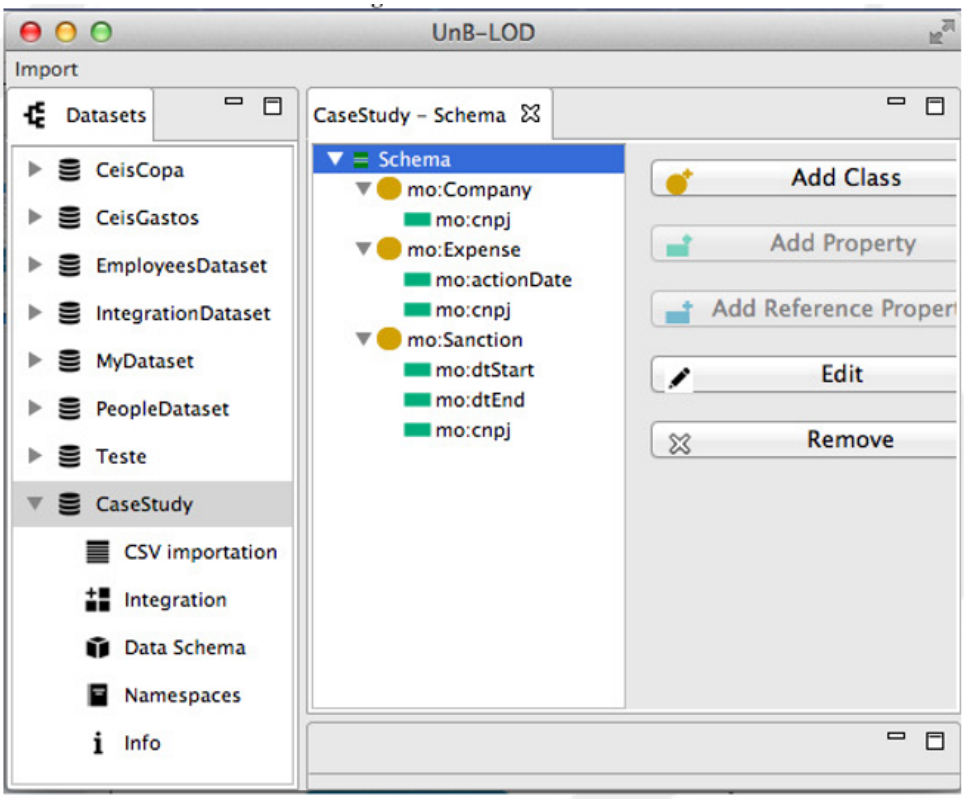


Figura 3.2: Interface GUI UnB-LOD (2014)
Fonte: [33]

esquema de dados em que foi criado um componente *Class* chamado "Client" e os componentes "Property" chamados "ex:name" e "ex:id". É possível verificar que a adição e a exclusão dos componentes é realizada através dos botões dispostos do lado direito da tela.

O UnB-LOD possui a vantagem de ter uma interface bastante simples e intuitiva para indexação semântica de dados abertos, proporcionando facilidade ao usuário ao construir esquemas de indexação que irão auxiliá-lo na integração quando os conjuntos de dados forem atualizados, contudo, possui a restrição de não interagir com outras ferramentas

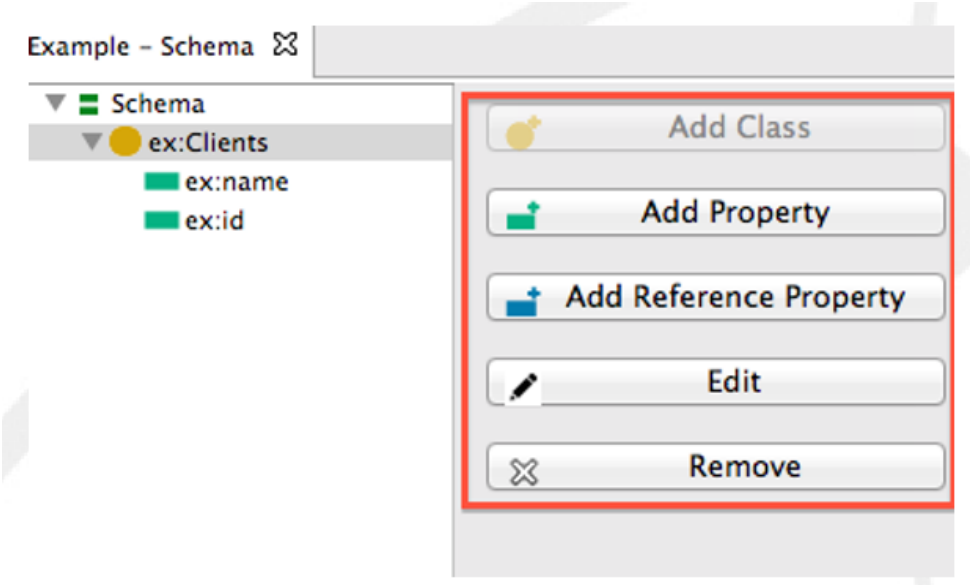


Figura 3.3: UnB-LOD Fase *Preparation* (2014)
Fonte: [33]

que possibilitem a recuperação dos dados ou a publicação. Outra questão é a plataforma utilizada, que restringe o uso a uma estação em que a ferramenta deve ser instalada, o que pode dificultar a integração de dados de usuários distintos.

Visto que foi apresentada a versão no estágio de prototipação, de acordo com a informação contidas no repositório da aplicação[1], o projeto foi descontinuado pelos autores. Propomos uma ferramenta que seja mais interativa, tanto com os usuários, com instituições que desejam publicar dados abertos conectados, quanto com outras ferramentas de publicação.

3.2 Projeto Dados Abertos Conectados - *Linked Open Data Project*(LOD)

Junto com o advento da Web dos Dados surgiu também um esforço da comunidade com o objetivo de identificar conjuntos de dados com licenças abertas e republicá-los, de modo a conectá-los uns aos outros usando RDF. Esse movimento ficou conhecido com *Linked Open Data Project* (LOD) ou Projeto Dados Abertos Conectados [7]. Iniciado em 2007, este projeto foi apoiado pelo W3C através do grupo de interesse *Semantic Web Education and Outreach* (SWEO).

Desde seu início, o mapeamento dos dados foi realizado de modo contínuo, sendo que atualmente foram mapeados cerca de 1163 *dataset* conectados no LOD [15]. Na Figura

[1] `https://github.com/marcusos/unblod`

3.4 é possível ver os conjuntos de dados mapeados pelo projeto LOD em setembro de 2008 quando já englobava 45 conjuntos de dados. O diagrama mais atual pode ser visto em `http://lod-cloud.net/`.

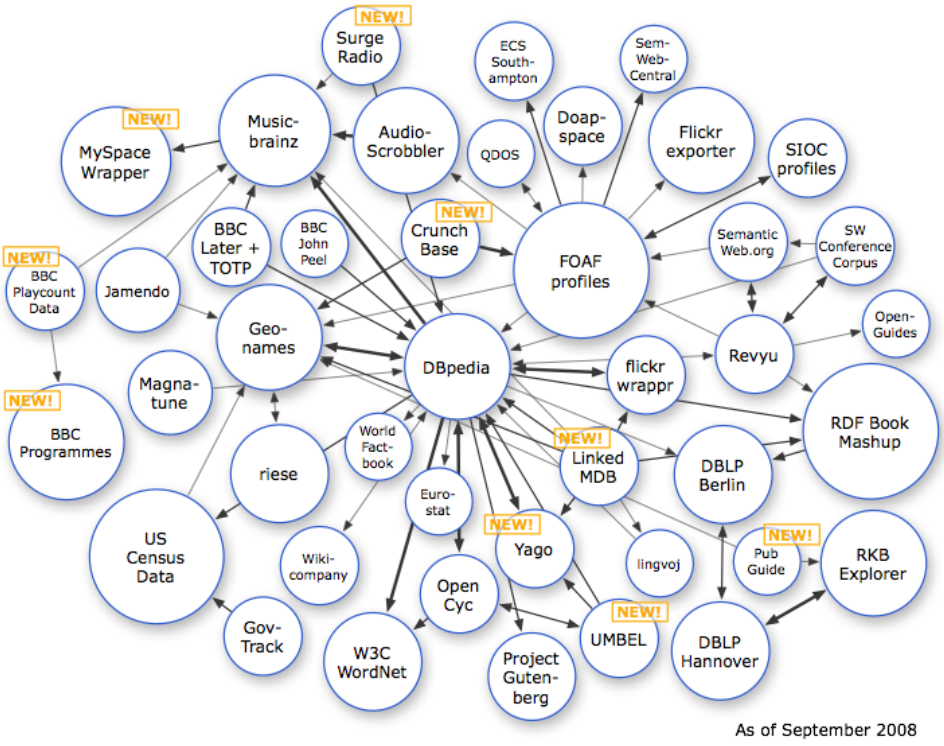


Figura 3.4: Diagrama de Rastreamento LOD (2008)
Fonte: `https://lod-cloud.net/versions/2008-09-18/lod-cloud.png`

Semelhante ao proposto pelo projeto LOD, esta pesquisa procura viabilizar um ambiente de dados conectados, contudo o foco deste trabalho é que esse ecossistema possa ser formado a partir de dados abertos governamentais, em que o incremento semântico de dados propiciará que os conjuntos de dados possam ser ligados uns aos outros organicamente.

3.3 DBgoldbr

O *Data Base Government Open Linked Data (DBgoldbr)* é uma ferramenta que visa a indexação semântica de fontes de dados abertos publicados, transformando-os em dados conectados [38]. Para conseguir este objetivo a equipe de pesquisadores propõe um modelo conceitual e um protótipo de ferramenta que realiza a indexação semântica das fontes de dados utilizando ontologias e triplas RDF. O resultado desta transformação é o enriquecimento da qualidade do dado que passa a ser classificado como dado 5 estrelas, proposto por Berners-Lee [4] como visto na Seção 2.2.

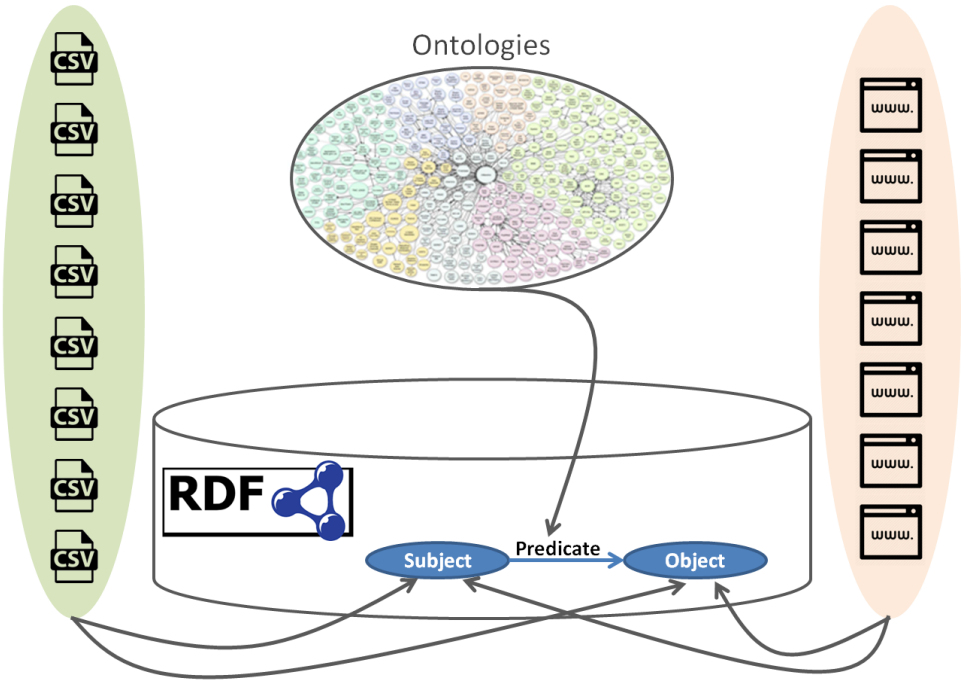


Figura 3.5: Visão Conceitual do DBgoldbr
Fonte: [38]

Conforme evidencia-se na Figura 3.5, na visão conceitual do DBgoldbr é construída a tripla RDF em que cada elemento da tripla é representado da seguinte maneira:

- **Sujeito**: CSV com dados extraídos da fonte;
- **Objeto**: Endereço do sítio institucional do órgão de origem do dado;
- **Predicado**: Vocabulário controlado, no qual deve ser utilizado preferencialmente ontologias.

A arquitetura lógica do DBgoldbr é dividida em duas grandes partes. A primeira parte consiste nos principais cadastros de dados que serão utilizados para a indexação semântica. Nesta parte, utiliza-se uma arquitetura tradicional de aplicação Java[2], utilizando o Sistema Gerenciador de Banco de Dados MYSQL[3], e é composto por um interface com função de gerenciar as informações sobre as Entidades Publicadoras, Fonte Publicadas e as Ontologias. A Figura 3.6 apresenta a primeira parte da arquitetura lógica.

Já a segunda parte da arquitetura lógica é responsável pela criação, armazenagem e recuperação das triplas RDF. A Figura 3.7 apresenta esta parte da arquitetura através de camadas. A primeira camada é uma *Java Application Client* e as demais camadas utilizam o *framework* Apache *Jena*[4] que possui APIs com o objetivo de processar arquivos

[2] `https://www.java.com/`
[3] `https://www.mysql.com/`
[4] `https://jena.apache.org/`

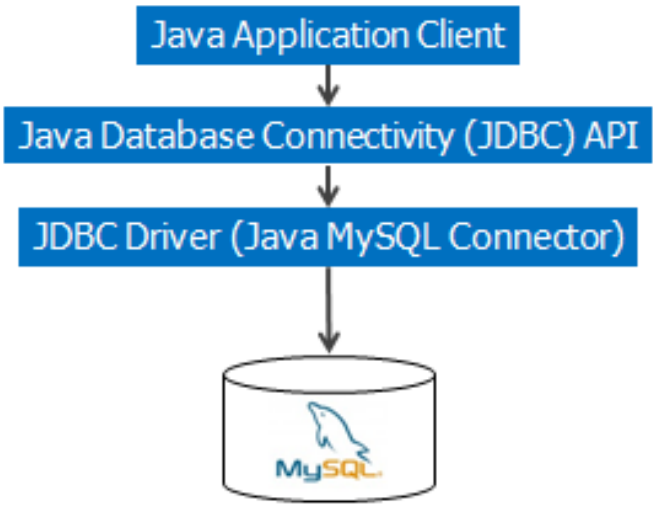


Figura 3.6: Primeira Parte da Arquitetura Lógica DBgoldbr
Fonte: [38]

RDF, que representam as triplas, arquivos *Ontology Web Language* (OWL), que são as ontologias e para o suporte as consultas SPARQL às triplas RDF. Existe também API com a função de viabilizar as inferências e o armazenamento de triplas. O DBgoldbr utiliza o componente TDB do *Jena* para persistir triplas RDF.

Atualmente a busca por dados abertos no Brasil é feita principalmente pelos motores de busca convencionais, como o Google, ou através do Portal Brasileiro de Dados Abertos[5][22]. Espera-se que com o uso do DBgoldbr para classificação semântica das fontes dos dados, possa enriquecer a qualidade dos resultados de buscas realizadas nos dados disponibilizados pelo governo brasileiro, sendo que serão realizadas a partir da consulta em um repositório de triplas RDF criado a partir da classificação realizada pelo DBgoldbr. A Figura 3.8 apresenta o estado atual da busca de dados abertos e o estado esperado com o DBgoldbr.

3.4 DBPedia

Segundo Auer et al. [3], o projeto DBPedia tem por objetivo construir uma base de conhecimento de grande escala e em multilinguagem, para isso, busca extrair dados estruturados dos conteúdos existentes da Wikipedia[6], possibilitando assim a conexão entre dados, e, como consequência, os seus artigos. A partir de técnicas da Web Semântica foi possível enriquecer consultas às informações expandindo para diversos assuntos relacionados, ge-

[5]http://dados.gov.br
[6]http:wikipedia.org

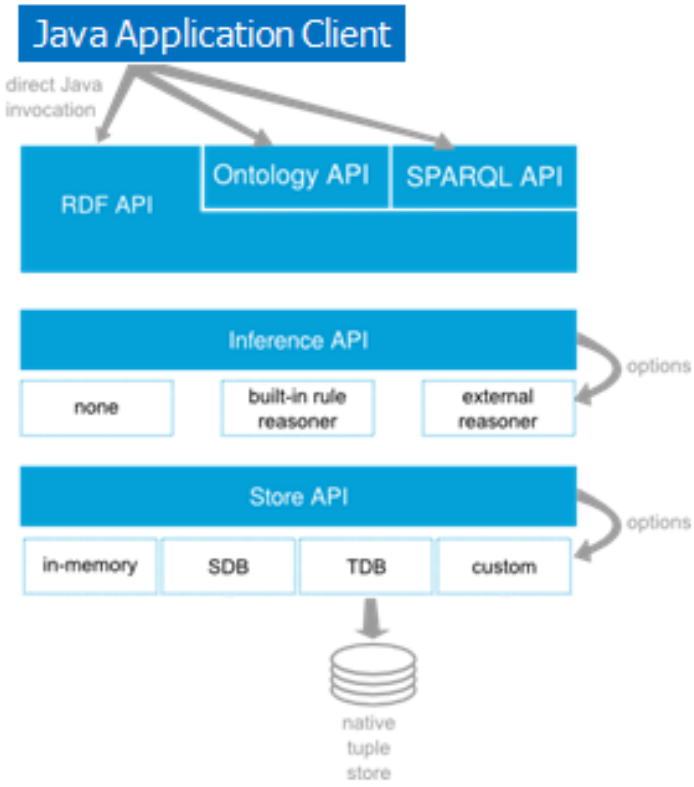


Figura 3.7: Segunda Parte da Arquitetura Lógica DBgoldbr
Fonte: [38]

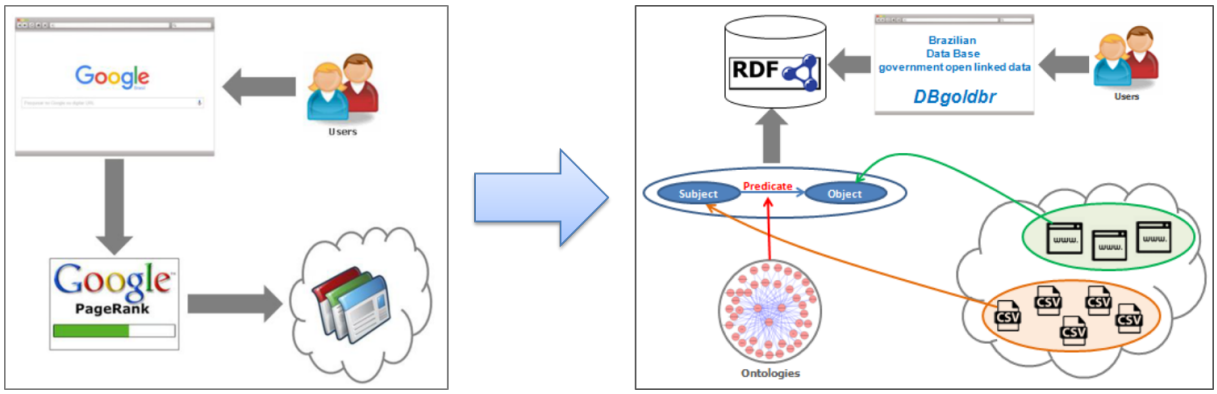


Figura 3.8: Evolução da Busca do Dados Aberto com DBgoldbr
Fonte: Adaptado de [38]

rando a integração dos dados, reconhecimento de entidade e domínios e classificação de documentos. Para a realização desta tarefa foram executadas as seguintes atividades:

- Foi desenvolvida uma estrutura de extração de informações a partir dos conteúdos publicados na Wikipédia convertendo em triplas RDF. Esses componentes compõem a base sobre a qual pesquisas adicionais sobre extração de informações, agrupa-

mento, gerenciamento de incertezas e processamento de consultas podem ser realizadas.

- É disponibilizado um grande conjunto de dados RDF dos conteúdos da Wikipédia que podem ser utilizados por uma grande variedades de aplicações.
- Também foram conectados os dados da DBPedia com outros conjuntos de dados abertos. A ligação da DBpedia com outros conjuntos de dados abertos em 2007 já possuía um total de 2 bilhões de triplas RDF.
- Interfaces e módulos de acesso foram desenvolvidos de modo que os conjuntos de dados pudessem ser acessados e conectados a outros sites.

De acordo com o site oficial da DBPedia[7], somente a sua versão inglesa da base de conhecimento que descreve 4,58 milhões de recursos, dos quais 4,22 milhões são classificados em uma ontologia consistente, incluindo 1.445.000 pessoas, 735.000 lugares (incluindo 478.000 lugares ocupados), 411.000 obras de arte (incluindo 123.000 álbuns de música, 87.000 filmes e 19.000 jogos de vídeo), 241.000 organizações (incluindo 58.000 empresas e 49.000 instituições educacionais), 251.000 espécies e 6.000 doenças.

Segundo [9], a DBPedia faz as seguintes contribuições para o desenvolvimento da Web de Dados:

- O desenvolvimento de uma estrutura que realiza a extração de conteúdos da Wikipédia e os converte em uma rica base de conhecimento de multi-domínio. É possível fazer o gerenciamento das atualizações do artigos publicados ao longo do tempo e, também, mapear os informações contidas nas caixa de informações de cada artigo, chamado *infobox*, para por meio de uma ontologia aumentar a qualidade dos dados;
- A definição de um identificador único na Web para cada entidade da DBPedia que garantiu que a partir deste identificador é possível utilizá-lo para conectar à um outro dado;
- A publicação de links RDF apontado da DBPedia para outras fontes de dados da Web, juntamente com o suporte aos editores de dados para configurarem seus links nas fontes de dados para DBPedia, fez surgir uma Web dos Dados em torno da DBPedia.

Dentre as ferramentas que a DBPedia disponibiliza, há o Virtuoso[8] que é uma plataforma onde oferece acesso transparente à suas fontes de dados por meio de consultas utilizando linguagem SPARQL[34]. Além disso oferece suas funcionalidades a serviço da

[7] `http://wiki.dbpedia.org/about`
[8] `https://dbpedia.org/sparql`

Web, podendo assim qualquer usuário criar aplicações que possam ser conectadas aos serviços da DBPedia.

A DBPedia foi um dos projetos que serviu de inspiração para esta pesquisa no que tange realizar o mapeamento de um conjunto de metadados de uma fonte específica de informação. No caso da DBPedia, o objetivo é mapear metadados de artigos garantindo assim a ligação de conteúdo correlato, o nosso é fazer isso com conjuntos de dados abertos.

Capítulo 4

Arquitetura Proposta

Neste capítulo é discutida a arquitetura proposta para implementação de um processo de publicação de dados abertos conectados. Na Seção 4.1 são apresentadas as camadas da arquitetura. A Seção 4.2 aborda o processo para extração dos dados nos bancos de dados e como são disponibilizados para publicação. Já na Seção 4.3 é apresentada a ferramenta UnBGOLD que realiza o processo de indexação semântica utilizando os metadados e ontologias e o processo para automatização da publicação. A Seção 4.4 trata da disponibilização dos conjuntos de dados para os usuários finais que podem ser publicados de maneira automatizada. A Seção 4.5 apresenta o que vem a ser o Catálogo de Conjuntos de Dados Abertos Conectados. Por fim, a Seção 4.6 aborda a interface de busca semântica no conjuntos de dados.

4.1 Arquitetura de Publicação de Dados

A arquitetura para publicação de dados abertos proposta tem por princípio estabelecer o gerenciamento de publicação de dados abertos, buscando o aumento da qualidade dos dados abertos. Como parâmetro de qualidade, utilizaremos a definição proposta por Berners-Lee [4] a partir do índice 5 Estrelas e também pela definição de Malamud et al. [27], em que os dados quanto mais atuais possuem maior valor. O aumento da qualidade é realizado através do enriquecimento semântico dos dados, o enriquecimento é possível através da definição de um vocabulário controlado que ofereça um significado semântico os dados. O enriquecimento semântico também viabiliza que os dados possam ser ligados a outros conjuntos de dados. Também possibilita que os conjuntos de dados possam ser publicados automaticamente, reduzindo a intervenção humana e garantindo a atualização dos dados em intervalos reduzidos. Também é objetivo a criação do Catálogo de Dados Abertos Conectados que consiste em um banco de dados onde é armazenado os metadados que identificam os conjuntos de dados que são publicados na arquitetura. Esses metadados

são descritos de forma semântica em triplas RDF, possibilitando que seja realizada busca em que o resultado é enriquecido semanticamente. A arquitetura dividida em três camadas que são:

- **Extração dos dados**: nesta camada são definidos os procedimentos de seleção, curadoria e extração dos dados, regras para acesso e disponibilização dos dados;
- **Indexação Semântica**: esta camada utiliza a ferramenta UnB - *Governamental Linked Open Data* (UnBGOLD) que oferece uma interface Web em que os Agentes Publicadores realizam a indexação semântica dos conjuntos de dados utilizando metadados e ontologias e determinam os parâmetros para publicação e atualizam o Catálogo de Dados Abertos Conectados;
- **Publicação dos Dados**: os dados são publicados pela ferramenta *Comprehensive Knowledge Archive Network* (CKAN) na qual ficarão disponíveis para os usuários finais utilizarem, além de uma interface de busca onde a pesquisa é realizada no banco de dados do Catálogo de Dados Abertos Conectados com o resultado enriquecido semanticamente.

Uma das características desta arquitetura é o desacoplamento entre as camadas finais, ou seja, a camada de extração pode ser substituída por qualquer outro serviço que ofereça os dados e a camada de publicação não é vinculada a uma instância CKAN específica. Desta forma é possível que a ferramenta possa publicar em diferentes instâncias. Esta característica é importante porque permite que o UnBGOLD tenha a flexibilidade necessária para ser utilizado por outros órgãos que estejam interessados nesta funcionalidade, desde que este órgão tenha definido as regras de acesso ao serviço de dados e mantenha uma instância CKAN habilitada para acesso via API do próprio CKAN. A Figura 4.1 apresenta a arquitetura de publicação de dados abertos.

Foi estipulado um processo simplificado que se deve seguir para publicação dos dados abertos. Inicialmente a Entidade Publicadora, a partir do plano de ação definido pelo seu PDA, seleciona os dados para abertura e faz a extração na fonte dos dados, gerando um arquivo em formato estruturado simples (CSV). O Agente Publicador ficará responsável por publicar os conjuntos de dados no portal de dados abertos da instituição em que ficará disponível para *download*, catalogar no PBDA e realizar a indexação semântica no UnBGOLD, além de definir os parâmetros para automatização da publicação. Os usuários finais poderão pesquisar por dados abertos em que os resultados serão enriquecidos por busca semântica através de consulta em SPARQL no banco de dados do Catálogo de Dados Abertos Conectados. A Figura 4.2 demonstra o processo proposto.

O desafio da indexação semântica consiste em duas partes:

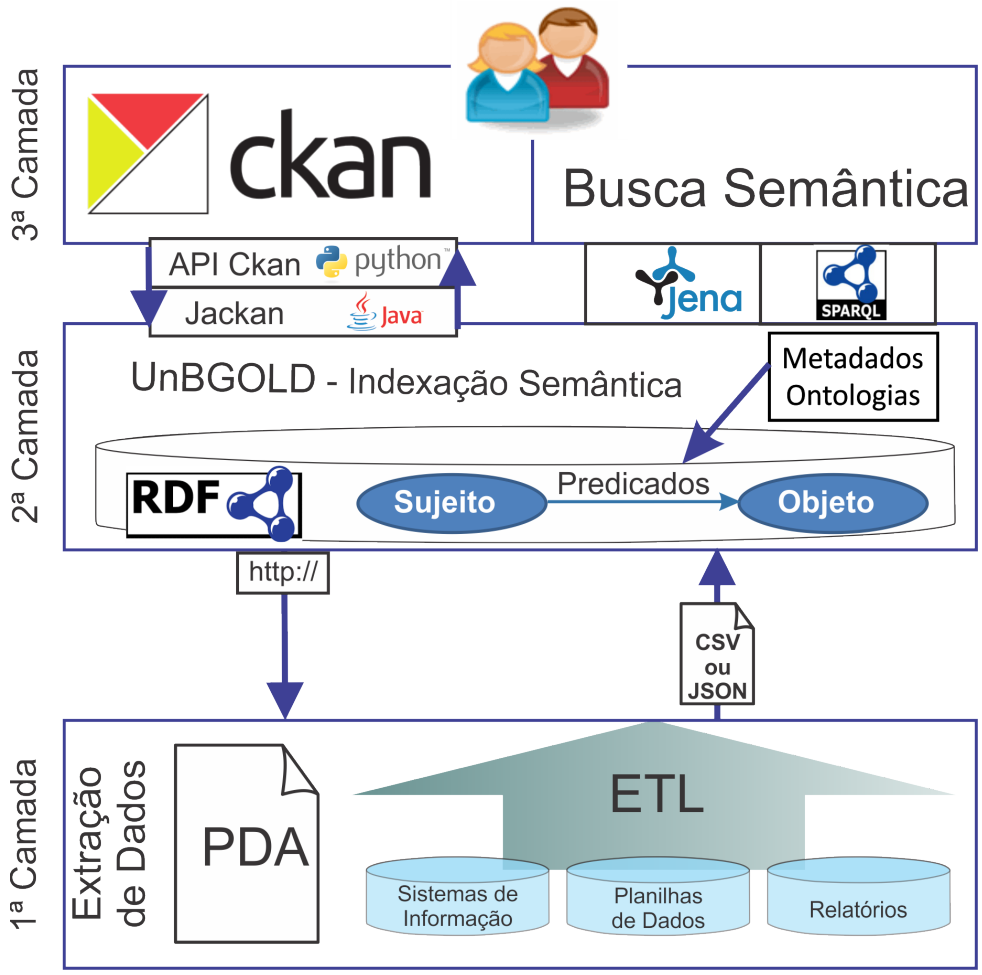


Figura 4.1: Arquitetura de Publicação de Dados Abertos
Fonte: Elaboração Própria

- Dar significado semântico aos dados que desejamos publicar, tornando-os assim dados abertos conectados;
- indexar a fonte de dados através da criação de um catálogo de conjuntos de dados conectados que irá conter as informações sobre os conjuntos publicados no qual será possível realizar pesquisas sobre as características dos dados e os resultados serão enriquecidos semanticamente.

As Seções 4.5 e 4.6 abordam como é realizada a catalogação semântica dos conjuntos de dados e a busca semântica respectivamente.

4.2 Extração de Dados

De acordo com o Decreto 8.777 de 2016, todos os órgãos da Administração Pública Federal devem publicar seu PDA que define como o órgão implementa a política de dados abertos

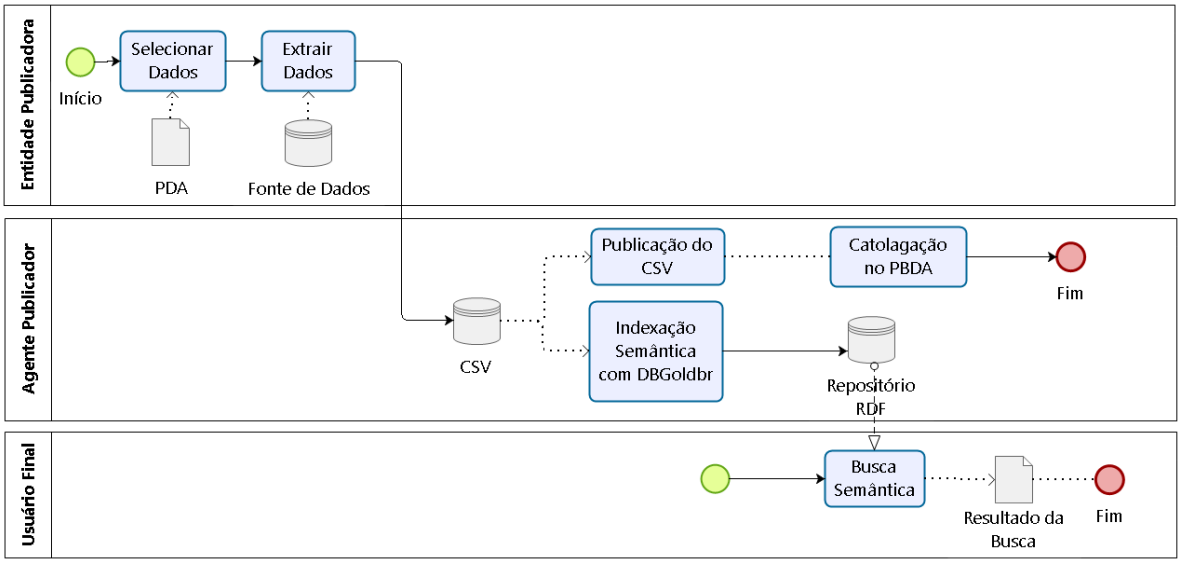


Figura 4.2: Processo de Publicação de Dados Abertos
Fonte: Elaboração Própria

definida pelo Governo Federal [13]. No PDA consta quais dados o órgão considera ser pertinente à publicação para a sociedade, o mapeamento sobre o domínio das informações, origem da fonte dados, quem são os responsáveis pelos dados e quais serão os atores que monitoram e dão sustentação à política de dados abertos. Geralmente a gestão do órgão define uma comissão permanente que terá a responsabilidade pela implantação da política, garantindo e fiscalizando para que o processo não seja interrompido. A fase de extração de dados deve aderir ao PDA na execução da atividade, no qual os dados são extraídos, limpos e validados por atores que realizam a curadoria, garantindo a sua integridade.

Cada órgão deve definir como irá realizar a extração dos dados. A arquitetura proposta define que os dados devem estar disponíveis na internet em formato estruturado aberto. Cada órgão tem a autonomia para decidir como irá implementar esta camada. Pode ser através de *web services* ou a simples disponibilização de arquivo CSV em um sítio institucional. Caso o órgão já tenha publicado o dado e só deseje a indexação semântica, pode ser utilizado o dado aberto já publicado através de sua URL.

4.2.1 Solução da UnB para camada de extração

Para a atividade de extração dos dados o CPD da UnB definiu que os dados para publicação devem passar por um processo de *Extract Transform Load* (ETL) e serem armazenados em um DW que seja disponível para consulta. Visto que o parque de sistemas de informações da UnB é bastante heterogêneo, optou-se por utilizar o DW como ponto

central de unificação das bases de dados, assim, no próprio processo de ELT já existe uma preparação dos dados, com curadoria da informação e garantindo a integridade dos dados.

A recuperação dos dados armazenados do DW é realizada por meio do barramento de serviços ErlangMS desenvolvido na UnB em meados de 2014, conforme apresentado na Seção 2.7, com o objetivo de unificar o acesso aos dados da universidade através de uma camada intermediadora entre componentes de software (denominados serviços) e as aplicações que consomem estes serviços [1]. Do ponto de vista da publicação dos dados abertos, a recuperação é transparente para a ferramenta UnBGOLD. A Figura 4.3 mostra como é a arquitetura de publicação proposta para UnB.

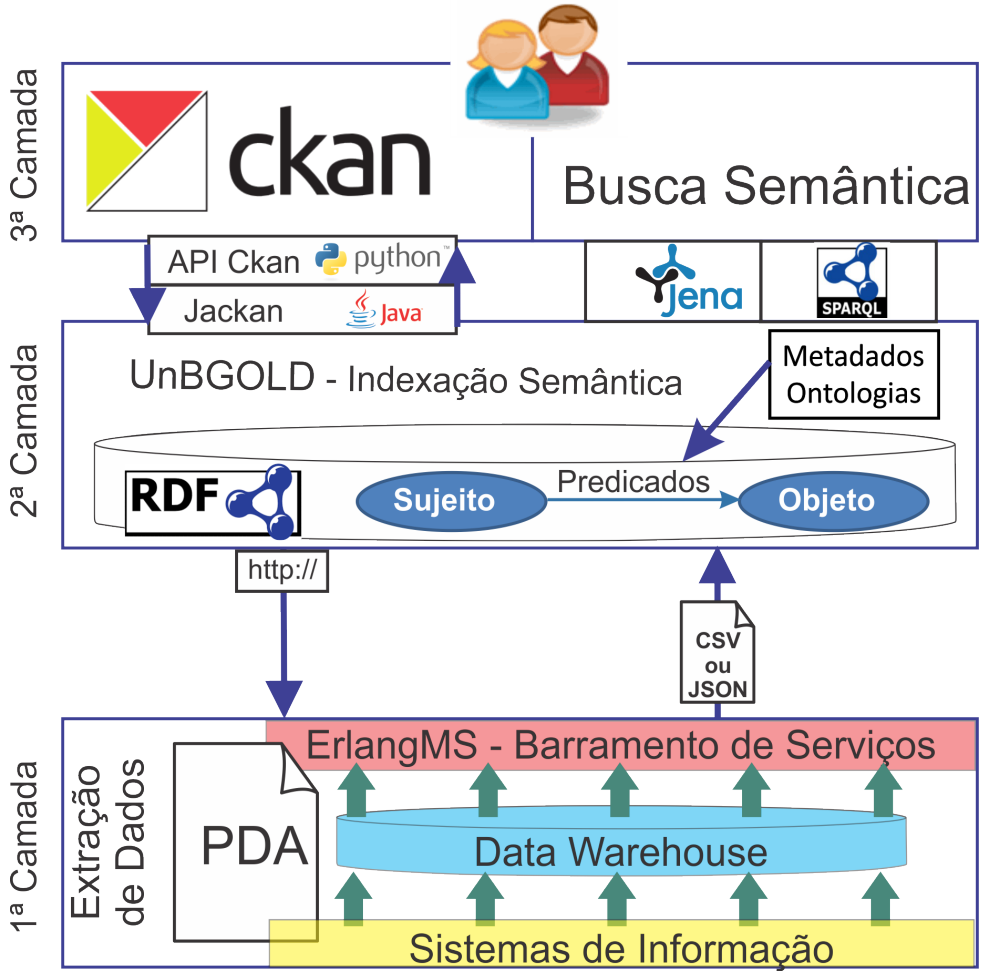


Figura 4.3: Arquitetura de Publicação de Dados Abertos da UnB
Fonte: Elaboração Própria

A Seção 4.3.3 aborda em detalhes como o Agente Publicador deve configurar a aplicação para recuperar os dados, tanto para indexação semântica quanto para a publicação dos dados.

4.3 UnBGOLD - UnB Governament Linked Open Data

O UnBGOLD é uma ferramenta que foi desenvolvida para auxiliar na publicação de dados abertos sendo a camada intermediária entre uma extração dos dados na suas fontes e a interface de consumo pelo usuário final, conforme podemos verificar na Figura 4.1. É no UnBGOLD o Agente Publicador configura os parâmetros de publicação e define os vocabulários que indexam os dados, quando a ferramenta publica automaticamente os dados, gerencia o Catálogo de Dados Abertos Conectados e oferece a interface de busca semântica ao usuários finais.

A Figura 4.1 apresenta a arquitetura de publicação de dados aberto da UnB, sendo que a primeira camada diz respeito à origem do dado em que se implementou uma solução através de um DW que já possui os dados limpos oriundos dos bancos de dados dos sistemas de informação da UnB. O gerenciamento da solicitação e entrega dos dados é feito por um barramento de serviços ErlangMS [1] que recebe uma solicitação através de uma requisição via HTTP e retorna os dados em formato estruturado, em formato aberto (CSV ou JSON).

A última camada diz respeito à interface de consumo final dos usuários. A UnB utiliza a solução CKAN[1] que é uma plataforma web desenvolvida pela *Open Knowledge Foundation* (OKF)[2] para publicação e compartilhamento de dados abertos. Além disso, é disponibilizada uma ferramenta de busca semântica nos dados já indexados.

4.3.1 Arquitetura

O UnBGOLD é uma aplicação web desenvolvida em linguagem Java, que utiliza diversas tecnologias conforme é apresentado na Figura 4.4.

Na camada de visão dos usuários é utilizada uma solução que integra *Angular*[3], *TypeScript*[4] e *Angular Material*[5], esta camada é acessada principalmente pelo Agente Publicador no momento em que gerencia os conjuntos de dados. Já os usuários finais acessam o UnBGOLD para utilização da ferramenta de busca semântica.

Para persistência dos dados de configuração e cadastros, a aplicação usa o sistema gerenciador de banco de dados MYSQL[6], sendo utilizado o *framework Hibernate* e a API JPA para fazer o mapeamento objeto-relacional e gerenciar a persistência dos dados. O

[1] https://ckan.org/
[2] https://okfn.org/
[3] `https://angular.io/`
[4] `https://www.typescriptlang.org/`
[5] `https://material.angular.io/`
[6] https://www.mysql.com/

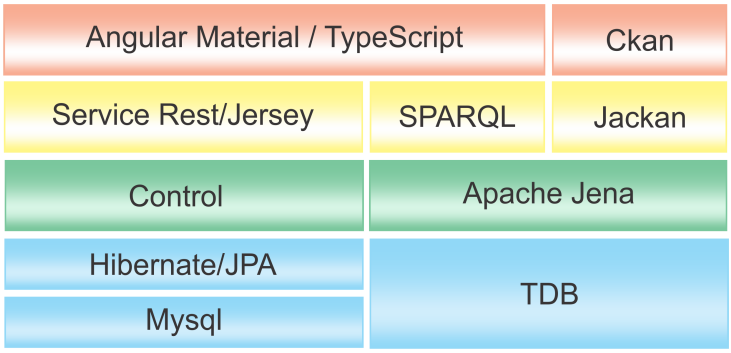


Figura 4.4: Arquitetura UnBGOLD
Fonte: Elaboração Própria

TDB[7], que é um componente do Apache Jena, é utilizado para o armazenamento de consultas e triplas RDF para salvar os dados já indexados.

Na camada de controle da aplicação, utilizamos a biblioteca Jersey para realizar o gerenciamento de serviços REST que alimentam os dados para a camada de visão. Além disso, usa-se a plataforma Apache Jena que é responsável pela manipulação das triplas RDF. A ferramenta disponibiliza uma interface de busca semântica que utiliza a linguagem SPARQL para pesquisar nas triplas RDF.

Por fim, a publicação dos conjuntos de dados é realizada através da ligação entre o UnBGOLD e a plataforma CKAN na qual a biblioteca em Java Jackan[8] se conecta à API do CKAN e gerencia os conjuntos de dados abertos. É preciso que o Agente Publicador informe quais são as chaves de acesso à instância do CKAN do órgão, onde a ferramenta realizará o gerenciamento das publicações.

4.3.2 Cadastros

Os principais focos do UnBGOLD é realizar a indexação semântica dos dados, automatizar a publicação e a busca semântica no Catálogo de Dados Abertos Conectados. Inicialmente é preciso realizar registros de informações que servirão de base para a realização dessas atividades, estes cadastros são:

- **Entidades Publicadoras**: Cadastros dos órgãos que desejam utilizar o UnBGOLD para auxiliar a publicação de dados abertos;
- **Agentes Publicadores**: Usuários que serão responsáveis por indexar e publicar os dados que são vinculados a entidades publicadoras;

[7]https://jena.apache.org/documentation/tdb/index.html
[8]`https://github.com/opendatatrentino/jackan`

- **Instâncias do CKAN**: Cadastrar as instâncias do CKAN que o UnBGOLD irá se comunicar para publicação dos dados;
- **Vocabulário**: Cadastro de Ontologias e Metadados e seus respectivos termos que serão utilizados pelos agentes publicadores para indexar os dados através do referenciamento da relação entre os dados.

4.3.3 Indexação e Publicação dos Dados

A interface oferecida para publicação de dados, foi definida de forma sequencial está dividida em quatro etapas, assim, o Agente Publicador deve inicialmente preencher as informações sobre os dados e depois realizar a indexação semântica em quatro etapas:

1. Publicação e Automatização
2. Informações Sobre os Dados
3. Vocabulário Controlado
4. Indexação Semântica

O detalhamento de cada etapa é realizada nas subseções a seguir.

Publicação e Automatização

A primeira etapa é a Publicação e Automatização. Nesta etapa o Agente Publicador preenche as informações necessárias para publicação, a Figura 4.5 apresenta esta tela. Inicialmente o Agente Publicador deve optar por automatizar a publicação, caso não opte por esta opção, terá a possibilidade de apenas realizar o *download* do arquivo RDF referente a indexação ao final do processo. Já para configurar a automatização, é necessário selecionar em qual instância CKAN das disponíveis deseja que os dados sejam publicados. A lista de instâncias CKAN que aparece como opção de seleção é referente as Entidades Publicadoras que o Agente Publicador tem vinculação, sendo que já deve estar previamente cadastrada.

Para automação é necessário informar a frequência com que a publicação ocorrerá, nesta opção o Agente Publicador deve selecionar uma das opções que pode ser diário, semanal, bimestral, semestral, semestre letivo e anual, além do horário que se deseja publicar. A responsabilidade que os dados sejam atuais é garantida pela implementação da camada de Extração de Dados, sendo que o UnBGOLD apenas faz a recuperação destes dados e não a atualização conforme detalhado na Seção 4.2.

A última opção desta tela é referente ao formato em que este dados serão publicados. É obrigatório selecionar pelo menos uma das opções apresentadas (CSV, JSON e/ou RDF).

Contudo, caso o Agente Publicador não opte pela opção RDF, não é necessário realizar os passos referentes a indexação semântica. Caso a publicação não esteja configurada para ser automática, a opção de publicação em RDF torna-se obrigatória, caso contrário o uso da ferramenta se tornaria irrelevante.

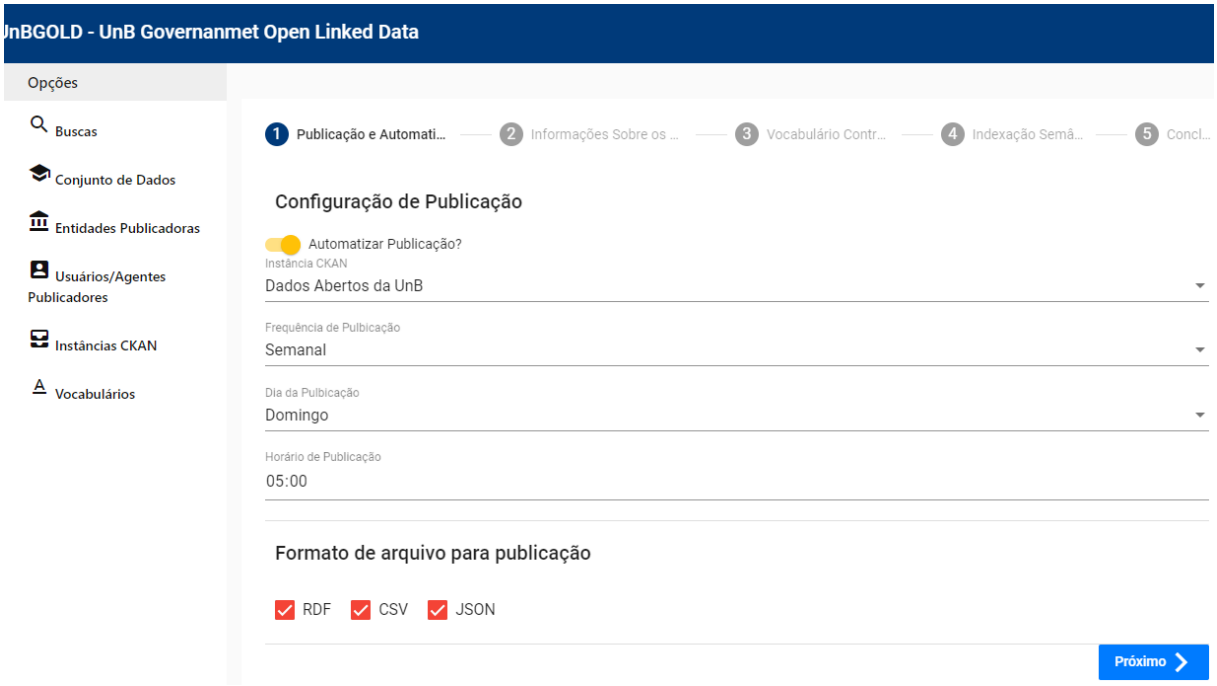


Figura 4.5: Etapa Publicação e Automatização
Fonte: Elaboração Própria

Informações Sobre os Dados

Na etapa "Informações Sobre os Dados" são preenchidas as informações que identificam estes dados, isto é feito informando os metadados referentes ao conjunto de dados. Figura 4.6.

A origem dos dados deve ser informada através do preenchimento do campo "Fonte de Dados", é preciso definir a origem dos dados por meio de uma URL por onde poderá ser realizado o *download* do arquivo que deverá estar disponível previamente. Também será possível informar parâmetros que serão utilizados na requisição HTTP. Eles poderão ser fixos, deixando explícito o valor que se deseja no parâmetro, ou temporal, ao qual o agente publicador deve informar uma das opções de valor temporal (dia, mês, semana, mês, semestre, ano...). O valor dos parâmetros que irão ser enviados serão referentes ao momento atual, assim se for "diário", o valor do parâmetro assume o dia da realização a requisição, por exemplo, se a requisição for realizada no dia 09 de novembro de 2018, o valor de um parâmetro temporal diário seria "09/11/2018". O UnBGOLD fará um teste de

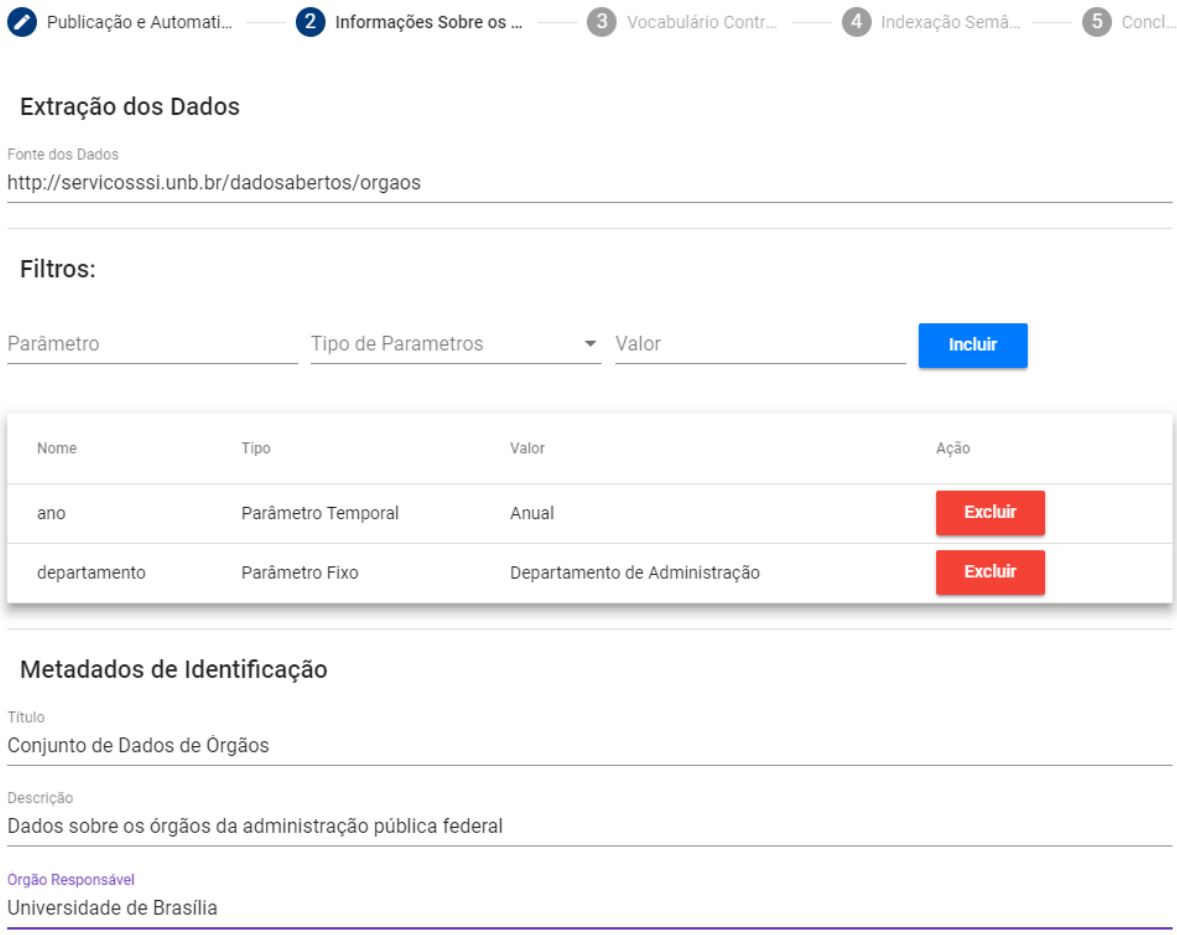


Figura 4.6: Etapa Informações Sobre o Dados
Fonte: Elaboração Própria

conexão acessando o conjunto de dados, caso funcione como o esperado, será apresentada mensagem de sucesso informando algumas características do conjunto de dados acessado. A Figura 4.6 apresenta a tela que contém o serviço de requisição dos dados e com dois parâmetros configurados de exemplo: um temporal e um fixo.

Em seguida são informadas as principais informações que identificam o conjunto de dados. Estes dados são padronizados a partir dos metadados definidos pela Cartilha de Publicação de Dados Abertos do Governo Federal[9]. Eles podem ser obrigatórios ou opcionais, sendo que alguns destes metadados são gerados automaticamente. Os metadados são utilizados para realizar a indexação semântica das fontes de dados. A Tabela 4.1 mostra quais são os metadados obrigatórios para catalogar o *dataset* e a Tabela 4.2 os metadados desejáveis, que são as informações que não são obrigatórias na catalogação no PBDA, mas que seria interessante informar este dados. Os metadados Autoria, Frequência de atualização, Referências e Vocabulário/ontologia são coletados nos dados de configura-

[9]http://dados.gov.br/pagina/cartilha-publicacao-dados-abertos

ção de publicação, já os demais metadados o Agente Publicador pode informar se achar conveniente.

Tabela 4.1: Metadados Obrigatórios no PBDA

Nome do Metadado	**Descrição**
Título	Nome do conjunto de dados
Descrição	Uma breve explicação sobre os dados
Catálogo origem	Página (URL) do órgão onde está publicado o conjunto de dados
Órgão responsável	Nome e sigla do órgão ou entidade responsável pela publicação do conjunto de dados
Categorias no VCGE	O Vocabulário Controlado de Governo Eletrônico é uma lista hierarquizada de assuntos do governo que utiliza termos comuns e é voltada para a sociedade. Para navegar e escolher as categorias acesse o VCGE em http://vocab.e.gov.br/2011/03/vcge
Recursos	Um conjunto de dados pode ser composto por mais de um arquivo de dados. O critério básico para separar vários recursos em mais de um conjunto de dados é a constatação de que eles divergem em vários metadados
Identificador	URL persistente que aponta para o recurso na Web
Formato	Formato do recurso. Ex.: XML, JSON, CSV, etc

Dentre os metadados obrigatórios destacamos o Vocabulário Controlado do Governo Eletrônico (VCGE), que é um vocabulário controlado com o objetivo de indexar informações (documentos, bases de dados, sites, etc) no âmbito do Governo Federal que possui uma lista de termos sobre diversos assunto no qual o governo atua [16]. O uso destes metadados tem a finalidade de categorizar o conjunto de dados por assunto e auxiliar no agrupamento dos dados.

Vocabulário Controlado

Na etapa de Vocabulário Controlado, é o momento em que são definidos quais serão os metadados e ontologias que serão utilizados para representar semanticamente os dados advindos da fonte. Salientamos que o vocabulário utilizado para o catálogo e os conjuntos de dados podem ser diferentes. É apresentada uma lista de metadados e ontologias cadastradas previamente, conforme já explicado no início na Subseção 4.3.2. É importante apontar que o Agente Publicador deve conhecer o conjunto de dados e também o vocabulário que deseja utilizar, assim ele irá selecionar o vocabulário que melhor se adéqua na descrição semântica dos dados. Como o cadastro de ontologias é teoricamente ilimitado, é importante que seja realizado uma seleção de qual vocabulário utilizar para facilitar a escolha dos termos. A Figura 4.7 apresenta esta tela.

Tabela 4.2: Metadados Desejáveis no PBDA

Nome do Metadado	**Descrição**
Etiquetas	Lista de palavras chaves relacionadas ao conjunto de dados, e que são úteis na classificação e busca dele
Autoria	Instituição ou pessoa responsável pela produção do recurso
Documentação	URL de documento que expõe detalhes sobre o conjunto de dados.
Cobertura geográfica	Localização ou região geográfica a que se referem os dados. Ex.: Recife.
Cobertura temporal	Data ou período à que referem os dados. Ex.: 03/2012.
Granularidade geográfica	Precisão geográfica da cobertura geográfica. Ex.: municipal
Granularidade temporal	Precisão temporal da cobertura temporal. Ex.: mês.
Frequência de atualização	Frequência temporal com que o conjunto de dados é atualizado
Referências	Relações com outros conjuntos de dados
Metodologia	Processo de criação dos dados
Vocabulário/ontologia	Documentos estruturados com metadados específicos do conjunto de dados

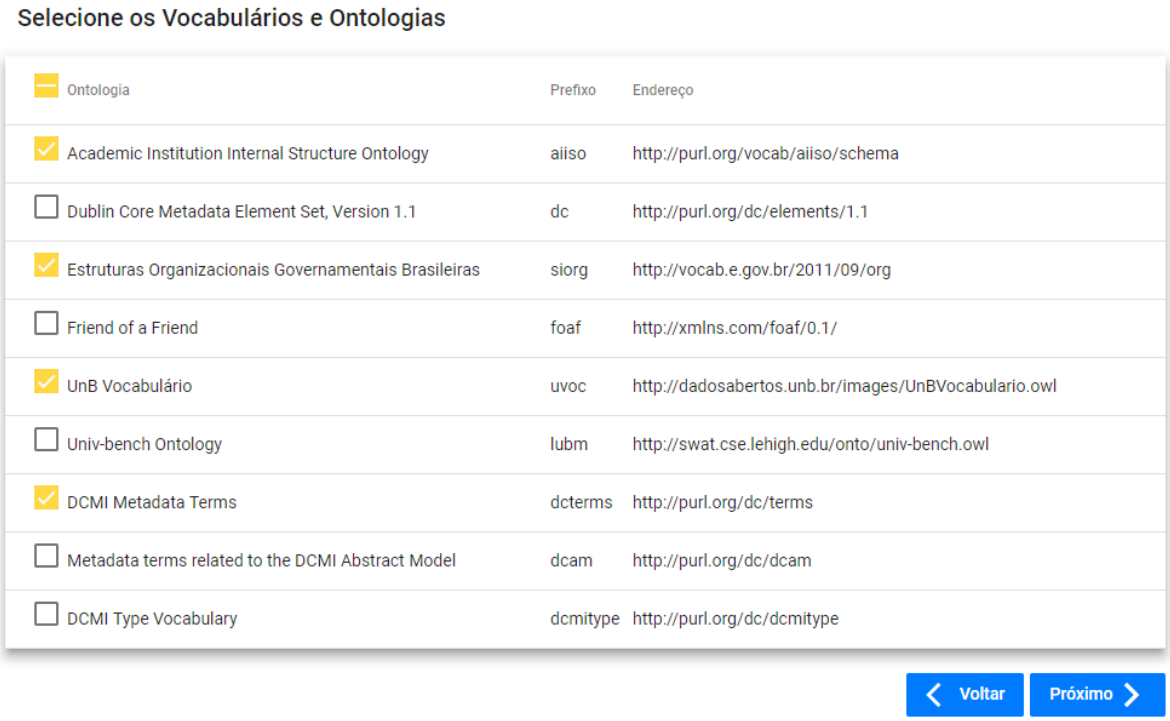


Figura 4.7: Etapa Vocabulário Controlado
Fonte: Elaboração Própria

Indexação Semântica

A última etapa para publicação dos dados é a "Indexação Semântica". Esta etapa consiste em dar um significado semântico aos dados por meio de termos do vocabulário controlado

originados das ontologias selecionadas e, para isso, é necessário transformar o conjunto de dados estruturados em triplas RDF. Supondo que os dados estejam em formato CSV, a transformação dos dados para dados conectados é realizada de maneira que as triplas do RDF serão formadas de modo que cada linha do CSV será um sujeito e que os dados contidos em cada coluna da linha serão os objetos. Um termo oriundo de um vocabulário controlado deve ser escolhido para ser o predicado formando a tripla RDF. A Figura 4.8 mostra um exemplo de como é realizada a transformação:

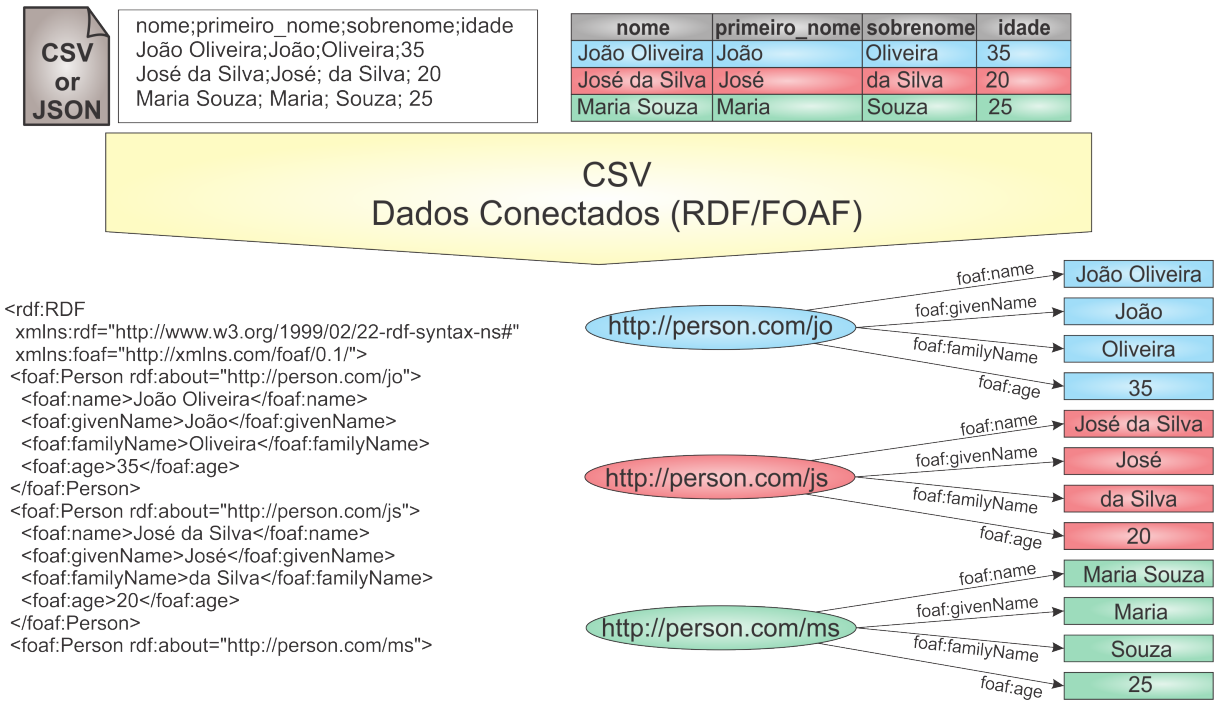


Figura 4.8: Transformação de CSV em RDF
Fonte: Elaboração Própria

Na parte superior da Figura 4.8 temos um conjunto de dados estruturados de exemplo, e neste caso é em formato CSV. Para melhor compreensão são apresentados os dados também em forma de tabela em que é possível verificar que existem quatro colunas (nome, primeiro_nome, sobrenome e idade) no conjunto dados e as três linhas com as informações que optou-se dividir em cores (azul, vermelho e verde). Selecionamos a ontologia *Friend of a Friend* (FOAF) para dar significado semântico aos dados no qual os vocabulários "foaf:name", "foaf:givenName", "foaf:familyName" e "foaf:age" descrevem os dados semanticamente. O resultado da transformação pode ser visto na parte inferior da Figura 4.8 onde temos a descrição em RDF/XML e também os grafos das triplas geradas. Das linhas do conjunto de dados originou-se um Recurso indexado com um URI, ou seja, o Sujeito da Triplas, assim, criamos 3 recursos identificados visualmente pela sua respectiva cor. Os objetos das triplas são formados pelos dados contidos em cada linha, assim, para

cada linha havia quatro informações que originaram seus respectivos objetos. Já a ligação entre sujeitos e objetos é realizada através do predicado que foi definido através do nosso vocabulário controlado.

Para dar significado semântico aos dados é necessário dividir em duas parte a etapa de indexação, na primeira parte são informados os parâmetros que definem o tipo daquele dado e a identificação de como será formado o sujeito da tripla através da URI. O tipo do dado é definido através da seleção de um dos termos advindos das ontologias ou metadados selecionados na etapa anterior, esta informação vai caracterizar a descrição semântica do tipo de informação que possui o conjunto de dados e ajudará na ligação com outros tipos de dados. Depois é necessário definir o sujeito do RDF, isto pode ser feito com a combinação de um campo texto com os dados de um campo advindo do conjunto de dados ou a simples definição de um campo que será o URI, contanto que o valor resultante seja uma URL válida e única. Na Figura 4.9 é possível visualizar a tela que representa esta etapa.

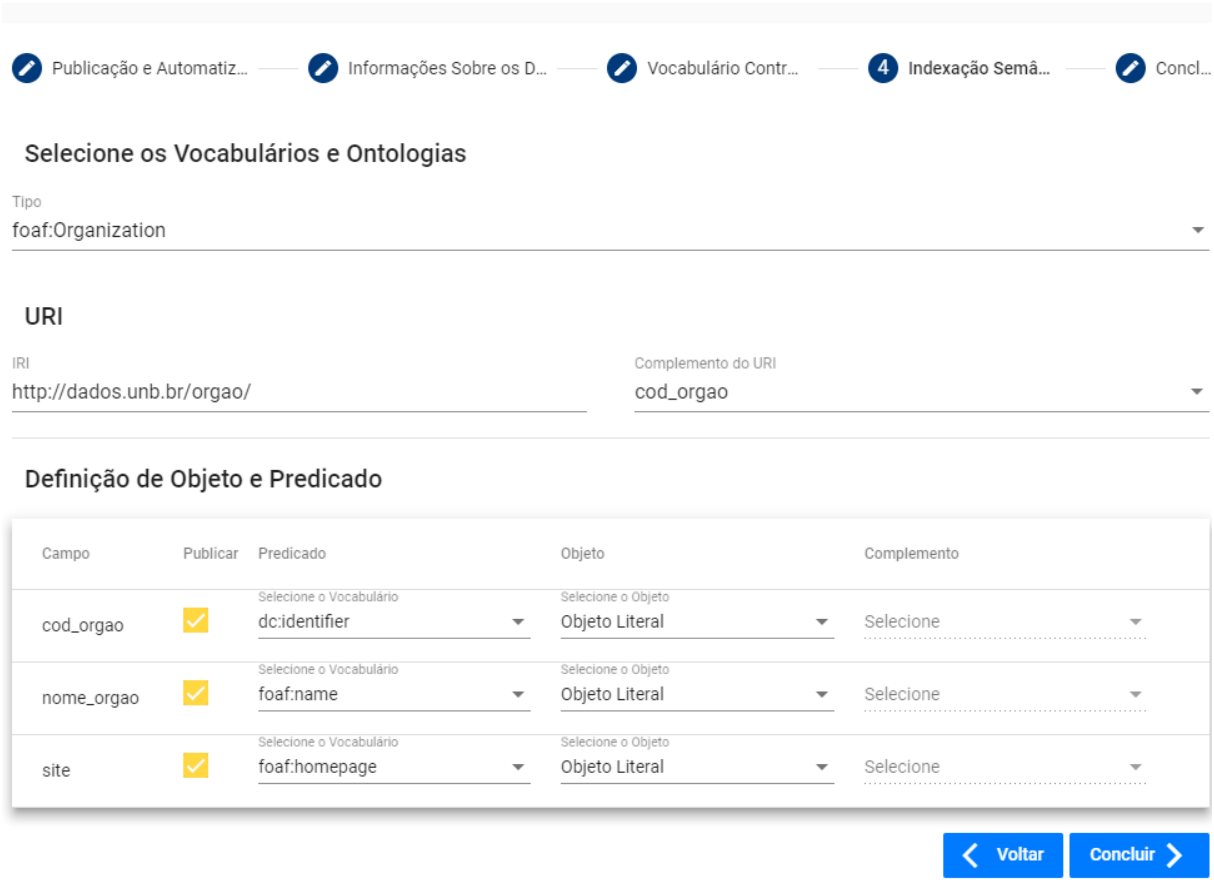


Figura 4.9: Etapa Indexação Semântica
Fonte: Elaboração Própria

A segunda parte consiste na definição dos predicados e objetos em que é configurado

a indexação semântica as linhas do conjunto dados. É apresentada a tabela de indexação com as listas das colunas do conjunto de dados conforme visto na Figura 4.9. As duas primeiras colunas da tabela são utilizadas para identificar o campo do conjunto de dados e habilitar os dados deste campo para indexação. Não é obrigatório indexar todos os campos, o Agente Publicador deve decidir, de acordo com o seu conhecimento sobre o conjunto de dados, se deve ou não indexar o campo por acreditar que o dado seja irrelevante para indexação ou se o dado pode vir a se tornar redundante a partir da conexão dos dados.

Na próxima coluna da tabela indexação é escolhido qual será o Predicado que descreve semanticamente a relação entre o sujeito e o objeto. Neste momento, o Agente Publicador irá selecionar qual vocabulário será utilizado para descrever a relação a partir do termos que existem nos metadados e/ou ontologias selecionados no passo anterior.

A quarta coluna da tabela é a definição do objeto da tripla, a primeira opção oferecida para escolha é que o Objeto seja um Literal, nesta escolha, dados bruto vindos do conjunto de dados assumirá o valor da entidade Objeto. Se o objeto não for um Literal, o Agente Publicador terá a opção de selecionar o conjunto de dados já cadastrado no UnBGOLD em que o predicado é igual ao tipo do dados do conjunto de dados. Ao selecionar o conjunto de dados são apresentados ao lado quais as opções para os campos do conjunto de dados. É neste momento que é feita a integração com conjunto de dados diferentes, pois a ferramenta vai procurar dentro das triplas RDF o recurso que se relaciona com aquele dado. Caso não encontre um recurso, o sujeito é gerado como um Literal. No Capítulo 5 é apresentado o estudo de caso explicando como é realizado e o resultado final.

4.4 Publicação de Dados

A comunicação entre UnBGOLD e CKAN é realizada através de uma API que o CKAN disponibiliza, em que é possível gerenciar os conjuntos de dados, inserindo, atualizando e excluindo conjuntos de dados [2]. Para fazer a comunicação com a API, agregamos ao UnBGOLD a biblioteca Jackan para automatização. Utilizaremos as APIs do CKAN que possibilitam realizar a importação de conjuntos de dados de outras aplicações, sendo que a fonte de dados será disponibilizada pela nossa ferramenta de extração e indexação. Esse processo irá realizar nova extração e indexação semântica e fará a atualização dos conjuntos de dados já cadastrados. Os parâmetros definidos na Seção 4.3.3 serão mantidos, a não ser que exista algum parâmetro temporal que será atualizado automaticamente.

Após o processo de publicação, o Catálogo de Conjuntos de Dados Abertos Conectados é atualizado automaticamente e armazenado em um banco de dados TDB e ficará disponível para consulta aos usuários como iremos discutir na Seção 4.6.

4.5 Catálogo de Conjuntos de Dados Abertos Conectados

O Catálogo de Conjunto de Dados Abertos Conectados é um banco de dados onde são armazenadas os metadados que expressam as características do conjunto de dados publicados de modo que também visa facilitar a pesquisa sobre os dados. A descrição destes dados de forma semântica é desejável, dado que também ficará disponível de forma conectada. Um dos objetivos deste trabalho é propor um conjunto de metadados para que possam identificar as características dos dados e descrevê-los semanticamente e através da definição um vocabulário controlado.

Para seleção de um vocabulário controlado optamos por utilizar a especificação DCMI. A DCMI define um conjunto de termos que pretende descrever objetos digitais existentes na Internet sendo que seus termos são classificados em Simples e Qualificado [23]. A especificação Simples possui 15 termos comuns, já o nível Qualificado possui diversos vocabulários que buscam trazer mais granularidade na descrição semântica das informações. A escolha dos termos está relacionada aos dados que serão indexados. Abaixo listamos os conjuntos de vocabulários escolhidos:

- *Dublin Core Metadata Element Set* (dc)
- *DCMI Metadata Terms* (dcterms)
- *Metadata terms related to the DCMI Abstract Model* (dcam)
- *DCMI Type Vocabulary* (dcmitype)
- *Friend of a Friend* (foaf)

Quando vamos referenciar um vocabulário, usualmente, é utilizado o prefixo ao qual ele representa, como é realizado nesta pesquisa. Dentre os vocabulários citados o único que não faz parte da especificação DCMI é a ontologia FOAF.

A definição dos dados parte inicialmente das informações que são preenchidas na Seção 4.3.3 que se baseia na Cartilha de Publicação de Dados Abertos. Além disso, definimos um conjunto de informações que podemos extrair da aplicação sem a necessidade de novas intervenções humanas. A Tabela 4.3 apresenta a relação dos metadados com seu respectivo vocabulário/termo.

A maioria dos metadados possui um valor único no catálogo, contudo, os metadados Formato, Referência, Etiquetas e Vocabulário, podem ocorrer mais vezes de uma vez. O metadado Referência, diz respeito se o conjunto de dados está ligado a outros. Como apresentado na Seção 4.3, é possível realizar a ligação entre os dados, assim este metadado é extraído automaticamente pela aplicação.

Tabela 4.3: Metadados e Vocabulário do Catálogo

Metadado	**Vocabulário**	**Termo**	**Tipo**
Identificador	dc	identifier	Literal
Título	dc	Title	Literal
Descrição	dc	description	Literal
Órgão Responsável	foaf	Organization	Literal
Vocabulário VCGE	dcam	VocabularyEncodingScheme	Recurso
Identificador URL	dcterms	source	Recurso
Formato	dcterms	FileFormat	Literal
Etiquetas (tags)	dc	subject	Literal
Frequência	dcterms	Frequency	Literal
Referência	dcterms	Relation	Recurso
Metodologia	dcterms	instructionalMethod	Literal
Vocabulário	dcam	VocabularyEncodingScheme	Recurso
Tipo	dc	type	Recurso
Data	dc	date	Literal
Data de Criação	dcterms	created	Literal
Linguagem	dcterms	language	Literal
Publicação	dcmitype	Dataset	Recurso

Definidas as informações e os vocabulários é necessário descrever estes dados semanticamente e isto é realizado através da construção de triplas RDF discutidas na Seção 2.5. Cada publicação tem um endereço na Internet específico onde ficará disponível para acesso aos usuários, assim, definimos que a URL de hospedagem dos dados será utilizada como IRI que indexa o nosso recurso na web, sendo assim o sujeito do catálogo. Os objetos serão descritos com as informações coletadas pelos sistemas, sendo que a Tabela 4.3 informa quais são literais e quais apontaram para outros recursos. As triplas formadas pelos metadados "Referência" serão associadas a ligação com outros conjuntos de dados, em que o Objeto será o Recurso referente às publicações existentes no catálogo de publicações. A Seção 5.8 apresenta o exemplo de como é realizada a catalogação.

4.6 Busca Semântica

A Busca Semântica consiste em disponibilizar uma interface para que os usuários possam pesquisar no Catálogo de Dados Abertos Conectados e que o resultado da busca terá um aumento na qualidade das respostas através do enriquecimento semântico. Para realização da consulta, utilizamos uma linguagem SPARQL.

SPARQL é uma linguagem desenvolvida para realização de consulta em dados estruturados em formato RDF. Um comando de busca SPARQL é estruturado em duas cláusulas: SELECT e WHERE. No SELECT são definidas as variáveis que são recuperadas no re-

sultado da consulta enquanto no WHERE identificamos as variáveis dentro das triplas e realizamos os filtros desejados [31].

Nossa pesquisa será estruturada de forma que o usuário informará uma cadeia de caracteres sobre o assunto que deseja pesquisar, que é chamado de termo, será empreendida uma busca sobre ocorrência do termo nos metadados existentes no catálogo. Encontrando um ou mais recursos disponíveis é apresentado o resultado em forma de uma lista com informações reduzidas, no qual o usuário poderá verificar com detalhes sobre o recurso. Na Seção 5.8 é apresentado um exemplo de pesquisa realizada e o detalhamento de um conjunto de dados.

Capítulo 5

Estudo de Caso

Este capítulo é destinado a realizar um estudo de caso sobre a publicação de alguns conjuntos de dados da UnB a partir da arquitetura proposta no capítulo anterior. O Capítulo inicia-se apresentando e descrevendo o modelo lógico da arquitetura proposta na Seção 5.1. Na Seção 5.2 é abordada a seleção dos conjuntos de dados para abertura e uma análise desses dados. A Seção 5.3 apresenta como é realizada a extração dos dados. A Seção 5.4 trata da seleção de um vocabulário controlado para descrever semanticamente os dados. A Seção 5.5 apresenta a realização da indexação semântica dos conjuntos de dados. A Seção 5.6 aborda a publicação dos dados para os usuários finais. Na Seção 5.7 discorre-se sobre o resultado do processo de publicação com os dados em formato RDF e se realiza, a ligação entre os dados. Por fim na Seção 5.8 temos a criação do Catálogo de Dados Abertos Conectados e como é realizada a busca semântica.

5.1 Modelo Lógico da Arquitetura

A Figura 5.1 apresenta o modelo lógico da arquitetura de publicação de dados abertos que possibilita que o processo possar ser automatizado e que os dados possam ser enriquecidos semanticamente tornando-os conectados.

A origem da publicação encontra-se nas Entidades Publicadoras que são instituições da administração pública que devem respeitar a legislação vigente [13] [12] e publicar seus dados em formato aberto para comunidade. Esses dados devem ser coletados dentro das bases de dados dos seus sistemas de informação. No caso da UnB, os dados são extraídos das bases de dados dos sistemas que informatizam seus processos, como, por exemplo, o SIPES, SIGRA, SIPPOS e SIEX ou qualquer outro sistema que vier a substituí-los. Os dados são limpos e tratados para que sejam importados para um DW onde ficarão disponíveis para consulta.

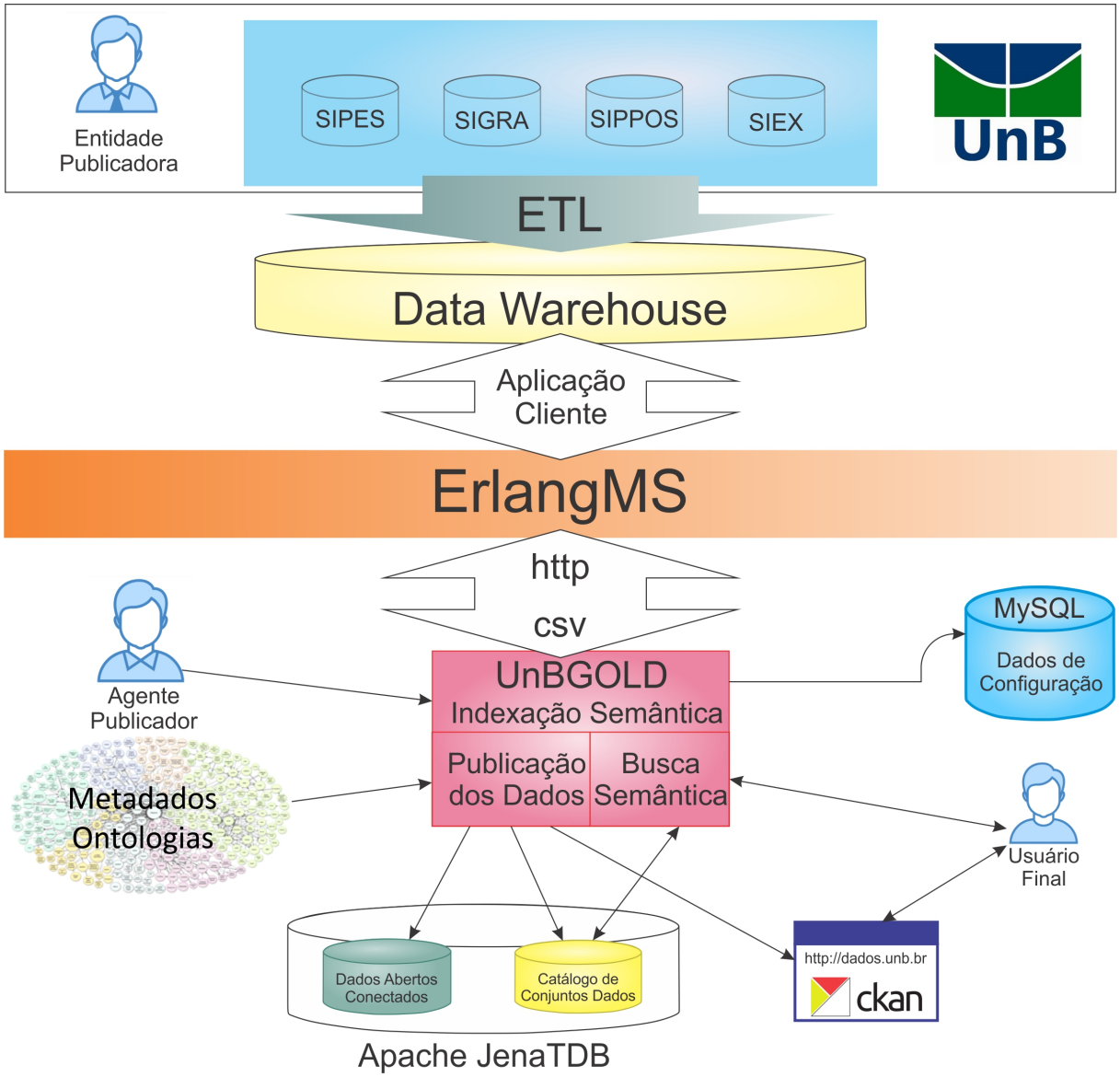


Figura 5.1: Modelo Lógico da Arquitetura de Publicação de Dados Abertos
Fonte: Elaboração Própria

A entrega dos dados será realizada através de requisição HTTP ao barramento de serviço ErlangMS, que envia a solicitação para uma aplicação cliente que extrai os dados do DW de acordo com a regras de negócio estabelecida pela Entidade Publicadora e retorna para o solicitante em formato já estruturado e aberto, no caso em CSV.

Recebido os dados, o Agente Publicador tratará os dados para consumo dos usuários finais através da ferramenta UnBGOLD . Para realizar a indexação semântica, o Agente Publicador seleciona um vocabulário composto de Metadados e Ontologias que possam descrever semanticamente os dados. Neste processo os dados indexados serão descritos em formato de triplas RDF e serão armazenados em um banco de dados TDB. Também será gerado um banco de dados com os metadados selecionados para descrever as princi-

pais características dos conjuntos de dados denominado Catálogo de Conjunto de Dados Abertos. Por fim, configuram-se os parâmetros que proporcionarão que os dados sejam publicados automaticamente no portal de dados abertos da UnB que é uma instância da plataforma CKAN. Enquanto os dados conectados são armazenados em bancos TDB, os dados de configuração são armazenados em um banco de dados relacional MySQL.

O Usuário Final consome os dados diretamente no portal dos dados abertos da UnB, onde poderá baixar os dados como bem entender e também é disponibilizada uma interface de busca no UnBGOLD que recupera informações, consultando no banco de dados as triplas RDF do catálogo.

5.2 Análise e Seleção dos Dados

Este estudo de caso utiliza conjuntos de dados da UnB que ela pretende abrir. Como a UnB ainda não publicou seu PDA[1], os presentes dados foram obtidos via solicitação formal ao CPD. Mesmo assim, uma das premissas desta arquitetura é que a seleção dos dados deve respeitar o PDA da instituição, então tomaremos por princípio que estes dados estão entre os dados candidatos para a abertura da UnB a partir do acesso a versão do PDA em fase de aprovação.

Como a UnB é um instituição de ensino superior, optamos por dados de cunho acadêmico que acreditamos ser de maior interesse público e, desta maneira, potencializar uma análise sobre a sua atividade fim.

A seguir temos a seleção dos conjuntos de dados e um breve explicação sobre eles.

- Departamentos: dados referentes aos departamentos internos da UnB;
- Cursos: dados referentes às cursos de graduação da UnB, estes cursos são ofertados pelos departamentos da UnB;
- Disciplinas: Dados referentes as disciplinas que são ofertadas pelos departamentos;
- Professores: Dados referentes ao corpo docente da UnB. Este conjunto de dados apresenta apenas informações que já são de domínio público, não apresentando dados pessoais que configurem quebra de privacidade;
- Ofertas: Os dados das ofertas são referentes às disciplinas que são oferecidas pelos departamentos em um determinado período letivo, obrigatoriamente elas possuem um professor vinculado.
- Fluxos de Curso: É denominado o fluxo de um curso as disciplinas que são obrigatórias para que um aluno passa concluir a curso.

[1] `http://paineis.cgu.gov.br/dadosabertos/index.htm`

As Tabelas 5.1, 5.2, 5.3, 5.4, 5.5 e 5.6 apresentam os campos referentes a cada conjunto de dados com uma breve descrição e a característica dos dados.

Tabela 5.1: Conjunto de Dados de Departamentos

Campos	**Descrição**	**Tipo de Campo**
COD_DEPARTAMENTO	Código identificador do Órgão	Numérico Inteiro
NOME_DEPARTAMENTO	Nome do Órgão	Carácter
SIGLA	Sigla do Cargo	Carácter
COD_ORGAO	Código identificador do Órgão	Numérico Inteiro

Tabela 5.2: Conjunto de Dados de Cursos

Campos	**Descrição**	**Tipo de Campo**
cod_curso	Código identificador do Curso	Numérico Inteiro
nome_curso	Nome da curso	Carácter
cod_departamento	Código identificador do departamento do curso	Numérico Inteiro
sigla_departamento	Sigla do departamento da curso	Carácter
nome_departamento	Nome do departamento do curso	Carácter

Tabela 5.3: Conjunto de Dados de Disciplinas

Campos	**Descrição**	**Tipo de Campo**
cod_disciplina	Código identificador da Disciplina	Numérico Inteiro
nome_disciplina	Nome da Disciplina	Carácter
cod_departamento	Código identificador do departamento da disciplina	Numérico Inteiro
sigla_departamento	Sigla do departamento da disciplina	Numérico Inteiro
nome_departamento	Nome do Departamento da disciplina	Carácter
creditos	Quantidade de crédito que possui a disciplina	Numérico Inteiro

Tabela 5.4: Conjunto de Dados de Professores

Campos	**Descrição**	**Tipo de Campo**
matricula	Matrícula do Professor	Numérico Inteiro
nome_professor	Nome do Professor	Carácter

Definidos os dados, inicialmente optamos por automatizar a publicação, isso implica em selecionar uma instância CKAN que está previamente cadastrada. O Agente Publicador deve ter vinculação com alguma Entidade Publicadora, na lista de opções serão visualizadas apenas as instâncias cadastradas para o seu órgão. Nossos conjuntos de dados serão publicados na instância do CKAN da UnB disponível no domínio `http://dados.unb.br`, no entanto, escolhemos para este estudo de caso utilizar uma instância criada especificamente para homologação do processo.

Tabela 5.5: Conjunto de Dados de Ofertas de Disciplinas

Campos	**Descrição**	**Tipo de Campo**
cod_oferta	Código identificador do Oferta	Numérico Inteiro
turma	Nome da Turma	Carácter
vagas_oferecidas	Quantidade de vagas oferecidas	Numérico Inteiro
periodo	Período letivo em que a disciplina é oferecida	Numérico Inteiro
matricula_professor	Matrícula do professor que ministra a disciplina	Numérico Inteiro
cod_disciplina	Código identificador da disciplina oferecida	Numérico Inteiro
cod_departamento	Código identificador do departamento	Numérico Inteiro

Tabela 5.6: Conjunto de Dados de Fluxo de Disciplinas de Cursos

Campos	**Descrição**	**Tipo de Campo**
sigla_departamento	Sigla do Departamento	Carácter
nome_departamento	Nome do Departamento	Carácter
cod_curso	Código identificador da Curso	Numérico Inteiro
nome_curso	Nome da Curso	Carácter
cod_habilitacao	Código identificador da Habilitação	Numérico Inteiro
nome_habilitacao	Nome da Habilitação	Carácter
cod_disciplina	Código identificador da disciplina	Numérico Inteiro
periodo	Periodo da disciplina no fluxo	Numérico Inteiro
prioridade	Prioridade da disciplina no fluxo	Numérico Inteiro

A análise dos dados demonstrou que o ideal é que o conjunto de dados de Professores seja publicado com frequência "Mensal", visto que o ingresso e saída de docentes que ocorre ao longo do mês, ao mesmo tempo que os demais conjuntos devem ser publicados com frequência "Período Letivo" por conta da característica dos dados que são atualizados no início de período letivo da universidade. Porém, para este estudo de caso, optou-se pela publicação de dados seja realizada de maneira "Diária" e o horário foi definido para as 5h00. Este horário não será exato, pois dependerá de fatores como a infraestrutura de processamento, tráfego na rede e fila de publicações, ou seja, este horário define quando será iniciado o processo de publicação e não a publicação em si.

O último item que devemos selecionar é quais serão os formatos dos dados que serão publicados. É aconselhado sempre publicar em um formato que seja de fácil manipulação por usuários leigos, assim decidimos por publicar em formato CSV e JSON para análises mais simplistas e em formato RDF para análise semântica.

5.3 Extração dos Dados

Na etapa de Informação dos Dados, o Agente Publicador irá definir os parâmetros necessários para recuperação dos dados e informar características importantes que serão úteis

para os usuários finais.

Na primeira parte iremos configurar a Extração dos Dados. Partimos da premissa que os serviços que disponibilizam os dados já estão ativos e já são considerados abertos, ou seja, não existe nenhum dispositivo que impeça ou dificulte seu acesso e processamento por máquina e está disponível através da Internet, assim, os dados são recuperados a partir de uma requisição HTTP realizada pela ferramenta UnBGOLD.

No estudo de caso, a disponibilização dos conjuntos de dados é realizada através do barramento de serviço ErlangMS conforme visto na subseção 4.2.1 a partir da arquitetura proposta para a UnB. Estes dados já devem vir em formato estruturado aberto (CSV, JSON ou XML), no nosso caso, configuramos o serviço para disponibilização dos dados em CSV.

Na Tabela 5.7 temos a descrição das URLs onde é acessado o serviço que disponibiliza os dados através do barramento.

Tabela 5.7: URL de Requisições

Conjunto de Dados	**URL**
Departamentos	http://servicos.unb.br/dadosabertos/departamentos
Cursos	http://servicos.unb.br/dadosabertos/cursos
Professores	http://servicos.unb.br/dadosabertos/professores
Disciplinas	http://servicos.unb.br/dadosabertos/disciplinas
Oferta	http://servicos.unb.br/dadosabertos/oferta
Fluxo	http://servicos.unb.br/dadosabertos/fluxo

Na Figura 5.2 é verificado como é realizado o roteamento da comunicação do ErlangMS. O Cliente do serviço, que no nosso caso seria a próxima camada da arquitetura, realiza uma requisição HTTP ao barramento na primeira via de comunicação através de uma url (http://servicos.unb.br/dadosabertos/oferta), já na segunda via, o barramento se comunica com a aplicação de serviço que acessa o *Data Warehouse* e retorna os dados solicitados.

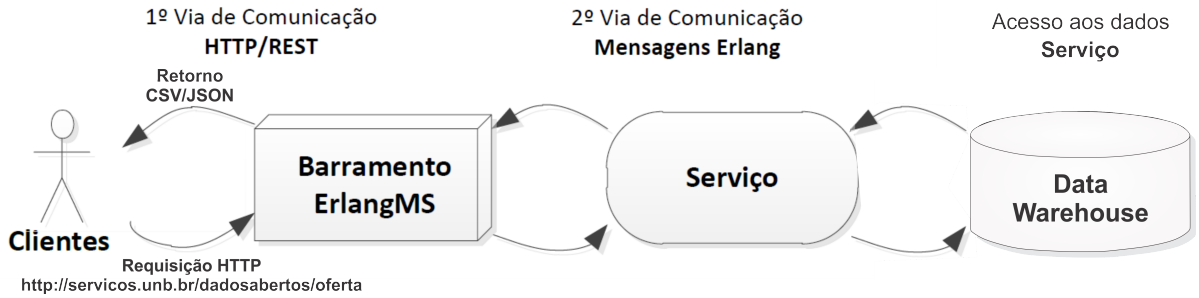


Figura 5.2: Esquema do Roteamento das Mensagens/Comunicação ErlangMS
Fonte: [17] Com Adaptações

No próximo passo são definidos os parâmetros que serão acrescidos na requisição. O conjunto de dados de Professores possui um parâmetro temporal chamado “mes” valor “mensal”, os demais conjuntos de dados possuem um parâmetro chamado “periodo” com valor temporal “periodo letivo”, visto que estes dados são dados acadêmicos que são atualizados a cada semestre letivo.

Na segunda parte desta etapa, são preenchidos os dados que informam as características do conjunto de dados. Conforme já visto na Tabela 4.1, optamos por utilizar o padrão definido pela Infraestrutura Nacional de Dados Abertos (INDA). Destacamos que os campos “Órgão Responsável”, “Catálogo Origem”, “Identificador” e “Formato” serão gerados automaticamente. No campo “Categoria VCGE”, o Agente Publicador deverá informar um dos termos oriundos deste vocabulário principal, visto a característica dos nossos conjuntos de dados, o termo Vocabulário Controlado do Governo Eletrônico (VCGE) que utilizamos para todos foi "Educação Superior". Na parte de metadados desejáveis, os campos “Autoria” e “Frequência de Atualização” já virão preenchidos com informações colhidas anteriormente, cabendo ao Agente Publicador a opção de retirá-las quando achar necessário, e optamos por deixá-las. Já os campos “Referência” e “Vocabulário/Ontologia” serão fornecidos com as informações que serão preenchidas nas próximas etapas.

5.4 Seleção de Vocabulário Controlado

Na etapa de Vocabulário Controlado, o Agente Publicador deve identificar quais são os vocabulários e ontologias que são utilizados para dar semântica aos dados. Este vocabulário será composto por um conjunto de metadados ou ontologias que possui termos que podem descrever as características de um dado. Cada conjunto de dados tem suas características específicas, o Agente Publicador deve, a partir dos seus conhecimentos sobre os dados, escolher os vocabulários que vai utilizar.

Para representação semântica de um conjunto de dados heterogêneos é possível utilizar uma ontologia já pré-existente que consiga contemplar a totalidade dos dados, desse modo a seleção de duas ou mais ontologias e/ou metadados se faz necessária. Dentre o levantamento de metadados e ontologias que possam ser utilizados para indexação semântica de dados de universidades públicas do estudo de caso, encontramos as ontologias *Global City Indicators Education Ontology* (GCIEO), *Lehigh University Benchmark* (LUBM) e *Academic Institution Internal Structure Ontology* (AIISO). Em comum, estes vocabulários apresentam uma especificação que visa descrever estruturas organizacionais das universidades, com seus departamentos, cursos, cursos e professores, que se adequa bem a alguns conjuntos de dados que queremos publicar, contudo, ainda estamos em busca dos voca-

bulários que visam descrever de forma mais específica o fluxo dos cursos e as ofertas de disciplinas. A solução para este problema foi a criação de uma vocabulário leve, que possa preencher as lacunas existentes nos dados que chamaremos de UnB *Vocabulary* (UVOC), cuja especificação está disponível no Apêndice A. Salientamos que este trabalho não tem por objetivo propor uma ontologia específica para as universidades, sendo uma solução paliativa até que um vocabulário mais completo seja incorporado. Para descrever os dados acadêmicos utilizamos então a LUBM e UVOC.

Também foram selecionados para este estudo de caso a ontologia FOAF para descrever dados relativos a pessoas, o conjunto de metadados *Dublin Core Metadata Element* (DC) para descrever recursos e o vocabulário VCGE já apresentando na Seção 5.3. A Tabela 5.8 apresenta quais vocabulários são utilizados na indexação semântica dos conjuntos de dados.

Tabela 5.8: Vocabulário Controlado dos Conjuntos De dados

	LUBM	**UVOC**	**FOAF**	**DC**	**VGCE**
Departamentos	X		X	X	X
Cursos	X		X		X
Disciplinas	X	X	X	X	X
Professores	X		X	X	X
Ofertas	X	X			X
Fluxo	X	X			X

5.5 Indexação Semântica

Como visto na Seção 4.3.3, a indexação semântica é realizada em duas etapas: inicialmente temos que selecionar um vocabulário que irá definir a característica do conjunto de dados, sendo que este será importante para identificar possíveis dados que possam ser conectados. Também é fundamental definir o URI do sujeito RDF das triplas. Na Tabela 5.9, é possível verificar quais foram os valores preenchidos para o tipo do conjunto de dados e enquanto na Tabela 5.10 é apresentada como foi criada a URI. No caso do URI, optamos em todos os conjuntos de dados, formá-los a partir de uma URL concatenada com um campo oriundo do próprio conjunto de dados.

A seguir é apresentado como foram preenchidas as informações dos conjuntos de dados para indexação semântica e integração:

- **Departamentos**: decidiu-se por publicar todos os campos, sendo que para os campos código, cod_departamento e nome_departamento tratados como objetos literais, utilizamos os termo dc:identifier e foaf:name;

Tabela 5.9: Tipo dos Conjuntos de Dados

Conjunto de dados	**Vocabulário**
Departamentos	lubm:Departament
Cursos	lubm:Course
Disciplina	uvoc:Disciplina
Professores	lubm:Professor
Ofertas	uvoc:oferta_disciplina
Fluxo	uvoc:fluxo_curso

Tabela 5.10: Formação dos URI dos Sujeitos

Conjunto de dados	**URI**
Departamentos	"http://dadosabertos.unb.br/departamentos/"+cod_departamento
Cursos	"http://dadosabertos.unb.br/cursos/"+cod_curso
Disciplina	"http://dadosabertos.unb.br/disciplina/"+cod_disciplina
Professores	"http://dadosabertos.unb.br/professores/"+matricula
Ofertas	"http://dadosabertos.unb.br/ofertadisicplinas/"+cod_oferta
Fluxo	"http://dadosabertos.unb.br/fluxo/"+cod_departamento

- **Cursos**: este conjunto de dados optou-se por descartar para publicação os campos sigla_departamento e nome_departamento, pois eles serão redundante a partir da integração dos dados. Os campos cod_curso e nome_curso utilizam os termos dc:identifier e foaf:name e são objetos literais. O campo cod_departamento é um objeto de ligação com o conjunto de dados de Departamento.

- **Professores**: este conjunto de dados tem seus dois campos publicados como objetos literais, sendo que matricula utiliza o termo uvoc:matricula e nome_professor o termo foaf:name.

- **Disciplinas**: no conjunto de dados de Disciplinas os objetos literais são cod_disciplina, nome_disciplina, créditos com os termos dc:identifier, foaf:name e uvoc:creditos. O campo cod_departamento fará ligação com o conjunto de dados de Departamento. Já os campos sigla_departamento e nome_departamento não serão publicados.

- **Ofertas**: os campos cod_oferta, turma, vagas_oferecidas e periodo serão literais com o termos dc:identifier, uvoc:turma, uvoc:vagas_oferecidas e uvoc:periodo_letivo. Os campos matricula_professor, cod_disciplina e cod_departamento são utilizados para integração com os conjuntos de dados e Professores, Disciplinas e Departamentos.

- **Fluxo**: Os campos nome_departamento e nome_curso não são publicados, sendo que sigla_departamento, cod_curso e cod_disciplina fazem a integração com os

conjuntos de dados de Departamentos, Cursos e Disciplinas. Os campos periodo e prioridade são objetos literais uvoc:periodo_ordem e uvoc:prioridade_disciplina.

A conexão dos dados é uma das funcionalidades mais importantes do UnBGOLD, desta maneira é possível garantir que os dados publicados não estejam disponíveis de maneira isolada propiciando, desta forma, a possibilidade de análise de dados através de técnicas mais sofisticadas. Para compreensão de como é realizada esta ação será analisado o exemplo de indexação do Conjunto de Dados de Oferta. A Figura 5.3 apresenta a tabela de indexação e as configurações realizadas neste exemplo.

Figura 5.3: Tabela de Indexação do Conjunto de Dados de Oferta de Disciplinas
Fonte: Elaboração Própria

As quatro primeiras linhas da tabela de indexação, destacadas em azul, selecionamos como literais, então o valor assumido pelo Sujeito será o dado explícito oriundo do conjunto de dados e a sua relação como o Sujeito será descrito pelo Predicado, que é um dos vocabulários escolhidos. As demais linhas, que estão destacadas em vermelho, não serão publicadas explicitamente, mas sim usadas de ponte para a ligação à outro conjunto de dados. Como exemplo, serão apresentados na Tabela 5.11 os dados referentes a uma linha do conjunto.

Todos os semestres são oferecidos pelos departamentos ofertas de vagas para que os alunos possam cursar as Disciplinas para concluírem os seus cursos, além deste conjunto de dados conter informações que descrevem a oferta (identificador oferta, turma, quantidade de vagas e período), também existem informações que estes dados se relacionam com outros conjuntos de dados. Neste caso já é identificada que disciplina é oferecida em várias oportunidades, então, uma das relações existentes seria com o Conjunto de Dados de Disciplinas. Há também o departamento que oferece a disciplina e também o professor que a ministra, define-se então que Oferta tem relação com esses três conjuntos:

Tabela 5.11: Exemplo de Dados do Conjunto de Oferta

Campo	**Valor**
cod_oferta	652174A
turma	A
vagas_oferecidas	40
periodo	20181
matricula_professor	149055
cod_disciplina	185221
cod_departamento	8591

Departamentos, Disciplinas e Professores. A análise dos dados mostra que os campos "cod_departamento", "matricula_professor", "cod_disciplina" são os valores que identificam os dados, assim, iremos utilizar estes dados para realizar a ligação.

A Figura 5.3 mostra que para o campo "matricula_professor" foi selecionado o vocabulário "lubm:Professor", no momento da seleção o UnBGOLD busca os demais conjuntos de dados que foram cadastrados onde o seu tipo semântico, ou o vocabulário selecionado para descrever o tipo do dados, seja o mesmo escolhido pelo Agente Publicador. Como apresentado na Tabela 5.9, o conjunto de Professores utiliza o mesmo vocabulário e assim habilita que o Agente Publicador possa selecionar este conjunto no campo Objeto e tentar realizar conexão, então o Agente Publicador tem a opção de manter este dados como um objeto Literal ou tentar estabelecer a ligação com os conjuntos de dados que são semanticamente relacionados através do vocabulário. Optando por realizar a ligação, a opção selecionada foi "Conjunto de Dados de Professores", e no campo "Complemento" da respectiva linha é habilitado uma lista dos nomes das colunas que o conjunto possui, na qual o Agente Publicador irá selecionar a responsável por identificar a ligação através da comparação entre os valores. Neste caso o Agente Publicador selecionou "matricula_professor" na opção complemento. Este processo funciona semelhante ao uso de chaves primaria e estrangeira em tabelas em bancos de dados. Esta configuração irá desencadear um processo no momento que começar a ação de publicação, ao chegar especificamente neste dados, o UnBGOLD tentará procurar um meio de ligar os dados comparando o valor dos dados. No exemplo temos que o valor referente a "matricula_professor' que está em oferta é "149055", neste caso ele irá procurar dentro da coluna do conjunto de Professores escolhido para ligação na tripla em que o valor seja exatamente igual, no caso "149055". A Figura 5.4 apresenta o resultado da ligação. Foi encontrado um recurso que possui diversas outras Triplas, então podemos dizer que existe um recurso do tipo (rdf:type) Professor, que o nome(foaf:name) é "LUIS FELIPE MIGUEL" e matrícula "149055" possui a relação de "Professor" com a Oferta.

Os demais campos de ligação seguem a mesma lógica já apresentada. Existe o campo

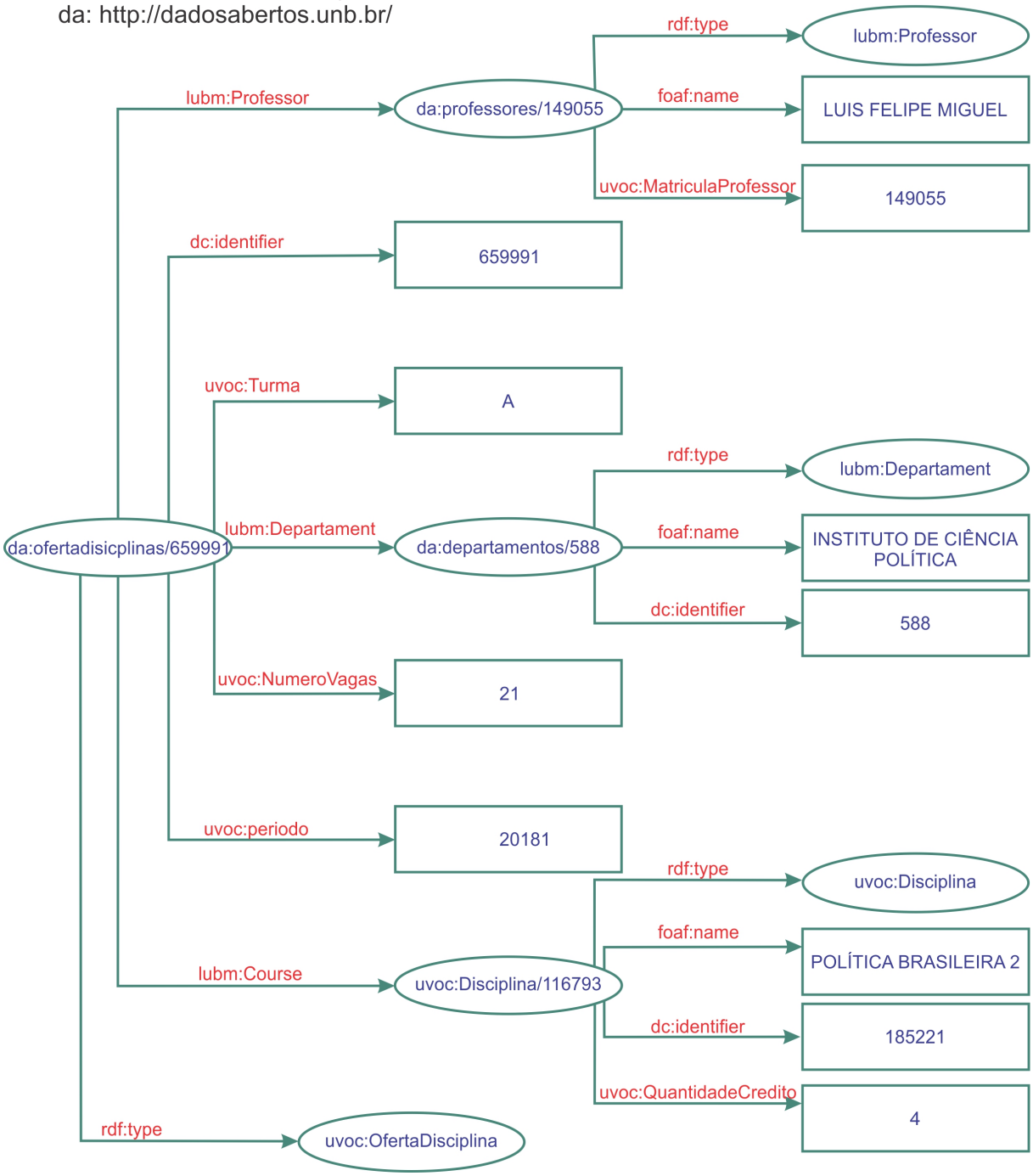


Figura 5.4: Grafos das Triplas do Conjunto de Dados de Ofertas e Ligações
Fonte: Elaboração Própria

"cod_departamento" que representa uma relação semântica com um conjunto de dados do tipo "lubm:Departament" e esta relação é realizada com o conjunto de Departamentos onde o campo "cod_departamento" que estabelece a ligação, assim, temos que neste campo da oferta o valor "588" que procura no campo "cod_departamento" no conjunto de dados se existe algum recurso que tenha este objeto literal. Já o campo "cod_disciplina" é pelo vocabulário "uvoc:Disciplina" e se relaciona com o conjunto de Disciplinas pelo seu campo "cod_disciplina". Na Figura 5.4 é possível ver claramente as ligações, nas quais os recursos são descritos em forma oval e os objetos literal em retângulos.

5.6 Publicação dos Dados

Há duas maneiras de gerar o arquivos RDF do conjunto de dados, manualmente, na qual o Agente Publicador solicita a geração e textitdownload do arquivo, e automatizada. Caso seja uma publicação automatizada, uma rotina de atualização verifica se existe alguma publicação pendente de ser publicada e em caso afirmativo, executa o processo de consulta, indexação e automaticamente pública na instância CKAN definida. Ao final da processo é disparado uma mensagem para o Agente Publicador com o relatório do processo de publicação.

5.7 Resultados

O produto gerado pelo processo é a publicação dos conjuntos de dados nos formatos indicados nas instâncias CKAN, sendo que, caso o Agente Publicador tenha configurado, publicará também os conjuntos de dados em formato indexado semanticamente RDF utilizando metadados e ontologias. Os conjuntos de dados podem ser publicados de forma integrada, onde será possível que os dados sejam conectados uns aos outros criando assim uma web de dados abertos. A conexão dos dados se dá identificando a relação entre recursos diferentes através de um dado em comum.

Para demonstrar como as triplas RDF foram criadas, vamos apresentar um exemplo de cada conjunto de dados. Nos Códigos 5.1, 5.2, 5.3, 5.4, 5.5 e 5.6 é possível verificar como os dados serão representados em formato RDF/XML, enquanto no Apêndice D são apresentadas a representação em grafos dos RDFs em que é possível visualizar com facilidade as triplas. Visto que cada conjunto de dados gera um número elevado de triplas, apresentamos apenas o resultado de um sujeito e seus objetos de cada conjunto de dados.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1#"
```

```
        xmlns:uvoc="http://dadosabertos.unb.br/ontologia#"
        xmlns:lubm="http://swat.cse.lehigh.edu/onto/univ-bench.owl#"
        xmlns:foaf="http://xmlns.com/foaf/0.1/#" >
  <lubm:Departament rdf:about="http://dadosabertos.unb.br/departamentos/108">
        <foaf:name>Departamento de Ciencia da Computacao</foaf:name>
        <dc:identifier>108</dc:identifier>
  </lubm:Departament>
</rdf:RDF>
```

Código 5.1: Código RDF/XML de Departamentos

,

```
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1#"
        xmlns:lubm="http://swat.cse.lehigh.edu/onto/univ-bench.owl#"
        xmlns:foaf="http://xmlns.com/foaf/0.1/#" >
  <lubm:Course rdf:about="http://dadosabertos.unb.br/cursos/370">
        <lubm:Department rdf:resource="http://dadosabertos.unb.br/departamentos/108"/>
        <foaf:name>Ciencia da Computacao</foaf:name>
        <dc:identifier>370</dc:identifier>
  </lubm:Course>
  <lubm:Course rdf:about="http://dadosabertos.unb.br/cursos/1341">
        <lubm:Department rdf:resource="http://dadosabertos.unb.br/departamentos/108"/>
        <foaf:name>Engenharia de Computacao</foaf:name>
        <dc:identifier>1341</dc:identifier>
  </lubm:Course>
</rdf:RDF>
```

Código 5.2: Código RDF/XML de Cursos

```
<rdf:RDF
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1#"
      xmlns:uvoc="http://dadosabertos.unb.br/ontologia#"
      xmlns:foaf="http://xmlns.com/foaf/0.1/#"
      xmlns:lubm="http://swat.cse.lehigh.edu/onto/univ-bench.owl#" >
  <lubm:Professor rdf:about="http://dadosabertos.unb.br/professores/1010590">
     <uvoc:MatriculaProfessor>1010590</uvoc:MatriculaProfessor>
     <foaf:name>JACIR LUIZ BORDIM</foaf:name>
  </lubm:Professor>
</rdf:RDF>
```

Código 5.3: Código RDF/XML de Professores

```
<rdf:RDF
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
```

```
      xmlns:dc="http://purl.org/dc/elements/1.1#"
      xmlns:uvoc="http://dadosabertos.unb.br/ontologia#"
      xmlns:lubm="http://swat.cse.lehigh.edu/onto/univ-bench.owl#"
      xmlns:foaf="http://xmlns.com/foaf/0.1/#" >
  <lubm:Course rdf:about="http://dadosabertos.unb.br/disciplinas/116793">
      <uvoc:QuantidadeCredito>4</uvoc:QuantidadeCredito>
      <foaf:name>INTRODUCAO A MICROINFORMATICA</foaf:name>
      <dc:identifier>116793</dc:identifier>
      <lubm:Department rdf:resource="http://dadosabertos.unb.br/departamentos/108"/>
  </lubm:Course>
</rdf:RDF>
```

Código 5.4: Código RDF/XML de Disciplina

```
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.org/dc/elements/1.1#"
     xmlns:uvoc="http://dadosabertos.unb.br/ontologia#"
     xmlns:lubm="http://swat.cse.lehigh.edu/onto/univ-bench.owl#">
  <uvoc:OfertaDisciplina rdf:about="http://dadosabertos.unb.br/ofertadisicplinas/659991"
    <uvoc:periodo>20181</uvoc:periodo>
    <uvoc:NumeroVagas>21</uvoc:NumeroVagas>
    <uvoc:Turma>A</uvoc:Turma>
    <lubm:Department rdf:resource="http://dadosabertos.unb.br/departamentos/108"/>
    <uvoc:Disciplina rdf:resource="http://dadosabertos.unb.br/disciplinas/116793"/>
    <lubm:Professor rdf:resource="http://dadosabertos.unb.br/professores/1103687"/>
    <dc:identifier>659991</dc:identifier>
  </uvoc:OfertaDisciplina>
</rdf:RDF>
```

Código 5.5: Código RDF/XML de Oferta de Disciplinas

```
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.org/dc/elements/1.1#"
     xmlns:uvoc="http://dadosabertos.unb.br/ontologia#"
     xmlns:lubm="http://swat.cse.lehigh.edu/onto/univ-bench.owl#">
  <uvoc:FluxoDisciplinas rdf:about="http://dadosabertos.unb.br/fluxo/396165221">
    <uvoc:PrioridadeDisciplinaFluxo>53</uvoc:PrioridadeDisciplinaFluxo>
    <uvoc:PeriodoDisciplinaFluxo>10</uvoc:PeriodoDisciplinaFluxo>
    <uvoc:Disciplina rdf:resource="http://dadosabertos.unb.br/disciplinas/165221"/>
    <lubm:Course rdf:resource="http://dadosabertos.unb.br/cursos/396"/>
    <lubm:Department rdf:resource="http://dadosabertos.unb.br/departamentos/140"/>
    <dc:identifier>396165221</dc:identifier>
  </uvoc:FluxoDisciplinas>
```

</rdf:RDF>

Código 5.6: Código RDF/XML de Fluxo

A Figura 5.5, que tem inspiração na Figura 3.4 do *Linked Open Data Project*, apresenta como os conjuntos de dados se conectam criando um contexto nos relacionamento das informações, assim, os dados adquirem um significado semântico entre eles. Em uma análise inicial, é possível afirmar que o conjunto de dados de Departamento tem um papel bastante destacado dentro deste contexto, no qual ele se conecta com a maioria dos demais conjuntos de dados.

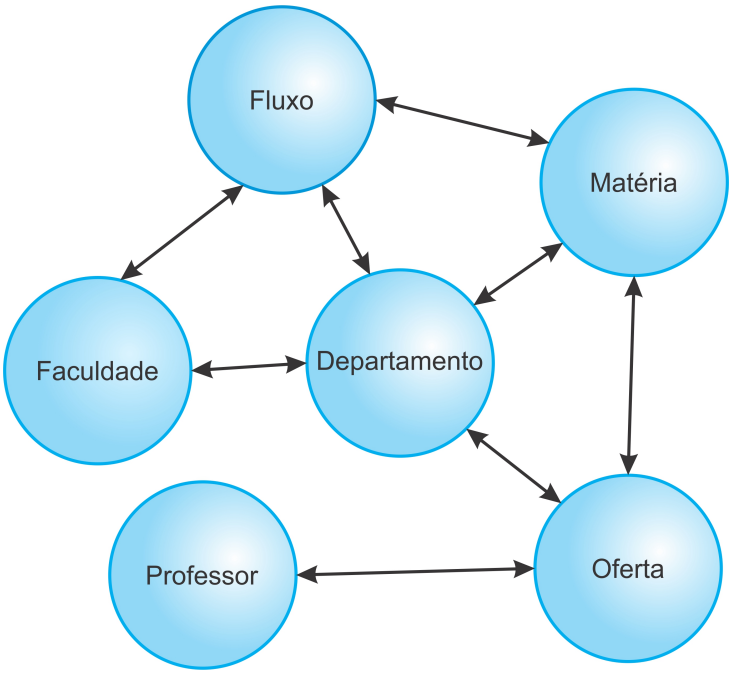


Figura 5.5: Diagrama de Rastreamento LOD-UnB
Fonte: Elaboração Própria

5.8 Catalogação e Busca Semântica

Toda vez que um conjunto de dados for publicado, será atualizado o Catálogo de Conjunto de Dados Abertos. Conforme a Seção 4.5, os dados serão descritos de forma semântica através de triplas RDF que são armazenadas no sistema gerenciador de banco de dados TDB. As triplas serão formadas pelo sujeito que identifica o conjunto de dados a que se refere a publicação. Os objetos e respectivos predicados serão gerados através do metadados existentes no sistema e de vocabulário já definido. O Código 5.7 apresenta o XML/RDF referente a catalogação do conjunto de dados de Oferta de Disciplinas.

<rdf:RDF

```
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dcmitype="http://purl.org/dc/dcmitype#"
   xmlns:vcge="http://vocab.e.gov.br/2011/03/vcge#"
   xmlns:dcam="http://purl.org/dc/dcam#"
   xmlns:dcterms="http://purl.org/dc/terms#"
   xmlns:dc="http://purl.org/dc/elements/1.1#"
   xmlns:foaf="http://xmlns.com/foaf/0.1/#">
<dcmitype:Dataset rdf:about="http://dados.unb.br/ofertaDisciplina">
  <dcterms:FileFormat>rdf</dcterms:FileFormat>
  <dc:subject>Departamanto</dc:subject>
  <dcterms:instructionalMethod>Dados gerados a partir da
    arquitetura de publicacao de dados
abertos da UnB</dcterms:instructionalMethod>
  <dc:identifier>7</dc:identifier>
  <dcterms:FileFormat>json</dcterms:FileFormat>
  <dcterms:language>pt</dcterms:language>
  <dc:description>Conjunto de dados de Oferta de Disciplinas</dc:description>
  <dcam:VocabularyEncodingScheme
    rdf:resource="http://swat.cse.lehigh.edu/onto/univ-bench.owl"/>
  <dcterms:Relation rdf:resource="http://dados.unb.br/disciplinas"/>
  <dc:subject>Aula</dc:subject>
  <dc:Title>Conjunto de dados de Oferta de Disciplinas</dc:Title>
  <dcmitype:Dataset rdf:resource="http://dados.unb.br/ofertaDisciplina"/>
  <dc:subject>Creditos</dc:subject>
  <dcterms:source rdf:resource="http://dados.unb.br/ofertaDisciplina"/>
  <dc:subject>Disciplina</dc:subject>
  <dcterms:FileFormat>csv</dcterms:FileFormat>
  <dcterms:Relation rdf:resource="http://dados.unb.br/professores"/>
  <dcam:VocabularyEncodingScheme
    rdf:resource="http://dadosabertos.unb.br/images/UnBVocabulario.owl"/>
  <foaf:Organization>Universidade de Brasilia</foaf:Organization>
  <dcterms:Relation rdf:resource="http://dados.unb.br/departamentos"/>
  <dcam:VocabularyEncodingScheme
       rdf:resource="http://vocab.e.gov.br/2011/03/vcge#educacao-superior"/>
  <dc:subject>Oferta</dc:subject>
  <dc:subject>Disciplina</dc:subject>
  <dcterms:created>2018-10-11 07:24:52.0</dcterms:created>
  <dc:date>2018-10-11 07:24:52.0</dc:date>
  <dc:type
   rdf:resource="http://dadosabertos.unb.br/images/UnBVocabulario.owl#
   oferta_disciplina"/>
  <dcterms:Frequency>Semestral</dcterms:Frequency>
</dcmitype:Dataset>
```

```
</rdf:RDF>
```

Código 5.7: Código RDF/XML do Cátalogo de Conjunto de Dados Abertos Conectados de Oferta de Disciplinas

O número de triplas geradas pelo conjunto de dados será variável de acordo com a quantidades de metadados que podem possuir mais de um valor, no Apêndice B é apresentado todo o código RDF/XML gerado pelos conjuntos de dados do nosso estudo de caso.

Após a criação do catálogo é possível realizar buscas pelos conjuntos de dados utilizando SPARQL. Não é objetivo desta funcionalidade realizar inferências semânticas mais complexas por não estar no escopo desta pesquisa. A busca é uma funcionalidade que é disponibilizada para os usuários finais que informam um termo que desejam pesquisar. Fizemos uma pesquisa sobre o termo "Graduação" e no Código 5.8 apresenta a consulta SPARQL que é executada no banco de Catálogo de Dados Abertos Conectados.

```
PREFIX dc: <http://purl.org/dc/elements/1.1#>
PREFIX dcterms: <http://purl.org/dc/terms#>
PREFIX dcam: <http://purl.org/dc/dcam#>
PREFIX dcmitype: <http://purl.org/dc/dcmitype#>
PREFIX vcge: <http://vocab.e.gov.br/2011/03/vcge#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/#>
SELECT  ?description
  ?title
        ?identifier
        ?source
        ?organization
        ?frequency
        ?frequency
        ?type
        ?date
        ?created
        ?language
        ?dataset
WHERE  { ?x dc:description ?description .
   ?x dc:Title ?title  .
         ?x dc:identifier ?identifier .
         ?x dcterms:source ?source .
         ?x foaf:Organization ?organization .
         ?x dcterms:Frequency ?frequency .
         ?x dc:type ?type .  ?x dc:date ?date .
         ?x dcterms:created ?created .
         ?x dcterms:language ?language .
         ?x dcmitype:Dataset ?dataset .
         ?x dcam:VocabularyEncodingScheme ?vocabulary .
```

```
    ?x dcterms:FileFormat ?fileFormat .
    ?x dc:subject ?subject .
    ?x dcterms:Relation ?relation .
    FILTER (regex(?title, "Graduacao", "i" ) ||
        regex(?description, "Graduacao", "i" ) ||
            regex(?identifier, "Graduacao", "i" ) ||
            regex(?source, "Graduacao", "i" ) ||
            regex(?organization, "Graduacao", "i" ) ||
            regex(?frequency, "Graduacao", "i" ) ||
            regex(?type, "Graduacao", "i" ) ||
            regex(?date, "Graduacao", "i" ) ||
            regex(?created, "Graduacao", "i" ) ||
            regex(?date, "Graduacao", "i" ) ||
            regex(?language, "Graduacao", "i" ) ||
            regex(?Dataset, "Graduacao", "i" ) ||
            regex(?vocabulary, "Graduacao", "i" ) ||
            regex(?fileFormat, "Graduacao", "i" ) ||
            regex(?subject, "Graduacao", "i" ) ||
            regex(?relation, "Graduacao", "i" )
        ) .
}
```

Código 5.8: Código SPARQL Para Consulta

Após a execução da busca o usuário irá receber uma lista com os conjuntos de dados que possuem conteúdo pesquisado em duas informações. Essa busca não visa fazer uma aprofundamento nas ligações das recursos, trazendo informações apenas nos recursos identificados no catálogo de conjuntos de dados. Na Figura 5.6 é apresentada a lista com as respostas. O usuário verá as informações simplificadas do conjunto de dados, mas ao clicar no item da lista será apresentada a versão completa sobre os dados que é demonstrada na Figura 5.7.

Ao solicitar o detalhamento do conjunto de dados o usuário visualiza todas as informações que existem no conjunto de dados que informamos na Seção 4.6, sendo que as informações com valores unitários são apresentadas no lado esquerdo da tela. Já o lado direito é dividido em caixas de informações, onde na primeira caixa denominada "Download" o usuário poderá baixar o conjunto de dados em um dos formatos disponíveis (definido pelo Agente Publicador) que estão hospedados na instância CKAN definida pela publicação. Na próxima caixa "Conjuntos de Dados" são apresentados outros Conjuntos de Dados que possuem conexão com o conjunto de dados que está sendo detalhado, sendo possível detalhá-lo também clicando sobre o nome do conjunto de dados. Na caixa "Marcadores" são listadas as *tags* utilizadas na marcação do conjunto de dados, onde é possível pesquisar outros conjuntos de dados que detêm o mesmo marcador. A última

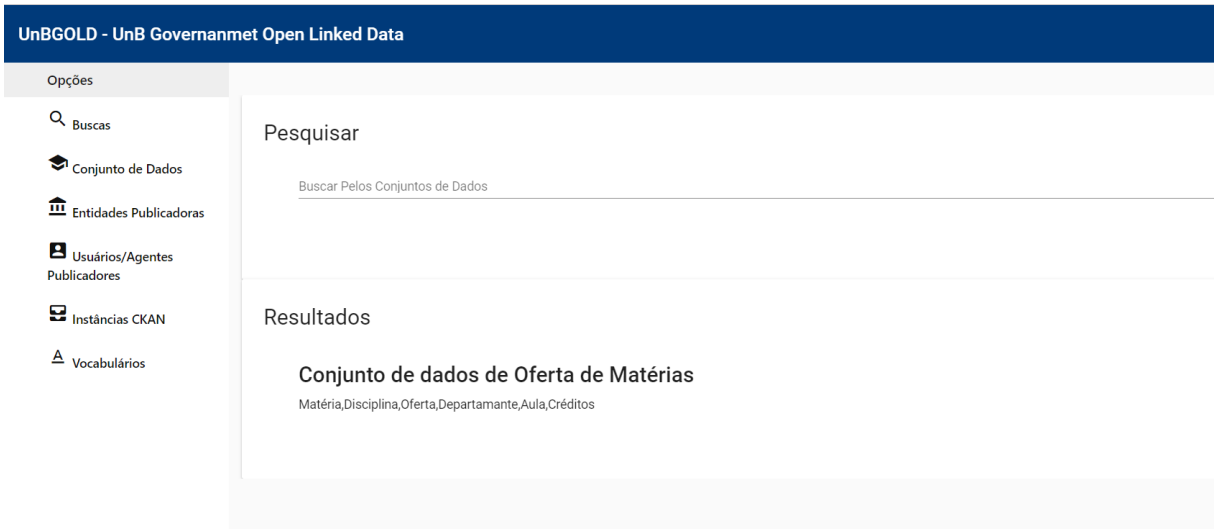


Figura 5.6: Interface de Busca Semântica
Fonte: Elaboração Própria

caixa de informação é a caixa “Vocabulário”, na qual consta a informação de quais foram os vocabulários usados na indexação semântica dos dados.

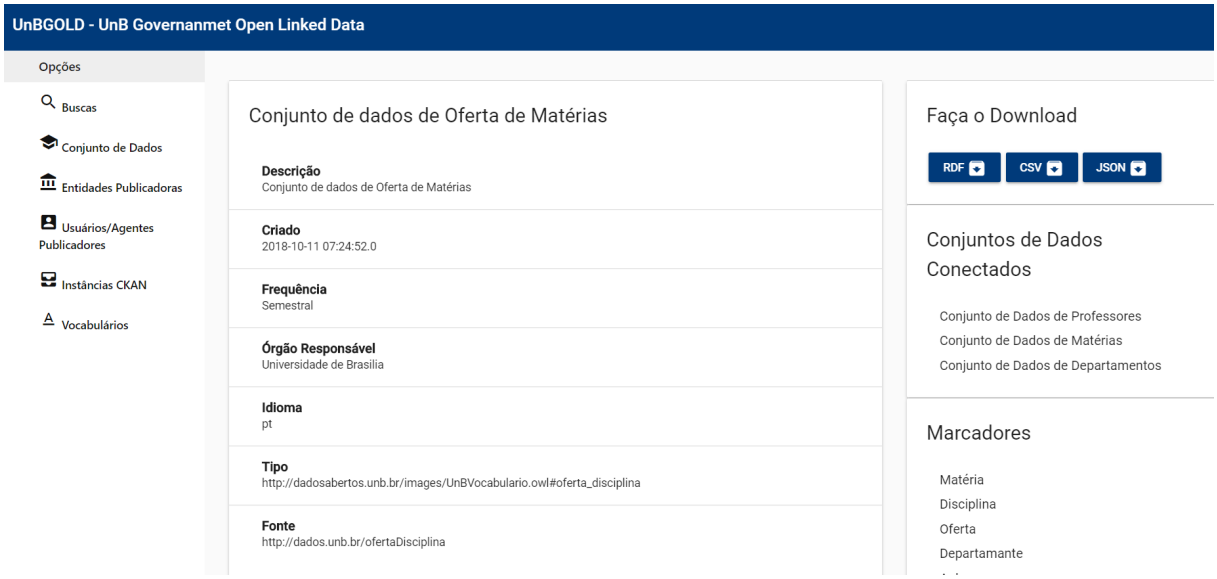


Figura 5.7: Detalhamento do Conjunto de Dados
Fonte: Elaboração Própria

Capítulo 6

Conclusão

A publicação de dados abertos nas instituições públicas brasileiras é uma realidade ao qual os órgãos devem adaptar-se sendo que não existe processo amplamente definido para que isso ocorra. A proposta apresentada neste trabalho busca contemplar a necessidade da UnB, estabelecendo o emprego de diversas tecnologias que buscam melhorar a atividade de publicação de dados ao mesmo tempo que busca aumentar a qualidade dos dados publicados através da indexação semântica, da atualização contínua dos dados por meio da automação da publicação e a da catalogação das fontes de dados descritas de forma semântica.

Neste passo, esta arquitetura é dividida em 3 etapas: 1) Extração dos Dados, 2) Automação e Indexação e 3) Publicação dos Dados. Para a etapa de extração dos dados, a escolha dos dados a serem publicados deve respeitar o PDA da instituição, sendo que, no caso da UnB, a curadoria, limpeza e disponibilização é feita através da combinação do armazenamento em um *Data Warehouse* e o acesso através do barramento de serviço ErlangMS. Para automação e indexação semântica foi desenvolvida a ferramenta UnB-GOLD, que busca os dados no serviço disponibilizado, utiliza um vocabulário controlado para indexar semanticamente os dados e gerencia uma interface para publicação automatizada. A última etapa é a publicação que é realizada na plataforma CKAN, desenvolvida especificamente como solução para publicação de dados abertos. Para realizar a integração entre as camadas, foi incorporada uma biblioteca que realiza a comunicação com a API do CKAN e possibilita o gerenciamento externo dos conjuntos de dados, proporcionando que seja realizada a publicação automatizada.

Com a automatização, é possível que os dados possam ser publicados em intervalos menores, garantindo que o dado seja o mais atual possível e ao passo que a indexação semântica viabiliza que os dados não sejam publicados isoladamente, oportunizando análises mais sofisticadas por meio da integração com outros conjuntos de dados. Com a indexação semântica das informação referente aos conjuntos de dados foi possível desenvolver

uma interface de pesquisa utilizando SPARQL em que os resultados são enriquecidos semanticamente, além de possibilitar que os usuários possam visualizar as conexões entre conjuntos de dados diferentes.

6.1 Trabalhos Futuros

Foi identificado também que para publicação de dados acadêmicos de nível superior não existe um vocabulário que representa semanticamente os diversos tipos de dados, visto que o domínio da informação é bastante amplo e que as ontologias encontradas buscam descrever principalmente os dados das estruturas organizacionais das universidades. Ao passo que existe um domínio maior de informações que são de interesse público não contemplados nas ontologias existentes, sendo necessário, futuramente, estabelecer um vocabulário padrão que vise uniformizar a publicação dos dados em diversas outras instituições de ensino superior proporcionando a melhor integração dos mesmo.

A interface de pesquisa apresentada neste trabalho não teve como objetivo realizar inferência semântica diretamente nos conjuntos de dados, mas apenas no Catálogo de Conjuntos de Abertos Abertos Conectados, assim, é possível desenvolver uma ferramenta de busca mais sofisticada onde seria utilizada inferência semântica para recuperação das informações.

6.2 Contribuições

Esta pesquisa teve as seguintes contribuições:

- **Acadêmica**: Foi disponibilização de um processo e de uma arquitetura que facilita de publicação de dados abertos descritos semanticamente;
- **Institucional**: A UnB terá seus dados abertos publicados com mais qualidade e automaticamente;
- **Tecnológica**: Implementado o software UnBGOLD que realiza a indexação de dados abertos tornando-os conectados além e automatizar a publicação na plataforma CKAN.

Referências

[1] Everton Agilar, Rodrigo Almeida, and Edna Canedo. A systematic mapping study on legacy system modernization. In *SEKE*, 2016. 30, 31

[2] CKAN Association. API guide, 2013. URL `https://docs.ckan.org/en/ckan-2.7.3/api/`. [Online; acessado em 10 de outubro de 2018]. 40

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBPedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007. 22

[4] Tim Berners-Lee. Desing Issues - Linked Data, July 2006. URL `https://www.w3.org/DesignIssues/LinkedData.html`. 6, 8, 20, 26

[5] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284:28–37, 2001. 5, 6

[6] Christian Bizer and Richard Cyganiak. Rdf 1.1 TriG-RDF dataset language-w3c recommendation, 2014. URL `https://www.w3.org/RDF/`. [Online; acessado em 20 de setembro de 2017]. 11

[7] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, pages 1265–1266. ACM, 2008. 19

[8] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227, 2009. 2, 7

[9] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBPedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165, 2009. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2009.07.002. URL `http://www.sciencedirect.com/science/article/pii/S1570826809000225`. The Web of Data. 24

[10] Willem Nico Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Universiteit Twente, 1997. 10

[11] Brasil. CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL DE 1988, October 1988. URL `http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm`. [Online; acessado em 20 de setembro de 2017]. 1

[12] Brasil. LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011., November 200. URL http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. [Online; acessado em 20 de setembro de 2017]. 1, 44

[13] Brasil. Decreto nº 8.777, de 11 de maio de 2016, May 2016. URL http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm. [Online; acessado em 20 de setembro de 2017]. 2, 29, 44

[14] Dan Brickley, R.V. Guha, and Brian McBride. RDF schema 1.1. URL https://www.w3.org/TR/rdf-schema/. [Online; acessado em 02 de outubro de 2018]. 11

[15] Anja; Abele Andrejs; McCrae John Cyganiak, Richard; Jentzsch. The linking open data cloud diagram, 2007. URL http://lod-cloud.net/. [Online; acessado em 10 de setembro de 2017]. 19

[16] Secretaria de Logística e Tecnologia da Informação. *Vocabulário Controlado de Governo Eletrônico*. Ministério do Planejamento, Orçamento e Gestão, https://www.governodigital.gov.br/transformacao/orientacoes/interoperabilidade/vocabulario-controlado-do-governo-eletronico-vcge, 2.1.0 edition, 2016. [Online; acessado em 05 de novembro de 2017]. 36

[17] Everton de Vargas Agilar. *Uma Abordagem Orientada a Serviços para a Modernização de Sistemas Legados*. PhD thesis, Universidade de Brasília, 2016. 14, 15, 49

[18] Metadata Iniciative Dublin Core. Metadata basics. URL http://dublincore.org/metadata-basics/. [Online; acessado em 20 de setembro de 2017]. 10

[19] Karl Dubost and Ivan Herman. State of the Semantic Web, 2008. URL https://www.w3.org/2008/Talks/0307-Tokyo-IH/Slides.pdf. [Online; acessado em 20 de setembro de 2017]. 6

[20] David Eaves. The Three Laws of Open Government Data, 2009. URL http://eaves.ca/2009/09/30/three-law-of-open-government-data/. [Online; acessado em 20 de setembro de 2017]. 13

[21] Thomas R Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5:199–220, 1993. 10

[22] Infraestrutura Nacional de Dados Abertos INDA. Perguntas mais frequentes, 2011. URL http://dados.gov.br/pagina/faq. [Online; acessado em 20 de setembro de 2017]. 22

[23] Dublin Core Metadata Initiative. Dcmi specifications, 1995-2018. URL http://dublincore.org/specifications/. [Online; acessado em 08 de outubro de 2017]. 41

[24] Open Knowledge International. What is Open Data? URL https://okfn.org/opendata/. [Online; acessado em 02 de agosto de 2017]. 12

[25] Seiji Isotani and IgIbert Bittencourt. *Dados Abertos Conectados: Em Busca da Web do Conhecimento*. Novatec Editora, 2015. 6, 7, 11

[26] Ralph Kimball. *The Data Warehouse Lifecycle Toolkit*. John Wiley & Sons, 2008. 10

[27] Carl Malamud, Tim O'Reilly, Greg Elin, et al. 8 Principles of Open Government Data, December 2007. URL `https://public.resource.org/8_principles.html`. [Online; acessado em 26 de julho de 2017]. 12, 26

[28] Riichiro Mizoguchi. Part 3: Advanced course of ontological engineering. In *Tutorial on Ontological Engineering*, volume 22, pages 193–220. New Generation Computing, 2004. 9

[29] Open Knowledge Fundation OKF. The Open Definition, 2015. URL `http://opendefinition.org/`. [Online; acessado em 26 de julho de 2017]. 12

[30] Durval Vieira Pereira and Carlos Henrique Marcondes. Modelagem e Representação Semântica de Dados Governamentais Abertos da Previdência Social Brasileira. Belo Horizonte - MG, October 2014. 2

[31] Eric Prud'hommeaux and Andy Seaborne. Sparql query language for rdf, 2018. URL `https://www.w3.org/TR/rdf-sparql-query/`. [Online; acessado em 08 de outubro de 2018]. 43

[32] Jose Eduardo Santarem Segundo. Tecnologías de la información y la comunicación para proporcionar datos abiertos en formato semántico. *Ibersid*, 7:33–40, 2013. ISSN 1888-0967. 1

[33] Marcus Oliveira Silva, Rommel Novaes Carvalho, Marcelo Ladeira, Henrique A. da Rocha, and Gilson Libório Mende. Unb-lod, a visual tool to work with linked open data. *LOD BRASIL Linked Open Data*, pages 43–53, 2014. 17, 18, 19

[34] OpenLink Software. Openlink virtuoso universal server documentation, 2018. URL `http://docs.openlinksw.com/virtuoso/`. [Online; acessado em 05 de novembro de 2017]. 24

[35] Terezinha Batista de Souza, Maria Elizabete Catarino, and Paulo Cesar dos Santos. Metadados: Catalogando dados na internet. *Transinformação-ISSNe 2318-0889*, 9 (2), 2012. 11

[36] Rudi Studer, V Richard Benjamins, and Dieter Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197, 1998. 10

[37] Arlene G Taylor. The organization or information. 2004. 10

[38] Sammohan; Holanda Maristela; Ishikawa Edson; Oliveira Edgard Victorino, Marcio; Chhetri. Transforming open data to linked open data using ontologies for information organization in big data environments of the brazilian government. *ISKO UK biennial conference 11th ? 12th*, 2017. 20, 21, 22, 23

[39] World Wide Web Consortium W3C et al. Best practices for publishing linked data, January 2014. URL https://www.w3.org/TR/ld-bp/. [Online; acessado em 22 de julho de 2017]. 7

Apêndice A

UnB Vocabulário

```
<?xml version="1.0" encoding="UTF-8"?>
<Ontology xmlns="http://www.w3.org/2002/07/owl#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xml:base="http://dadosabertos.unb.br/images/UnBVocabulario.owl"
ontologyIRI="http://dadosabertos.unb.br/images/UnBVocabulario.owl">
    <Prefix name="" IRI="http://dadosabertos.unb.br/images/UnBVocabulario.owl" />
    <Prefix name="cc" IRI="http://creativecommons.org/ns#" />
    <Prefix name="dc" IRI="http://purl.org/dc/terms/" />
    <Prefix name="uvoc" IRI="http://dadosabertos.unb.br/images/UnBVocabulario.owl" />
    <Prefix name="owl" IRI="http://www.w3.org/2002/07/owl#" />
    <Prefix name="rdf" IRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#" />
    <Prefix name="xml" IRI="http://www.w3.org/XML/1998/namespace" />
    <Prefix name="adms" IRI="http://www.w3.org/ns/adms#" />
    <Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#" />
    <Prefix name="vann" IRI="http://purl.org/vocab/vann/" />
    <Annotation>
        <AnnotationProperty abbreviatedIRI="adms:relatedDocumentation" />
        <Literal xml:lang="pt-br"
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">http://dad
        osabertos.unb.br/images/UnBVocabulario.pdf</Literal>
    </Annotation>
    <Annotation>
        <AnnotationProperty abbreviatedIRI="vann:preferredNamespaceUri" />
        <Literal
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">http://dad
        osabertos.unb.br/images/UnBVocabulario.owl#</Literal>
    </Annotation>
    <Annotation>
        <AnnotationProperty abbreviatedIRI="vann:preferredNamespacePrefix" />
        <Literal
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">#</Literal
        >
    </Annotation>
    <Annotation>
        <AnnotationProperty abbreviatedIRI="dc:description" />
        <Literal xml:lang="en"
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">Este é
        ontologia leve desenvolvida para dar suporte para publicação de dados abertos
        conectados em instituições de ensino superior públicas brasileiras</Literal>
    </Annotation>
    <Annotation>
        <AnnotationProperty abbreviatedIRI="dc:title" />
        <Literal xml:lang="en"
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">UnB
        Vocabulário</Literal>
    </Annotation>
    <Annotation>
        <AnnotationProperty abbreviatedIRI="dc:creator" />
        <Literal xml:lang="en"
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">Luiz C B
        Martins (luizmartins@unb.br)</Literal>
    </Annotation>
    <Annotation>
        <AnnotationProperty abbreviatedIRI="cc:license" />
        <Literal
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">#</Literal
        >
    </Annotation>
    <Annotation>
        <AnnotationProperty abbreviatedIRI="rdfs:label" />
        <Literal xml:lang="en"
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">UnB
        Vocabulário</Literal>
    </Annotation>
    <Annotation>
        <AnnotationProperty abbreviatedIRI="owl:versionInfo" />
        <Literal xml:lang="en"
        datatypeIRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral">0.1 22
        Out 2018</Literal>
    </Annotation>
```

```
49      <Declaration>
50         <Class IRI="#Disciplina" />
51      </Declaration>
52      <Declaration>
53         <Class IRI="#periodo" />
54      </Declaration>
55      <Declaration>
56         <Class IRI="#quantidade_credito" />
57      </Declaration>
58      <Declaration>
59         <Class IRI="#numero_vagas" />
60      </Declaration>
61      <Declaration>
62         <Class IRI="#oferta_disciplina" />
63      </Declaration>
64      <Declaration>
65         <Class IRI="#turma" />
66      </Declaration>
67      <Declaration>
68         <Class IRI="#fluxo_materias" />
69      </Declaration>
70      <Declaration>
71         <Class IRI="#periodo_materia_fluxo" />
72      </Declaration>
73      <Declaration>
74         <Class IRI="#prioridade_materia_fluxo" />
75      </Declaration>
76   </Ontology>
```

Apêndice B

Catálogo de Dados Abertos Conectados

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dcmitype="http://purl.org/dc/dcmitype#"
    xmlns:vcge="http://vocab.e.gov.br/2011/03/vcge#"
    xmlns:dcam="http://purl.org/dc/dcam#"
    xmlns:dcterms="http://purl.org/dc/terms#"
    xmlns:dc="http://purl.org/dc/elements/1.1#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/#">
  <dcmitype:Dataset rdf:about="http://dados.unb.br/materias">
    <dc:subject>Conteúdo</dc:subject>
    <dc:Title>Conjunto de dados de Matérias</dc:Title>
    <dcam:VocabularyEncodingScheme rdf:resource="http://xmlns.com/foaf/0.1/"/>
    <dc:date>2018-10-09 17:24:09.0</dc:date>
    <dcterms:created>2018-10-09 17:24:09.0</dcterms:created>
    <dcterms:Relation>
      <dcmitype:Dataset rdf:about="http://dados.unb.br/departamentos">
        <dc:date>2018-09-15 11:15:23.0</dc:date>
        <dc:description>Conjunto de dados de departamento</dc:description>
        <dcterms:source rdf:resource="http://dados.unb.br/departamentos"/>
        <dc:subject>Graduação</dc:subject>
        <dc:identifier>1</dc:identifier>
        <dcterms:FileFormat>rdf</dcterms:FileFormat>
        <dcterms:instructionalMethod>Dados gerados a partir da arquitetura de
        publicação de dados abertos da UnB</dcterms:instructionalMethod>
        <foaf:Organization>Universidade de Brasilia</foaf:Organization>
        <dc:subject>Ensino Superior</dc:subject>
        <dc:subject>Universidade</dc:subject>
        <dcterms:Relation>
          <dcmitype:Dataset rdf:about="http://dados.unb.br/faculdades">
            <dcterms:FileFormat>rdf</dcterms:FileFormat>
            <dc:date>2018-09-25 20:15:15.0</dc:date>
            <dc:subject>Universidade</dc:subject>
            <dc:description>Conjunto de Dados de Faculdades</dc:description>
            <dc:Title>Conjunto de Dados de Faculdades</dc:Title>
            <dcam:VocabularyEncodingScheme
            rdf:resource="http://swat.cse.lehigh.edu/onto/univ-bench.owl"/>
            <dcterms:source rdf:resource="http://dados.unb.br/faculdades"/>
            <dcterms:Relation rdf:resource="http://dados.unb.br/departamentos"/>
            <dcterms:language>pt</dcterms:language>
            <dc:subject>Cursos</dc:subject>
            <dcterms:Frequency>Semanal</dcterms:Frequency>
            <dcam:VocabularyEncodingScheme
            rdf:resource="http://vocab.e.gov.br/2011/03/vcge#educacao-superior"/>
            <dcam:VocabularyEncodingScheme
            rdf:resource="http://purl.org/dc/elements/1.1"/>
            <dc:subject>Graduação</dc:subject>
            <dcterms:Relation>
              <dcmitype:Dataset rdf:about="http://dados.unb.br/fluxo">
                <dc:Title>Conjunto de dados de Fluxo de Matérias</dc:Title>
                <dc:identifier>8</dc:identifier>
                <dc:type
                rdf:resource="http://dadosabertos.unb.br/images/UnBVocabulario.owl#flu
                xo_materias"/>
                <dcam:VocabularyEncodingScheme rdf:resource="http://unb.br/literal"/>
                <dc:subject>Fluxo</dc:subject>
                <dcterms:FileFormat>json</dcterms:FileFormat>
                <dcterms:source rdf:resource="http://dados.unb.br/fluxo"/>
                <foaf:Organization>Universidade de Brasilia</foaf:Organization>
                <dc:description>Conjunto de dados de Fluxo de
                Matérias</dc:description>
                <dc:subject>Faculdade</dc:subject>
                <dcterms:language>pt</dcterms:language>
                <dcterms:Relation rdf:resource="http://dados.unb.br/materias"/>
                <dcterms:Relation rdf:resource="http://dados.unb.br/departamentos"/>
                <dc:date>2018-10-16 12:48:06.0</dc:date>
                <dcam:VocabularyEncodingScheme
                rdf:resource="http://vocab.e.gov.br/2011/03/vcge#educacao-superior"/>
                <dcam:VocabularyEncodingScheme
                rdf:resource="http://dadosabertos.unb.br/images/UnBVocabulario.owl"/>
                <dcmitype:Dataset rdf:resource="http://dados.unb.br/fluxo"/>
                <dc:subject>Curso</dc:subject>
                <dcterms:FileFormat>csv</dcterms:FileFormat>
```

```
                    <dcterms:created>2018-10-16 12:48:06.0</dcterms:created>
                    <dcterms:instructionalMethod>Dados gerados a partir da arquitetura
                    de publicação de dados abertos da UnB</dcterms:instructionalMethod>
                    <dcterms:Frequency>Semestral</dcterms:Frequency>
                    <dcam:VocabularyEncodingScheme
                    rdf:resource="http://swat.cse.lehigh.edu/onto/univ-bench.owl"/>
                    <dcterms:FileFormat>rdf</dcterms:FileFormat>
                    <dcterms:Relation rdf:resource="http://dados.unb.br/faculdades"/>
                  </dcmitype:Dataset>
                </dcterms:Relation>
                <dc:identifier>2</dc:identifier>
                <dc:subject>Ensino Superior</dc:subject>
                <foaf:Organization>Universidade de Brasilia</foaf:Organization>
                <dcmitype:Dataset rdf:resource="http://dados.unb.br/faculdades"/>
                <dc:type rdf:resource="http://purl.org/vocab/aiiso/schema#Faculty"/>
                <dc:subject>Faculdade</dc:subject>
                <dcterms:FileFormat>json</dcterms:FileFormat>
                <dcterms:FileFormat>csv</dcterms:FileFormat>
                <dcterms:instructionalMethod>Dados gerados a partir da arquitetura de
                publicação de dados abertos da UnB</dcterms:instructionalMethod>
                <dcterms:created>2018-09-25 20:15:15.0</dcterms:created>
                <dcam:VocabularyEncodingScheme rdf:resource="http://xmlns.com/foaf/0.1/"/>
              </dcmitype:Dataset>
            </dcterms:Relation>
            <dc:subject>Departamento</dc:subject>
            <dcterms:Relation rdf:resource="http://dados.unb.br/fluxo"/>
            <dcam:VocabularyEncodingScheme
            rdf:resource="http://vocab.e.gov.br/2011/03/vcge#educacao-superior"/>
            <dcterms:FileFormat>json</dcterms:FileFormat>
            <dc:subject>Faculdade</dc:subject>
            <dcterms:Relation rdf:resource="http://dados.unb.br/materias"/>
            <dcmitype:Dataset rdf:resource="http://dados.unb.br/departamentos"/>
            <dcterms:Relation>
              <dcmitype:Dataset rdf:about="http://dados.unb.br/ofertaDisciplina">
                <dcterms:FileFormat>rdf</dcterms:FileFormat>
                <dc:subject>Departamante</dc:subject>
                <dcterms:instructionalMethod>Dados gerados a partir da arquitetura de
                publicação de dados abertos da UnB</dcterms:instructionalMethod>
                <dc:identifier>7</dc:identifier>
                <dcterms:FileFormat>json</dcterms:FileFormat>
                <dcterms:language>pt</dcterms:language>
                <dc:description>Conjunto de dados de Oferta de Matérias</dc:description>
                <dcam:VocabularyEncodingScheme
                rdf:resource="http://swat.cse.lehigh.edu/onto/univ-bench.owl"/>
                <dcterms:Relation rdf:resource="http://dados.unb.br/materias"/>
                <dc:subject>Aula</dc:subject>
                <dc:Title>Conjunto de dados de Oferta de Matérias</dc:Title>
                <dcmitype:Dataset rdf:resource="http://dados.unb.br/ofertaDisciplina"/>
                <dc:subject>Créditos</dc:subject>
                <dcterms:source rdf:resource="http://dados.unb.br/ofertaDisciplina"/>
                <dc:subject>Disciplina</dc:subject>
                <dcterms:FileFormat>csv</dcterms:FileFormat>
                <dcterms:Relation>
                  <dcmitype:Dataset rdf:about="http://dados.unb.br/professores">
                    <dc:identifier>6</dc:identifier>
                    <dc:subject>Faculdade</dc:subject>
                    <dcam:VocabularyEncodingScheme
                    rdf:resource="http://dadosabertos.unb.br/images/UnBVocabulario.owl"/>
                    <dc:description>Conjunto de dados de professores</dc:description>
                    <dc:subject>Professor</dc:subject>
                    <dcmitype:Dataset rdf:resource="http://dados.unb.br/professores"/>
                    <dc:subject>Docente</dc:subject>
                    <dcterms:FileFormat>rdf</dcterms:FileFormat>
                    <dc:Title>Conjunto de dados de professores</dc:Title>
                    <dcterms:instructionalMethod>Dados gerados a partir da arquitetura
                    de publicação de dados abertos da UnB</dcterms:instructionalMethod>
                    <dc:type
                    rdf:resource="http://swat.cse.lehigh.edu/onto/univ-bench.owl#Professor
                    "/>
                    <dcterms:Relation
                    rdf:resource="http://dados.unb.br/ofertaDisciplina"/>
                    <dcterms:language>pt</dcterms:language>
```

```
                    <dcterms:created>2018-10-10 14:39:44.0</dcterms:created>
                    <dcterms:FileFormat>json</dcterms:FileFormat>
                    <dcterms:source rdf:resource="http://dados.unb.br/professores"/>
                    <dcam:VocabularyEncodingScheme
                    rdf:resource="http://xmlns.com/foaf/0.1/"/>
                    <dcam:VocabularyEncodingScheme
                    rdf:resource="http://vocab.e.gov.br/2011/03/vcge#educacao-superior"/>
                    <foaf:Organization>Universidade de Brasilia</foaf:Organization>
                    <dc:date>2018-10-10 14:39:44.0</dc:date>
                    <dcterms:Frequency>Semestral</dcterms:Frequency>
                    <dc:subject>Universidade</dc:subject>
                    <dc:subject>Graduação</dc:subject>
                    <dcterms:FileFormat>csv</dcterms:FileFormat>
                    <dc:subject>Ensino Superior</dc:subject>
                  </dcmitype:Dataset>
                </dcterms:Relation>
                <dcam:VocabularyEncodingScheme
                rdf:resource="http://dadosabertos.unb.br/images/UnBVocabulario.owl"/>
                <foaf:Organization>Universidade de Brasilia</foaf:Organization>
                <dcterms:Relation rdf:resource="http://dados.unb.br/departamentos"/>
                <dcam:VocabularyEncodingScheme
                rdf:resource="http://vocab.e.gov.br/2011/03/vcge#educacao-superior"/>
                <dc:subject>Oferta</dc:subject>
                <dc:subject>Matéria</dc:subject>
                <dcterms:created>2018-10-11 07:24:52.0</dcterms:created>
                <dc:date>2018-10-11 07:24:52.0</dc:date>
                <dc:type
                rdf:resource="http://dadosabertos.unb.br/images/UnBVocabulario.owl#oferta_
                disciplina"/>
                <dcterms:Frequency>Semestral</dcterms:Frequency>
              </dcmitype:Dataset>
            </dcterms:Relation>
            <dc:type rdf:resource="http://purl.org/vocab/aiiso/schema#Department"/>
            <dcterms:language>pt</dcterms:language>
            <dcterms:Frequency>Semestral</dcterms:Frequency>
            <dcam:VocabularyEncodingScheme rdf:resource="http://xmlns.com/foaf/0.1/"/>
            <dcterms:FileFormat>csv</dcterms:FileFormat>
            <dc:subject>Instituto</dc:subject>
            <dcam:VocabularyEncodingScheme
            rdf:resource="http://swat.cse.lehigh.edu/onto/univ-bench.owl"/>
            <dc:Title>Conjunto de dados de departamento</dc:Title>
            <dcterms:created>2018-09-15 11:15:23.0</dcterms:created>
            <dcam:VocabularyEncodingScheme
            rdf:resource="http://purl.org/dc/elements/1.1"/>
            <dc:subject>Pós-Graduação</dc:subject>
          </dcmitype:Dataset>
        </dcterms:Relation>
        <dcam:VocabularyEncodingScheme rdf:resource="http://purl.org/dc/elements/1.1"/>
        <dc:subject>Ensino Superior</dc:subject>
        <dcterms:Frequency>Semestral</dcterms:Frequency>
        <dc:identifier>4</dc:identifier>
        <dcam:VocabularyEncodingScheme
        rdf:resource="http://dadosabertos.unb.br/images/UnBVocabulario.owl"/>
        <dc:subject>Curso</dc:subject>
        <dcterms:FileFormat>csv</dcterms:FileFormat>
        <dc:subject>Matéria</dc:subject>
        <dcterms:Relation rdf:resource="http://dados.unb.br/ofertaDisciplina"/>
        <dc:subject>Disciplina</dc:subject>
        <dcterms:Relation rdf:resource="http://dados.unb.br/fluxo"/>
        <dcam:VocabularyEncodingScheme
        rdf:resource="http://vocab.e.gov.br/2011/03/vcge#educacao-superior"/>
        <dcterms:language>pt</dcterms:language>
        <dc:type rdf:resource="http://purl.org/vocab/aiiso/schema#Course"/>
        <dc:description>Conjunto de dados de Matérias</dc:description>
        <dcterms:FileFormat>rdf</dcterms:FileFormat>
        <dcterms:FileFormat>json</dcterms:FileFormat>
        <dcterms:source rdf:resource="http://dados.unb.br/materias"/>
        <dc:subject>Universidade</dc:subject>
        <dcmitype:Dataset rdf:resource="http://dados.unb.br/materias"/>
        <dcam:VocabularyEncodingScheme rdf:resource="http://unb.br/literal"/>
        <dcterms:instructionalMethod>Dados gerados a partir da arquitetura de publicação
        de dados abertos da UnB</dcterms:instructionalMethod>
```

```
186       <dcam:VocabularyEncodingScheme
          rdf:resource="http://swat.cse.lehigh.edu/onto/univ-bench.owl"/>
187       <foaf:Organization>Universidade de Brasilia</foaf:Organization>
188     </dcmitype:Dataset>
189   </rdf:RDF>
190
```

Apêndice C

Artigo Publicado no The 10th International Conference on Management of Digital EcoSystems (MEDES'18)

UnBGOLD: UnB government open linked data: semantic enrichment of open data tool

Full Text: PDF Get this Article

Authors:
Luiz C. B. Martins University of Brasilia (UnB), Brasília - DF, Brasil
Márcio C. Victorino University of Brasilia (UnB), Brasília - DF, Brasil
Maristela Holanda University of Brasilia (UnB), Brasília - DF, Brasil
George Ghinea Brunel Univerity London, London, UK
Tor-Morten Grønli Kristiania University College, Oslo, Norway

2018 Article

Bibliometrics

· Citation Count: 0
· Downloads (cumulative): 10
· Downloads (12 Months): 10
· Downloads (6 Weeks): 10

Published in:

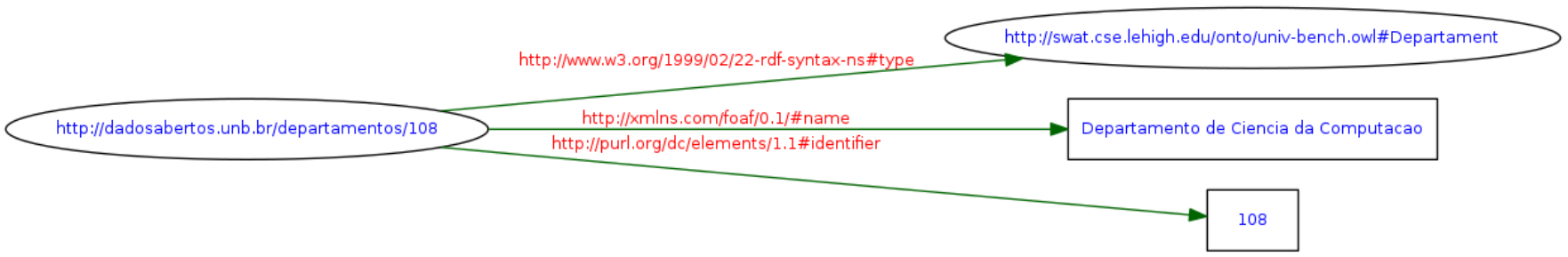
· Proceeding
MEDES '18 Proceedings of the 10th International Conference on Management of Digital EcoSystems
Pages 1-6

Tokyo, Japan — September 25 - 28, 2018
ACM New York, NY, USA ©2018
table of contents ISBN: 978-1-4503-5622-0 doi>10.1145/3281375.3281394

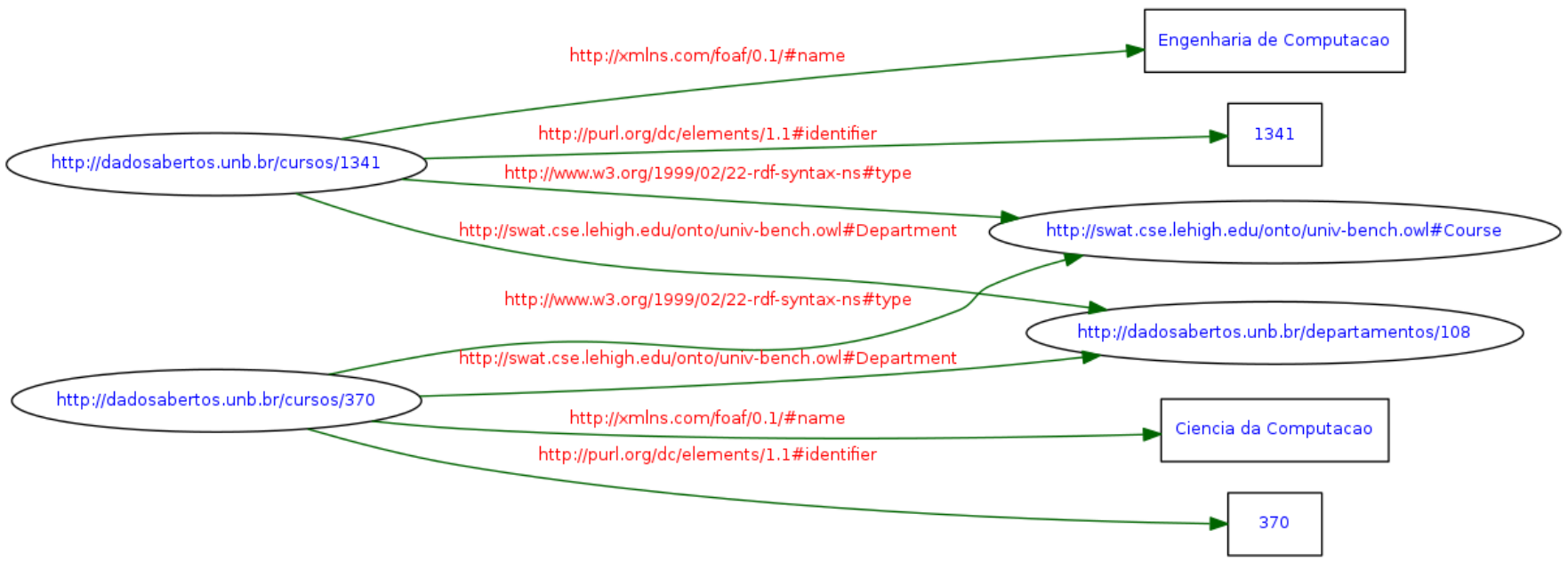
In accordance with current legislation designed to make public management more efficient and transparent, Brazilian Federal agencies have adhered to an open data publication policy, despite the challenge presented by datasets being published collectively rather than in isolation. Aiming to facilitate this process, this article presents the UnBGOLD, which adresses the need to connect the data in order to facilitate the publication of open semantically enhanced data. It is a tool that couples the architecture of open data publishing of the University of Brasilia and makes it possible to transform datasets into linked open data utilizing metadatas and ontologies in RDF formats, aside from making it possible for the data to be published automatically on the CKAN platform.

Apêndice D

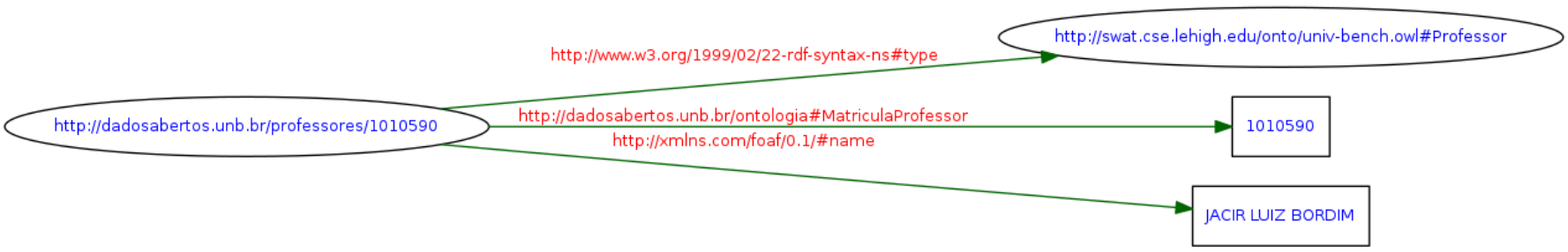
Representação em Grafos dos RDF



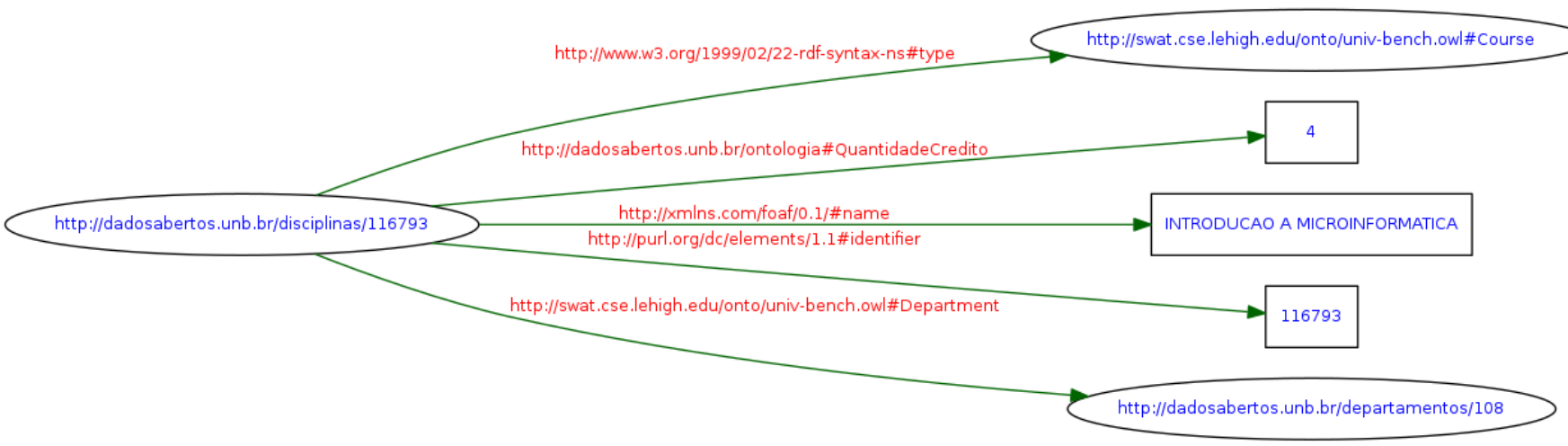
Grafos das triplas do conjunto de dados de departamentos



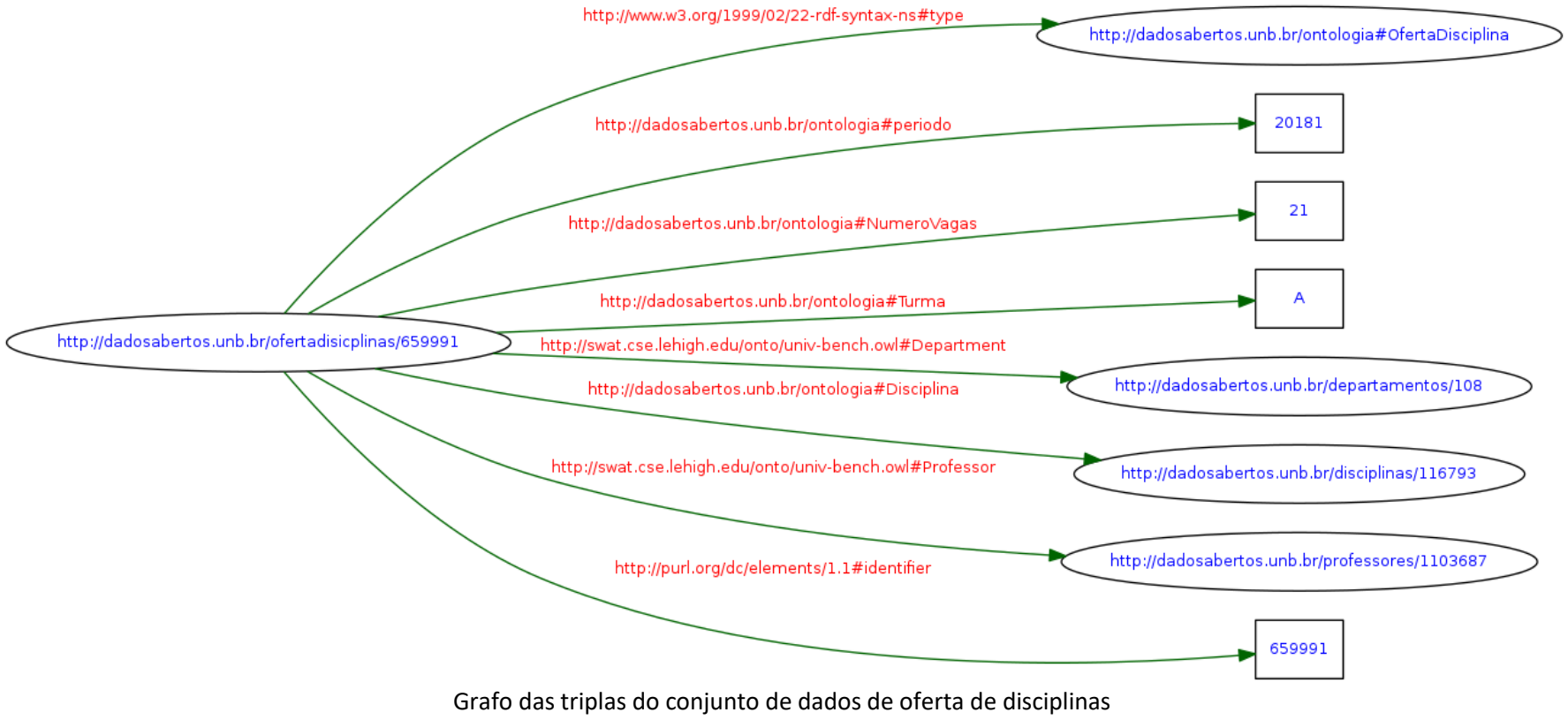
Grafo das triplas do conjunto de dados de cursos



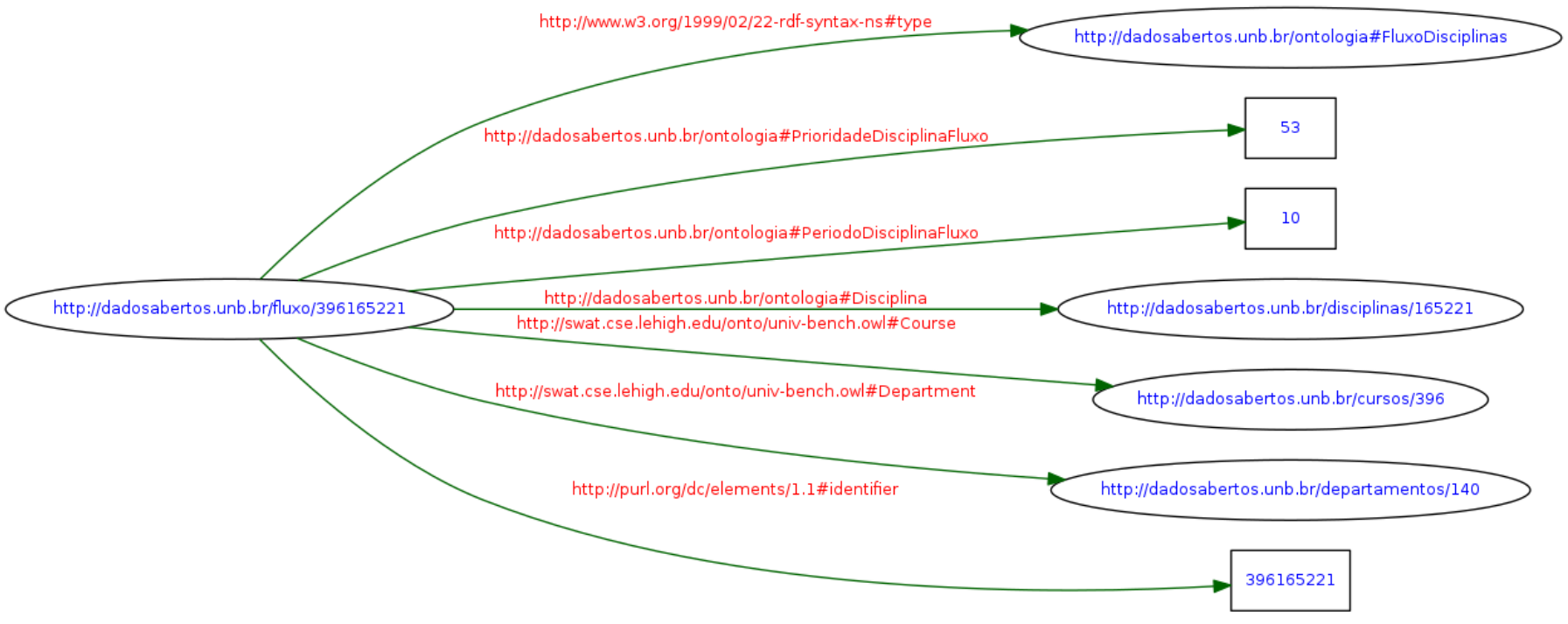
Grafo das triplas do conjunto de dados de Professores



Grafo das triplas do conjunto de dados de disciplinas



Grafo das triplas do conjunto de dados de oferta de disciplinas



Grafo das triplas do conjunto de dados de Fluxo de Disciplinas nas Cursos