EDUARDO PAZ SERAFIM

# COLLECTMED: EXTRAÇÃO E REUSO DE CONHECIMENTO COLETIVO PARA O REGISTRO ELETRÔNICO EM SAÚDE

João Pessoa

2011

EDUARDO PAZ SERAFIM

# COLLECTMED: EXTRAÇÃO E REUSO DE CONHECIMENTO COLETIVO PARA O REGISTRO ELETRÔNICO EM SAÚDE

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal da Paraíba como requisito parcial para obtenção do título de Mestre em Informática.

Área de concentração:
Sistemas de Computação.

Orientador:
Prof. Dr. Gustavo Henrique Matos Bezerra Motta

João Pessoa

2011

À minha família. Seu apoio incondicional me guia.

# Agradecimentos

Aos meus pais, João Neto e Vitória, principais responsáveis pela minha educação agradeço pelos grandes exemplos de dedicação e empenho.

À minha irmã, Ana Caroline, eu agradeço pelo carinho e apoio, mesmo durante tempos difíceis, me ajudando a crescer e amadurecer.

À Aline, minha namorada, que demonstrou muita paciência, sempre com palavras encorajadoras e sinceras quando precisei.

Aos amigos, Thiago e Lamara, que tenho certeza, estarão sempre presentes em minha vida como verdadeiros irmãos.

Ao orientador, Professor Gustavo Motta, acima de qualquer coisa, agradeço pela confiança depositada desde o início do meu trabalho no LArqSS.

Aos amigos de graduação, pós e "Quintas", Alysson, Arthur, Bruno, Daniel, Everaldo, João Paulo, Jonâtas, Moisés, Pizzol, Rodrigo, Targino e Tati, pelos momentos de descontração e companheirismo genuinamente essenciais ao longo destes anos.

Um agradecimento especial a Gustavo Cavalcanti (in memoriam) que foi para todos ao seu redor um enorme exemplo de força, confiança e perseverança.

Aos companheiros de trabalho no LArqSS, Andrea, Brunna, Duílio, Hélio, Hugo, João, Luciano, Renan e Walber, que me auxiliaram muito além das suas atribuições, tanto no ambiente do laboratório, assim como fora dele.

Por fim, meu agradecimento à FINEP pelo auxílio financeiro que possibilitou a realização deste trabalho.

# Sumário

# Lista de Figuras e Gráficos

# Lista de Tabelas

# Lista de Acrônimos

| | |
|---|---|
| AM | *Aprendizagem de máquina* |
| API | *Application Programming Interface* |
| ARFF | *Attribute-Relation File Format* |
| CASE | *Computer Aided Software Engineering* |
| CDS | *Clinical Decision Support* |
| CRISP-DM | *Cross Industry Standard Process for Data Mining* |
| CTI | *Central de Tratamento Intensivo* |
| DM | *Data Mining* |
| DSS | *Decision Support System* |
| EAV | *Entity Attribute Value* |
| EHR | *Electronic Health Record* |
| EJB | *Enterprise Java Beans* |
| ER | *Entidade Relacionamento* |
| FEM | *Ficha de Evolução Médica* |
| FINEP | *Financiadora de Estudos e Projetos* |
| GNU | *GNU´s Not Unix* |
| GPL | *General Public License* |
| HULW | *Hospital Universitário Lauro Wanderley* |
| IDE | *Integrated Development Enviroment* |
| IRI | *Uniform Resource Identifier* |
| JPA | *Java Persistence API* |
| JSF | *JavaServer Faces* |
| LArqSS | *Laboratório de Arquitetura e Sistemas de Software* |
| MACA | *Middleware de Autenticação e Controle de Acesso* |
| OLAP | *Online Analytical Processing* |
| ORM | *Object-Relational Mapping* |
| OWL | *Web Ontology Language* |
| POJO | *Plain Old Java Object* |
| RES | *Registro Eletrônico em Saúde* |
| SDSS | *Specific Decision Support System* |
| UFPB | *Universidade Federal da Paraíba* |
| UML | *Unified Modeling Language* |
| UTI | *Unidade de Terapia Intensiva* |
| XML | *eXtensible Markup Language* |

# Resumo

SERAFIM, E. P. **CollectMed: Extração e Reuso de Conhecimento Coletivo para o Registro Eletrônico em Saúde**. 2011. 119 p. Dissertação (Mestrado) – Departamento de Informática, Universidade Federal da Paraíba, João Pessoa, 2011.

Diversos avanços tecnológicos ocorridos nos últimos anos fizeram com que os Sistemas de Registro Eletrônico em Saúde (RES) se consolidassem como uma alternativa viável para substituir, progressivamente e com eficiência, o uso dos registros de saúde em papel. Os benefícios encontrados são associados ao uso de métodos de apoio à decisão clínica, disponibilidade dos dados, facilidade na busca por informações, entre outras vantagens inerentes ao uso de sistemas computadorizados. Entretanto, existem ainda, muitos desafios e pesquisas para fazer com que todo o potencial desses sistemas seja utilizado. Por exemplo, a quantidade de dados clínicos que os sistemas de RES armazenam, é muito elevado. Diversos interesses poderiam ser beneficiados, caso houvesse uma ferramenta capaz de realizar uma análise automatizada, ou semi-automatizada (como é mais comumente encontrada), para buscar padrões úteis no conjunto de dados armazenados no sistema.

Diversos trabalhos apontam que os esforços realizados no campo de aprendizado automático alcançam ótimos resultados em diversas áreas, inclusive para informações clínicas. Porém, o esforço necessário ainda é elevado, aumentando o tempo dedicado ao planejamento e execução, assim como altos custos e necessidade de grande volume de dados para o processamento. Este trabalho, associado ao sistema de apoio à decisão do OpenCTI busca reduzir, significativamente,o esforço necessário para promover tanto o reuso de informações clínicas a partir do aprendizado automático, quanto o desenvolvimento de mecanismos de apoio à decisão clínica a um baixo custo.

O presente trabalho, busca oferecer tal benefício aos usuários de sistemas de RES, por meio de um mecanismo simples, porém amplo, de análise dos dados clínicos armazenados nos bancos de dados dos RES. Essa análise será realizada por meio de uma metodologia de extração de conhecimento, utilizando algoritmos de inteligência coletiva ou *data mining*, passando por etapas de busca, seleção, pré-processamento, modelagem, avaliação e aplicação destas informações extraídas dos sistemas. A partir disso, mecanismos de apoio à decisão clínica dos RES, poderão utilizar o arcabouço oferecido pelo CollectMed para promover, com mais facilidade e precisão, recuperação de informações mais apuradas a respeito das condições clínicas específicas sobre seus pacientes, de acordo com o que já foi registrado por profissionais de saúde em casos clínicos semelhantes persistidos no RES.

Palavras-chave: Sistemas Computadorizados de Registros Médicos, Inteligência Coletiva, Extração de Informação, Aprendizagem de Máquina.

# Abstract

SERAFIM, E. P. **CollectMed: Extração e Reuso de Conhecimento Coletivo para o Registro Eletrônico em Saúde.** 2011. 119 p. Dissertação (Mestrado) – Departamento de Informática, Universidade Federal da Paraíba, João Pessoa, 2011.

Several technological advances during recent years provided that the Electronic Health Record systems (EHR) became a solidified and viable alternative to replace progressively and efficiently, the use of health records on paper. The benefits found are associated with the use of methods for clinical decision support (CDS), data availability, ease in finding information, among other advantages inherent in computerized systems use. However, there are still many challenges and research to get the full potential of such systems. For example, the amounts of clinical data for EHR storage are very high. Several interests might benefit if there was a tool capable of performing an automated analysis, or more commonly found, semi-automated, useful for search patterns in the data set stored in the system.

Several studies indicate that efforts in the field of machine learning achieve great results in various areas including clinical information. However, the effort required is still high, increasing the time spent with planning and processing, with high costs and large amounts of data needed for processing. This work, in association with the OpenCTI's CDS seeks to significantly reduce the amount of effort necessary to promote both the reuse of clinical information from the automatic learning, and the development of mechanisms for clinical decision support with low cost.

This study seeks to offer those benefits to users of EHR systems, through a simple mechanism, but extensive, for analysis of clinical data stored in clinical databases. This analysis is performed using a methodology of knowledge extraction algorithms using collective intelligence or data mining, through steps of search, selection, preprocessing, modeling, evaluation and application of the information extracted from these systems. From this, mechanisms for clinical decision support of EHR, may use the framework offered by CollectMed to promote with greater ease and precision, more accurate information regarding specific medical conditions on their patients, according to what has already been registered by health professionals in similar cases using the EHR.

Keywords: Medical Records Systems, Ontologies, Collective Intelligence, Information Extraction, Machine Learning.

Capítulo

# 1

# Introdução

"A resposta certa, não importa nada: o essencial é que as perguntas estejam certas."

*Mário Quintana*

Este trabalho visa propor uma solução que contribua para extração e promoção do reuso de informações mantidas pelo registro eletrônico em saúde (RES) por meio de mecanismos de extração de conhecimento coletivo e *data mining* (DM). Neste capítulo, apresentamos as principais questões que motivaram o desenvolvimento do trabalho, assim como o seu objetivo, as justificativas relacionadas e metodologia aplicada ao longo do projeto.

## 1.1 Motivação

O RES surgiu como alternativa ao uso de registros de saúde em papel e, em pouco tempo, motivou o desenvolvimento de diversas tecnologias inovadoras neste novo campo. Algumas destas tecnologias têm como objetivo auxiliar os usuários no processo de levantamento de informações clínicas, facilidade para recuperação de registros antigos e aumento da disponibilidade desses dados em diversos ambientes, independentemente de limites geográficos.

Entre outras vantagens inerentes à automatização de sistemas em ambientes complexos, tal como é o ambiente de prestação de serviços de saúde, o advento do RES causou também um aumento significativo no volume dos dados clínicos armazenados pelas organizações de saúde. Tal aumento impõe desafios relacionados a encontrar formas simples e eficientes de coletar informações nessa grande base de dados formada com o uso do RES, e posteriormente dispor de conhecimento obtido

dessas informações para os usuários, e ainda mais importante, de forma transparente. Isso permitiria que os usuários manipulassem dados clínicos, obtidos por meio do emprego de grande esforço computacional, sem saber ao menos da existência de tais mecanismos complexos.

Adicionalmente aos avanços citados, uma das principais características e benefícios do RES é servir como meio onde podem ser aplicados mecanismos e técnicas de auxílio à decisão clínica (*Clinical Decision Support*, CDS). Por exemplo, métodos de validação de dados ou sugestões auxiliam os usuários, evitando que pequenos erros de digitação ocorram ou mesmo indicando condutas terapêuticas. Muitos desses métodos poderiam ser beneficiados caso houvesse uma forma simples de resgatar informações contextualizadas e pertinentes do conhecimento clínico existente na base de dados do RES (GREENES, 2007a).

Em linhas gerais, tais métodos de apoio à decisão são criados com o auxílio de profissionais especializados no domínio de aplicação e, na sua composição, diversas regras do negócio são desenvolvidas por inserção direta no código fonte da aplicação. Por exemplo, um mecanismo de validação em um RES, referente a um campo que registre a frequência cardíaca média do paciente, é usualmente codificado com estruturas de desvios condicionais onde o valor médio aceitável foi declarado explicitamente no código fonte da aplicação. Um método de CDS desenvolvido dessa forma poderia apenas ser aproveitado em um ambiente restrito, onde as regras aplicadas no método de apoio à decisão sejam válidas. Em um ambiente de cuidado à saúde, entretanto, essa rigidez não é adequada. As diversidades de quadros clínicos que podem ser encontrados, durante o processo de auxílio à saúde, demandam uma solução mais dinâmica e flexível, onde as condições de desvios sejam adaptadas ao contexto.

Dessa forma, o desenvolvimento e implantação de métodos de apoio à decisão clínica só poderá ser realmente efetivo se conseguir levar em consideração não apenas os dados manipulados indistintamente, mas também em qual setor do hospital ele será aplicado, a natureza do problema clínico do paciente, entre outras variáveis do ambiente. Exemplificando, um método que verifique se o peso registrado de um paciente está entre 0 Kg e 150 Kg não é de completa utilidade em uma UTI neonatal. Caso o usuário, por engano, digitasse 15 Kg, enquanto o valor que ele realmente pretendia registrar fosse 1.5 Kg, esse método não seria capaz de identificar o possível engano e aconselhar o usuário a verificar se o valor inserido está correto. É im-

portante evitar esse tipo de erro, que levaria a sérias consequências para o neonato, cuja prescrição depende diretamente do peso indicado. Caso fosse possível verificar em tempo de execução se o valor digitado está de acordo com o restante dos valores registrados usualmente naquele setor do hospital, especificamente, ou em outro que possua pacientes com um quadro clínico semelhante, seria possível garantir flexibilidade ao método de CDS. Um único método, capaz de realizar tal verificação durante a sua execução, poderia ser aplicado em diversos setores que utilizam o sistema, e teria o seu comportamento ajustado de acordo com o contexto, no exemplo, o grupo de pacientes de uma UTI neonatal.

Além de dados numéricos, o mesmo princípio pode ser levado em consideração para informações textuais. Existem termos de saúde, medicamentos ou procedimentos que são utilizados, geralmente, em um contexto comum. Por exemplo, caso um paciente sofra de problemas cardíacos, os resultados de medicamentos para aquela condição clínica podem ser filtrados para sugerir, primeiramente, aqueles que são prescritos quando os pacientes possuem esse tipo de doença, de forma que as sugestões oferecidas pelos métodos de CDS sejam mais efetivas.

Para representar as informações que estarão disponíveis no RES, existem esforços que se baseiam em arquétipos e ontologias de domínio (OPENEHR FOUNDATION; SPÄTH, 2010; LEZCANO, 2011; BRASS, 2010), onde estão modelados os diversos conceitos de saúde. O uso desse tipo de abordagem oferece generalidade ao RES ao fazer com que novos conceitos de saúde e documentos sejam adicionados sem necessidade de empregar esforço de programação e manutenção do modelo de dados. Em contra partida, existe um aumento no nível de complexidade necessário para disponibilizar este arcabouço genérico, assim como para realizar buscas sobre a base de dados.

Associada ao aumento gradual no volume de dados armazenados, a complexidade do modelo de persistência é também incrementada em decorrência do uso das ontologias, impossibilitando análise não automatizada das informações. A existência de uma ferramenta que auxilie na consulta e extração desse conhecimento para reuso, automaticamente, é fundamental. De outro modo, todos esses dados serão de pouca utilidade para a atenção à saúde do paciente, ou mesmo a organização de saúde que é responsável por sua guarda devendo mantê-los por tempo indeterminado (CONSELHO FEDERAL DE MEDICINA, 2007), e incorrendo em elevadas

despesas de manutenção de sistemas, espaço de armazenamento, entre outros custos relacionados.

## 1.2 Objetivo

O objetivo deste trabalho é desenvolver, testar e aplicar o CollectMed (C*ollective* M*edical* D*ata*), uma ferramenta para extração e aplicação de conhecimento coletivo em um RES apoiado em ontologias, de forma que a descoberta e aplicação dessas informações sejam efetivadas de forma simples. Objetiva-se que as informações extraídas com base no conhecimento coletivo dos usuários sejam reutilizadas por meio de sugestões contextualizadas integradas ao RES, ou que essas possam ser utilizadas em outros métodos de CDS elevando o seu grau de generalidade e de reuso. Como objetivos específicos no desenvolvimento desse trabalho têm-se:

Objetivo 1. Desenvolver uma ferramenta que permita consultar uma base de dados clínicos baseada em ontologias. A seleção desses dados deve ocorrer de acordo com a escolha do usuário, objetivando compor um conjunto de dados a partir do qual serão executados algoritmos de inteligência coletiva e/ou *data mining*;

Objetivo 2. Avaliar métodos de apoio à decisão clínica que possam se beneficiar de mecanismos de extração de conhecimento coletivo e mineração de dados;

Objetivo 3. Utilizar um ambiente flexível capaz de utilizar diversos algoritmos de mineração de dados e/ou de inteligência coletiva;

Objetivo 4. Avaliar e disponibilizar o conhecimento extraído de forma que seja possível realizar consultas de forma simples e transparente sobre estes modelos;

Objetivo 5. Aplicação da solução desenvolvida, onde as informações extraídas de um RES baseado em ontologias sejam utilizadas em métodos CDS simulando o uso à beira do leito, de modo a demonstrar a aplicabilidade dos métodos desenvolvidos no Col-

lectMed. Utilizaremos para isso o RES do sistema OpenCTI[1], doravante denominado simplesmente OpenCTI.

## 1.3 Justificativa

Como foi apresentado nas seções anteriores, o volume de dados mantido pelo RES é potencialmente muito extenso. Isso impossibilita uma análise não automatizada dos mesmos (MITNITSKI, 2003). Adicionalmente, existe demanda desses dados nos métodos de CDS para oferecer uma melhor contextualização e por consequência, uma demanda por melhores resultados em sugestões de preenchimento para o usuário final do sistema. Portanto como justificativas para a realização dos objetivos específicos desse trabalho podemos citar:

Justificativa 1. Desenvolver uma ferramenta capaz de consultar uma base de dados clínicos baseada em ontologias oferece possibilidade de realizar consultas sem necessidade de conhecer a estrutura do banco de dados clínico em profundidade. Os dados selecionados irão servir como base para a execução dos algoritmos de *data mining*. Além disso, possibilitará aos usuários, extrair dados para diversas pesquisas que são comumente executadas em ambientes hospitalares, principalmente em um hospital-escola;

Justificativa 2. Com a determinação de métodos de apoio à decisão clínica a serem beneficiados por meio de mecanismos de extração de conhecimento coletivo e *data mining,* é possível montar um arcabouço com um bom nível de abstração que possibilite grande utilidade dos mecanismos e modelos criados a partir da ferramenta desenvolvida;

Justificativa 3. É esperado que novos algoritmos de *data mining* sejam desenvolvidos frequentemente, oferecendo melhor desempenho em relação aos seus antecessores (WITTEN, 2005). Em virtude disso, visa-se desenvolver um ambiente flexível, onde novos métodos, não disponíveis no momento de concep-

1 OpenCTI: Software de uma Central de Telemedicina para Apoio à Decisão Médica em Medicina Intensiva. Projeto financiado pela FINEP nº 01.08.0533.00.

ção da ferramenta, possam ser incorporados por ela, com o menor esforço possível;

Justificativa 4. Ao oferecer uma forma simples de disponibilização das informações adquiridas pelo processo de *data mining*, visa-se ampliar a reutilização dos dados previamente persistidos no banco de dados clínico;

Justificativa 5. Ao utilizar os dados extraídos diretamente à beira de leito, esperamos que eles se fizessem úteis justamente no momento mais crítico do processo de prestação de cuidados à saúde (GREENES, 2007a). Por meio dessas informações disponibilizadas de acordo com o contexto onde é aplicado e juntamente com informações relativas ao estado de saúde do paciente, podem ser evitados erros no preenchimento das informações clínicas e obter melhores resultados também em sugestões de preenchimento.

## 1.4 Metodologia

Para compor a metodologia de trabalho, foi efetuada uma revisão bibliográfica das soluções e tecnologias disponíveis para auxiliar na solução do problema que se propôs resolver. Em seguida realizou-se a avaliação das tecnologias fundamentais e a escolha do processo de desenvolvimento e dos princípios que alicerçam este trabalho.

### 1.4.1 Licença de software

O CollectMed faz parte da solução proposta para o OpenCTI, desta forma, obedece à mesma licença de software aplicada, ou seja, é um software *open source*, sob licença GPL.

### 1.4.2 A abordagem adotada: tecnologias fundamentais

Como tecnologias fundamentais que serão utilizadas ao longo do trabalho, foi escolhida a linguagem de programação Java (ORACLE, 2006a), por possuir ampla disseminação, ser independente de plataforma, e oferecer um grande número de APIs e *frameworks* implementados e disponíveis nessa linguagem. Com isso visa-se reduzir

o tempo de desenvolvimento e integração com outras tecnologias que serão necessárias para a execução do projeto.

Como ferramenta para documentar o sistema a ser desenvolvido, escolhemos a forma de diagramação UML (*Unified Modeling Language*) (OBJECT MANAGEMENT GROUP, 2006) por sua simplicidade associada ao mesmo tempo com um alto poder de expressividade e aceitação, assim como *JavaDoc* (ORACLE, 2010b) para construir a documentação do código fonte produzido. Mais detalhes serão apresentados sobre aspectos tecnológicos envolvidos podem ser encontrados na subseção 1.4.5.

### 1.4.3 Processo de desenvolvimento

Para a construção do CollectMed, o processo de desenvolvimento escolhido foi o desenvolvimento evolucionário (SOMMERVILLE, 2007). Nesse processo, temos como objetivo trabalhar alternando atividades de especificação projeto, implementação e validação. Com o objetivo de construir um produto que atenda aos requisitos previamente definidos, é realizada uma especificação inicial, em seguida um protótipo é projetado e desenvolvido, e então avaliado. Esse processo é repetido por meio de refinamentos sucessivos até alcançar os requisitos iniciais ou outros definidos durante o desenvolvimento.

### 1.4.4 Princípios

Durante o desenvolvimento da ferramenta CollectMed, alguns pontos serão levados em consideração. Esses são importantes durante o processo de desenvolvimento por apontar determinados caminhos que se deve trilhar para manter o trabalho de acordo com essa metodologia. São eles:

- **Qualidade** – Prima-se pela qualidade do produto a ser desenvolvido, dessa forma é evitado o retrabalho para corrigir problemas deixados pela falta de cuidado no desenvolvimento;

- **Reuso** – Sempre que houver uma solução já consolidada, com alto nível de aceitação e qualidade, essa será reutilizada no desenvolvimento;

- **Manutenibilidade** – Trabalhar de forma que o código do sistema seja de fácil manutenção, tento em vista que o tempo de manutenção de um software é responsável pela maior parcela no seu ciclo de vida de desenvolvimento.

## 1.4.5 APIs e Frameworks

Durante o desenvolvimento da ferramenta CollectMed, tornou-se mandatória a integração com o OpenCTI para simplificar o posterior uso da ferramenta por parte dos seus usuários futuros. Desta forma, o CollectMed compartilha das principais tecnologias que são utilizadas pelo projeto OpenCTI, dentre as quais, podemos destacar as seguintes tecnologias.

### 1.4.5.1 *Framework* Seam

No CollectMed, o *framework* Seam é utilizado para promover algumas funções que simplificam a atividade de desenvolvimento de uma aplicação web. Algumas das características que o *framework* Seam oferece e são aproveitadas no CollectMed dizem respeito à presença de um *container* inversão de controle, integração entre JSF e EJB 3.0, uso abrangente de anotações em detrimento de configurações em XML para a aplicação, gerência de contexto de persistência e, finalmente, testes de integração simulando interações com o usuário, utilizando os frameworks de teste JUnit ou TestNG.

### 1.4.5.2 JPA e Hibernate

Com o intuito de manter independência em relação aos fornecedores de soluções para a camada de persistência, utilizamos a JPA (*Java Persistence API)* que faz parte da especificação da tecnologia Java para persistências de POJOs (*Plain Old Java Objects*). Assim, uma possível substituição da implementação pode ser realizada sem necessidade de refatorar grande parte do código dedicado à persistência. Embora a especificação JPA seja bastante abrangente, a implementação oferecida pelo Hibernate possui alguns facilitadores, além de cobrir a especificação JPA.

A função, portanto, do conjunto JPA/Hibenate é oferecer todo o suporte para o mapeamento objeto-relacional (ORM, ou Object Relational Mapping) em Java. A escolha do Hibernate frente a outras implementações da especificação JPA, por exemplo, TopLink (ORACLE, 2011), OpenJPA (APACHE, 2010) ou EclipseLink (ECLIPSE FOUNDATION, 2011), se deu pelo fato da implementação Hibernate ser

altamente disseminada, testada e validada por muitas aplicações e ser uma biblioteca de código aberto.

#### 1.4.5.3 Middleware de Autorização e Controle de Acesso (MACA)

Para oferecer maior controle sobre o uso da aplicação, apenas usuários autênticos do sistema podem utilizar e administrar o CollectMed. O controle de acesso de usuários e autorização de suas ações dentro do sistema é realizado por meio do serviço oferecido pelo Middleware de Autorização e Controle de Acesso, MACA (MOTTA, 2004), através de uma API padronizada, independente de plataforma e linguagem de programação. No MACA é implementado um modelo controle de acesso é baseado em papéis (CABP), provendo escalabilidade para usuários e recursos administrados, onde os diversos papéis podem receber autorizações diferenciadas, viabilizando a definição de políticas de controle de acesso.

### 1.4.6 Ferramentas utilizadas

Durante o desenvolvimento o CollectMed, foi utilizado um conjunto de ferramentas abaixo listadas e descritas brevemente. Estas ferramentas apresentaram-se de valia para auxiliar o desenvolvimento e documentação do CollectMed.

#### 1.4.6.1 Eclipse Jboss Tools

Ambiente integrado de desenvolvimento (IDE) de código aberto Eclipse, integrado com um conjunto de "*plug-ins*" selecionados para promover o desenvolvimento de aplicações Web com o framework Seam e servidor de aplicação JBoss. Dentre os *plug-ins* que foram utilizados no IDE, destacam-se:

1. Seam Dev Tools (JBoss, 2010a) – promover a criação de novos projetos com o *framework* Seam, configurando suas dependências e com ferramentas wizards para instanciação de novos componentes para a aplicação. O Seam Dev Tools usado encontra-se sob versão 3.1.0. GA;

2. JBoss Server Manager (Jboss, 2010b) - utilizado em sua versão 2.1.0.GA para configurar, iniciar, reiniciar ou interromper o servidor de aplicação em modo normal ou debug, além de realizar *deploy* do código no servidor, oferecer acesso aos *logs* do sistema, entre outras operações;

3. TestNG (2010)– plug-in para criação, execução e monitoramento de testes unitários e de integração das classes desenvolvidas para a aplicação CollectMed.

#### 1.4.6.2 Astah Community

A ferramenta CASE (*Computer Aided Software Engineering*) (SOMMERVILLE, 2007) Astah Community (CHANGE VISION, 2010) é utilizada para a modelagem em linguagem UML do sistema. O Astah Community é de uso livre, portanto compatível com a metodologia aplicada no projeto.

#### 1.4.6.3 Weka

O Weka oferece uma extensa coleção de algoritmos de aprendizagem de máquina (*machine learning*) que podem ser aplicados, através da sua ferramenta, diretamente a um conjunto de dados selecionados ou utilizados a partir de código Java, disponibilizado por sua API. Como o Weka é possível utilizar diversas atividades de pré-processamento, classificação, regressão, agrupamento, formação de regras de associação e visualização dos dados e resultados obtidos. Assim como a ferramenta, o código fonte é disponível sob licença de software Gnu GPL, condição necessária para o seu uso junto ao código do CollectMed.

Os algoritmos presentes no Weka foram testados e validados no uso em diversas aplicações desde o seu lançamento inicial, consequentes correções e ampliações, e desta forma, apresentam-se como alternativa segura e consolidada, viabilizando o seu uso em detrimento do desenvolvimento de implementações próprias de algoritmos de aprendizagem de máquina. A versão do Weka 3.6.3 foi utilizada neste trabalho.

### 1.4.7 Ambiente de produção

Durante o desenvolvimento deste trabalho foram utilizados recursos do Laboratório de Arquitetura e Sistemas de Software (LArqSS) do Departamento de Informática filiado à Universidade Federal da Paraíba. O ambiente conta com estações de trabalho, servidores, impressora, dispositivos de armazenamento, roteadores, entre outros recursos. Equipamentos estes adquiridos com recursos do projeto OpenCTI, financiado pela FINEP. O laboratório contava, em 2011, com uma equipe de aproximadamente 10 colaboradores, entre coordenador, pesquisadores e estagiários, em

sua maioria dedicada ao projeto OpenCTI ou subprojetos relacionados, como é o cenário do projeto CollectMed.

## 1.5 Estrutura do trabalho

Esta dissertação encontra-se estruturada de acordo com a seguinte lista de capítulos e respectivos objetivos:

- No segundo capítulo, “Fundamentação teórica”, são encontradas definições dos principais conceitos que são utilizados no escopo deste trabalho e serão levantadas, também, questões relacionadas à extração de conhecimento em bancos de dados e outros aspectos relevantes para a resolução do problema apresentado;
- No terceiro capítulo, “CollectMed”, serão apresentados alguns requisitos da solução tecnológica através da qual buscamos alcançar os objetivos levantados na seção 1.2 e também como esta solução foi desenvolvida durante o trabalho realizado;
- O quarto capítulo, “Resultados”, concentram-se informações sobre os efeitos alcançados com o desenvolvimento deste trabalho;
- No quinto capítulo, “Considerações finais” serão levantadas discussões sobre os resultados alcançados, a indicação de trabalhos futuros proporcionados a partir deste trabalho.

Capítulo

# 2

# Fundamentação teórica

> "Talvez os problemas filosóficos sejam difíceis não porque sejam divinos, irredutíveis, sem sentido ou ciência rotineira, mas porque a mente do Homo sapiens não dispõe do equipamento cognitivo para resolvê-los. Somos organismos, e não anjos, e nossa mente é um órgão, e não um conduto para a verdade".
>
> *Steven Pinker*

Este capítulo procede apresentando os principais conceitos e fundamentos que serão utilizados ao longo da dissertação. Tais embasamentos abordados dizem respeito aos registros eletrônicos em saúde, sistemas de apoio à decisão, extração de informação e aprendizagem automatizada de máquina, assim como o registro eletrônico em saúde OpenCTI, juntamente com suas características de persistência, modelo semântico baseado em ontologia, geração de interface e suporte ao apoio à decisão. Estes conceitos são de suma importância para entendimento do contexto onde o projeto CollectMed se aplica, a sua definição e as características que deve apresentar para que sejam alcançados os objetivos deste trabalho.

## 2.1 Registros eletrônicos em saúde

Por sistemas de registros eletrônicos em saúde (RES), se podem entender aqueles sistemas que ofereçam mecanismos para entrada e recuperação de dados clínicos, ordens eletrônicas de medicamentos e compartilhamento das informações entre profissionais que objetivam prestar algum tipo de cuidado à saúde de um paciente ou população (GUNTER, 2005).

O advento dos registros eletrônicos em saúde proporcionou um aumento na quantidade de dados clínicos disponíveis, dessa foi forma, é possível acelerar diversas pesquisas relacionadas à saúde, impactando diretamente no nível quantitativo de conhecimento e informações disponíveis para os profissionais de saúde (HOFFMANM, 2008). Armazenamento físico de documentos com registros de saúde é bastante problemático e custoso. Manter as mesmas informações eletronicamente reduz os custos em relação à manutenção de registros em papel, promove maior disponibilidade das informações e facilidade para o compartilhamento do registro de saúde. Além disso, muitas vezes registros manuscritos são associados a problemas de legibilidade, ocasionando erros ou atrasos importantes no cuidado à saúde dos pacientes (INSTITUTE OF MEDICINE, 1999).

Embora existam vantagens associadas à aplicação de sistemas para o RES, também é possível encontrar problemas decorrentes do uso dos RES. A utilização de tais sistemas pode levar à adição de uma nova categoria de erros e, em consequência, problemas no cuidado à saúde dos pacientes. Santell (2004) indica o uso de sistemas automatizados de prescrição como causa de erro em 84% de 500 hospitais e instituições de saúde que participaram da sua pesquisa. Realizando uma analogia com os problemas à saúde introduzidos durante o atendimento médico, essa nova categoria originada pelo uso de comunicação e registro eletrônico de informações de saúde foi denominada por Weiner et al. (2007) como *e-iatrogenesis* ou *technological iatrogenesis*.

## 2.2 Sistemas de apoio à decisão

Em diversos campos de aplicações, por exemplo, economia, comércio, administração, onde sistemas computadorizados são utilizados para automatizar atividades de registro, armazenamento e recuperação de informações, podem ser encontrados sistemas especializados em auxiliar os usuários a tomar decisões com base em dados e cenários apresentados. Tais aplicações são chamadas de sistemas de apoio à decisão (*Decision Support Systems*, DSS) ou sistemas de suporte à decisão (POWER, 2002).

Alguns termos encontrados na literatura e soluções de mercado podem ser considerados tipos de sistemas de apoio à decisão, por exemplo, *business intelligence,* sistemas colaborativos, *data mining, data warehousing*, gestão de conhecimento ou mesmo *on-line analytical processing* (OLAP). Power (2002) trata sistemas de apoio à decisão como sistemas computacionais interativos que auxiliam os seus usu-

ários a utilizarem as capacidades de comunicação, dados, documentos, conhecimento e modelos computacionais para resolver problemas e realizar decisões. Vale salientar que DSS são sistemas auxiliares, ou seja, não são desenvolvidos com intenção de substituir os tomadores de decisão, que além de capacitados para realizar as ações, são, acima de tudo, responsáveis pelas decisões tomadas, com ou sem o uso de sistemas de apoio à decisão.

### 2.2.1 Gerência de conhecimento clínico

Um estudo realizado por SITTIG (2010) aponta algumas características que permitem comparar sistemas de apoio à decisão em relação às funcionalidades que se relacionam à gerência de conhecimento clínico neste escopo. As principais características levantadas por SITTIG (2010) são a existência de:

1. *Equipe multidisciplinar responsável por criar e manter o conteúdo clínico* – tal equipe é composta por médicos, enfermeiros, fisioterapeutas, nutricionistas, assim como, analista de sistemas, desenvolvedores de software. A multidisciplinaridade torna-se importante neste contexto para atribuir diferentes atividades às pessoas mais capacitadas que se seja possível dentro da equipe, simplificando e especializando as atividades.

2. *Repositório de conhecimento clínico com interface web:* Em contraste com a implementação de regras e conhecimento clínico voltado ao apoio à decisão diretamente no código das aplicações, a existência de um repositório onde tal conhecimento esteja disponível para consultas e visualização por parte dos seus usuários facilita a manutenção e disseminação facilitada desde conhecimento. Deve ser possível aos usuários, partir dessa ferramenta, consultar qual é o comportamento e as indicações das ferramentas de apoio à decisão clínica sobre um determinado conceito biomédico.

3. *Ferramenta online, colaborativa e interativa:* Permitindo aos desenvolvedores e usuários utilizar-se de uma comunicação síncrona ou assíncrona (em tempo real ou não) onde seja possível discutir os benefícios e problemas relacionados às intervenções e sugestões dos mecanismos de CDS existentes no sistema. Tal ferramenta pode ser formada por um *chat*, fórum de discussão, vídeo-conferência, entre outras.

4. *Ferramenta disponível para controlar os conceitos clínicos:* Diversas terminologias e ontologias existem para dar suporte a estes conceitos (SNOMED, LOINC, ICD-9, entre outras). Além de manter conhecimento clínico para os conceitos existentes dentro do ambiente de saúde utilizando uma terminologia comum, é importante também ser capaz de criar e dar manutenção aos conceitos existentes no contexto do registro eletrônico em saúde.

   SITTIG (2010) aponta ainda, como uma característica desejável aos sistemas de gerenciamento de conhecimento clínico, que estes deveriam também apresentar, seria o processamento e reuso dos dados a fim de promover aprendizado a partir do banco de dados formado pelas informações clínicas dos pacientes. Entretanto, nenhuma das organizações, envolvidas no estudo realizado pelo autor, promoviam o uso deste tipo de ferramenta.

### 2.2.2 Arquitetura de sistemas de apoio à decisão

Diversos autores em sistemas de apoio à decisão utilizam em seus trabalhos definições de um conjunto de componentes básicos e características que um sistema de apoio à decisão deve possuir, tais definições fundamentais são apresentadas por Sprague e Carlson (1982). Embora o trabalho possua quase 30 anos desde a sua publicação, os autores retratam de forma precisa as demandas e requisitos dos sistemas de apoio à decisão até a atualidade. De acordo com Sprague e Carlson, sistemas de apoio à decisão podem ser subdivididos em 3 níveis:

5. *Specific Decision Support System* (SDSS), ou sistema de suporte à decisão específico – em relação à arquitetura dos sistemas de apoio à decisão, este elemento é quem executa tarefas. Cada elemento SDSS é dedicado a um problema isolado, esta especialização dos DSS promove melhor gerenciamento das atividades de suporte à decisão, permitindo aos tomadores de decisão lidar com diferentes grupos de problemas relacionados à tomada de decisão;

6. *Decision Support System Generators,* ou criadores de sistemas de suporte à decisão – são pacotes de software desenvolvidos para promover a criação de SDSS simplificadamente, com baixos custos e rapidamente. Com o uso dos *Decision Support System Generators* é possível realizar modelagem de SDSS e visualizar relatórios dos modelos criados;

7. DSS *Tools,* ou ferramentas de sistemas de suporte à decisão – utilizados para construir os SDSS, estas ferramentas são algoritmos, modelos estatísticos, relatórios de dados ou quaisquer outras técnicas que possibilitem a construção dos SDSS. As ferramentas de DSS são o nível fundamental para sistemas de suporte à decisão.

Algumas características levantadas por Sprague e Carlson devem ser levadas em consideração para a criação de um processo de desenvolvimento de SDSS, a fim de alcançar um sistema de apoio à decisão adaptativo e flexível. Em primeiro lugar, desenvolvimento focado em subproblemas, característica essa que leva a uma segunda, foco em SDSS pequenos, mas de fato usáveis. Com estes dois primeiros pontos é proposto que grandes problemas, para os quais tomadores de decisão desejam apoio, sejam subdivididos em problemas menores, com menor número de variáveis e, por conseqüência, soluções individuais mais simplificadas e possivelmente mais exatas, portanto de maior utilidade para os tomadores de decisão.

Outras características apontadas dizem respeito à construção de planos de refinamento ou modificação dos SDSS e mantê-los em constante avaliação. Ademais, construir e disponibilizar sistemas de apoio à decisão é apenas um passo em direção ao objetivo de prover suporte à decisão. Devem-se construir também mecanismos de manutenção por meio de atualizações planejadas, acompanhamento de suas execuções, nível de satisfação dos usuários com os resultados, entre outras métricas que possam ser aplicadas.

### 2.2.3 Apoio à decisão clínica

É categorizada como sistemas de apoio à decisão clínica (CDS, *Clinical Decision Support*) uma grande variedade de aplicações, indo desde simples procedimentos de checagem de erros, validação da entrada dos dados, até mecanismos sofisticados para monitoração contínua de resultados de exames e sinais vitais de pacientes, a formulação de hipóteses de diagnóstico diferencial, entre outros.

Alguns fatores são determinantes para o sucesso de aplicações de CDS, são eles: aproximação dos casos específicos dos pacientes, indicando sugestões ou informações que sejam válidas para aquele contexto específico; obter alta integração com o sistema (sem envolver o usuário em outras atividades para que o mecanismo de CDS atue com eficiência); e, principalmente, oferecer as informações diretamente

no ponto de cuidado à saúde, ou seja, à beira do leito do paciente (GREENES, 2007a).

Objetivando tornar possível a disponibilidade de informações aos tomadores de decisão por meio de sistemas de apoio à decisão clínica, é necessário que os dados contidos nos bancos de dados do registro eletrônico em saúde sejam analisados por ferramentas e técnicas de extração de informações e aprendizagem de máquina. Na próxima seção, os conceitos relacionados e principais técnicas utilizadas com estes objetivos são apresentados.

## 2.3 Extração de informação e aprendizagem de máquina

Em linhas gerais, a extração de informação envolve a extração de tipos predefinidos de informações a partir de texto (MEYSTRE, 2008), ou seja, dados não estruturados. Entretanto, para este trabalho, se considera que a extração de informação possui o mesmo objetivo da sua definição encontrada na literatura, contudo, com um mote diferenciado, que nesse caso são os dados semi estruturados do RES, especificamente os encontrados no registro eletrônico em saúde do OpenCTI.

A aprendizagem de máquina (AM) ou *machine learning* consiste em utilizar ou desenvolver técnicas computacionais sobre o processo de aprendizado e construir sistemas capazes de adquirir conhecimento de forma automática ou semi-automatizada (WITTEN, 2005). É importante ressaltar que objetivamos não controlar o conhecimento utilizado na geração dos dados, especificamente, o conhecimento clínico empregado para diagnosticar, realizar observações ou leitura de dados clínicos dos pacientes, mas podermos construir modelos computacionais que representem uma boa aproximação dos dados disponíveis para consulta. A AM é por sua natureza multidisciplinar. A criação de modelos utiliza-se de teorias estatísticas e matemáticas juntamente com teorias da ciência da computação para possibilitar a execução e representação destas tarefas e modelos.

Alguns métodos podem ser utilizados para alcançar o objetivo da extração de informação, ou aprendizagem de máquina, no contexto do CollectMed. Apresentamos a seguir alguns dos que estão de acordo com a metodologia aplicada neste trabalho, e que sejam apropriados para o uso sobre informações clínicas.

### 2.3.1 Inteligência coletiva

Segaran (2007) afirma que o objetivo da inteligência coletiva é colecionar dados de contribuintes independentes com objetivos em comum e, a partir disso, construir novas conclusões baseadas em levantamentos estatísticos dos grupos analisados. Essa definição geral pode ser aplicada especificamente para a extração de informação sobre dados clínicos. Profissionais de saúde (usuários do RES), que registram as informações clínicas dos seus pacientes, atuam como os chamados contribuintes independentes. No que diz respeito aos objetivos dos usuários, na definição de Segaran (2007), a correspondência pode ser feita com o objetivo do próprio RES, ou seja, registrar as informações clínicas para auxiliar no cuidado ao paciente, que deve ser executado pelos participantes do sistema. As disposições dos usuários do ambiente hospitalar, assim como dos pacientes, caracterizam os grupos que devem ser analisados com o objetivo de extrair o conhecimento coletivo mantido no sistema.

### *2.3.2 Data mining*

Witten e Frank (2005) definem *Data Mining* (DM) como o processo utilizado para descobrir padrões em dados. Esse processo, ainda segundo Witten e Frank, deve ser automatizado, ou como mais frequentemente encontrado, semi-automatizado. Esses padrões encontrados devem ser significantes, de forma que proporcionem alguma vantagem em decorrência da sua composição. Witten e Frank citam que uma possível vantagem encontrada é de ordem econômica, já que o processo de DM é bem desenvolvido e alcança um maior interesse em aplicações comerciais de sistemas com arquitetura e modelo de dados bem definidos.

O processo de DM é utilizado, predominantemente, em aplicações que manipulam um grande volume de dados, onde não é possível fazer uma análise manual dos dados do sistema. Adicionalmente, temos que em aplicações de saúde, tal como são os RES, o conjunto de dados utilizado possui um maior dinamismo quando comparado com as aplicações onde tradicionalmente o processo de DM é aplicado e bem sucedido.

O modelo de processo CRISP-DM (CHAPMAN, 2000) serve como princípios gerais para implementações de aplicações que envolvam mineração de dados. O CRISP-DM divide o processo de DM em quatro níveis. No primeiro nível, o processo é dividido em fases correspondentes a um conjunto de tarefas genéricas, pertencen-

tes ao segundo nível da divisão. Tem-se, portanto, um modelo genérico o suficiente para cobrir a todos os possíveis cenários no âmbito de atividades de *data mining*.

O modelo genérico montado demanda uma descrição de como as atividades genéricas serão executadas de acordo com as situações específicas. O terceiro nível trata justamente dessas atividades especializadas. Já no quarto nível, encontra-se a instanciação dos processos, correspondente a um registro das ações realizadas, decisões, entradas e saídas dos processos da instanciação proposta pelas atividades mais gerais. A Figura 1 representa essa divisão em níveis do processo CRISP-DM.

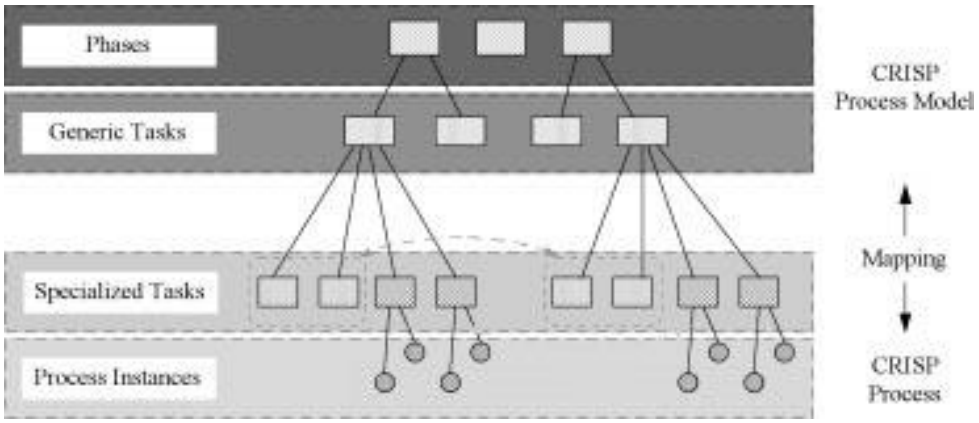


Figura 1: Divisão do modelo CRISP-DM em quatro níveis, fonte: Chapman (2000)

O modelo de referência CRISP-DM apresenta seis fases para o processo de DM, a saber: Entendimento do negócio (*Business Understanding*), Entendimento dos dados (*Data Understanding*), Preparação dos dados (*Data Preparation*), Modelagem (*Modeling*), Avaliação (*Evaluation*) e Aplicação (*Deployment*). Cada uma dessas fases possui suas atividades genéricas e resultados associados pré-definidos no modelo de referência. A Figura 2 mostra as fases do processo DM, acima citadas, assim como suas interações e é seguida por uma breve introdução das funções de cada uma das fases.

- Entendimento do negócio - Visa esclarecer os objetivos do projeto e quais são os requisitos no ponto de vista das regras de negócio. Em seguida, nessa fase, deve ser desenvolvida uma definição do problema a ser resolvido através de data mining e também de um planejamento para alcançar os objetivos definidos.
- Entendimento dos dados - Essa fase trata de obter familiarização com o conjunto de dados a ser processado. Quais são suas características intrínsecas,

problemas que podem influenciar na qualidade dos dados e formular hipóteses a partir deste primeiro contato com o conjunto de dados.

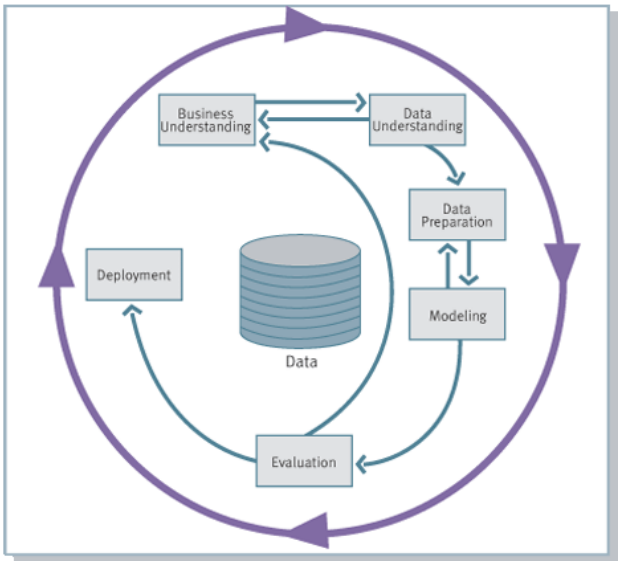


Figura 2: Ilustração do modelo CRISP-DM, fonte: Chapman (2000)

- Preparação dos dados - Nesta fase agrupam-se todas as atividades referentes à construção do data set a ser utilizado pelas ferramentas de modelagem. Por exemplo, são atividades da fase de preparação: seleção de tabelas, registros e atributos, transformação e limpeza dos dados.
- Modelagem - Existe grande variedade de técnicas que podem ser utilizadas na fase de modelagem. Durante esta etapa, deve ser escolhido um método adequado de acordo com o objetivo apontado nas primeiras fases e também realizados os ajustes necessários para calibrar os parâmetros desses métodos para obter melhores resultados no processo.
- Avaliação - A fase de avaliação formaliza o processo de testes que asseguram a qualidade do modelo encontrado na fase de modelagem. É de extrema importância por garantir que os modelos criados na fase anterior estejam de acordo com certos limiares de sucesso definidos previamente. Caso os modelos não possuam o nível de qualidade estabelecido, eles devem ser descartados ou reformulados para obter os resultados esperados.
- Aplicação - A partir da consolidação do modelo, realizada na fase anterior, o conhecimento obtido durante a execução de todas as fases do processo de DM deve ser colocado em prática. Geralmente, essa fase é caracterizada pela aplicação dos modelos criados dentro do processo de tomada de decisão da organização.

## 2.4 O registro eletrônico em saúde do OpenCTI

O CollectMed está sendo desenvolvido em paralelo ao projeto OpenCTI, contando com a infra-estrutura do Laboratório de Arquitetura e Sistemas de Software (LArqSS), do Departamento de Informática (DI), na Universidade Federal da Paraíba (UFPB). O OpenCTI será um software livre, baseado em padrões abertos, que dará apoio à decisão para o cuidado ao paciente internado em UTIs de hospitais distantes de centros de referência em saúde, que geralmente não contam com médicos especialistas em medicina intensiva ou outras especialidades relacionadas.

Entre as principais responsabilidades da central, destaca-se a gestão de informações clínicas e colaboração dos membros das equipes de saúde (local e remota). A gestão de informações clínicas irá coletar dados clínicos específicos para medicina intensiva e manterá registro eletrônico em saúde dos pacientes internados em UTIs de hospitais usuários dos serviços da central de telemedicina. Ferramentas colaborativas da central permitirão a interação dos membros da equipe e o compartilhamento de informações do RES visando auxiliar, à distância, condutas diagnósticas e terapêuticas.

### 2.4.1 Estruturação do RES

Um dos principais objetivos no uso dos RES é representar as narrativas dos profissionais de saúde acerca do estado clínico dos seus pacientes. Nessa categoria, várias informações podem ser incluídas, a saber: a) histórico de saúde familiar, b) histórico de saúde pessoal, c) exames físicos, d) notas de evolução, e) exames clínicos, entre outros. Desta forma, tem-se uma variedade de documentos que podem ser gerados, e que demandam diferentes níveis de detalhamento para sua representação.

Metodologias distintas estão disponíveis para promover a representação de conhecimento em saúde. Tais metodologias variam seu nível de estruturação entre duas formas de representação:

1. Representação em linguagem natural ou texto livre – considerada como forma ideal para representar as narrativas no domínio de aplicações em saúde, pois propicia elevado grau de expressividade. O uso de texto livre em linguagem natural apresenta desvantagens inerentes para a execução de rotinas

computacionais de pesquisa de dados e extração de informações, devido à ambiguidade associada às narrações em linguagem natural;

2. Representação completamente estruturada – forma ideal para promover pesquisas de dados e extração de informação. O uso de registros completamente estruturados é qualificado pelo uso de estruturas de dados para modelagem de todas as informações presentes na narrativa dos usuários, limitando, portanto, o grau de expressividade disponível.

Ao definir uma metodologia para modelagem dos documentos e informações clínicas do RES, é realizada uma escolha entre poder de expressividade e eficiência para busca e recuperação dos dados. A partir desta escolha, será definida a capacidade do RES em representar com mais ou menos detalhes os conceitos de saúde envolvidos. Em seu trabalho, LOS (2006) apresenta que o ideal seria montar um arcabouço sem limitações no tocante aos detalhes que a narrativa possa conter e que, ao mesmo tempo, estruture os dados, possibilitando realizar pesquisas com mais eficiência quando comparado às pesquisas em textos completamente não estruturados. O OpenCTI adota essa metodologia para concepção do seu modelo de dados, entretanto, utilizando uma perspectiva diferenciada em relação ao trabalho de LOS (2006). A seção a seguir apresenta com mais detalhes os principais conceitos que foram adotados para o OpenCTI, neste sentido.

### 2.4.2 RES baseado em ontologias

Os documentos de saúde mencionados na seção anterior são compostos por diversos conceitos biomédicos que, a partir de uma estruturação e propósito definidos, compõem os documentos manipulados pelo sistema. No OpenCTI, os conceitos biomédicos passíveis de observação por parte dos usuários, ou seja, presentes nos documentos, estão descritos em uma ontologia de conceitos biomédicos (NÓBREGA, 2010), descritos na linguagem OWL (*Web Ontology Language*) (DEAN, 2004). Na Figura 3, é apresentada uma representação resumida da ontologia dos conceitos que estruturam os dados no OpenCTI.

Os conceitos biomédicos (*BiomedicalConcept*) utilizados pelo OpenCTI podem surgir no sistema em duas formas distintas, conceitos biomédicos abstratos (*AbstractBiomedicalConcept*) ou conceitos biomédicos concretos (*ConcreteBiomedicalConcept*). Os chamados conceitos abstratos não possuem valores associados,

eles são agrupamentos semânticos, que organizam os conceitos biomédicos concretos alinhados a eles. Por outro lado, conceitos concretos podem ser subdivididos em duas categorias, qualitativos (*QualitativeConcreteBiomedicalConcept*) ou quantitativos. Os conceitos concretos qualitativos representam observações, como sinais e sintomas, que denotam um sentido por si. Por exemplo, conceitos relativos à intensidade: "forte", "grave", "bom", entre outros. Já os conceitos quantitativos representam os valores registrados para os conceitos associados, podendo ter uma unidade associada, na figura representada pelo conceito *Unit*. Os conceitos biomédicos podem ter associados a eles uma lista de exclusão mútua (*MutualExclusionList*), ou mesmo conceitos equivalentes (*EquivalentBiomedicalConcept*).

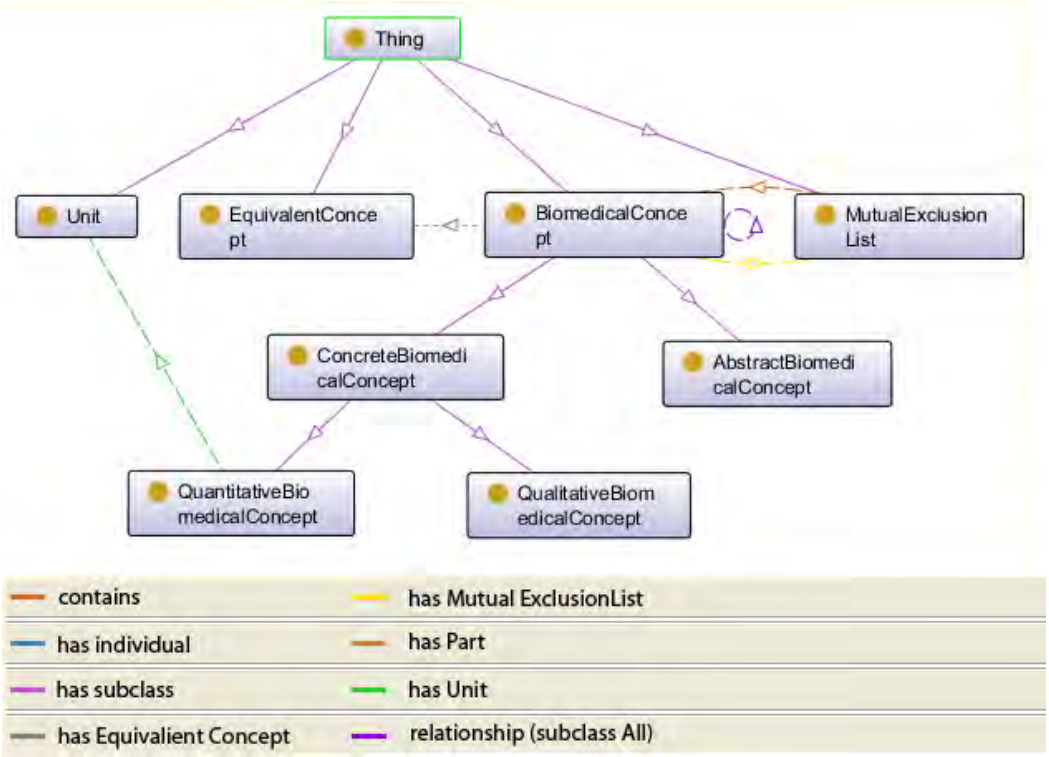


Figura 3: Representação da ontologia criada para os conceitos biomédicos do OpenCTI.

Representando a estrutura dos diferentes documentos de saúde manipulados pelo OpenCTI, a equipe de desenvolvimento do RES definiu mapeamentos da ontologia dos documentos para uma ontologia de conceitos biomédicos, proporcionando o reuso dos conceitos em contextos distintos. Com isso, tem-se que os documentos de saúde (embora possam ser vistos como um agrupamento de conceitos biomédicos) são ortogonais à estes. Chamamos de arquétipos tais agrupamentos de conceitos no âmbito de um documento. Um mesmo conceito pode ser utilizado em

vários documentos, bem como alterações realizadas sobre a estrutura dos documentos não devem influenciar na base de conceitos biomédicos.

Uma vez definido como os conceitos de saúde serão representados na ontologia do OpenCTI, é necessário modelar como os documentos que utilizam esses conceitos estão organizados. Na Figura 4, os conceitos utilizados estão representados utilizando notação UML.

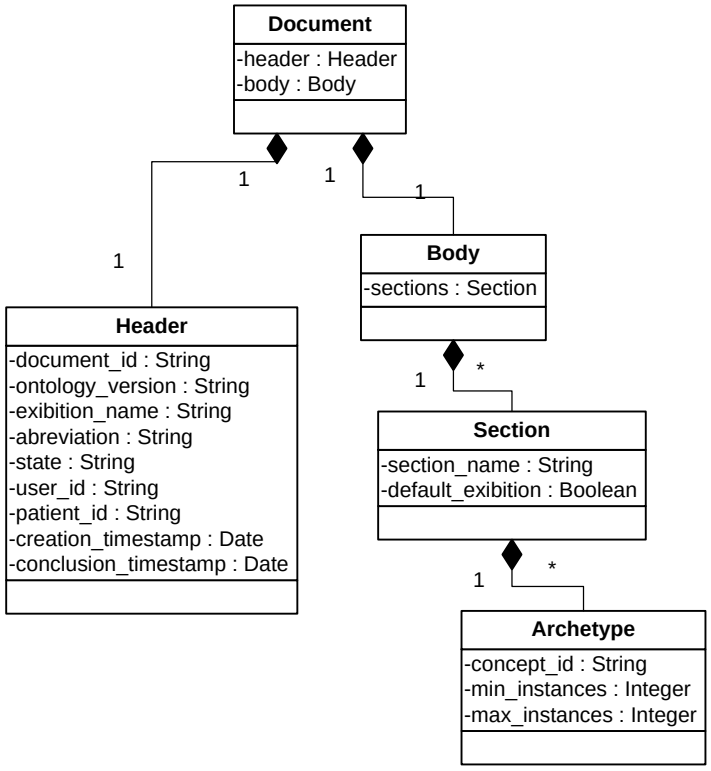


Figura 4: Diagrama de classes UML da estrutura dos documentos

Um documento no OpenCTI (conceito *Document*) é composto por um cabeçalho (*Header*), um corpo (*Body*) que é formado por uma ou diversas seções (*Section*). Cada um desses conceitos possui atributos que os descrevem. São nas seções que os arquétipos (conceito *Archetype*) estão localizados, e é nesse ponto que é realizada a ligação com os conceitos biomédicos modelados.

### 2.4.3 Persistência de dados no OpenCTI

Como apresentado anteriormente, o OpenCTI é um sistema de relativa complexidade. No que diz respeito à persistência dos dados, diversos requisitos devem ser cobertos para dar suporte às características do sistema. Existe a necessidade de arma-

zenar diversas informações que irão sofrer pouca ou nenhuma modificação na sua estrutura ao longo do tempo. Nessa categoria estão incluídas, por exemplo, dados relativos ao estoque de insumos, relações de leitos que o sistema irá gerir, entre outros. As tecnologias disponíveis para desenvolver essa categoria de problemas de persistência já são conhecidas e amplamente disseminadas. O OpenCTI trata dessa categoria de entidades com uma modelagem relacional tradicional dos dados (SILBERSCHATZ, 2010). Por outro lado, o conjunto de dados clínicos que compõem o RES é altamente dinâmico, volátil e esparso (JOHNSON, 1996). Avanços em pesquisas relacionadas à saúde fazem com que novas informações sejam coletadas e ao longo do tempo, os dados previamente modelados podem receber novos atributos ou talvez não sejam mais úteis para os usuários do RES. Desta forma, é necessário um mecanismo de persistência de dados capaz de se adaptar a esse dinamismo característico do ambiente de saúde.

Na modelagem relacional tradicional, os atributos de uma entidade do modelo relacional são representados como colunas de uma tabela do banco de dados físico. Primeiramente, devido ao dinamismo da estrutura dos dados clínicos, em um modelo relacional tradicional, seria preciso realizar refatoração de código do sistema frequentemente em decorrência dos ajustes no modelo de dados, refletindo diretamente em altos custos de manutenção em um sistema crítico como um RES típico.

Outro aspecto importante diz respeito à característica que os dados clínicos possuem de ser esparsos. Um documento de saúde pode possuir entre dezenas e centenas de informações que poderiam ser representadas, entretanto não existe uma obrigatoriedade em relação ao preenchimento dessas informações e, frequentemente, apenas uma pequena parcela desses dados é efetivamente utilizada. Ao utilizar uma tabela relacional tradicional para armazenar essas informações, teríamos um grande desperdício em termos de espaço de armazenamento, pois diversas colunas dos registros não irão armazenar dados reais do sistema. O diagrama exibido na Figura 5 apresenta o modelo ER desenvolvido para o OpenCTI (DUARTE, 2010). A modelagem referente aos dados clínicos do OpenCTI foi desenvolvida a com base na abordagem de persistência EAV (*Entity Atributte Value*) descrita no trabalho de DINU e colaboradores (2007).

Para tratar das consultas realizadas sobre a base de dados do OpenCTI, é necessário um módulo capaz de responder as solicitações de usuários, ou ferramentas associadas ao sistema com a presença de alguns parâmetros de busca pré-

definidos de tal forma que a persistência do OpenCTI poderá responder a solicitações de dados armazenados de acordo com alguns argumentos presentes na requisição. Os parâmetros que restringem a busca de dados disponíveis no OpenCTI são encontrados na Tabela 1.

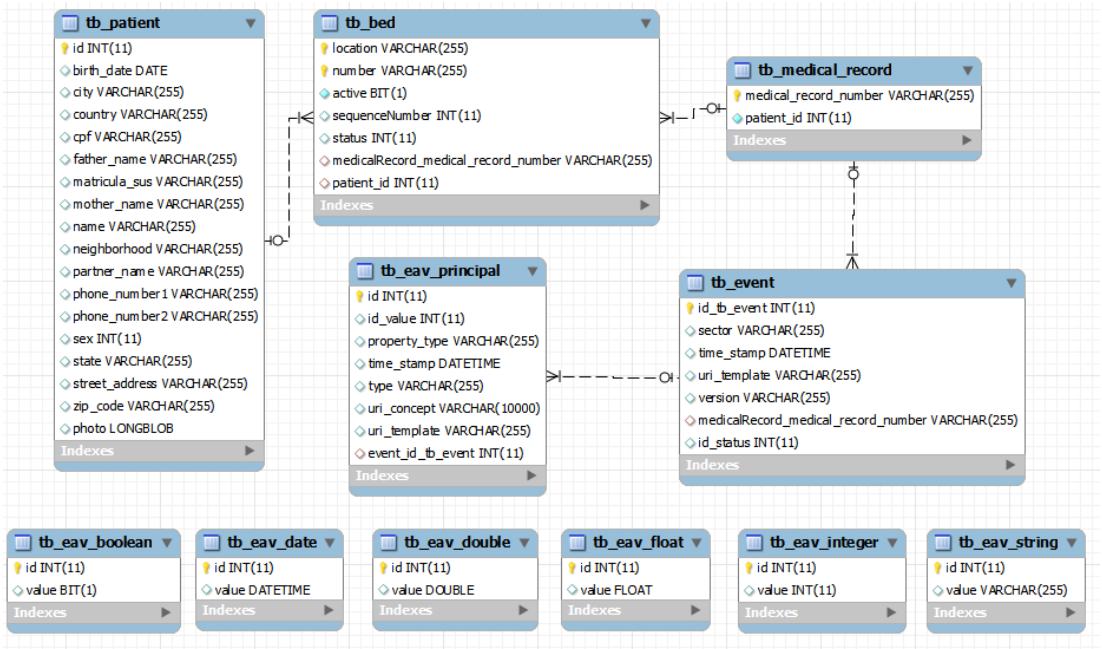


Figura 5: Modelo de Entidade Relacionamento do OpenCTI

Tabela 1: Parâmetros para busca de dados no OpenCTI

| **Parâmetro** | **Exemplos** |
|---|---|
| Conceitos biomédicos | Peso, altura, idade, frequência cardíaca, motivo de internação; |
| Documentos de saúde | Ficha de evolução médica, Ficha de evolução de enfermagem, Documento de óbito; |
| Setores do hospital | Setor de cardiologia, unidade de terapia intensiva cardiológica, ambulatório, pré-operatório; |
| Data inicial | 01/01/2011; |
| Data final | 31/01/2011; necessariamente maior que a data inicial; |

Por exemplo, um usuário poderia requisitar dados relacionados aos conceitos de frequência cardíaca, pressão arterial média, idade e índice APACHE (*Acute Physiology and Chronic Health Evaluation*) (KNAUS, 1985) dos pacientes atendidos na CTI pediátrica e CTI adulto do Hospital Lauro Wanderley (HULW). Dados

esses que estejam contidos em uma Ficha de Evolução Médica (FEM), armazenados entre os dias 01/05/2010 até 31/05/2010.

### 2.4.4 Geração dinâmica de interfaces

Com um modelo de dados adaptável às demandas dos usuários em nível de persistência e camada de domínio, torna-se necessário o desenvolvimento de uma camada de apresentação igualmente flexível e configurável. No OpenCTI, os diversos componentes de apresentação utilizados na interface com o usuário são escolhidos, configurados e instanciados em tempo de execução, oferecendo a flexibilidade necessária para compatibilidade com o restante do modelo de dados (DUARTE, 2011).

As ontologias de documentos e de conceitos biomédicos são interpretadas durante a geração dinâmica da interface para determinação e associação de componentes de interface disponíveis e que possam apresentar de forma eficiente o modelo de dados descrito nestas ontologias. É previsto ainda que esta metodologia seja utilizada para promover a visualização dos documentos de saúde utilizando diversos tipos de dispositivos, tais como *handhelds* ou *tablets*.

Com o arcabouço oferecido pela geração dinâmica de interface com usuários do OpenCTI, é habilitada uma integração e manutenção simplificada dos componentes de interface com uma arquitetura desenvolvida para o OpenCTI com intuito de promover o uso de agentes de CDS. A seguir é apresentada de forma sucinta esta arquitetura e seus objetivos.

### 2.4.5 MultiPersOn-CDS

Objetivando endereçar suporte à decisão clínica de forma genérica e adaptável para diversos cenários de uso, igualmente configurável, gerenciável e personalizável por uso de ontologias, o OpenCTI conta com framework intitulado MultiPersOn-CDS (*Framework* Multipropósito, Personalizável, baseado em Ontologia, utilizando agentes contextuais especializados para o apoio à decisão clínica) (PIZZOL, 2010). Tal framework oferece mecanismos para criação e gerenciamento de agentes de CDS que dispõem de sensores e atuadores para manipular dados dos conceitos de saúde existentes no OpenCTI, de acordo com objetivos e implementações específicas destes agentes. O modelo possibilita a criação de agentes de diversos tipos, variando sua complexidade e recursos necessários de acordo com a implementação específica uti-

lizada para descrever o comportamento do agente. A Figura 6, a seguir, apresenta a arquitetura desenvolvida para dar suporte ao framework MultiPersOn-CDS.

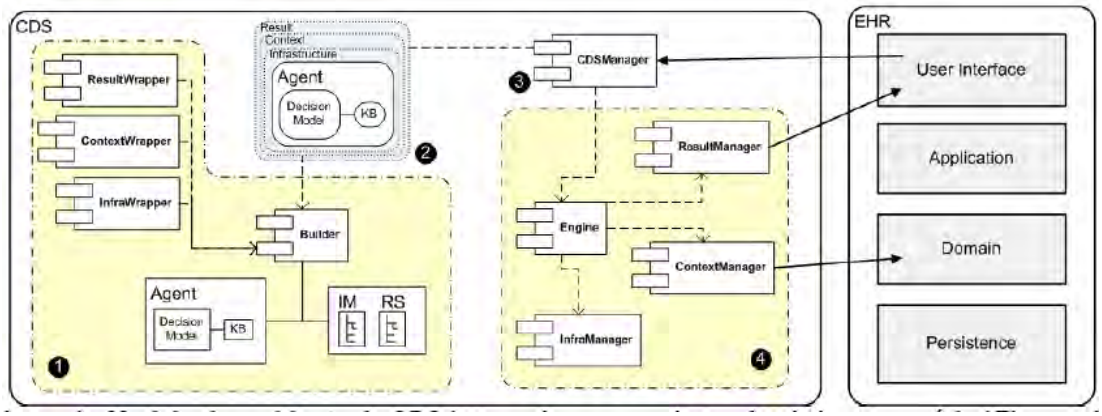


Figura 6: Arquitetura do MultiPersOn integrado à arquitetura em camadas de um RES, fonte: Pizzol (2010)

Na Figura 6, o item 1 agrupa componentes empregados para realizar a instanciação de novos agentes no ambiente de CDS. É realizada uma interpretação de arquivos que descrevem um novo agente a ser criado e, a partir disto, são associados: recursos de infraestrutura, (persistência e intercomunicação entre agentes) por meio do componente *InfraWrapper;* recursos do contexto de dados, para oferecer acesso aos conceitos de saúde do RES, através do componente *ContextWrapper* e; recursos para interação com o usuário, componente *ResultWrapper*.

Os agentes criados ficam em estado de espera, aguardando possíveis solicitações, ilustrado como o item 2 da Figura 6. Essa solicitação se dá por meio do componente *CDSManager,* que é encarregado de atuar como *listener* dos eventos lançados pela interface com o usuário, retirar o agente de CDS correspondente do estado de espera e repassar ao mesmo controle sobre os recursos necessários para sua atuação, itens agrupados de número 4. Uma vez encerrada a atuação do agente, o mesmo volta para estado de espera, encerrando um ciclo de execução.

Desta forma, os agentes de CDS utilizados possibilitam o uso de diversos recursos existentes no OpenCTI, além daqueles que são disponibilizados pelo ambiente de CDS, abstraindo questões de implementação e facilitando o desenvolvimento de novos agentes de suporte à decisão clínica.

Este ambiente de criação e execução de agentes de CDS é importante no contexto deste trabalho, pois o MultiPersOn-CDS será o principal cliente dos serviços disponibilizados pelo CollectMed a partir da sua API de consultas aos modelos selecionados, levando as informações coletadas pela ferramenta diretamente aos usuários por meio dos agentes de CDS criados.

## 2.5 Considerações finais

Ao final desse capítulo, é possível visualizar os princípios teóricos que guiam o desenvolvimento deste trabalho, desde as definições relacionadas aos RES, extração de conhecimento, até a forma como se entende o conceito de apoio à decisão, e como se aplica ao trabalho em questão. Além das questões de caráter teórico, apresentamos também o RES ao qual este trabalho está associado, o OpenCTI, com suas características fundamentais de persistência, modelo de dados, geração de interface e arquitetura de agentes de CDS. O estudo e familiarização com uma metodologia de extração de conhecimento e aprendizagem de máquina é importante para a definição de etapas de um processo que deverá ser seguido também pelo CollectMed, apresentado e desenvolvido nos capítulos subsequentes.

Capítulo

# 3
# CollectMed

"O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo. Mesmo não atingindo o alvo, quem busca e vence obstáculos, no mínimo fará coisas admiráveis."

*José de Alencar*

O objetivo deste capítulo é apresentar os requisitos gerais da ferramenta CollectMed e a solução do problema apresentado nos capítulos anteriores, assim como dos requisitos adicionais levantados neste capítulo.

Buscando alcançar os seus objetivos, o CollectMed integra um sistema de suporte à decisão para o OpenCTI. Mais especificamente, o CollectMed age como um criador de sistema de suporte à decisão, tal qual a definição apresentada na seção 2.2.2. Fazendo-se presente durante a concepção da solução, os sistemas de apoio à decisão específicos (SDSS), manipulados no CollectMed e denominados apenas como "modelos". Neste sentido, foi desenvolvido um processo e arquitetura que oferece suporte à criação, manutenção e consultas sobre estes modelos. As subseções que se seguem apresentam os requisitos levantados para o CollectMed; o processo de criação e gerenciamento dos modelos criados; e a arquitetura desenvolvida para a solução.

## 3.1 Requisitos

Com o arcabouço oferecido pelo OpenCTI, em relação à modelagem e persistência dos dados clínicos, é possível realizar diversas atividades que objetivam agregar mais

valor à aplicação. Ou seja, existe condição para buscar novas características que melhorem o sistema em termos de usabilidade, percepção de inteligência e reatividade para o usuário. Como mecanismo para oferecer tais características, pode-se utilizar da extração de informação baseada em dados provenientes do próprio sistema ou de outros semelhantes.

Nas seções iniciais, 1.1 e 1.2, apresentamos o problema em linhas gerais. Como citado, o objetivo é que os usuários da ferramenta possam promover o reuso de conhecimento clínico, presente nas bases de dados do RES, mesmo com pouco conhecimento em extração de conhecimento. O desenvolvimento desse objetivo leva a requisitos que guiam o desenvolvimento do CollectMed, a saber:

a) Manutenção simplificada de modelos;

b) Baixo acoplamento com o modelo de dados clínicos; por fim,

c) Disponibilidade para consultas ao modelo.

### 3.1.1 Manutenção simplificada de modelos

Uma grande dificuldade ao utilizar ferramentas de extração de conhecimento e aprendizagem de máquina é o elevado nível de especialidade necessário para sua manipulação e as técnicas associadas. Entendemos que o processo de extração de conhecimento deve ser simples e intuitivo para de fato tornar-se efetivo, de outra forma, caso seja necessário muito esforço e tempo para executar o processo de extração de conhecimento o sistema poderia encontrar resistência dos próprios usuários na sua utilização. Deve ser possível aos usuários, a partir de breve treinamento no uso da ferramenta desenvolvida, criar modelos capazes de identificar padrões no modelo de dados do RES e disponibilizá-los para uso em métodos e/ou agentes de CDS. Reduzindo a necessidade de um profissional altamente especializado que seja dedicado à criação e elaboração desses modelos.

Uma vez criados com auxílio do CollectMed, os modelos devem ser passíveis de manutenção. Para tanto, a ferramenta deve prover os mecanismos necessários para que os modelos possam ser ajustados, recriados, ter sua atuação no sistema suspensa, ou mesmo excluída, de acordo com as necessidades dos usuários do sistema e por consequência implantando requisitos de sistemas de apoio à decisão clínica, como apresentado na seção 2.2.

### 3.1.2 Baixo acoplamento com o modelo de dados clínicos

Adicionalmente à dificuldade encontrada para realizar etapas referentes à extração de conhecimento propriamente dito, o CollectMed propõe-se a manipular dados clínicos, que, como foi mencionado em seções anteriores, são dinâmicos e voláteis. Greenes (2007b) aponta que sistemas de apoio à decisão devem ter um baixo acoplamento com o modelo de dados clínicos utilizado, de outra forma, seria necessário aumentar o esforço para manter a ferramenta em concordância com as alterações com o modelo de dados clínicos dinâmicos do RES, e por este motivo, adota-se o baixo acoplamento em nível de modelo de dados como requisito do CollectMed.

### 3.1.3 Disponibilidade para consultas ao modelo

Apenas a criação simplificada dos modelos a ser promovida pelo CollectMed não é suficiente para possibilitar amplo reuso do conhecimento presente nos RES. A existência dos modelos permite que sejam utilizados de diversas maneiras, por exemplo, seria possível desenvolver agentes de CDS específicos que manipulam diretamente informações que esses modelos representam. No entanto, o trabalho relacionado à programação dos agentes de CDS pode ser reduzido substancialmente ao oferecer, por exemplo, acesso aos modelos gerenciados pelo CollectMed por meio de uma API de execução de requisições.

## 3.2 Processo de criação de ferramentas de apoio à decisão clínica

Visando aumentar a produtividade e eficiência na criação e manutenção dos modelos a partir do CollectMed e baseado na definição de processos com etapas e tarefas de mineração de dados, assim como apresentado na seção 2.3.2, foi delineado para o CollectMed um processo para criação de ferramentas de apoio à decisão clínica com a fixação de usuários, etapas e atividades.

O processo de criação de ferramentas de apoio à decisão no contexto do CollectMed possui três etapas, a saber: Análise, Desenvolvimento e Implantação. Estas etapas são executadas em sequência e envolvem as categorias de usuários definidos na interação com o CollectMed, apresentados a seguir.

### 3.2.1 Usuários do CollectMed

Durante a concepção do CollectMed, foi levada em consideração a participação de usuários em três diferentes níveis de interação com a ferramenta, a saber: gerente de suporte à decisão, desenvolvedores de métodos de apoio à decisão clínica e usuários do RES OpenCTI.

#### 3.2.1.1 Gerentes de suporte à decisão

O primeiro tipo de usuário aqui descrito executa o papel *Gerente de suporte à decisão*. Suas ações dizem respeito à criação, gerenciamento e manutenção dos SDSS criados. A Figura 7 mostra um diagrama de caso de uso com as atividades para o ator *Gerente de suporte à decisão*, são elas, "Criar Modelo", "Visualizar Modelo", "Atualizar Modelo" e "Excluir Modelo".

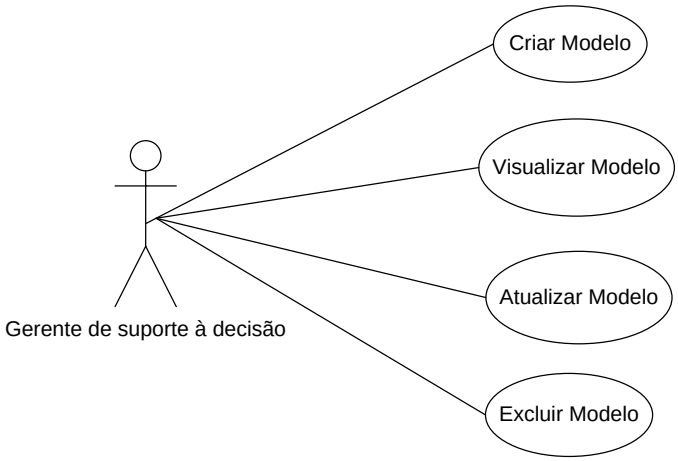


Figura 7: Diagrama de caso de uso para o ator Gerente de apoio à decisão

Além dos casos de uso apresentados no diagrama, é também responsabilidade deste usuário realizar ações de planejamento para garantir que os modelos estão sendo utilizados e, principalmente, que os mesmos são de fato úteis aos usuários dos demais níveis de interação. No capítulo seguinte encontram-se descritas em mais detalhes as funcionalidades do sistema que são executadas pelos gerentes de apoio à decisão.

Para executar o papel de gerente de suporte à decisão, os usuários poderiam, preferencialmente, fazer parte do corpo de profissionais de saúde que utilizam o RES em suas atividades regulares (médicos, enfermeiros, fisioterapeutas, nutricio-

nistas) e, desta forma, ser capaz de identificar problemas que possam ser auxiliados através de métodos de apoio à decisão automatizada. Como a criação de modelos é composta de atividades de natureza multidisciplinar, é importante que os gerentes de suporte à decisão obtenham familiaridade também com o uso de ferramentas dedicadas à mineração de dados, que em associação ao conhecimento dos documentos e conceitos de saúde aplicados no RES este usuário será capaz de extrair melhores resultados dos modelos criados utilizando o CollectMed.

#### 3.2.1.2 Desenvolvedores de métodos de apoio à decisão clínica

Atuando em um maior nível de abstração em relação à ferramenta CollectMed, quando comparado aos gerentes de suporte à decisão, os desenvolvedores de métodos de apoio à decisão clínica são os responsáveis por manipular o resultado do trabalho de criação de modelo através de chamadas à API de serviço oferecida pelo CollectMed. Trabalhando em conjunto com os gerentes de suporte à decisão e usufruindo do framework MultiPersOn do OpenCTI, os desenvolvedores de métodos de apoio à decisão clínica usarão os modelos criados e colocarão resultados obtidos com o uso dos modelos à disposição dos usuários do RES, que correspondem aos usuários finais do esforço e processamento realizado pelo CollectMed.

#### 3.2.1.3 Usuários do RES OpenCTI

O terceiro tipo de usuários relacionados ao CollectMed são os que utilizam o RES OpenCTI e são beneficiados com o uso dos modelos criados pelos gerentes de suporte à decisão. Por não atuarem diretamente sobre a construção dos modelos ou sua manutenção através da ferramenta criada ou no desenvolvimento dos métodos de apoio à decisão, estes usuários podem sequer tomar conhecimento da infraestrutura, existente tanto no RES quanto no CollectMed, que oferece suporte à tomada de decisão clínica através do reuso de conhecimento clínico por meio do CollectMed ou outras ferramentas e *frameworks* que o OpenCTI utilize neste sentido.

### 3.2.2 Etapas e Atividades do processo de criação

Baseado em processos utilizados para realização de KDD (*Knowledge Discovery in Databases*) (Fayyad, Piateskyp-Shapiro & Smyth, 1996) foi desenvolvido um processo composto por etapas e atividades que são descritas a seguir. Este processo atua como um delineador geral, e deve ser especializado de acordo com necessidades específicas de organizações que o utilize. O desenvolvimento do CollectMed dá suporte

para execução das atividades existentes em cada uma das etapas. Diversas atividades apresentadas são executadas pelos usuários do CollectMed em conjunto, em virtude da natureza multidisciplinar do processo.

#### 3.2.2.1 Etapa de Análise

Durante esta etapa, os usuários concentram-se em identificar e obter maior entendimento do problema que necessita apoio à decisão e obtenção de informações que auxiliarão na resolução do mesmo. Na Figura 8, são apresentadas as atividades relacionadas a esta etapa do processo de criação de métodos de apoio à decisão.

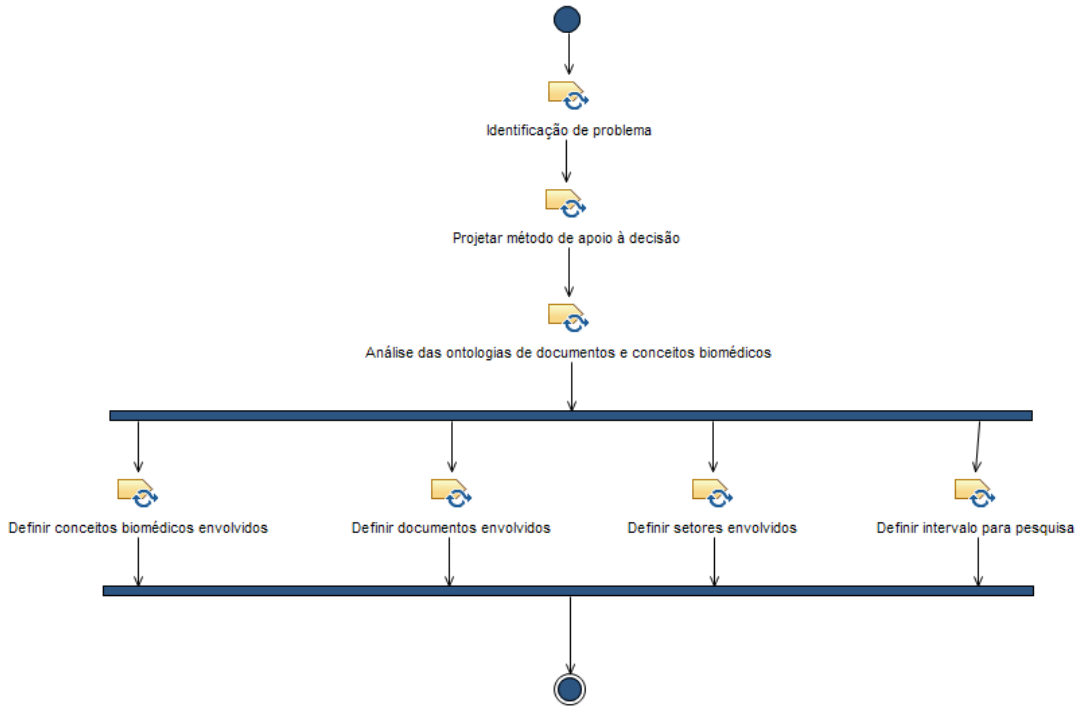


Figura 8: Atividades da etapa de Análise

##### 3.2.2.1.1 Identificação do problema

Em termos gerais, qualquer usuário do sistema pode identificar um problema relacionado ao conteúdo dos documentos que poderiam ser auxiliados por meio de métodos de apoio à decisão clínica. Esta atividade objetiva solicitar ao gerente de apoio à decisão clínica que o problema levantado seja analisado e que seja realizado um levantamento da viabilidade da criação de um mecanismo de CDS equivalente. De acordo com determinações do gerente de suporte à decisão, esta atividade pode ser

formalizada por uma solicitação formal, para posterior acompanhamento e documentação da solicitação.

#### 3.2.2.1.2 Projetar método de apoio à decisão

Juntamente com desenvolvedores de métodos de CDS, o gerente de apoio à decisão deve projetar, juntamente com os desenvolvedores, o funcionamento do método de apoio à decisão a ser criado. Nesta atividade, podem ser definidos diversos recursos a ser aplicados no desenvolvimento do mecanismo, a equipe de trabalho, prazos para implantação, objetivos gerais e específicos do mecanismo de CDS, entre outras atividades de projeto e planejamento.

Como linha geral, é preferível que o problema identificado seja subdividido em subproblemas menores sempre que possível, assim como apresentado na seção 2.2.1, facilitando a sua composição e possibilitando a obtenção de melhores resultados com o método de apoio à decisão criado.

#### 3.2.2.1.3 Análise das ontologias de documentos e conceitos biomédicos

Dando prosseguimento a definição do mecanismo CDS, é necessário analisar as ontologias de documentos e conceitos biomédicos criadas e mantidas no OpenCTI para definir quais conceitos podem ser utilizados para a resolução do problema. Uma análise detalhada desta atividade é vital para a construção de modelo de qualidade, pois é a base para o restante do processo.

Decorrente desta análise sobre as ontologias, o gerente de suporte à decisão poderá definir os conceitos biomédicos e documentos de saúde que irão compor os dados, além de setores e intervalos de abrangência das pesquisas, representados na Figura 8 pelas atividades: “Definir conceitos biomédicos envolvidos”, “Definir documentos envolvidos”, “Definir setores envolvidos” e “Definir intervalo para pesquisa”.

### 3.2.2.2 Etapa de desenvolvimento

Com a finalização da etapa de análise, deve-se ter conhecimento do problema que se deseja atingir com o desenvolvimento de um método de apoio à decisão clínica e as informações que condicionam sua solução. As etapas seguintes dão continuidade ao desenvolvimento através da pesquisa, pré-processamento, treinamento e avaliação do modelo construído. Na Figura 9, são ilustradas as atividades desta etapa.

#### 3.2.2.2.1 Pesquisa dos dados clínicos

O CollectMed tem por objetivo promover reuso de informações clínicas contidas nas bases de dados do sistema OpenCTI. Com a posse de informações obtidas nas atividades anteriores, é possível uma seleção adequada destas informações. O gerente de apoio à decisão clínica deve fazer tal seleção utilizando os recursos disponibilizados pelo OpenCTI juntamente aos do CollectMed, obtendo ao final da atividade uma base de dados relativa aos conceitos de saúde selecionados, encontrados no documentos de saúde persistidos.

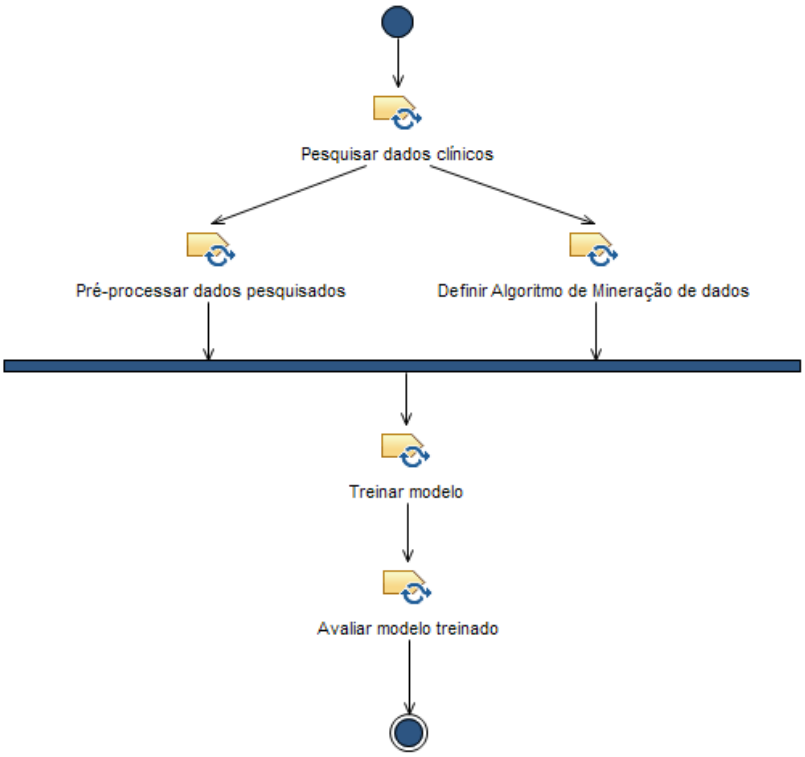


Figura 9: Atividades da etapa de desenvolvimento

Em geral, a quantidade de registros indicada para a composição da base de dados de treinamento deve ser de, pelo menos, dez vezes maior que a quantidade de atributos (conceitos de saúde) (GREENES, 2007c). Exemplificando, quando a pesquisa utilizar quatro conceitos de saúde, a base de dados deve conter pelo menos quarenta registros para que os resultados possam ser satisfatórios, e no caso da pesquisa conter uma quantidade de conceitos dependentes igual a dez, a base de dados correspondente a pesquisa deve ser de no mínimo cem registros.

#### 3.2.2.2.2 Pré-processar dados pesquisados

Mesmo que apenas uma pequena fração da grande massa de dados persistida no RES seja resultante da atividade anterior, não é possível garantir que os dados apresentem-se de forma íntegra e adequados para a composição de mecanismos automatizados de extração de informação e aprendizagem de máquina. É, portanto, necessária a execução de atividades de pré-processamento sobre o conjunto de dados.

A fim de realizar o pré-processamento, algumas abordagens e técnicas podem ser utilizadas neste objetivo, assim como para simplificar as etapas consecutivas do processo de extração de conhecimento e aprendizagem de máquina.

Quanto menos atributos forem conhecidos para um determinado registro, menor será a precisão dos algoritmos de *data mining* durante a criação dos modelos. Caso muitos dos registros do *data set* original possuírem uma baixa taxa de população (relação entre a quantidade de atributos conhecidos e a quantidade total de atributos), provavelmente o modelo não será preciso o suficiente para alcançar bons resultados. Dessa forma, a etapa de pré-processamento poderia eliminar os registros que não alcancem uma taxa de população mínima exigida, e manter apenas aqueles que carreguem consigo mais informações, elevando a qualidade dos modelos obtidos a partir do CollectMed.

No *data set* obtido na pesquisa, alguns dos registros podem incluir valores muito elevados ou muito baixos para alguns dos seus atributos, destacando um registro frente ao restante dos valores encontrados para o mesmo atributo, no restante dos registros do *data set*. A presença desses valores podem influenciar o resultado dos modelos, diminuindo a capacidade de identificar padrões mais discretos. Portanto, pode ser necessário excluir os registros que possuam valores extremos, quando eles ocorrerem em pequeno número (caracterizando exceções), para oferecer uma maior uniformidade ao modelo. A aplicação desse método deve ser utilizado com cautela, pois esses valores podem representar padrões importantes, e ao excluí-los, eles não seriam levados em consideração para a criação dos modelos de *data mining*.

Em diversas ocasiões, dados que representam as mesmas informações serão armazenados com diferentes unidades ou escalas de medidas. Torna-se necessário para estes casos, realizar transformação dos dados, mantendo uma unidade de me-

dida uniforme para as instâncias presentes no conjunto de dados e consequentemente evitando interpretações equivocadas dos padrões por parte dos modelos.

#### 3.2.2.2.3 Definir algoritmo de mineração de dados

Em paralelo ao pré-processamento dos dados clínicos é necessária a definição de um algoritmo para a modelagem das informações contidas no conjunto de dados selecionado. A avaliação de um algoritmo adequado é de suma importância para a qualidade dos resultados obtidos na etapa seguinte.

#### 3.2.2.2.4 Treinar modelo

Com posse de uma base de dados de treinamento e um algoritmo definido, é possível dar início à atividade de treinamento do modelo. Através do qual, o novo modelo será capaz de responder a solicitações dos usuários a respeito das informações ali presentes.

#### 3.2.2.2.5 Avaliar modelo treinado

Como medida cautelar, a atividade de avaliação deve ser realizada, por meio de testes de validação para garantir que os modelos selecionados consigam oferecer sugestões de qualidade para o problema proposto. Caso isto não aconteça, é indicado que o fluxo de trabalho seja desviado para as atividades iniciais, revisando cada etapa realizada, em busca de melhoras sobre a seleção dos dados, pré-processamento ou mesmo escolha do algoritmo de mineração de dados.

### 3.2.2.3 Etapa de Implantação

Última etapa do processo de criação de modelos de apoio à decisão, a etapa de implantação diz respeito ao desenvolvimento, integração e testes dos métodos de CDS que utilizam os serviços proporcionados pelo CollectMed. As atividades desta etapa são ilustradas na Figura 10.

#### 3.2.2.3.1 Desenvolver Agente de CDS

Como foi visto, as etapas executadas anteriormente tratam, em sua maioria, da construção da solução de apoio à decisão clínica por parte do CollectMed. Entretanto, sem o desenvolvimento de um agente de CDS correspondente utilizando o *frame-*

*work* MultiPersOn, os usuários do RES não possuem acesso aos serviços oferecidos pelo CollectMed.

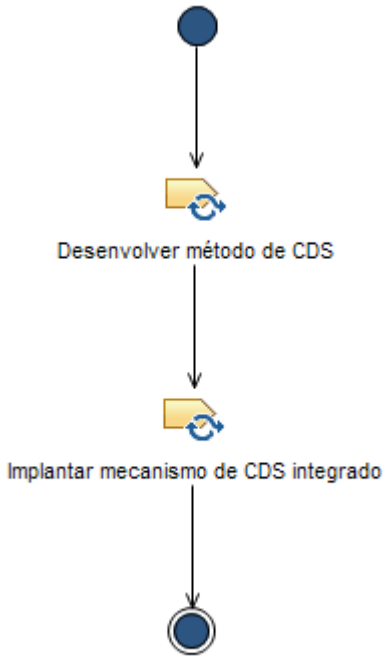


Figura 10: Atividades da etapa de implantação

Esta atividade, portanto, dedica-se ao desenvolvimento ou configuração de agentes de apoio à decisão que interceptem os dados dos usuários do OpenCTI, consultem o CollectMed, e em seguida retornem aos usuários as sugestões baseadas em informações clínicas presentes nas bases de dados do RES.

#### 3.2.2.3.2 Implantar mecanismo de CDS integrado

A última atividade do processo de criação de modelos de apoio à decisão é realizada pelo gerente de apoio à decisão para que o agente criado na atividade anterior, depois de realizadas as atividades de validação e testes, seja aplicado pelos usuários do RES no seu dia-dia, proporcionando o auxílio à decisão clínica à beira do leito.

## 3.3 Arquitetura da solução

Para dar suporte às etapas e atividades descritas no processo de criação de ferramentas de apoio à decisão, foi desenvolvida a arquitetura do CollectMed, assim como sua ferramenta de administração, o *CollectMed Admin*. A Figura 11 apresenta

um diagrama simplificado da arquitetura do CollectMed e como ele é integrado ao OpenCTI e ao módulo de gerência de CDS.

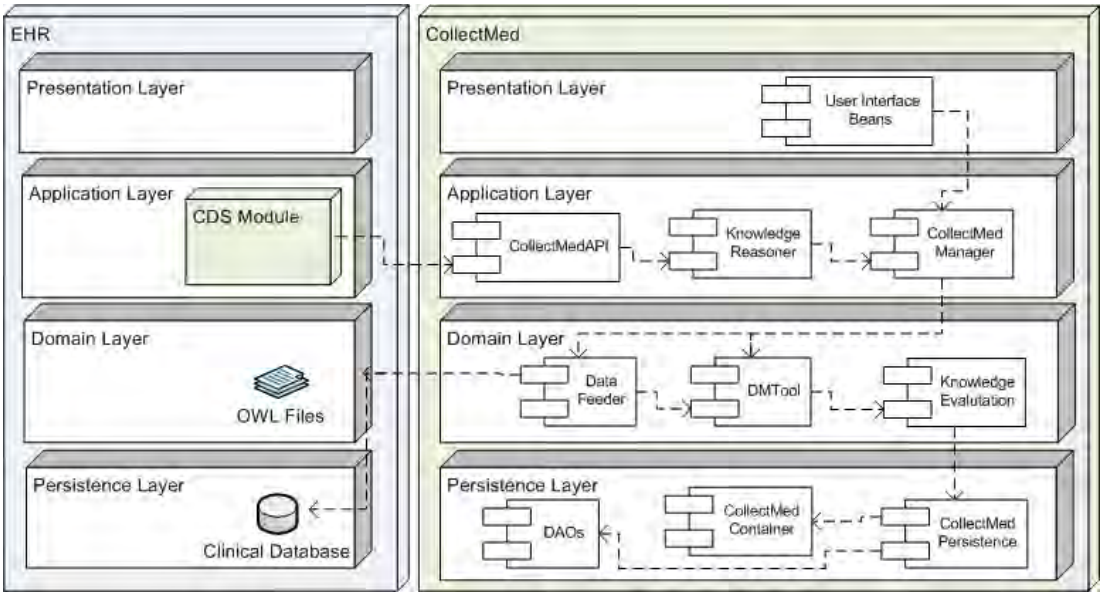


Figura 11: Arquitetura de camadas do OpenCTI e CollectMed com principais componentes.

À esquerda da figura está representado o RES OpenCTI, com sua arquitetura dividida em quatro camadas, persistência, domínio, aplicação e apresentação. Na camada de persistência, estão armazenados os dados clínicos de acordo com a metodologia apresentada na seção 2.4.3. Na camada de domínio estão componentes utilizados para a manipulação destes dados em estado transiente, instanciação de novos documentos e conceitos baseados em ontologias (padrão OWL), assim como apresentado na seção 2.4.2. Já na camada de aplicação, encontram-se principalmente as regras de negócio desenvolvidas para a aplicação, determinando como os dados da camada de domínio serão acessados, incluindo, portanto, o módulo de CDS. Por fim, temos a camada de apresentação do OpenCTI, a qual é delegada responsabilidade de interpretar as ontologias que descrevem os documentos e conceitos, criar páginas e formulários correspondentes a estes documentos. O detalhamento da arquitetura do OpenCTI em seus componentes integrantes não é relevante para o contexto deste trabalho, desta forma, optou-se por não abordá-los em detalhes nesta descrição.

Do lado direito da Figura 11 temos a arquitetura do CollectMed dividida também em quatro camadas, com os seus componentes apresentados. A listagem que segue abaixo descreve os componentes arquiteturais utilizados no desenvolvi-

mento do CollectMed, organizados de acordo com a camada aos quais se encontram na arquitetura.

Da camada de persistência do CollectMed, destacamos dois componentes: *CollectMedContainer*, utilizado para acessar os modelos criados e persistidos, importante para recuperação do estado dos modelos pré-existentes quando o sistema é iniciado; e os componentes *DAOs*, que reúnem classes e padrões responsáveis pela persistência das metainformações relacionadas aos modelos criados com no CollectMed.

Na camada de domínio, os principais componentes são: *DataFeeder*, responsável por buscar dados do RES OpenCTI, sejam metainformações necessárias para descrever os modelos ou dados clínicos utilizados para realizar o treinamento e avaliação dos modelos criados; componente *DMTool*, executa atividades de treinamento de novos modelos, assim como instanciação de modelos existentes; e o componente *KnowledgeEvaluation*, sucede a etapa de treinamento ao realizar avaliação dos modelos recém criados. Estas atividades são realizadas com interferência e suporte do usuário, sendo portanto uma atividade semi-automatizada.

Participando da camada de aplicação do CollectMed, o componente *CollectMedManager* atua como elemento central na arquitetura do CollectMed, gerenciando o funcionamento do restante dos componentes, através da ligação entre os mesmos. A execução de consultas de apoio à decisão utilizando o CollectMed se dá utilizando a API disponibilizada pelo componente *CollectMedAPI*, e executada de fato no componente *KnowledgeReasoner*.

A camada de apresentação é dedicada aos usuários que interagem com o CollectMed através da sua ferramenta de administração (CollectMed Admin). Nesta camada se fazem presentes as páginas de administração, criação e manutenção dos modelos com seus *beans* de controle.

Nas seções seguintes, são apresentados com mais detalhes o funcionamento da ferramenta e seus componentes, partindo do seu elemento fundamental (*CollectMedDecisionSupport*), seleção dos dados clínicos, pré-processamento dos dados, treinamento dos modelos, persistência, e execução de consultas sobre os modelos criados.