INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
IFCE CAMPUS ARACATI
COORDENADORIA DE CIÊNCIA DA COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JOSÉ RENATO DA SILVA FREITAS

MODELO DE INTEGRAÇÃO DE DADOS BASEADO EM ONTOLOGIA E *LINKED DATA* PARA O CÁLCULO DO RISCO DE ÓBITO MATERNO-INFANTIL

ARACATI-CE
2019

JOSÉ RENATO DA SILVA FREITAS

MODELO DE INTEGRAÇÃO DE DADOS BASEADO EM ONTOLOGIA E *LINKED DATA* PARA O CÁLCULO DO RISCO DE ÓBITO MATERNO-INFANTIL

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE - Campus Aracati, como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Orientador (a): Prof. Dr. Antonio Mauro Barbosa de Oliveira

Aracati-CE
2019

Dados Internacionais de Catalogação na Publicação
Instituto Federal do Ceará - IFCE
Sistema de Bibliotecas - SIBI
Ficha catalográfica elaborada pelo SIBI/IFCE, com os dados fornecidos pelo(a) autor(a)

F862m Freitas, José Renato da Silva.
MODELO DE INTEGRAÇÃO DE DADOS BASEADO EM ONTOLOGIA E LINKED DATA PARA O CÁLCULO DO RISCO DE ÓBITO MATERNO-INFANTIL / José Renato da Silva Freitas. - 2019.
61 f. : il.

Trabalho de Conclusão de Curso (graduação) - Instituto Federal do Ceará, Bacharelado em Ciência da Computação, Campus Aracati, 2019.
Orientação: Prof. Dr. Antonio Mauro Barbosa de Oliveira.

1. Integração de dados. 2. Ontologia. 3. Linked Data. 4. Óbito materno- infantil. I. Titulo.

JOSÉ RENATO DA SILVA FREITAS

MODELO DE INTEGRAÇÃO DE DADOS BASEADO EM ONTOLOGIA E *LINKED DATA* PARA O CÁLCULO DO RISCO DE ÓBITO MATERNO-INFANTIL

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE - Campus Aracati, como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Aprovada em 23 de abril de 2019

BANCA EXAMINADORA

Prof. Dr. Antonio Mauro Barbosa de Oliveira (Orientador)
Instituto Federal do Ceará

Prof. Me. Odara Sena dos Santos Feitosa
Instituto Federal do Ceará

Prof. Me. Diego Rocha Lima
Instituto Federal do Ceará

DEDICATÓRIA

Dedico este trabalho aos meus pais, principalmente à minha gentil e doce mãe, Maria Lucinda. Aos meus irmãos, com honras ao Júnior. Dedico também à minha esposa, Eliene Maia e aos meus filhos, Manuel Neto e Ravi.

AGRADECIMENTOS

Agradeço, primeiramente, ao criador de todas as coisas. Agradeço-o pela vida com saúde e por todas as oportunidades de felicidade. A todos os professores que contribuíram para minha formação acadêmica, do Jardim I ao nível superior. Muito obrigado!

Às agências de fomento CNPQ, FUNCAP e ao programa PIBIC pelo incentivo à pesquisa. Ao IFCE por levar ensino de qualidade e profissionalização ao interior do estado.

Agradeço à Lucélia Ribeiro, Cleilton Lima e Charlys Pinheiro do Instituto Atlântico10, e à Dra. Ivana Barreto, da Fundação Oswaldo Cruz (Fiocruz), que muito contribuíram com as heurísticas dos riscos de óbito materno e infantil. À professora Vânia Vidal que conduziu a pesquisa sobre ontologia. Este trabalho teve o apoio da FINEP e FUNCAP, no âmbito do Programa de Incentivo à Interiorização e Inovação Tecnológica - BPI, FUNCAP, edital nº 09/2015.

Agradecimento especial ao meu orientador e amigo, professor Mauro Oliveira. Tenha certeza, professor Mauro, você conseguiu fazer nosso mundo melhor com sua Escola pra Valer. Dás o máximo de si. À professora dessa disciplina, Carina Oliveira, que incrivelmente sempre esteve disponível para ajudar. Suas orientações, críticas construtivas e estímulos fazem jus a sua fama de excelência no ensino.

Agradeço ainda ao meu amigo e irmão de ciência, Bacharel Oton Braga. Obrigado pelas sugestões, apoio e por todas as discussões, questionamentos e descobertas.

Agradecimento mais que especial à minha esposa, Eliene Maia e aos meus filhos Manuel Neto e Ravi. Sou grato por toda paciência, amor e companheirismo. Vocês são meus pilares, pois, trabalhar, estudar, ser pai e marido é uma verdadeira batalha.

RESUMO

Tomar boas decisões de governança é um desafio constante em qualquer atividade profissional, não sendo diferente na área da saúde. Na Saúde Pública, por exemplo, devido à complexidade dos problemas enfrentados pelos profissionais, uma análise de dados deve ser realizada em diversas bases relacionadas a um problema a fim de encontrar soluções. Contudo, tomando como exemplo o SUS cujas bases de dados não se comunicam e a maioria é distribuída e heterogênea, o processo de análise dos dados dessas bases é massivo e dispendioso quando se utiliza ferramentas clássicas. Portanto, são necessários mecanismos computacionais mais elaborados para integrar dados e extrair informações relevantes que auxiliem profissionais de saúde na tomada de decisão. Para tal, as Tecnologias de Informação e Comunicação surgem como facilitadores no processo de integrar dados e extrair informações. Nesse contexto, este trabalho propõe um modelo inteligente baseado em Ontologia e *Linked Data* que realiza a integração de dados clínicos e socioeconômicos de algumas bases de dados do SUS, fornecendo uma visão homogeneizada dos dados anteriormente isolados. Além disso, esse modelo calcula a probabilidade do risco de óbito materno e infantil. Para tal, desenvolveu-se uma ontologia denominada Ontologia de Risco. Essa ontologia foi modelada a partir das heurísticas de especialistas em saúde materno-infantil. Por fim, os resultados obtidos pelo modelo proposto dão subsídios para os profissionais de saúde tomar decisões mais efetivas, seja para traçar estratégias de prevenção ou para solucionar problemas de saúde materno-infantil.

Palavras-chaves: Integração de dados. Ontologia. *Linked Data*. Óbito materno-infantil.

ABSTRACT

Making good governance decisions is a challenge for any professional activity, not being different in healthcare area. In public health, for example, due to the complexity of problems faced by professionals, a data analysis must be performed on several databases related to a problem in order to find solutions. However, using as an example SUS - whose databases do not communicate among themselves, being most of them distributed and heterogeneous - the process of analyzing the data from these databases is massive and expensive using classic tools. Therefore, more elaborate computational mechanisms are needed to integrate data and extract relevant information to assist health professionals in decision making. To this end, Information and Communication Technologies appear as facilitators in the process of integrating data and extracting information. In this context, this research proposes an intelligent pattern based on Ontology and Linked Data that performs the integration of clinical and socioeconomic data from some SUS databases, providing a homogenized view of previously isolated data. In addition, this pattern calculates the probability of maternal and infant death risk. For that, an ontology called Risk Ontology was developed. This ontology was patterned from the heuristics of maternal and child health specialists. Finally, the results obtained by the proposed pattern allow health professionals to make more effective decisions, either to develop prevention strategies or to solve maternal and child health problems.

Keywords: Data integration. Ontology. Linked Data. Maternal and child death.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

RDF	*Resource Description Framework*
SPARQL	*SPARQL Protocol and RDF Query Language*
HTML	*Hypertext Markup Language*
HTTP	*Hypertext Transfer Protocol*
TIC	Tecnologia de Informação e Comunicação
SUS	Sistema Único de Saúde
DATASUS	Departarmento de Informática do SUS
SINASC	Sistema de Informações sobre Nascidos Vivos
SIM	Sistema de Informação sobre Mortalidade
SINAN	Sistema de Informação de Agravos de Notificação
SBC	Sistemas Baseados em Conhecimento
GISSA	Governança Inteligente em Sistemas de Saúde
FINEP	Financiadora de Estudos e Projetos
SAD	Sistema de Apoio a Decisão
W3C	*World Wide Web Consortium*
IRI	*Internationalized Resource Identifier*
FOAF	Friend of a Friend
XML	*Extensible Markup Language*
URI	*Uniform Resource Identifier*
LOD	*Linked Open Data*
URL	*Uniform Resource Locator*
JSON	*JavaScript Object Notation*
CSV	*Comma-separated Values*
SILK	*Link Specification Language*

SIEVE	*Linked Data Quality Assessment and Fusion*
D2RQ	*Database to RDF Platform*
R2RML	*RDB to RDF Mapping Language*
OWL	*Ontology Web Language*

SUMÁRIO

1 INTRODUÇÃO

Importantes organizações como o Banco Mundial e as Nações Unidas têm adotado o conceito de governança em suas diretrizes. Porém, tomar boas decisões de governança é um desafio constante em qualquer atividade profissional, não sendo diferente na área da saúde pública. O conceito de governança está relacionado a uma série de pendências ligadas à administração pública e gestão de políticas públicas. Seu objetivo é explorar e alavancar o conhecimento de um domínio para melhorar o desempenho administrativo e a democratização no processo de tomada de decisão (OLIVEIRA, 2014). Nesse sentido, a Tecnologia da Informação e Comunicação (TIC) tem se mostrado um importante pilar para a obtenção de governança.

Na área da saúde, por exemplo, as TIC's atuam como facilitadoras dos processos de saúde oferecendo ferramentas que ajudam os profissionais de saúde a realizarem melhores diagnósticos. Tais ferramentas também podem auxiliar na gestão dos serviços prestados e gestão de informação em saúde (SENA; MAURO, 2014). Contudo, a gestão da informação passa pelo problema da fragmentação dos serviços de saúde. Essa fragmentação se dá pela inexistência ou dificuldade de manter a integração de dados em sistemas de saúde. Nesse contexto, um dos maiores desafios do Departamento de Informática do SUS (DATASUS) é a integração dos seus dados, dispersos em diversas fontes. (PIERRO, 2011).

1.1 Motivação

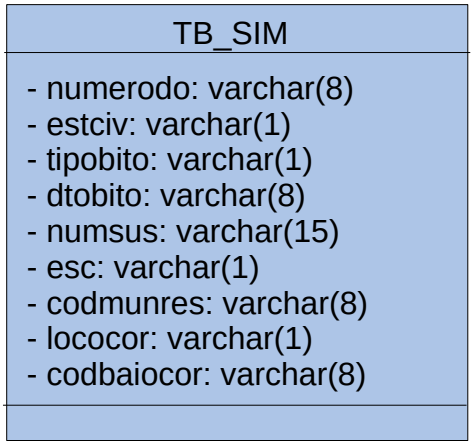
Dado a interdependência entre os diversos domínios envolvidos em sistemas de saúde pública (clínico, epidemiológico, administrativo, normativo e gestão do conhecimento), profissionais de saúde precisam analisar a relação entre os dados provenientes desses domínios (ANDRADE, 2012). Essa análise permite descobrir as prováveis causas de um determinado problema de saúde. Além disso, ajuda a traçar melhores estratégias, seja para prevenir ou solucionar problemas.

No Brasil, os dados de saúde pública são gerados pelo Sistema Único de Saúde (SUS). Por isso, o Ministério da Saúde criou o DATASUS com objetivo de manter os dados de saúde pública e criar ferramentas de *software* para informatizar todas as suas atividades. A partir desses dados, os profissionais de saúde podem medir indicadores a fim de analisar a situação de saúde, fazer comparações e avaliar mudanças ao longo do tempo (SOARES; ANDRADE; CAMPOS, 2001). Além disso, esses dados podem subsidiar a tomada de decisão baseada em evidências além de

elaborar programas e ações em saúde (BRASIL, 2008).

Problemas enfrentados por profissionais de saúde geralmente são complexos, exigindo uma análise de dados em diversas bases distintas a fim de resolvê-los (FARINELLI; ALMEIDA, 2014). No SUS, tais bases armazenam dados de sistemas que não se comunicam e tão pouco trocam informações, uma vez que possuem esquemas de base de dados e terminologias diferentes. O Sistema de Informações de Nascidos Vivos (SINASC[1]), o Sistema de Informações de Mortalidade (SIM[2]) e o Sistema de Informação de Agravos de Notificação (SINAN[3]), por exemplo, são sistemas mantidos pelo DATASUS que geram dados de saúde pública, porém suas bases são distribuídas e heterogêneas. A Figura 1 mostra parte do esquema da base de dados SIM.

Figura 1 – Esquema parcial da base de dados SIM.



Fonte: DATASUS.

As nomenclaturas usadas para os atributos da tabela “TB_SIM” na Figura 1, não são intuitivas. Desse modo, para um ser humano, profissional de saúde ou não, decifrar os significados desses atributos é um problema. Fazendo-se um esforço intelectual, pode-se, talvez, deduzir que o atributo “tipobito” represente o tipo do óbito. Que o atributo “dtobito” signifique a data do óbito, porém seu tipo fora definido como *varchar(8)* e não *date* ou *datetime*, normalmente utilizados para armazenar datas. Entretanto, entender os significados dos atributos “lococor” e “codbaiocor” é um desafio.

Dessa forma, o processo para a realização da análise dos dados das diversas bases disponibilizadas pelo DATASUS, com as atuais ferramentas, ainda é massivo e dispendioso para os tomadores de decisão. Nesse contexto, as bases relacionadas ao domínio de um problema devem ser integradas, sintática e semanticamente, ge-

[1] http://datasus.saude.gov.br/sistemas-e-aplicativos/eventos-v/sinasc-sistema-de-informacoes-de-nascidos-vivos

[2] http://datasus.saude.gov.br/sistemas-e-aplicativos/eventos-v/sim-sistema-de-informacoes-de-mortalidade

[3] http://sinan.saude.gov.br/sinan/login/login.jsf

rando um único conjunto homogêneo de dados. Assim, esse único conjunto de dados proporciona uma visão mais ampla do domínio do problema. Contudo, o processo de integração de dados não é trivial. Faz-se necessário, então, mecanismos computacionais mais elaborados, capazes de integrar dados e extrair informações relevantes que auxiliem gestores de saúde na tomada de decisão.

Tais mecanismos computacionais como os Sistemas Inteligentes (SI) são a melhor opção para tornar explícito o conhecimento de um domínio (REZENDE, 2003; VASCONCELOS, 2018). A partir dos dados, os Sistemas Inteligentes podem extrair informações relevantes e, com isso, tornar mais ágil e efetivo o processo de tomada de decisão. Ainda no contexto de SI, existem, além de outros, os Sistemas Baseados em Conhecimento (SBC). Tais Sistemas Baseados em Conhecimento são excelentes mecanismos para gerar explicações a partir de um modelo simbólico racional (REZENDE, 2003). Além disso, sistemas computacionais baseados em conhecimento que fazem uso de ontologias e tecnologias da Web Semântica são eficazes na integração de fontes de dados heterogêneos e distribuídos (HU et al., 2014).

No contexto de uso de Ontologia e SI, o Sistema de Governança Inteligente de Saúde (GISSA) auxilia, por meio de indicadores, dashboards e mapa de calor, na tomada de decisão em ambientes de saúde pública. Inspirado no framework LARIISA LARIISA (OLIVEIRA et al., 2010), o GISSA foi custeado pela Financiadora de Estudos e Projetos do Ministério de Ciência, Tecnologia e Inovação (FINEP) e desenvolvido pelo Instituto Atlântico[4]. O GISSA atende uma demanda do Programa Rede Cegonha do Ministério da Saúde, cujo objetivo é preservar a saúde da mãe e do bebê durante toda a gestação, puerpério[5] e nos primeiros anos de vida (SILVA et al., 2017).

A carência de mais atenção à saúde da mãe e do bebê reflete consideravelmente na mortalidade infantil. A Tabela 1 mostra uma taxa de 29,02 óbitos a cada mil nascidos no ano 2000 e uma redução para 13,82 óbitos a cada mil nascidos em 2015 (IBGE, 2018). Ainda assim, o problema do óbito infantil é um desafio para os gestores de saúde, pois um dos agravos é o número alarmante de óbitos neonatais. Cerca de 70% dos óbitos neonatais são registrados nas quatro primeiras semanas após o nascimento da criança (ANGULO-TUESTA; SANTOS; NATALIZI, 2016).

Uma das estratégias para melhorar os cuidados em saúde materno-infantil é baseada em critério epidemiológico, taxa de mortalidade infantil, razão de mortalidade materna e densidade populacional (BRASIL, 2018). Nesse contexto, este trabalho propõe elaborar um modelo inteligente baseado em Ontologia e *Linked Data* (Dados Conectados) que realiza a integração de dados clínicos e socioeconômicos do SUS e calcula a probabilidade do risco de óbito materno e infantil.

[4] http://www.atlantico.com.br/
[5] https://pt.wikipedia.org/wiki/Puerpério

Tabela 1 – Decrescimento da Taxa de mortalidade infantil no Brasil.

Ano	Taxa	Ano	Taxa
2000	29,02	2001	27,48
2002	26,04	2003	24,68
2004	23,39	2005	22,18
2006	21,04	2007	19,98
2008	18,99	2009	18,07
2010	17,22	2011	16,43
2012	15,69	2013	15,02
2014	14,40	2015	13,82

Fonte: Instituto Brasileiro de Geografia e Estatística - IBGE.

A integração de dados clínicos e sociais materno-infantil do SUS e a criação da Ontologia de Risco baseada em heurísticas dos especialistas em saúde são as principais contribuições desse trabalho. Enfatiza-se que essas heurísticas foram realizadas pelos profissionais de saúde que fizeram parte do projeto GISSA. A partir de suas experiências e pesquisas, esses profissionais determinaram alguns riscos de óbito materno e infantil, assim como definiram pesos para cada risco. Além disso, estipularam as faixas de riscos. Os valores definidos pelas heurísticas são detalhados na Seção 4.6.1.

Nesse contexto, esse modelo contribuirá para o módulo de inteligência do GISSA, calculando a probabilidade de óbito materno-infantil, fortalecendo a gestão de conhecimento e apoiando profissionais e tomadores de decisão no SUS.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver um modelo de integração de dados baseado em Ontologia e *Linked Data* para o cálculo do risco de óbito materno-infantil.

1.2.2 Objetivos Específicos

A construção do modelo aqui proposto passa por uma série de passos bem definidos, pautados nos seguintes objetivos específicos:

1. Definir a arquitetura do modelo.
2. Analisar as heurísticas dos especialistas.
3. Construir a Ontologia de Risco.

4. Formalizar o cálculo do risco de óbito.

1.3 Produção Científica

No decurso desse trabalho de conclusão de curso, o seguinte artigo científico foi aceito e publicado:

- **Freitas, R.**, Rocha, C., Braga, O., Lopes, G., Monteiro, O., Oliveira, M. **Using Linked Data in the Data Integration for Maternal and Infant Death Risk of the SUS in the GISSA Project.** In: Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web. ACM, 2017. p. 193-196.

1.4 Organização do Trabalho

Os capítulos subsequentes organizam este trabalho da seguinte maneira. O Capítulo 2 apresenta as tecnologias empregadas nesse trabalho e explica os fundamentos teóricos necessários para entender a proposta. No Capítulo 3 são feitas análises de trabalhos com objetivos similares ao deste ou que tenham relação com esta proposta. Tentou-se focar nos trabalhos que utilizam as mesmas tecnologias, porém não foi quesito obrigatório. Já o Capítulo 4 apresenta a proposta do modelo detalhando sua arquitetura e construção. Os resultados alcançados são detalhados no Capítulo 5, assim como os experimentos e contribuições. Por fim, no Capítulo 6 são expostas as considerações finais deste trabalho e as perspectivas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

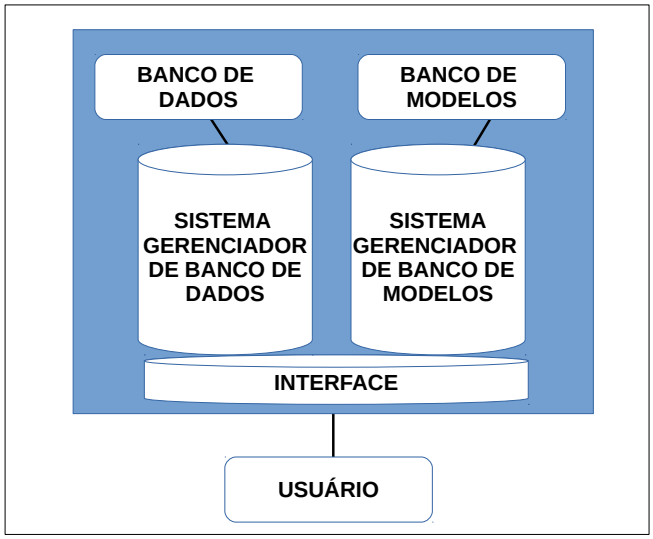
A utilização de TIC's mais emergentes foram essenciais para a implementação do modelo proposto neste trabalho. Vale ressaltar que algumas dessas tecnologias não são apresentadas nas grades curriculares normais. Assim, este capítulo apresenta uma visão geral das tecnologias utilizadas.

2.1 Sistemas de Apoio a Tomada de Decisão

Sistemas de Apoio a Tomada de Decisão ou simplesmente Sistema de Apoio a Decisão (SAD) é um tipo de Sistema de Informação que fornece subsídios úteis aos gestores das organizações e apresenta um conjunto de soluções baseadas em cenários que envolvem o processo de tomada de decisão (HEINZLE; GAUTHIER; FIALHO, 2017). Entre esses subsídios estão os recursos que permitem comparar, analisar, simular e apoiar a escolha de uma ou mais alternativas.

Na literatura, são encontradas várias arquiteturas para um SAD. Entretanto, em boa parte, pode-se observar a existência, no mínimo, dos três componentes seguintes: Banco de Dados; Banco de Modelos & Ferramentas analíticas e Interface com o usuário. A Figura 3 mostra a arquitetura de um SAD proposta por (SPRAGUE; WATSON; JR, 1991). Essa arquitetura apresenta, além dos três componentes básicos, os gerenciadores de bancos de dados e de banco de modelos.

Figura 2 – Arquitetura do SAD definido por *Sprague*.



Fonte: (SPRAGUE; WATSON; JR, 1991).

2.2 Ontologia

Ontologia é um termo de origem grega que tem um sentido especial em organização da informação. É o ramo da Filosofia que estuda do mundo como ele é, o estudo do ser ou da realidade (MORAIS; AMBRÓSIO, 2007). Na área da Ciência da Computação o termo Ontologia é usado na Representação do Conhecimento, um subcampo da Inteligência Artificial (IA), desde 1960 (ALMEIDA, 2014).

No trabalho "O que é uma Ontologia?" (de título original "What is an Ontology?"), Tom Grubber define Ontologia como a especificação de uma conceitualização. Em outras palavras, para Tom Grubber, Ontologia é uma descrição de conceitos e relacionamentos, semelhante a uma especificação formal de um programa (GRUBER, 1993).

Já (BORST, 1997) interpreta Ontologia como uma especificação formal e explícita de uma conceitualização compartilhada. Nesse sentido, adotamos as definições de Tom Grubber e Borst que convergem em uma representação formal e inequívoca do conhecimento através do conjunto de conceitos, suas relações e propriedades. A representação formal significa que a Ontologia é legível tanto por homem quanto por computador e inequívoca porque traduz que os conceitos não são ambíguos.

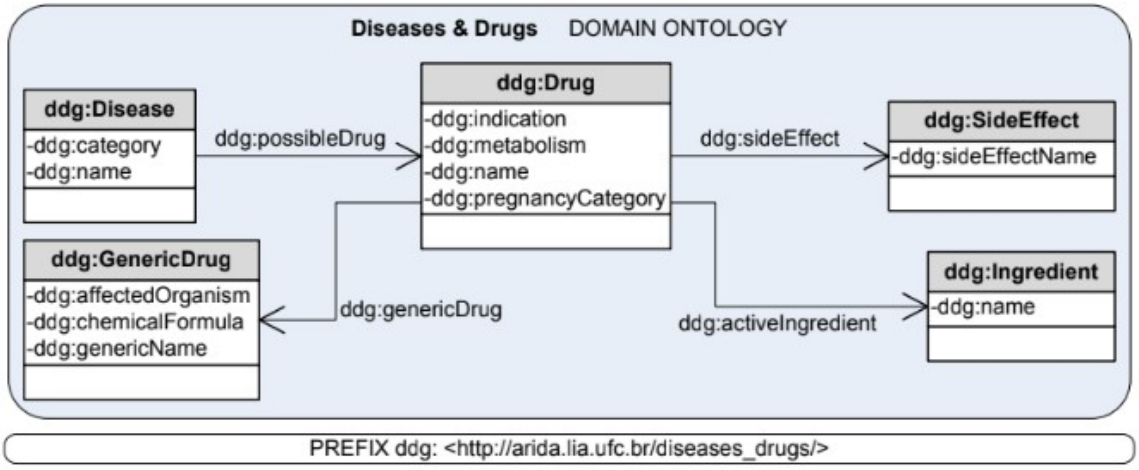
As ontologias, geralmente, são compostas por classes, propriedades, relações e indivíduos. As classes representam as entidades do domínio (os objetos categorizados) e as propriedades são as características dessas entidades. Já as relações são as ligações entre as classes (tipo de interação entre as entidades) e os indivíduos são as instâncias das classes (representam elementos específicos, ou seja, os dados de uma classe) (ALMEIDA; BAX, 2003). Nesse sentido, o propósito essencial da construção de ontologias é permitir compartilhamento e reutilização do conhecimento (MORAIS; AMBRÓSIO, 2007)

A Figura 3 apresenta um exemplo de uma ontologia. Essa ontologia foi modelada em (MAGALHÃES et al., 2012) e denominada Doenças & Drogas (*Diseases & Drugs* - D&D). Tipificada como Ontologia de Domínio, D&D utiliza o prefixo "ddg" para facilitar na construção do seu vocabulário.

O objetivo da ontologia D&D é representar o conhecimento cujos possíveis medicamentos são indicados para o tratamento de uma doença. Para isso, a ontologia D&D representa as classes: Doença (*ddg:Disease*), Droga (*ddg:Drug*), Efeito Colateral (*ddg:SideEffect*), Medicamento Genérico (*ddg:GenericDrug*) e Ingrediente (*ddg:Ingredient*). Observa-se, ainda, que nessa representação existem quatro relacionamentos entre essas classes. A classe *ddg:Drug* se relaciona com as clas-

ses *ddg:SideEffect*, *ddg:Ingredient* e *ddg:GenericDrug* por meio das propriedades "efeito colateral" (*ddg:sideEffect*), "ingrediente ativo" (*ddg:activeIngredient*) e "medicamento genérico" (*ddg:genericDrug*), respectivamente. Já a classe *ddg:Disease* relaciona-se com a classe *ddg:Drug* por meio da propriedade "possível medicamento" (*ddg:possibleDrug*).

Figura 3 – Exemplo Ontologia.



Fonte: (MAGALHÃES et al., 2012).

2.2.1 Utilização de Ontologias

O fato das ontologias também serem legíveis por computador permite um leque de possibilidades de seu uso, dentre os quais destacamos o uso nas seguintes áreas da Ciência da Computação: gestão do conhecimento, comércio eletrônico, sistemas de recomendação, processamento de linguagens naturais, recuperação da informação na Web e Web Semântica.

2.2.2 Tipos de Ontologias

As ontologias são classificadas de acordo com suas características, levando em consideração o grau de formalismo, aplicação, conteúdo ou função (MORAIS; AMBRÓSIO, 2007). Dessa forma, as ontologias podem ser:

- **Quanto ao nível de formalismo**
  - Altamente informais (linguagem natural);
  - Semi-informais (linguagem natural estruturada);

  - Semiformais (linguagem artificial definida formalmente);
  - Rigorosamente formais (termos são definidos com semântica formal, teoremas e provas).

- **Quanto à aplicação**
  - De autoria neutra (aplicativo é descrito em uma língua e convertido para o uso em outros sistemas);
  - De especificação (baseada em uma ontologia de domínio, utilizada para documentação e manutenção no desenvolvimento de softwares);
  - De acesso comum à informação (vocabulário é inacessível. a ontologia torna a informação possível compreensível compartilhado dos termos).

- **Quanto ao conteúdo**
  - Terminológicas (representam termos para modelar o conhecimento de um domínio);
  - De informação (especificam estruturas de registros de um banco de dados);
  - De modelagem de conhecimento (definem conceitualizações do conhecimento;
  - De aplicação (contém as definições necessárias para modelar conhecimento em uma aplicação);
  - De domínio (expressam conceitualizações específicas de domínio);
  - De representação (explicam as conceitualizações que estão por trás dos formalismos de representação do conhecimento).

- **Quanto a sua função**
  - Ontologias Genéricas (descrevem conceitos mais amplos, como elementos da natureza, tempo, coisas, eventos, processos ou ações, independente de um problema específico ou domínio particular);
  - Ontologias de Domínio (descrevem conceitos e vocabulários relacionados a domínios específicos, medicina ou computação, por exemplo. Este é o tipo de ontologia mais comum);
  - Ontologias de Tarefas (descrevem atividades que podem contribuir na resolução de problemas, por exemplo, processos de vendas ou diagnóstico);
  - Ontologias de Aplicação (descrevem conceitos que dependem tanto de um domínio particular quanto de uma tarefa específica);
  - Ontologias de Representação (explicam as conceituações que fundamentam os formalismos de representação de conhecimento, procurando tornar claros os compromissos ontológicos embutidos nestes formalismos).

2.3 *Framework* de Descrição de Recurso

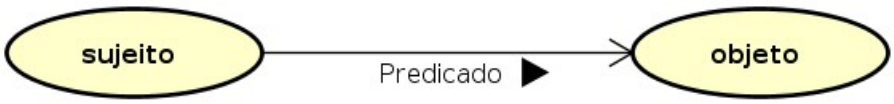
O Consórcio da Rede Mundial de Computadores (mais conhecido em inglês como *World Wide Web Consortium* - W3C) é uma comunidade formada por organizações internacionais que trabalham em conjunto para desenvolver padrões da Web a fim de explorar todo o potencial da Internet (W3C, 2018).

Na busca de desenvolver tais padrões para Web, o W3C criou o framework de descrição de recurso (em inglês, *Resource Description Framework* - RDF). O RDF permite representar formalmente qualquer objeto tornando-o legível por computador. Nesse contexto, é importante compreender a riqueza de expressividade que o RDF proporciona para descrever e representar informações na Web.

Usado para representar e conectar dados na Web, o RDF é um modelo de dados baseado em grafo. Grafo é uma estrutura composta por um conjunto de elementos chamados “nós” ou “vértices” e por um conjunto de pares de “nós”, não ordenados, denominados “arestas”. As arestas podem ter orientação e um valor associado (FEOFILOFF, 2018).

A estrutura base do RDF é no formato de tripla. Cada tripla consiste de um **sujeito**, **predicado** e **objeto** e um conjunto de triplas é chamado de grafo RDF (W3C, 2014a; W3C, 2014b). A Figura 4 é um exemplo de um grafo RDF com dois nós (sujeito, objeto) e uma aresta (predicado) constituindo uma única tripla.

Figura 4 – Estrutura base de tripla RDF.

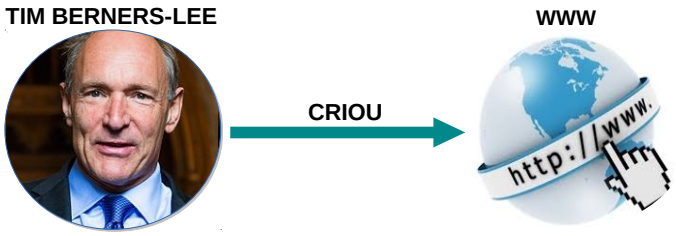


Fonte: Elaborado pelo autor.

No universo dos grafos RDF, um nó pode ser um literal, um nó vazio ou um identificador de recurso internacionalizado (em inglês, *Internationalized Resource Identifiers* - IRI). O IRI é o método de identificar de maneira única um objeto na Web. Esses objetos identificados são denominados recursos. Dessa forma, qualquer objeto pode ser um recurso, seja ele uma entidade, documento, número, data, imagem, *string*, entre outros.

Intuitivamente, a Figura 5 induz a seguinte declaração: “Tim Berners-Lee criou o WWW”. Essa declaração, seguindo o padrão de tripla RDF (sujeito, predicado, objeto), representa uma relação entre dois recursos. Nesse caso, `<TIM BERNERS-LEE>` e `<WWW>` são os recursos dessa relação e representam, respectivamente, o sujeito e objeto da tripla RDF. Ao passo que, `<CRIOU>` é a propriedade do recurso `<TIM`

Figura 5 – Exemplo de recursos em tripla RDF.



Fonte: Elaborado pelo autor.

`BERNERS-LEE>` e representa o predicado da tripla. O valor de um objeto pode ser vazio, um valor literal ou outro recurso, como é nesse caso.

A boa prática do RDF diz que todo recurso deve ser referenciado por um IRI. Seguindo essa boa prática, a Tabela 2 é um exemplo de uso de IRI para representar a tripla `<TIM BERNERS-LEE> <CRIOU> <WWW>`.

Tabela 2 – Exemplo de uso de IRI.

	Nó	IRI
Sujeito	<Tim Berners-Lee>	<http://exemplo.edu/recurso/Tim-Berners-Lee>
Predicado	<criou>	<http://exemplo.edu/propriedade/criou>
Objeto	<www>	<http://exemplo.edu/recurso/WWW>

Fonte: Elaborado pelo autor.

Apesar do exemplo da Tabela 2 ser válido, é recomendado reutilizar vocabulários já consolidados na representação de dados na Web como, por exemplo, o FOAF[1], DBPedia[2], WikiData[3], Dublin Core[4] e outros para estabelecer os IRI's. Esses vocabulários são formados por um conjunto de palavras ou termos em IRI's definidos por uma pessoa ou organização. A Tabela 3 mostra a reutilização dos vocabulários *WikiData* e *DBPedia* para identificar a tripla `<TIM BERNERS-LEE> <CRIOU> <WWW>`.

Tabela 3 – Exemplo de IRI com reutilização de vocabulários.

	Nó	IRI
Sujeito	<Tim Berners-Lee>	<http://www.wikidata.org/entity/Q80>
Predicado	<criou>	<http://dbpedia.org/ontology/developer>
Objeto	<www>	<http://dbpedia.org/page/World_Wide_Web>

Fonte: Elaborado pelo autor.

[1] http://xmlns.com/foaf/spec/
[2] https://wiki.dbpedia.org/
[3] https://www.wikidata.org/wiki/Wikidata:Main_Page
[4] http://dublincore.org/

2.3.1 Serialização RDF

Dados RDF podem ser escritos em diversos formatos, conhecidos como serializações. Na época do surgimento do RDF, as linguagens de programação tinham mais suporte para a Linguagem de Marcação Extensível (em inglês, *Extensive Markup Language* - XML). Esse maior suporte foi o motivo para o RDF/XML se tornar a primeira notação usada para RDF. Porém, essa notação é difícil para leitura humana. Por conta dessa dificuldade, outras notações surgiram proporcionando uma leitura mais fácil para as pessoas. Entre essas novas notações, as mais utilizadas são: *Turtle*, *N-Triple* e *JSON-LD* (LAUFER, 2015).

Para exemplificar a serialização RDF, usaram-se os dados relacionais das canções exibidas na Tabela 4. Nessa serialização, cada canção é um recurso, cada coluna é uma propriedade e cada célula é o valor do objeto na tripla RDF.

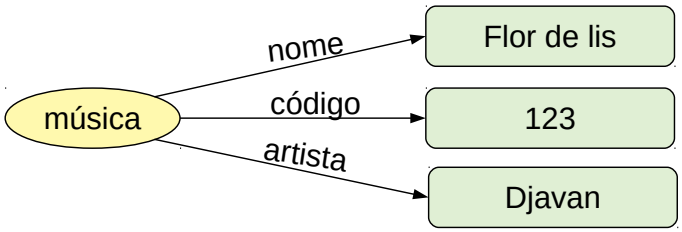
Tabela 4 – Tabela de canções.

código	nome	artista
123	Flor de Lis	Djavan
101	Non, Je ne regrette rien	Édith Piaf
212	Nessun Dorma	Pavarotti

Fonte: https://en.wikipedia.org/.

A Figura 6 mostra um grafo RDF com três triplas de um mesmo recurso da Tabela 4. A partir desse grafo criou-se o exemplo de notação *Turtle* reutilizando os vocabulários *FOAF*, *Music Ontology*[5] e *Dublin Core*.

Figura 6 – Grafo com três triplas de um mesmo recurso (música).



Fonte: Elaborado pelo autor.

A sintaxe da serialização *Turtle* é construída por sentenças. Cada sentença contém três elementos que definem uma tripla RDF e é concluída por um ponto final (LAUFER, 2015). Seguindo esse preceito, a Figura 7 apresenta as três triplas da canção “Flor de lis” em notação *Turtle*. A primeira tripla diz respeito à identificação da canção, reutilizando o URI `<http://purl.org/dc/elements/1.1/indentifier>`

[5] http://musicontology.com/

do vocabulário *Dublin Core*. A segunda tripla refere-se ao nome da canção, reutilizando o vocabulário *FOAF* por meio do URI `<http://xmlns/com/foaf/0.1/name>`. Por fim, a última tripla desse exemplo reutiliza a URI `<http://purl.org/ontology/mo/MusicArtist>`, do vocabulário *Music Ontology* para descrever o artista que interpreta a canção.

Figura 7 – Notação *Turtle*.

```
<http://exemplo.edu/#flor-de-lis>
  <http://purl.org/dc/elements/1.1/identifier> "123" .

<http://exemplo.edu/#flor-de-lis>
  <http://xmlns.com/foaf/0.1/name> "Flor de lis" .

<http://exemplo.edu/#flor-de-lis>
  <http://purl.org/ontology/mo/MusicArtist> "Djavan" .
```

Fonte: Elaborado pelo autor.

Considerando que um documento RDF pode conter uma grande quantidade de triplas, a notação *Turtle* da Figura 7 deixaria os arquivos muito extensos. Utilizar prefixos para os IRI's dos vocabulários e agrupar as propriedades de um recurso ajudam a manter esses documentos mais compactos e a notação *Turtle* mais legível.

Considerando que um documento RDF pode conter n triplas e que as notações em *Turtle* podem deixar esses arquivos muito extensos, a utilização de prefixos para os IRI's dos vocabulários e o agrupamento de propriedades de um recurso ajuda a manter esses documentos mais compactos e a notação *Turtle* mais legível. Nesse sentido, a Figura 8 apresenta as mesmas três triplas da canção "Flor de lis" de forma mais enxuta. Na primeira linha, estabelece-se o endereço base do recurso. Nas linhas seguintes, definem-se todos os prefixos dos vocabulários que serão utilizados. Logo em seguida, descreve-se, entre "<>", o recurso desejado. As propriedades de cada recurso são agrupadas por meio do sinal de pontuação ";". Percebe-se que, neste formato, a escrita das triplas é mais simples e para alguns, mais legível. Assim, este estilo de notação é o escolhido e usado neste trabalho.

2.4 *Linked Data*

A aplicação do conceito de *Linked Data* vem crescendo consideravelmente. Prova disso está no aumento significativo das conexões entre dos dados em RDF disponibilizados na Web. O projeto Dados Abertos Conectados (*Linked Open Data* - LOD), por exemplo, publica vários conjuntos de dados como RDF na Web e define

Figura 8 – Notação *Turtle* compacta.

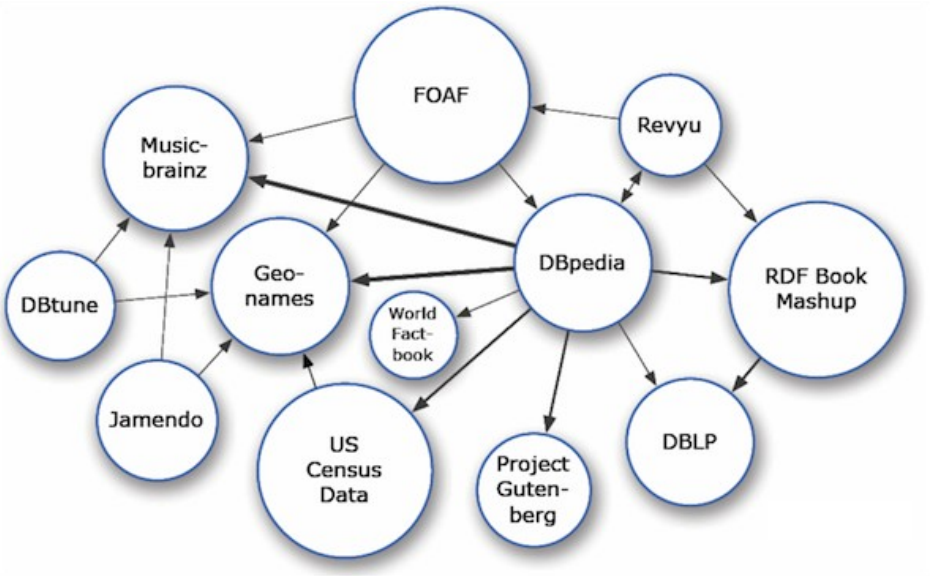
```
@base <http://exemplo.edu/>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-Syntax-ns#>
@prefix foaf: <http://xmlns.com/foaf/0.1/name>
@prefix mo: <http://purl.org/ontology/mo/MusicArtist>
@prefix dc: <http://purl.org/dc/elements/1.1/>

<#flor-de-lis>
     dc:identifier   “123” ;
     foaf:name       "Flor de lis" ;
     mo:MusicArtist  "Djavan" .
```

Fonte: Elaborado pelo autor.

links semânticos entre itens de fontes distintas. Esses dados são apresentados na nuvem do LOD em forma de diagrama.

A Figura 9 mostra o início do diagrama de *Linked Data* do LOD. Em maio de 2007 a nuvem do LOD era formada por 12 conjuntos de dados. Já em novembro de 2018, essa nuvem acumulara um total de 1231 conjuntos de dados conectados. A Figura 10 exibe o diagrama do LOD de 2018.

Figura 9 – Nuvem do Linked Open Data em 2007.

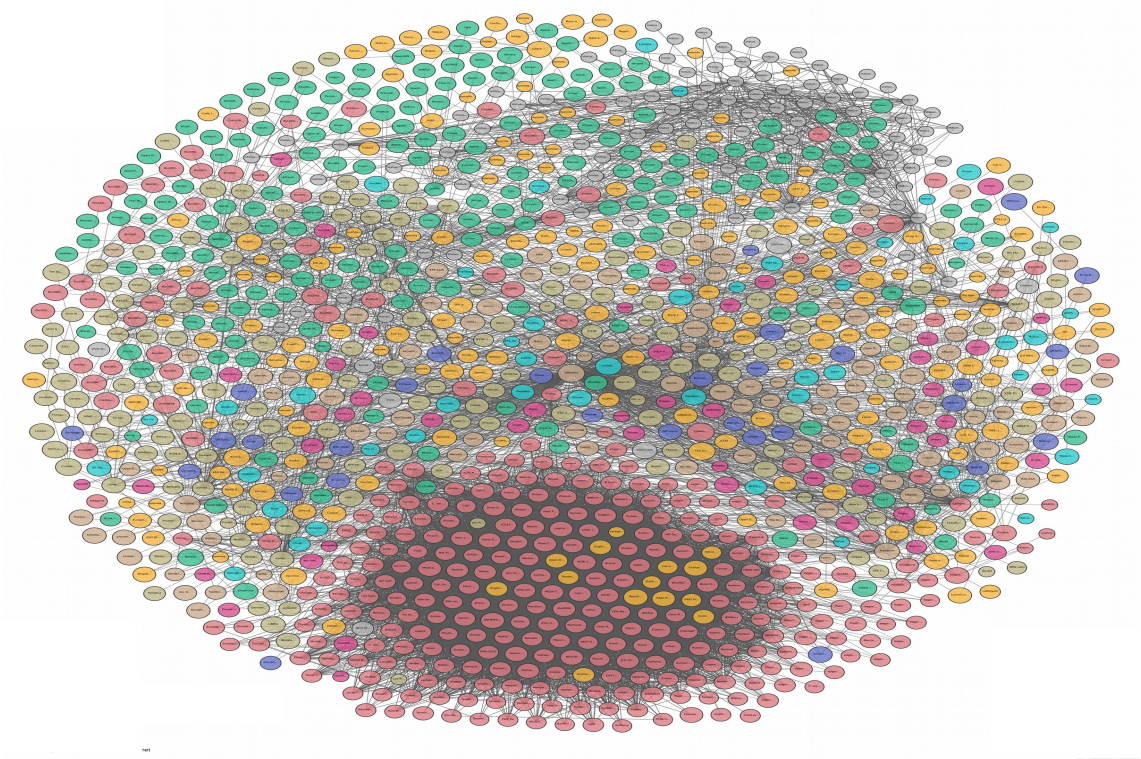


Fonte: https://lod-cloud.net/versions/2007-05-01/lod-cloud.png

Um dos primeiros documentos sobre *Linked Data* foi apresentado por Tim Berners-Lee em 2006. O documento em questão é intitulado “Design Issues: Linked Data” [6]. Nesse documento, Tim Berners-Lee acredita que a Web Semântica não é apenas sobre publicar dados na Web, mas é sobre criar *links* de um dado para outros permitindo que pessoas ou máquinas explorem esses dados (BERNERS-LEE, 2006).

[6] https://www.w3.org/DesignIssues/LinkedData.html

Figura 10 – Nuvem do Linked Open Data em 2018.



Fonte: https://lod-cloud.net/versions/2018-11-26/lod-cloud.png

O conceito de *Linked Data* pode ser definido como um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na Web, com significados bem definidos e legíveis por máquina (BIZER; HEATH; BERNERS-LEE, 2011). Essas práticas, fundamentadas em tecnologias web, usam o Identificador de Recurso Uniforme (*Uniform Resource Identifier* - URI) para identificar uma única entidade ou objeto no mundo através do Protocolo de Transferência de Hipertexto (*Hypertext Transfer Protocol* - HTTP). Além disso, tais práticas adotam o uso de *links* semânticos para conectar diferentes fontes de dados na Web. Essas fontes podem estar em lugares diferentes ou serem sistemas distintos de uma mesma organização, por exemplo..

Diferente da web de documentos, que é a web atual, cujos *links* são relacionamentos entre documentos hipertextos ou hipermídias escritos em HTML, a Web de Dados usa *links* que conectam objetos descritos em RDF. Dessa maneira, para que seja possível o desenvolvimento da Web de Dados, as seguintes regras são estabelecidas: (BERNERS-LEE, 2006):

1. Usar URI como nome para objetos;
2. Usar URI's em HTTP para que as pessoas possam procurar esses nomes;

3. Quando alguém procura um URI, forneça informações úteis;
4. Inclua *links* para outros URI's para que pessoas possam descobrir mais objetos.

A primeira regra é bem clara e entendida por quem usa tecnologias da Web e *Linked Data*: se não usar URI, não se pode nomear tão pouco conectar coisas na Web. A segunda regra tem a ver com entender que os URI's HTTP são nomes, não endereços, que facilitam a pesquisa de objetos na Web. Fornecer informações úteis significa que deve ser usado o RDF para descrever o URI procurado, de acordo com a terceira regra. Por fim, a quarta regra diz que é necessário conectar dados de outros lugares para que possamos encontrar todos os tipos de coisas.

A maneira mais simples de conectar dados na Web é usar um URI que aponta para outro. Para tal, pode-se escrever um arquivo RDF com URL base e identificadores locais, como por exemplo:

```
1 @base <http://exemplo.com/freitas>
2 @prefix fam: <http://exemplo.com/ontologias/familia\#>
3
4 <\#renato>
5     fam:temFilho <\#manuel> <\#ravi> .
```

Código 2.1 – RDF de exemplo do recurso Renato

A respeito do Código 2.1, na linha 1 é descrito a URL base para o documento RDF. Logo em seguida, na linha 2, é descrito o prefixo "fam" para referenciar o vocabulário exemplo. Esse vocabulário exemplo é uma ontologia que representa a estrutura familiar. Dando prosseguimento, dentre as linha 4 e 5 tem-se uma tripla afirmando que o recurso `<#renato>` tem os filhos `<#manuel>` e `<#ravi>`. Após o RDF acima ser publicado na Web, os URI's <http://exemplo.com/freitas#renato>, <http://exemplo.com/freitas#manuel> e <http://exemplo.com/freitas#ravi>, podem ser usados por qualquer pessoa ou máquina para referenciar os recursos Renato, Manuel e Ravi e obter informações. Em outro arquivo RDF, por exemplo, poderia ter o seguinte conteúdo:

```
1 @base <http://exemplo.com/maia>
2 @prefix fam: <http://exemplo.com/ontologias/familia\#>
3 @prefix freitas: <http://exemplo.com/freitas>
4
5 <\#eliene>
6     fam:temFilho <freitas\#ravi> .
```

Código 2.2 – RDF de exemplo do recurso Eliene

O documento RDF apresentado no Código 2.2 também apresenta a URL base dos recursos e os prefixos para os vocabulários reutilizados, nas linhas 1 a 3. É possível observar, na linha 5, a reutilização do URI `fam:temFilho` para referenciar uma propriedade. Essa propriedade está conectando `<#eliene>`, que é um recurso do segundo arquivo RDF, ao recurso `<freitas#ravi>` do arquivo RDF do Código 2.1. Nesse exemplo simples, ao criar o *link* entre os recursos, pode-se observar que os recursos Renato e Eliene são os pais do recurso Ravi. Só é possível extrair tal informação após conectarem-se os dados dos dois documentos RDF. Nesse contexto, dados que podem se integrar a outros dados e consequentemente formar novos conhecimentos, demonstram que *Linked Data* é uma área de relevância a ser explorada.

2.5 SPARQL

Os dados em RDF publicados na Web podem ser disponibilizados como grafos em memória, arquivos de texto ou *RDF Stores*. *RDF Stores* são bancos de dados apropriados para o armazenamento e acesso a grafos RDF e suas tripas (LAUFER, 2015; CAVALCANTE, 2017). Esses bancos de dados podem ser do tipo *Triple Stores* ou *Quad Stores*. Entre os diversos *RDF Stores* citados na literatura, o Virtuoso[7] e o Fuseki[8] são fortemente recomendados em aplicações para a Web Semântica (CAVALCANTE, 2017).

Nesse contexto, a linguagem de consulta RDF (em inglês, *SPARQL Protocol and RDF Query Language* - SPARQL) é recomendada pelo W3C para consultas semânticas aos RDF Stores, para recuperar e manipular dados em RDF (PRUD'HOMMEAUX; HARRIS; SEABORNE, 2013).

No intuito de permitir a recuperação e manipulação de dados, diversos *RDF Stores* oferecem pontos de acesso na Web que aceitam o SPARQL (protocolo e linguagem). Esses pontos de acesso SPARQL são denominados Pontos Finais. No entanto, na literatura, a denominação "Pontos Finais" em português não é habitual. O termo mais usual para esse tipo de ponto de acesso, em inglês, é chamado de *endpoints*. Dessa forma, esse trabalho adotará tal nomenclatura. *Endpoints* podem ser específicos ou genéricos. *Endpoints* específicos, só permitem consultas a um determinado conjunto de dados. Já os genéricos, como o *Open Link Software*[9], possibilita consultas a partir da especificação da localização dos grafos RDF desejados (LAUFER, 2015).

[7] http://virtuoso.openlinksw.com/
[8] https://jena.apache.org/documentation/serving_data/
[9] http://demo.openlinksw.com/sparql

2.5.1 Consulta SPARQL

A maioria das consultas SPARQL é formada por um conjunto básico de padrões de triplas. Esses padrões de triplas seguem o mesmo padrão das triplas RDF, a não ser pelo fato de que cada sujeito, predicado e objeto pode ser uma variável. Nesse caso, quando os termos de um grafo RDF podem ser substituídos pelas variáveis de uma consulta SPARQL, o resultado dessa consulta corresponde a um subgrafo RDF do grafo original.

A fim de mostrar um exemplo simples de consulta SPARQL, o grafo RDF de canções, apresentado na Figura 8, foi utilizado como dados para consulta. A partir desses dados, a consulta escrita no Código 2.3 tem a finalidade de descobrir qual é o artista que interpreta a canção “Flor de lis”. Essa consulta consiste em duas cláusulas: *SELECT* e *WHERE*. A cláusula *SELECT* determina as variáveis que aparecerão nos resultados. Nesse exemplo, apenas a variável “?nome” foi criada. Em SPARQL, os nomes das variáveis são iniciados pelo ponto de interrogação “?”. Já a cláusula *WHERE* define o padrão de triplas que buscam corresponder ao grafo de dados RDF. Nesse exemplo, o conjunto básico de triplas é formado por apenas uma tripla cuja variável “?nome” está na posição do objeto da tripla. O resultado desse exemplo pode ser visto na Tabela 5. Em SPARQL, o retorno de uma consulta é um conjunto de tuplas (e não triplas) e podem ser definidos nos formatos HTML, XML, JSON, CSV (LAUFER, 2015).

Consulta SPARQL:

```
1 PREFIX mo: <http://purl.org/ontology/mo/>
2 SELECT ?nome
3 WHERE
4 {
5   <http://exemplo.edu/flor-de-lis> mo:MusicArtist ?nome .
6 }
```

Código 2.3 – Exemplo simples de consulta SPARQL.

Resultado:

Tabela 5 – Resultado da consulta SPARQL do Código 2.3.

nome
"Djavan"

Fonte: Elaborado pelo autor.

O SPARQL permite consultas mais complexas que podem ser feitas utilizando filtros e operadores. Por todos os seus recursos, o SPARQL pode ser considerado uma tecnologia chave da Web Semântica

2.6 Web Semântica

O *World Wide Web*, ou simplesmente Web, surgiu da necessidade de compartilhamento de documentos entre computadores em um espaço global. Esse compartilhamento foi possível através da criação da linguagem de marcação de hipertexto (*Hipertext Markup Language* - HTML), do protocolo de transferência de hipertexto (*Hipertext Transfer Protocol* - HTTP) e do identificador de documentos (*Uniform Resource Locator* - URL). Essa Web também é chamada de Web 1.0 cuja interação do usuário é quase inexistente. Nessa Web 1.0, clicar nos *links*, que levam os usuários para outro documento, é basicamente a única interação disponível.

Com o passar do tempo, a Web passou a exibir conteúdo dinâmico em vez de apenas páginas de documentos estáticos. Nascia, assim, a Web 2.0. Na Web 2.0, além de conteúdo dinâmico, várias aplicações são permitidas. Nessa Web, a interação do usuário é potencializada, visto que várias aplicações precisam de entradas de dados.

Nesse contexto, a Web 2.0 passa a ser um espaço com várias informações valiosas que podem auxiliar em diversas pesquisas, como por exemplo: cura de doenças; previsões no mercado financeiro e na tomada de decisão. Porém, essas informações não estão estruturadas de forma que um computador possa ler, compreender e processá-las.

Como o computador pode processar uma quantidade enorme de dados bem mais rápido que o homem, de forma mais precisa, e a Web é um gigantesco repositório de dados, Tim Berners-Lee propôs evoluir a Web 2.0 para a Web de Dados.

A Web de Dados ou Web Semântica cria inúmeras oportunidades para a integração semântica dos próprios dados, motivando o desenvolvimento de novos tipos de aplicações e ferramentas (ISOTANI; BITTENCOURT, 2015). A Tabela 6 mostra as principais diferenças entre a Web de Documentos e a Web Semântica.

A comparação apresentada na Tabela 6 diz que a Web de Documentos equivale a um sistema global de arquivos enquanto a Web de Dados equipara-se a um banco de dados global. Os documentos são os principais objetos da Web de Documentos ao passo que na Web Semântica os recursos é que são. Dessa forma, documentos têm apenas *links* para outros documentos ou partes deles enquanto recursos usam *links* para encontrar quaisquer outros recursos, inclusive documentos. Além disso, a Web de Documentos tem semântica implícita, ou seja, necessita de mecanismo para compreensão. Já a Web Semântica, como o próprio nome já diz, tem semântica de explícita. Por fim a Web de Documentos está delineada para consumo humano, à medida que a Web Semântica é dedicada, primordialmente, para consumo

Tabela 6 – Comparação Web de Documentos e Web Semântica.

	Web de Documentos	Web Semântica
Analogia a	Um sistema global de arquivos	Um banco de dados global
Objetos primários	Documentos	Coisas
Links entre	Documentos ou partes dele	Coisas (incluindo documentos)
Grau de estrutura em objetos	Relativamente baixo	Alto
Semântica de conteúdos e *links*	Implícito	Explícito
Desenhado para	Consumo humano	Máquina primeiro, humanos depois

Fonte: http://tomheath.com/slides/2008-04-beijing-linked-data-principles-and-state-of-the-art.pdf.

por máquina.

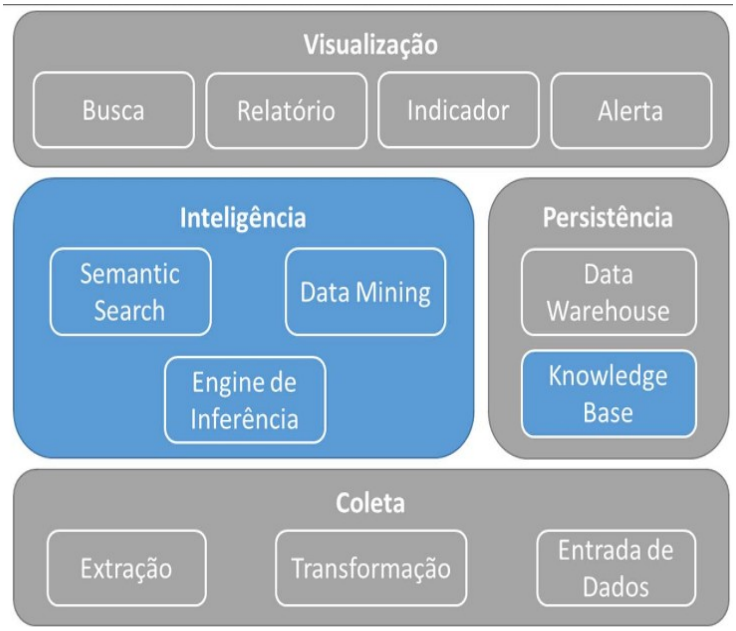
2.7 GISSA

O GISSA é uma plataforma de computação que provê inteligência de governança na tomada de decisão nos cinco domínios clássicos da área de Gestão em Saúde: clínico epidemiológico, técnico administrativo, normativo, gestão compartilhada e gestão do conhecimento. A partir dos dados coletados dos sistemas de informação em saúde no âmbito do SUS e Cartão Nacional de Saúde (CNS), o GISSA usa seu mecanismo de inteligência para analisar e transformar os dados em informações integradas (OLIVEIRA, 2015).

A Figura 11 exibe os quatro núcleos da arquitetura do GISSA: Coleta, Persistência, Inteligência e Visualização. Cada núcleo é formado por um conjunto de componentes e permite a coleta, integração, transformação, inferência e visualização de informações fornecendo aos usuários finais fatos e dados necessários às diversas decisões relacionadas à saúde pública.

O núcleo Coleta é responsável pela a extração, transformação e entrada de dados das fontes utilizadas. O núcleo Visualização provê a interface com o usuário: *dashboards*, relatórios e indicadores, por exemplo. Já o núcleo Persistência armazena as informações das fontes de dados utilizadas pelo GISSA. Nesse núcleo estão presentes os componentes: Armazém de Dados (*Datawarehouse*) e Base de Conhecimento (*Knowledge Base*). Por fim, o núcleo Inteligência conta com os componentes de Mineração de Dados (*Data Mining*), Busca Semântica e Motor de Inferência. Esses componentes são partes integrantes da gestão do conhecimento do GISSA.

Figura 11 – Arquitetura GISSA.



Fonte: (OLIVEIRA, 2015).

Como prova de Conceito, o GISSA foi implantado no município de Tauá, no Ceará, para atender uma demanda do Programa Rede Cegonha do Ministério da Saúde. Esse programa atua sobre os cuidados materno e infantil.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta alguns trabalhos que fizeram uso de *Linked Data*, Ontologias, tecnologias da Web Semântica e outras tecnologias emergentes com o objetivo de realizar a integração de dados e auxiliar no processo de tomada de decisão.

Dados abertos conectados ou, em inglês, *Linked Open Data* (LOD) para auxiliar profissionais de saúde foi a tecnologia adotada em (KOZáK et al., 2013). Esse trabalho apresentou uma integração de dados referente a informações de medicamentos comercializados na República Tcheca. Nessa integração, as fontes de dados utilizadas foram selecionadas de acordo com as principais necessidades de conhecimentos dos médicos para o atendimento. Tais necessidades foram prontamente encontradas a partir da aplicação de um questionário. Esse questionário foi submetido a quarenta e três (43) médicos de especialidades diferentes e dentre os conhecimentos necessários apontados pelos médicos, estão as várias informações sobre medicamentos.

Após obter o resultado dos questionários, foram identificadas as fontes de dados que supririam tais necessidades dos médicos e que fossem substancialmente fáceis de integrar. Nesse sentido, a essas fontes de dados, (KOZáK et al., 2013) aplicara os princípios *Linked Data* e Processamento de Linguagem Natural (PLN) para realizar a integração.

Além da integração, uma aplicação foi desenvolvida para consumir os dados integrados. Essa aplicação tem por objetivo otimizar o tempo do médico em relação a consultas sobre medicamentos, servindo como uma ferramenta de suporte à tomada de decisão, ajudando a reduzir erros nas prescrições médicas. O resultado foi um conjunto de dados ligados disponibilizado para consultas via SPARQL ao *endpoint*: `http://linked.opendata.cz/sparql`. Há uma pretensão dos autores em disponibilizar a aplicação desenvolvida na Internet ou implantá-la em intranets de hospitais.

Embora tenha usado os princípios *Linked Data* esse trabalho não mostrou ou deixou implícito o uso de ontologia de domínio ou ontologia de aplicação para representar o conhecimento que o médico necessita ou para fazer algum tipo de mapeamento semântico entre bases de dados, ontologias fontes e o próprio *dataset* que ele gerou.

O trabalho de (JUNG; YOON, 2016) implementou um modelo de integração de dados usando a Linguagem de Ontologia da Web (*Ontology Web Language* - OWL) a fim de criar uma base inteligente de conhecimento biomédico. Para essa integra-

ção foram coletados registros de saúde e dados relacionados a doenças usando o OPENAPI Sua arquitetura combina registros médicos eletrônicos e dados de saúde oriundos de dispositivos pessoais de monitoramento. Para a análise sobre essa base de conhecimento, algoritmos de Aprendizado de Máquina (em inglês, *Machine Learning*) foram usados para aprender a classificar a fim de proporcionar suporte nas tomadas de decisão clínica.

Apesar de esse trabalho ter seguido o padrão HL7 e OWL para realizar a interoperabilidade semântica, o mesmo não fez uso dos princípios *Linked Data* recomendados pelo W3C.

Já o sistema proposto em (HU et al., 2014), baseado em *Linked Data*, faz uma integração entre os dados dos sistemas de informações de hospitais, que registram casos de tratamento de câncer e dados abertos no campo de ciência da vida, (*Linked Life Data* - LLD)[1]. Depois da integração, usa algoritmo de seleção para encontrar casos de tratamento de câncer com base na similaridade de classificação do paciente. Esse modelo ajuda médicos a filtrarem históricos dos casos, semanticamente.

Um grande diferencial do nosso trabalho para o trabalho proposto por (HU et al., 2014) é o mapeamento das ontologias fontes exportadas. Apesar de esse trabalho usar o D2RQ Server para converter o banco de dados relacional para RDF ele não faz o mapeamento semântico das colunas das tabelas. Assim, se uma coluna tiver uma nomenclatura que não pode ser compreendida torna-se difícil uma integração semântica com qualidade.

Em (SENA; MAURO, 2014) foi proposto uma plataforma de integração de dados para o LARIISA. Essa plataforma permite a integração de várias bases de dados de saúde onde problemas de governança estão envolvidos. Dessa forma, essa integração sugere o aumento na qualidade da inferência do LARIISA quando relacionado à gestão em saúde. Tal integração foi proposta usando *Linked Data Mashup* (LDM).

O artigo aborda a má qualidade dos serviços e governança de saúde no Brasil, apesar dos esforços empenhados. Além disso, justifica o uso de TIC's como facilitadores no processo de prover saúde na perspectiva de tomada de decisão. Propõe um modelo que determina como os dados deveriam ser submetidos para o LARIISA, sejam esses dados oriundos de sistemas exclusivos, de sistemas externos ou ainda de Dados Abertos. Além disso, o modelo proposto especifica como o LARISSA infere sobre os dados obtidos.

Para a integração foram utilizadas duas fontes de dados de dengue: o Sistema de Monitoramento Diário de Agravo[2] e o Observatório da Dengue[3]. Apesar de

[1] http://linkedlifedata.com/about
[2] http://tc1.sms.fortaleza.ce.gov.br/simda/index
[3] https://www.observatorio.inweb.org.br/dengue/conteudo/sobre

promissor, o modelo proposto por (SENA; MAURO, 2014) faz uso de uma fonte não oficial (o Observatório da Dengue) e não conta com uma interface amigável.

Outro trabalho com *Linked Data Mashup* (LDM) é encontrado em (MAGALHÃES et al., 2012). Esse trabalho apresenta uma arquitetura de LDM baseada em *Linked Data Mashup Services* (LIDMS). Frisa-se LIDMS como serviços Web que combinam e integram dinamicamente dados de múltiplas fontes e retornam o resultado no padrão *Linked Data*. Na arquitetura proposta existe um componente para execução eficiente de consultas federadas sobre *Linked Data*. As principais contribuições desse trabalho são: (i) a definição de um processo para geração de consultas federadas capazes de realizar a integração de *Linked Data*; (ii) a definição e implementação de um ambiente para execução eficiente dessas consultas.

Além disso, esse trabalho realizou um estudo de caso na área de saúde que integra informações sobre doenças e medicamentos. Esse estudo de caso é um *Mashup* de dados a partir das fontes *Diseasome*, *DailyMed*, *DrugBank*, *DBpedia* e *Sider* denominado Doenças e Drogas (DD). Essas fontes estão estruturadas e disponíveis publicamente na Web. Contudo, o tempo de respostas das consultas ainda precisa ser melhorado, pois esse trabalho não conta um *RDF Store* pra visões materializadas.

O trabalho em (SILVA et al., 2017) apresentou um analisador baseado em Mineração de Dados para emitir alertas inteligentes em sistemas de saúde. Para a atuação desse analisador, foram utilizados os dados integrados das bases de dados do DATASUS: SIM e SINASC. O acesso a essas bases foi feito via TABWIN (um software exclusivo para o sistema operacional Windows disponibilizado pelo DATASUS para tabulação de dados e conversão para SQL).

Para realizar a integração, o atributo “numerodn” foi estabelecido para encontrar o relacionamento entre as bases de dados, visto que esse atributo está presente em ambas. Após a integração e análises, os autores determinaram 16 atributos para o aprendizado do modelo inteligente. Esses atributos foram escolhidos por apresentaram completude em seus dados. Os alertas inteligentes foram implementados através de um protótipo de *software* formado por uma *interface* de entrada de dados da mãe e da criança e pelo modelo inteligente gerado. O modelo inteligente classifica cada entrada exibindo a probabilidade de óbito.

Apesar da relevante contribuição desse trabalho, ressalta-se que fora usado *scripts* SQL na integração dos dados. Contudo, consultas SQL se tornam complexas em cenários cujos vários conjuntos de dados estão presentes. Além disso, o SQL não descrevem dados de forma semântica.

Embora a maioria dos trabalhos discorridos acima tenha usado Ontologia e

Linked Data, este trabalho de conclusão se sobressai por utilizar uma ontologia de risco desenvolvida a partir das heurísticas dos especialistas em saúde. Ou seja, o modelo deste trabalho foi embasado por profissionais da área atribuindo credibilidade.

A Tabela 12 exibe a comparação dos trabalhos relacionados apresentados neste capítulo e destaca a contribuição deste trabalho como a melhor solução.

Figura 12 – Tabela dos trabalhos relacionados.

TRABALHOS	LINKED DATA	ONTOLOGIA	HEURÍSTICAS
Desenvolvimento de Linked Data Mashups com o uso de LIDMS (2012).	SIM	SIM	NÃO
Linked Open Data for Health Profissional (2013).	SIM	SIM	NÃO
A Linked Data Based Decision Support System for Cancer Treatment (2014).	SIM	SIM	NÃO
A Data Integration Model for a Decision Making System in Health (2014).	SIM	SIM	NÃO
LAIS, um Analisador Baseado em Classificadores para a Geração de Alertas Inteligentes em Saúde (2017).	NÃO	NÃO	NÃO
Nossa proposta	**SIM**	**SIM**	**SIM**

Fonte: Elaborado pelo autor.

4 MODELO PROPOSTO

Ter uma visão homogênea dos dados de fontes isoladas e heterogêneas não é trivial. Por isso, esse trabalho propõe um modelo computacional mais eficiente de integração de dados que calcula a probabilidade de óbito materno e infantil. Esse capítulo apresenta a arquitetura básica do modelo, bem como o seu processo de construção.

A construção do modelo proposto neste trabalho é fundamentada em Ontologia e *Linked Data*. Portanto, seguiram-se as especificações de materialização apresentadas em (LOPES; VIDAL; OLIVEIRA, 2016a; LOPES; VIDAL; OLIVEIRA, 2016b). Essa materialização resulta em uma visão homogeneizada dos dados, da qual é utilizada para realizar inferências. Dessa forma, a elaboração desse modelo decorre por cinco etapas, a seguir:

1. Selecionar as fontes de dados.
2. Extrair os dados das fontes selecionadas, possivelmente heterogêneos, e transformá-los em grafos RDF.
3. Identificar os *links* semânticos entre as fontes de dados.
4. Fundir as representações do mesmo objeto em fontes distintas em uma visão homogeneizada.
5. Realizar consultas parametrizadas a fonte integrada de dados usando o vocabulário da Ontologia de Domínio e obter o cálculo da probabilidade do risco de óbito materno-infantil.

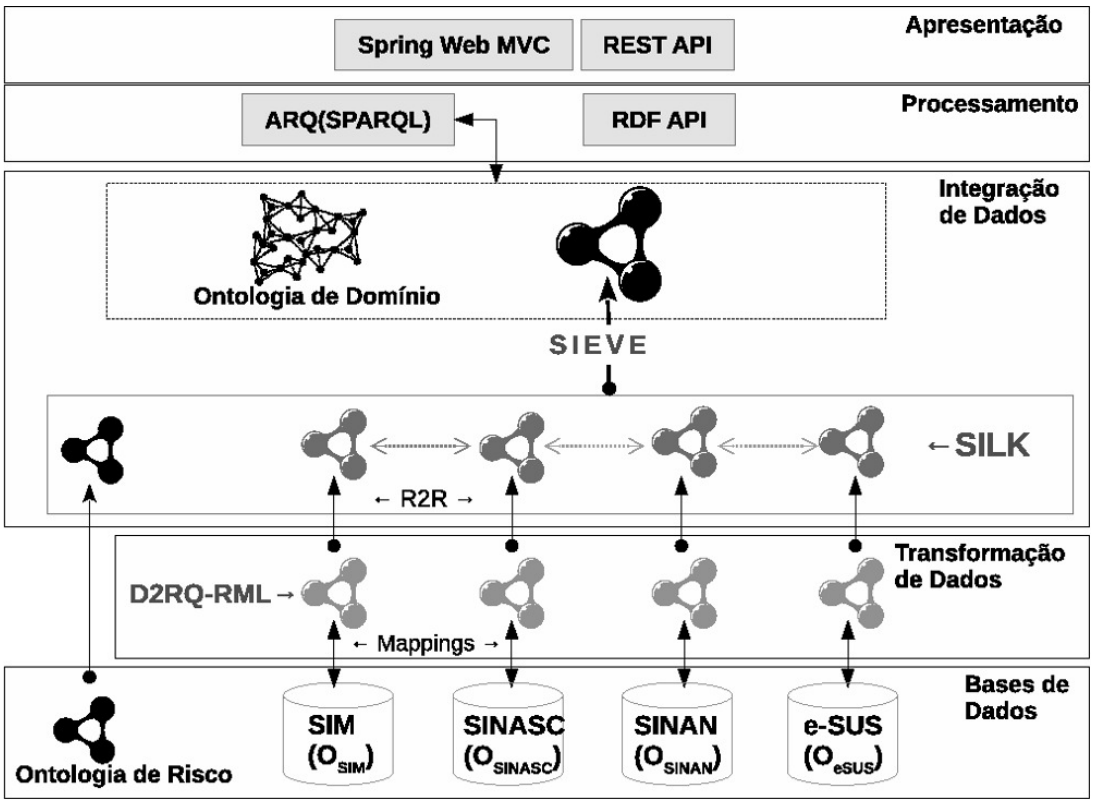
4.1 Arquitetura

Descrevendo partir da camada mais abaixo na arquitetura, a Camada Bases de Dados é formada pelas bases de dados selecionadas e pela Ontologia de Risco. A arquitetura do modelo proposto está exibida na Figura 13 e estruturada em cinco (5) camadas.

Na Camada Transformação de Dados é realizado o acesso e a conversão das bases de dados relacionais para RDF. Para esse fim, são utilizados os *frameworks* D2RQ[1] e R2RML[2]. A Camada Integração de Dados é responsável por realizar a resolução de identidade, ou seja, interligar as fontes RDF descobrindo uma mesma

[1] http://d2rq.org/
[2] https://www.w3.org/TR/r2rml/

Figura 13 – Arquitetura do modelo baseado em Ontologia e Dados Conectados.



Fonte: Elaborado pelo autor.

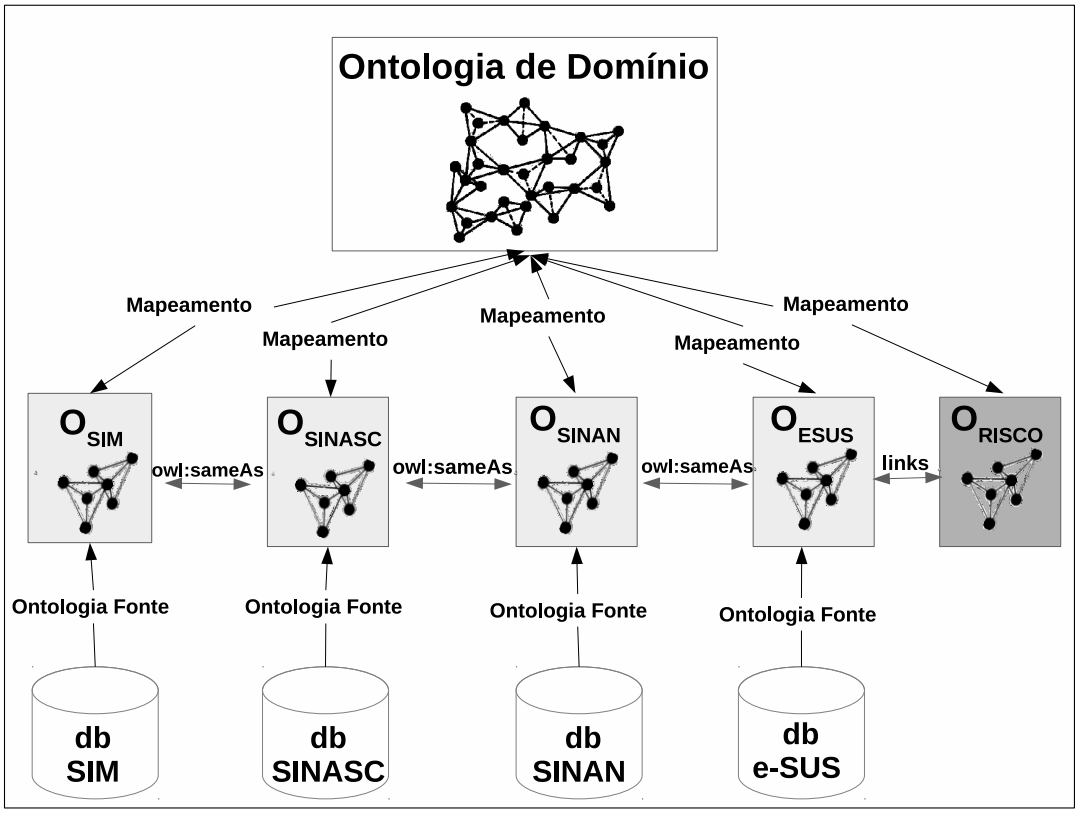
entidade em fontes distintas. Na Camada Processamento de Dados são realizadas consultas parametrizadas à fonte integrada de dados. Nessa camada, também são realizados os cálculos dos riscos de óbitos materno e infantil. Por fim, na Camada Apresentação, uma aplicação pode consumir as informações inferidas pela Camada Processamento de Dados.

As cinco camadas da arquitetura serão descritas com mais detalhes nas secções seguintes, incluindo as tecnologias utilizadas no processo.

4.2 Base de Conhecimento

Todo modelo computacional baseado em Ontologia deve possuir uma base de conhecimento. Uma Base de Conhecimento é geralmente composta por uma ontologia de domínio, uma ontologia de aplicação e regras de inferência. Assim, para a base de conhecimento do modelo proposto foram desenvolvidas uma Ontologia de Domínio e uma Ontologia de Risco (Figura 14). A Ontologia de Domínio e a Ontologia de Risco foram, respectivamente, denominadas O_D e O_{Risco}. A O_D representa uma demanda de governança em saúde e a O_{Risco} representa as heurísticas dos especialistas. Essas ontologias são detalhadas nas subseções a seguir.

Figura 14 – Base de Conhecimento e Mapeamentos das fontes de dados.



Fonte: Elaborado pelo autor.

4.2.1 Ontologia de Domínio

A ontologia de domínio O_D dentro da Base de Conhecimento é uma ontologia de referência, isto é, ela especifica todos os conceitos e axiomas necessários ao modelo proposto. Contém o vocabulário geral que é utilizado para integrar os dados das fontes, exportados em RDF, através dos mapeamentos e *links* semânticos. Além disso, a O_D é o guia que permite a realização de consultas SPARQL.

4.2.2 Ontologia de Risco

A O_{Risco} também faz parte da Base de Conhecimento. Ela foi desenvolvida a partir das heurísticas dos especialistas em saúde materno-infantil. Na Figura 15 é mostrada parte da O_{Risco}, englobando os riscos materno e infantil nos domínios clínico e social.

Nesses domínios estão os fatores de risco relevantes para o cálculo da probabilidade para risco de óbito materno-infantil. Além dos riscos clínico e social, também são descritos na O_{Risco} os eventos relacionados à gestação e ao parto.

Assim, a Ontologia de Risco O_{Risco} representa uma coleção de riscos, tais como: Uma mãe que tem baixa escolaridade, que não recebe auxílio financeiro do governo (riscos sociais), que teve rubéola (risco clínico), etc. Considerando que alguns riscos maternos têm influência direta no bebê, essa correlação também está representada na O_{Risco}. Por exemplo, se uma mãe teve tétano neonatal, se o parto foi induzido, se a gestação foi múltipla, então o risco de óbito do bebê aumenta consideravelmente. A O_{Risco} possui 53 riscos e cada risco tem um peso. Esses pesos foram definidos pelos especialistas, mediante o relato de suas experiências e pesquisas, em conformidade com a gravidade do risco.

A Figura 15 apresenta o grafo da O_{Risco} com alguns riscos clínicos do bebê. O nó inicial desse grafo é a classe padrão *owl:Thing* que é gerada na criação de ontologias utilizando a ferramenta PROTÉGÉ[3]. Em seguida, da classe *owl:Thing* derivam-se duas subclasses: Risco dos Eventos Parto e Gestação (*ClinicalRiskOfEvents*) e Risco de Óbito (*RiskOfDeath*). Da classe Risco de Óbito, derivam-se as subclasses Risco Clínico de Óbito (*ClinicalRiskOfDeath*) e Risco Social de Óbito (*SocialRiskOfDeath*). Da mesma forma, a classe Risco Clínico de Óbito tem duas subclasses: Risco Clínico de Óbito Infantil (*ClinicalRiskOfInfantDeath*) e Risco Clínico de Óbito Materno (ClinicalRiskOfMaternalDeath). Por fim, a classe de Risco Clínico de Óbito Infantil apresenta os riscos: `Baixo peso ao nascer` (*low_weight_at_birth*), `Apgar crítico` (*critical_apgar*), `Baixo apgar de` (*low_apgar_from*), `Prematuro em menos de 22 semanas` (*premature_less_than_22_weeks*), `Prematuro de 22 a 27 semanas` (*premature_from_22_to_27_weeks*), `Prematuro de 28 a 31 semanas` (*premature _from _28_to_31_weeks*) e `Prematuro de 32 a 36 semanas` (*premature_from_32 _to_36_weeks*).

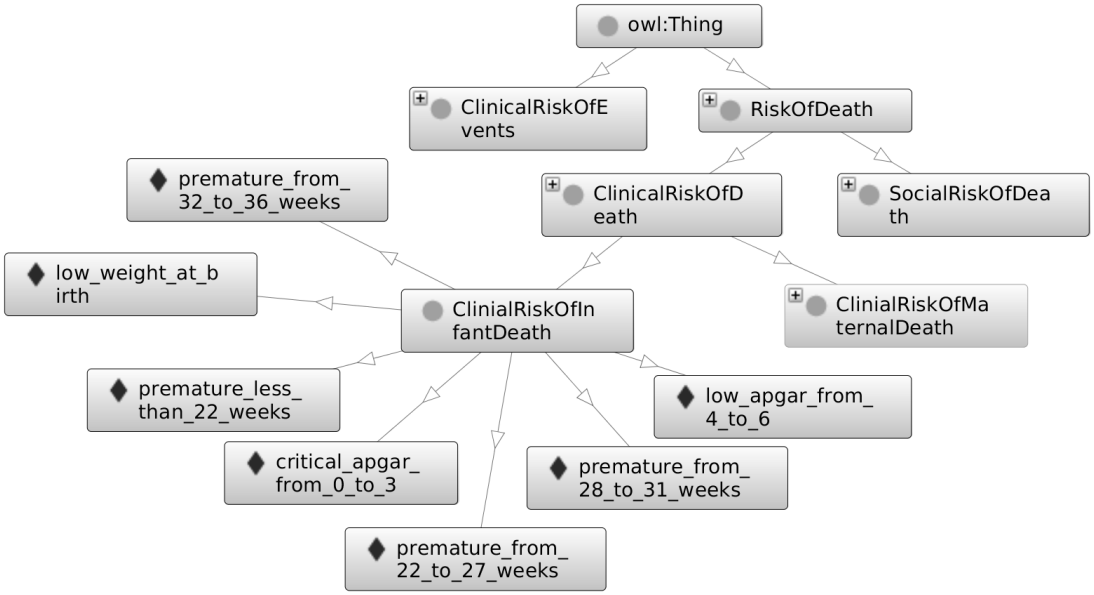
4.3 Seleção das Fontes de Dados

As fontes de dados selecionadas foram criadas e são mantidas pelo DATASUS. Elas estão diretamente ligadas ao contexto de saúde materno e infantil. Salienta-se que a escolha das fontes foi baseada nas experiências dos especialistas em saúde materno-infantil. Assim, essas fontes são essenciais para a proposta deste trabalho.

A primeira fonte selecionada advém do Sistema de Informações de Nascidos Vivos (SINASC[4]). Esta fonte reúne informações epidemiológicas sobre os nascimentos no território brasileiro. Já a fonte de dados do Sistema de Informações de Mortalidade (SIM[5]), segunda selecionada, persiste os dados sobre mortalidade, também em nível nacional. Com base nas informações existentes na fonte SIM, pode-se re-

[3] https://protege.stanford.edu/
[4] http://www2.datasus.gov.br/DATASUS/index.php?area=060702
[5] http://www2.datasus.gov.br/DATASUS/index.php?area=060701

Figura 15 – Parte da Ontologia de Risco.



Fonte: Elaborado pelo autor.

alizar análise da situação, planejar e avaliar ações e programas na área subsidiando todas as esferas da gestão na saúde pública. Além disso, proporciona a construção de indicadores de saúde, análise estatística, epidemiológicas e sociodemográficas.

A terceira fonte de dados utilizada foi o Sistema de Informação de Agravos de Notificação (SINAN[6]). Essa fonte oferece dados clínicos por meio da notificação e investigação de casos de doenças e agravos que constam da lista nacional de doenças de notificação compulsória.

A quarta e última fonte de dados selecionada foi o SUS Eletrônico (e-SUS[7]). Essa fonte dispõe informações clínicas, sociais e econômicas mediante registro da situação de saúde individualizado do cidadão via Cartão Nacional de Saúde.

Cada fonte de dados F_i selecionada é descrita por uma ontologia fonte O_{Fi}. Isso significa que cada base de dados é representada por uma ontologia por meio de um mapeamento. Esse mapeamento é explicado na Seção 4.4 e exemplificado na Figura 14.

4.4 Transformação dos Dados

Para transformar as bases de dados selecionadas em grafo RDF utilizaram-se duas ferramentas recomendadas pela comunidade da Web Semântica[8].

6 http://www2.datasus.gov.br/DATASUS/index.php?area=0203id=29878153
7 http://datasus.saude.gov.br/projetos/50-e-sus
8 https://www.w3.org/2001/sw/wiki/D2RQ

D2RQ A *A Database to RDF Mapper* (D2RQ) é uma plataforma para acessar bases de dados relacionais e proporcionar uma visão dos dados no formato RDF. Essa plataforma roda em máquina virtual Java 1.5 ou superior e oferece suporte para alguns bancos de dados como: *PostgreSQL*, *MySQL*, *SQL Server* e *Oracle Database*.

A plataforma D2RQ acessa as bases de dados utilizando a ferramenta generate-mapping. Essa, por sua vez, lê o script com as configurações de acesso à base de dados e realiza o processo que cria o arquivo de mapeamento no formato R2RML. Além disso, o D2RQ interpreta os arquivos de mapeamentos do R2RML e gera os RDF populados. A Figura 16 mostra um *script* de utilização da ferramenta *generate-mapping*. Nesse *script* são passados os seguintes argumentos: `-r2rml`, indicando o padrão do mapeamento adotado; `-d`, o drive do banco de dados; `-u`, o usuário do banco de dados; `-p`, a senha do usuário; o `jdbc` seguido da URL do banco de dados e após o sinal “>” o `$arquivo_r2rml$`, nomeando o arquivo de saída.

Figura 16 – *Script* de acesso à base de dados.

```
generate-mapping --r2rml -d $nome_drive_classe$ -u
$usuario_bd$ -p $senha_bd$ $jdbc:url_bd$ >
$arquivo_r2rml$
```

Fonte: Elaborado pelo autor.

R2RML O *framework RDB to RDF Mapping Language* (R2RML) é uma linguagem que faz o mapeamento customizado das bases de dados relacionais para o formato RDF.

Para realizar o mapeamento, definiu-se uma coluna-chave que identifica um registro (tupla) na tabela do banco de dados para ser o URI do sujeito da tripla, de acordo com o padrão RDF. Da base de dados SIM, por exemplo, escolheu-se o atributo “numerodo”. Esse atributo representa o identificador da declaração de um óbito. Já da base de dados SINASC, escolheu-se o atributo “numerodn” que representa o identificador da declaração de um nascimento. As demais colunas selecionadas foram mapeadas para serem as propriedades do referido URI e seus valores literais foram mapeados para serem os objetos na tripla.

A Figura 17 exibe um documento R2RML customizado. Esse documento foi gerado a partir da execução da ferramenta *generate-mapping* e customizado manualmente para garantir o mapeamento semântico. A customização se deu pela escolha das colunas das tabelas dos bancos de dados relacionais e pela utilização do vocabulário da O_D.

Junto à notação `rr:sqlQuery` do documento customizado, realizou-se a seleção das colunas NUMERODN, PESO e APGAR1 da tabela tbSinasc. Além disso, mapeou-se a coluna NUMERODN para ser o sujeito das triplas, seguindo o padrão RDF, por ser o identificador de um registro na tabela. Esse mapeamento pode ser visto na notação `rr:subjectMap`. Já o mapeamento dos predicados foi realizado com a notação `rr:predicateObjectMap`. Observa-se que as demais colunas foram mapeadas como predicados das triplas. Por fim, os objetos das triplas foram mapeados para serem os valores das respectivas colunas, porém, receberam novas descrições que correspondem a um termo do vocabulário da O_D . A coluna PESO, por exemplo, passou a ser representada pelo termo "peso_ao_nascer", referenciado pelo prefixo da O_D, "gissa".

Figura 17 – Documento R2ML Customizado.

```
map:Sinasc a rr:TripleMap;
rr:logicalTable [
   rr:sqlQuery """
      SELECT  NUMERODN, PESO, APGAR1 FROM tbSinasc""";
   ];
   rr:subjectMap [ rr:class gissa:Sinasc; rr:template
"Sinasc/{\"NUMERODN\"}"; ];
   rr:predicateObjectMap [
      rr:predicate gissa:peso_ao_nascer;
      rr:objectMap [ rr:column "\"PESO\"" ];
   ];
   rr:predicateObjectMap [
      rr:predicate gissa:apgar1;
      rr:objectMap [ rr:column "\"APGAR1\"" ];
   ];
.
```

Fonte: Elaborado pelo autor.

As ferramentas D2RQ e R2RML trabalham em conjunto para realizar tal transformação na Camada Transformação de Dados. Nesse sentido, após a customização do documento R2RML foi necessário gerar os documentos RDF's. Para esse fim, utilizou-se a ferramenta *dump-rdf* da plataforma D2RQ. A Figura 18 apresenta o *script* utilizado para gerar um documento RDF. A esse *script* foram atribuídos os argumentos para acesso à base de dados, `-u`, indicando o usuário; `-p`, para a senha; `-f`, determinando o formato do documento e `-j`, para a URL do banco de dados. Antes do sinal ">" foi indicado o arquivo R2RML customizado a ser interpretado pelo D2RQ e após o sinal ">" foi posto o nome do arquivo de saída.

Figura 18 – Script dump-rdf.

```
dump-rdf -u $usuario_bd$ -p $senha_bd$ -f
$formato_rdf$ -j $jdbc:url_bd$ $arquivo_r2rml$ >
$arquivo_rdf_saida$
```

Fonte: Elaborado pelo autor.

4.5 Integração dos dados

Usou-se o *framework SILK Link Specification Language* (SILK) para a descoberta de *links* semânticos. O SILK é uma linguagem de especificação de *links* no padrão XML para identificar os relacionamentos semânticos entre entidades dentro dos arquivos RDF gerados na transformação dos dados (VOLZ et al., 2009).

O Código 4.1 apresenta a estrutura básica da especificação de *links* semânticos entre as fontes SIM e SINASC usadas na construção do modelo proposto. Entre as linhas 2 e 6 denominou-se todos os prefixos que referenciam as URI's das fontes de dados. Para tal, usou-se nas *tags* `<Prefix>` os atributos "id" para identificar unicamente uma fonte e "namespace" para sinalizar o endereço URI da fonte de dados. Nas linhas 8 a 15 foram configurados os `<DataSources>` especificando o arquivo e o formato de cada fonte de dados. Nesse caso, os arquivos especificados foram: "sim.nt" e "sinasc.nt", ambos com o formato *N-Triple*.

```
1  <Silk>
2    <Prefixes>
3      <Prefix id="rdfs"namespace="http://www.w3.org/2000/01/rdf-schema#"/>
4      <Prefix id="sim" namespace="http://gissa.org/ontology/sim#"/>
5      <Prefix id="sinasc" namespace="http://gissa.org/ontology/sinasc#"/>
6    </Prefixes>
7    <DataSources>
8      <DataSource id="sim" type="file">
9        <Param name="file" value="sim.nt"/>
10       <Param name="format" value="N-TRIPLE"/>
11     </DataSource>
12     <DataSource id="sinasc" type="file">
13       <Param name="file" value="sinasc.nt"/>
14       <Param name="format" value="N-TRIPLE"/>
15     </DataSource>
16   </DataSources>
17   <Interlinks>
18     <Interlink id="pessoa">
19       <LinkType>owl:sameAs</LinkType>
20       <SourceDataset dataSource="sim" var="a">
21         <RestrictTo>
22           ?a ?p ?v .
23         </RestrictTo>
24       </SourceDataset>
25       <TargetDataset dataSource="sinasc" var="b">
26         <RestrictTo>
```

```
27            ?a ?p ?v .
28          </RestrictTo>
29        </TargetDataset>
30        <LinkageRule linkType="owl:sameAs">
31          <Aggregate id="unnamed_1" required="false" weight="1" type="min">
32            <Compare metric="levenshteinDistance" threshold="0.0" index="true">
33              <TransformInput id="unnamed_2" function="lowerCase">
34                <Input path="?a/sim:numerodn"/>
35              </TransformInput>
36              <TransformInput id="unnamed_3" function="lowerCase">
37                <Input path="?b/sinasc:numerodn"/>
38              </TransformInput>
39            </Compare>
40          </Aggregate>
41        </LinkageRule>
42        <Filter/>
43      </Interlink>
44    </Interlinks>
45    <Outputs>
46      <Dataset id="gissa" type="file">
47        <Param name="file" value="linkGissa.nt"/>
48        <Param name="format" value="N-TRIPLE"/>
49      </Dataset>
50    </Outputs>
51  </Silk>
```

Código 4.1 – Estrutura Silk.

Configurou-se também, na linha 19, a *tag* `<LinkType>` com o valor *owl:sameAs*. Esse valor diz que um recurso (um objeto) em uma fonte de dados alvo é, semanticamente, o mesmo recurso em uma fonte origem. Entre as linhas 20 e 29 definiu-se a fonte de dados de origem e a fonte de dados de destino, configurando-se, respectivamente as tags *tags* `<SourceDataset>` e `<TargetDataset>` com os "ids" definidos no `<Prefix>`.

Estabeleceu-se ainda, dentro da *tag* `<LinkageRule>`, na linha 32, as regras de comparação por meio da propriedade `<Compare metric="levenshteiDistance" threshold="1">`, verificando as entradas `<Input path="?a/sim:numerodn"/>` e `<Input path="?b/sinasc:numerodn"/>`. Por fim, nas linhas 45 a 50, configurou-se os parâmetros `<Param name= "format" value= "ntriples"/>` e `<Param name= "file" value= "linkGissa.nt"/>` para definir o nome e formato de saída do arquivo. Nesse caso o nome do arquivo é "linkGissa" e o formato é ".nt" (N-triple).

Nesse processo de integração de dados existe a etapa de fusão dos dados. Para realizar essa fusão utilizou-se o *framework* de avaliação de qualidade *Linked Data* e de fusão chamado SIEVE[9] (MENDES; MÜHLEISEN; BIZER, 2012). Por meio de um arquivo SIEVE, especificaram-se os requisitos para definir a qualidade e os métodos para resolver conflitos permitindo a fusão dos dados.

[9] http://sieve.wbsg.de/

```
1  <Sieve xmlns="http://www4.wiwiss.fu-berlin.de/ldif/">
2      <Prefixes></Prefixes>
3      <QualityAssessment name="Recent and Reputable is Best">
4          <AssessmentMetric id="sieve:recency"></AssessmentMetric>
5          <AssessmentMetric id="sieve:reputation"></AssessmentMetric>
6      </QualityAssessment>
7
8      <Fusion name="Fusion strategy" description="The idea is to use values from datasets to
           improve the quality of data.">
9          <Class name="gissa:Pessoa">
10             <Property name="gissa:cns">
11                 <FusionFunction class="KeepFirst" metric="sieve:recency"/>
12             </Property>
13             <Property name="gissa:nome">
14                 <FusionFunction class="KeepFirst" metric="sieve:recency"/>
15             </Property>
16             <Property name="gissa:dataNascimento">
17                 <FusionFunction class="KeepFirst" metric="sieve:reputation"/>
18             </Property>
19         </Class>
20     </Fusion>
21 </Sieve>
```

Código 4.2 – Estrutura SIEVE.

O Código 4.2 mostra a estrutura básica de um arquivo SIEVE usado na construção do modelo proposto. Na linha 2, definem-se os prefixos, do mesmo modo que foram definidos os prefixos do arquivo SILK. Nas linhas 3 a 6 estabelecem-se as métricas para a qualidade dos dados. Por último, nas linhas 8 a 20 estipulam-se as regras de fusão dos dados. Cada propriedade (`<Property>`) pode ter um tipo de função de fusão (`<FusionFunction>`) diferente de acordo com o tipo de dado. O resultado dessa fusão apresenta-se em forma de arquivo do tipo *N-Triple*, uma serialização RDF.

4.6 Cálculo do Risco

A partir dos riscos identificados nas heurísticas dos especialistas em saúde, calcula-se a probabilidade de óbito materno e infantil sobre dois principais domínios: clínico e social.

4.6.1 Heurísticas

Heurística é um termo que tem origem no grego antigo e significa “encontrar”, “descobrir”, “inventar”, “obter”. Assim, a capacidade humana de descobrir, inventar e resolver problemas mediante experiência própria ou observada é denominada heurística (WIKIPéDIA, 2019).

Com base nas experiências e pesquisas dos especialistas em saúde materno-infantil, foi possível identificar atributos que representam risco de morte para um indi-

víduo, mãe e bebê, durante o período gestacional e puerpério. Tais identificações deu origem às heurísticas utilizadas no trabalho aqui proposto.

Nessas heurísticas, alguns riscos clínicos e/ou sociais presentes no indivíduo/gestante/mãe influenciam diretamente no cálculo do risco de óbito do indivíduo/descendente/bebê. Além disso, essas heurísticas preveem os riscos da gestação e parto. Ambos também impactam diretamente na vida do bebê. Os riscos de gestação e parto podem ser: parto foi induzido; gestação única, dupla ou múltipla; parto cesariano e parto realizado fora de estabelecimento de saúde. Paralelamente, existem os riscos sociais. Estes sempre pertencentes à gestante/mãe.

Com o intuito de uma melhor visualização das heurísticas, todos os riscos foram expostos em tabelas. A Tabela 7 mostra todos os possíveis riscos clínicos maternos e seus respectivos pesos (definidos nas heurísticas) que uma mãe pode apresentar. Da mesma forma, enquanto a Tabela 8 exibe os riscos sociais maternos, a Tabela 9 apresenta os riscos infantis. A soma de todos os riscos totalizam 53 atributos entre clínico, social e evento. Os valores de zero a vinte (0 – 20) para os pesos assim como os percentuais indicando os intervalos das faixas de risco foram determinados pelos especialistas.

Tabela 7 – Riscos Clínicos Materno.

Risco	Peso	Risco	Peso
Consumir álcool	15	Fumar	10
Parto cesariano	10	Ter diabetes	20
Portador de doença cardíaca	20	Portador de doença renal	20
Sobrevivente de AVC	20	Sobrevivente de infarto	20
Ter câncer	20	Ter hanseníase	10
Ter hipertensão	20	Ter mais de 35 anos	15
Ter tuberculose	10	Gestação com menos de 3 pré-natais	10
Usar drogas	20	Portador de H1N1	10
Soropositivo	10		

Fonte: Elaborado pelo autor.

Tabela 8 – Riscos Sociais Materno .

Risco	Peso	Risco	Peso
Baixa Renda	15	Baixa escolaridade	15
Situação de trabalho: Desempregado	10	Situação de trabalho: Não trabalha	10
Está em situação de rua	20	Faz uso de drogas	20
É mãe solteira	10	Vítima de violência	15

Fonte: Elaborado pelo autor.

Tabela 9 – Riscos Infantil.

Risco	Peso	Risco	Peso
Prematuro: menos de 22 semanas	20	Prematuro: de 22 a 27 semanas	20
Prematuro: de 28 a 31 semanas	15	Prematuro: de 32 a 36 semanas	10
Baixo peso ao nascer	15	Gestação menos de 3 pré-natais	20
Apgar baixo de 4 a 6	15	Apgar crítico de 0 a 3	20
Parto fora de estab. de saúde	20	Realizou uma cesárea	5
Realizou duas cesáreas	5	Realizou três cesáreas	10
Realizou mais de três cesáreas	15	Gestação dupla	10
Gestação múltipla	15	Parto induzido	10
Parto com cesariana marcadas	10	Teve uma perda fetal	5
Teve duas perdas fetais	10	Teve mais de duas perdas fetais	20
Parto cesariano	10	Portador de sífilis	15
Portador de sífilis congênita	15	Portador de difteria	20
Portador de rubéola congênita	15	Possui tétano neonatal	20
Possui rotavírus	5	Deficiência intelectual cognitiva	15

Fonte: Elaborado pelo autor.

4.6.2 Cálculo da Probabilidade

Uma mãe ou um bebê podem ser classificados em baixo, médio ou alto risco, considerando percentual de risco calculado para cada indivíduo. As faixas de valores para classificar um indivíduo são:

1. De 0% a 10% corresponde a baixo risco;
2. De 11% a 20% o indivíduo é classificado em risco intermediário;
3. Acima de 20% é classificado em alto risco.

Essas faixas levam em consideração a quantidade de riscos existentes em um indivíduo, visto que para se atingir o critério de alto risco são necessários vários riscos presentes em um indivíduo. Vale ressaltar que essas faixas de valores também são frutos das heurísticas dos especialistas.

Em seguida, é descrito como é realizado o cálculo do fator de risco de óbito materno e infantil nos domínios clínico e social. Primeiramente, o cálculo do percentual de óbito materno considerando os fatores sociais é dado por:

Definição 1 $\forall$ *mãe* m, $\exists$ *um conjunto de riscos* R_m *que é um subconjunto de* TR_m. TR_m *é o conjunto de todos os riscos que* m *pode apresentar. A quantidade de riscos em* R_m *é representada por* k *cujo* $0 \leq k \leq n$.

Definição 2 *Cada risco* $r_i \in TR_m$ *tem um peso* $0 \leq w \leq 20$, *onde* $0 < i \leq n$ *e* n *é o número de riscos em* TR_m.

Desta forma, encontra-se o total do risco social de m pela Equação 4.1.

$$RiscoSocialDaMae(m) = \sum_{i=1}^{k} f(r_i),\ r_i \in R_m \tag{4.1}$$

$$f(r_i) = \left\{ \begin{array}{ll} PesoDoRiscoSocial(r_i), & se\ r_i \in TR_m \\ 0, & caso\ contrário \end{array} \right\}$$

Para obter o máximo de pesos dos riscos que poderiam estar presentes em uma mãe, efetua-se a Equação 4.2.

$$MaxRiscoSocialMaterno() = \sum_{i=1}^{n} PesoDoRiscoSocial(r_i),\ r_i \in TR_m \tag{4.2}$$

A probabilidade de óbito materno considerando os riscos sociais é dada pela Equação 4.3.

$$ProbabObitoRiscoSocial(m) = \frac{RiscoSocialDaMae(m)}{MaxRiscoSocialMaterno()} \tag{4.3}$$

O segundo passo é realizar os mesmos cálculos para encontrar a probabilidade de óbito materno, observando os riscos de fatores clínicos. Nesse sentido, encontra-se o total do risco clínico de m pela Equação 4.1, porém, substituindo a função $PesoDoRiscoSocial(r_i)$ pela função $PesoDoRiscoClinico(r_i)$. Além disso, atribui-se o resultado obtido ao $RiscoClinicoDaMae(m)$. O $MaxRiscoClinicoMaterno()$ é encontrado pela Equação 4.2 trocando no somatório a função $PesoDoRiscoSocial(r_i)$ pela função $PesoDoRiscoClinico(r_i)$. Assim, a probabilidade do risco de óbito materno levando em consideração os fatores clínicos se dá pela razão $\frac{RiscoClinicoDaMae(m)}{MaxRiscoClinicoMaterno()}$.

Por último, para calcular o risco de óbito infantil considera-se a influência de fatores de risco da mãe do bebê e dos eventos que o envolvem diretamente. O resultado do $RiscosDaMae(m)$ é a fusão entre $RiscoSocialDaMae(m)$ e $RiscoClinicoDaMae(m)$, tal como, o $MaxRiscoMaterno()$ é a soma do $MaxRiscoSocialMaterno()$ e $MaxRisco-ClinicoMaterno()$. O valor do $RiscosDosEventos(m)$ compreende a soma dos pesos dos riscos de parto e gestação existentes em m e o $MaxRiscoEventos()$ representa a soma dos pesos de todos os riscos de eventos deste modelo.

Para toda criança c, o $RiscosDaCrianca(c)$ é obtido ao realizar o somatório dos pesos dos riscos presente em c. O valor de $MaxRiscoInfantil()$ é o total dos pesos de todos os riscos clínicos que uma criança pode apresentar. Dessa maneira, a probabilidade P de óbito infantil é encontrada pela Equação 4.4.

$$P = \frac{(RiscosDaCrianca(c) + RiscosDaMae(m) + RiscosDosEventos(m))}{(MaxRiscoInfantil() + MaxRiscoMaterno() + MaxRiscoEventos())} \quad (4.4)$$

Todos os cálculos deste modelo são efetuados na Camada 5 da sua arquitetura. Para realizar consultas sobre os dados integrados e exibi-los, usa-se a API Jena, um *framework* desenvolvido em linguagem Java para construir aplicações para Web Semântica e *Linked Data*. A API Jena usa protocolos SPARQL e o vocabulário da O_D nos *scripts* de consultas.

5 RESULTADOS

Este capítulo apresenta um experimento sobre as heurísticas desenvolvidas pelos especialistas em saúde materno e infantil e os resultados alcançados com o modelo desenvolvido.

Para a realização do experimento analisou-se os dados de cinco mães e cinco bebês, obtidos do DATASUS. Das cinco mães, constatou-se que cinco eram fumantes, três tiveram uma gestação com menos de três pré-natais, uma consumia álcool e uma já realizou parto cesariano. Esses dados estão devidamente apresentados na Tabela 10. Essa tabela contém os riscos que estavam presentes nas mães no momento do experimento. Percebe-se que algumas mães apresentaram mais de um risco clínico.

Tomando por base os dados da Tabela 7, o "máximo de peso" possível para o risco clínico das cinco mães selecionadas para o experimento é 1300. Esse valor é encontrado ao realizar a multiplicação da quantidade de mães que estão em avaliação pela soma dos pesos de todos os riscos clínicos maternos. Já o valor "peso acumulado" das cinco mães em avaliação é 105. Esse valor é obtido somando-se os valores da coluna "Peso x Quantidade" da Tabela 10. Assim, o percentual de risco clínico materno é dado pela razão $\frac{peso\ acumulado}{maximo\ de\ peso}$, cujo resultado nesse experimento fora 8,08%. Então, para esse grupo de 5 mulheres a classificação equivaleu a baixo risco.

Tabela 10 – Experimento - Risco Clínico Materno.

Risco	Peso	Quantidade	Peso x Quantidade
Consumir álcool	15	1	15
Fumar	10	5	50
Parto cesariano	10	1	10
Gestação com menos de 3 pré-natais	10	3	30
Quantidade de mães	5	Total dos pesos	105

Fonte: Elaborado pelo autor.

O experimento também foi realizado sobre a perspectiva do risco social materno. Para tanto, os mesmos passos para obter o percentual do risco clínico materno foram seguidos a fim de encontrar os valores "peso acumulado" e "máximo de peso" para o risco social materno. Dessa vez, observando os dados das Tabelas 8 e 11, o resultado configurou-se em médio risco, já que o experimento fixou uma razão de 10,43%. A Tabela 11 exibe os riscos sociais presentes nas mães do experimento. Das cinco mães, duas apresentaram o risco Baixa Renda, duas portavam o risco Baixa

Escolaridade e uma não apresentou risco.

Tabela 11 – Experimento - Risco Social Materno.

Risco	Peso	Quantidade	Peso x Quantidade
Baixa Renda	15	2	30
Baixa Escolaridade	15	2	30
Quantidade de mães	5	Total dos pesos	60

Fonte: Elaborado pelo autor.

Tabela 12 – Experimento do Risco Clínico com 5 bebês.

Risco	Peso	Quantidade	Peso x Quantidade
Prematuro: menos de 22 semanas	20	2	40
Gestação com menos de 3 pré-natais	15	3	45
Baixo peso ao nascer	20	3	60
Apgar crítico: 0 a 3	20	1	20
Parto cesariano	10	1	10
Portador de sífilis	15	2	30
Total			205

Fonte: Elaborado pelo autor.

Utilizando a mesma forma de cálculo das duas etapas anteriores e levando em consideração os dados das Tabelas 9 e 12, o experimento chegou ao resultado 14,64% para o risco clínico infantil. Esse percentual representa um risco de óbito de nível médio. Em resumo, a Tabela 13 mostra a probabilidade de óbito das mães e bebês de acordo com os riscos apresentados em relação a todos os riscos possíveis.

Tabela 13 – Resumo do Experimento.

Nível do Risco	Clínico Materno	Social Materno	Clínico Infantil
Baixo <= 10%	8,08%	-	-
10% < Médio <= 20%	-	10,43%	14,64%
Alto > 20%	-	-	-

Fonte: Elaborado pelo autor.

Além do experimento realizado, o modelo de integração de dados e cálculo de probabilidade de risco de óbito materno e infantil proposto neste trabalho obteve outros resultados significativos, tanto acadêmicos quanto profissional.

O primeiro resultado alcançado foi uma publicação de um artigo científico intitulado *"Using Linked Data in the Data Integration for Maternal and Infant Death Risk of the SUS in the GISSA Project"* cuja abordagem chave foi aplicar o modelo proposto neste trabalho (FREITAS et al., 2017). Aprovado no XXIII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia[1]), esse artigo foi apresentado em outubro de 2017

[1] https://webmedia.org.br/2017/pt/artigos-aceitos/

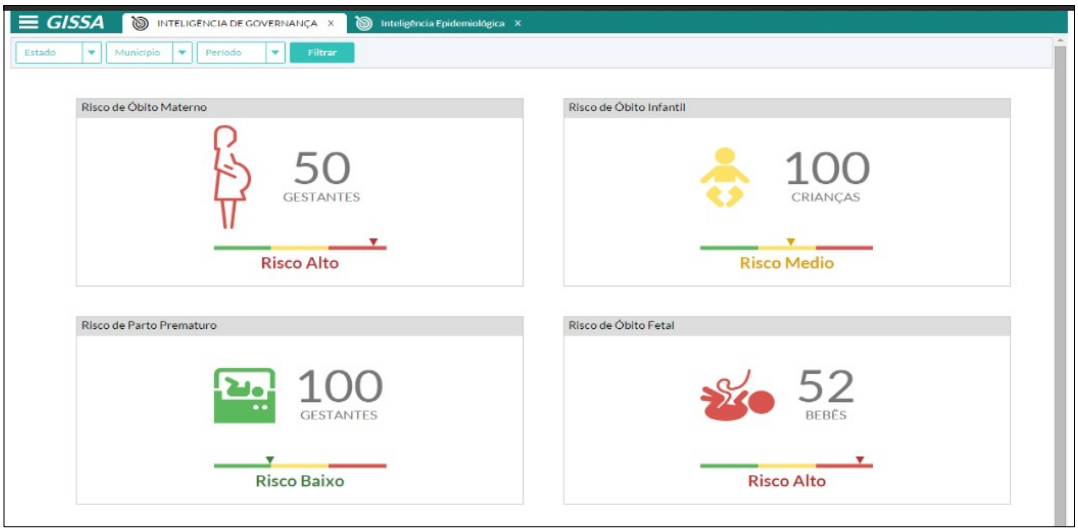
na cidade de Gramado, Rio Grande do Sul, Brasil. Pode-se encontrá-lo na plataforma *ACM Digital Library*[2] por meio do *link* https://doi.org/10.1145/3126858.3131606.

Uma segunda contribuição é a aplicação deste modelo como parte integrante do módulo de inteligência da plataforma GISSA. Por se tratar de um sistema de saúde que auxilia gestores a melhorar a vida das pessoas, os benefícios do GISSA, em sua essência, retribuem ou revertem todos os investimentos realizados em bolsas de pesquisa. Além disso, aplicado no âmbito empresarial, o GISSA está sendo comercializado pela empresa cearense Avicena[3]. Tal feito mostra a relevância da contribuição deste projeto.

A Figura 19 mostra um *dashboard* do GISSA exibindo os níveis dos riscos de óbito materno e infantil. Na parte superior esquerda da Figura 19, exibe-se um quantitativo de 50 gestantes apresentando um alto risco. Na parte superior direita da mesma figura, exibe-se uma quantidade de 100 crianças indicado risco de nível médio. Ainda na mesma figura, na parte inferior esquerda, tem-se o total de 100 gestantes em de risco de parto prematuro que apresentam risco de nível baixo. Por fim, na parte inferior direita, tem-se 52 bebês em risco de óbito fetal a risco de nível alto.

Essas informações são obtidas por meio do modelo aqui apresentado funcionando dentro do GISSA. Assim, esse *dashboard* proporciona ao gestor uma visão mais efetiva da situação de saúde materno-infantil do seu município.

Figura 19 – *Dashboard* GISSA.



Fonte: GISSA.

[2] https://dl.acm.org/
[3] https://avicena.in/

6 CONCLUSÃO

Este trabalho apresentou um modelo baseado em Ontologia e *Linked Data* que integra bases de dados relacionais isoladas que possuem esquemas heterogêneos, aplicando as metodologias desenvolvidas em (LOPES; VIDAL; OLIVEIRA, 2016b; LOPES; VIDAL; OLIVEIRA, 2016a).

O modelo desenvolvido realizou a integração de dados de saúde do SUS sob a óptica do problema do óbito materno-infantil. Além disso, forneceu a probabilidade de risco de óbito materno-infantil. Atualmente, este modelo está implantado no módulo de inteligência do GISSA sendo uma das principais engrenagens no suporte à governança inteligente.

Contudo, esse modelo não é o único mecanismo para obter a probabilidade de óbito materno e infantil. O trabalho de (SILVA et al., 2017), por exemplo, foi desenvolvido com o mesmo objetivo de predição, porém, utilizando Mineração de Dados para aprender sobre um conjunto de dados de 16 atributos oriundos da integração das bases de dados SIM e SINASC. Paralelamente, esse trabalho também contribui no módulo de inteligência do projeto GISSA provendo alertas para risco de óbito materno e infantil.

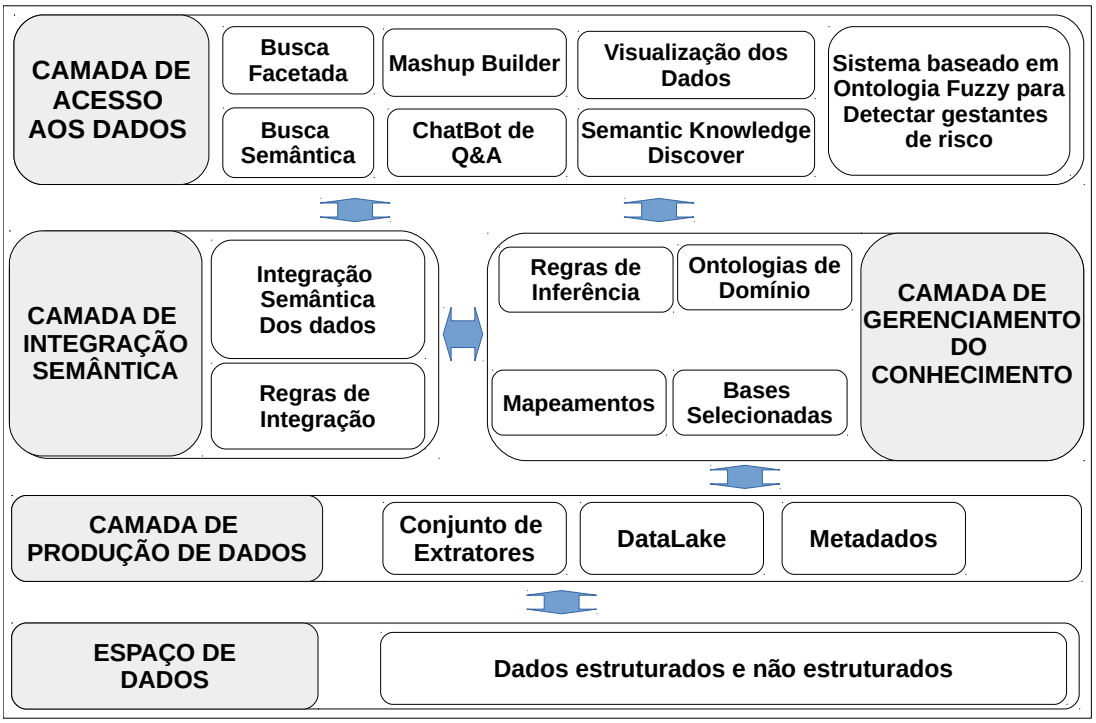
Mesmo com essas contribuições, o GISSA ainda busca a definição de seu modelo de inteligência. Nesse sentido, este trabalho e o trabalho com Mineração de Dados de (SILVA et al., 2017) são relevantes para essa definição. Assim, a expectativa é que o modelo final de inteligência do *framework* GISSA seja espelhado em um tipo híbrido cujo mecanismo determinante seja o apresentando neste trabalho.

Outra contribuição relevante deste trabalho é na construção de um Portal Semântico que está sendo realizada pelo Grupo de Ontologias do projeto GISSA e liderada pela professora Vânia Vidal da Universidade Federal do Ceará. A Figura 20 ilustra a arquitetura proposta para tal Portal Semântico, cujo modelo aqui desenvolvido colabora diretamente na Camada de Integração Semântica e na Camada de Gerenciamento de Conhecimento.

Durante o desenvolvimento deste trabalho surgiram alguns desafios. Um dos mais árduos foi decifrar e compreender as nomenclaturas dos atributos das bases de dados de saúde para realizar os mapeamentos e criar a Ontologia de Risco. Essas tarefas só foram realizadas graças às expertises dos especialistas em saúde.

Outro desafio foi o acesso aos dados para realizar o experimento. Os dados maternos e infantis obtidos para efetuar o experimento vieram em grupos e não individuais. Dessa forma, de acordo com o experimento realizado, se uma mãe ou bebê

Figura 20 – Arquitetura do Portal Semântico do GISSA.



Fonte: Projeto GISSA — Grupo de Ontologias.

apresentar vários riscos, o grupo ainda poderia ser classificado como baixo risco, porém, este indivíduo poderia estar na classificação de alto risco e, assim, não receberia a atenção adequada.

Apesar da probabilidade do risco de óbito materno-infantil ter sido validada pelos especialistas em saúde, ainda não existe uma validação matemática para os resultados da O_{Risco}. Nesse sentido, tem-se como proposta de trabalho futuro realizar uma validação matemática da O_{Risco} fazendo avaliações individuais. Essa validação assemelha-se à Matriz de Confusão na validação de algoritmos de aprendizagem de máquina.

REFERÊNCIAS

ALMEIDA, M. B. Uma abordagem integrada sobre ontologias: Ciência da informação, ciência da computação e filosofia. *Perspectivas em Ciência da Informação*, v. 19, n. 3, p. 242–258, 2014. Citado na página 21.

ALMEIDA, M. B.; BAX, M. P. **Uma visão geral sobre ontologias:** pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Ciência da Informação, Brasília*, SciELO Brasil, v. 32, n. 3, p. 7–20, 2003. Citado na página 21.

ANDRADE, L. O. Monteiro de. **Inteligência de Governança para apoio à Tomada de Decisão**. *Ciência & Saúde Coletiva*, Associação Brasileira de Pós-Graduação em Saúde Coletiva, v. 17, n. 4, 2012. Citado na página 15.

ANGULO-TUESTA, A.; SANTOS, L. M. P.; NATALIZI, D. A. **Impact of health research on advances in knowledge, research capacity-building and evidence-informed policies:** a case study on maternal mortality and morbidity in brazil. *Sao Paulo Medical Journal*, SciELO Brasil, v. 134, n. 2, p. 153–162, 2016. Citado na página 17.

BERNERS-LEE, T. ***Linked Data***. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 07 dez. 2018. Citado 2 vezes nas páginas 28 e 29.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. **Linked data:** the story so far. In: *Semantic services, interoperability and web applications: emerging concepts*. [S.l.]: IGI Global, 2011. p. 205–227. Citado na página 29.

BORST, W. N. **Construction of engineering ontologies for knowledge sharing and reuse**. Centre for Telematics and Information Technology (CTIT), Netherlands, 1997. Citado na página 21.

BRASIL. ***O DATASUS***. [S.l.]: Ministério da Saúde, 2008. Disponível em: <http://www2.datasus.gov.br/DATASUS/index.php?area=01>. Acesso em: 02 dez. 2018. Citado na página 16.

BRASIL. ***Rede Cegonha***. 2018. Disponível em: <http://dab.saude.gov.br/portaldab/ape_redecegonha.php>. Acesso em: 07 dez. 2018. Citado na página 17.

CAVALCANTE, G. M. L. ***MAURA:*** *Um Framework baseado em Mediador Semântico para construção eficiente de Linked Data Mashups*. Dissertação (Mestrado) — Instituto Federal do Ceará, Departamento de Telemática, PPGCC, 4 2017. Citado na página 31.

FARINELLI, F.; ALMEIDA, M. **Interoperabilidade semântica em sistemas de informação de saúde por meio de ontologias formais e informais:** um estudo da norma openehr. *XVII Encontro Nacional de Pesquisa em Ciência da Informação*, v. 17, n. 1, 2014. Citado na página 16.

FEOFILOFF, P. ***O que é um grafo?*** 2018. Disponível em: <https://www.ime.usp.br/ pf/algoritmos_em_grafos/aulas/grafos.html>. Acesso em: 13 mar 2019. Citado na página 24.

FREITAS, R. et al. **Using Linked Data in the Data Integration for Maternal and Infant Death Risk of the SUS in the GISSA Project**. In: *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*. New York, NY, USA: ACM, 2017. (WebMedia '17), p. 193–196. ISBN 978-1-4503-5096-9. Disponível em: <http://doi.acm.org/10.1145/3126858.3131606>. Citado na página 55.

GRUBER, T. ***What is an Ontology***. 1993. Disponível em: <https://pdfs.semanticscholar.org/08ca/030f827f38bf1ac17b5886adefd3cc2d3264.pdf>. Acesso em: 03 dez. 2018. Citado na página 21.

HEINZLE, R.; GAUTHIER, F. A. O.; FIALHO, F. A. P. **Semântica nos sistemas de apoio a decisão:** o estado da arte. *Revista da UNIFEBE*, v. 1, n. 8, p. 225–248, 2017. Citado na página 20.

HU, J. et al. **A Linked Data Based Decision Support System for Cancer Treatment**. In: *2014 Enterprise Systems Conference*. [S.l.: s.n.], 2014. p. 39–44. Citado 2 vezes nas páginas 17 e 37.

IBGE, B. ***Brasil em Síntese***. 2018. Disponível em: <https://brasilemsintese.ibge.gov.br/populacao/taxas-de-mortalidade-infantil.html>. Acesso em: 24 nov. 2018. Citado na página 17.

ISOTANI, S.; BITTENCOURT, I. I. ***Dados Abertos Conectados***. São Paulo: Novatec Editora, 2015. Citado na página 33.

JUNG, Y.; YOON, Y. I. **Data integration for clinical decision support**. In: IEEE. *Ubiquitous and Future Networks (ICUFN), 2016 Eighth International Conference on*. [S.l.], 2016. p. 164–166. Citado na página 36.

KOZáK, J. et al. **Linked Open Data for Healthcare Professionals**. In: *Proceedings of International Conference on Information Integration and Web-based Applications & Services*. New York, NY, USA: ACM, 2013. (IIWAS '13), p. 400:400–400:409. ISBN 978-1-4503-2113-6. Disponível em: <http://doi.acm.org/10.1145/2539150-.2539195>. Citado na página 36.

LAUFER, C. ***Guia de Web Semântica***. São Paulo: Novatec Editora, 2015. Citado 3 vezes nas páginas 26, 31 e 32.

LOPES, G.; VIDAL, V.; OLIVEIRA, M. **A Framework for Creation of Linked Data Mashups:** a case study on healthcare. In: *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web*. New York, NY, USA: ACM, 2016. (Webmedia '16), p. 327–330. ISBN 978-1-4503-4512-5. Disponível em: <http://doi.acm.org/10.1145/2976796.2988213>. Citado 2 vezes nas páginas 40 e 57.

LOPES, G.; VIDAL, V. M. P.; OLIVEIRA, M. **Construção de Linked Data Mashup para Integração de Dados da Saúde Pública**. In: *SBBD*. [S.l.: s.n.], 2016. p. 145–150. Citado 2 vezes nas páginas 40 e 57.

MAGALHÃES, R. P. et al. **Desenvolvimento de Linked Data Mashups com o uso de LIDMS**. In: *SBBD (Short Papers)*. [S.l.: s.n.], 2012. p. 217–224. Citado 3 vezes nas páginas 21, 22 e 38.

MENDES, P. N.; MÜHLEISEN, H.; BIZER, C. **Sieve:** linked data quality assessment and fusion. In: ACM. *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. [S.l.], 2012. p. 116–123. Citado na página 48.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Ontologias:** conceitos, usos, tipos, metodologias, ferramentas e linguagens. *Universidade Federal de Goiás*, 2007. Citado 2 vezes nas páginas 21 e 22.

OLIVEIRA, A. M. B. de. **Cloud LARIISA:** a context-aware framework for a public health environment based on cloud computing concept. *Conexões-Ciência e Tecnologia*, v. 8, n. 3, 2014. Citado na página 15.

OLIVEIRA, M. ***Projeto GISSA:*** *Meta Física 3 – Atividade 3.1 Definir Modelo de Inteligência de Gestão na Saúde*. 2015. Disponível em: <https://amauroboliveira.files.wordpress.com/2015/11/gi2s_ia_gissa_plano_trabalho.pdf>. Acesso em: 11 nov. 2018. Citado 2 vezes nas páginas 34 e 35.

OLIVEIRA, M. et al. **A context-aware framework for health care governance decision-making systems:** a model based on the brazilian digital tv. In: IEEE. *World of Wireless Mobile and Multimedia Networks (WoWMoM), 2010 IEEE International Symposium on a*. [S.l.], 2010. p. 1–6. Citado na página 17.

PIERRO, B. de. ***Dados sobre saúde precisam de integração***. 2011. Disponível em: <http://www.advivo.com.br/materia-artigo/dados-sobre-saude-precisam-de-integracao>. Acesso em: 05 dez. 2018. Citado na página 15.

PRUD'HOMMEAUX, E.; HARRIS, S.; SEABORNE, A. **SPARQL 1.1 Query Language**. *W3C Recommendation*, 2013. Citado na página 31.

REZENDE, S. O. ***Sistemas inteligentes:*** *fundamentos e aplicações*. [S.l.]: Editora Manole Ltda, 2003. Citado na página 17.

SENA, O.; MAURO, O. **A Data Integration Model for a Decision Making System in Health**. IEEE Healthcom, 2014. Citado 3 vezes nas páginas 15, 37 e 38.

SILVA, C. et al. **LAÍS, um Analisador Baseado em Classificadores para a Geração de Alertas Inteligentes em Saúde**. In: . [S.l.]: XXXV Simpósio Brasileiro de Redes de Computadores (SBRC) - I Workshop de Computação Urbana (CoUrb), 2017. p. 1–13. Citado 3 vezes nas páginas 17, 38 e 57.

SOARES, D. A.; ANDRADE, S. M. d.; CAMPOS, J. J. B. d. **Epidemiologia e indicadores de saúde**. *Bases da saúde coletiva. Londrina: Ed. UEL*, p. 183–210, 2001. Citado na página 15.

SPRAGUE, R. H.; WATSON, H. J.; JR, R. H. S. ***Sistema de Apoio à Decisão:*** *colocando a teoria em prática*. [S.l.: s.n.], 1991. Citado na página 20.

VASCONCELOS, A. C. T. **Sistema inteligente de apoio à decisão:** um estudo sobre algoritmo genético e lógica fuzzy. Niterói, 2018. Citado na página 17.

VOLZ, J. et al. **Silk-A Link Discovery Framework for the Web of Data**. *LDOW*, v. 538, 2009. Citado na página 47.

W3C. ***RDF - Resource Description Framework***. 2014. Disponível em: <https://www.w3.org/RDF/>. Acesso em: 07 dez. 2018. Citado na página 24.

W3C. ***RDF 1.1 Concepts and Abstract Syntax***. 2014. Disponível em: <https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225>. Acesso em: 09 dez. 2018. Citado na página 24.

W3C. ***About W3C***. 2018. Disponível em: <https://www.w3.org/Consortium/>. Acesso em: 09 dez. 2018. Citado na página 24.

WIKIPéDIA. ***Heurística***. 2019. Disponível em: <https://pt.wikipedia.org/wiki/Heurística>. Acesso em: 13 mar 2019. Citado na página 49.