UNIVERSIDADE ESTADUAL DO CEARÁ
CENTRO DE CIÊNCIAS E TECNOLOGIA - CCT
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
PRÓ-REITORIA DE PESQUISA, DE INOVAÇÃO E PÓS-GRADUAÇÃO – PRPI
MESTRADO PROFISSIONAL EM COMPUTAÇÃO APLICADA

ODARA SENA DOS SANTOS FEITOSA

UM PROCESSO DE INTEGRAÇÃO DE DADOS PARA UM SISTEMA INTELIGENTE DE SAÚDE

FORTALEZA – CEARÁ

2015

ODARA SENA DOS SANTOS FEITOSA

UM PROCESSO DE INTEGRAÇÃO DE DADOS PARA UM SISTEMA INTELIGENTE DE SAÚDE

Dissertação submetida à Coordenação do Curso de Mestrado Profissional em Computação Aplicada da Universidade Estadual do Ceará e do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial para a obtenção do título de mestre em computação aplicada.

Orientador: Prof. Dr. Antônio Mauro Barbosa de Oliveira.

Coorientador: Prof. Dr. Anilton Salles Garcia.

FORTALEZA – CEARÁ

2015

Dados Internacionais de Catalogação na Publicação
Universidade Estadual do Ceará
Sistemas de Bibliotecas

Feitosa, Odara Sena dos Santos
Um Processo de Integração de Dados para um Sistema Inteligente de Saúde [recurso eletrônico] / Odara Sena dos Santos Feitosa. - 2015.
1 CD-ROM: il.; 4 3⁄4 pol.

CD-ROM contendo o arquivo no formato PDF do trabalho acadêmico com 110 folhas, acondicionado em caixa de DVD Slim (19 x 14 cm x 7 mm).

Dissertação (mestrado profissional) – Universidade Estadual do Ceará, Centro de Ciências e Tecnologia, Mestrado Profissional em Computação Aplicada, Fortaleza, 2015.
Área de concentração: Redes de Computadores
Orientação: Prof. Ph.D. Antônio Mauro Barbosa de Oliveira.
Coorientação: Prof. Dr. Anilton Salles Garcia.

1. Ontologias. 2. Web Semânntica. 3. *Linked Data*. 4. Integração de Dados. 5. Sistemas Inteligentes de Saúde. I. Título.

Odara Sena dos Santos Feitosa

UM PROCESSO DE INTEGRAÇÃO DE DADOS PARA UM SISTEMA INTELIGENTE DE SAÚDE

Dissertação apresentada ao Curso de Mestrado Profissional em Computação Aplicada da Universidade Estadual do Ceará, como requisito parcial para a obtenção do grau de Mestrado em Computação.

Defesa em: **15/05/2015**

BANCA EXAMINADORA

Antonio Mauro Barbosa de Oliveira, PD. DSc. (IFCE)
Presidente (Orientador)

Luiz Odorico Monteiro de Andrade, Dsc. (UFC)
Membro Externo

Marcos José Negreiros Gomes, PD. DSc. (UECE)
Membro Interno

Dedico este trabalho à minha família.

AGRADECIMENTOS

Primeiramente, à minha mãe Antoniêta, que mesmo não estando mais entre nós, fez um grande trabalho em me educar.
À minha Tia Antonilda e Vó Regina, que me tornaram a pessoa que sou hoje.
Ao meu pai, Jorge, que nunca deixou de acreditar no meu sucesso.
Ao meu esposo, Thiago, por toda a paciência, amor e dedicação.
Ao meu orientador, Prof. Ph.D. Mauro Oliveira, por nunca duvidar do meu trabalho e por ser hoje o meu espelho profissional e pessoal.
Ao Prof. Dr. Anilton Garcia, pelas orientações e disponibilidade, sempre.
À Deus, pelos belíssimos planos traçados para a minha vida.
À todos os meus familiares e amigos, em especial aos amigos feitos no MPCOMP, Evandro, Pablo e Reivel, pelo apoio e união.
E à toda equipe MPCOMP.

“Torna-te aquilo que és.”

(Friedrich Nietzsche)

RESUMO

O uso de Tecnologias da Informação e Comunicação (TIC's) aplicadas à saúde vem despontando nas últimas décadas, principalmente por viabilizar uma melhor gestão do conhecimento e interpretação das informações integradas, funções fundamentais para suporte à tomada de decisão. Ao encontro dessa tendência destaca-se o LARIISA, uma plataforma inteligente que objetiva auxiliar à tomada de decisão na área de saúde, utilizando conceitos de *context-aware*, ontologias e inteligências de governança de saúde para apoiar a realização da inferência pelo sistema. No intuito de ampliar o poder de inferência do LARIISA se faz necessário um enriquecimento da base de conhecimento do *framework,* que pode ser conseguido através da integração de dados relacionados à saúde já existentes. Dessa forma, este trabalho apresenta um processo de integração de dados para o LARIISA, capaz de lidar com fontes de dados heterogêneas, independentes e distribuídas, como, por exemplo, dados oriundos dos sistemas mantidos pelo Ministério da Saúde, ou por outras esferas do governo, resultando em uma base de conhecimento mais rica, aumentando assim seu poder de inferência. Os resultados deste trabalho são obtidos através da utilização dos conceitos de dados *linkados* (*Linked Data*) e ontologias, que se destacam no contexto da Web Semântica.

Palavras-chave: Ontologias. *Web* Semântica. *Linked Data*. Integração de Dados. Sistemas Inteligentes de Saúde.

ABSTRACT

The use of information and communication technologies (ICT's) applied to health has been rising in recent decades, primarily by enabling better management of knowledge and interpretation of the integrated information, fundamental functions to support decision-making. To meet this trend stands out the LARIISA, an intelligent platform that aims to assist decision making in the area of health, using concepts of context-aware, ontologies and intelligences of health governance to support the realization of the inference by the system. In order to enlarge the power of LARIISA inference is necessary to enrich the knowledge base of the framework, which can be achieved through the integration of existing health-related data. Thus, this work presents a data integration process for the LARIISA, capable of dealing with heterogeneous data sources, independent and distributed, such as data from the systems maintained by the Ministry of Health, or by other spheres of Government, resulting in a richer knowledge base, thereby increasing its power of inference. The results of this work are obtained through the use of linked data concepts (Linked Data) and ontologies, which stand out in the context of the Semantic Web.

Key words: *Ontology. Semantic Web. Linked Data. Data Integration. Intelligent Health Systems.*

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ANVISA	Agência Nacional de Vigilância Sanitária
APAC	Autorização de Procedimentos de Alta Complexidade
API	*Application Programming Interface*
ARiDa	*Advanced Research in Database*
CNES	Cadastro Nacional de Estabelecimentos de Saúde
CSV	*Comma-Separated Values*
DATASUS	Departamento de Informática do Sistema Único de Saúde
EHR	*Electronic Health Record*
ETL	*Extract, Transform, Load*
FUNCEME	Fundação Cearense de Meteorologia e Recursos Hídricos
HTML	*HyperText Markup Language*
HTTP	*HypeText Transfer Protocol*
IDH	Índice de Desenvolvimento Humano
KTA	*Knowledge-To-Action*
LARIISA	Laboratório de Redes Inteligentes e Integradas de Saúde
LIDMS	*Linked Data Mashup Services*
OLAP	*On-line Analytical Processing*
OMS	Organização Mundial da Saúde
OQL	*Object Query Language*
OWL	*Web Ontology Language*
PEP	Prontuário Eletrônico do Paciente
PNI	Programa Nacional de Imunizações
PNUD	Programa das Nações Unidas para o Desenvolvimento
QEF-LD	*Query Evaluation Framework - Linked Data*
RDF	*Resource Description Framework*
RDFS	*Resource Description Framework Schema*
SCNS	Sistema do Cartão Nacional de Saúde

SGBD	Sistema de Gerenciamento de Banco de Dados
SIA	Sistema de Informações Ambulatoriais
SIAB	Sistema de Informação da Atenção Básica
SIH	Sistema de Internações Hospitalares
SIM	Sistema de Informações de Mortalidade
SIMDA	Sistema de Monitoramanto Diário de Agravos
SINASC	Sistema de Informações de Nascidos Vivos
SINAN	Sistema de Informação de Agravos de Notificação
SIS	Sistemas de Informação de Saúde
SISA	Sistema de Saúde Adaptado-ao-contexto de Gestão de Saúde
SOA	*Service-Oriented Architecture*
SOAP	*Simple Object Access Protocol*
SPARQL	*Sparql Protocol And RDF Query Language*
SQL	*Structured Query Language*
SUS	Sistema Único de Saúde
UDDI	*Universal Description, Discovery and Integration*
UFC	Universidade Federal do Ceará
UFRN	Universidade Federal do Rio Grande do Norte
UNICEF	*United Nations Children's Fund*
URI	*Uniform Resource Identifier*
WSDL	*Web Services Description Language*
XML	*eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

O uso de Tecnologias da Informação e Comunicação (TIC's) aplicadas à saúde vem despontando nas últimas décadas. O termo *e-health* está sendo usado para representar essa integração das TIC's com a saúde e medicina. Várias são as pesquisas sobre o tema, como é evidenciado pelos periódicos dedicados a sua discussão.

Telemedicina e *home care,* são algumas subáreas que se destacam como principais tendências de pesquisa nessa área. Mas os sistemas de informação desenvolvidos para saúde despontam sobre todas as outras, tendo os sistemas de *EHR* (*Eletronic Health Records*; em português, Registro Eletrônico de Saúde - RES) como os mais significativos (BATH, 2008; HÄYRINEN; SARANTO; NYKÄNEN, 2008).

Os prontuários eletrônicos são focados tanto no paciente, já que podem sumarizar todo o seu histórico médico, como alergias, resultados de exames, tratamentos realizados, doenças crônicas, etc., como podem ser úteis para propósitos de governança em saúde, sendo fontes de informações para um gestor, com a geração de informações agregadas, clínicas e administrativas, promovendo grande impacto e benefício na melhoria da eficácia, eficiência, segurança, e qualidade da prática de saúde (MASSAD; MARIN; AZEVEDO, 2003).

No Brasil, a ineficaz gestão é muitas vezes citada como sendo a principal causa da péssima situação na qual se encontra a saúde pública, onde usuários podem esperar anos para conseguir uma cirurgia eletiva. A dificuldade de grande parte dos gestores na tomada de decisão nas três esferas de governo se deve a vários fatores, dos quais se destacam: o baixo nível de cobertura das informações; o retardo entre os eventos de coleta e análise das informações; e a baixa confiabilidade dessas informações.

Muitos trabalhos reforçam que a aplicação de Sistemas de Informação em Saúde (SIS) na área pública possibilitam uma gestão de saúde mais eficiente, por conseguir viabilizar uma melhor gestão do conhecimento e interpretação das informações integradas, funções fundamentais para suporte à tomada de decisão.

Pode-se destacar os trabalhos de Santos e Gutierrez (2008) e Montenegro *et al.* (2013).

O projeto LARIISA vem ao encontro dessa tendência, com uma proposta de realizar inferência de informações contextuais de saúde e auxiliar a tomada de decisão do gestor de saúde. A plataforma se baseia em aplicações de saúde do tipo *context-aware*, ou seja, que levam em consideração informações que possam ser usadas para caracterizar a situação de uma entidade (DEY, 2000).

Em termos práticos, ser uma plataforma *context*-aware significa dizer que as inferências realizadas pelo LARIISA podem ser obtidas a partir de informações geradas em um curto período de tempo e não só a partir de um histórico longo de informações. Por exemplo, a partir de uma aplicação móvel, integrada ao LARIISA, que faça usuários relatarem casos de suspeita de dengue, o LARIISA poderá inferir sobre um possível foco de dengue em um determinado bairro, fazendo com que o gestor de posse dessas informações possa tomar providências a respeito, como intensificar as campanhas no local ou aumentar a frequência de carros fumacê[1].

Além de *context-aware,* o LARIISA baseia sua inteligência nos 5 domínios do modelo de inteligência de governança que, segundo Andrade (2012), são necessários para um gestor de saúde realizar as melhores decisões: (i) inteligência de gestão do conhecimento; (ii) inteligência normativa; (iii) inteligência clínica-epidemiológica; (iv) inteligência administrativa; e (v) inteligência compartilhada.

Para que o LARIISA possa realizar inferências relevantes, que possam contribuir com uma boa governança de saúde, defende-se neste trabalho a ideia de que não só informações contextuais são necessárias para a construção de uma base de conhecimento em saúde significativa. Para ampliar o poder decisório do *framework* é importante que contenha uma base de conhecimento de saúde com dados que representem os 5 domínios de inteligência citados no parágrafo anterior. Mas como construir essa base de conhecimento?

No contexto brasileiro, o DATASUS, departamento de informática do Ministério da Saúde, responsável por coletar, processar e disseminar informações de saúde pública, também poderia ser um fornecedor de informações para o LARIISA, já que o departamento mantém diversos sistemas de saúde, como o Sistema de Informações Ambulatoriais (SIA), Sistema de Informações Hospitalares

[1] Inseticida pulverizado em ultra baixo volume

(SIH), Cadastro Nacional de Estabelecimentos de Saúde (CNES), Sistema de Informações sobre Mortalidade (SIM), Sistema de Informações de Nascidos Vivos (SINASC), entre outros (DATASUS, 2014).

Trabalhos como o de Junior (2009) e de Leão *et al.* (2004) mostram que a grande limitação dos sistemas de informação mantidos pelo DATASUS é o fato de suas bases de dados não se integrarem, resultado de cada sistema ter sido desenvolvido para atender uma demanda específica, diminuindo a possibilidade de uma gestão central e local mais efetiva. O próprio ex-diretor do DATASUS, Augusto César Gadelha Vieira, confirma:

> A sistematização de dados sobre a saúde enfrenta hoje, no Brasil, dois problemas de organização. O primeiro é a dificuldade para integrar todos os sistemas do maior banco de dados na área, o DATASUS, ligado ao Ministério da Saúde. O segundo, o entrave da redundância, ou seja, muitas informações colhidas pelo governo sendo repetidas em vários outros bancos de dados e que poderiam ser enxugadas. A constatação é do diretor do DATASUS, Augusto César Gadelha Vieira [...] (PIERRO, 2011, p. 1).

É sabido que essa integração e interoperabilidade são aspectos prioritários para o SUS, como demonstra a iniciativa de implantação do Sistema Cartão Nacional de Saúde (SCNS) e do desenvolvimento de uma plataforma de arquitetura orientada a serviços (barramento de serviços de saúde). O SCNS não se trata apenas da distribuição de cartões aos usuários do SUS. O primeiro aspecto que ele prioriza é o da identificação única desses usuários, atacando o problema da redundância de informações. Após o processo de identificação concluído, o projeto do SCNS atuará como um registro eletrônico do paciente, armazenando todo o seu histórico médico, podendo ser acessado em qualquer lugar do planeta.

Outras fontes de dados também podem ser relevantes para criar a base de conhecimento do LARIISA, como dados abertos disponíveis na Internet, dados de aplicações diversas de saúde, e quaisquer outros dados que podem vir a ser relevantes e que não sejam necessariamente ligados à saúde.

1.1 CARACTERIZAÇÃO DO PROBLEMA

Considerando o contexto exposto, observa-se o desafio enfrentado. O LARIISA se propõe a fornecer inteligências de governança na tomada de decisão na saúde considerando sua base de conhecimento construída a partir de informações de contexto. Os trabalhos realizados do LARIISA até o momento só apresentaram a

construção dessa base de conhecimento a partir de informações coletadas de aplicações desenvolvidas especificamente para a plataforma (ANTUNES, 2011; GARDINI *et al.*, 2013; PINHEIRO, TACIANO *et al.*, 2011).

No entanto, o LARIISA também pode considerar dados outros que possam contribuir para inferências mais significativas. Isso significa ter de considerar dados de sistemas de saúde já existentes, como os dados oriundos dos sistemas mantidos pelo Ministério da Saúde, ou por outras esferas do governo.

Na prática, não só dados de saúde que podem ser considerados, mas sim quaisquer dados que possam se mostrar relevantes para auxiliar a tomada de decisão. Dados contextuais de uma localidade, como índices de precipitação, temperatura, podem se mostrar úteis, por exemplo, para identificar possíveis áreas endêmicas. Cabe ressaltar que não é necessário a construção de provedores de contexto específicos para coletar essas informações, muitas delas já estão disponíveis.

Sendo assim, fica claro a necessidade do LARIISA de ser capaz de lidar com diferentes bases de dados. Para tal deve ter a capacidade de não apenas acessar esses dados ou considerá-los de forma isolada, mas também integrando esses dados, obtendo assim informações mais ricas e consequentemente possibilitando tomadas de decisões mais eficientes. Este trabalho, portanto, lida diretamente com a problemática de integração de dados heterogêneos, independentes e distribuídos no LARIISA.

1.2 OBJETIVOS GERAL E ESPECÍFICOS

1.2.1 Objetivo Geral

O objetivo geral da dissertação é especificar um processo de integração de dados para o LARIISA, capaz de lidar com fontes de dados heterogêneas, independentes e distribuídas, resultando em uma base de conhecimento mais rica, aumentando assim seu poder de inferência. O processo proposto possibilitará a um projetista, a integração de dados para a plataforma ser realizada sobre demanda, ou seja, novos dados poderão ser integrados ao LARISSA conforme a necessidade.

Isso é possível com a utilização dos conceitos de dados *linkados* (*Linked Data*) e ontologias (já considerada na arquitetura do LARIISA), que se destacam no contexto da Web Semântica. A Web Semântica considera o conceito de migrar a

“Web de documentos” para uma “Web de Dados”, assim, ao invés do aglomerado de páginas e links entre elas que existem hoje, tem-se um aglomerado de dados, conectados entre si, disponíveis da Internet.

A Web semântica possibilita interligar significados de palavras e consegue atribuir um significado (sentido) aos conteúdos publicados na Internet, de modo que seja perceptível tanto pelo humano como pelo computador. Ela fornece tecnologias para efetivamente publicar, recuperar e descrever dados distribuídos na Web. A integração de dados em grande escala é provavelmente um dos melhores casos de uso para as tecnologias de Web semântica, pois há vários aspectos da Web semântica que a tornam adequada para a integração de dados de fontes distribuídas e heterogêneas (HEATH; BIZER, 2011).

Essa integração de dados na Web Semântica é alcançada através do uso de *Linked Data*. *Linked Data* refere-se a um conjunto de melhores práticas para publicar e interligar dados estruturados na Web. São elas: (i) Uso de URI’s como nome para as coisas, (ii) Uso de URI’s em HTTP, para que as pessoas possam procurar esses nomes, (iii) Uso de padrões como RDF e SPARQL, (iv) Inclusão de links para outras URI’s, para que se possa descobrir mais coisas (BAUER; KALTENBÖCK, 2012; BERNERS-LEE, 2006).

Desse modo, *Linked Data* tem o potencial de facilitar o acesso aos dados semanticamente relacionados, estabelecendo conexões explícitas entre conjuntos de dados distintos a fim de facilitar sua integração e fornecendo, portanto, um novo cenário à integração de dados.

Também é utilizado uma modelagem conceitual baseada em ontologias que permitirá criar um único vocabulário, considerando os diversos conceitos existentes das fontes de dados que serão integradas. Essa abordagem se mostra oportuna, uma vez que diversas pesquisas demonstram a eficiência da utilização de ontologias como abordagem de integração de dados.

Com isso, o processo proposto permite integrar dados, sejam esses dados provindos de aplicações construídas exclusivamente para fornecer informações à plataforma, sejam dados provindos de sistemas de saúde já existentes que queiram fornecer dados para o LARIISA, sejam dados de saúde abertos já disponíveis no formato *Linked Data*.

1.2.2 Objetivos Específicos

Como objetivos específicos tem-se:

a) Identificar os principais desafios da integração de dados no contexto do SUS;
b) Analisar as abordagens de integração de dados existentes e a compatibilidade com a plataforma LARIISA;
c) Seleção de tecnologias para a definição de um processo de integração de dados para o LARIISA;
d) Definição de cenários de aplicação do processo de integração de dados desenvolvido;
e) Aplicação da proposta no projeto GISSA (Projeto FINEP – 2015).

1.3 CONTRIBUIÇÕES

A principal contribuição deste trabalho está na especificação do processo de integração de dados para o LARIISA. Com esse processo definido, a integração de dados no LARIISA pode ser feita com maior rapidez e qualidade, saindo do processo empírico. Isso permite à plataforma a criação de uma rica base de conhecimento, o que resulta em realização de inferências mais eficientes. Isso se deve ao fato do processo considerar para integração as mais diversas fontes de dados, sejam essas fontes oriundas de sistemas legados, sejam fontes de dados disponíveis na Web, fontes de dados disponíveis em tabelas, etc.

Outra contribuição importante está na apresentação de uma arquitetura que define o ambiente de execução do esquema mediado criado pelo processo proposto, o que torna o esquema mediado disponível para as aplicações LARIISA como um serviço Web. Assim, os resultados deste trabalho serve tanto o desenvolvimento de uma aplicação para o LARIISA, que objetiva apenas a integração de umas poucas bases de dados, para resolver alguma questão específica da saúde, quanto para aprimorar a base de conhecimento da plataforma, que envolve a integração de um número bem maior de bases de dados.

1.4 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação possui seis capítulos mais as considerações finais. Não sendo mais necessário tratar deste capítulo introdutório, os demais são delineados a seguir.

O Capítulo 2 – Fundamentação Teórica – apresenta uma síntese dos assuntos mais relevantes que servem de fundamentação para o entendimento dos demais capítulos desta dissertação. Ele expõe os principais conceitos e ferramentas relacionados a integração de dados, Web Semântica, Ontologias, RDF e SPARQL.

O Capítulo 3 – Trabalhos Relacionados – trata dos principais trabalhos que abordaram a integração de dados heterogêneos e distribuídos. São destacadas funcionalidades e desvantagens de cada uma delas.

O Capitulo 4 – Percurso Metodológico – é apresentado o percurso metodológico através do qual se construiu o presente trabalho: as escolhas, a definição do objeto de pesquisa e os principais conceitos utilizados.

O Capítulo 5 apresenta o processo de integração de dados proposto para o LARIISA utilizando a tecnologia de *Linked Data*, levando em consideração fontes de dados heterogêneas.

Por fim, o Capítulo 6 tece conclusões sobre o trabalho e apresenta possíveis trabalhos futuros para dar prosseguimento ao que foi obtido até aqui.

2 FUNDAMENTAÇÃO TEÓRICA

A plataforma LARIISA, desde que foi concebida em 2010, já resultou em diversos trabalhos, desde aplicações desenvolvidas para captação de informações de contexto (como condições de saúde de pacientes e número de casos epidemiológicos detectados por agentes de saúde) (ANTUNES, 2011; PINHEIRO, TACIANO *et al.*, 2011), até outras abordagens conceituais (como a utilização de redes bayesianas, como ferramenta de inferência) (SANTOS, IVOMAR; TELES; OLIVEIRA, 2013).

Neste trabalho é incorporado um novo conceito para o LARIISA. Com a evolução da Web atual, onde ela vem deixando de ser um aglomerado de documentos interligados para se tornar um aglomerado de dados vinculados, dos mais variados domínios, surge um novo cenário para a integração de dados. Considerando a grande demanda do LARIISA por dados integrados, vê-se a utilização desses novos conceitos como facilitador do processo de inferência.

Neste capítulo são apresentados esses conceitos, essenciais para o entendimento do contexto ao qual foi fundamentada esta evolução do LARIISA, assim como a teoria necessária para a compreensão dos capítulos seguintes.

2.1 LARIISA

O LARIISA foi concebido para ser um sistema capaz de auxiliar na tomada de decisão em saúde a partir da inferência de informações obtidas de aplicações que interagem com o usuário. Essas aplicações devem ser desenvolvidas considerando a obtenção de dados de contexto, enriquecendo o poder de inferência do sistema.

Inicialmente concebido para ter como principal provedor de contexto softwares desenvolvidos para a TV Digital Brasileira, suportando aplicações de tele monitoramento, procedimentos médicos remotos, etc., trabalhos seguintes comprovaram que a plataforma LARIISA teria inúmeras outras possibilidades de obter esses dados contextuais.

A explosão dos dispositivos e da Internet móveis cria um ambiente perfeito para novos provedores de contexto, além da TV Digital Brasileira anteriormente definida na plataforma. Os smartphones, com seus sensores já

embutidos, podem ser utilizados para coleta e envio de informações sobre as condições de saúde de seus usuários. Essas informações seriam encaminhadas para o LARIISA e posteriormente aos profissionais de saúde, com o intuito de oferecer melhorias à coordenação das ações e eficácia dos procedimentos de detecção/tratamento remoto de doenças.

Além da vantagem de ser uma plataforma sensível a contexto, o LARIISA também melhora qualitativamente a tomada de decisão, pois leva em consideração as inteligências de governança de saúde (ANDRADE, 2012), sobre o que é necessário conhecer para se realizar uma boa governança em saúde.

A Figura 1 ilustra a modelagem do LARIISA fundamentada nos 5 domínios de inteligência em governança de saúde: (i) inteligência de gestão do conhecimento; (ii) inteligência normativa; (iii) inteligência clínica-epidemiológica; (iv) inteligência administrativa; e (v) inteligência de informações compartilhadas. Observa-se com a figura que a plataforma permite obter dados diversos, dados esses que podem representar um domínio de inteligência isolado ou permear todos os domínios. Com a integração de todos esses dados, o LARIISA poderá realizar uma inferência muito mais significativa, pois os dados utilizados para a realização de inferências são dados significativos para a governança em saúde e, consequentemente resulta numa boa tomada de decisão do gestor. Nas seções seguintes, são tratados com mais detalhes esses domínios de inteligência.

Figura 1 - Sistema de Governança em Saúde do LARIISA

Fonte: ANDRADE (2012).

2.1.1 Sistema De Governança Em Saúde

2.1.1.1 Inteligência De Gestão Do Conhecimento

A gestão do conhecimento, independente da área, vem sendo bastante valorizada, principalmente sendo vista como um diferencial competitivo para uma gestão institucional de sucesso. Segundo Davenport & Prusak (1998), a gestão do conhecimento pode ser vista como uma série de ações gerenciais constantes e sistemáticas que facilitam os processos de criação, registro e compartilhamento do conhecimento nas organizações. Nos tempos atuais, com o avanço das tecnologias de informação e comunicação, o conhecimento passou a ser considerado o diferencial competitivo das organizações que pretendem ter longevidade e sucesso.

Na área da saúde, a gestão do conhecimento possibilita uma composição de estratégicas e práticas para identificar, criar e representar experiências de cuidados de saúde, por qualquer que seja o ator envolvido, dos profissionais de saúde ao gestor, passando pelas experiências do próprio paciente.

Nesse contexto, o LARIISA acaba por facilitar essa gestão do conhecimento, pela grande diversidade de informações que é capaz de reunir. O LARIISA também terá de suportar a inserção desse conhecimento de alguma forma.

Tendo claro a importância da gestão do conhecimento para qualquer tipo de governança e, principalmente no contexto de governança para a saúde, o LARIISA não poderia deixar de considerar a obtenção de dados resultantes da gestão de conhecimento em saúde.

2.1.1.2 Inteligência Normativa

O LARIISA precisa agregar e considerar as normativas relacionadas à saúde que devem ser consideradas por um gestor na tomada de decisão. Por exemplo, para elaboração de uma estratégia de combate à dengue, o gestor não deve ter ciência apenas da legislação diretamente ligada ao combate à dengue, mas ele deve ter ciência de artigos da constituição federal (como a questão da inviolabilidade da casa de um indivíduo, sobre a admissão de agentes comunitários de saúde,...); de artigos do código penal referentes à crimes contra a saúde pública

(como as penas relacionadas ao causador de epidemias, infração por impedir medida sanitária preventiva, etc.); e outras mais.

Sendo assim, o LARIISA considera em suas inferências informações relacionadas à legislação, pertinentes ao domínio de conhecimento em que a plataforma está inserida.

2.1.1.3 Inteligência Clínica-Epidemiológica

A inteligência clínica-epidemiológica diz respeito às informações referentes ao conhecimento dos processos saúde-doença, ao conhecimento dos protocolos envolvidos em rotinas clínicas, dos processos de investigação de eventos de saúde pública que possam representar uma ameaça para a saúde pública, ou seja, de uma forma geral, refere-se às informações diretamente ligadas ao conhecimento de saúde.

Um exemplo da importância dessa inteligência para a governança de saúde pode estar relacionado na atuação do LARIISA provendo ao gestor decisões mais assertivas na contenção de uma situação endêmica.

2.1.1.4 Inteligência Administrativa

Como o próprio nome já sugere, essa inteligência para a governança relaciona-se ao conhecimento dos processos de gestão administrativa na área da saúde. Pode estar relacionado, por exemplo, à gestão da alocação de profissionais de saúde para atingir um determinado objetivo, considerando treinamento e o tempo necessário para atingir esse objetivo; pode estar relacionado também à gestão de recursos de saúde considerando períodos endêmicos.

2.1.1.5 Inteligência De Gestão Compartilhada

A gestão compartilhada diz respeito à participação social no contexto da saúde. Por exemplo, pode-se citar a atuação da população no compartilhamento de informações nas mídias sociais, como, quando informam que estão com uma determinada condição clínica, ou quando informam sobre focos de dengue (SILVA *et al.*, 2011).

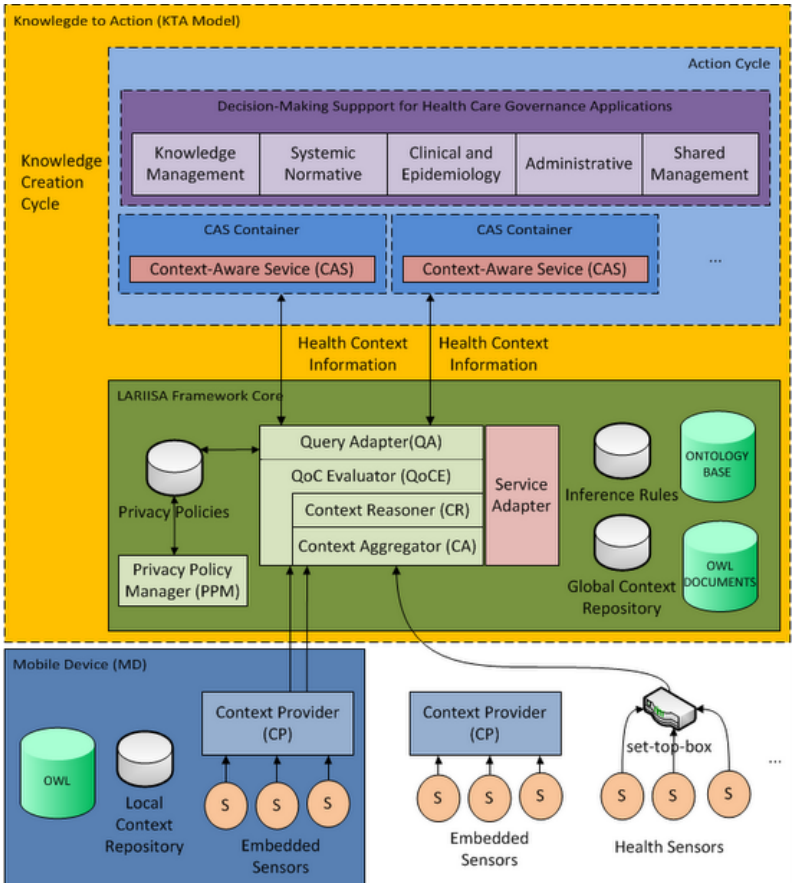
A inteligência de gestão compartilhada também diz respeito à gestão de saúde compartilhada das três esferas de governo. É importante para o gestor ter plena compreensão sobre as suas competências frente aos sistema de saúde brasileiro, seja ele de responsabilidade do município, estado ou união.

2.1.2 Arquitetura Do Lariisa

A arquitetura do LARIISA possui como alicerces os conceitos de *context-aware* e de inteligências de governança em saúde. O primeiro grande desafio era sobre a diferença entre o processo de detecção de contexto (*knowledge)* e como esse contexto afetaria as aplicações relacionadas (*Action*). Para isso, o *framework* se vale do modelo *Knowledge-To-Action* (KTA) (GRAHAM *et al.*, 2006). A

Figura 2 a seguir ilustra a arquitetura do LARIISA.

Figura 2 - Arquitetura do LARIISA e Aplicações de Tomada de Decisão em Governança



Fonte: (OLIVEIRA et al., 2010).

É possível observar, através da

Figura 2, que o conceito de ontologias já era considerado na arquitetura do LARIISA, mas nunca foi explorado no contexto da integração de dados, além da representação por ontologias do conceito de informações contextuais de saúde. Vale destacar aqui o componente da arquitetura relacionado com o conceito de ontologias:

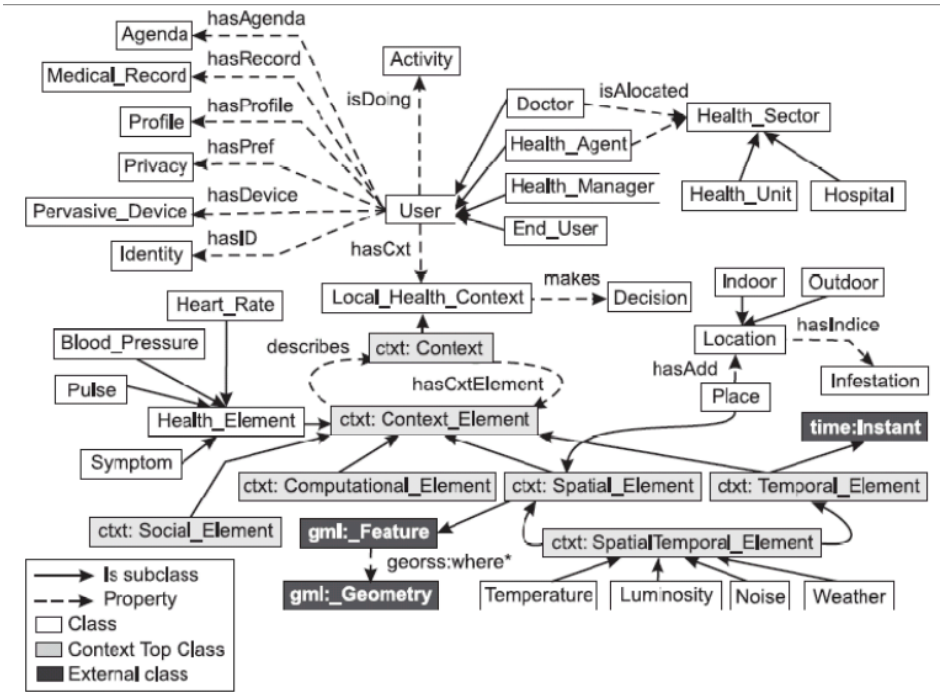
- *Ontology Base*: Esse componente provê a gerência do conhecimento dentro do LARIISA. Ele armazena ontologias, suas instâncias, regras de inferência e derivação. Permite manipulação e recuperação do conhecimento pelo *Context Reasoner*.

Como dito, as ontologias são utilizadas na arquitetura do LARIISA para representar o conceito de informações contextuais em saúde. Nesse sentido, são definidos os conceitos de *Local Health Context*, que descreve a situação de qualquer entidade interagindo com o sistema de governança, como pacientes, gestores de saúde, agentes de saúde, etc. Essas informações são utilizadas para definir regras de decisões locais de saúde e para compor o *Global Health Context.* O *Global Health Context* descreve informações de alto nível geradas a partir das informações de contexto local entregues pelo *Local Health Context*. Por exemplo, o *Global Health Context* descreve o número de casos de dengue confirmados numa região, durante um determinado período de tempo. A Figura 3 e a Figura 4 a seguir ilustram as ontologias de *Local Health Context* e *Global Health Context*.

Como já mencionado, alguns trabalhos já foram desenvolvidos referentes ao projeto LARIISA. Merece destaque o trabalho de Pinheiro *et al.* (2011), tomando como partida o trabalho do Diga-Saúde (SANTOS, 2011), que propõe um prontuário eletrônico capaz de captar informações de contexto do paciente em sua própria casa. O trabalho de Frota (2011) construiu uma aplicação para a inclusão e exclusão de provedores de contexto ao LARIISA. O trabalho de Antunes (2011), chamado de SISA, apresenta um protótipo de aplicação para o LARIISA no contexto de agravos de dengue. Esse protótipo possui um módulo para captar informações locais de contexto e um módulo Web de visualização das informações globais de contexto.

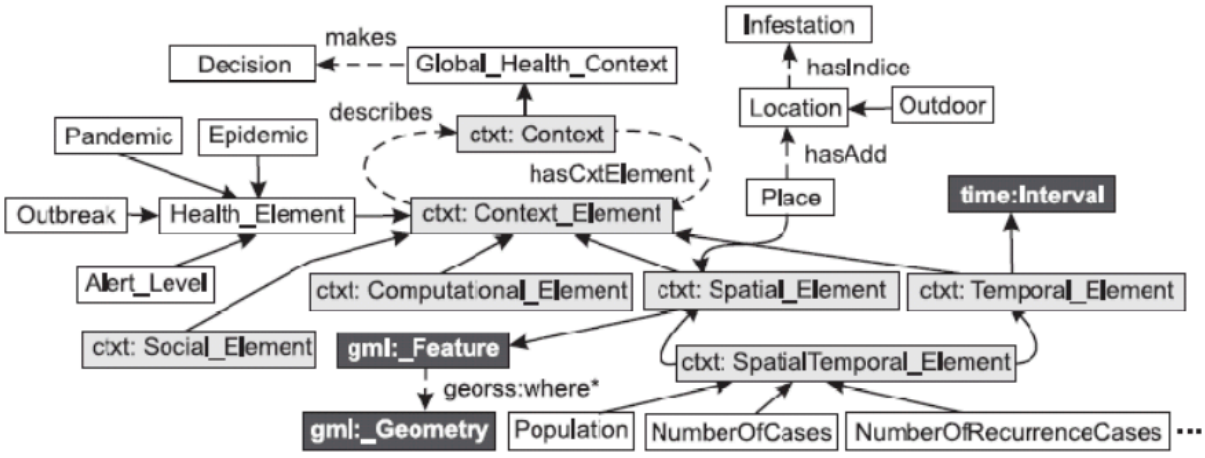
O trabalho de Alcântara (2012), que apresentou uma ferramenta para construção de aplicações para o LARIISA. O trabalho de Teles (2013) propôs uma nova arquitetura para o LARIISA, onde ele considera a utilização de redes bayesianas como motor de inferência para o *framework*.

Figura 3 - *Local Health Context*



Fonte: (OLIVEIRA *et al.*, 2010).

Figura 4 - *Global Health Context*

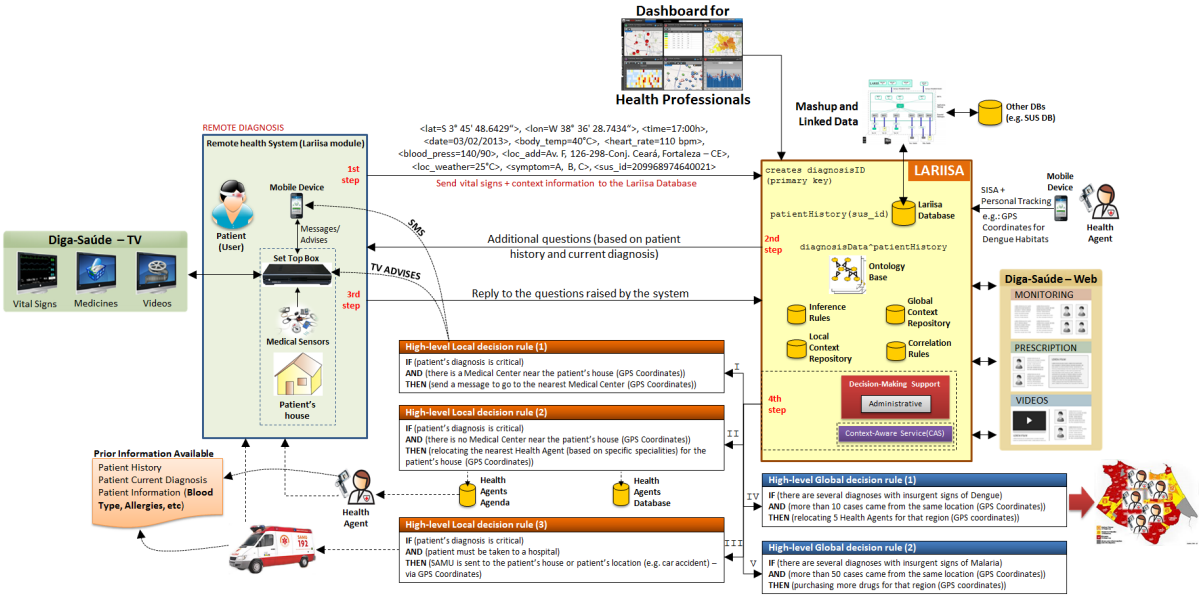


Fonte: (OLIVEIRA *et al.*, 2010).

A Figura 5 a seguir, mostra um fluxo de utilização do LARIISA, em um cenário de *home care,* onde são considerados os resultados dos trabalhos mencionados. Ela mostra a interação de um usuário com o *framework* através do envio de seus dados de saúde, e suas ações resultantes após inferência desses

dados pelo LARIISA, podendo ser o envio de um agente de saúde ou de uma ambulância no contexto de saúde local, e em um contexto de saúde global, a compra de mais medicamentos ou a realocação de agentes de saúde para a região. Têm-se então os seguintes passos:

Figura 5 - Cenário de Utilização do LARIISA



Fonte: GARDINI et al. (2013).

Passo 1 – Em verde, tem-se a representação da captura de dados contextuais de saúde do usuário de uma aplicação do LARIISA. Vê-se a possibilidade de utilização de vários provedores de contexto, como o uso da TV Digital (PINHEIRO, TACIANO *et al.*, 2011; SANTOS, 2011) e do *smartphone* (FROTA, 2011).

Passo 2 – Em amarelo, tem-se a visão do *framework* LARIISA e suas interações ao receber as informações de saúde contextuais do usuário. Nesse cenário, já observa-se a base de conhecimento do LARIISA (*LARIISA Database*) sendo formada por dados integrados (*Mashup and Linked Data* – objeto deste trabalho), inclusive dados esses provindos de outras fontes, como os bancos de dados do SUS. Após inferências feitas, o *framework* solicita ao usuário informações adicionais.

Observa-se também a troca de informações do LARIISA com suas aplicações, como o SISA (ANTUNES, 2011) e o DIGA-Saúde (SANTOS, 2011).

Neste cenário, O SISA fornece informação contextual dos agentes de saúde, a localização. O DIGA-Saúde apresenta dados de saúde global para o gestor considerando os dados do usuário que o LARIISA acabou de processar.

Passo 3 e 4 – Após o envio das respostas às informações adicionais solicitadas pelo LARIISA ao usuário (passo 3), o framework possui 3 possíveis decisões a tomar em nível de saúde local (passo 4):

a) Decisão 1: enviar uma mensagem ao usuário com as coordenadas da unidade de saúde que ele deve ir, se seu diagnóstico for crítico e existir uma unidade de saúde próxima à caso do usuário;

b) Decisão 2: enviar um agente de saúde especialista à casa do paciente, se seu diagnóstico for crítico e não existir uma unidade de saúde próxima à caso do usuário (nesse caso, o agente de saúde escolhido será determinado pelas informações de contexto providas pelo SISA);

c) Decisão 3: enviar um SAMU (Serviço de Atendimento Médico de Urgência) à casa do usuário, se o seu diagnóstico for crítico e se for necessário a sua ida a um hospital.

Já em nível de saúde global, o LARIISA possui duas decisões a tomar (passo 4):

a) Decisão 1: realocar 5 agentes de saúde para a região no entorno da casa do usuário, se há vários diagnósticos na cidade com sinais de dengue e se há mais de 10 casos confirmados nessa região;

b) Decisão 2: planejar a compra de mais remédios para atender à região, se há vários diagnósticos na cidade com sinais de malária e se há mais de 50 casos confirmados nessa região.

Em resumo, consegue-se observar na Figura 5 todo o fluxo de funcionamento do *framework* LARIISA, em um contexto de saúde local, onde é possível inferir ações a serem realizadas considerando o estado de saúde do usuário e de todo o seu histórico (através de uma base de conhecimento integrada); e em um contexto de saúde global, onde os dados desse usuário influenciam em decisões de gestão.

2.2 ONTOLOGIAS

Na ciência da computação, define-se as ontologias simplesmente como um processo formal para representar conhecimento. As ontologias nomeiam e definem tipos, propriedades e relacionamentos de entidades em um domínio específico. Considera-se as entidades como conceitualizações desse domínio.

As ontologias foram desenvolvidas dentro do contexto de Inteligência Artificial para facilitar o compartilhamento de conhecimento e seu reuso. Desde então o estudo de ontologias vem sendo aplicado em diversos outros campos de estudo como gerência de compartilhamento, comércio eletrônico, recuperação de informações, sistemas cooperativos de informação, e mais recentemente no contexto de Web Semântica. Essa popularização muito se deve ao que as ontologias prometem fazer: um entendimento comum e compartilhado de algum domínio que pode ser comunicado entre pessoas e sistemas de aplicação (FENSEL, 2001).

Em Gruber (1993) é apresentada uma definição mais formal e aceita pela comunidade acadêmica: “Ontologia é uma especificação formal e explícita de uma conceitualização compartilhada”. ‘Conceitualização’ significa ter um modelo abstrato de um fenômeno que identifica conceitos relevantes deste. ‘Explícito’ significa que as definições de nomenclaturas são não-ambíguas; ‘formal’ significa passível de ser processada automaticamente; e ‘compartilhada' representa o conhecimento consensual de um domínio.

2.2.1 Classificação Das Ontologias Baseado Na Granularidade Que Representam

Com relação ao escopo dos objetos descritos pela ontologia, pode-se ter a seguinte classificação (ROUSSEY *et al.*, 2011):

a) Ontologias de Aplicação: são ontologias de um domínio específico, que representa o ponto de vista único, de um usuário ou desenvolvedor. Podem ser consideradas como uma especialização das ontologias de domínio.

b) Ontologias de Domínio: representam o ponto de vista de um fenômeno compartilhado por um grupo de usuários.

c) Ontologias de Referência: representam um padrão utilizado por diferentes grupos de usuários. Esse tipo de ontologia é resultado da integração de várias ontologias de domínio, geralmente construída para captura os conceitos e relações centrais do domínio.

d) Ontologias gerais: esse tipo de ontologia não é dedicado a um domínio específico. Possuem conhecimentos gerais. Exemplo: conceito de lugar, organização, pessoa, etc.

e) Ontologias de Alto Nível: são ontologias genéricas aplicáveis a vários domínios. Definem conceitos básicos como objetos, processos, relações, eventos, etc.

2.2.2 Componentes Das Ontologias

Uma ontologia consiste de um número de diferentes componentes, com seus nomes podendo variar dependendo da linguagem ontológica utilizada. Apesar disso, os componentes básicos de uma ontologia, listados a seguir, são compartilhados pela maioria das linguagens:

a) Classe (ou Conceito): descrevem conceitos em um domínio. Um conceito representa um grupo de diferentes *indivíduos*, que compartilham características comuns, podendo ser mais ou menos específicos.

b) Relações (conhecido também como **Predicado** ou **Propriedade):** usada para estabelecer um relacionamento entre dois termos. O primeiro termo deve ser um conceito que represente o domínio (*domain*) da relação; e o segundo termo deve ser um conceito que represente o contradomínio (*range*) da relação. O contradomínio de uma propriedade também pode ser um tipo de dado primitivo como *string*, *decimal* ou *boolean*. E uma relação pode ter sub-relações**.**

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

A representação do conhecimento de ontologias deve ser realizada por linguagens específicas, normalmente variantes da lógica descritiva que provêm alta expressividade e raciocínio. RDFS e OWL são exemplos de linguagens utilizadas.

2.3 WEB SEMÂNTICA

A ideia da Web Semântica surgiu em 2001 da publicação de Berners-Lee, Hendler e Lassila (2001) onde se defendia a extensão da Web atual, baseada em publicação de documentos, para uma Web baseada na semântica das informações, construindo assim uma Web de dados. Os dados podem ser acessados usando *Uniform Resource Identifier* (URI's) e estar relacionados uns com os outros da mesma forma que os documentos já são.

A Web Semântica descreve uma Web qualitativamente diferente da atual, onde, ao invés de se ter conteúdos estruturados sendo exibidos e lidos por pessoas, tem-se informações semanticamente "inteligíveis" por programas, os chamados "agentes de software". A Web Semântica tem como finalidade atribuir um significado aos conteúdos publicados na Internet de modo que seja perceptível tanto pelo humano quanto pelo computador.

Neste sentido, diversas sãs as tecnologias necessárias para possibilitar que as máquinas passem a entender o significado das coisas. As ontologias, descritas na seção anterior, é uma delas, possibilitando definir conceitos e relações relacionados à uma área de conhecimento específico.

Já o conceito de *Linked Data* é considerado o "coração" da Web Semântica, permitindo integração de dados em larga escala na Web. Mais especificamente, pode-se definir *Linked Data* como um termo utilizado para descrever as melhores práticas para expor, compartilhar e conectar pedaços de dados, informação e conhecimento na Web, utilizando URI's e *Resource Description Framework* (RDF).

O RDF é a base para a publicação e *linkagem* de dados. O RDF provê uma linguagem para modelagem de dados baseado na ideia de construir asserções sobre recursos na forma de tuplas, sujeito-predicado-objeto. Com isso, todos os dados da Web Semântica estariam sobre o padrão RDF. Para consultar esses dados em RDF, utiliza-se como ferramenta a linguagem de consulta SPARQL.

2.4 LINKED DATA

Em resumo, *Linked Data* descreve um método de publicação de dados estruturados de modo que possam ser interligados e se tornem mais úteis. Ele se

baseia em tecnologias padrão Web, tais como HTTP e URIs, mas ao invés de usá-las para servir páginas web para leitores humanos, estende-se o compartilhamento de informações de uma forma que possa ser lido automaticamente por computadores. Isso permite que os dados de diferentes fontes sejam conectados e consultado (BERNERS-LEE; HEATH; BIZER, 2009).

A proposta de dados abertos interligados oferece grande potencial ao conectar recursos informacionais através de links semânticos, links que são significativos também para programas. Ao contrário, links convencionais nada mais são (além de uma eventual etiqueta textual significativa para usuários humanos) que meios para que programas navegadores, a partir de um recurso, acessem outro, sem explicitar qual o significado da ligação entre os recursos. Sendo significativos para programas, links semânticos podem ser processados de forma mais rica por eles, explorando e enriquecendo cognitivamente o significado (legível por máquina) da ligação entre ambos os recursos (MARCONDES, 2012).

Berners-Lee (2006) delimitou um conjunto de "regras" para a publicação de dados na Web de forma que todos os dados publicados tornem-se parte de um espaço único de dados globais que ficaram conhecidos como “Princípios Linked Data”. São eles:

1. Usar URIs como nomes para as coisas;
2. Usar URIs de modo que as pessoas possam procurar esses nomes;
3. Quando alguém procurar uma URI, fornecer informações úteis, usando os padrões (RDF, SPARQL);
4. Incluir links para outras URIs, para que se possa descobrir mais coisas.

Como visto, o conceito de *Linked Data* apoia-se em um pequeno conjunto de padrões já bem estabelecidos e amplamente utilizados na Web: um mecanismo de identificação global e único (URI’s), um mecanismo de acesso universal (HTTP), o modelo de dados RDF, e a linguagem de consulta SPARQL para acesso aos dados.

2.5 RDF

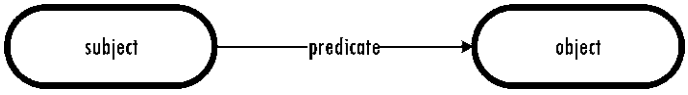
RDF foi originalmente criado em 1999 pelo W3C como um padrão para codificação de metadados. Com o advento da Web Semântica, em 2004 o W3C

lançou uma nova especificação do RDF ampliando seu significando, passando a representar qualquer recurso e suas relações existentes no mundo real.

A ideia básica do RDF é usar um modelo abstrato para decompor informação/conhecimento em pequenos pedaços, com algumas regras simples sobre o significado de cada pedaço desse, sendo um método simples e flexível para representar qualquer fato, não deixando de ser estruturado para que aplicações consigam operar esse conhecimento (YU, 2011).

Esse “modelo abstrato” possui os seguintes componentes principais: (i) Assertiva; (ii) Sujeitos e objetos; (iii) predicado. As assertivas podem ser expressas pela forma *Sujeito-Predicado-Objeto*, sempre nessa ordem, onde o sujeito e o objeto são ‘coisas’ do mundo real, e o predicado é o nome de uma relação que conecta essas duas coisas. Uma assertiva pode ser representada pela estrutura de um grafo, como representado na Figura 6.

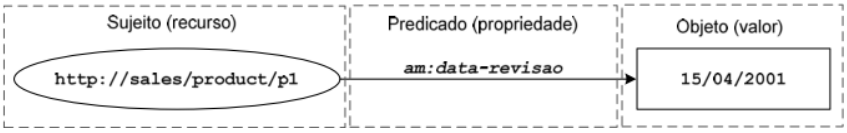
Figura 6 - Representação de um grafo (ou tripla) RDF



Fonte: (YU, 2011).

O *sujeito* pode ser representado por uma URI ou um *blank node* (nó em branco). O *predicado* é representado apenas por uma URI e o *objeto* pode ser representado por uma URI, *blank node*, ou um dado literal. Na Figura 7 a seguir, tem-se um exemplo de uma assertiva sendo representada por um grafo. O produto *p1,* representando o sujeito da assertiva, que é identificado pela URI *http://sales/product/p1*, possui uma propriedade *data-revisão* que possui o valor “15/04/2001”. O prefixo *am* é utilizado como sinônimo para o espaço de nomes, identificado na URI *http://amazon.com/schema/*, no qual o predicado *data-revisão* fora definido.

Figura 7 - Exemplo de um grafo RDF



Fonte: (PINHEIRO, JOÃO CARLOS, 2011).

Para armazenar as assertivas RDF (também podendo ser chamadas de tuplas RDF) existem os RDF *store* ou RDF *data store*, construídos especialmente para este fim. Bancos de dados relacionais podem ser usados para armazenar assertivas RDF, possibilitando a execução das funções básicas de adicionar, deletar, atualizar e localizar um registro, mas não obtém um desempenho apropriado se comparado a um sistema otimizado para operar com RDF. A Figura 8 lista os principais RDF *data store* disponíveis no mercado atual e a linguagem de implementação.

Figura 8 - Exemplos de implementações de bancos de dados RDF

RDF data store name	implementation language	home page
4store	C	http://www.4store.org
ARC	C	http://arc.semsol.org
Joseki	Java	http://www.joseki.org
Redland	C	http://librdf.org
Sesame	Java	http://www.openrdf.org
Virtuoso	C	http://virtuoso.openlinksw.com

Fonte: (YU, 2011).

2.6 SPARQL

SPARQL é a linguagem padrão da Web Semântica para recuperação de informações contidas em grafos RDF, sendo não só uma linguagem de consulta, mas também um protocolo usado para enviar consultas e recuperar resultados através do protocolo HTTP.

SPARQL provê quatro diferentes formas de consulta: (i) *SELECT query;* (ii) *ASK query*; (iii) *DESCRIBE query*; e (iv) *CONSTRUCT query*. Todas as formas de consulta são baseadas em padrão de triplas e padrão de grafos.

Assim como o conceito de tripla do RDF, o padrão de triplas no SPARQL segue o mesmo princípio. A diferença é que no padrão de triplas do SPARQL podem ser inseridas variáveis. Qualquer um ou todos de uma vez só, *sujeito*, *predicado* e

objeto, podem ser variáveis em um padrão de tripla SPARQL, que são identificadas pelo caractere *?* procedido de um nome de variável.

Figura 9 - Representação de um padrão de tripla SPARQL

```
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
<http://danbri.org/foaf.rdf#danbri> foaf:name ?name.
```

Fonte: (YU, 2011).

O padrão de tripla apresentada na

Figura 9 pode ser lido da seguinte maneira: “ache o valor da propriedade *foaf:name* definido por um recurso RDF identificado por *http:// danbri.org/foaf.rdf#danbri*.”

O conceito de padrão de grafo SPARQL é similar ao padrão de triplas visto. Um grafo SPARQL é um conjunto de padrões de triplas delimitadas por *{* e *}* utilizadas para selecionar recursos de um dado grafo RDF. Da mesma forma que as variáveis podem se fazer presentes nos padrões de triplas SPARQL, elas também podem aparecer nos grafos SPARQL. Nesse caso, cabe observar que se uma variável aparece em múltiplas triplas SPARQL, seus valores deverão ser os mesmos em todas elas.

Figura 10 - Exemplo de uma padrão de grafo SPARQL

```
{
   ?who foaf:name ?name.
   ?who foaf:interest ?interest.
   ?who foaf:knows ?others.
}
```

Fonte: (YU, 2011).

O padrão de grafo apresentado na Figura 10, tenta encontrar qualquer recurso RDF que possui todas as 3 propriedades definidas (*foaf:name*, *foaf:interest* e *foaf:knows*).

Das formas de consulta SPARQL apresentadas, a mais comum é *SELECT query.* Como visto na Figura 11, uma *SELECT query* inicia-se com a diretiva *BASE* seguida de um número arbitrário de assertivas *PREFIX*. Essas duas partes iniciais são opcionais e servem para abreviações de URI.

A cláusula *SELECT* especifica quais variáveis, ou item de dados, precisam ser recuperados da consulta e posteriormente exibidos. A cláusula *FROM* especifica ao SPARQL contra que grafo a pesquisa deve ser conduzida. A cláusula *WHERE* contém o padrão de grafo que especifica o resultado desejado. A última parte geralmente é chamada de modificadores de consulta, que possui como principal propósito reorganizar as consultas.

A Figura 12 apresenta um exemplo de uma consulta SPARQL onde é possível ver todos os componentes da consulta. Essa consulta resultará em todas as propriedades com seus respectivos valores que o recurso “Danbri” possui.

Figura 11 - Estrutura de uma *SELECT Query*

```
# base directive
BASE <URI>

# list of prefixes
PREFIX pref: <URI>
...

# result description
SELECT...

# graph to search
FROM ...

# query pattern
WHERE {
    ...
}

# query modifiers
ORDER BY...
```

Fonte: (YU, 2011).

Figura 12 - Exemplo de Consulta SPARQL

```
1: base <http://danbri.org/foaf.rdf>
2: prefix foaf: <http://xmlns.com/foaf/0.1/>
3: select *
4: from <http://danbri.org/foaf.rdf>
5: where
6: {
7:   <#danbri> ?property ?value.
8: }
```

Fonte: (YU, 2011)

2.6.1 Sparql Endpoint

Um SPARQL *endpoint* pode ser entendido como uma interface através da qual usuários, humanos ou aplicações, realizam consultas a um banco de dados RDF, utilizando a linguagem SPARQL, obtendo a resposta apropriada.

SPARQL *endpoints* podem ser classificados como *genéricos* ou *específicos*. Os SPARQL *endpoints genéricos* trabalham com qualquer conjunto de dados RDF. Os específicos, como o próprio nome sugere, apenas com conjuntos de dados RDF específicos.

Eles têm sido grande aliado para o sucesso e disseminação da utilização da linguagem SPARQL, pois possibilitam a execução de consultas de dados

disponíveis no padrão de Linked Data na Web. Alguns exemplos são DBpedia[2], Data.Gov[3], dentre outros (PINHEIRO, 2011).

2.7 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo, apresentou-se os principais conceitos necessários para o completo entendimento deste trabalho. Além disso, foi descrita uma contextualização do LARIISA, destacando os conceitos mais relevantes para o presente trabalho de sua arquitetura: as inteligências de governança de saúde e a sua base de dados de ontologias.

Foi apontado também conceitos importantes relacionados as tecnologias envolvidas na abordagem de integração de dados que utilizaremos na proposta apresentada neste trabalho, como ontologias, Web semântica, *Linked Data*, RDF e SPARQL.

[2] http://dbpedia.org/sparql
[3] environment.data.gov.uk/lab/sparql.html

3 TRABALHOS RELACIONADOS

Este capítulo tem como objetivo mostrar trabalhos encontrados que focaram na problemática da integração de dados de saúde pública, principalmente no Brasil. Esses trabalhos resolvem a questão utilizando soluções distintas: (i) Integração de dados de saúde baseada na criação de um novo esquema de dados; (ii) Integração de dados de saúde baseada na utilização de Web Services; (iii) Integração de dados de saúde baseada na abordagem de *Data Warehouse*; (iv) Integração de dados de saúde baseada em ontologias. Espera-se com esse estudo, obter subsídios para a formulação de contribuições relevantes para esta dissertação analisando a desvantagens de cada abordagem.

3.1 INTEGRAÇÃO DE DADOS DE SAÚDE BASEADA NA CRIAÇÃO DE UM NOVO ESQUEMA DE DADOS

Junior (2009) apresenta em seu trabalho uma fundamentação teórica para um sistema de integração de dados de saúde, a nível municipal, desenvolvido a partir do ano de 2005 no DATASUS, o *Integrador*[4]. O *Integrador* foi uma iniciativa do DATASUS de estudar uma possível padronização das informações que eram capturadas, tratadas e armazenadas nos seus SIS (SIA, SIH, APAC, SIM, SINASC, SINAN, PNI, entre outros).

O *Integrador* teve seu desenvolvimento realizado de forma empírica, utilizando-se dos levantamentos de necessidades junto aos gestores municipais, e das experiências da equipe de desenvolvimento.

O *Integrador* tinha como objetivo ser um ponto central onde os gestores municipais pudessem ter acesso aos dados de saúde advindos dos SIS utilizados, mas dados totalmente integrados, auxiliando esses gestores na tomada de decisão de saúde.

Com isso, a estratégia do *Integrador* foi construir uma nova base de dados, construída considerando um novo vocabulário, fruto de uma operação de avaliação dos dados dos SIS envolvidos, para se conseguir uma padronização. O trabalho de padronização não se limitou ao banco de dados do próprio sistema, mas

[4] http://integrador.datasus.gov.br/INTEGRADOR/index.php

também a padronização dos layouts de aquisição de dados dos sistemas envolvidos, a fim de evitar a despadronização futura.

Pinto (2006) apresentou em seu trabalho estratégias metodológicas para utilização e integração de bancos de dados e sistemas nacionais de informação em saúde para permitir a análise de políticas de saúde através de um observatório de saúde. É defendido o processo de integração de dados a nível estrutural como estratégia de avaliação de saúde. Em seu trabalho, ele avalia cada conjunto de banco de dados a ser integrado, procurando identificar as possibilidades de integração e qual o potencial desses bancos integrados em conseguir proporcionar uma avaliação da política de saúde.

O trabalho de Geremias, Jacobsen e Pereira (2013) faz uma análise das dificuldades encontradas na criação de um sistema, no âmbito da cidade de Florianópolis, que integra dados de saúde de sistemas locais e dos sistemas mantidos pelo Ministério da Saúde. Para a criação do sistema, observa-se as mesmas premissas existentes nos trabalhos anteriores, baseados na criação de uma única base de dados, destacando-se: (a) evitar a captação de informações duplicadas e o (b) uso de informações padronizadas, como preconizadas pelo Ministério da saúde.

Os trabalhos aqui relacionados possuem como desvantagens:

1. **Alto custo de implantação e complexidade:** criar um novo esquema de dados, com dados integrados, principalmente se considerado um quantidade considerável de bases de dados, como visto nos trabalhos elencados, é um processo muito custoso, principalmente de tempo. Há a necessidade de se padronizar o vocabulário de dados, verificar duplicidade de dados, fazer carreamento dos dados no novo esquema, etc.;
2. **Pouca escalabilidade:** adicionar novos esquemas de dados a um esquema de dados que já passou por um processo de integração pode significar refazer todo o processo;
3. **Necessidade de mineração de dados para a descoberta de conhecimento:** a extração de conhecimento de um esquema de dados necessita ser feito em um processo a parte.

3.2 INTEGRAÇÃO DE DADOS DE SAÚDE BASEADA NA UTILIZAÇÃO DE WEB SERVICES

Spazziani e Nardon (2004) propõem uma arquitetura com o objetivo de solucionar a problemática da integração de sistemas complexos e de grande porte na área de saúde. Além da integração de dados propriamente dita, é defendido que tarefas que sejam semelhantes ou até mesmo idênticas, não precisam ser reimplementadas, devem ser desenvolvidas de tal forma que possam se tornar componentes compartilhados.

As técnicas apresentadas nesse trabalho foram aplicadas na construção do sistema de informação da Secretaria Municipal de Saúde de São Paulo (SMS-SP), que integra o sistema de captura do atendimento, com o Cartão Nacional de Saúde, o Cadastro Nacional de Estabelecimentos de Saúde, o sistema de Autorização de Procedimentos de Alta Complexidade e com um sistema de agendamento universal.

A técnica é baseada na identificação das funcionalidades que precisam existir no sistema construído e mapeá-las com os sistemas legados que vão ser integrados. A integração de dados que precisa existir entre os componentes e os sistemas que possuem os dados necessários para sua execução é realizada via *Web Services.*

Pires e Ruiz (2010) apresentam em seu trabalho a problemática da interoperabilidade terminológica entre as aplicações médicas, apresentando uma solução para a interoperabilidade entre sistemas. É criado, então, um *Web Services* que utiliza o *UMLS MetaThesaurus*, um tesauro específico da área médica. Esse *Web Service* criado deverá ser acessado por aplicações médicas que desejem obter essa interoperabilidade terminológica.

Já a tese de Hira (2012) apresenta uma proposta de um arcabouço de Saúde Digital, que possibilita a integração e interoperabilidade de serviços de saúde ou sistemas de informação entre várias instituições de saúde, tendo o registro como elemento principal de convergência de informação do paciente, possibilitando ações de cuidado ao paciente de forma integrada, para efetiva atenção do paciente e de seu bem-estar. A interoperabilidade do ambiente é conseguida também via *Web Services.*

Os trabalhos aqui relacionados possuem como desvantagens:

1. **Não acesso aos dados brutos:** abordagens de integração baseadas em *Web Services* promovem uma interoperabilidade entre sistemas, e não uma integração de dados propriamente dita (como mostrado no Capítulo 4). Dessa forma, um serviço Web fornecerá respostas apenas àquilo para o qual ele foi desenvolvido. O não acesso aos dados brutos pode impossibilitar o processo de extração de conhecimento dos dados.
2. **Dependência do provedor do serviço:** os clientes dos serviço apenas obtém acesso àquilo que o provedor do serviço pode oferecer. Caso o cliente necessite de uma nova informação que o serviço atualmente não ofereça, o provedor poderá se negar a reescrever o serviço para atender aos requerimentos de um cliente específico. Outra desvantagem com relação ao provedor do serviço diz respeito a disponibilidade do serviço. Se por algum motivo o serviço se tornar indisponível, não há como ocorrer o processo de integração.
3. **Semântica:** a integração via *web services* não garante que diferentes serviços tenham a mesma semântica associada a seus dados.

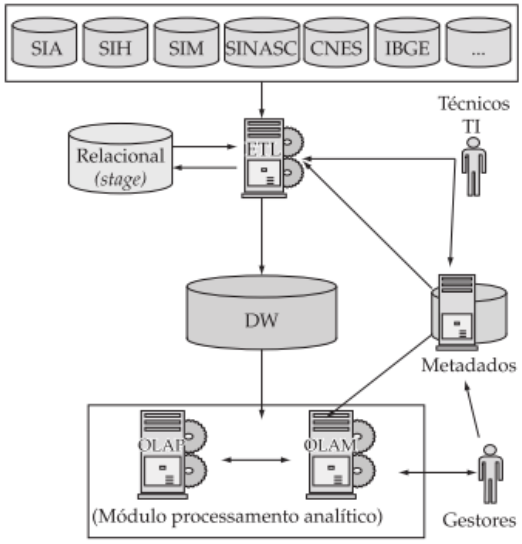
3.3 INTEGRAÇÃO DE DADOS DE SAÚDE BASEADA NA ABORDAGEM DE DATA WAREHOUSE

Santos e Gutierrez (2008) apresentam um trabalho de integração de dados de saúde baseada em *Data Warehouse*, o *MINERSUS*. O *MINERSUS* é um ambiente computacional para a produção de informação analítica por meio da mineração das bases de dados dos sistemas de informação do SUS, tendo sua utilidade avaliada por meio de uma pesquisa de campo que permitiu a interação do usuário com o ambiente.

A Figura 13 apresenta a arquitetura do *MINERSUS*, que não difere da arquitetura padrão do processo de modelagem de um *Data Warehouse.* Ela foi desenhada partindo de algumas premissas que devia partir o sistema, considerando os desafios encontrados na implantação de uma ferramenta analítica para a área da saúde pública (SANTOS, RICARDO S.; GUTIERREZ, 2004; SANTOS, RICARDO S *et al.*, 2006).

Moraes (1998) desenvolveu um protótipo de informação ambulatorial para o SUS, baseado em *data warehouse*, como parte integrante de uma dissertação de mestrado. Seu trabalho teve como objetivo mostrar as vantagens obtidas no processo decisório depois da implantação de um *data warehouse.*

Figura 13 - Arquitetura do Ambiente Computacional do *MINERSUS*



Fonte: (SANTOS, RICARDO DA SILVA; GUTIERREZ, 2008)

Os trabalhos aqui relacionados possuem como desvantagens:

1. **Desatualização dos dados no repositório:** considerando fontes de dados muito dinâmicas, que sofrem atualizações constantemente, a abordagem de *data warehouse* pode não ser a mais indicada. O *data warehouse* possibilita uma rápida resposta de consulta, mas as respostas podem não considerar a situação real das fontes de dados.
2. **Alto custo de implantação e manutenção:** todo o processo de extração de dados das fontes, transformação e carregamento desses dados para o repositório demanda tempo e possui um alto custo, tanto para implementar quanto para manter, principalmente se considerarmos fontes dinâmicas e em grande número.
3. **Alto custo de escalabilidade:** o alto custo de manutenção citado acima, a grande quantidade de dados de um *warehouse* são variáveis que tornam a escalabilidade dessa abordagem um processo altamente custoso.

3.4 INTEGRAÇÃO DE DADOS DE SAÚDE BASEADA EM ONTOLOGIAS

Medeiros, Oliveira e Sousa (2011) apresenta uma proposta de um sistema Web usando ontologias para auxiliar na busca e recuperação de informações de saúde, a *OntS* (Ontologia de Saúde). O sistema é desenvolvido na linguagem *JAVA*, baseado em regras e combina o uso de ontologias para facilitar a integração de bases de dados heterogêneas (SINAN e SIAB), além de permitir a disseminação do conhecimento.

Inicialmente a *OntS* se propõe a representar os conceitos relacionados à vigilância epidemiológica e os agravos de notificação compulsória, e desenvolver uma ferramenta para auxiliar aos usuários na busca e recuperação de informações com relação aos agravos. Contudo, a tendência é que essa base de conhecimentos cresça, de forma a abranger novos conceitos já armazenados nas bases de dados da secretaria de saúde.

A *OntS* foi desenvolvida utilizando a metodologia *Methontology* (FERNANDEZ; GOMEZ-PEREZ; JURISTO, 1997) e a ferramenta *Protégé*[5]. Os conceitos e as relações entre os mesmos foram obtidos através de entrevistas com experts no domínio, de levantamento bibliográfico e de alguns dados disponíveis em portais. Após a modelagem dos conceitos e relações, faz-se a instanciação, realizada de forma automática com a geração de código OWL a partir das bases de dados dos sistemas de informação em saúde.

A ferramenta é composta de uma interface de usuário, um módulo de inferências, e um gerador de instâncias. A interface de usuário é onde o usuário interage para elaborar sua consulta, exibindo também o resultado. O módulo de inferências é a parte responsável por fazer a análise na base de conhecimento. O módulo de instanciação é responsável pela geração de instâncias para a base de conhecimento, conforme citado anteriormente.

Gubiani, Port e Ornellas (2003) apresentaram um ambiente que permite a interoperabilidade de informações do Prontuário Eletrônico do Paciente baseada na interoperabilidade semântica. Metadados descrevem o significado das informações

[5] Protégé é um editor de ontologies *open-source* e um *framework* para manipulação de bases de conhecimento. Acessível através de http://protege.stanford.edu

heterogêneas e distribuídas do PEP de acordo com vocabulários específicos especialmente para descrever o PEP, e agentes utilizam estas descrições para proporcionar a busca e a manipulação das informações do PEP.

Nardon e Jr. (2004) demonstraram como a representação do conhecimento baseado em ontologias pode ser utilizada na integração de sistemas heterogêneos de saúde. Foram usadas as tecnologias RDF e DAML+OIL para representar a informação e como mecanismo de inferência foi utilizado um Sistema de Bancos de Dados Dedutivos.

Lopes, Andrade e Wangenheim (2011) desenvolveram uma ontologia capaz de representar um vocabulário único relacionado aos atendimentos de emergência. Com essa ontologia, torna-se possível que diferentes sistemas de saúde conversem entre si com o intuito de levantarem informações relevantes de pacientes, possibilitando um atendimento ágil e de qualidade.

Os trabalhos aqui relacionados possuem como desvantagens:

1. **Necessidade de um especialista para definição semântica:** para cada domínio de conhecimento se faz necessário a presença de um especialista para definição dos conceitos da ontologia;
2. **Necessidade de relacionamento entre tesauros:** na falta de um tesauro único que padroniza os conceitos de uma determinada área de conhecimento, se faz necessário fazer um relacionamento entre os diversos termos.

3.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo objetivou apresentar os principais trabalhos relacionados à integração de dados no domínio da saúde, utilizando abordagens diferenciadas: (i) Integração de dados de saúde baseada na criação de um novo esquema de dados; (ii) Integração de dados de saúde baseada na utilização de Web Services; (iii) Integração de dados de saúde baseada na abordagem de Data Warehouse; (iv) Integração de dados de saúde baseada em ontologias.

Para cada abordagem de integração, foram relacionadas as suas desvantagens. A análise dessas desvantagens é crucial para que se possa avaliar a abordagem de integração que melhor se encaixa ao LARIISA. Para a realização de

um trabalho relevante é necessário que sejam apresentadas menores desvantagens das que foram apresentadas nos trabalhos pesquisados.

Considerando o objetivo deste trabalho, de integrar dados heterogêneos, independentes e distribuídos, relacionados à saúde, para que o LARIISA aumente seu poder de inferência, a plataforma se torna mais sensível a determinadas variáveis, que devem ser consideradas e evitadas.

A escalabilidade é uma característica forte ao LARIISA. Como a plataforma deve trabalhar com fontes diversas, o processo de integração proposto deve permitir facilmente a integração por demanda. Abordagens de integração de difícil escalabilidade ou difícil manutenção devem ser evitadas.

Outra característica importante são as informações contextuais com as quais a plataforma trabalha, ou seja, o LARIISA realiza suas inferências tanto com dados históricos, quanto com dados que representem uma situação no tempo real. Abordagens de integração que possuem uma logística frequente de atualização com suas fontes, ou em tempo real, são as mais indicadas.

É importante observar que os trabalhos aqui relacionados também tiveram a finalidade de reforçar a necessidade de integração de dados no contexto da saúde pública e das dificuldades encontradas no processo. Com esse estudo prévio, espera-se obter subsídios para a formulação de contribuições relevantes para o objeto de estudo deste trabalho.

4 PERCURSO METODOLÓGICO

Neste capítulo é apresentado o percurso metodológico através do qual é construído o presente trabalho: as avaliações feitas sobre as abordagens de integração de informações, a escolha da abordagem a ser utilizada na definição do processo e as análises dos resultados obtidos por grupos de pesquisa que possuem como objeto a integração via *Linked Data*.

4.1 INTRODUÇÃO

Como visto no Capítulo 2, na arquitetura do LARIISA, definida em Oliveira *et al.* (2010), as ontologias são responsáveis pela representação das informações de contexto, formando uma base ontológica. Essa abordagem visa facilitar questões de interoperabilidade, compartilhamento e até mesmo as inferências do sistema.

Demonstra-se ao final deste capítulo que o uso de *Linked Data* se mostra a melhor abordagem de integração no contexto do LARIISA a ser utilizado no processo definido. A definição do uso de ontologias na arquitetura do LARIISA para representar as informações de contexto, também embasa a conclusão obtida, já que a tecnologia de *Linked Data* baseia-se na representação do conhecimento por ontologias. No entanto, para se chegar a essa conclusão, se fez necessário uma avaliação sobre as possibilidades tecnológicas existentes que podem agir sobre a problemática da integração de dados do LARIISA com outros sistemas e/ou outros dados de saúde.

4.2 A PROBLEMÁTICA DA INTEGRAÇÃO DE INFORMAÇÕES

A questão da integração de informações é um assunto há muito tempo discutido e pesquisado. É um dos principais entraves e desafios para o avanço da ciência em diversos campos, onde grupos de cientistas estão de forma isolada coletando dados e tentando colaborar uns com os outros, assim como se torna um desafio e entrave para os governos que desejam, por exemplo, possuir uma gestão integrada e efetiva entre seus diversos órgãos.

A problemática da integração de dados permeia principalmente na forma em que os dados são estruturados e armazenados. Vários tipos de heterogeneidade

são encontrados quando se quer extrair informações contidas em mais de uma fonte. Segundo Sheth (1999), podemos encontrar as seguintes heterogeneidades: (i) **Heterogeneidade de Sistemas**; (ii) **Heterogeneidade Estrutural**; (iii) **Heterogeneidade Sintática**, e; (iv) **Heterogeneidade Semântica**.

A **heterogeneidade de sistemas** está relacionada às diferenças entre arquiteturas de sistemas, tipo de hardware, sistemas operacionais, modelo de dados, etc. **Heterogeneidade estrutural** está relacionada à esquema de dados distintos. Por exemplo, esquemas de dados orientados à objetos suportam generalização e outros, não.

A **heterogeneidade sintática** ocorre quando há diferenças no formato de representação de um dado. Por exemplo, uma base de dados pode representar o atributo “sexo” como “F” ou “M”, enquanto em outra base de dados seja representado por 0 ou 1. Já a **heterogeneidade semântica** diz respeito ao significado dos dados armazenados, como por exemplo, a palavra “manga” pode ter o significado de uma fruta em uma base de dados e de parte de uma camisa em outra.

Com o passar dos anos, com os desafios que se apresentavam no contexto de cada época, cada tipo de heterogeneidade de dados citada passa por uma fase diferente de busca de soluções. Nos anos 80, os trabalhos estavam concentrados em se alcançar a interoperabilidade de sistemas, principalmente devido à diferenças em SGBD’s, com trabalhos que abordavam a heterogeneidade sintática e estrutural.

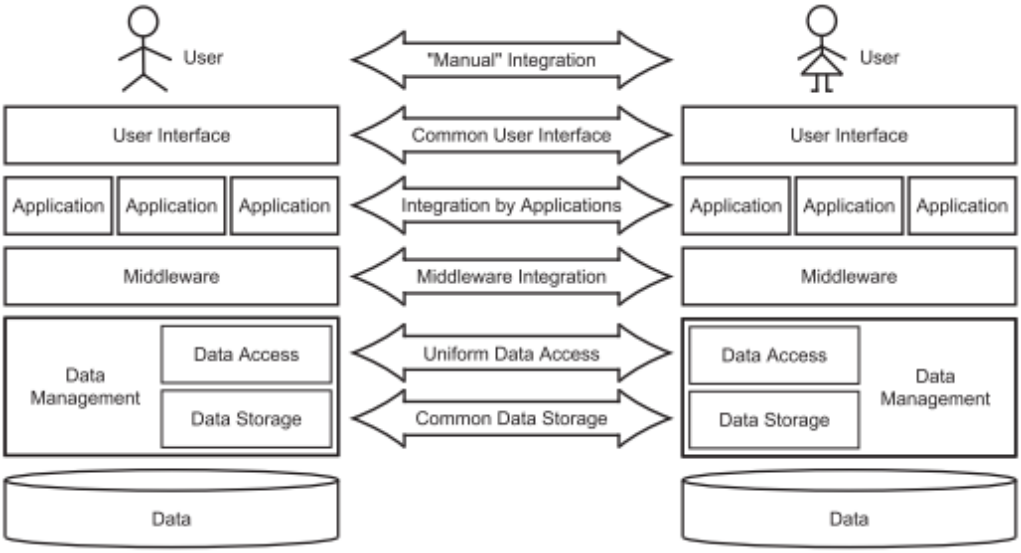
Mais tarde, houve duas grandes tendências que trouxeram grandes oportunidades para a interoperabilidade e operação de dados: (i) a grande diversidade de tipos de dados, dos estruturados aos semiestruturados, às mídias digitais, e; (i) a proliferação da Web. Teve-se a exploração dos meta-dados; da criação de padrões; da aceitação da Internet como padrão para a interconexão entre sistemas; da evolução de infraestruturas e middlewares que suportassem sistemas distribuídos (RMI, CORBA). Interoperabilidades sintáticas incluíram, por exemplo, a formatação e exportação de dados que fossem suportados por padrões como HTML, MPEG-1, etc.; a nível estrutural, padrões para modelagem de dados surgiram, como ANSI SQL e UML; interoperabilidade estrutural e uma limitada interoperabilidade semântica foram alcançadas através do *Dublin Core* e outros padrões de meta-dados.

Atualmente, com o progresso da interconectividade global, a escala do problema mudou de uns poucos bancos de dados para milhões de informações, que geram outros tantos milhões de informações. Estratégias que dependam do acesso baseado em palavras-chave ou envolvam apenas componentes de dados estruturais ou representacionais apresentam geralmente um resultado pobre e impreciso. Os desafios principais a serem enfrentados estão no nível semântico, onde os usuários esperam que os sistemas de informação os ajude não somente a nível de dados, mas a nível de informação, aumentando assim seus níveis de conhecimento.

Já com relação ao nível de arquitetura em que se trabalha a integração, (ZIEGLER; DITTRICH, 2007) apresentam 6 abordagens gerais:

i. a **integração manual**, onde os usuários precisam lidar com diferentes interfaces de usuários e linguagens de consulta para realizar a integração das informações que deseja;
ii. a integração facilitada por uma **interface comum ao usuário**: tem-se a facilidade de se ter uma interface comum para o usuário (como, por exemplo, um *web browser*) para que ele possa colher a informação, mas como na integração manual, o usuário será o responsável pela homogeneização e integração dos dados;
iii. **integração por aplicações**: abordagem que usa aplicações integradoras que acessam vários conjuntos de dados e retornam resultados integrados para o usuário;
iv. **integração por middleware**: aplicações com funções reutilizáveis geralmente utilizadas para resolver aspectos específicos de integração;
v. **acesso uniforme aos dados**: neste caso, tem-se uma integração lógica dos dados. Os dados locais distribuídos podem manter sua autonomia ;
vi. **armazenamento de dados comum**: os dados são fisicamente integrados e armazenados em novo local.

Figura 14 - Abordagens gerais de integração em diferentes níveis de arquitetura



Fonte: (ZIEGLER; DITTRICH, 2007)

Cada abordagem mencionada por Ziegler e Dittrich, (2007) suporta diferentes técnicas, como, por exemplo: (i) **integração por *web services,*** que realiza integração de sistemas através de componentes de softwares na Internet, representando uma abordagem uniforme de acesso aos dados para a realização de uma *integração manual* ou de uma *integração por aplicações;* (ii) **integração por esquema único de dados,** que representa uma solução de integração do tipo *manual*; (iii) **integração por data warehouses**, que realiza um *armazenamento de dados comum;* (iv) **integração baseada em mediadores**, que representa uma solução do tipo *acesso uniforme aos dados*, disponibilizando um ponto único de consulta a várias fontes de dados.

Neste trabalho, são investigados e comparadas as quatro técnicas de integração citadas. Na integração por mediadores, avalia-se especificamente a utilização de ontologias como forma de mediação visto que a interoperabilidade semântica se mostra como principal tendência atualmente, como já citado. Tenta-se avaliar, então, se essa tendência também é verdadeira para as demandas do LARIISA.

4.3 INTEGRAÇÃO POR WEB *SERVICES*

A integração realizada a nível de aplicação, ou integração de sistemas, facilita a comunicação entre sistemas, independente da linguagem de programação utilizada para desenvolvê-lo ou da plataforma sobre a qual o sistema executa. Um contexto muito comum é quando se precisa integrar aplicações legadas e autônomas para suportar processos de negócios, e os desenvolvedores não possuem conhecimento sobre como os aplicativos funcionam ou sobre como usar tecnologias antigas e/ou proprietárias.

Os trabalhos realizados por Hansen, Madnick e Siegel (2003) e Umapathy e Purao (2010) apresentam uma relação dos Web Services como uma tecnologia útil para superar esses desafios da integração de sistemas.

> Seguindo os princípios SOA, os Web Services provêm mecanismos que possibilitam aplicações comunicarem umas com as outras independente de linguagens de programação e plataforma. Eles provêm uma interface padrão para aplicativos e protocolos de Internet comunicarem com serviços disponíveis (UMAPATHY; PURAO, 2010, tradução nossa).

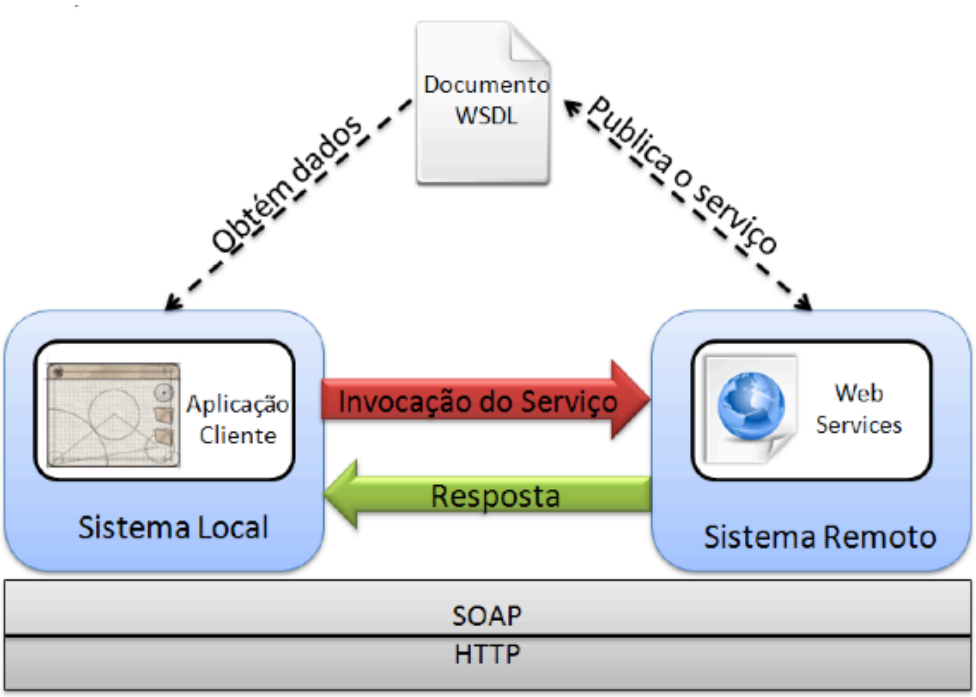
ALONSO et al. (2004) conceitua Web Services como uma forma de expor as funcionalidades de um sistema de informação, fazendo-o disponível através de padrões Web. O uso de padrões é a chave para a interoperabilidade, consequentemente facilita a integração de sistemas. Tais padrões são: (i) dados são trocados entre sistemas utilizando o padrão **XML**; (II) **SOAP** é usado para enviar e receber documentos XML; (iii) Serviços de integração são especificados utilizando **WSDL**; (iv) Um registro de todos os serviços é publicado utilizando **UDDI**. Na Figura 15 a seguir, é apresentada a arquitetura dos Web Services.

Apesar da frequente utilização de *Web Services* na integração de sistemas no domínio da saúde e das vantagens na utilização de *Web Services* como tecnologia de integração, há também desvantagens, que para as características de integração exigidas pelo LARIISA, podem não ser interessantes. Dentre elas cabe destacar:

- **Dependência do provedor de serviços:** o principal objetivo de se integrar dados heterogêneos ao LARIISA, é dar maiores subsídios ao processo de inferência da plataforma. Quanto mais acesso a dados o LARIISA possuir, melhor poderá ser a inferência realizada. A limitação do uso de Web Services, nesse sentido, é que os clientes do serviço

apenas obtém acesso aquilo que o provedor do serviço pode oferecer. Muito diferente de quando se obtém dados brutos para integração, como acontece com outras abordagens. Caso o cliente necessite de uma nova informação que o serviço atualmente não ofereça, o provedor poderá se negar a reescrever o serviço para atender aos requerimentos de um cliente específico;

Figura 15 - Arquitetura dos Web Services



Fonte: (CASTAÑEDA, 2011)

- **Questões semânticas:** a integração via *web services* não garante que diferentes serviços tenham a mesma semântica associada a seus dados. Problemas simples de semântica podem estar relacionados a diferentes padrões de unidades numéricas utilizadas, como por exemplo, uma empresa de telecom que requisite de suas filiais a largura de banda consumida por um determinado cliente. O *Web* Service de uma das filiais pode responder em Mbps, outro *web service* de outra filial pode responder em bps, ou seja não existe um padrão estabelecido. Uma abordagem capaz de resolver essa questão é a utilização de um mediador de contexto que identificaria e resolveria

potenciais conflitos semânticos entre o cliente e o provedor do serviço Web (HANSEN; MADNICK; SIEGEL, 2003).

Mesmo avaliando as desvantagens de utilização da integração via *Web Services* para o LARIISA, pode haver situações em que sua utilização não estaria de um todo descartado. *Web Services* pode ser utilizado como forma de obtenção de dados de uma determinada fonte de dados. Por exemplo, uma fonte de dados do SUS pode oferecer como único método de obtenção de dados um Web *Service.* Nesse caso, o LARIISA utiliza esse *Web Service* de terceiro apenas para captar dados e realiza a integração de dados com outras abordagens.

4.4 INTEGRAÇÃO POR CRIAÇÃO DE UM ESQUEMA DE DADOS ÚNICO

Essa abordagem objetiva, através da criação de um único esquema de dados, substituir os diversos esquemas existentes que se encontram pulverizados em sistemas de informação diversos. O esquema de dados resultante da integração é necessário, por exemplo, para representar os requerimentos de uma aplicação. Essa abordagem de integração de dados foi encontrada principalmente em trabalhos relacionados ao SUS, que como já citado, possui uma situação peculiar de ter vários SIS que não conversam uns com os outros. Essa característica acaba por resultar em vários dados duplicados, que foram modelados de uma forma em um sistema e de outra forma em outro sistema, mas que significam a mesma coisa.

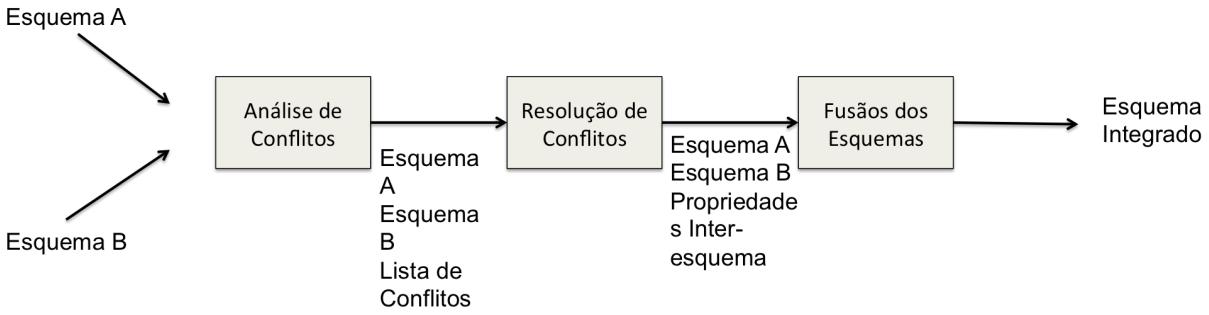
Nessa abordagem, o processo de integração dos esquemas envolve três principais estágios:

- Análise de conflitos: nessa etapa, diferenças entre os esquemas são identificadas. Por exemplo, conceitos similares que são representados em diferentes formas;
- Resolução de conflitos: nessa etapa, os conflitos identificados na etapa anterior são resolvidos. Por exemplo, um método único de representação de conceitos semelhantes deverá ser decidido. Esse processo pode envolver a discussão dos problemas com os usuários ou corrigindo erros nos esquemas.

- Fusão dos esquemas: nessa etapa, os esquemas são fundidos em um esquema único usando as decisões feitas durante a etapa de resolução de conflitos.

A Figura 16 ilustra o processo.

Figura 16 - Processo de Integração por Criação de um esquema único de dados



Fonte: Elaborado pelo autor.

Como já comentado, foram encontrados alguns trabalhos que utilizaram essa abordagem para realizar a integração de informações na área da saúde, como mencionado no capítulo 3.

Esse tipo de abordagem de integração possui muitas desvantagens. Senão, veja-se:

- Um alto custo de implantação, principalmente se for considerado uma quantidade grande de bases de dados a ser integrada;
- Uma abordagem de difícil escalabilidade, em situações onde há a previsão de integração constante de outros esquemas de dados;
- O fato dos dados já estarem adequadamente manipulados pelos sistemas legados, e algumas aplicações necessitariam ser reescritas para trabalhar com este novo sistema integrado, o que não é considerado prático;
- As heterogeneidades semânticas nessa abordagem de integração também não são resolvidas;

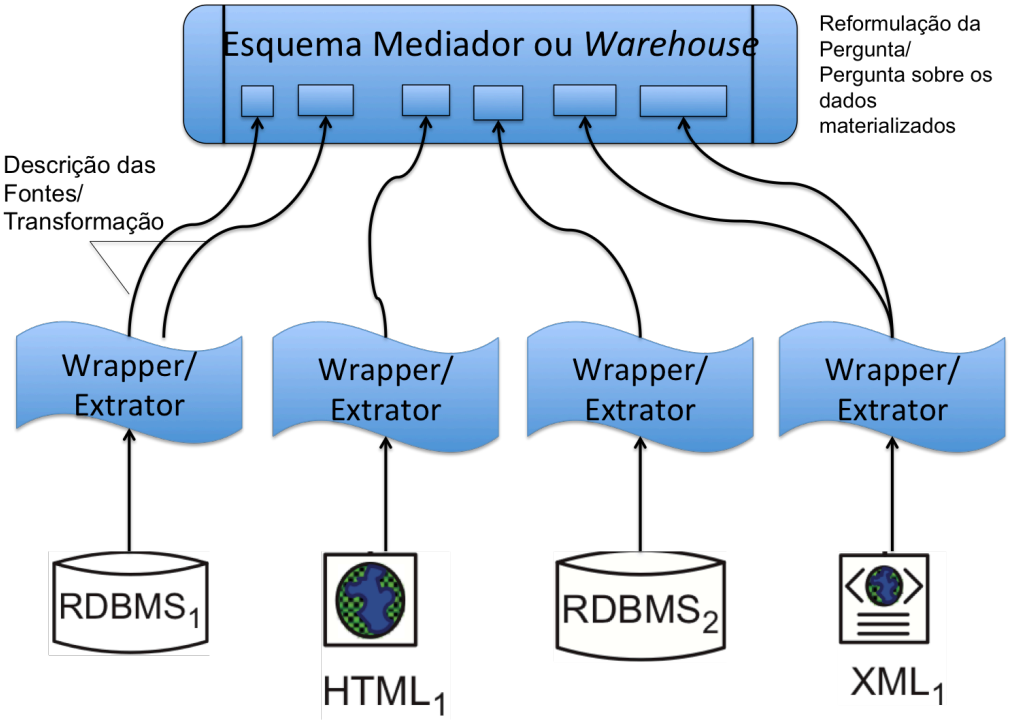
Essa abordagem de integração é mais utilizada em situações onde se já tem definido quais esquemas de dados serão integrados, sem nenhuma expectativa de integrar esquemas de dados posteriormente.

Essas características se tornam incompatíveis para a realidade do LARIISA, onde espera-se a integração de diversas fontes de dados, independente se elas acontecerão em um mesmo momento ou de forma escalável. Integração por criação de um esquema único de dados é uma abordagem mais apropriada quando se tem o objetivo que o esquema de dados integrado sirva de base de dados para uma aplicação específica.

4.5 ABORDAGEM MATERIALIZADA E VIRTUALIZADA DE INTEGRAR DADOS

Existem várias possibilidades de abordagens para integração de dados, mas de forma geral, pode-se colocar que a maioria dos sistemas optam pela abordagem **materializada** ou **virtualizada**. Na abordagem materializada os dados são carregados de fontes individuais e materializados em um banco de dados físico, os *data warehouses*, para onde são direcionadas as consultas. Na abordagem virtualizada, os dados continuam em suas fontes e são acessados em tempo de execução das consultas, através de um esquema mediador.

Figura 17 - Arquitetura básica de um sistema de propósito geral de integração de dados



Fonte: Adaptado de Doan et al. (2012).

A Figura 17 mostra os componentes lógicos de ambas as abordagens de integração de dados. Iniciando pela parte inferior da figura, tem-se as **fontes de dados**. Essas fontes de dados podem variar desde o tipo de modelo de dados até ao tipo de consulta que elas suportam. Podem ser, por exemplo, dados do tipo relacional, XML ou quaisquer outros dados estruturados.

Acima das fontes de dados estão os ***wrappers*** para a abordagem virtualizada, ou **extrator** na abordagem materializada. Esse elemento está relacionado à programas cujo papel é comunicar com as fontes de dados, mandando consultas, recebendo as respostas e possivelmente aplicando até alguma transformação nessas respostas.

Os usuários interagem com o sistema integrado de dados através de um único esquema, chamado de **esquema mediador**. O esquema mediador é construído para a aplicação e apenas contém os aspectos de domínio relevantes da aplicação. Dessa forma, não se faz necessário conter todos os atributos presentes nas fontes de dados. Na abordagem virtualizada, o esquema mediador não armazena nenhum dado, o que não acontece na abordagem materializada. Na abordagem virtualizada, o esquema mediador é puramente lógico, utilizado apenas para receber as consultas dos usuários (ou aplicações) empregadas para o sistema de dados integrados.

Os **descritores das fontes de dados** (ou **transformações** na abordagem materializada) podem ser considerados os principais elementos da arquitetura de dados integrados. Eles conectam o esquema mediador aos esquemas de fontes de dados. Esses descritores especificam as propriedades das fontes que o sistema precisa saber em ordem para usar esses dados. Os principais elementos desses descritores são os **mapeamentos semânticos** realizados entre os dicionários de dados das fonte de dados e do esquema mediador.

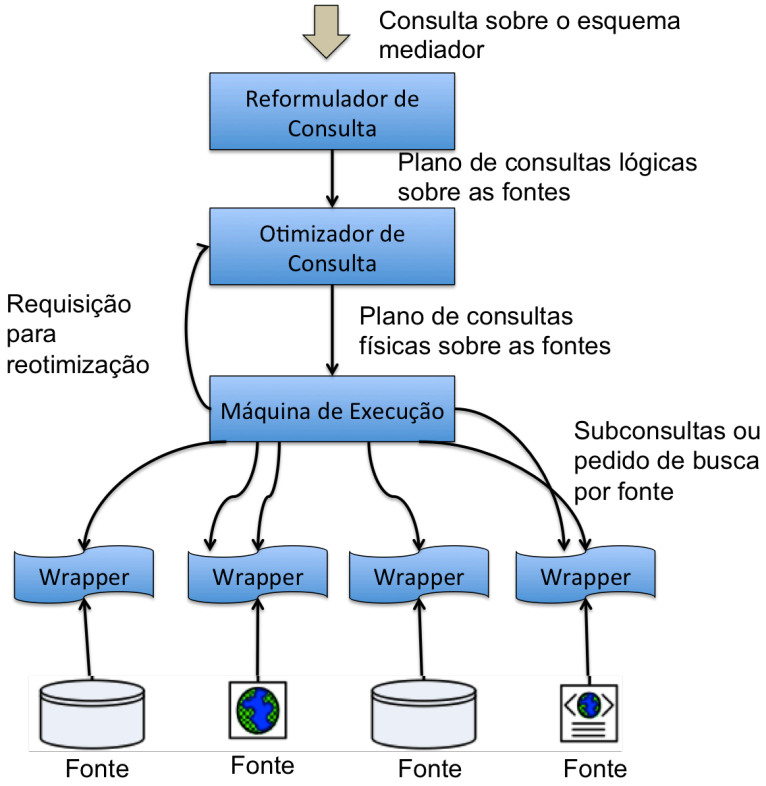
O processamento de consultas que ocorre nesse esquema difere de um banco de dados tradicional de duas maneiras: primeiro as consultas que chegam ao esquema mediador precisam ser reformuladas para que cheguem as fontes de dados; segundo, a execução das consultas precisa estar preparada para se adaptar, já que o plano de execução da consulta pode mudar à medida que a consulta está sendo executada. A Figura 18 ilustra o processo de consultas em um sistema integrado de dados.

4.5.1 Integração Por *Data Warehouse*

Como citado por Doan, Halevy e Ives (2012) *data warehouse* é uma das arquiteturas de integração de dados mais utilizadas por empresas que precisam manter um histórico de seus dados, para auditoria ou análise, e para auxiliar a tomada de decisão, através de aplicações *Online Analytic Processing* (OLAP) – aplicações que auxiliam a tomada de decisão a partir das características de dados integrados -, por exemplo.

Nesse modelo, todos os dados necessários são traduzidos para um esquema alvo e copiado para um sistema gerenciador de banco de dados (que pode ser do tipo paralelo ou distribuído). Os dados passam por uma rotina de atualização, a depender de cada estratégia de manutenção. Definir um *data warehouse* envolve duas tarefas principais: desenvolver o esquema central de banco de dados e o designe físico da arquitetura, e definir o conjunto de operações de Extrair/Transformar/Carregar (ETL – Extract/Transform/Load). A Figura 19 apresenta como se estabelece a arquitetura de um *data warehouse*.

Figura 18 - Processamento de consultas em um sistema integrado de dados

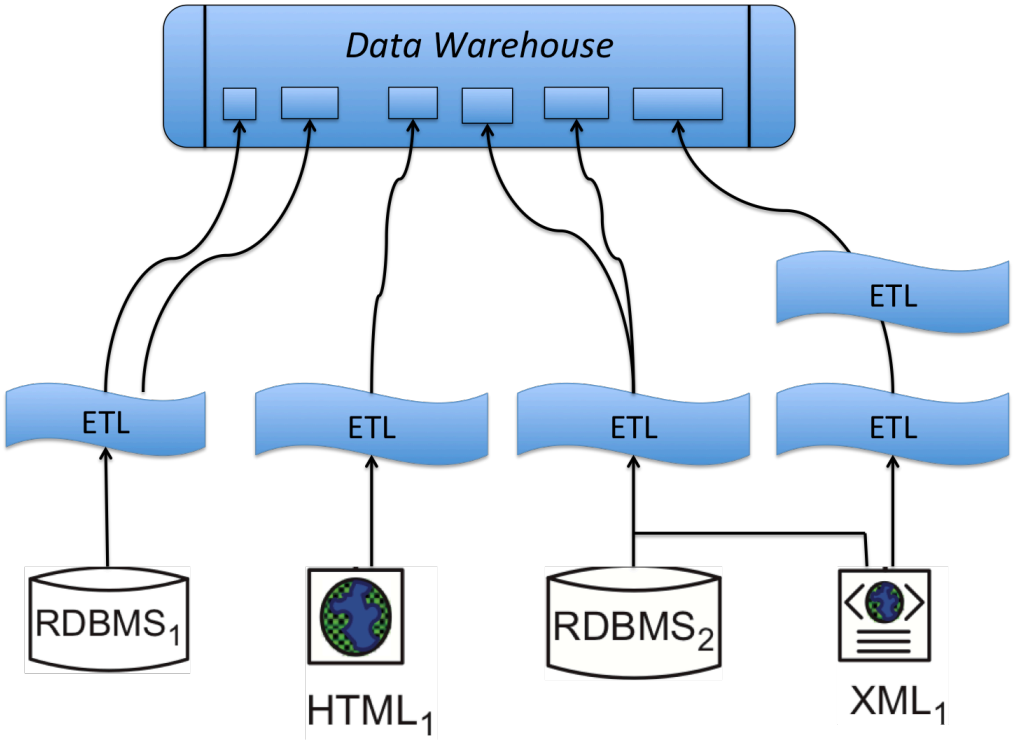


Fonte: Adaptado de Doan et al. (2012).

As fontes de dados podem ser sistemas de arquivos, banco de dados, documentos HTML, documentos XML, entre outras. Conectado com cada fonte de dados estão os ETL's, componentes de software cuja função é a extração de dados das fontes, transformação desses dados conforme regras de negócios e por fim a carga dos dados no *data warehouse.* Os usuários submetem suas consultas diretamente ao *data warehouse* e ali são processadas, não havendo nenhuma interação com as fontes de dados. A depender da implementação do *data warehouse* podem haver também monitores, que detectam automaticamente modificações nas fontes, repassando as alterações relevantes.

Como citado por Gupta e Mumick, (1995) e Widom (1995) um dos principais problemas a ser considerado na arquitetura de *data wareshouses* diz respeito à manutenção do repositório de dados, estando sempre consistente com os dados das fontes, considerando que essas fontes de dados continuem ativas, podendo assim sofrer alterações. Neste caso, existem duas abordagens para a manutenção da consistência:

Figura 19 - Componentes Lógicos de um *Data Warehouse*



Fonte: Adaptado de Doan et al. (2012).

- **Rematerialização de visão:** o conteúdo do *data warehouse* é descartado e a visão é novamente materializada com os novos dados das fontes. É considerado um método muito custoso uma vez que elimina todo o repositório e cria um novo;
- **Manutenção incremental:** as alterações nas fontes de dados são propagadas incrementalmente para o *data warehouse* (BATISTA, 2003).

Apesar da existência de vários cases de utilização de *data warehouses* como abordagem de integração de dados na área da saúde, como visto no Capítulo 3, essa abordagem possui características não muito vantajosas para o contexto do LARIISA, como descrito a seguir:

- **Atualização dos dados no repositório:** como já citado previamente, a manutenção do repositório é um dos principais problemas na abordagem de *data* warehouse. Considerando o dinamismo das fontes de dados que serão integradas ao LARIISA, assim como o grande número de fontes de dados utilizadas, a abordagem de *data warehouse* pode não ser a mais indicada. O *data warehouse* possibilita uma rápida resposta de consulta, mas as respostas nunca são condizentes a situação real das fontes de dados. No contexto onde existem fontes de dados que mudam constantemente com o tempo e em alto número, manter o repositório de dados sempre atualizado, torna-se um grande desafio, principalmente porque todo o processo de extração de dados das fontes, transformar e carregar para o repositório demanda tempo e também possui um alto custo, tanto para implementar quanto para manter;
- **Essência da abordagem de *data warehouse* para dados históricos*:*** como mencionado no primeiro parágrafo desta seção, a abordagem de integração de dados por *data warehouse* atende, em sua maioria, necessidades específicas, geralmente de empresas que desejam manter dados históricos de suas transações para realização de análise. No contexto do LARIISA, as fontes de dados a serem integradas não necessariamente se constituem de dados históricos, mas de fontes de dados que armazenem informações de contexto, que para a realização

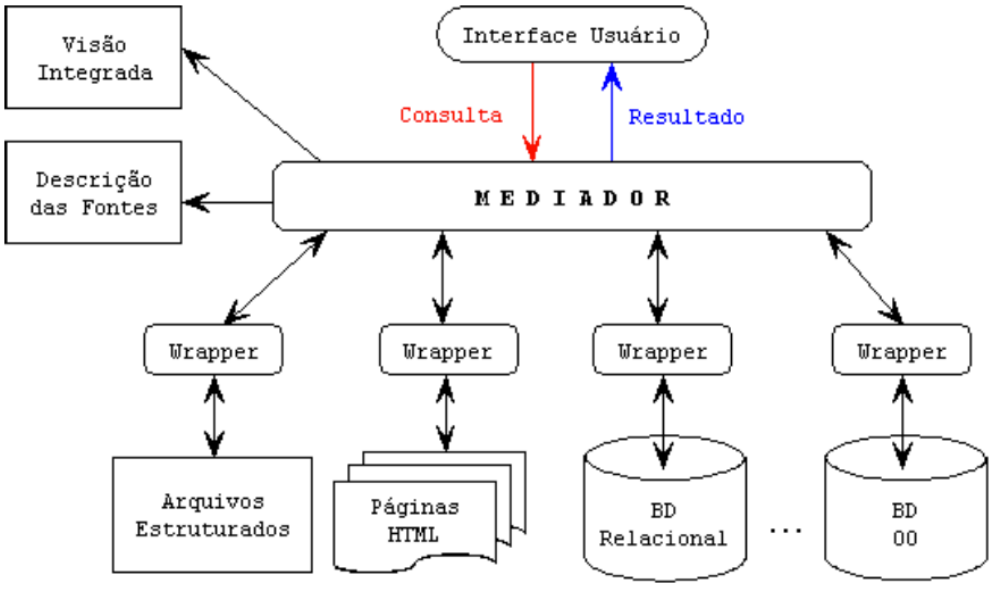
de inferência pode ser necessário uma análise histórica ou apenas informações de outros provedores de contexto integradas.

4.5.2 Integração Baseada Em Mediadores

De forma geral, a arquitetura baseada em mediadores se assemelha a arquitetura básica de um sistema integrador mostrado na Figura 17. No trabalho de Wiederhold (1992), os mediadores são definidos como a interface entre as aplicações de usuários e as fontes de dados, tornando as aplicações independentes dessas fontes.

Como reforçado na Figura 20, o mediador possui uma visão integrada das fontes de dados, assim como uma descrição dessas fontes. O usuário, a partir de uma interface, submete consultas ao mediador, que por sua vez decompõe essas consultas em sub-consultas menores direcionadas às fontes de dados. Essas sub-consultas devem passar pelos *wrappers,* que realizam a tradução dessas sub-consultas no formato e linguagem que as fontes individuais suportam. Quando as fontes individuais respondem às sub-consutas, os *wrapper* também possuem a função de traduzir tais respostas em um modelo de dados comum, que é compreendido pelo mediador.

Figura 20 - Arquitetura de Integração de Dados Baseada em Mediadores



Fonte: (BATISTA, 2003)

A arquitetura de mediadores possui como principal problemática o custo e o tempo elevados para se processar consultas em tempo real, onde se fixam as principais pesquisas na área.

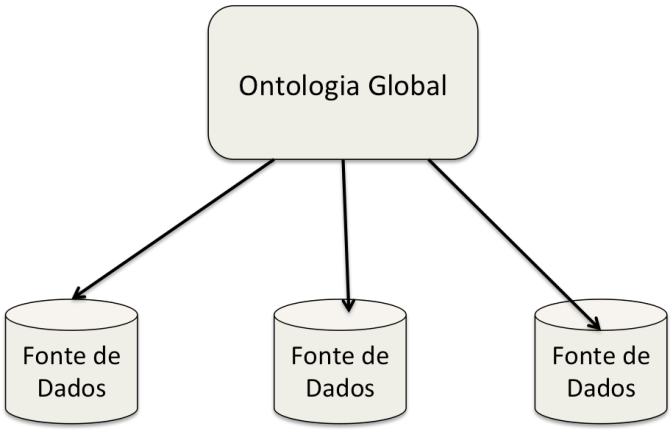
4.5.2.1 Integração De Dados Baseada Em Mediadores Utilizando Ontologias

Muitos aspectos da integração de dados também podem ser vistos através da lógica da representação do conhecimento, com a modelagem de relacionamentos entre fontes de dados ou sobre fontes de dados e esquema mediador, tendo as ontologias como principal ferramenta para realização dessa modelagem.

Através de ontologias é possível descrever explicitamente a semântica das fontes de informação, possibilitando a identificação e associação dos conceitos semanticamente correspondentes entre as fontes. Conforme apresentado por Wache et al. (2001), existem ainda três possíveis formas diferentes de realizar a descrição semântica dessas fontes. Como visto a seguir:

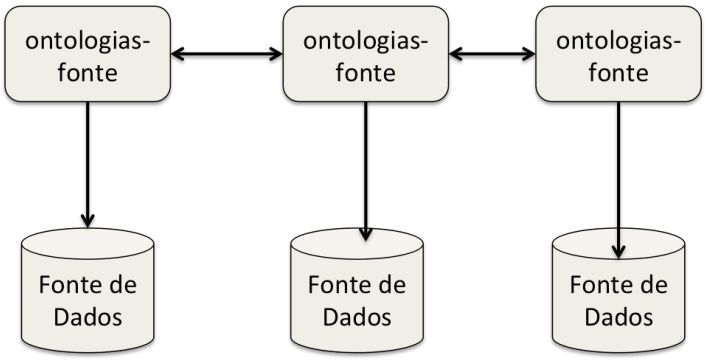
- **Ontologia única –** essa abordagem utiliza uma ontologia global que provê um vocabulário compartilhado para a especificação da semântica do domínio que pertencem as fontes. É um tipo de abordagem utilizada quando as fontes de informação compartilham da mesma visão em um domínio, inclusive a nível de granularidade. Essa abordagem é muito susceptível a mudanças nas fontes de dados que afetem a conceitualização do domínio representado pela ontologia. Tais mudanças podem significar alterações na ontologia global, assim como nos mapeamentos feitos entre as outras fontes de dados;
- **Múltiplas Ontologias –** nessa abordagem cada fonte de dados é descrita por sua própria ontologia, o que facilita os casos de mudança, quando há a necessidade de alteração, adição ou remoção de fontes. Por outro lado, a falta de um vocabulário comum torna extremamente difícil a comparação entre essas ontologias-fonte. Como forma de superar esse problema, há a necessidade de se especificar também mapeamentos entre ontologias, que identifica termos semanticamente correspondentes entre as ontologias-fonte. A Figura 22, representa a abordagem baseada em múltiplas ontologias descrita.

Figura 21 - Abordagem de descrição semântica por ontologia única



Fonte: Adaptado de (Wache *et al.*,2001).

Figura 22 - Abordagem de descrição semântica por múltiplas ontologias

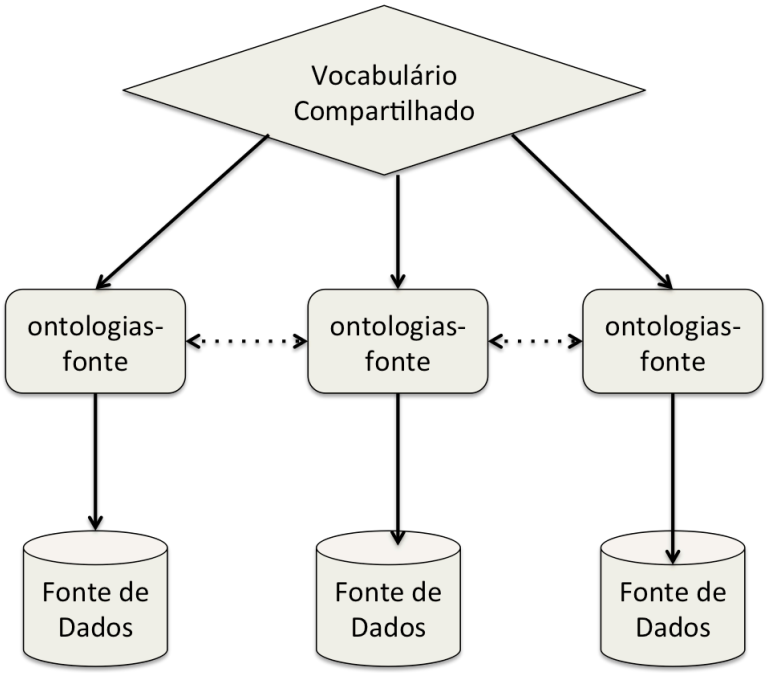


Fonte: Adaptado de (WACHE *et al.*, 2001).

- **Abordagem Híbrida –** essa abordagem intenciona superar os problemas das duas abordagens anteriores. É similar a abordagem de múltiplas ontologias, onde cada fonte é descrita por sua própria ontologia, mas como na abordagem de ontologia única, cada ontologia fonte é construída sob um vocabulário único compartilhado. Esse vocabulário compartilhado, que também pode ser descrito por uma ontologia, contém vocabulários primitivos que definem o domínio. As ontologias-fonte, que possuem termos mais complexos, são construídas a partir do vocabulário compartilhado. Como vantagem dessa abordagem híbrida, destaca-se a possibilidade de facilmente adicionar novas fontes. A desvantagem é que as ontologias preexistentes não podem ser reutilizadas com facilidade, pois todas

as ontologias-fonte devem referir-se a um vocabulário comum. A Figura 23 a seguir ilustra essa abordagem híbrida.

Figura 23 - Abordagem híbrida de descrição semântica baseada em ontologias.



Fonte: Adaptado de (WACHE *et al.*, 2001)

As ontologias se destacam não só na descrição semântica de fontes de dados, mas Wache et al. (2001) explica que podem ser encontrados outros papéis para as ontologias. Por exemplo, as ontologias podem ser utilizadas como **modelo de consulta** e para **verificação** das descrições de integração, que podem ter sido definidas tanto por usuários como por algum processo automatizado.

No contexto do LARIISA, a abordagem de integração por interoperabilidade semântica, ou seja, através de ontologias, é a mais interessante, analisando-se as abordagens de integração apresentadas até aqui, e as características intrínsecas à plataforma. Como já discutido no Capítulo 3, o LARIISA exige alta escalabilidade, baixo custo de manutenção e base de dados integrada sempre atualizada com as fontes de dados. Além do fato do LARIISA já utilizar ontologias na modelagem das informações de contexto, tem-se nessa abordagem uma ampla utilização por parte da comunidade acadêmica, até por conta do crescimento dos conceitos da *Web semântica*, onde as ontologias são tendências de uso na hora de modelar os conceitos de um determinado domínio.

4.6 A ENTRADA DO CONCEITO DE *LINKED DATA* AO LARIISA

Como parte do processo investigativo, foram realizados encontros ao grupo de pesquisa ARIDA (*Advanced Research in DAtabase*), instalado no LIA (Laboratórios de Pesquisa em Ciência da Computação da UFC). Nas ocasiões fomos apresentados aos trabalhos da Prof. Dra. Vânia Maria Ponte Vidal[6], que possui grande atuação nas áreas de integração semântica, integração de dados na web e descoberta de conhecimento em dados de mobilidade.

Nesses encontros, surgiu a oportunidade de conhecer as possibilidades que a tecnologia de *Linked Data* pode proporcionar à integração de dados na área da saúde, já considerando que a utilização de ontologias seria a abordagem mais recomendada para o LARIISA. Roberval Mariano[7], integrante do grupo de pesquisa, e orientando de doutorado da Prof.ª Vânia, apresentou sua pesquisa, relacionado a integração de dados governamentais abertos no domínio de compras municipais com o intuito de fiscalizar arbitrariedades legais, por exemplo, fornecedores considerados inidôneos fornecendo para a administração pública.

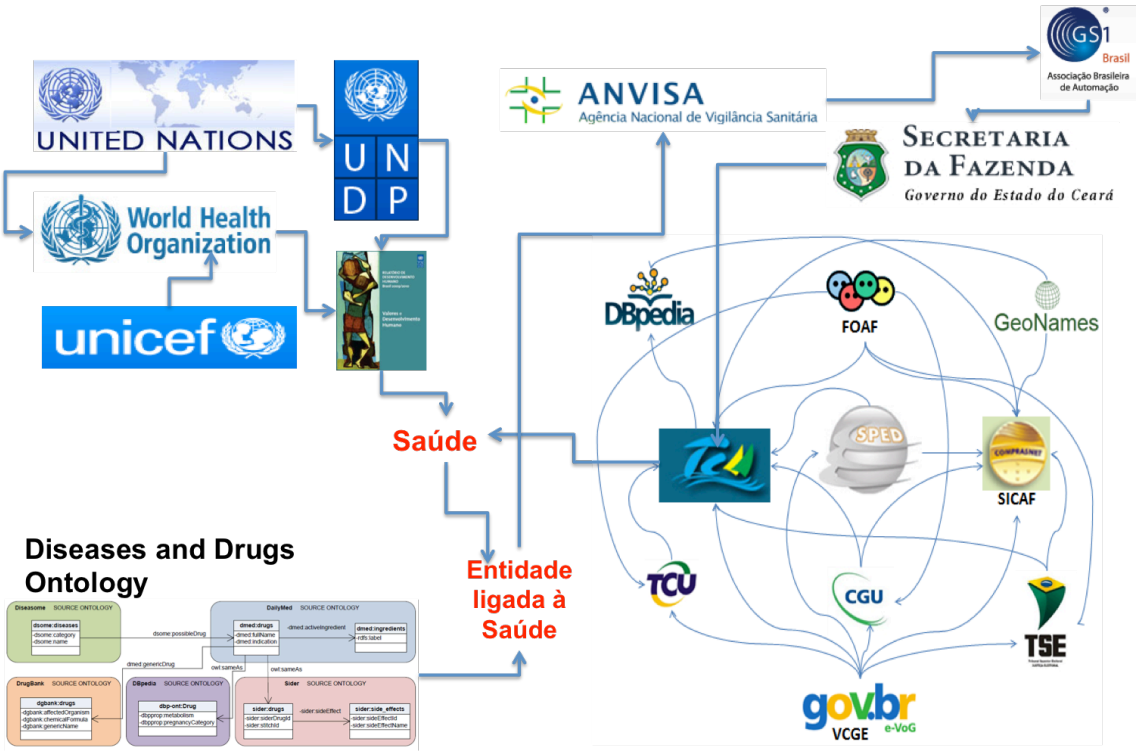
Relacionando seu trabalho com os objetivos do LARIISA, Mariano apresentou um exemplo de integração de fontes de dados no domínio de saúde via *Linked Data* obtendo, dessa forma, informações que poderiam instruir melhor decisões do gestor de saúde. Essas fontes de dados poderiam ser: Nações Unidas, OMS, UNICEF, Ministério da Saúde, Secretarias de Saúde (estado e município) e outras, que já fornecem dados abertos. Outras fontes não diretamente ligadas à saúde também poderiam ser integradas, como a ANVISA, que disponibiliza, para consulta, lista de preços de medicamentos para compras públicas, que contém o teto de preço pelo qual entes da Administração pública podem adquirir medicamentos, e a GS1, com seus padrões de identificação de produtos.

A Figura 24, mostra a relação das fontes citadas no exemplo proposto por Mariano. Nessa integração, tem-se: as fontes relacionadas diretamente com as entidades ligadas à saúde ou ao desenvolvimento humano (UNICEF, UNITED NATIONS, World Health Organization, PNUD), que pode fornecer parâmetros e/ou indicadores, assim como outras informações de saúde de esfera mundial; uma modelagem ontológica que relaciona doenças e suas drogas; a ANVISA que

[6] http://lattes.cnpq.br/9431229866203038
[7] http://lattes.cnpq.br/9428387943972577

disponibiliza preços de remédios para a administração pública; a GS1, que através de seus códigos, pode-se identificar o remédio, e suas apresentações (cartela ou caixa, por exemplo); e demais fontes já utilizadas para fiscalização de compras municipais.

Figura 24 - Possibilidades de Integração de fontes abertas no domínio de saúde



Fonte: Informação verbal[8].

Nesse cenário, pode-se obter respostas às questões:

1. A administração pública está adquirindo remédios adequados às doenças? (Através da ontologia de drogas e doenças pode-se saber quais drogas ou medicamentos estão aptos a tratar uma determinada doença);
2. O preço do remédio adquirido pela administração pública está condizente com a sua apresentação? (Através das informações da ANVISA e GS1 pode-se verificar se a administração pública não está comprando uma cartela de remédios pelo preço de uma caixa);

[8] Informação fornecida por R. Mariano, como resultado do seu trabalho de doutorado ainda em conclusão.

3. A administração pública está adquirindo remédios a um preço adequado? (Através das informações da ANVISA pode-se verificar se a administração pública não está adquirindo remédios superfaturados).

A partir dessa nova ótica, das possibilidades que o *Linked Data* apresenta*,* considera-se essa tecnologia em conjunto com uma abordagem baseada em mediadores e ontologias, como a melhor estratégia de integração de dados para o LARIISA, mesmo cientes que no contexto do LARIISA as fontes de dados podem não estar disponibilizadas em um padrão aberto, o que não incapacita a utilização da tecnologia, como mostrado no Capítulo 5.

4.7 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou os caminhos percorridos para se obter os objetivos definidos por esta dissertação. Primeiramente, foi realizada uma investigação sobre os principais métodos de integração de informações, considerando as heterogeneidades de dados. Após análise das técnicas pesquisadas, que envolveram integração via *Web Services,* integração por criação de um esquema único de dados, integração por *data warehouse* e integração por utilização de mediadores, conclui-se que a técnica de integração pela utilização de mediadores é a mais adequada para utilizarmos no processo proposto, considerando o contexto de aplicação no LARIISA, com destaque para a realização de mediação através de ontologias.

Considerando a necessidade de se aumentar a expertise na integração de dados baseada em ontologias, foram abordadas também as visitas realizadas ao ARIDA como parte do processo investigativo. Através do conhecimento de trabalhos relevantes dentro da área de integração semântica, integração de dados na web e descoberta de conhecimento em dados de mobilidade, foi possível conhecer a tecnologia de *Linked Data*, que também utiliza ontologias para mediar a integração, e as vantagens em sua utilização, através de exemplos de aplicabilidade da tecnologia ao LARIISA.

No próximo capítulo é apresentada a descrição do modelo de integração proposto.

5 PROCESSO DE INTEGRAÇÃO DE DADOS PARA O LARIISA

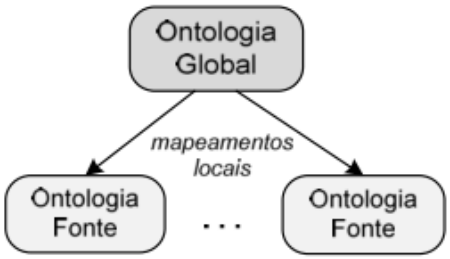
Este capítulo apresenta a especificação do processo de integração de dados proposto para o LARIISA, utilizando uma abordagem baseada em mediadores, utilizando ontologias e *Linked Data*, levando em consideração fontes de dados heterogêneas.

5.1 ARQUITETURA DE MEDIAÇÃO DE TRÊS NÍVEIS BASEADO EM ONTOLOGIAS PARA INTEGRAÇÃO DE DADOS NO PADRÃO *LINKED DATA*

Segundo Sacramento et al. (2010), há duas principais arquiteturas para integração de dados baseado em ontologias para especificação dos mapeamentos entre ontologias: de dois níveis e de três níveis.

A arquitetura de dois níveis (Figura 25), possui como componentes: a ontologia de domínio ou global, que contém os termos essenciais de um domínio; as ontologias-fonte, que descrevem as fontes de dados usando uma linguagem de ontologia; e os mapeamentos, que especificam a correspondência entre as ontologias-fonte e a ontologia de domínio.

Figura 25 - Arquitetura de dois níveis baseada em ontologias

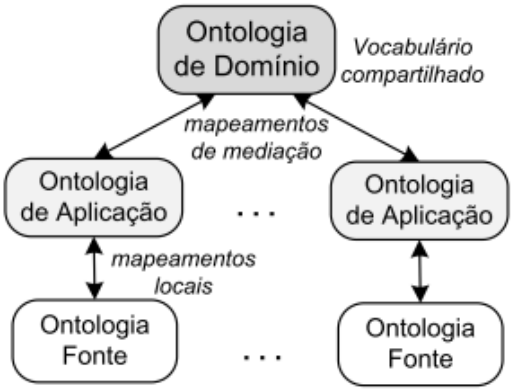


Fonte: (PINHEIRO, JOÃO CARLOS, 2011).

A arquitetura de três níveis (Figura 26), possui como componentes, além da ontologia de domínio e das ontologias-fonte: ontologias de aplicação, que descrevem as ontologias-fonte usando um subconjunto do vocabulário da ontologia global; o mapeamento que especifica as correspondências entre as ontologias de aplicação e a ontologia global; e o mapeamento que especifica as correspondências entre as ontologias-fonte e as ontologias de aplicação (mapeamentos locais).

Na arquitetura de dois níveis, a ontologia de domínio é usada apenas para especificar o esquema de mediação. Assim, para permitir a descoberta, recuperação e integração de dados, o usuário tem de definir os mapeamentos. Esses mapeamentos são considerados heterogêneos, pois as ontologias não compartilham o mesmo vocabulário e são estruturalmente heterogêneas.

Figura 26 - Arquitetura de três níveis baseada em ontologias



Fonte: (PINHEIRO, JOÃO CARLOS, 2011).

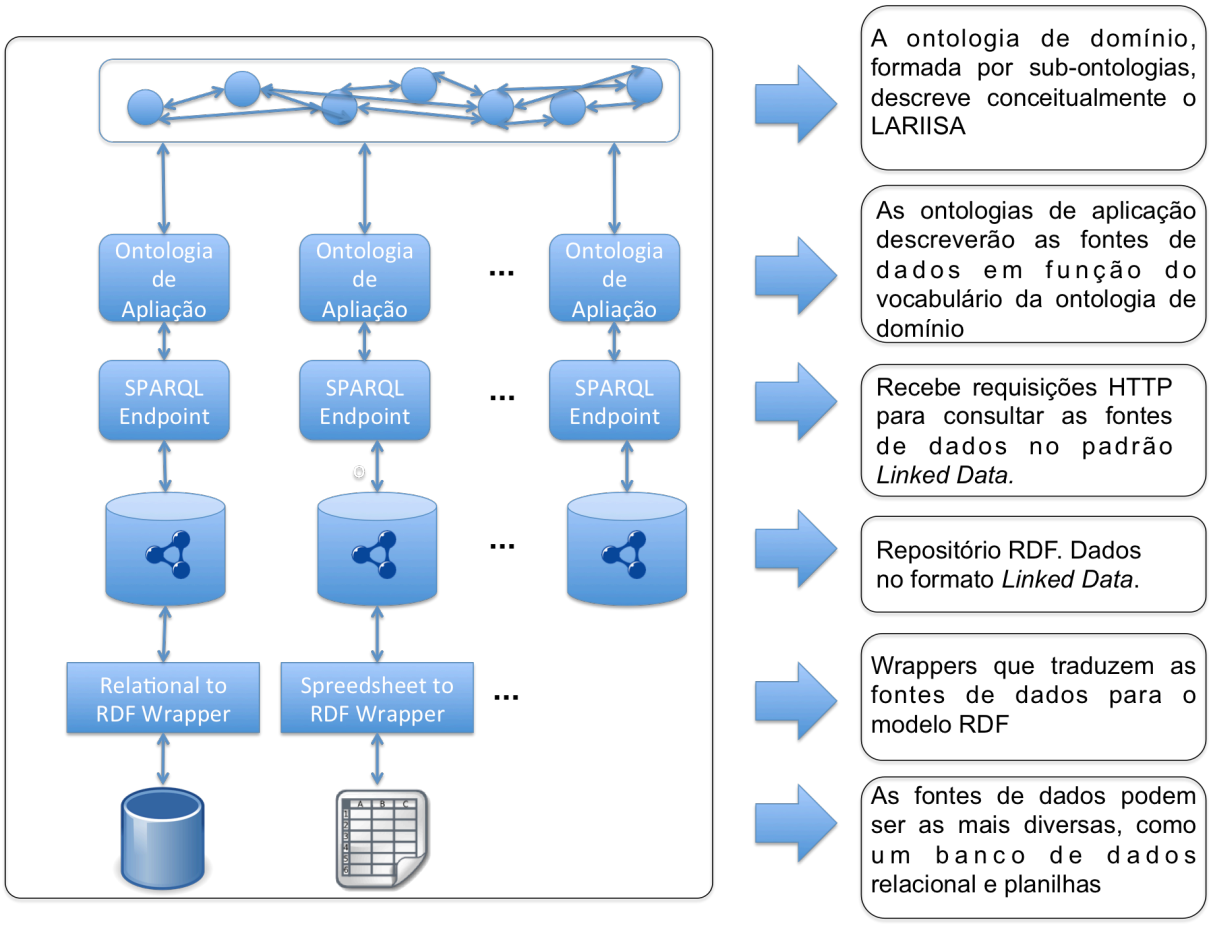
Na arquitetura de três níveis, a ontologia de domínio é usada tanto para especificar o esquema de mediação como um vocabulário compartilhado. Visto que as ontologias de aplicação são fragmentos de uma ontologia de domínio, os mapeamentos entre elas são considerados mapeamentos homogêneos.

A arquitetura de três níveis é vantajosa se comparada a arquitetura de dois níveis pela inserção das ontologias de aplicação, que simplificam a definição dos mapeamentos de mediação, facilitando assim o processo de reformulação de consulta, bom como a integração de dados. Os mapeamentos são relativamente simples e diretos. O processo de integração proposto neste trabalho para o LARIISA adota a arquitetura de três níveis.

A Figura 27 apresenta a aplicação da arquitetura de três níveis ao LARIISA. São destacadas as fontes de dados a serem integradas, que no LARIISA, podem ser de qualquer tipo, como fontes de dados relacionais ou planilhas, inclusive fontes de dados já no padrão *Linked Data*. Em caso de fontes de dados que não estejam no padrão *Linked Data*, é preciso que exista um *wrapper* específico que traduza esses dados para RDF.

O repositório RDF representa as ontologias fonte da arquitetura de três níveis. Ele pode ser referente a dados originalmente já no padrão *Linked Data* ou ser resultado da tradução de um tipo de fonte para RDF. Existem *wrappers* que traduzem um tipo de fonte para RDF *on-the-fly*, ou seja, os dados são convertidos apenas quando solicitados, eliminando a necessidade de replicar os dados em um repositório RDF dedicado.

Figura 27 - Arquitetura de integração de 3 níveis aplicada ao LARIISA



Fonte: Elaboração do autor.

A ontologia de domínio descreve conceitualmente o LARIISA e as consultas direcionadas ao sistema são todas em função do vocabulário definido. Criar uma representação conceitual completa do LARIISA é uma tarefa bastante complexa, isso porque o LARIISA foi pensado para ser uma plataforma totalmente escalável, onde os dados vão sendo integrados de acordo com a necessidade. Sendo assim, como apresentado na Figura 27, considera-se que a ontologia de

domínio do LARIISA é construída considerando uma abordagem de subdomínios, isto é, a ontologia de domínio “LARIISA” é definida por sub-ontologias, e mapeamentos inter-ontologias, que responderão à questões específicas.

As ontologias de aplicação, como já mencionado, descrevem as fontes de dados através de um subconjunto do vocabulário da ontologia de domínio. Dessa forma, os processos de reformulação de consultas se tornam mais simples, pois as ontologias de aplicação são baseadas nas mesmas primitivas, e nenhuma inferência é necessária.

5.2 ESPECIFICAÇÃO DO PROCESSO DE INTEGRAÇÃO DE DADOS

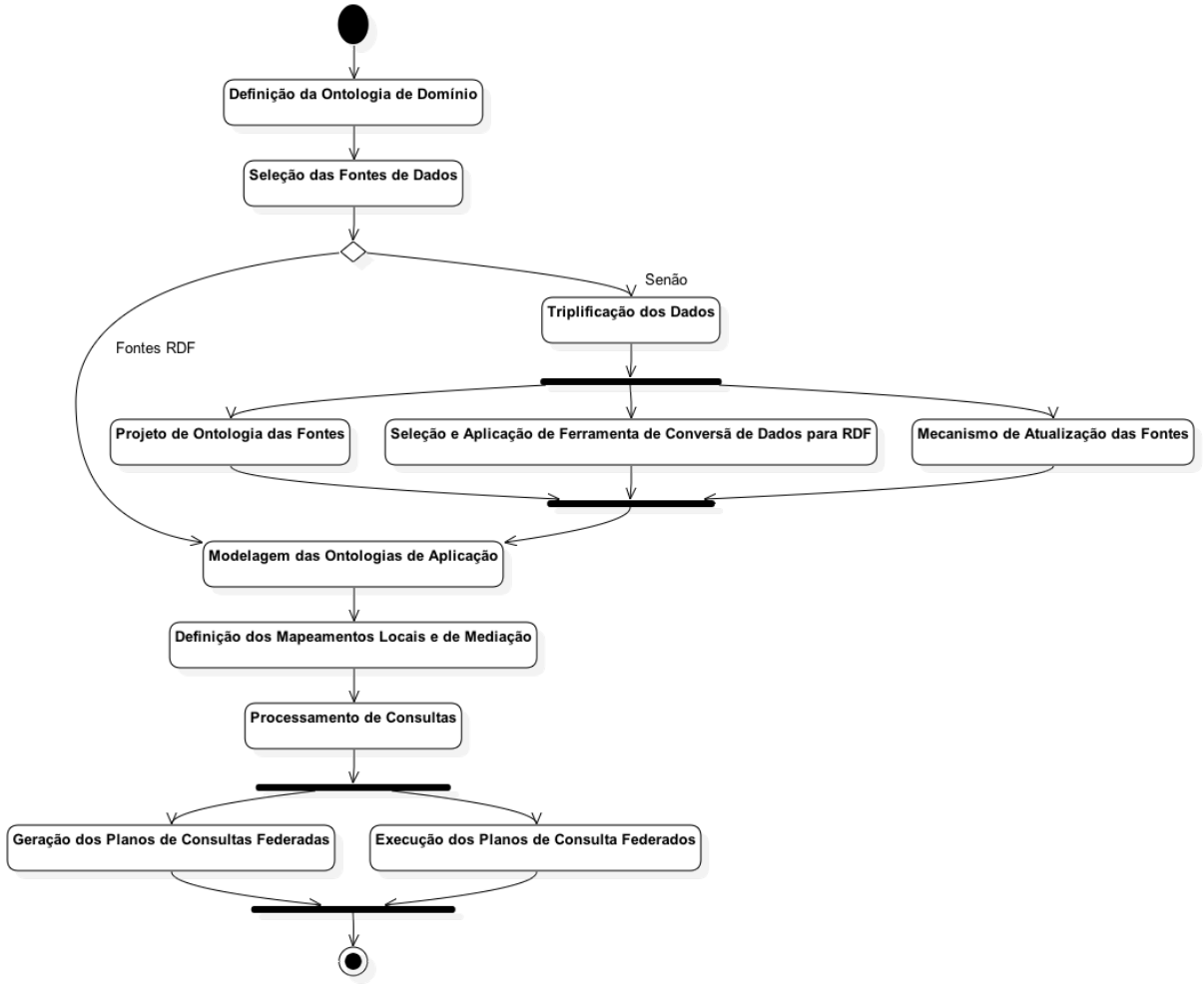
O processo de integração de dados de saúde para o LARIISA, considerando fontes heterogêneas e distribuídas, proposto neste trabalho, é composto por 6 passos a saber: (i) Definição da ontologia de domínio; (ii) Seleção das fontes de dados; (iii) Triplificação dos dados (se os dados não estiverem no padrão RDF); (iv) Modelagem das ontologias de aplicação; (v) Definição dos mapeamentos locais e de mediação; e o (vi) Processamento de consultas. A Figura 28 a seguir ilustra o processo.

O processo proposto pode ser utilizada em dois cenários dentro do LARIISA:

Cenário 1: aplicação do processo para realizar a integração de bases de dados definidas, armazenando as ontologias de domínios e de aplicação criadas na base de conhecimento do LARIISA, ou seja, no componente *Ontology Base* da arquitetura. Sendo assim, o processo será aplicado sempre que existir a necessidade de aumentar a base de conhecimento do LARIISA.

Cenário 2: aplicação do processo para suportar uma demanda específica, uma aplicação do LARIISA, que necessite da integração de uma base de dados definida para responder à esta demanda. Neste caso, a aplicação submeterá consultas ao esquema mediado criado, que estará disponível através de um serviço Web (vide seção 5.3).

Figura 28 - Diagrama de Atividades que define o processo proposto de integração de dados para o LARIISA



Fonte: Elaboração do autor.

5.2.1 Definição Da Ontologia De Domínio

O primeiro procedimento do processo especificado é a construção da ontologia de domínio que responderá às questões demandadas. A ontologia de domínio pode ser criada através de um processo manual, com a ajuda de um editor de ontologias, como o *protégé*, e um especialista do domínio dado, seguindo uma metodologia de construção de ontologias. Noy e Mcguinness (2001) sugerem as seguintes etapas para a construção de ontologias:

1. Determinar o domínio e escopo da ontologia;
2. Considerar o reuso de ontologias existentes:

3. Enumerar termos importantes na ontologia;
4. Definir as classes e a hierarquia de classes;
5. Definir as propriedades de classes-slots;
6. Definir as facetas dos slots;
7. Definir instâncias.

A ontologia de domínio também pode ser criada através de um processo semiautomático ou automático, conhecido por *Ontology Learning*, que extrai ontologias completas de textos de linguagem natural. Buitelaar e Magnini (2005) propõem os seguintes passos incrementais para esse processo:

1. Extração de termos relevantes e seus sinônimos de um corpo textual para um domínio de destino;
2. Identificação de conceitos;
3. Derivação de uma hierarquia dos conceitos anteriormente identificados;
4. Identificação das relações não taxonômicas entre os conceitos;
5. Ajuste da ontologia com novas instâncias, conceitos e propriedades (População da ontologia);
6. Descoberta de novas regras e relações axiomáticas entre conceitos e propriedades.

Como mencionado na seção anterior, a ontologia de domínio pode estar relacionada a uma demanda específica de integração de dados para servir uma aplicação do LARIISA, ou pode estar relacionado a todo o domínio de conhecimento da plataforma.

5.2.2 Seleção Das Fontes De Dados

O próximo passo, após criado a Ontologia de Domínio, é a seleção de fontes relevantes para o problema. As fontes de dados poderão estar disponíveis na Web de forma “aberta”, mas não necessariamente publicadas no padrão *Linked Data*. Sendo assim, temos os seguintes cenários:

- **Cenário 1: fontes de dados no padrão *Linked Data.*** Este é o melhor cenário. As fontes de dados estão disponíveis publicamente

na Web, no padrão *Linked Data.* Cada fonte tem associado a ela um *SPARQL endpoint*.

- **Cenário 2: fontes de dados publicadas na Web sem estarem no padrão *Linked Data.*** Neste caso, os dados estão disponíveis da Web em outros formatos, como CSV, XML, planilhas, e outros. Esses dados, depois de obtidos, deverão passar por um *wrapper* específico que traduza os dados do seu formato original para RDF.
- **Cenário 3: fontes de dados não publicadas na Web, mas disponíveis a partir de um *Web Service.*** Neste caso, os dados são obtidos a partir de um Web Service, mas como no cenário 2 podem estar em diferentes formatos. Deverão passar por um *wrapper* específico que traduza os dados do seu formato original para RDF.
- **Cenário 4: fontes de dados não publicadas nem disponíveis através de um *Web Service.*** Se os dados não estiverem disponíveis publicamente na *Web* ou a partir de um *Web Service*, provavelmente será necessário a obtenção dos dados diretamente com o mantenedor, o que pode significar a obtenção de um esquema de dados relacional. Como no cenário 2 e 3, os dados devem ser submetidos a um wrapper específico que traduza os dados relacionais para RDF.

5.2.3 Triplificação Dos Dados

Exceto em fontes de dados do Cenário 1, discutido na seção anterior, onde os dados já estão no formato RDF, publicados, todos os outros cenários exigem que os dados sejam triplificados. Consequentemente, a triplificação dos dados exigirá a manutenção de componentes que integram a infraestrutura de *Web* Semântica: o *RDF Store (*ou *triple store)*, que como já citado em seções anteriores, é um banco de dados adequado para armazenar e obter dados de triplas; e o RDF *Query Engine*, o mecanismo de busca que fornece a capacidade de recuperar informações de um RDF *Store*, utilizando uma linguagem de consulta, o SPARQL.

No trabalho de Cifuentes-Silva, Sifaqui e Labra-Gayo (2011) encontra-se um processo mais detalhado de como criar e disponibilizar *Linked Data*. O trabalho considera os seguintes passos: (i) contextualização, (ii) projeto da ontologia, (iii)

modelagem do grafo RDF, (iv) implementação dos *endpoint* SPARQL, (v) implementação do grafo RDF, (vi) atualização do serviço gráfico e (vii) opcional visualização dos grafos.

Diversas ferramentas e métodos também são citadas pelo W3C[9] capazes de realizar a triplificação de dados para RDF. Para este trabalho é considerada a seguinte sequência de passos para a conversão dos dados:

(i) Projeto de Ontologia das Fontes. As fontes de dados precisam ser descritas através de uma ontologia. São as "Ontologias Fonte" descritas na arquitetura de três níveis (Figura 26). Um ponto importante é a criação de uma URI para a publicação da ontologia.

(ii) Seleção e aplicação de ferramenta de conversão de dados para RDF.

(iii) Mecanismo de atualização das fontes. É preciso estabelecer um processo de atualização onde os grafos RDF publicados possam refletir as atualizações das fontes de dados. Porém, o estabelecimento desse processo irá depender do tipo de ferramenta utilizada para a conversão dos dados em RDF. Se a ferramenta escolhida for baseada em uma abordagem virtualizada, a própria ferramenta estará responsável pela atualização dos grafos. Porém, se a ferramenta utilizar uma abordagem materializada, a rotina de atualização deve ser bem definida.

A depender da estratégia de negócio entre o LARIISA e os mantenedores das fontes de dados, o processo de triplificação pode ser responsabilizado tanto ao LARIISA, quanto aos mantenedores. A observar a tendência em disponibilizar dados abertos, em RDF, é de se esperar que logo não seja necessário o LARIISA manter uma infraestrutura para os dados triplificados das fontes. Consequentemente, este passo se resumirá ao acesso do projeto das ontologias-fonte pelo processo de integração proposto.

5.2.4 Modelagem Das Ontologias De Aplicação

Após a triplificação das fontes, é hora de gerar as ontologias de aplicação, que são obtidas após a realização das correspondências entre cada ontologia fonte e a ontologia de domínio. O vocabulário de uma ontologia de aplicação é composto

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

de classes e propriedades que são o subconjunto da ontologia de domínio combinada à ontologia fonte.

O trabalho de Sacramento et al. (2010) descreve uma estratégia para geração automática de ontologias de aplicação, considerando um conjunto de ontologias fonte, a ontologia de domínio e os resultados das correspondências entre cada ontologia fonte e a ontologia de domínio, que pode ser utilizada como abordagem para esta etapa do processo.

5.2.5 Definição Dos Mapeamentos Locais E De Mediação

Após criada as ontologias de aplicação, passa-se para a etapa de definição dos mapeamentos locais e de mediação. Os mapeamentos consistem de uma etapa importante para o esquema de mediação, como citado na seção introdutória deste capítulo, fazendo o papel de “cola” entre a ontologia de domínio, ontologias de aplicação e ontologias fonte.

Nesta etapa também podem ser usadas diversas metodologias e/ou ferramentas para a definição desses mapeamentos, sejam elas manuais ou automáticas[10].

5.2.6 Processamento De Consultas

Depois de definido o esquema conceitual do método de integração, com a criação das ontologias de domínio, de aplicação e os mapeamentos locais e de mediação, deve-se definir como se dará o processo de consulta sob esse esquema mediado.

Para isso, utiliza-se neste trabalho o método para o processamento de consultas proposto por Pinheiro (2011), baseado em dois módulos principais: um módulo responsável por gerar planos de execução de consultas federadas (consultas que serão realizadas em múltiplas fontes de dados) e um módulo responsável pela execução desses planos.

[10] http://wiki.opensemanticframework.org/index.php/Ontology_Tools#Ontology_Mapping

5.2.6.1 Geração Dos Planos De Consultas Federadas

A estratégia de geração dos planos de execução de consultas federadas consiste resumidamente de três passos:

1. Tradução. Uma consulta SPARQL Q é submetida ao esquema mediador, expressada em termos de ontologia de domínio, e é transformada numa árvore que representa a estrutura da consulta.

2. Reformulação. A consulta Q é reformulada, baseada nos mapeamentos de mediação, em um conjunto de sub-consultas sob as ontologias de aplicação. Cada sub-consulta é então reformulada, baseado nos mapeamentos locais, em termos de consulta sobre as ontologias fonte.

3. Otimização. Este passo objetiva otimizar ainda mais as sub-consultas que serão feitas sobre os dados RDF, de forma a minimizar o tempo de resposta e a quantidade de dados transferidos dos SPARQL *endpoints* para o mediador.

5.2.6.2 Execução Dos Planos De Execução De Consultas Federadas

Após a geração dos planos de execução de consultas eles são então executados. Um dos principais desafios de se realizar consultas federadas é torna-las eficientes. Para isso, buscam-se algoritmos que possam reduzir o volume de dados que são repassados às fontes e que realizem o processamento paralelo de consultas. Tem-se como trabalhos futuros, a definição dos melhores algoritmos de execução a serem utilizados no processo proposto, considerando as peculiaridades do LARIISA. Para este trabalho, utiliza-se a *engine* de execução do *framework Apahe Jena*[11], que suporta o processamento de consultas federadas.

Tanto Pinheiro (2011), como Magalhães (2012), apresentam algoritmos de execução que se mostram de maior desempenho se comparado a *engine* do Jena. Veremos na seção 6.2 (*Trabalhos Futuros*) que o algoritmo de execução é um campo vasto a ser explorado.

[11] https://**jena**.apache.org/

5.3 ARQUITETURA PARA APLICAÇÕES WEB LARIISA

Nos casos em que a base de conhecimento gerada pelo processo proposto servir uma demanda específica, uma aplicação Web LARIISA, como descrito no Cenário 2, seção 5.2, é apresentada uma arquitetura que define a relação entre a aplicação Web, o esquema mediado de dados e as fontes de dados RDF (Figura 29). Considera-se que o esquema mediador ficará disponível para a aplicação na forma de um serviço Web.

Diferente do proposto por Pinheiro (2011), os planos de execução de consultas não serão gerados dinamicamente, mas serão definidos a priori, em tempo de projeto. Em resumo, o processo descrito na seção 5.2 será aplicado até a etapa de geração dos planos de consulta federadas (descrita na seção 5.2.6.1). As fontes de dados integradas para o LARIISA terão, em sua maioria, como objetivo resolver necessidades específicas, justificando a não necessidade de abrir o esquema mediado para qualquer consulta posta pelo usuário.

A arquitetura apresentada na Figura 29 foi apresentada por Magalhaes, (2012) e foi adaptada considerando as peculiaridades do LARIISA. A grande vantagem dessa arquitetura é que não é necessário o uso de um mediador, em tempo de execução, para realizar a geração do plano de consulta federada.

A **Aplicação** tem como função enviar parâmetros necessários para o **Processador de Consulta,** através de uma URI, bem como exibir os resultados quando recebido. A Aplicação envia os seguintes parâmetros ao Processador de Consulta: (i) p_1, que identifica a consulta que será utilizada e (ii) [p_2, p_3...p_i], lista de parâmetros de entrada que serão necessários para a execução dessa consulta.

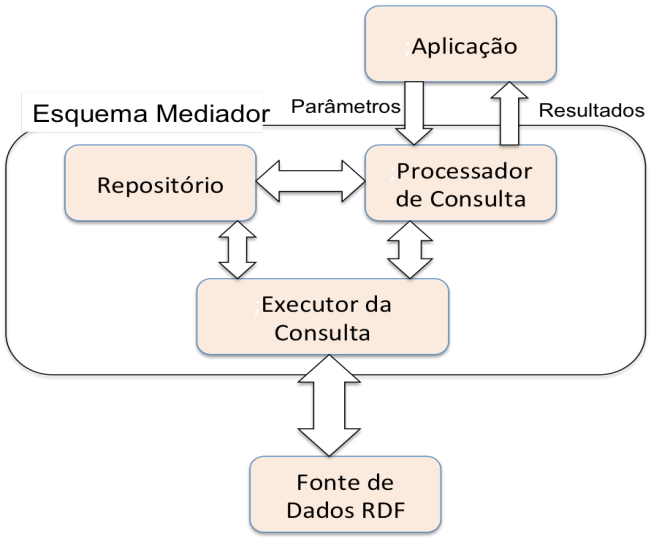
O **Processador de Consulta** verifica a conformidade dos parâmetros recebidos pela aplicação. Verifica, por exemplo, se todos os parâmetros necessários à execução da consulta indicada se fazem presentes. Se algum parâmetro não condizer com o esperado, ou na falta de parâmetros, uma mensagem para a aplicação é retornada.

De acordo com a identificação da consulta recebida, o Processador de Consulta verifica no **Repositório** qual será o plano de consulta federada a ser executada pelo Executor de Consulta e repassa para ele.

O Processador de Consulta também é responsável por converter para o formato de saída esperado pela aplicação os resultados da consulta que vão sendo

obtidos. No trabalho de Magalhães (2012), o formato de saída suportados são o XML e o JSON. Para este trabalho, temos como trabalho futuro definir quais formatos de saídas são os mais adequados para o LARIISA.

Figura 29 - Arquitetura que define a relação entre a aplicação, o esquema mediado de dados e as fontes de dados RDF



Fonte: Adaptado de Magalhães (2012).

O **Executor da Consulta** é o componente mais importante do esquema. Ele é responsável pela execução do plano de consulta identificado pelo *Processador de Consulta* sob as fontes de dados. O Executor de Consulta corresponde a última etapa do processo, como definido na seção anterior. A escolha do algoritmo do Executor de Consulta, ou a necessidade de desenvolvimento de um novo, define a *performance* com que essas consultas serão executadas. Este algoritmo deverá suportar consultas federadas. Como visto na seção 5.2.6.2, os planos de consulta federados podem ser executados pela própria *engine* do *framework Jena*.

O **Repositório** armazena todos os planos de execução de consultas federadas, que foram geradas após a aplicação do processo, e são armazenados em XML; armazena os mapeamentos locais e de mediação, e; armazena meta-dados sobre as fontes de dados, uma forma de registrar as fontes de dados que compõem o esquema mediado. Os metadados são armazenados sob o vocabulário VoiD (*Vocabulary of Interlinked Datasets*), vocabulário padronizado pelo W3C para publicação de metadados sobre conjunto de dados disponíveis como Linked Data.

Sendo assim, para a implementação do serviço como descrito na arquitetura, é necessário primeiramente gerar os planos de execução de consultas federadas (seção 5.2.6.1) e convertê-los em XML, para serem, então, armazenados no Repositório.

5.4 APLICAÇÃO DO PROCESSO PROPOSTO

Considere o seguinte cenário: a gestão municipal de saúde mantém um sistema de monitoramento de agravos de dengue. A partir de informações obtidas dos estabelecimentos de saúde, consegue-se realizar um monitoramento semanal, desde os casos notificados de dengue até os casos mais graves.

Através da vigilância epidemiológica da doença é possível se fazer uma avaliação de sua intensidade, colaborando para orientar o gestor na tomada de ações, assim como, avaliar as medidas que vem sendo tomadas, permitindo, por exemplo, a otimização do uso dos recursos disponíveis para controle da doença.

Apenas a obtenção do número de casos de dengue não é suficiente para uma gestão eficiente de controle da doença. É preciso conhecer os fatores que influenciam na incidência de casos, possibilitando, assim, que a gestão possa trabalhar mais efetivamente com a prevenção e combate da doença. Por exemplo, através da vigilância epidemiológica constata-se que um determinado bairro aumentou em 20% os casos notificados de dengue. Mas porque desse aumento? Houve incidências de chuva no bairro? Problemas de saneamento? Houve denúncias de focos de dengue pela população do bairro? Ou seja, perguntas que representam a necessidade de investigação do porque do aumento de número de casos no bairro. Para responder a essas perguntas o gestor precisa ter os dados de contexto da região integrados aos dados de vigilância epidemiológica para responder a essas perguntas.

Considerando o cenário descrito, surge a demanda de desenvolvimento de uma aplicação que considera integrar dados para responder à questão específica. Através do processo de integração proposto neste trabalho, a plataforma LARIISA se torna capaz de responder à demanda, além de contribuir para a sua própria base de conhecimento. Tem-se, então, a aplicação do processo na integração de dados, como posto no **Cenário 2,** tornando o esquema mediado

acessível através de um serviço para a aplicação (utilizando a arquitetura definida na seção anterior).

5.4.1 Definição Da Ontologia De Domínio

Para construir a ontologia de domínio que responderá às questões impostas pelo gestor, é necessário que um especialista indique quais os conceitos que são abordados na problemática, e principalmente indicar quais os fatores ambientais e socioeconômicos influenciam para o aumento de casos de dengue em uma região.

- Para ilustrar essa etapa, a ontologia de domínio do cenário foi criada através de um processo manual, utilizando o editor de ontologias *protégé*, como mostra a

Figura 30. É uma ontologia que relaciona informações contextuais do bairro (IDH, precipitação, densidade demográfica, denúncias de focos de dengue) que podem estar relacionados com o aumento de casos de dengue registrados.

A relação dos conceitos de IDH, precipitação, densidade demográfica com o número de casos de dengue foi embasada nos trabalhos de Flauzino, Souza-Santos e Oliveira (2011) e Mondini e Chiaravalloti (2007). Como já mencionado na seção 5.1, a ontologia de domínio construída servirá tanto como esquema mediador, quanto como vocabulário compartilhado.

5.4.2 Seleção Das Fontes De Dados

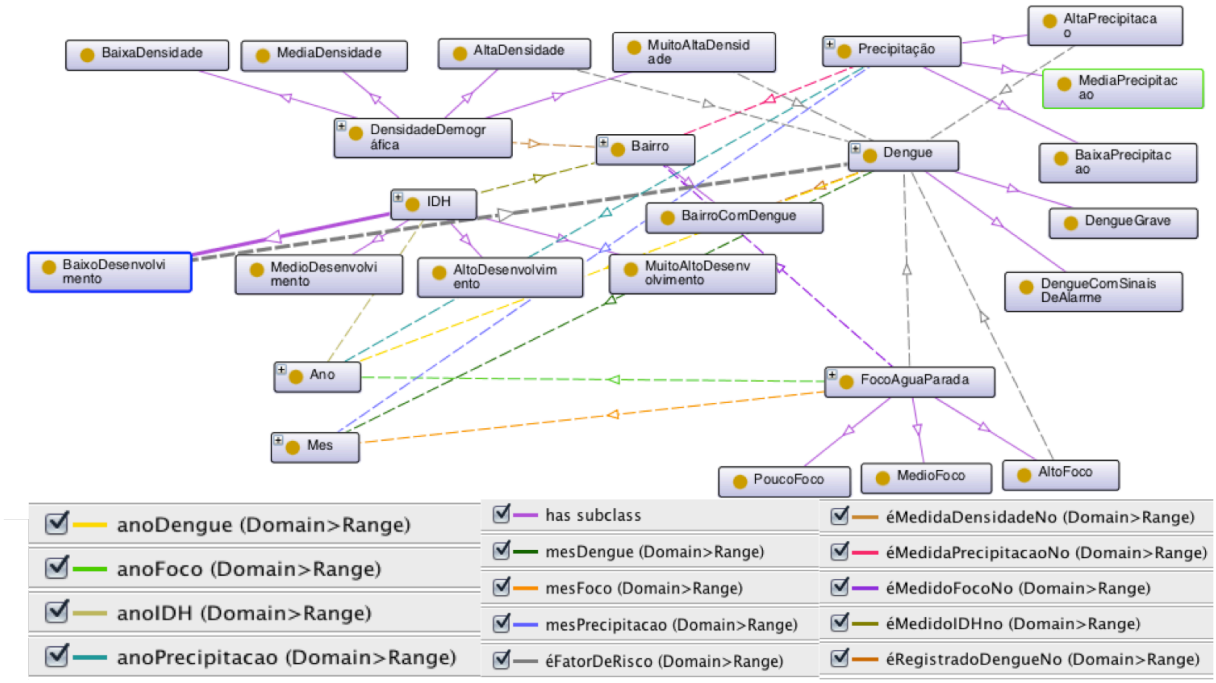
Para o cenário descrito, foram selecionados as seguintes fontes de dados:

- **Dados do Sistema de Monitoramento Diário de Agravos (SIMDA[12]) da Secretaria Municipal de Saúde de Fortaleza**. O sistema disponibiliza semanalmente os casos de dengue notificados e confirmados no nível de granularidade por estabelecimento. Os dados podem ser visualizados na própria *Web,* como é possível exportar os

[12] http://tc1.sms.fortaleza.ce.gov.br/simda/dengue/mes

dados em formato de planilha ou documento de texto. Através dessa fonte de dados pode-se obter o registro de dengue de cada bairro.

Figura 30 - Ontologia de Domínio resultante do Passo 1 do Processo de Integração Proposto



Fonte: Elaboração do autor.

- **Dados das Secretarias Regionais de Fortaleza[13].** Cada secretaria mantém informações sobre os bairros que administra como IDH, população e área (o que nos permite obter a densidade demográfica). Esses dados podem ser obtidos diretamente da *Web*, de forma manual, ou diretamente com as regionais em formado de planilha.
- **Dados da Fundação Cearense de Meteorologia e Recursos Hídricos (Funceme)[14].** Através do site da Funceme é possível ter acesso aos dados pluviométricos diários por município. O problema dessa fonte de dados é que os registros pluviométricos não são feitos por bairro, mas por 7 estações, localizadas nos bairros do Pici, Mondubim, Castelão, Água Fria, Messejana, Aldeota e Edson Queiroz. Os dados podem ser exportados em formato de planilha.

[13] http://www.fortaleza.ce.gov.br/regionais
[14] http://www.funceme.br/index.php/areas/tempo/chuvas-diarias-municipios

- **Dados sobre focos de dengue.** Em Fortaleza, o cidadão que deseja fazer esse tipo de denúncia, deve entrar em contato diretamente com a regional do bairro, que agenda uma visita para verificação da área denunciada. No Rio Grande do Norte foi desenvolvido pela UFRN o Observatório da Dengue[15], um aplicativo *Web* onde moradores podem denunciar focos de mosquitos. Para os gestores públicos, o conjunto de denúncias feitas pelos cidadãos forma um mapa que possibilita visualizar onde há suspeitas da doença e onde existem mosquitos transmissores.

5.4.3 Triplificação Dos Dados

No caso do cenário descrito, todas as fontes selecionadas estão no formato de planilhas, o que significa ser necessário converter as planilhas em grafos RDF. Utiliza-se, então, a ferramenta XLWrap (LANGEGGER; WOS, 2009), por possuir todos os componentes para a publicação dos dados, e pela sua simplicidade em realizar os mapeamentos de conversão.

Para este trabalho é considerada a seguinte sequência de passos para a conversão dos dados:

(i) Projeto de Ontologia das Fontes. A

A

[15] http://www.telessaude.ufrn.br/observatoriodadengue/

Figura 32 apresenta os dados obtidos do SIMDA, em formato de planilha; o Quadro 1 apresenta o mapeamento feito, de acordo com as orientações da ferramenta; e a Figura 33 mostra os dados já convertidos, sendo obtidos através do *endpoint SPARQL* provido pela ferramenta.

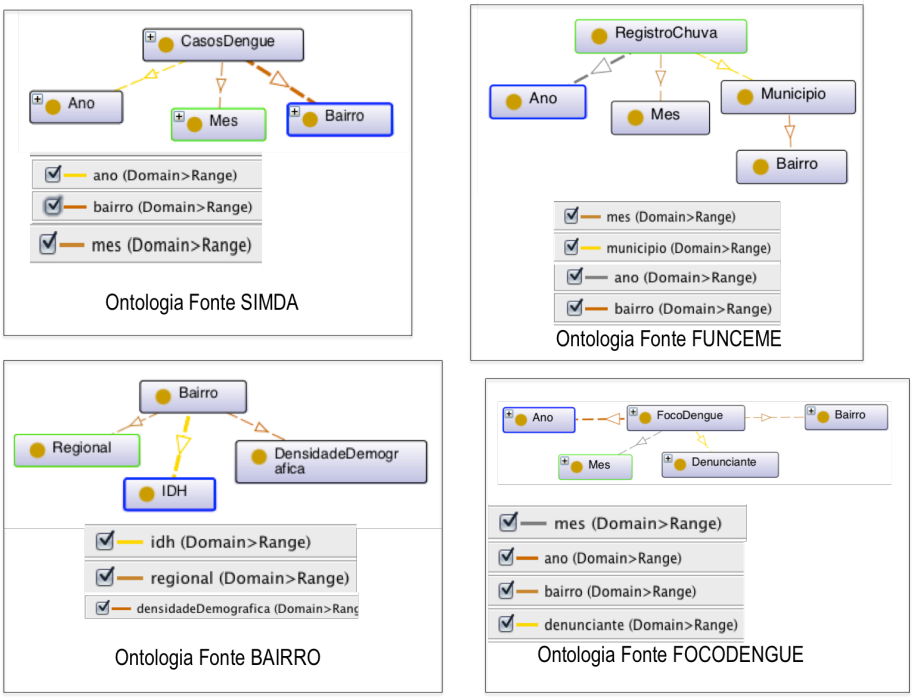
Figura 31 representa as ontologias fonte criadas. Na impossibilidade de obter uma URI válida para o exemplo, é considerado que a ontologia das fontes, assim como suas instâncias, foram publicadas sob o prefixo <http://example.org>. Outro detalhe importante é que as ontologias criadas podem utilizar vocabulário parecido com o vocabulário da ontologia de domínio, visto que, no caso específico do exemplo, não é utilizada nenhuma ontologia já publicada.

(ii) Seleção e aplicação de ferramenta de conversão de dados para RDF. Para o cenário, como já citado, foi escolhida a ferramenta *XLWrap*, que converte dados armazenados em planilhas para grafos RDF. Através de um mapeamento definido entre a planilha e o vocabulário do passo anterior, os grafos são criados pelo *wrapper* tornando-os disponíveis em um *endpoint SPARQL*.

A

Figura 32 apresenta os dados obtidos do SIMDA, em formato de planilha; o Quadro 1 apresenta o mapeamento feito, de acordo com as orientações da ferramenta; e a Figura 33 mostra os dados já convertidos, sendo obtidos através do *endpoint SPARQL* provido pela ferramenta.

Figura 31 - Exemplos de ontologias fonte



Fonte: Elaboração do autor.

(iii) Mecanismo de atualização das fontes. No exemplo utilizado, como grande parte das fontes de dados estão no formato de planilhas, e que no domínio as demais informações estarão em função dos dados de casos de dengue, que são atualizados semanalmente, as fontes publicadas no formato RDF deverão passar por uma rotina de atualização semanal.

Figura 32 - Dados obtidos do SIMDA, referente ao número de casos de dengue registrados por ano, por mês, por bairro, em formato de planilha

BAIRRO	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ	TOTAL
MONDUBIM	19	51	93	103	126	182	114	93	33	17	9	5	845
BOM JARDIM	25	25	42	88	105	134	115	91	28	21	9	6	689
MESSEJANA	14	26	92	135	119	99	87	52	20	16	5	8	673
CANINDEZINHO	8	45	74	115	150	106	84	47	12	7	9	3	660
PREFEITO JOSE WALTER	17	28	44	41	46	75	88	72	15	13	5	8	452
JANGURUSSU	10	15	64	73	66	74	57	42	16	10	8	4	439
BARROSO	8	19	46	81	63	65	63	38	19	14	2	6	424
PLANALTO AIRTON SENNA	9	11	35	41	52	93	84	62	11	12	11	3	424
IGNORADO	14	25	35	52	35	40	44	54	18	11	7	4	339
PASSARE	12	23	29	32	47	60	53	34	23	7	8	0	328
SERRINHA	6	17	31	50	40	39	29	40	17	8	7	6	290
JARDIM DAS OLIVEIRAS	24	16	14	42	33	32	63	37	12	6	5	4	288
LAGOA REDONDA	6	17	38	71	48	49	27	12	2	3	3	1	277
SIQUEIRA	3	17	30	33	39	69	37	16	7	5	5	4	265
VILA MANOEL SATIRO	4	14	20	39	24	58	42	39	10	8	4	3	265
RODOLFO TEOFILO	5	5	11	36	44	58	41	33	19	4	2	4	262
ANTONIO BEZERRA	13	7	11	28	39	39	32	32	16	20	14	6	257
BOM SUCESSO	8	19	21	31	45	28	30	44	10	3	5	9	253
PARQUE SANTA ROSA	8	15	12	31	31	49	64	32	6	3	2	0	253
PICI	9	8	8	24	26	38	58	46	20	4	4	2	247
PALMEIRAS	5	17	25	41	32	28	31	43	12	5	1	6	246
JOAO XXIII	8	10	16	34	56	39	32	18	12	5	2	7	239
MARAPONGA	7	11	20	24	26	47	38	31	19	6	5	4	238

2014 2013 +

Fonte: Elaboração do autor.

Figura 33 - Resultado da triplificação dos dados considerando a fonte SIMDA

```
SELECT DISTINCT * WHERE {
  ?s ?p ?o
}
LIMIT 20
```

Results: Browse Go! Reset

SPARQL results:

s	p	o
<http://example.org/casosDengue_2014_JAN_BOM+JARDIM>	<http://example.org/casosDengue>	11
<http://example.org/casosDengue_2014_JAN_BOM+JARDIM>	rdf:type	<http://example.org/CasosDengue>
<http://example.org/casosDengue_2014_JAN_BOM+JARDIM>	<http://example.org/bairro>	<http://pt.dbpedia.org/page/BOM JARDIM>
<http://example.org/casosDengue_2014_JAN_BOM+JARDIM>	<http://example.org/year>	<http://dbpedia.org/resource/2014>
<http://example.org/casosDengue_2014_JAN_BOM+JARDIM>	<http://example.org/mes>	"JAN"
<http://example.org/casosDengue_2014_JAN_MONDUBIM>	<http://example.org/casosDengue>	8
<http://example.org/casosDengue_2014_JAN_MONDUBIM>	rdf:type	<http://example.org/CasosDengue>
<http://example.org/casosDengue_2014_JAN_MONDUBIM>	<http://example.org/bairro>	<http://pt.dbpedia.org/page/MONDUBIM>
<http://example.org/casosDengue_2014_JAN_MONDUBIM>	<http://example.org/year>	<http://dbpedia.org/resource/2014>
<http://example.org/casosDengue_2014_JAN_MONDUBIM>	<http://example.org/mes>	"JAN"
<http://example.org/casosDengue_2014_JAN_MESSEJANA>	<http://example.org/casosDengue>	21
<http://example.org/casosDengue_2014_JAN_MESSEJANA>	rdf:type	<http://example.org/CasosDengue>
<http://example.org/casosDengue_2014_JAN_MESSEJANA>	<http://example.org/bairro>	<http://pt.dbpedia.org/page/MESSEJANA>
<http://example.org/casosDengue_2014_JAN_MESSEJANA>	<http://example.org/year>	<http://dbpedia.org/resource/2014>
<http://example.org/casosDengue_2014_JAN_MESSEJANA>	<http://example.org/mes>	"JAN"
<http://example.org/casosDengue_2014_JAN_CANINDEZINHO>	<http://example.org/casosDengue>	7
<http://example.org/casosDengue_2014_JAN_CANINDEZINHO>	rdf:type	<http://example.org/CasosDengue>
<http://example.org/casosDengue_2014_JAN_CANINDEZINHO>	<http://example.org/bairro>	<http://pt.dbpedia.org/page/CANINDEZINHO>
<http://example.org/casosDengue_2014_JAN_CANINDEZINHO>	<http://example.org/year>	<http://dbpedia.org/resource/2014>
<http://example.org/casosDengue_2014_JAN_CANINDEZINHO>	<http://example.org/mes>	"JAN"

Fonte: Elaboração do autor.

Quadro 1 - Mapeamento realizado para a fonte de dados SIMDA considerando as orientações da ferramenta de conversão XLWrap da planilha da Figura 32

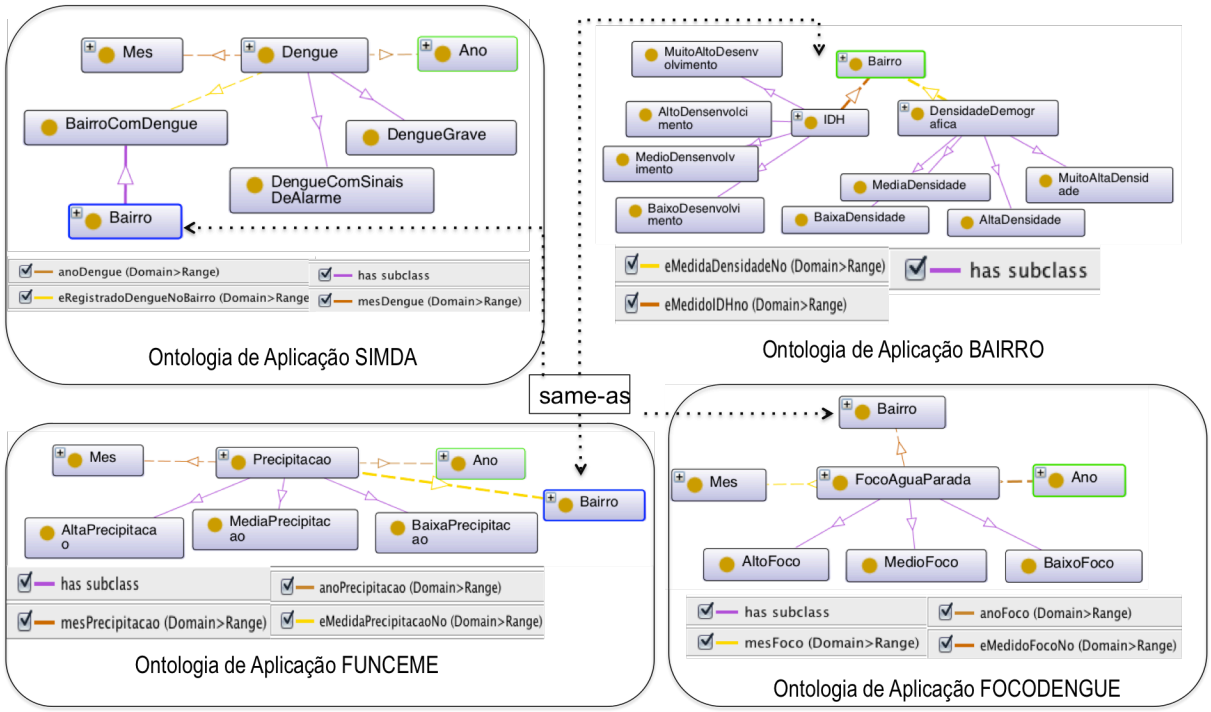
```
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .
@prefix ex:                <http://example.org/> .
@prefix dc:   <http://purl.org/dc/elements/1.1/> .
@prefix xl:                <http://purl.org/NET/xlwrap#> .
@prefix scv:   <http://purl.org/NET/scovo#> .
@prefix :     <http://myApplication/configuration#> .

# mapping
{ [] a xl:Mapping ;
    xl:offline "false"^^xsd:boolean ;
    xl:template [
            xl:fileName "mappings/files/casos-de-dengue.xls" ;
            xl:sheetNumber "0" ;
            xl:templateGraph :CasosDengue ;
            xl:transform [
                    a rdf:Seq ;
                    rdf:_1 [
                            a xl:RowShift ;
                            xl:restriction "A2; B2:M2" ;
                            xl:breakCondition "A2 == 'TOTAL'" ;
                            xl:steps "1" ;
                    ];
                    rdf:_2 [
                            a xl:ColShift ;
                            xl:restriction "B1; B2:B120"^^xl:Expr ;
                            xl:breakCondition "B1 == 'TOTAL'" ;
                            xl:steps "1" ;
                    ];
                    rdf:_3 [
                            a xl:SheetShift ;
                            xl:restriction "#1.*"^^xl:Expr ;
                            xl:repeat "2"
                    ];
            ]
    ] .
}
:CasosDengue {
    [ xl:uri "'http://example.org/casosDengue_' & URLENCODE(SHEETNAME(A1) & '_' & B1 & '_' &
A2)"^^xl:Expr ] a ex:CasosDengue ;
    ex:bairro    "URI('http://pt.dbpedia.org/page/' & A2)"^^xl:Expr ;
    ex:year     "DBP_YEAR(SHEETNAME(A1))"^^xl:Expr ;
    ex:mes   "B1"^^xl:Expr ;
    ex:casosDengue   "B2"^^xl:Expr .
}
```

Fonte: Elaboração do autor.

5.4.4 Modelagem Das Ontologias De Aplicação

É importante observar que no caso do cenário em questão, por se tratar de um domínio não tão complexo, as ontologias de aplicação foram criadas manualmente. No trabalho de Sacramento et al., (2010) é apresentada uma estratégia para geração automática das ontologias de aplicação, como já citado. A Figura 34 mostra as ontologias de aplicação obtidas para o nosso exemplo.

Figura 34 - Ontologias de Aplicação resultantes do Passo 4 do processo de integração



Fonte: Elaboração do autor.

5.4.5 Definição Dos Mapeamentos Locais E De Mediação

O Quadro 2 apresenta o modelo dos mapeamentos locais (define classes e propriedades da ontologia de aplicação em termos de suas ontologias fonte) e o Quadro 3 o modelo dos mapeamentos de mediação (define classes e propriedades da ontologia de domínio em termos das ontologias de aplicação).

Para distinguir os espaços de nomes usam-se os seguintes prefixos “simda:”, “funceme:”, “bairro:” e “focos:” para se referir aos vocabulários das

ontologias **fonte;** SIMDA, FUNCEME, BAIRRO e FOCODENGUE, respectivamente; “sim_oa:”, “fcm_oa:”, “b_oa:” e “fd_oa” para se referir aos vocabulários das ontologias de **aplicação** SIMDA, FUNCEME, BAIRRO e FOCODENGUE, respectivamente; e “lariisa:” para se referir ao vocabulário da ontologia de **domínio**.

Quadro 2 - Modelo de Mapeamentos Locais

OF SIMDA – OA SIMDA
Classes:
sim_oa:Dengue(d) ⇐ simda:CasosDengue(d)
sim_oa:Bairro(b) ⇐ simda:Bairro(b)
sim_oa:Ano(a) ⇐ simda:Ano(a)
sim_oa:Mes(m) ⇐ simda:Mes(m)
Propriedades:
sim_oa:mesDengue(d, m) ⇐ simda:mes(d,m)
sim_oa:anoDengue(d,a) ⇐ simda:ano(d,a)
sim_oa:eRegistradoDengueNoBairro(d,b) ⇐ simda:bairro(d,b)
OF FUNCEME – OA FUNCEME
Classes:
fcm_oa:Precipitacao(p) ⇐ funceme:RegistroChuva(p)
fcm_oa:Bairro(b) ⇐ funceme:Bairro(b)
fcm_oa:Ano(a) ⇐ funceme:Ano(a)
fcm_oa:Mes(m) ⇐ funceme:Mes(m)
Propriedades:
fcm_oa:mesDengue(p,m) ⇐funceme:mes(p,m)
fcm_oa:anoDengue(p,a) ⇐funceme:ano(p,a)
fcm_oa:eMedidaPrecipitacaoNo(p,b) ⇐funceme:bairro(p,b)
OF BAIRRO – OA BAIRRO
Classes:
b_oa:DensidadeDemografica(dd) ⇐bairro: DensidadeDemografica(dd)
b_oa:IDH(i) ⇐bairro:IDH(i)
b_oa:Bairro(b) ⇐bairro:Bairro(b)
Propriedades

b_oa:eMedidoIDHno(b,i) ⇐bairro:idh(b,i)
b_oa:eMedidaDensidadeNo(b,dd) ⇐bairro:densidadeDemografica(b,dd)
OF FOCODENGUE – OA FOCODENGUE
Classes
fd_oa:FocoAguaParada(fd) ⇐focos:FocoDengue(fd)
fd_oa:Bairro(b) ⇐focos:Bairro(b)
fd_oa:Mes(m) ⇐focos:Mes(m)
fd_oa:Ano(a) ⇐focos:Ano(a)
Propriedades
fd_oa:mesFoco(fd,m) ⇐ focos:mes(fd,m)
fd_oa:anoFoco(fd,a) ⇐ focos:ano(fd,a)
fd_oa:eMedidoFocoNo(fd,b) ⇐ focos:bairro(fd,b)

Fonte: Elaboração do autor.

Quadro 3 - Modelo de mapeamentos de mediação

Ontologia de Domínio – Ontologias de Aplicação
Classes:
lariisa:Bairro(b) ⇐ sim_oa:Bairro(b);fcm_oa:Bairro(b);b_oa:Bairro(b);fd_oa:Bairro(b)
lariisa:Ano(a) ⇐ sim_oa:Ano(b);fcm_oa:Ano(a);b_oa:Ano(a);fd_oa:Ano(a)
lariisa:Mes(m) ⇐ sim_oa:Mes(m);fcm_oa:Mes(m);b_oa:Mes(m);fd_oa:Mes(m)
...
Propriedades:
lariisa:anoDengue(d,a) ⇐ sim_oa:anoDengue(d,a)
lariisa:eMedidaPrecipitacaoNo(p,b) ⇐ fcm_oa:eMedidaPrecipitacaoNo(p,b)
lariisa:eRegistradoDengueNo(d,b) ⇐ sim_oa:eRegistradoDengueNoBairro(d,b)
...

Fonte: Elaboração do autor.

O *framework R2R* (BIZER; SCHULTZ, 2010) consiste de uma linguagem de mapeamento e uma *API Java* para descobrir e processar mapeamentos, e foi utilizado neste exemplo. Mas como descrito na seção 5.2.5, qualquer metodologia ou ferramenta pode ser utilizada. A Quadro 4 apresenta o mapeamento entre a

ontologia fonte SIMDA (sim_oa) e sua respectiva ontologia de aplicação, realizado com o *framework R2R.*

Quadro 4 - Mapeamentos entre a ontologia fonte SIMDA e sua respectiva ontologia de aplicação utilizando o *framework R2R*

```
@prefix r2r: <http://www4.wiwiss.fu-berlin.de/bizer/r2r/> .
@prefix p: < http://example.org/lariisa/mapaementos > .
@prefix sim_oa: < http://example.org/sim_oa.
@prefix simda: < http://example.org/simda.

# mapeamento das classes da Ontologia Fonte SIMDA para a Ontologia da Aplicação SIMDA
p:simoaToSimda
 a r2r:ClassMapping ;
r2r:targetPattern "?s rdf:type sim_oa:Dengue" ;
r2r:sourcePattern "?s rdf:type simda:CasosDengue" .
r2r:targetPattern "?s rdf:type sim_oa:Bairro" ;
r2r:sourcePattern "?s rdf:type simda:Bairro" .
r2r:targetPattern "?s rdf:type sim_oa:Ano" ;
r2r:sourcePattern "?s rdf:type simda:Ano" .
r2r:targetPattern "?s rdf:type sim_oa:Mes" ;
r2r:sourcePattern "?s rdf:type simda:Mes" .

# mapeamento das propriedades da Ontologia Fonte SIMDA para a Ontologia da Aplicação SIMDA
p:simoaToSimda
 a r2r:ClassMapping ;
r2r:targetPattern "?s rdf:Property sim_oa:mesDengue" ;
r2r:sourcePattern "?s rdf:Property simda:mes" .
r2r:targetPattern "?s rdf:Property sim_oa:anoDengue" ;
r2r:sourcePattern "?s rdf:Property simda:ano" .
r2r:targetPattern "?s rdf:Property sim_oa:eRegistradoDengueNoBairro" ;
r2r:sourcePattern "?s rdf:Property simda:bairro"
```

Fonte: Elaboração do autor.

5.4.6 Processamento Das Consultas Sob O Esquema Mediado

O *Apache Jena*[16] e *Sesame*[17] são *frameworks* Java para o desenvolvimento de aplicações de *Linked Data* e *Web Semântica*. As duas ferramentas possibilitam o processamento de consultas federadas, necessário para a implementação do nosso mediador, mas o *Sesame* apenas suporta a linguagem RDF, não OWL. Para implementar o processo proposto utilizou-se o *framework Jena*, assim como fizeram Pinheiro (2011) e Magalhaes (2012).

[16] https://**jena**.apache.org/
[17] http://rdf4j.org

A geração dos planos de execução de consulta é demonstrada considerando o trabalho de Pinheiro (2011), como já mencionado. Apesar da arquitetura mostrada na Figura 29, considerar consultas pré-definidas, ou seja, os planos de consultas já estarão armazenados no *Repositório,* esse passo define a geração desses planos de execução, e será realizado em tempo de projeto.

Suponha a consulta sobre a ontologia de domínio *LARIISA* do exemplo utilizado. Suponha uma consulta que retorne IDH, precipitação, densidade demográfica e número de casos de dengue, representada na Quadro 5.

Quadro 5 - Exemplo de Consulta SPARQL submetida ao Mediador Criado

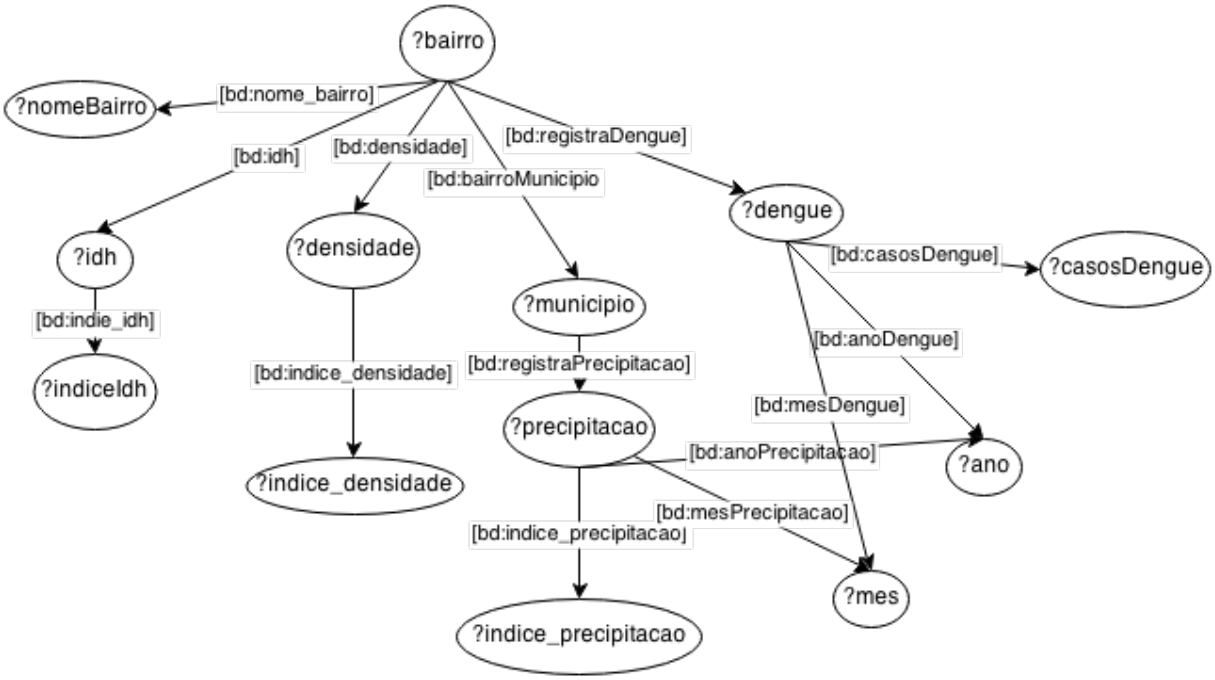
```
PREFIX lariisa: <http://lariisa#>

SELECT ?indiceIdh ?indice_precipitacao ?indice_densidade ?casosDengue
WHERE {
        ?bairro bd:nome_bairro ?:nomeBairro ;
        ?bairro bd:idh ?idh ;
        ?idh bd:índice_idh ?indiceIdh ;
        ?bairro bd:densidade ?densidade ;
        ?densidade bd:indice_densidade ?indice_densidade
        ?bairro bd:bairroMunicipio ?municipio
        ?municipio bd:registraPrecipitacao ?precipitacao
        ?precipitacao bd:mesPrecipitacao ?:mes
        ?precipitacao bd:anoPrecipitacao ?:ano
        ?precipitacao bd:índice_precipitacao ?indice_precipitacao
        ?bairro bd:registraDengue ?dengue
        ?dengue bd:mesDengue ?:mes
        ?dengue bd:anoDengue ?:ano
        ?dengue bd:casosDengue ?casosDengue
}
```

Fonte: Elaboração do autor.

Essa consulta possui como parâmetros **?:nomeBairro**, **?:mes** e **?:ano** e serão enviados pela aplicação ao mediador, assim como a aplicação envia um

parâmetro que identifica a consulta da Quadro 5. Em tempo de projeto, submete-se a consulta ao processo de geração dos planos de execução de consultas federadas detalhado na seção anterior. A consulta ao passar pela etapa de *tradução* é transformada numa árvore que representa a estrutura da consulta sobre a ontologia de domínio, conforme a Figura 35.

Figura 35 - Árvore de consulta gerada ao submeter a consulta da Quadro 5 ao processo de *tradução*



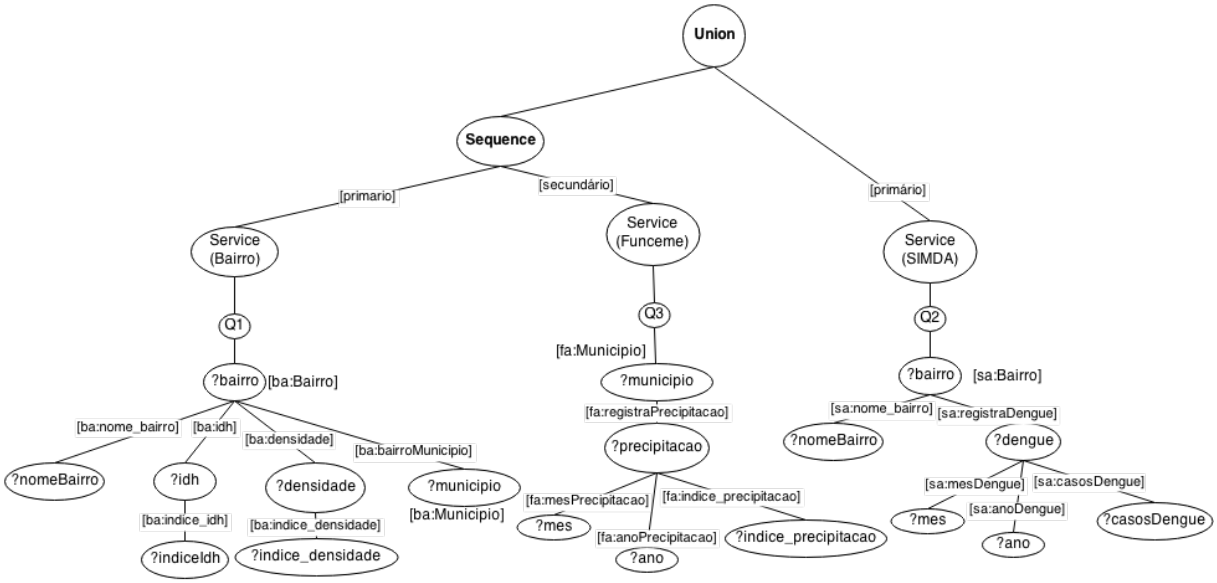
Fonte: Elaboração do autor.

Ao passar pela etapa de reformulação da consulta, a árvore gerada na etapa de tradução é reformulada. Para isso, primeiro, selecionam-se as ontologias de aplicação cujo vocabulário contém o conceito primário da consulta, identificadas como *primário* na **Error! Reference source not found.**. Para cada novo conceito encontrado nas ontologias primárias, cria-se uma sub-consulta sobre uma ontologia de aplicação, que serão denominadas de *secundárias.* O plano final da consulta consiste na união dos planos de consulta reformulada nas ontologias primárias. Caso tenha sido identificada mais de uma ontologia primária, o operador Union será o nó-raiz.

A **Error! Reference source not found.** mostra a árvore resultante da reformulação de consulta. Para o cenário descrito, as ontologias de aplicação

BAIRRO e SIMDA foram as ontologias de aplicação primárias selecionadas, por conterem em seu vocabulário o conceito primário ?bairro. A ontologia de aplicação FUNCEME, por conter um novo vocabulário, foi definida como sendo a secundária.

Figura 36 - Árvore reformulada após a etapa de reformulação de consulta. Cada operador *Service* contém sub consultas sobre as ontologias de aplicação



Fonte: Elaboração do autor.

O plano de execução é composto por 3 sub consultas Q1, Q2 e Q3 (vide Quadro 6). Em seguida cada sub consulta é reescrita conforme os mapeamentos locais da Quadro 2, em termos de uma consulta sobre o esquema das fontes de dados. As sub consultas Q1', Q2' e Q3' são apresentadas na Quadro 7. O plano de consulta federado em álgebra SPARQL gerado em memória é apresentado na Figura 37. Então é convertida e armazenada no *Repositório* sendo atrelada a ela um identificador.

O plano de consulta gerado pode ser executado pela *engine* de execução (*Executor de Consulta*) do próprio Jena/ARQ[18], já que a API suporta consultas federadas; pode ser executado utilizando a própria estratégia de execução de consultas federadas proposta no trabalho de (PINHEIRO, 2011), que através de resultados experimentais evidenciaram melhor desempenho se comparado a *engine* do JENA/ARQ; ou pode ser executado também pela módulo de processamento de consultas federadas QEF-LD (*Query Evaluation Framework - Linked Data*)

[18] *Query Engine* do *framework Jena* que suporta a linguagem de consulta RDF, SPARQL

apresentado no trabalho de Magalhaes (2012). Veremos na seção de *Trabalhos Futuros* que o algoritmo do *Executor de Consulta* é um campo vasto a ser explorado.

Quadro 6 - Consultas SPARQL sobre as ontologias de aplicação

Q1	**SELECT** ?nomeBairro ?indiceIdh ?indice_densidade ?municipio **WHERE{** ?bairro ba:nome_bairro ?nomeBairro ?bairro ba:idh ?idh ?idh ba:índice_idh ?indiceIdh ?bairro ba:densidade ?densidade ?densidade ba:índice_densidade ?indice_densidade ?bairro ba:bairroMunicipio ?municipio **}**	**Q2**	**SELECT** ?nomeBairro ?casosDengue **WHERE {** ?bairro sa:nome_bairro ?nomeBairro ?bairro sa:registraDengue ?dengue ?dengue sa:mesDengue **?:mes** ?dengue sa:anoDengue **?:ano** ?dengue sa:casosDengue ?casosDengue **}**
Q3	**SELECT** ?indice_precipitacao **WHERE {** ?municipio fa:registraPrecipitacao ?precipitacao ?precipitacao fa:mesPrecipitacao **?:mes** ?precipitacao fa:anoPrecipitacao **?:ano** ?precipitacao fa:índice_precipitacao ?indice_precipitacao **}**		

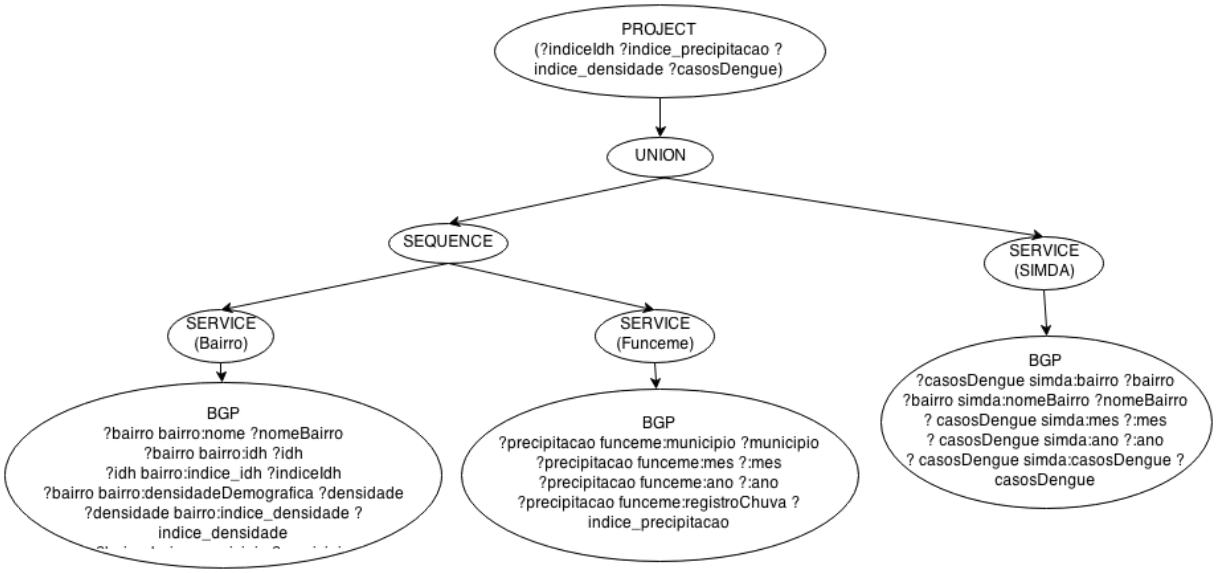
Fonte: Elaboração do autor.

Quadro 7 - Consultas SPARQL sobre as ontologias fontes

Q1'	**SELECT** ?nomeBairro ?indiceIdh ?indice_densidade ?municipio **WHERE{** ?bairro bairro:nome ?nomeBairro ?bairro bairro:idh ?idh ?idh bairro:índice_idh ?indiceIdh ?bairro bairro:densidadeDemografica ?densidade ?densidade bairro:índice_densidade ?indice_densidade ?bairro bairro:municipio ?municipio **}**	**Q2'**	**SELECT** ?nomeBairro ?casosDengue **WHERE {** ?casosDengue simda:bairro ?bairro ?bairro simda:nomeBairro ?nomeBairro ? casosDengue simda:mes **?:mes** ? casosDengue simda:ano **?:ano** ? casosDengue simda:casosDengue ?casosDengue **}**
Q3'	**SELECT** ?indice_precipitacao **WHERE {** ?precipitacao funceme:municipio ?municipio ?precipitacao funceme:mes **?:mes** ?precipitacao funceme:ano **?:ano** ?precipitacao funceme:registroChuva ?indice_precipitacao **}**		

Fonte: Elaboração do autor.

Figura 37 - Consulta federada em álgebra SPARQL.



Fonte: Elaboração do autor.

5.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foi apresentado o processo de integração de dados no contexto do projeto LARIISA, considerando a arquitetura de 3 níveis baseada em ontologias apresenta por Sacramento et al. (2010). O processo constitui-se de seis passos, a saber: (i) Definição da ontologia de domínio; (ii) Seleção das fontes de dados; (iii) Triplificação dos dados (se dados não estiverem no padrão RDF); (iv) Modelagem das ontologias de aplicação; (v) Definição dos mapeamentos locais e de mediação; e o (vi) Processamento de consultas. A etapa de triplificação dos dados pode ser de responsabilidade tanto do LARIISA, quanto do mantenedor da fonte de dados, a depender da estratégia de negócio.

Também foi apresentado a arquitetura que define o ambiente de execução das aplicações LARIISA. O esquema mediado de dados estará disponível para as aplicações como um serviço Web e possui os seguintes componentes: (i) Processador de Consulta; (ii) Repositório, e; (iii) Executor de Consulta.

Para exemplificar o processo proposto, foi definido um cenário de uma demanda de um gestor de saúde, na área da dengue, que envolvia a integração de fontes de dados distintas. Aplicou-se, então, o processo proposto, etapa a etapa,

com fins de se obter uma aplicação que interagisse com o esquema mediado, como proposto pela arquitetura apresentado na seção 5.3.

No próximo capítulo, são feitas as considerações finais, assim como expectativas de trabalhos futuros.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

6.1 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma proposta de processo de integração de dados para o LARIISA, considerando fontes independentes, heterogêneas e distribuídas. As tecnologias cernes do processo são as relacionadas à Web Semântica, as ontologias e *Linked Data.* Para corroborar a utilização dessas tecnologias realizou-se um estudo sobre as possibilidades tecnológicas de integração de dados que poderia ser utilizado pela plataforma, apontando para um processo de integração baseado em mediadores construídos por ontologias.

Como destacado no Capítulo 2, a grande maioria dos trabalhos desenvolvidos no contexto do LARIISA focaram em aplicações que captavam informações de contexto e as entregavam à plataforma para realização de inferência. Este trabalho é alicerçado na inteligência que a plataforma pode entregar ao gestor de saúde, a partir de uma base de conhecimento criada da integração de dados de saúde ou relacionados, sejam dados provindos de aplicações desenvolvidas a partir do LARIISA, sejam dados provindos de sistemas de informação de saúde já existentes.

Com o estudo também foi possível concluir que a heterogeneidade semântica dos dados se mostra como principal desafio para as soluções de integração, visto a escala de informações geradas e consumidas atualmente. Para o LARIISA, abordagens de integração que resolvem às questões semânticas dos dados também se mostraram as mais indicadas.

Prova da importância em promover a interoperabilidade semântica no LARIISA, foi o recente projeto apoiado pelo Fundo de Inovação Tecnológica (FIT) da Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (Funcap), titulado NextSaúde. O projeto é voltado em aplicar o LARIISA para o desenvolvimento de soluções inovadoras que promovam a interoperabilidade semântica entre sistemas legados e os novos sistemas desenvolvidos, especificamente no âmbito do Sistema Único de Saúde (SUS). O NextSaúde será um dos projetos onde será aplicado o processo de integração apresentado neste trabalho.

O estudo das abordagens de integração também contribuiu para observar as compatibilidades e incompatibilidades de cada abordagem com a plataforma LARIISA. Foram elas: (i) Integração por *Web Services*; (ii) integração por criação de esquema único de dados; (iii) Integração por *Data Warehouse,* e*;* (iv) Integração por Mediadores.

Uma contribuição importante deste trabalho está relacionado à independência que se proporciona ao LARIISA dos dados que podem compor a sua base de conhecimento. Pode-se afirmar que quaisquer dados/informações, que estejam dispostas na Web, em forma de planilhas, arquivos de texto ou qualquer outro formato; que estejam dispostos em sistemas diretamente ligados à saúde pública ou não; que estejam dispostos em bases de dados relacionais ou não, ou seja, qualquer informação, pode compor a base de conhecimento do LARIISA.

Isso fica evidenciado quando se avalia o cenário da contribuição que o projeto objetiva entregar à saúde pública. A burocracia e o custo que se teria de transformar os dados da saúde pública em conhecimento para o LARIISA, considerando a enorme pulverização de SIS existentes, e consequentemente começar a entregar inteligência de governança de saúde, parecia algo muito distante. A nova cultura de transparência de informações, onde mais e mais setores do governo disponibilizam informações na Web, facilita muito a utilização do processo proposto neste trabalho. Isso pode ser constatado ao observar que o cenário onde se aplicou o processo proposto foi construído somente com dados abertos disponibilizados em sites oficiais na Web. Não foi preciso ter acesso diretamente à base de dados dos sistemas que produziam essas informações, apenas acessar a página Web das instituições. Mesmo ciente da limitação desses dados, essa informação é válida para se observar a tendência dos dados abertos.

Este trabalho considerou trabalhos relevantes para a construção do processo proposto, como o de Pinheiro (2011) e Magalhães (2012). Utilizou-se a arquitetura de mediação de três níveis, ao invés da arquitetura de dois níveis, pois a arquitetura simplifica a definição dos mapeamentos de mediação, facilitando assim o processo de reformulação de consulta, bom como a integração de dados. Também foi definido que a ontologia de domínio do LARISSA pode ser formada por sub-ontologias de domínio, especializadas em um determinado conhecimento, interligadas por mapeamentos interontologias.

Também foi apresentada uma arquitetura, adaptada de Magalhães (2012), que define um ambiente de execução para o esquema mediado. A arquitetura considera o esquema mediado como um serviço Web, possibilitando as aplicações terem acesso a esse esquema através de requisições.

Essa arquitetura possibilita ao LARIISA ser utilizado, por exemplo, como solução a uma demanda específica de integração de dados da saúde pública. Como já mostrado no Capítulo 1, e no Capítulo 3 (trabalhos relacionados), qualquer esfera da saúde pública brasileira, seja ela municipal, estadual ou federal, sofre com a pulverização de dados da saúde, o que favorece para um gestão pobre e pouco eficiente. O LARIISA, como abordagem de integração de dados de saúde, fugindo de iniciativas isoladas de um município ou de um estado, poderia transformar a saúde pública no país, não só porque a plataforma poderia realizar a tarefa de integração de dados, sem interferir na dinâmica de captação de dados já existente, mas também por ter meios da própria plataforma realizar análises desses dados, indicando ao gestor as decisões mais acertadas, ou servindo como segunda opinião frente a um conselho de profissionais especializados.

Este trabalho também está contribuindo para o desenvolvimento do projeto GISSA (Governança Inteligente de Sistemas de Saúde), projeto aprovado em 2015 para financiamento pela FINEP (Financiadora de Estudos e Projetos). Este projeto objetiva aplicar a plataforma LARIISA em um domínio específico, o da Rede Cegonha[19]. O GISSA é baseado em tecnologia OLAP, para fornecimento de *Business Intelligence* (BI), e em Ontologias e RDF para fornecimento de inteligência através de inferências. O processo apresentado neste trabalho será aplicado para construir uma base ontológica que possa possibilitar as inferências do sistema.

Em resumo, se for analisado a problemática apresentada no Capítulo 1, assim como os objetivos e contribuições que desejava-se alcançar, conclui-se que o trabalho foi relevante, pois:

- Realizou-se um estudo da problemática de **interoperabilidade semântica** no projeto LARIISA (Seção 4.2);
- Especificou-se um processo de **integração de dados** para o LARIISA (Seção 5.2);
- Definiu-se um cenário de aplicação do processo proposto (Seção 5.4);

[19] http://dab.saude.gov.br/portaldab/ape_redecegonha.php

- Apresentou-se uma arquitetura **de integração de dados** para o LARIISA (Seção 5.3);
- Aplicação do trabalho em projetos LARIISA, integrando P&D (Pesquisa e Desenvolvimento):
  - Projeto **NextSAUDE** (FUNCAP);
  - Projeto **GISSA** (Rede Cegonha - FINEP).

6.2 TRABALHOS FUTUROS

Este trabalho, de forma alguma, exauriu as pesquisas a serem realizadas em torno do assunto, o que demonstra a complexidade do assunto pesquisado. Como expectativas para trabalhos futuros pode-se citar:

a) **Modelagem de uma ontologia de domínio para o LARIISA.** A modelagem do conhecimento de saúde dentro do contexto do LARIISA é um campo a ser bastante explorado e bastante complexo. Pela complexidade da área de estudo, que envolve a utilização das melhores técnicas de criação de ontologias para a saúde, seja utilizando abordagens automáticas ou manuais, da utilização de conhecimento especializado, foi definido no Capítulo 5 que a ontologia de domínio “LARIISA” seria o conjunto de sub-ontologias (e seus relacionamentos) que modelassem um determinado nível de conhecimento na área da saúde, como vigilância epidemiológica, *home care*, doenças e drogas, gestão de recursos, etc. A modelagem dessas ontologias, envolvendo as melhores técnicas, pode ser explorado em trabalhos futuros.

b) **Definição de um algoritmo para execução das consultas federadas.** Como destacado no capítulo 5, este trabalho não possuiu como objetivo chegar a esse nível de profundidade na definição do processo. O plano de execução de consultas federadas gerado poderia ser submetido ao próprio *JENA/ARQ*, como ao algoritmo proposto por Pinheiro (2011), Magalhaes (2012) ou algum outro. É necessário que, como trabalho futuro, se possa atestar sobre os

melhores algoritmos de execução de consultas federadas no contexto do LARIISA, ou até a proposição ou evolução dos já existentes.

c) **Desenvolvimento de um protótipo utilizando o processo proposto no contexto da saúde pública.** Também é necessário a continuidade deste trabalho com o desenvolvimento de um protótipo que se valha do processo aqui descrito, avaliando tanto a disponibilização do esquema mediado como serviço, como a avaliação do conhecimento e inferências realizadas pelo LARIISA após a modelagem das ontologias de domínio.
d) **Definição da melhor abordagem de publicação dos esquemas mediados como serviços.** É necessário realizar um estudo capaz de identificar qual a melhor abordagem de publicação desses serviços no contexto do LARIISA, se utilizando SOAP ou REST.

Como é possível constatar, são várias as direções em que se pode expandir e melhorar o trabalho apresentado, cujos resultados têm possibilidade de aplicação imediata, o que, por sinal, era um objetivo que também tinha-se em mente ao iniciá-lo.

REFERÊNCIAS

ALCÂNTARA, Taciano Pinheiro de Almeida. **Paola:** Uma Plataforma Para O Desenvolvimento De Aplicações Baseadas em Ontologias para o Projeto Lariisa. 2012. 124 f. Dissertação (Mestrado Profissional em Computação Aplicada) - Centro de Ciências e Tecnologia, Universidade Estadual do Ceará, Fortaleza, 2012.

ALONSO, Gustavo *et al.* Web Services. In:_____. **Web Services:** Concepts, Architectures and Applications. Heidelberg: Springer, 2004. p 123-149. Disponível em: <http://ahvaz.ist.unomaha.edu/azad/temp/softarch/04-alonso-webservices-server-architecture-soa.pdf>. Acesso em: 23 ago. 2014. .

ANDRADE, Luiz Odorico Monteiro De. Inteligência de Governança para apoio à Tomada de Decisão. **Ciência & Saúde Coletiva,** v. 17, n. 4, p. 829–832 , abr. 2012. Disponível em: <http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S1413-81232012000400003&lng=pt&nrm=iso&tlng=pt>. Acesso em: 30 out. 2013.

ANTUNES, Franciano. **SISA:** Uma aplicação sensível ao contexto para agravos de Dengue : uma prova de conceito do projeto Lariisa. 2011. 123 f. Dissertação (Mestrado Profissional em Computação Aplicada) - Centro de Ciências e Tecnologia, Universidade Estadual do Ceará, Fortaleza, 2011.

BATH, Peter. A. Health informatics: current issues and challenges. **Journal of Information Science,** v. 34, n. 4, p. 501–518 , 13 jun. 2008. Disponível em: <http://jis.sagepub.com/cgi/doi/10.1177/0165551508092267>. Acesso em: 17 jan. 2014.

BATISTA, Maria da Conceição Moraes. **Otimização de acesso em um sistema de integração de dados através do uso de caching e materialização de dados**. 2003. 128 f. Dissertação (mestrado) - Pós-Graduação em Ciência da Computação, Recife, Universidade Federal de Pernambuco, 2003.

BAUER, Florian; KALTENBÖCK, Martin. **Linked Open Data : The Essentials**. Viena: edition mono/monochrom, 2012.

BERNERS-LEE, Tim. *Linked Data - Design Issues*. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 15 maio 2014.

BERNERS-LEE, Tim; HEATH, Tom; BIZER, Christian. Linked Data - The Story So Far. **International Journal on Semantic Web and Information Systems,** v. 5, n. 3, p. 1–22, 2009.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. **The Semantic Web**: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Disponível em: <http://www.dke.univie.ac.at/semanticweb/history/ws0405/BeHL01.ps>. Acesso em: 29 abr. 2014.

BIZER, Christian; SCHULTZ, Andreas. The R2R framework: Publishing and discovering mappings on the web. In: 1st International Workshop on Cosuming Data, 2010, Shangai. **CEUR Workshop Proceedings** Shangai: 2010.

BUITELAAR, Paul; MAGNINI, Bernardo. **Ontology Learning from Text**: An Overview. IOS Press, 2005. p. 3–12. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.70.3041>. Acesso em: 17 abr. 2015.

CASTAÑEDA, William Alberto Cruz. **NOVO PARADIGMA DA ENGENHARIA CLÍNICA NA INTEGRAÇÃO DE TIC´s PARA CRIAÇÃO DE AMBIENTES UBIQUOS E DE INTEROPERABILIDADE NA SAÚDE.** 2011. 128 f. Dissertação (mestrado) - Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal De Santa Catarina, 2011.

CIFUENTES-SILVA, Francisco; SIFAQUI, Christian; LABRA-GAYO, Jose Emilio. Towards an architecture and adoption process for Linked Data technologies in Open Government contexts: A case study for the Library of Congress of Chile. In: 7th International Conference on Semantic Systems, 2011, Graz, Austria. **Proceedings of the 7th International Conference on Semantic Systems**, New York: ACM, 2011. p.79–86.

DATASUS. *Departamento de Informática do SUS*. Disponível em: <http://datasus.saude.gov.br/>. Acesso em: 13 mar. 2014.

DAVENPORT, Thomas H.; PRUSAK, Laurence. **Conhecimento empresarial: como as organizações gerenciam o seu capital intelectual**. 4 ed. Rio de Janeiro: Campus, 1998. 237 p.

DEY, Anind K. **Providing Architectural Support for Building Context-Aware Applications**. 2000. 188 f. Tese (Doutorado) - Georgia Institute of Technology, Georgia, 2000.

DOAN, AnHai; HALEVY, Alon; IVES, Zachary. **Principles of Data Integration**. Waltham, MA: Elsevier, 2012. 497 p.

FENSEL, Dieter. **Ontologies:** A Silver Bullet for Knowledge Management and Eletronic Commerce. Hilderberg, Berlim: Springer, 2001.

FERNANDEZ, Mariano; GOMEZ-PEREZ, Asuncion; JURISTO, Natalia. **METHONTOLOGY :** From Ontological Art Towards Ontological Engineering. Disponível em: <http://oa.upm.es/5484/1/METHONTOLOGY_.pdf>. Acesso em: 2 de junho de 2014.

FLAUZINO, Regina Fernandes; SOUZA-SANTOS, Reinaldo; OLIVEIRA, Rosely Magalhães De. Indicadores socioambientais para vigilância da dengue em nível local. **Saúde e Sociedade** v. 20, n. 1, p. 225–240 , mar. 2011. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-12902011000100023&lng=en&nrm=iso&tlng=pt>. Acesso em: 7 out. 2014.

FROTA, João Batista Bezerra. **Proposta de solução de integração de provedores de contexto ao sistema lariisa**. 2011. 78 f. Dissertação (Mestrado Profissional em Computação Aplicada) - Centro de Ciências e Tecnologia, Universidade Estadual do Ceará, Fortaleza, 2011.

GARDINI, Leonardo M. *et al.* Clariisa, a context-aware framework based on geolocation for a health care governance system. **2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)** n. Healthcom, p. 334–339 , out. 2013. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6720696>. Acesso em: 14 de julho 2014.

GEREMIAS, Marcos Aurélio; JACOBSEN, Alessandra de Linhares; PEREIRA, Juliana. Superação dos Desafios na Integração dos Sistemas de Informação em Saúde na Secretaria Municipal de Saúde de Florianópolis. **Coleção Gestão da Saúde Pública - Volume 8**. Florianópolis: Fundação Boiteux, 2013. 1 v. p. 146–160.

GRAHAM, Ian D *et al.* Lost in knowledge translation: time for a map? **The Journal of continuing education in the health professions,** v. 26, n. 1, p. 13–24 , jan. 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16557505>. Acesso em: 16 jul. 2014.

GRUBER, Thomas R. A translation approach to portable ontology specifications. **Knowledge Acquisition** v. 5, n. 2, p. 199–220 , jun. 1993. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1042814383710083>. Acesso em: 29 jul. 2014.

GUBIANI, Juçara Salete; PORT, Rafael; ORNELLAS, Marcos Cordeiro D. Interoperabilidade Semântica do Prontuário Eletrônico do Paciente. In: Simpósio de Informática da Região Centro do RS, 2003. **Anais do II Simpósio de Informática da Região Centro do RS**, Santa Maria, RS: UNIFRA, 2003.

GUPTA, Ashid; MUMICK, Inderpal Singh. Maintenance of Materialized Views : Problems , Techniques , and Applications. **IEEE Data Eng. Bull.** v. 18, n. 2, p. 3–18 , 1995.

HANSEN, Mark; MADNICK, Stuart; SIEGEL, Michael. Data Integration using Web Services. In: VLDB 2002 Workshop EEXTT and CAiSE 2002 Workshop DTWeb on Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web-Revised Papers, 2002. **Proceedings of the VLDB 2002 Workshop EEXTT and CAiSE 2002 Workshop DTWeb on Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web-Revised Papers**, 2003, Londres, UK: Springer-Verlag. p.165–182. Disponível em: <http://web.mit.edu/smadnick/www/wp/2002-14.pdf>. Acesso em: 23 ago. 2014.

HÄYRINEN, Kristiina; SARANTO, Kaija; NYKÄNEN, Pirkko. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. **International journal of medical informatics** v. 77, n. 5, p. 291–304 , maio 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17951106>. Acesso em: 23 jan. 2014.

HEATH, Tom; BIZER, Christian. **Linked Data: Evolving the Web into a Global Data Space**. Morgan & Claypool, 2011. 136 p. Disponível em: <http://linkeddatabook.com/editions/1.0/#htoc79>. Acesso em: 15 maio 2013.

HIRA, Adilson Yuuji. **Saúde Digital :** Novo Paradigma Da Convergência Das Tecnologias De Informação Para A Área Da Saúde. 2012. 244 f. Tese (Doutorado) - Departamento de Engenharia de Sistemas Eletrônicos, Escola Politécnica da Universidade de São Paulo, 2012.

JUNIOR, Wilson Coelho de Souza. **Integração de sistemas de informações em saúde.** Uma proposta de solução para a melhoria da qualidade na gestão do SUS. 2009. 150 f. Dissertação (Mestrado Profissional em Saúde Pública) - Fundação Oswaldo Cruz, 2009.

LANGEGGER, Andreas; WOS, Wolfram. XLWrap – Querying and Integrating Arbitrary Spreadsheets with SPARQL. **Proceedings of the 8th International Semantic Web Conference,** Berlim: Springer, 2009. 359-374 p. .

LEÃO, Beatriz de F *et al.* O Desafio de Integrar os Sistemas de Informação em Saúde. **Anais do IX Congresso Brasileiro de Informática em Saúde**, Ribeirão Preto: SBIS, 2004. p.1–6.

LOPES, Paula M A; ANDRADE, Rafael; WANGENHEIM, Aldo Von. Uma Ontologia para o Atendimento Emergencial de Pacientes. **Anais do XI Congresso Brasileiro de Informática em Saúde**, Campos do Jordão: SBIS, 2011. p.1–6.

MAGALHÃES, REGIS PIRES. **UM AMBIENTE PARA PROCESSAMENTO DE CONSULTAS FEDERADAS EM LINKED DATA MASHUPS**. 2012. 118 f. Dissertação (mestrado) - Departamento de Computação, Universidade Federal Do Ceará, 2012.

MARCONDES, Carlos Henrique. *"Linked data" – dados interligados - e interoperabilidade entre arquivos, bibliotecas e museus na web*. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**. Santa Catarina: UFSC, 2012, v. 17, n. 34. Disponível em: <http://www.periodicos.ufsc.br/index.php/eb/article/view/24410>. Acesso em: 20 nov. 2012. , 9 ago. 2012

MASSAD, Eduardo; MARIN, Heimar de Fátima; AZEVEDO, Raymundo Soares De (Orgs.). **O PRONTUÁRIO ELETRÔNICO DO PACIENTE NA ASSISTÊNCIA , INFORMAÇÃO E CONHECIMENTO MÉDICO**. São Paulo: H. de F. Marin, 2003. 213 p.

MEDEIROS, Wilma Maria da Costa; OLIVEIRA, Luiz Affonso H. Guedes De; SOUSA, Lirisnei Gomes De. Uso de Ontologias para Acesso a Informações de Saúde Armazenadas em Bases de Dados Heterogêneas. **Anais do XI Congresso Brasileiro de Informática em Saúde**, Campos do Jordão: SBIS, 2011.

MONDINI, Adriano; CHIARAVALLOTI, Francisco Neto. Variáveis socioeconômicas e a transmissão de dengue. **Revista de Saúde Pública** v. 41, n. 6, p. 923–930, dez. 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102007000600006&lng=en&nrm=iso&tlng=pt>. Acesso em: 7 out. 2014.

MONTENEGRO, Lívia Cozer *et al.* Sistema de informação como instrumento de gestão: perspectivas e desafios em um hospital filantrópico. **Journal of Health**

Informatics v. 5, n. 1 , 29 mar. 2013. Disponível em: <http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/203>. Acesso em: 18 abr. 2014.

MORAES, Rodrigo Leal De. **Sistemas de Data Warehouse: Estudo e Aplicação na Área da Saúde**. 1998. Dissertação (mestrado) - UFRGS, 1998.

NARDON, Fabiane Bizinella; JR, Lincoln de A Moura. Knowledge Sharing and Information Integration in Healthcare using Ontologies and Deductive Databases. **MEDINFO.** Amsterdan: IOS PRESS, 2004. p.62–66.

NOY, Natalya F.; MCGUINNESS, Deborah L. *Ontology Development 101: A Guide to Creating Your First Ontology*. Disponível em: <http://arion.csd.uwo.ca/courses/CS9626a/papers_files/ontology-tutorial-noy-mcguinness.pdf>. Acesso em: 22 abr. 2015. , 2001

OLIVEIRA, Mauro *et al.* A context-aware framework for health care governance decision-making systems: A model based on the Brazilian Digital TV. **2010 IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)** p. 1–6 , jun. 2010.

PIERRO, Bruno De. *Dados sobre saúde precisam de integração | Brasilianas.Org*. Disponível em: <http://www.advivo.com.br/materia-artigo/dados-sobre-saude-precisam-de-integracao>. Acesso em: 18 maio 2013.

PINHEIRO, João Carlos. **Processamento de Consulta em um Framework baseado em Mediador para Integração de Dados no Padrão de Linked Data**. Universidade Federal do Ceará, 2011.

PINHEIRO, Taciano *et al.* PELADA – UM PRONTUÁRIO ELETRÔNICO LARISSA-DATASUS, PARA UMA PLATAFORMA SENSÍVEL AO CONTEXTO. In: X Encontro de Pesquisa e Pós-Graduação do IFCE. Fortaleza: IFCE, 2011.

PINTO, Luiz Felipe Da Silva. **ESTRATÉGIAS DE INTEGRAÇÃO E UTILIZAÇÃO DE BANCOS DE DADOS NACIONAIS PARA AVALIAÇÃO DE POLÍTICAS DE SAÚDE NO BRASIL**. 2006. 224 f. Tese (doutorado) - Departamento de Ciências Sociais, Escola Nacional de Saúde Pública Sergio Arouca, 2006.

PIRES, Daniel Facciolo; RUIZ, Evandro Eduardo Seron. Interoperabilidade terminológica em sistemas de informação em saúde : problemas e soluções com a UMLS. **Journal of Health Informatics** v. 2, n. 2, p. 34–42 , 2010.

ROUSSEY, Catherine *et al.* An Introduction to Ontologies and Ontology Engineering. **Ontologies in Urban Development Projects**. Advanced Information and Knowledge Processing. London: Springer London, 2011. 1 v. p. 9–39. Disponível em: <http://link.springer.com/10.1007/978-0-85729-724-2>. Acesso em: 18 out. 2014. 978-0-85729-723-5.

SACRAMENTO, Eveline R *et al.* Towards Automatic Generation of Application Ontologies. **Journal of Information and Data Management** v. 1, n. 3, p. 535–550 , 2010.

SANTOS, Ivomar; TELES, Germanno; OLIVEIRA, Mauro. Interface do mecanismo de apoio à decisão baseado em redes bayesianas para a plataforma LARIISA. In: VI Congresso Tecnológico Infobrasil Ti & Telecom, Fortaleza, 2013.

SANTOS, MARCOS EDUARDO DA SILVA. **DIGA SAÚDE - UMA PROPOSTA DE SISTEMA DE APOIO A SERVIÇOS DE HOME CARE BASEADO NO MODELO BRASILEIRO DE TV DIGITAL**. 2011.

SANTOS, Ricardo da Silva; GUTIERREZ, Marco Antônio. MINERSUS – Ambiente computacional para extração de informações para a gestão da saúde pública por meio da mineração dos dados do SUS MINERSUS – A computational. **Revista Brasileira de Engenharia Biomédica** v. 24, n. 2, p. 77–90 , 2008.

SANTOS, Ricardo S *et al.* Data Warehouse para a Saúde Pública : Estudo de Caso SES-SP. **Anais do X Congresso Brasileiro de Informática em Saúde**, Florianópolis, SC: SBIS, 2006. p.53–58.

SANTOS, Ricardo S.; GUTIERREZ, Marco Antônio. Implementações de Data Warehouse na Área da Saúde. **Anais do IX Congresso Brasileiro de Informática em Saúde**, Ribeirão Preto: SBIS, 2004. p.125–130.

SHETH, Amit P. Changing Focus on Interoperability in Information Systems:From System, Syntax, Structure to Semantics. In: GOODCHILD, Michael *et al.* (Orgs.). **Interoperating Geographic Information Systems**. New York: Springer US, 1999. p. 5–29.

SILVA, Ismael S *et al.* Observatório da Dengue : Surveillance based on Twitter Sentiment Stream Analysis. In: SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS, 26., Florianópolis, 2011.

SPAZZIANI, Bianca De Oliveira; NARDON, Fabiane Bizinella. Componentização e Integração de Sistemas de Informação em Saúde de Grande Porte. In: Congresso Brasileiro de Informática em Saúde, 9., 2004. **Anais...**, Ribeirão Preto, SP: SBIS, 2004.

TELES, Germanno Gurgel do Amaral. **UM MECANISMO DE APOIO À TOMADA DE DECISÃO EM AGRAVO DE DENGUE BASEADO EM DADOS PROBABILÍSTICOS**. 2013. 101 f. Dissertação (Mestrado Profissional em Computação Aplicada) - Centro de Ciências e Tecnologia, Universidade Estadual do Ceará, Fortaleza, 2013.

UMAPATHY, Karthikeyan; PURAO, Sandeep. Systems Integration and Web Services. **IEEE Computer Society** v. 43, n. 11, p. 91–94 , 2010.

WACHE, H. *et al.* Ontology-Based Integration of Information — A Survey of Existing Approaches. In: IJCAI-01 Workshop: Ontologies and Information, 2001, Seattle, USA. **Anais...** Seattle, USA: 2001. p.108–117.

WIDOM, Jennifer. Research Problems in Data Warehousing. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 4., 1995, Baltimore, USA. **Anais...** New York, USA: ACM, 1995. p.25–30. Disponível em: <http://ilpubs.stanford.edu:8090/91/1/1995-24.pdf>. Acesso em: 1 set. 2014.

WIEDERHOLD, Gio. Mediators in the Architecture of Future Information Systems. **The IEEE Computer Magazine,** n. 3 , 1992. p 38-49.

YU, Liyang. **A Developer's Guide to the Semantic Web**. New York: Springer, 2011.

ZIEGLER, Patrick; DITTRICH, Klaus R. Data Integration — Problems, Approaches, and Perspectives. In: KROGSTIE, John; OPDAHL, Andreas Lothe; BRINKKEMPER, Sjaak (Eds.). **Conceptual Modelling in Information Systems Engineering**. Heidelberg: Springer, 2007. p. 39–58. Disponível em: <http://www.uniriotec.br/~tanaka/SAIN/ZieglerSolvberg2007.pdf>. Acesso em: 19 ago. 2014.