GENÔMICA TRANSLACIONAL: INTEGRANDO DADOS CLÍNICOS E BIOMOLECULARES

Newton Shydeo Brandão Miyoshi

DISSERTAÇÃO APRESENTADA

AO

PROGRAMA INTERUNIDADES EM BIOINFORMÁTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA

OBTENÇÃO DO GRAU DE MESTRE

EM

CIÊNCIAS

Área de Concentração: Bioinformática

Orientador: Prof. Dr. Joaquim Cezar Felipe

- Ribeirão Preto, 06 de fevereiro de 2013 -

GENÔMICA TRANSLACIONAL: INTEGRANDO DADOS CLÍNICOS E BIOMOLECULARES

Este exemplar corresponde à versão original da dissertação de Mestrado devidamente corrigida e defendida pelo aluno Newton Shydeo Brandão Miyoshi e aprovada pela comissão julgadora.

Ribeirão Preto, 06 de fevereiro de 2013.

Banca Examinadora:

- Prof. Dr. Joaquim Cezar Felipe (orientador) - FFCLRP-USP
- Prof. Dr. João Eduardo Ferreira - IME-USP
- Prof. Dr. Wilson Araújo da Silva Junior - FMRP-USP

"Nossa maior fraqueza está em desistir.

O caminho mais certo para vencer é tentar mais uma vez"

Thomas Edison

Aos meus pais, Newton e Suely

À minha irmã, Stephanie

À minha namorada, Juliana

Por todo amor, dedicação e paciência

AGRADECIMENTOS

Agradeço a Deus por mais esta realização e pelas oportunidades e desafios que me fazem progredir sempre.

Aos meus pais, Newton e Suely, e minha irmã Stephanie, por todo apoio, amor e sacrifícios que fazem para tornar meus sonhos realidade. Agradeço também a todos meus familiares que compartilham e comemoram comigo mais esta etapa de minha vida, me apoiando sempre.

À minha namorada Juliana, pelo carinho, ternura, companheirismo e incentivo de todos os dias, sem os quais, com certeza, não teria realizado este trabalho. Também por contribuir essencialmente durante a revisão da dissertação.

Aos colegas e amigos de laboratório, Marlon, Gisele, Yuri, Rômulo, Lulu, Amanda e Ana Lívia, que compartilharam comigo esses quase 3 anos de mestrado.

Ao meu orientador, Joaquim, por mais esta oportunidade de trabalho, pela confiança, paciência e por todos os conselhos dados não só relativamente ao projeto mas quanto a carreira acadêmica.

Ao professor Wilson e a equipe do BiT, em especial ao Daniel, por contribuírem de forma essencial para elaboração do projeto, pelo disponibilização dos dados e pela ajuda constante.

Á Patrícia por ser sempre muito atenciosa e paciente nos auxiliando, sempre com muita boa vontade, nas questões administrativas e burocráticas.

Ao Programa Interunidades em Bioinformática por todo apoio durante o mestrado.

A CAPES pelo apoio financeiro.

SUMÁRIO

LISTA DE FIGURAS

RESUMO

A utilização do conhecimento científico para promoção da saúde humana é o principal objetivo da ciência translacional. Para que isto seja possível, faz-se necessário o desenvolvimento de métodos computacionais capazes de lidar com o grande volume e com a heterogeneidade da informação gerada no caminho entre a bancada e a prática clínica. Uma barreira computacional a ser vencida é o gerenciamento e a integração dos dados clínicos, sócio-demográficos e biológicos. Neste esforço, as ontologias desempenham um papel essencial, por serem um poderoso artefato para representação do conhecimento. Ferramentas para gerenciamento e armazenamento de dados clínicos na área da ciência translacional que têm sido desenvolvidas, via de regra falham por não permitir a representação de dados biológicos ou por não oferecer uma integração com as ferramentas de bioinformática. Na área da genômica existem diversos modelos de bancos de dados biológicos (tais como AceDB e Ensembl), os quais servem de base para a construção de ferramentas computacionais para análise genômica de uma forma independente do organismo de estudo. Chado é um modelo de banco de dados biológicos orientado a ontologias, que tem ganhado popularidade devido a sua robustez e flexibilidade, enquanto plataforma genérica para dados biomoleculares. Porém, tanto Chado quanto os outros modelos de banco de dados biológicos não estão preparados para representar a informação clínica de pacientes. Este projeto de mestrado propõe a implementação e validação prática de um *framework* para integração de dados, com o objetivo de auxiliar a pesquisa translacional integrando dados biomoleculares provenientes das diferentes tecnologias "*omics*" com dados clínicos e sócio-demográficos de pacientes. A instanciação deste *framework* resultou em uma ferramenta denominada IPTrans (*Integrative Platform for Translational Research*), que tem o Chado como modelo de dados genômicos e uma ontologia como referência. Chado foi estendido para permitir a representação da informação clínica por meio de um novo Módulo Clínico, que utiliza a estrutura de dados entidade-atributo-valor. Foi desenvolvido um *pipeline* para migração de dados de fontes heterogêneas de informação para o banco de dados integrado. O *framework* foi validado com dados clínicos provenientes de um Hospital Escola e de um banco de dados biomoleculares para pesquisa de pacientes com câncer de cabeça e pescoço, assim como informações de experimentos de *microarray* realizados para estes pacientes. Os principais requisitos almejados para o *framework* foram flexibilidade, robustez e generalidade. A validação realizada mostrou que o sistema proposto satisfaz as premissas, levando à integração necessária para a realização de análises e comparações dos dados.

Palavras-chave: Bancos de Dados Biológicos, Integração de Dados, Pesquisa Translacional, Ontologias

ABSTRACT

The use of scientific knowledge to promote human health is the main goal of translational science. To make this possible, it is necessary to develop computational methods capable of dealing with the large volume and heterogeneity of information generated on the road between bench and clinical practice. A computational barrier to be overcome is the management and integration of clinical, biological and socio-demographics data. In this effort, ontologies play a crucial role, being a powerful artifact for knowledge representation. Tools for managing and storing clinical data in the area of translational science that have been developed, usually fail due to the lack on representing biological data or not offering integration with bioinformatics tools. In the field of genomics there are many different biological databases (such as AceDB and Ensembl), which are the basis for the construction of computational tools for genomic analysis in an organism independent way. Chado is a ontology-oriented biological database model which has gained popularity due to its robustness and flexibility, as a generic platform for biomolecular data. However, both Chado as other models of biological databases are not prepared to represent the clinical information of patients. This project consists in the proposal, implementation and validation of a practical framework for data integration, aiming to help translational research integrating data coming from different "omics" technologies with clinical and socio-demographic characteristics of patients. The instantiation of the designed framework resulted in a computational tool called IPTrans (Integrative Platform for Translational Research), which has Chado as template for genomic data and uses an ontology reference. Chado was extended to allow the representation of clinical information through a new Clinical Module, which uses the data structure entity-attribute-value. We developed a pipeline for migrating data from heterogeneous sources of information for the integrated database. The framework was validated with clinical data from a School Hospital and a database for biomolecular research of patients with head and neck cancer. The main requirements were targeted for the framework flexibility, robustness and generality. The validation showed that the proposed system satisfies the assumptions leading to integration required for the analysis and comparisons of data.

Key-words: Biological Databases, Data Integration, Translational Research, Ontologies

1. INTRODUÇÃO

1.1. CONTEXTUALIZAÇÃO

A medicina translacional pode ser definida como a aplicação dos resultados das pesquisas científicas, especialmente aqueles provenientes das tecnologias "*omics*" na melhoria dos processos de saúde e doença (WOOLF, 2008). Esta nova área de pesquisa busca reduzir a distância entre a bancada e a prática clínica, num desafio com muitas barreiras a serem vencidas, sendo que uma das mais difíceis e importantes está relacionada à natureza da informação.

A natureza dos dados clínicos é diferente da natureza dos dados moleculares, embora eles estejam intimamente relacionados. Dados clínicos podem variar desde informações gerais tais como idade, peso, altura, antecedentes familiares até informações específicas de uma especialidade médica, como o estadiamento de um tumor, incluindo medicamentos e exames laboratoriais. Dados biomoleculares são aqueles gerados pelas tecnologias de biologia molecular tais como *microarray*, sequenciamento de DNA e os dados de sequenciadores de nova geração.

Quando se investigam mecanismos complexos responsáveis pelo surgimento dos processos patológicos, é necessária uma análise global considerando os diferentes níveis de informação. Para tornar essa análise uma realidade, dois grandes aspectos, em relação aos dados, devem ser bem definidos e melhorados: armazenamento e análise. Nesse contexto, é necessária também a definição de uma plataforma computacional e um modelo de dados capaz de armazenar, representar e integrar a informação clínica e a informação biomolecular de forma consistente. A partir de um modelo formal e bem estruturado é possível projetar métodos consistentes de análises computacionais.

Na área de pesquisa translacional existem poucas plataformas computacionais para armazenamento e recuperação de dados clínicos. Slim-Prim (*Scientific Laboratory Information Management – Patient-care Research Information Management*) é um sistema de dados integrado para coletar, arquivar e distribuir dados de pesquisas

básicas e clínicas. O Slim-Prim é mantido pela Universidade do Tenessi e possui uma versão gratuita e de código livre, denominada PRIME (*Protected Research Information Management Environment*) (VIANGTEERAVAT et al., 2009). Embora os autores de Slim-Prim e PRIME afirmem que estes sistemas permitem o gerenciamento de dados de *microarray*, de sequências de DNA e outros tipos de dados moleculares, eles não fornecem integração com nenhuma ferramenta de bioinformática. Esses tipos de dados são tratados como um tipo de dado genérico em um ambiente de gerenciamento de conteúdo.

STRIDE **(***Stanford Translational Research Integrated Database Environment*), desenvolvido na Universidade de Stanford, é uma plataforma computacional baseada em padrões para apoiar a pesquisa clínica e translacional (LOWE et al., 2009). Esta plataforma é constituída por três componentes principais: (i) um *data warehouse* clínico, baseado no HL7 RIM (*Health Level Seven - Reference Information Model*); (ii) um modelo semântico baseado em ontologias tais como SNOMED, ICD, RxNorm e (iii) um *framework* para desenvolver aplicativos de gerenciamento de pesquisa. Atualmente os autores não planejam disponibilizar o STRIDE para ser implementado fora de Stanford.

Um dos Centros Nacionais para Computação Biomédica (NCBC – *National Center for Biomedical Computing*) do NIH (*National Institute of Health*), localizado nos EUA, denominado I2B2 (*Informatics for Integrating Biology and the Bedside*) é responsável por desenvolver aplicativos para gerenciar dados clínicos na era genômica (I2B2, 2011)**.** O I2B2 Hive é um framework composto de módulos de software para auxiliar computacionalmente a pesquisa clínica (MURPHY et al., 2007)**.** Cada módulo de software é denominado uma 'Célula' e cada Célula pode comunicar-se com outra por Serviços Web. Os principais módulos são responsáveis pelo armazenamento de dados, gerenciamento de ontologias e gerenciamento de identidade entre outros. Embora o I2B2 Hive seja uma ferramenta poderosa e escalável para gerenciar a informação clínica, ela não possui um módulo para representar ou analisar dados biomoleculares tais como *microarray* ou dados de sequência moleculares.

Na área da genômica existem diversos modelos de bancos de dados biológicos tais como AceDB, Ensembl e Chado. Os modelos são a base para construir ferramentas computacionais para análise genômica de forma independente do organismo estudado.

AceDB (*A C.elegans Database*) (STEIN; THIERRY-MIEG, 1999) é um dos modelos pioneiros para bancos de dados biológicos. É um Sistema Gerenciador de Banco de Dados hierárquico e foi inicialmente construído para auxiliar em pesquisas sobre *C. elegans*, porém posteriormente foi adaptado para outros organismos. É baseado em uma abordagem integrativa e pode ser utilizada para representar muitos outros tipos de informação, até aquelas não relacionadas à biologia.

Ensembl (HUBBARD et al., 2002) foi inicialmente desenvolvido para apoiar pesquisas relacionadas ao projeto genoma humano e atualmente oferece suporte para mais de 45 genomas de espécies diferentes. Possui um conjunto de ferramentas computacionais tais como EnsMart (KASPRZYK et al., 2004) que é um *data warehouse* biológico para integrar e consultar dados biológicos.

Um modelo de banco de dados biológico que tem ganhado popularidade entre grupos de pesquisas que estudam diferentes organismos é o Chado (MUNGALL; EMMERT; THE FLYBASE CONSORTIUM, 2007). Chado é uma plataforma robusta, flexível e genérica que pode ser adaptada para auxiliar a pesquisa em diversos organismos. Define um esquema modular de um banco de dados relacional que pode ser adaptado e estendido. Uma característica essencial do Chado, que difere de outros modelos de bancos de dados biológicos, é o fato de ser orientado a ontologias. Ontologias são artefatos informacionais estruturados, utilizados para representação, padronização e integração do conhecimento em diferentes domínios. Ontologias variam desde simples vocabulários utilizados para padronização de termos, até completos modelos conceituais que permitem inferência e descoberta de conhecimento (RUBIN; SHAH; NOY, 2008). Chado, assim como outros modelos de bancos de dados biológico, não possui um módulo para armazenar e representar informações clínicas e sócio demográficas.

É nesse contexto que o presente trabalho está inserido, objetivando o projeto e implementação de um framework computacional que agregue, de uma forma consistente,

dados clínicos e biomoleculares. Com isso, é possível o desenvolvimento de métodos de análises computacionais para serem aplicados no campo da medicina translacional. Uma ontologia de referência foi utilizada para garantir a padronização e permitir o futuro desenvolvimento de ferramentas genéricas para análise de dados.

Este trabalho de mestrado está vinculado ao Projeto "Oncogenômica Aplicada à Terapia de Carcinomas de Cabeça e Pescoço", que tem como objetivo a realização de pesquisas conjuntas voltadas para análise de mecanismos genéticos e epigenéticos responsáveis por regular o transcriptoma e o secretoma em carcinomas de cabeça e pescoço e que estão relacionadas à busca de biomarcadores de diagnóstico, prognósticos e que possam ser usados como alvos terapêuticos (CNPq processo 559809/2009-3). Este é um projeto colaborativo, coordenado pelo Prof Dr Wilson Araújo da Silva Junior, do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto, da Universidade de São Paulo (FMRP-USP), que compreende 20 pesquisadores de 9 Grupos de pesquisas diferentes. Dentre esses grupos, 7 pertencem à FMRP-USP, 1 grupo pertence à Universidade Federal do Pará (UFPA) e o outro é vinculado ao Hospital Santa Tereza de Guarapuava. O projeto é composto por 6 subprojetos, cada um abordando determinado aspecto da oncogenômica aplicada à terapia de carcinomas de cabeça e pescoço. Em todos os subprojetos serão utilizadas amostras de carcinoma de cabeça e pescoço em estágio inicial e avançado.

1.2. MOTIVAÇÃO E OBJETIVOS

Este projeto de mestrado teve como objetivo geral o estudo e o projeto de métodos computacionais para integração e análise de dados clínicos e biomoleculares. Como objetivos específicos, podemos citar: (i) projetar um *framework* computacional para representar, armazenar e integrar informação clínica e sócio demográfica com dados provenientes de experimentos biológicos, tais como *microarray*; (ii) desenvolver uma metodologia de integração de diferentes bancos de dados clínicos para o *framework* proposto a partir da utilização de uma ontologia de referência e (iii) instanciar o framework em uma ferramenta computacional para o domínio da oncogenômica.

Como caso de uso para teste do *framework*, foi realizada integração dos dados do projeto "Oncogenômica aplicada à terapia de carcinoma de cabeça e pescoço". Por meio desse *framework*, é possível integrar dados de sequência moleculares, dados de expressão gênica de *microarrays* e dados de biomaterias com as características clínicas e sócio-demográficas de pacientes que forneceram amostras para geração dos exames laboratoriais.

1.3. ORGANIZAÇÃO DA DISSERTAÇÃO

A presente dissertação está organizada da seguinte forma. No Capítulo 2 são descritos os fundamentos teóricos que serviram de base para o desenvolvimento deste trabalho. No Capítulo 3 são discutidos os trabalhos correlatos. O *framework* de integração é descrito no Capítulo 4. A ferramenta IPTrans, implementada a partir do *framework*, é descrita em detalhes no Capítulo 5. No Capítulo 6 é demonstrada a aplicação da ferramenta em dois projetos distintos, validando a sua utilização. Conclusões, possíveis limitações e trabalhos futuros são discutidos no Capítulo 7.

2. FUNDAMENTOS TEÓRICOS

2.1. CONSIDERAÇÕES INICIAIS

Neste capítulo serão apresentados os conceitos teóricos que serviram de base para o projeto e implementação do *framework* de integração de dados clínicos e biomoleculares. Inicialmente é discutido o problema de integração de dados com uma visão geral sobre diferentes metodologias. Em seguida, será discutido o modelo de dados Entidade-Atributo-Valor que foi o modelo utilizado como suporte para a base de dados integradora. Posteriormente é apresentada uma visão geral sobre ontologias que foram adotadas como modelo de representação do conhecimento, finalizando com uma descrição sobre *microarrays* que foi a tecnologia de biologia molecular utilizada neste projeto.

2.2. INTEGRAÇÃO DE DADOS

Segundo Lenzerini (LENZERINI, 2002) integração de dados é o problema de combinar dados residindo em diferentes fontes e fornecer ao usuário uma visão unificada desses dados. A integração normalmente ocorre em dois diferentes cenários. O primeiro é quando se sabe exatamente as questões que se quer responder e quais são os dados disponíveis. O segundo ocorre quando queremos retirar o máximo de informações a partir dos dados e não conhecemos todas as fontes a serem integradas, por exemplo, quando buscamos achar correlações ou descobrir novos conhecimentos a partir de técnicas de *data mining*. Projetos de integração de dados clínicos normalmente envolvem o segundo tipo de cenário, por exemplo, identificar surtos de epidemias, correlação entre doenças e fatores de riscos, causalidade entre doenças, ou identificação de grupo de pacientes baseados em diversos fatores.

De acordo com Bernstein(BERNSTEIN; HAAS, 2008) existem diferentes arquiteturas para resolver o problema de integração de dados, como, por exemplo, um *Data Warehouse*, que é um banco de dados voltado à consolidação de dados oriundos de diferentes fontes. Neste caso, os dados passam por um processo de limpeza para excluir duplicação de informação, corrigir erros de digitação, normalização, entre outras alterações, normalmente

com o auxílio de uma ferramenta de ETL (*Extraction, Transformation and Load*). Com o Data Warehouse é possível sumarizar e analisar grandes volumes de dados. Uma outra arquitetura é a Integração Virtual de Dados que fornece ao usuário ou às aplicações a ilusão de que os dados estão sendo consolidados em uma fonte única, porém sem realmente ocorrer a carga desses dados em um banco único. Para isso é fornecido um esquema mediador por meio do qual são realizadas as consultas. O sistema mediador traduz a consulta feita em termos do esquema mediador para consultas específicas para os esquemas das fontes de dados que estão sendo integradas. Uma outra arquitetura é o Mapeamento de Mensagens, onde utiliza-se um *middleware* orientado a mensagens que auxilia na integração de diferentes aplicações por meio da troca de mensagens entre elas. Esta arquitetura normalmente é implementa de duas formas diferentes, denominadas: Integração de Aplicações Corporativas (EAI – *Enterprise Application Integration*), onde se utiliza um *broker* para traduzir mensagens de um sistema para outro; Barramento de Serviços Corporativos (EBS – *Enterprise Service Bus*) onde as mensagens são trocadas utilizando um protocolo comum, por exemplo, serviços web, dispensando o uso de um *broker*.

A seguir, será detalhada a arquitetura de integração que consiste de um esquema global e um conjunto de fontes heterogêneas. O esquema global fornece uma visão unificada e integrada das fontes de dados. Um processo crucial é realizar o mapeamento dos esquemas fontes de dados para o esquema global.

2.2.1. CAMADAS DE INTEGRAÇÃO

Segundo Brazhnik(BRAZHNIK; JONES, 2007) a integração ocorre em quatro grandes camadas: fontes de dados, ED (elementos de dados), conjuntos de dados e valores de dados. Incluindo integração de conceitos, modelos, vocabulários controlados, métodos de aquisição, frequência de atualizações assim como unidades e formatos dos registros.

Outro aspecto importante de integração de fontes de dados são os metadados. Do ponto de vista da integração, existem dois tipos de metadados: um é utilizado para identificação única de uma instância, enquanto o outro é utilizado para representar informações adicionais sobre o ambiente.

Do ponto de vista de integração de dados, os EDs podem ser classificados em três tipos: chaves de integração, ED informativo e ED auxiliar. O objetivo da integração é obter e analisar os EDs informativos, pois são eles que contém a informação necessária para geração do conhecimento, tais como sinais, sintomas, idade, grau do tumor, etc. Porém só estes elementos não são suficientes para que a integração ocorra. Para construir um arcabouço de integração é necessário identificar as chaves de integração que irão fazer a ligação entre os elementos de fontes de dados diferentes. Uma chave de integração consiste em um conjunto de EDs focais que estão presentes em ambas as fontes de dados e que juntos identificam univocamente uma mesma entidade nas diferentes fontes. EDs auxiliares normalmente estão associados a regras de negócios e são utilizadas para lidar com ambiguidades ou exceções. EDs auxiliares também podem ser utilizados para validação de algoritmos de combinação automática baseada nas chaves de integração.

2.2.2. Padrões e Formatos

Pode-se caracterizar a informação clínica em três dimensões: intra-institucional, inter-institucional e temporal(BRAZHNIK; JONES, 2007). A dimensão intra-institucional consiste em dados relacionados a diferentes setores dentro de uma mesma instituição em saúde como: radiologia, laboratório, farmácia, ambulatórios, enfermarias, entre outros. A dimensão inter-instituicional consiste em dados vindos de diferentes instituições em saúde, por exemplo, postos de saúde, hospitais, clínicas, laboratórios, órgãos governamentais. Neste caso os padrões de informação em saúde tais como HL7 (*Health Level 7*), CDISC (*Clinical Data Interchange Standards Consortium*), DICOM (*Digital Imaging and Communication in Medicine*) e SNOMED (*Systematized Nomenclature of Medicine*) são importantes para auxiliar na interoperabilidade semântica e sintática. A dimensão temporal consiste nas mudanças realizadas a partir da passagem do tempo que incluem desde alterações nas condições de saúde de um paciente até a criação ou modificação de novos procedimentos e equipamentos. Novas tecnologias em biomedicina são continuamente criadas, avanços na área de bioinformática também têm criado uma nova gama de dimensões no domínio biomédico que, por sua vez, necessitam o desenvolvimento de novos métodos e tecnologias para um mapeamento e integração eficiente e robusta.

Padrões e formatos devem ser definidos não só para os EDs mas também para os conceitos. Uma mensagem HL7 define todos os EDs e sua posição em uma mensagem. Para permitir a combinação de EDs de diferente fontes de dados é necessária uma definição explícita dos EDs utilizando padrões e unidades de medidas. Existem diferentes técnicas para o mapeamento do tipo de dado entre diferentes bancos de dados, como por exemplo o mapa de Torque(DEVAKI, 2004).

O processo de aquisição de dados consiste em obter os dados diretamente da fonte de estudo ou de múltiplas fontes. Os métodos para aquisição variam pelas características das fontes. As fontes podem variar conforme o seu projeto funcional, disponibilidade dos dados e o processo de aquisição dos dados.

Quanto às disponibilidades dos dados, as fontes podem ser classificadas como públicas ou privadas. Existem diversos bancos públicos como GenBank, UniProt, GEO e a obtenção dos dados pode variar desde total acesso ao banco de dados relacional até arquivos em formatos próprios.

As fontes também podem ser classificadas em primárias ou secundários de acordo com o método de aquisição da informação. Fontes primárias são aquelas em que os dados primeiramente são armazenados. Nestes casos a entrada de dados pode ser manual (recepcionista, enfermeira) ou automática (leitores de código de barra). Fontes de dados secundárias são aquelas que obtém os dados a partir das fontes primárias. Nos bancos secundários os dados normalmente são obtidos de três formas diferentes: agentes, *pull* de dados e arquivos de texto. Agentes são configurados para obter dados de uma fonte em um intervalo definido de tempo. *Pull* de dados são consultas personalizadas que são realizadas em bancos de dados externos. Arquivos de texto normalmente são obtidos quando não é possível um acesso direto ao banco de dados, por medidas de segurança.

2.2.3. Apresentação dos Dados Integrados

Data marts são criados a partir dos dados limpos e validados presentes no BD Principal para cada tipo de usuário. O propósito dos *data marts* é oferecer suporte a ferramentas e visões específicas para os usuários. Um importante aspecto relacionado a

dados de pesquisa é a de-identificação. Neste processo dados de paciente precisam ser mascarados para não permitir a identificação dos mesmo sem perder o valor da informação, ou seja, mantendo aquilo que é importante para a pesquisa. O processo de anonimização dos dados ou mascaramento consiste em trocar os valores reais dos EDs por valores codificados ou aleatórios. Quais EDs devem ser mascarados dependem do objetivo da pesquisa.

O processo de integração não consiste somente em apresentar todos os dados ao usuário. Para manter a consistência da informação é necessário que os modelos que representam as fontes de dados estejam mapeados entre si. A integração consiste em mapear conceitos, os modelos de doenças, assim como esquemas internos das fontes de dados. Esse processo de mapeamento dos esquemas é denominado combinação de esquemas (*schema matching*). A maioria dos dados clínicos está armazenada em bancos de dados heterogêneos que utilizam diferentes nomes e diferentes modelos de dados sendo necessário um mapeamento direto de sistema a sistema para permitir a integração. Poucos sistemas de informação em saúde são construídos utilizando um modelo de dados comum. Integração com outros domínios do conhecimento incluindo a genômica, proteômica e imagens traz novos desafios ao processo de integração de dados.

2.2.4. Combinação de Esquemas

Combinação de esquemas (*schema matching*) (CASANOVA et al., 2007) é o processo de encontrar mapeamentos entre conceitos de um esquema fonte e conceitos de um esquema alvo, relacionando-os semanticamente. A combinação de esquemas é importante tanto para *data warehousing* quanto para mediação de consultas (*query mediation*). A mediação de consultas usa um mediador para traduzir consultas, formuladas em termos de um esquema comum, para termos correspondentes dos esquemas fontes. Segundo Casanova, existem três principais abordagens para combinação de esquemas: sintático, semântico e *a priori*.

Abordagem Sintática

A abordagem sintática busca encontrar similaridades baseada em aspectos sintáticos tais como: tipo de dado dos atributos, nome dos atributos, som, etc. Essas técnicas baseiam-se

no princípio de que a similaridade sintática corresponde à similaridade semântica dos termos sendo mapeados.

ABORDAGEM SEMÂNTICA

A abordagem semântica geralmente tenta detectar como objetos do mundo real são representados em diferentes bancos de dados utilizando-se de aspectos semânticos. Um exemplo seria analisar o conteúdo de tabelas em fontes diferentes para tentar mapear os conceitos mais gerais. Normalmente mostram bons resultados com esquemas simples.

ABORDAGEM *A PRIORI*

Tanto a abordagem sintática ou semântica podem ser classificadas como *a posteriori,* no sentido de tentar combinar esquemas de banco de dados pré-existentes. Casanova propõe uma abordagem *a priori* na qual é definido, inicialmente, um conjunto de padrões ou ontologias que vão guiar o processo de combinação de esquemas. Estes padrões ou ontologias funcionam como um esquema global, neste caso a combinação é feita *a priori*, pois a combinação posterior de dois esquemas fontes é feita de forma trivial quando um esquema global já está definido.

Por meio de algoritmos de combinação de esquemas, são definidos mapeamentos entre esquemas diferentes. Mapeamentos semânticos são utilizados para definir como dados de uma fonte são traduzidos para dados de uma outra fonte, ou alternativamente, como uma consulta a uma fonte pode ser reescrita para uma consulta equivalente em outra (VIDAL et al., 2009).

Os mapeamentos podem ser classificados de acordo com a sua acurácia de três maneiras diferentes: *sound* (parecidos), exatos e completos. Mapeamentos são do tipo *sound* quando a *view* definida por ele satisfaz os elementos correspondentes do esquema mediador, porém pode haver outros dados que não foram abrangidos pela *view*. Mapeamentos são dito completos quando nem todos os dados que a *view* provê satisfazem o elemento do esquema mediador, porém todos os dados que satisfazem o elemento são providos pela *view*. A *view* é dita exata quando os dados providos por ela satisfazem o elemento no esquema mediador.

A utilização de ontologias é uma possível abordagem para o caso em que a heterogeneidade semântica é grande devido à grande quantidade de atributos e/ou flexibilidade de geração de novos atributos, aumentando a variação entre esquemas de um mesmo domínio.

2.3. MODELO ENTIDADE-ATRIBUTO-VALOR

O modelo EAV (Entidade-Atributo-Valor) tem como base as listas de associações que representam a informação relacionada a qualquer objeto utilizando-se de pares atributo-valor. As listas de associações foram criadas na década de 1950 em linguagens de programação tais como LISP ou SIMULA67. Atualmente a maioria das linguagens de programação oferecem suporte a criação de listas de associações, denominadas *hashs, maps* e dicionários. Muitos modelos de representação da informação tais como XML (*eXtensible Markup Language*) e RDF (*Resource Description Framework*) estão relacionados com pares atributo-valor. Em bancos de dados relacionais os atributos da relação precisam ser atômicos e não podem conter grupos ou valores repetidos (primeira forma normal – 1FN) e dessa forma os pares atributo-valor se tornam triplas com a entidade (o indíviduo que está sendo descrito, representado por algum tipo de identificador único).

Em bancos de dados relacionais, os atributos de uma classe são tradicionalmente representados por colunas em uma tabela. Esta forma de modelagem é interessante quando a quantidade de atributos é fixa sendo cada instância, representada em uma linha na tabela, possui valores para a maioria ou todos os atributos. Para os casos em que uma classe possui um grande número de atributos e cada instância possivelmente terá poucos atributos não-nulos a modelagem atributo-coluna não é adequada. Dados que mostram essa característica, ou seja, que possuem uma discrepância relevante entre o número de potenciais atributos e a quantidade real de atributos, são caracterizados como **esparsos**. Os dados também podem ser caracterizados como **voláteis** quando a quantidade de atributos são variáveis, ou seja, novos atributos são adicionados com o tempo e outros podem se tornar obsoletos.

2.3.1. *ROW MODELLING* PARA ESPARSIDADE E VOLATILIDADE

Row Modelling é o processo de projetar uma tabela de forma que cada registro *(row)* representa um ou mais fatos relacionados a uma entidade. Fatos adicionais são armazenados utilizando registros adicionais. *Row Modelling* trata os problemas de esparsidade e volatilidade e deveria ser utilizado em qualquer dessas duas condições.

Modelo Entidade-Atributo-Valor é uma generalização de *Row Modelling*, onde uma única tabela (ou um conjunto de tabelas) é utilizada para armazenar todos os fatos que possuem a propriedade de volatilidade e esparsidade dentro de um banco de dados. A vantagem da utilização do modelo EAV é a flexibilidade e simplicidade quanto à representação de uma grande quantidade de atributos. Existem alguns casos em que o modelo EAV é preferivelmente aplicado em relação ao modelo colunar:

- Atributos são heterogêneos com respeito ao tipo de dado;
- Muitas classes precisam ser representadas e sua quantidade pode variar continuamente, porém a quantidade de instâncias para cada classe é pequena mesmo se os atributos são poucos e não esparsos. Neste caso, o esquema seria composto por um grande número de relações (tabelas) com poucos registros em cada, e o esquema estaria em constante mudança;
- Existência de classes híbridas, ou seja, que possuem alguns atributos pouco esparsos e fixos e outros atributos variáveis e esparsos. Os atributos não esparsos são representados da forma convencional enquanto que os outros são representados em uma tabela EAV.

2.3.2. REPRESENTANDO ENTIDADE, ATRIBUTO E VALOR

Em bancos de dados biológicos a entidade é comumente representada utilizando-se uma tabela “Objetos”, na qual é registrado todo “objeto” do banco de dados. Campos comuns à tabela objeto são o nome, descrição e a classe a que este objeto se refere. Alguns bancos que utilizam essa estratégia são o Chromosome 19 database e NCBI Entrez. Em bancos clínicos a entidade normalmente representa um “Evento Clínico” que comumente é composto por um código identificador do paciente, um valor temporal (data e hora) em que o evento ocorreu e outros atributos dependentes da aplicação como protocolos de estudo.

Valores em uma tabela EAV são comumente armazenados utilizando-se o tipo de dado "*string*" (comumente o tipo *text* na maioria dos SGBDs). A vantagem deste método é a simplicidade de implementação e utilização, sendo que a desvantagem é não ser capaz de utilizar efetivamente as técnicas de indexação para dados intrinsecamente numéricos ou do tipo data. Existem algumas alternativas para contornar essa desvantagem, como: (i) criar múltiplas colunas para cada tipo de dado e uma coluna identificadora do tipo real do valor; ou (ii) utilizar tabelas EAV separadas para cada tipo de dado e, na definição do atributo definir o tipo do valor.

A representação dos atributos pode ser tratada de forma similar a um vocabulário controlado. Uma descrição detalhada dos atributos é importante para a definição de uma infraestrutura de gerenciamento de metadados. Podemos classificar os metadados dos atributos de acordo com a sua função:

- Validação: incluem o tipo de dado do atributo, restrições como valor máximo e mínimo, valor padrão ou se aceita valores nulos ou não;
- Apresentação: define a forma como o atributo vai ser apresentado ao usuário, assim como textos alternativos e possivelmente suporte à internacionalização;
- Agrupamento: atributos comumente estão agrupados a algum contexto, por exemplo, algum tipo de estudo, formulário ou exame. Este tipo de metadados define como os atributos são agrupados e como eles se relacionam dentro do agrupamento.

Sistemas de EAV trocam a simplicidade na estrutura física de dados por uma maior complexidade em relação aos metadados. Dessa forma, um importante componente de um sistema EAV é o componente de gerenciamento de metadados. O sistema continuamente acessa o componente de metadados para operações como apresentação, validação e consultas *ad hoc*. Estes tipos de metadados são denominados "metadados ativos" e este tipo de software é denominados "*metadata-driven*" *software*.

Minimamente, os sistemas EAV oferecem suporte a metadados dos atributos. Sistemas que lidam com diferentes tipos de entidade também devem oferecer suporte a

metadados de classes. Metadados também auxiliam na configuração automática das aplicações, por exemplo:

- *Skip Logic:* alguns atributos ou campos são dinamicamente disponibilizados de acordo com valores de outros atributos precedentes, por exemplo, no caso de um formulário com questões dependentes;
- Formulas Pré-Definidas: os valores de alguns atributos podem ser computados ou são influenciados pelos valores de outros atributos;
- Listas Dinâmicas: alguns campos são baseados em listas dinâmicas, com opções sendo definidas de acordo com escolhas prévias.

2.3.3. INTEGRIDADE DOS DADOS EM SISTEMAS EAV

No projeto convencional de banco de dados a integridade é mantida utilizando-se restrições definidas como expressões em SQL aplicadas a uma coluna ou um conjunto de colunas de uma tabela. Em sistemas EAV as restrições a atributos e classes são definidos nos metadados, de forma que o desafio está na implementação dessas restrições.

Sistemas EAV, assim como sistemas convencionais, são implementados utilizando N-camadas. A camada inferior, denominada camada de dados ou *back-end*, é basicamente o banco de dados; a última camada, denominada camada de apresentação ou *front-end* é a camada de interface com o usuário e as camadas intermediárias são onde reside a lógica de negócios e seu número é variável, de acordo com cada aplicação.

Dessa forma, as restrições em sistemas EAV podem ser executadas em qualquer uma das camadas do sistema. Na camada de dados, essas restrições podem ser executadas a partir de gatilhos em SQL. A vantagem seria a garantia de execução desses gatilhos, porém há a necessidade de utilizar linguagem procedural para definição dos gatilhos já que é necessária uma interpretação dos metadados para correta aplicação das restrições. As restrições também podem ser executadas na camada intermediária, onde são avaliadas antes de ocorrer qualquer inserção ou modificação na camada inferior. Na camada de apresentação essas restrições também podem ser geradas automaticamente a partir dos metadados. Em aplicações web, normalmente são geradas restrições definidas em

Javascript que fazem a checagem dos valores antes dos dados serem enviados ao servidor, fornecendo um retorno instantâneo para o usuário. Porém algumas medidas de segurança devem ser aplicadas para garantir a integridade da informação nestes casos.

A operação que consiste na conversão de dados do formato EAV para formato colunar é denominada pivoteamento *(pivoting)*. Algumas situações em que essa operação é aplicada são: visualização ou edição de uma grande quantidade de dados para uma única entidade, análises estatísticas e geração de interfaces para consultas *ad hoc*, entre outros. A operação reversa, transformar dados no formato colunar para formato EAV é comumente realizada durante um processo de migração de dados.

2.3.4. Estratégias para Otimização de Consultas

O conceito do mecanismo do pivoteamento é uma série de junções do tipo *FULL OUTER* na tabela EAV. Essas operações de FULL OUTER JOINS são realizadas de forma que os atributos sejam colocados lado a lado, no formato colunar, e ao mesmo tempo permita a representação de valores nulos para alguns atributos. A eficiência do modelo EAV é diretamente proporcional à necessidade de realizar o pivoteamento dos dados. Consultas que exigem essa transformação são relevantemente menos eficientes quando se compara esta abordagem com o uso do modelo de dados colunar tradicional. Existem alguns métodos mais eficientes que podem ser aplicados dependendo da operação a ser realizada posteriormente e que demanda o pivoteamento. Pode-se caracterizar 3 tipos diferentes de operações: operações centradas na entidade, consultas *ad hoc* (centradas no atributo) e extração em massa de dados.

- Operações Centradas na Entidade: são operações que envolvem a seleção de entidades em particular e retornam todos os atributos associados (ex: todos os detalhes de um determinado paciente em um determinado período de tempo). Segundo Chen (CHEN et al., 2000) este tipo de operação é executada significantemente mais rápido em sistemas EAV. Isto acontece porque em sistemas convencionais a informação sobre uma entidade está dispersa em várias tabelas no banco de dados enquanto que no modelo EAV ela fica centralizada. A otimização deste tipo de consulta consiste na criação de índices em colunas da tabela de

paciente (colunas com informação demográfica, tais como nome e data de nascimento), na tabela de entidade (id) e no id dos atributos na tabela EAV.

- Consultas *ad hoc* (centradas no Atributo): Este tipo de operação envolve selecionar uma quantidade de entidades a partir dos atributos. As expressões de consulta são compostas por critérios booleanos baseando-se nos valores desses atributos na tabela EAV. A consulta é executada selecionando grupos de entidades de acordo com cada atributo e depois combinando esses grupos utilizando de operadores lógicos como união e intersecção. Esse tipo de operação é executada mais lentamente em sistemas EAV do que em sistemas convencionais(CHEN et al., 2000), porém esses tipo de consulta normalmente é importante para termos de pesquisa, e nesses casos não precisam ser executadas em tempo real. Dois métodos para execução dessa operação podem ser realizados: (i) gerar uma única e grande consulta em SQL que busca obter todos os dados em um único passo ou; (ii) gerar várias pequenas consultas em SQL e depois realizar junção delas. Segundo os autores em (CHEN et al., 2000), o segundo método obteve um melhor desempenho em relação ao primeiro.
- Extração de Dados em Massa: essa operação consiste normalmente em obter todos os dados de todas as entidades para um determinado contexto ou período de tempo (ex: obter dados clínicos e demográfico de todos os pacientes de um determinado estudo). Existem dois métodos que podem ser aplicados para extração de dados em massa em um modelo EAV: Extração com Antecedência ou Extração por Demanda.
    - Extração com Antecedência (Visões Materializadas): Uma visão materializada consiste em uma relação definida a partir de resultados pré-computados de uma consulta em SQL;
    - Extração por Demanda (Estruturas de Dados na Memória): Nesta técnica, estruturas de dados tais como *hashs* e *arrays* são utilizados para realizar o pivoteamento de dados em demanda (por exemplo, para cada estudo) e posteriormente estes dados são escrito em disco.

2.4. ONTOLOGIAS

Cada vez mais as ontologias vão se popularizando dentro da área da bioinformática e também das biociências em geral (BODENREIDER; STEVENS, 2006). Segundo Gruber (GRUBER, 1993), ontologia é definida como a especificação explícita de uma conceitualização, ou seja, é a representação formal dos conceitos de um domínio, a parte da realidade que nos interessa. Construir uma ontologia pode resumir-se em definir os conceitos (classes) e os relacionamentos entre estes conceitos. Porém, ainda existe muita confusão em relação ao que de fato é uma ontologia. Diversos artefatos são denominados ontologias, mas podemos classificá-los, em grau crescente de complexidade, como sendo (RUBIN; SHAH; NOY, 2008):

- *Terminologias ou vocabulários controlados*: uma lista de conceitos com termos léxicos correspondentes e com descrições textuais de seu significado. São organizados normalmente de maneira hierárquica e utilizados para indexação ou para registro em um banco de dados. Exemplo: *Gene Ontology* (GO);
- *Modelos de informação*: são modelos que definem um determinado domínio de interesse de maneira organizada. Nesses modelos são descritos como os conceitos se relacionam entre si. Ex: *Microarray Gene Expression Object Model* (MAGE-OM);
- *Ontologias completas*: são representações formais do conhecimento de um determinado domínio. Os conceitos e os relacionamento entre estes conceitos são expressos em termos de axiomas em alguma lógica bem definida.

As ontologias específicas da área da biomedicina são denominadas bio-ontologias. A mais famosa entre as bio-ontologias existentes é, certamente, a *Gene Ontology* (GO). A GO é definida como um vocabulário controlado e estruturado do domínio da biologia molecular e celular que padroniza os termos utilizados para descrever as funções dos genes e dos produtos gênicos em qualquer organismo (ASHBURNER et al., 2000). A GO permite a padronização da informação a partir de três ramificações de diferentes domínios da biologia molecular e celular: Função Molecular, Processo Biológico e Componente Celular. O mapeamento da interação entre termos de uma mesma ontologia e entre ontologias diferentes é feito por meio de um conjunto de 6 diferentes relacionamentos: *"is-a",*

"part_of", "regulates", "positively-regulates", "negatively-regulates" e *"has_part"* (CONSORTIUM, 2010). Cada um desses relacionamentos possui semântica bem definida e permite o controle da qualidade do mapeamento ontológico assim como a realização de consultas para responder hipóteses biológicas. É possível obter as ontologias no formato OBO (*Open Biomedical Ontology*), porém existem diversas ferramentas para conversão automática do formato de arquivo OBO em um arquivo OWL (*Web Ontology Language*). As partes componentes da GO continuam a crescer e a se modificar a cada nova versão, melhorando a abrangência e a especificidade do conhecimento mapeado. Atualmente existem mais de 28 mil termos distribuídos nas três ramificações que fazem parte da GO (CONSORTIUM, 2010).

Existem diversas ontologias utilizadas para integração de dados na área da medicina. Serão discutidos em detalhes três ontologias: SNOMED CT, uma ontologia que procura abranger todos os aspectos da área médica; Translational Medicine Ontology, específica do domínio da medicina translacional e a ACGT Master Ontology, que é específica do domínio da oncogenômica.

SNOMED CT – *SYSTEMATIZED NOMENCLATURE OF MEDICINE CLINICAL TERMS*

Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) surgiu a partir da união do SNOMED RT (*Reference Terminology*) criado pelo Colégio de Patologistas Americanos (CAP – *College of American Pathologist*) e CTV3 (*Clinical Terms Version 3*) desenvolvido pelo Serviço Nacional de Saúde do Reino Unido (IHTSDO, 2012). SNOMED CT é uma terminologia clínica que compreende um conjunto de conceitos, termos e relacionamentos, com o objetivo de representar informação em saúde. Atualmente o SNOMED CT é mantido e distribuído pela associação sem fins lucrativos IHTSDO (*International Health Terminology Standards Development Organization*).

SNOMED CT inclui mais de 311.000 conceitos ligados por aproximadamente 1.360.000 relacionamentos. Estes conceitos não estão restritos à área médica e compreendem domínios mais gerais como localizações geográficas e contexto social. Os conceitos ainda são organizados em 19 hierarquias com escopos bem definidos (Figura 1): achado clínico, procedimento, entidade observável, estrutura anatômica, organismo,

substância, produto biológico/farmacêutico, espécime, conceito espacial, conceito de acoplamento (*linkage*), fenômeno físico, evento, localização geográfica ou ambiente, contexto social, contexto explícito com situação, escala e estadiamento, objeto físico, valor qualificador e artefato de registro.

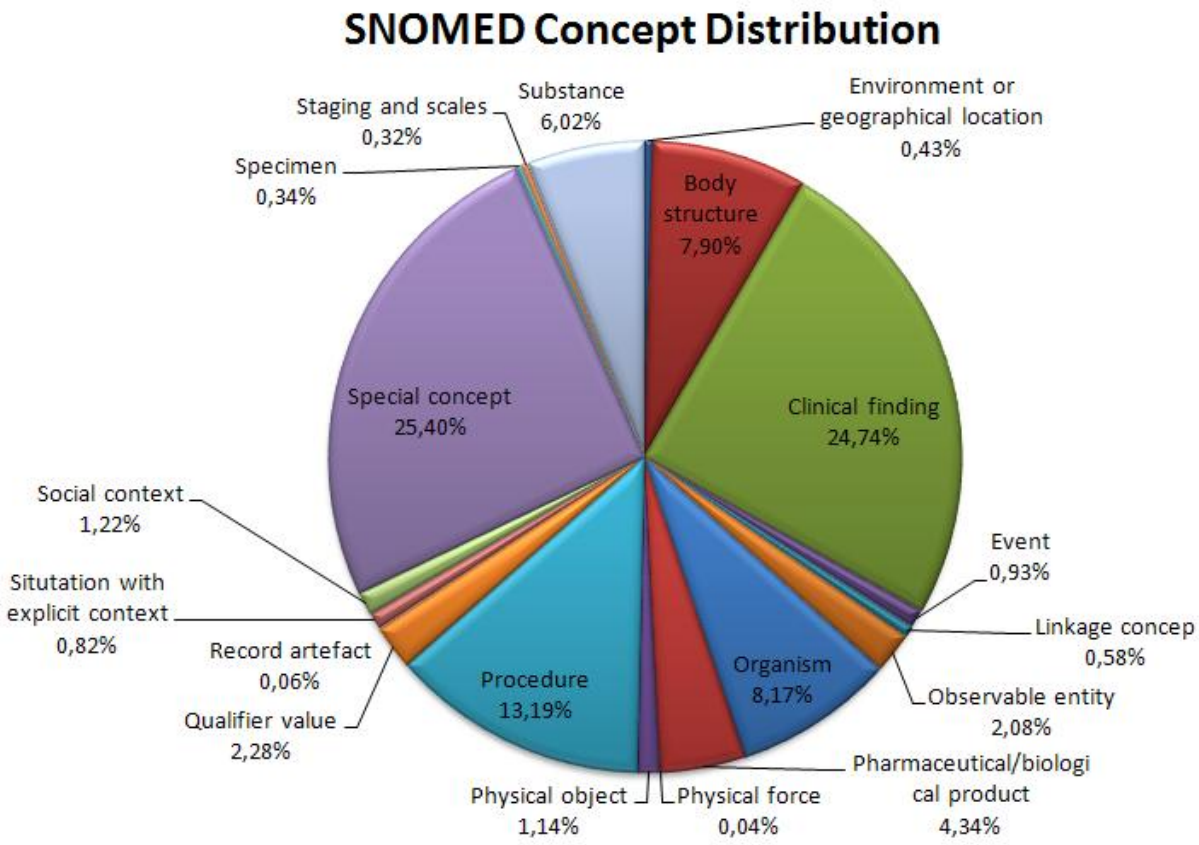

FIGURA 1. DISTRIBUIÇÃO DOS CONCEITOS DA SNOMED NAS HIERARQUIAS (DADOS DE 2011)

Os três componentes básicos do SNOMED são: conceitos, descrições e relacionamentos. Um conceito é um significado clínico que possui um identificador único (ConceptID), é formalmente definido a partir de seus relacionamentos e é descrito a partir de um conjunto de termos que formam os componentes do tipo 'descrição'. Um conceito também possui um FSN (*Fully Specified Name*) que é um identificador alternativo ao ConceptID porém legível e facilmente entendível. Uma descrição é composta por um termo ou um conjunto de termos e é identificada pelo DescriptionID podendo estar associada com um determinado conceito. Os relacionamentos definem logicamente os conceitos e formam a estrutura da SNOMED. O relacionamento principal é o *´is-a´* que define o supertipo de

cada conceito criando a hierarquia da terminologia. Existem outros 50 relacionamentos que definem os atributos de cada conceito como por exemplo *´finding-site´*, *´associated-with´* e *´has-intent´*.

Embora a IHTSDO, mantenedora da SNOMED, seja uma organização sem fins lucrativos, para utilizar essa ontologia é necessário obter uma licença. Seu uso é gratuito para países que são membros do IHTSDO e países considerados pouco desenvolvidos. O Brasil não se enquadra em nenhuma das duas categorias, porém é possível requisitar uma licença gratuita para fins acadêmicos.

TMO – TRANSLATIONAL MEDICINE ONTOLOGY

O Grupo de Interesse em Web Semântica da W3C (*World Wide Web Consortium*) denominado HCLS (*Health Care and Life Sciences*) busca desenvolver, distribuir e apoiar o uso de tecnologias da Web Semântica no domínio da biomedicina(W3C, 2010). Uma das forças-tarefa desse grupo é denominada *Translational Medicine*(W3C, 2012) e procura demonstrar a utilização das tecnologias da Web Semântica para promover a área da medicina translacional. Dois dos principais esforços dessa força-tarefa são a TMO (*Translational Medicine Ontology*) e a TMKB (*Translational Medicine Knowledge Base*).

A *Translational Medicine Ontology* (LUCIANO et al., 2011) compreende um conjunto de conceitos e relacionamentos de alto nível tanto do domínio clínico quanto biológico (Figura 2). A TMKB consiste de um conjunto de mapeamentos entre a TMO e outras ontologias e terminologias, além de possuir dados no formato RDF abrangendo descobertas científicas no desenvolvimento de fármacos, importantes tanto para pesquisa quanto para prática clínica.

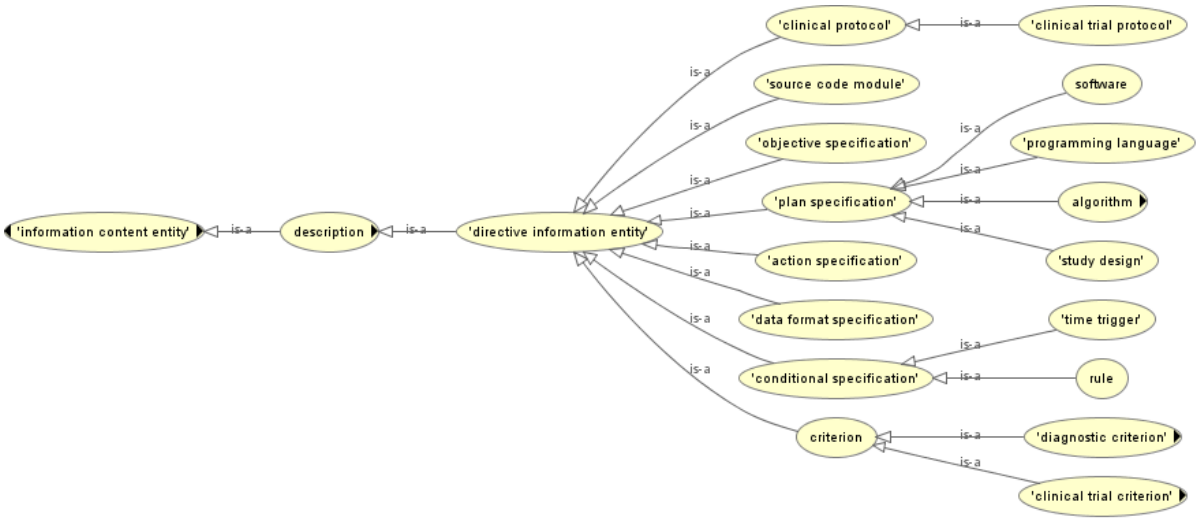

FIGURA 2. PARTE DA TRANSLATIONAL MEDICINE ONTOLOGY

A TMO utiliza 3 ontologias externas como base para seu desenvolvimento, são elas: BFO (*Basic Formal Ontology*), RO (*Relationship Ontology*) e IAO (*Information Artifact Ontology*). A BFO foi utilizada como ontologia de alto nível, fornecendo um conjunto de termos gerais que foram estendidos para o domínio da medicina translacional. RO define um conjunto de relacionamentos que foram utilizados na construção da TMO. IAO é uma ontologia de entidades de informação e define conceitos como ´símbolo´, ´título´, ´autor´, ´endereço´ e foi utilizada para realizar anotações para os conceitos da TMO. A TMO difere da SNOMED pois define conceitos apenas do domínio da medicina translacional, não abrangendo termos mais gerais.

Oitenta classes foram criadas representando materiais (proteína, molécula, linhagens celulares), processos (diagnóstico, estudo, intervenção), qualitativo, papéis (paciente, alvo) e entidades de informação (dosagem, mecanismo de ação, sinal/sintoma, histórico familiar). A TMO foi construída usando Protégé 4.0.2 e é distribuída no formato OWL2 (*Web Ontology Language*).

A TMKB possui um conjunto de mapeamentos realizados manualmente entre a 60 conceitos da TMO e 223 conceitos equivalentes em outras 40 ontologias incluindo SNOMED CT, ACGT Master Ontology, Gene Ontology e Sequence Ontology. Esses mapeamentos foram realizados e são disponibilizados no NCBO BioPortal e da UMLS. Um conjunto de datasets em RDF também compõe o TMKB. Esses *datasets* proveem dados sobre pesquisas

clínicas, informações farmacogenômicas, associações entre genes e doenças além de dados científicos sobre medicamentos. Esses dados foram obtidos de bancos de dados públicos como Trials.gov, DailyMed, Diseasome e DrugBank.

A TMO é distribuída gratuitamente sob a licença Creative Commons 3 e é continuamente mantida e desenvolvida pela força-tarefa em medicina transcional do HCLS IG da W3C com o objetivo de ser uma ontologia geral do domínio da pesquisa translacional.

ACGT MASTER ONTOLOGY

O ACGT (*Advancing Clinico-Genomic Trials on Cancer*) é um projeto financiado pela União Europeia com o objetivo de desenvolver uma infraestrutura de serviços computacionais que permita a execução de *workflows* científicos no contexto de ensaios clínicos multicêntricos e pós-genômicos(MARTIN et al., 2008). Os principais resultados do projeto ACGT são: desenvolvimento do ACGT Master Ontology (MO), desenvolvimento de uma infraestrutura técnica denominada ACGT Platform e uma aplicação para gerenciamento de ensaios clínicos denominada ObTiMA (*Ontology-based Trial Management Application*).

A ACGT *Master Ontology* foi publicada em 2007 e vem sendo continuamente expandida desde então (Figura 3). O seu desenvolvimento foi guiado e revisado por pesquisadores de dois ensaios clínicos preexistentes que estudavam câncer de mama e nefroblastoma conduzidos pela Sociedade Internacional de Oncologia Pediátrica. Dessa forma o escopo da ACGT MO é a área da pesquisa em câncer. Essa ontologia também pode ser considerada uma ontologia de aplicação, pois aborda conceitos relacionados ao gerenciamento de dados clínicos e de pesquisa na área de oncologia, representando também aspectos administrativos assim como relacionados a terapia e a parte laboratorial.

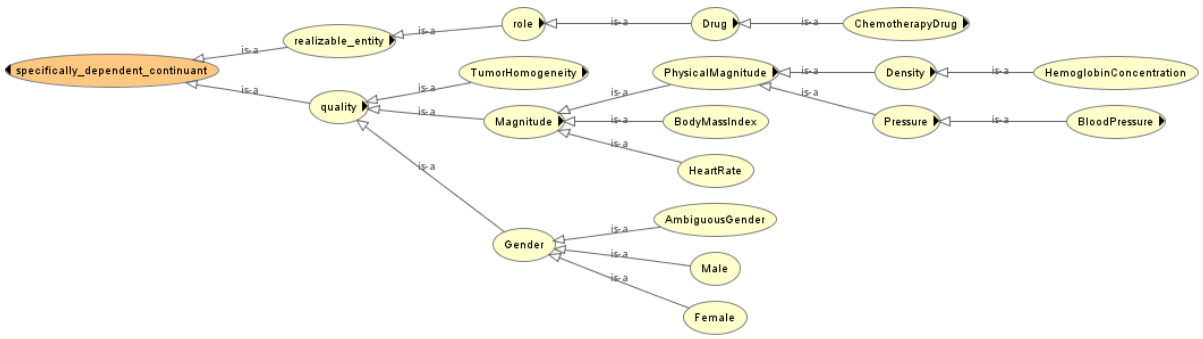

FIGURA 3. PARTE DA ACGT MASTER ONTOLOGY

A BFO foi utilizada como ontologia de alto nível da ACGT Master Ontology facilitando a sua integração com outras ontologias do domínio da saúde que estendam a BFO. A RO (*Relation Ontology*) também foi utilizada como fonte de relacionamentos entre conceitos da ACGT Master Ontology.

ACGT Platform é uma infraestrutura computacional desenvolvida para auxiliar pesquisadores envolvidos em ensaios clínicos em câncer nas tarefas de gerenciamento e análise das informações. Utiliza-se da ACGT MO como um modelo conceitual na definição da semântica da informação gerenciada. Serviu de base para construção de outras ferramentas computacionais como o ObTiMA.

ObTiMA é uma ferramenta computacional que utiliza a ACGT MO por meio da ACGT Platform para auxiliar pesquisadores no projeto e condução de ensaios clínicos em câncer. Os principais componentes do ObTiMA são: Trial Builder e Patient Data Management System. O Trial Builder permite a construção de Fichas Clínicas de Pesquisa (do inglês CRF - *Case Report File*). O Patient Data Management System auxilia na coleta e no gerenciamento dos dados do paciente no período em que a pesquisa está sendo realizada.

O ACGT MO sendo peça fundamental da arquitetura, é composto por 1667 conceitos, 288 relacionamentos e 15 propriedades de dados. Foi implementado utilizando a linguagem OWL na ferramenta Protégé. Esta ontologia é distribuída gratuitamente e está integrada no BioPortal do NCBO (*National Center of Biomedical Ontologies*).

2.5. MICROARRAY

Microarray (microarranjo) é uma tecnologia desenvolvida no início da década de 1990, cujo principal objetivo é detectar e quantificar o nível de expressão gênica de uma célula em um determinado estado biológico (RUSSELL; MEADOWS; RUSSELL, 2009). A tecnologia de *microarray* consiste em uma lâmina dividida em minúsculos compartimentos (*spots*) sendo que em cada um desses compartimentos está armazenada uma determinada sequência de nucleotídeos que representa univocamente uma sequência biológica, ou seja, um gene, um exon, um microRNA, etc. O tipo mais comum de *microarray* é o de DNA ou DNA-array, onde a molécula estudada é o gene. Em cada compartimento (*spot*) está armazenada uma sequência de nucleotídeos que representa um determinado gene. Existem plataformas comerciais em que é possível representar o genoma inteiro de um organismo, ou seja, todos os genes de um organismo em uma única lâmina.

Em um experimento típico, o RNA total é extraído da célula na condição biológica estudada, por exemplo, câncer, em seguida são copiados para cDNAs (DNA complementar) por ser uma molécula mais estável e então são marcados com um fluoróforo (normalmente Cy3 ou Cy5). O cDNA da célula tumoral é incubado juntamente com a lâmina e, pela propriedade de complementaridade de bases, o cDNA marcado com o fluoróforo se liga com a sequência de nucleotídeos correspondente na lâmina (hibridização). Em seguida essa lâmina é levada para processo de lavagem onde são retiradas as moléculas que não se ligaram e outros resíduos. A lâmina é então analisada por um scanner que gera uma imagem (normalmente em formato TIFF). A imagem gerada é composta por um conjunto de pontos luminosos onde cada ponto representa um *spot*, e a intensidade de luminosidade representa a quantidade de cDNAs que se ligaram às sequências armazenadas naquele spot. Dessa forma a intensidade de cada *spot* determina o nível de expressão do gene que é representado pela molécula armazenada naquele *spot*. Alguns *spots* estão apagados, ou seja, não aparecem na imagem, isso significa que o gene representado por aquela molécula não está sendo expresso, pois nenhum cDNA se ligou às moléculas daquele *spot*. O sinal obtido de cada *spot* é processado, quantificado e normalizado e no fim é possível obter um

valor numérico que representa o nível de expressão de cada gene naquele determinado estado biológico.

A partir da comparação do nível de expressão dos genes em células normais e em células tumorais, por exemplo, é possível determinar o conjunto de genes que estão diferencialmente expressos e que serão associados com o aparecimento do câncer. Com isso, é possível definir novos métodos diagnósticos, novos tratamentos e novo alvos para futuros fármacos. A tecnologia de *microarray* é uma das mais bem sucedidas e consolidadas para geração de dados biomoleculares.

2.6. CONSIDERAÇÕES FINAIS

Neste capítulo apresentamos os principais conceitos teóricos envolvidos no desenvolvimento deste projeto de mestrado. Abordamos quatro grandes temas: integração de dados, modelo entidade-atributo-valor, ontologias e microarrays. Os fundamentos abordados na seção de integração de dados serviram de base para o desenvolvimento da metodologia de integração e do mapeamento de conceitos. Parte do *framework* foi inspirado no modelo entidade-atributo-valor, para permitir a integração de informações heterogêneas e uma maior flexibilidade na representação da informação. Também foi feito uso de ontologias como metadados. Os dados biomoleculares utilizados como teste para integração foram dados de microarray.

3. TRABALHOS CORRELATOS

3.1. CONSIDERAÇÕES INICIAIS

Neste capítulo serão abordados trabalhos relacionados com o presente projeto de mestrado. Especificamente, serão discutidas três plataformas que oferecem suporte à pesquisa translacional: I2B2 (*Integrating Biology and the Bedside*), STRIDE (*Stanford Translational Research Integrated Database Environment*) e Slim-Prim (*Scientific Laboratory Information Management and Patient-care Research Information Management system*).

3.2. I2B2 – INTEGRATING BIOLOGY AND THE BEDSIDE

Os Centros Nacionais de Informática Biomédica (NCBC – *National Center for Biomedical Compute*) são projetos cooperativos financiados pelo Fundo Comum do NIH (*National Institute of Health*) com o objetivo de formar uma infraestrutura computacional para pesquisa biomédica (NIH, 2012). Um NCBC relacionado à integração da informação clínica é o I2B2 (*Integration Biology and the Bedside*).

A plataforma I2B2(MURPHY et al., 2010) foi desenvolvida para permitir que pesquisadores da área médica tenham acesso a ferramentas necessárias para integrar dados clínicos com dados de pesquisa. Modularidade é uma de suas características principais. A plataforma como um todo é denomina ´*Hive*´, sendo composta por módulos individuais denominados ´*Cells*´ (Figura 4). Cada *Cell* é responsável por uma determinada funcionalidade dentro da plataforma, encapsulando a lógica de negócios e tornando a solução fracamente acoplada. Atualmente existem 19 *Cells* sendo 6 que compõe o núcleo da Hive responsáveis pelas funcionalidades básicas: gerenciamento de ontologias, projetos, identidade, processamento de texto, repositório de arquivos, repositório de dados e workflow. Cada *Cell* se comunica com a outra por meio de serviços web e mensagens em XML que especificam certas propriedades comuns a *Cells* e outras utilizadas nas tarefas administrativas de enviar, receber e processar as mensagens.

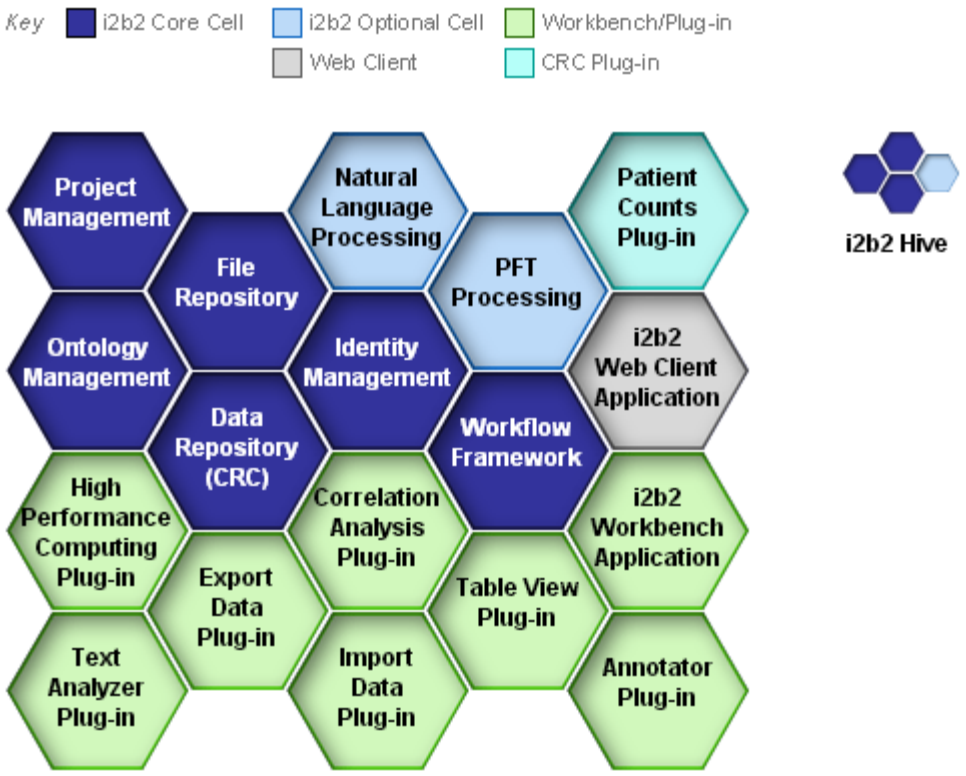

FIGURA 4. I2B2 HIVE

Na implementação das principais *Cells* foi utilizada a linguagem de programação Java sendo o *backend* escrito usando as especificações Java J2EE e o *frontend* utilizando a biblioteca SWT (*Standard Widget Toolkit*) que mantém uma apresentação padronizada em diferentes sistemas operacionais. Para permitir a comunicação entre as *Cells* usando as tecnologias REST ou SOAP foi utilizado Apache Axis2. Para mapear objetos em Java para XML e vice-versa foi utilizada a biblioteca JAXB (*Java Architecture for XML Binding*)(MENDIS et al., 2007).

A *Cell* responsável por armazenar os dados clínicos dos pacientes é denominada CRC (*Clinical Research Chart*). Nesta *Cell* os dados clínicos são armazenados em um Sistema Gerenciador de Banco de Dados Relacional (SGBD Relacional). O modelo de dados utilizado no CRC é baseado no "esquema estrela" típico de Data Warehouse, cuja estrutura possui uma tabela ´fato´ central onde cada linha representa um único fato. Os dados no CRC são representados no formato EAV e, neste caso, cada fato representa uma observação

relacionada a um paciente, sendo o tipo da observação um código que pertence a um vocabulário controlado.

A plataforma I2B2 é bastante flexível devido a sua arquitetura baseada em componentes (*Cells*). A utilização do formato de dados EAV a torna capaz de representar a heterogeneidade da informação clínica. Porém, atualmente o I2B2 não possui um modelo de dados genômicos, nem uma forma de representar a informação biomolecular, não possuindo também uma integração com ferramentas de bioinformática. Uma outra características é que o CRC, a *Cell* que armazena os dados clínicos, somente permite a utilização dos Sistemas Gerenciadores de Bases de Dados (SGBDs) Oracle e Microsoft SQL Server, pois o processo de ETL (Extração-Transformação-Carga) é definido utilizando ferramentas desses sistemas proprietários.

3.3. STRIDE

O Centro de Informática Clínica de Stanford (*Stanford Center for Clinical Informatics* – SCCI) criou em 2003 a plataforma computacional baseada em padrões denominada STRIDE (*Stanford Translational Research Integrated Database Environment*). Os principais objetivos do STRIDE são: (i) permitir o acesso eficiente a dados clínicos com o objetivo de pesquisa; (ii) fornecer soluções robustas para gerenciamento de dados e (iii) disponibilizar em um nível empresarial o gerenciamento de biomateriais.

Toda informação do STRIDE está armazenada em um banco de dados relacional administrado pelo SGBD Oracle 11g e está organizada em uma arquitetura de n-camadas (Figura 5). Os dados são armazenados seguindo o formato EAV (Entidade-Atributo-Valor) e são representados em uma estrutura de dados orientada a objetos utilizando o HL7 RIM (*Reference Information Model*). STRIDE também possui um MPI (*Master Person Index*) restrito ao SUMC (*Stanford University Medical Center*) que busca representar univocamente pacientes e biomateriais. A informação é logicamente distribuída em 3 bancos de dados diferentes: (1) *data warehouse* clínico, (2) banco de dados de pesquisa e (3) banco de dados de biomaterial, sendo os dois últimos compostos por vários pequenos bancos de dados

específicos de projetos. A camada semântica é composta por um conjunto de terminologias, vocabulários controlados e ontologias tais como SNOMED, ICD 9 e RxNorm que servem de base diversos serviços, como: integração de dados, recuperação de informação utilizando conceitos, entrada de dados padronizada, etc. A camada de aplicação utiliza um conjunto de clientes Java Swing que se comunicam com uma camada de serviços organizados em um Arquitetura Orientada a Serviços (SOA – *Service Oriented Architecture*) para gerenciar dados demográficos, clínicos e de pesquisa.

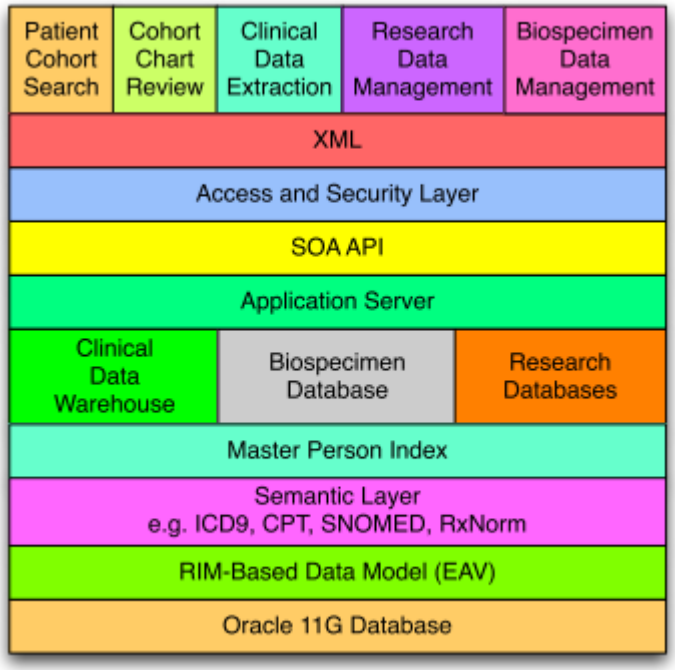

FIGURA 5. ARQUITETURA DO STRIDE

O *Data Warehouse* Clinico (DWC) do STRIDE foi inicialmente populado com dados de dois sistemas legados do SUMC em um processo de ETL. Atualmente a carga do DWC é feita em tempo real por meio de mensagens HL7 a partir de vários sistemas ligados ao SUMC. Foi desenvolvida a ferramenta de consulta visual *Anonymous Patient Cohort Discovery Tool*. Por meio dessa ferramenta pesquisadores podem buscar e definir coortes de interesse.

STRIDE vem sendo gradualmente desenvolvido desde 2003, buscando estabelecer uma plataforma baseada em padrões para área de pesquisa translacional. Fornecem uma estrutura computacional para desenvolvimento de aplicativos específicos para diversos

projetos de pesquisa. Atualmente os autores estão considerando a utilização do REDCap (HARRIS et al., 2009) para pequenos projetos de pesquisa devido à complexidade do STRIDE. Apesar do STRIDE utilizar padrões de informação em saúde e tecnologias portáteis como Java, ele foi desenvolvido exclusivamente para ser utilizado junto aos aplicativos do SUMC da Stanford e, segundo os autores, atualmente não há nenhum interesse de que ele venha a ser disponibilizado para fora de Stanford.

3.4. SLIM-PRIM

Slim-Prim (*Scientific Laboratory Information Management and Patient-care Research Information Management system*) (VIANGTEERAVAT et al., 2009) é o sistema computacional desenvolvido na Universidade do Tenessi para coletar e compartilhar dados e melhores práticas com o objetivo de acelerar o progresso na ciência biomédica.

Para permitir a inclusão de diferentes tipos de dados o Slim-Prim utiliza um sistema de gerenciamento de conteúdo denominado ContentNOW. ContentNOW usa da técnica de *row modelling* como forma de armazenamento dos dados, permitindo a representação de dados cujo formato não se conhece *a priori*. Segundo os autores, essa técnica aliada ao processo de mapeamento de conceitos mostrou-se promissora para agregação de dados em uma base integrada.

Slim-Prim emprega um sistema apoiado por metadados para controlar os diferentes tipos de informação que são integrados no sistema. Para isso são utilizados padrões de informação, ontologias e vocabulário controlados como SNOMED e CID10. O sistema de metadados também é utilizado para que os pesquisadores desenvolvam formulários que sirvam como entrada de dados para seus projetos de pesquisa (Figura 6) (VIANGTEERAVAT et al., 2009).

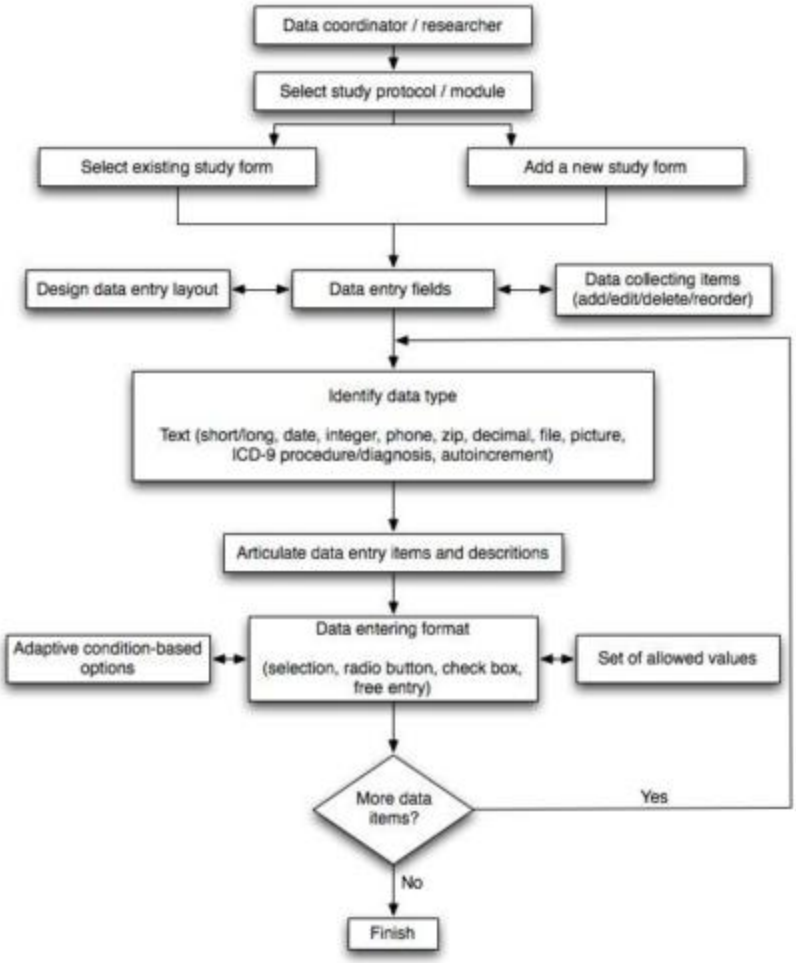

FIGURA 6. *WORKFLOW* PARA CRIAÇÃO DE UM FORMULÁRIO DE PESQUISA

Slim-Prim também possui uma ferramenta online para triagem de candidatos a pesquisas clínicas. Nessa ferramenta os pesquisadores definem um conjunto de critérios de interesse e um formulário online fica disponível para que candidatos acessem e forneçam informação de forma anônima. A ferramenta é responsável por analisar as informações e realizar a triagem informando aos pesquisadores aqueles candidatos que correspondem aos critérios de interesse. Todos os dados do Slim-Prim estão armazenados pelo SGBD Oracle 11g.

Slim-Prim é um software proprietário exclusivo da Universidade do Tenessi, porém existe uma versão de código aberto do Slim-Prim denominada PRIME (*Protected Research Information Management Environment*) (VIANGTEERAVAT et al., 2011). PRIME foi desenvolvido na linguagem de programação PHP e utiliza-se do banco de dados MySQL.

Tanto PRIME quanto Slim-Prim ainda não possuem uma integração com ferramentas de bioinformática e não permitem a representação de dado biomoleculares tais como sequências genômicas e *microarrays*.

3.5. CONSIDERAÇÕES FINAIS

Neste capítulo discutimos três sistemas computacionais para apoio à pesquisa translacional. O I2B2 é o mais disseminado entre os três discutidos, segue uma arquitetura baseada em serviços com um o modelo de integração baseado em *Data Warehouse*, porém não possui uma forma de representar informação biomolecular e não integra com ferramentas de bioinformática. O STRIDE, embora permita uma maior integração entre dados clínicos, dados de amostras biológica e biomoleculares, foi exclusivamente construído para ser utilizado com as bases de dados de Stanford. O Slim-Prim, a versão de código aberto do PRIME, é um sistema gerenciador de conteúdo para área da pesquisa translacional. A desvantagem do Slim-Prim é tratar os dados biomoleculares como dados genéricos não possuindo, também, maior integração com outras ferramentas e padrões de arquivo da área de bioinformática.

4. Proposta de um Framework de Integração

4.1. Considerações Iniciais

Para permitir a integração de dados clínicos e biomoleculares foi projetado um *framework* conceitual de integração. Neste capítulo serão abordados os diversos módulos e aspectos que compõe o framework assim como metodologias para integração da informação e o *pipeline* para migração de dados de bancos de dados legados para o *framework*.

4.2. Componentes do Framework

O *framework* proposto é composto pelos seguintes componentes: módulo de integração de esquemas, módulo de importação dos dados, módulo de consulta integrada, módulo de resolução de entidades, ontologia de referência e camada de compatibilidade (Figura 7).

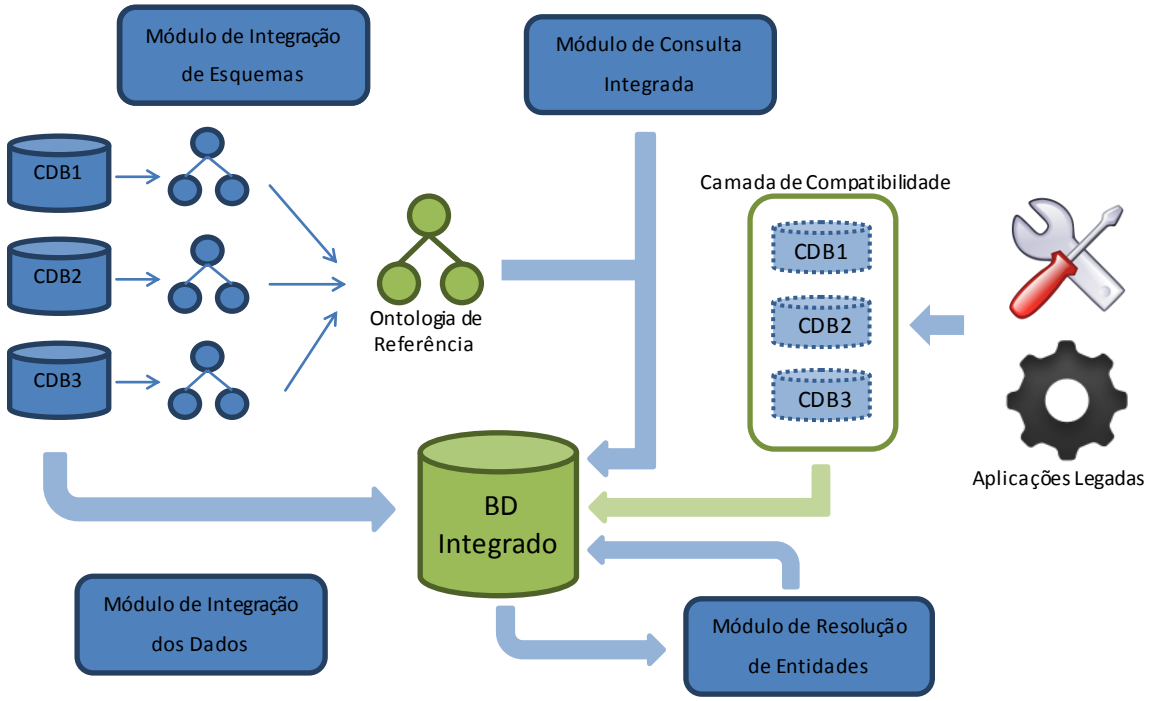

FIGURA 7. FRAMEWORK DE INTEGRAÇÃO

Fontes de Informação

As fontes de informação são o local originário dos dados a serem integrados. Fontes de informação podem variar desde arquivos de textos estruturados, arquivos com valores

separados por vírgula, bancos de dados relacionais, arquivos XML, ou então arquivos seguindo algum padrão de informação.

MODELOS DAS FONTES DE INFORMAÇÃO CLÍNICA

Estes modelos são responsáveis por representar a estrutura das fontes de informação clínica. São os esquemas das fontes de informação. Sua representação dependerá do tipo da fonte de informação clínica podendo ser, por exemplo, um esquema XML se a fonte for um arquivo XML ou o esquema lógico do banco de dados se a fonte for um banco de dados relacional.

MÓDULO DE INTEGRAÇÃO DE ESQUEMAS

O módulo de integração de esquemas é responsável por combinar os esquemas das fontes de dados com o esquema global, ou seja, é o módulo que realiza o mapeamento entre os modelos das fontes de dados e a ontologia de referência. O módulo de integração de esquema deve oferecer suporte para mapeamento manual, semi-automático e automático. Neste módulo estão armazenados os diferentes algoritmos de combinação de esquemas.

ONTOLOGIA DE REFERÊNCIA

A ontologia de referência é uma ontologia utilizada para integrar a informação clínica de diferentes modelos de Bancos de Dados Clínicos. Ela é composta por uma ou mais ontologias de domínio. Esta ontologia é utilizada como um arcabouço conceitual onde a informação é integrada por meio de mapeamentos entre conceitos de um modelo de banco de dados clínico e a ontologia de referência. A função da ontologia de referência é atuar como um esquema global ao qual todos os esquemas das fontes são mapeados.

BANCO DE DADOS INTEGRADO

O banco de dados integrado é o local onde serão armazenados os dados clínicos extraídos das fontes de dados. Deve ser capaz de representar a informação biológica, como, por exemplo, sequências moleculares (DNA, RNA, proteína, microRNAs, etc) e resultados de análise de expressão; assim como a informação clínica e sócio demográfica como, por exemplo, idade, peso, altura e informações mais específicas como tamanho do tumor, estadiamento, características da metástase, etc.

Módulo de Importação de Dados

O módulo de importação de dados é responsável por realizar a carga do banco de dados integrado a partir das fontes de informação. O módulo é composto por diferentes tipos de interface de importação cada interface para um tipo diferente de fonte. Também é responsável por armazenar o esquema de fonte de informação.

Módulo de Resolução de Entidade

O módulo de resolução de entidade realiza o processo de deduplicação dos dados, ou seja, identifica se existem dados duplicados para uma mesma entidade realizando a combinação dessas informações. Este processo é realizado após a importação dos dados e a combinação dos esquemas.

Camada de Compatibilidade

A camada de compatibilidade consiste em um conjunto de visões realizadas em cima do banco de dados integrado, as quais representam a estrutura original das fontes de dados clínico. A vantagem de criar uma camada de compatibilidade é facilitar a consulta e a adaptação de ferramentas computacionais legadas, ou seja, que foram projetadas originalmente para a fonte de dados clínicos. Por meio da camada de compatibilidade é possível realizar a integração das aplicações legadas.

Módulo de Consulta Integrada

Este módulo é responsável por permitir realizar consultas na base de dados integrada a partir da ontologia de referência. As consultas são definidas inicialmente em termos da ontologia de referência. A partir do mapeamento realizado pelo módulo de combinação de esquema, essa consulta é expandida para conter os elementos das fontes de dados. Após ter sido realizada a consulta, os resultados são agrupados de acordo com os termos da ontologia de referência.

4.3. CONSIDERAÇÕES FINAIS

Neste capítulo apresentamos um *framework* conceitual para integração de dados baseado em ontologias. A abordagem de integração proposta utiliza-se de um esquema mediator, representado pela ontologia de referência, para integrar os esquemas das fontes de dados através do mapeamento dos modelos conceituais para a ontologia de referência. Também propusemos os módulo de resolução de entidades, de integração de esquemas, de consulta integrada e de importação de dados assim como a utilização da camada de compatibilidade para adaptação de ferramentas legadas. Embora o *framework* tenha sido projetado para área da pesquisa translacional, ele é genérico o suficiente para ser implementado em outros domínios de interesse.

5. IPTRANS: *INTEGRATIVE PLATFORM FOR TRANSLATIONAL RESEARCH*

5.1. CONSIDERAÇÕES INICIAIS

A partir do *framework* projetado foi implementada a ferramenta computacional IPTrans (*Integrative Platform for Translational Research*). Neste capítulo será discutida a arquitetura da plataforma assim como as tecnologias utilizadas durante o seu desenvolvimento. Tanto o *framework* quanto a ferramenta foram desenvolvidos para serem flexíveis o suficiente para gerenciar os diferentes tipos de dados clínicos e biomoleculares, porém, durante o desenvolvimento e principalmente na determinação da ontologia de referência, focamos a sua aplicação no domínio específico da oncogenômica, mais precisamente em câncer de cabeça e pescoço.

5.2. ARQUITETURA DA PLATAFORMA

A plataforma IPTrans foi desenvolvida em uma arquitetura de quatro camadas: camada de dados, camada semântica, camada de aplicação e camada de apresentação (Figura 8). Na camada de dados, o modelo Chado está sendo utilizado como o modelo de dados genômico. Como parte desta proposta, foi criado um novo módulo para representar a informação clínica (*Clinical Module* – CM). A camada semântica é composta pelas ontologias de domínio, utilizadas para anotar a informação biológica, pelos esquemas conceituais das bases de dados clínicas que representam a estrutura dessas fontes de dados e pela ontologia de referência que atua como um framework conceitual. A camada de aplicação é composta por um conjunto de módulos (bibliotecas de software) responsáveis pelo gerenciamento das bases de dados clínicas, ontologias e do processo de integração. A interação entre o sistema e o usuário é feita por meio da camada de apresentação, e para isso foi utilizado o Framework MVC (*Model-View-Controller)* Catalyst.

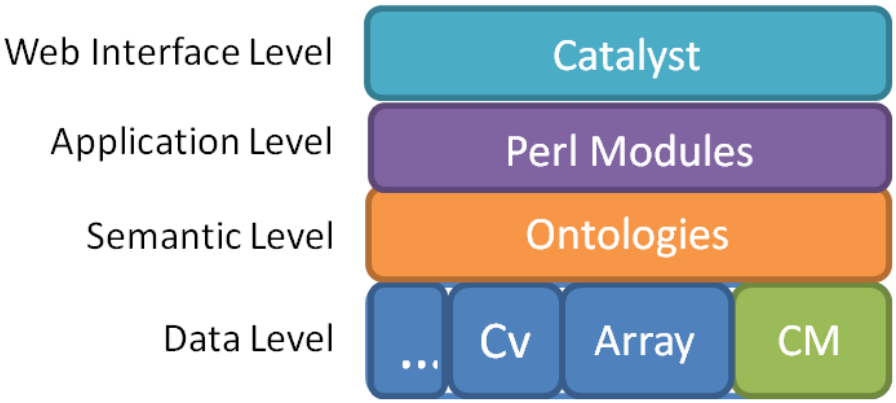

FIGURA 8. CAMADAS QUE COMPÕE A PLATAFORMA DE INTEGRAÇÃO DE DADOS CLÍNICOS E BIOMOLECULARES.

Também está sendo proposta uma metodologia de migração para ser aplicada em bancos de dados clínicos legados, assim como um processo de mapeamento ontológico que permite a padronização dos dados, integração e desenvolvimento de ferramentas de análises genéricas.

CAMADA DE DADOS

A camada de dados é a responsável pelo armazenamento da informação. Ela é composta pelo sistema gerenciador de bancos de dados relacional que implementa dois modelos de dados genômico e clínico. O modelo de dados Chado foi escolhido para representar dados biomoleculares, mas não possibilita a representação de dados clínicos. Para isso foi proposto o Módulo Clínico que é implementado nesta camada para representar dados clínicos e sócio-demográficos.

CHADO

Mungal, Emmert e o grupo FlyBase propuseram um esquema modular baseado em ontologias para representar informação biológica denominado Chado (MUNGALL; EMMERT; THE FLYBASE CONSORTIUM, 2007). Chado é um esquema de banco de dados relacional que pode ser utilizado como base para qualquer grupo de pesquisa em genômica. Compõe o projeto GMOD (*Generic Model Organism Database*) (O'CONNOR et al., 2008) e atualmente é utilizado por diversos grupos de pesquisas tais como: XenDB, ParameciumDB, AphidBase, BeetleBase, dentre outros.

Chado é composto por 18 módulos. Cada módulo é definido como um conjunto de tabelas, gatilhos e funções responsáveis por gerenciar informações de um subdomínio da genômica. Desses 18 módulos, 5 são os principais, fazendo parte do núcleo do Chado. Por meio desses módulos, Chado permite representar diversas informações biomoleculares tais como: sequências moleculares (*Sequence Module)*, experimentos de *microarrays* (*MAGE Module)*, resultados de análise de expressão (*Expression Module),* entre outros. *Chado* é extensível, pois permite a incorporação de novos módulos e, se necessário, a alteração de módulos existentes.

Um diferencial do Chado em relação a outros modelos de bancos de dados genéricos, tais como AceDB e Ensembl, discutidos anteriormente, é o fato do Chado fazer uso intensivo de ontologias. Ontologias têm um papel central no Chado, pois toda informação armazenada deve estar relacionada a alguma ontologia ou vocabulário controlado. Algumas ontologias já vêm incorporadas ao Chado tais como a Sequence Ontology que é utilizada pra descrever tipos de sequências moleculares e o OBO Relation Ontology que é uma ontologia utilizada pra descrever relacionamentos. Mas é possível incorporar novas ontologias descritas em OWL (*Web Ontology Language*) ou no formato OBO (*Open Biomedical Ontologies*).

Existem ferramentas computacionais que são compatíveis com bancos de dados Chado. Essas ferramentas são na maioria disponibilizadas pelo grupo GMOD. Podemos citar o GBrowse e o Apollo. Chado também permite a incorporação de outras ferramentas por meio da criação de *Bridge Layers* ("Camadas Ponte") que são visões construídas que funcionando como camadas de compatibilidade com outras ferramentas. Chado não possui um módulo para armazenar informações clínicas e sócio demográficas.

CAMADA SEMÂNTICA

A Camada Semântica é composta por um conjunto de ontologias, mais os esquemas conceituais dos bancos de dados. De acordo com Rubin et al. (RUBIN; SHAH; NOY, 2008), podem ser classificadas como ontologias uma grande variedade de artefatos informacionais, tais como: terminologias, thesaurus, vocabulários controlados até modelos conceituais

formalmente definidos que permitam a realização de inferência. Os componentes da Camada Semântica podem ser classificadas em três tipos:

- Esquema Conceitual de Bancos de Dados Clínico: esses modelos descrevem a estrutura de um banco de dados clínico. Para isso, é necessário representar as tabelas e suas colunas correspondentes. O esquema é representado como uma hierarquia simples e possui três níveis: o primeiro nível é elemento raiz, um elemento genérico que, por convenção, possui o nome do banco de dados clínico; o segundo nível é composto por conceitos que representam as tabelas; no terceiro nível são representadas as colunas interligadas com suas respectivas tabelas (Figura 9).
- Ontologias de Domínio: ontologias que representam conceitos de um domínio de interesse específico, por exemplo: CID 10, SNOMED, Gene Ontology. A ontologia de referência, utilizada para integrar a informação clínica, pode ser composta por uma ou mais ontologias de domínio.
- Ontologia de Referência: é uma ontologia utilizada para integrar a informação clínica de diferentes Modelos de Bancos de Dados Clínicos. Ela é composta por uma ou mais ontologias de domínio. Essa ontologia é utilizada como um arcabouço conceitual onde a informação é integrada utilizando mapeamentos ontológicos entre conceitos de um modelo de banco de dados clínico e a ontologia de referência.

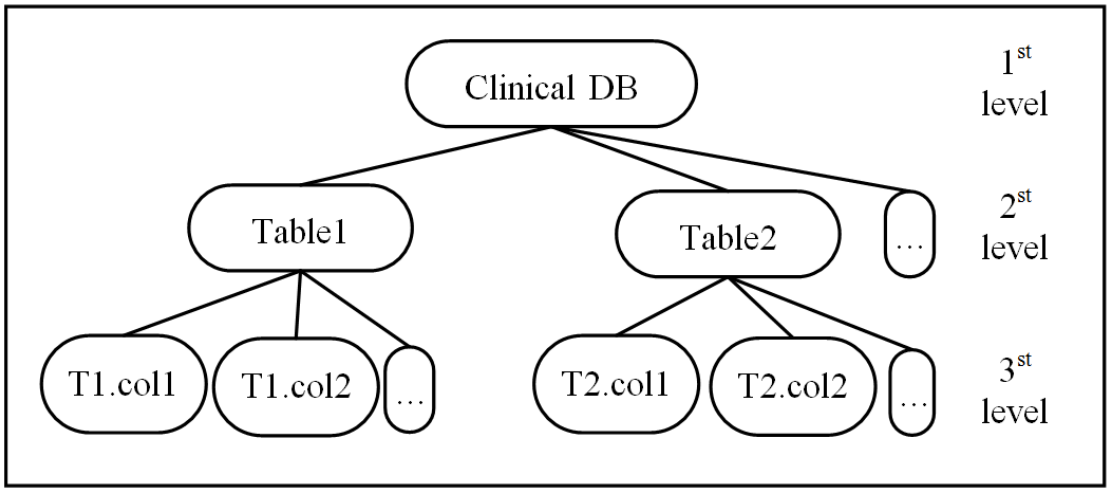

FIGURA 9. ESQUEMA CONCEITUAL DE UM BANCO DE DADOS CLÍNICO

CAMADA DE APLICAÇÃO

A camada de aplicação é composta por um conjunto de módulos (bibliotecas de software) responsáveis por criar, atualizar, recuperar e gerenciar a informação. Como linguagem de programação foi utilizado Perl. Os principais módulos implementados são: o Módulo de Importação de Dados, Módulo de Combinação de Esquemas e o Módulo de Busca Integrada. Além disso, estendemos o pacote Bio::Chado::Schema para se adaptar ao novo Módulo Clínico desenvolvido. Este pacote é responsável por fazer o mapeamento objeto-relacional.

CAMADA DE APRESENTAÇÃO

Catalyst é o Framework MVC (*Model View Controller)* em Perl para construção de aplicativos web (JOHN, 2009). É possível projetar e implementar aplicativos web de uma forma modular, testável e de fácil manutenção. Para implementar a camada de apresentação foi utilizado Catalyst. A interação entre o usuário e o sistema é realizada pela interface web. Além de permitir o gerenciamento da informação clínica e sócio demográfica, essa aplicação também permite o gerenciamento de projetos, usuários, ontologias, plataformas e experimentos de *microarrays* assim como amostras biológicas.

MÓDULO CLÍNICO PROPOSTO

O modelo de dados proposto neste trabalho é um novo módulo do Chado exibido na Figura 10. A semântica do dado clínico armazenado nesse módulo é dada por uma ontologia armazenada no módulo *Controlled Vocabulary* (CV) do Chado. Esta ontologia pode ser qualquer ontologia do domínio biomédico que represente os conceitos relativos aos dados clínicos.

O módulo clínico proposto é suficientemente genérico para representar bancos de dados clínicos de pesquisas. As tabelas em rosa são do módulo '*General*' do Chado, as tabelas em verde pertencem ao módulo *'Controlled Vocabulary'* e as tabelas em azul compõe o Módulo Clínico.

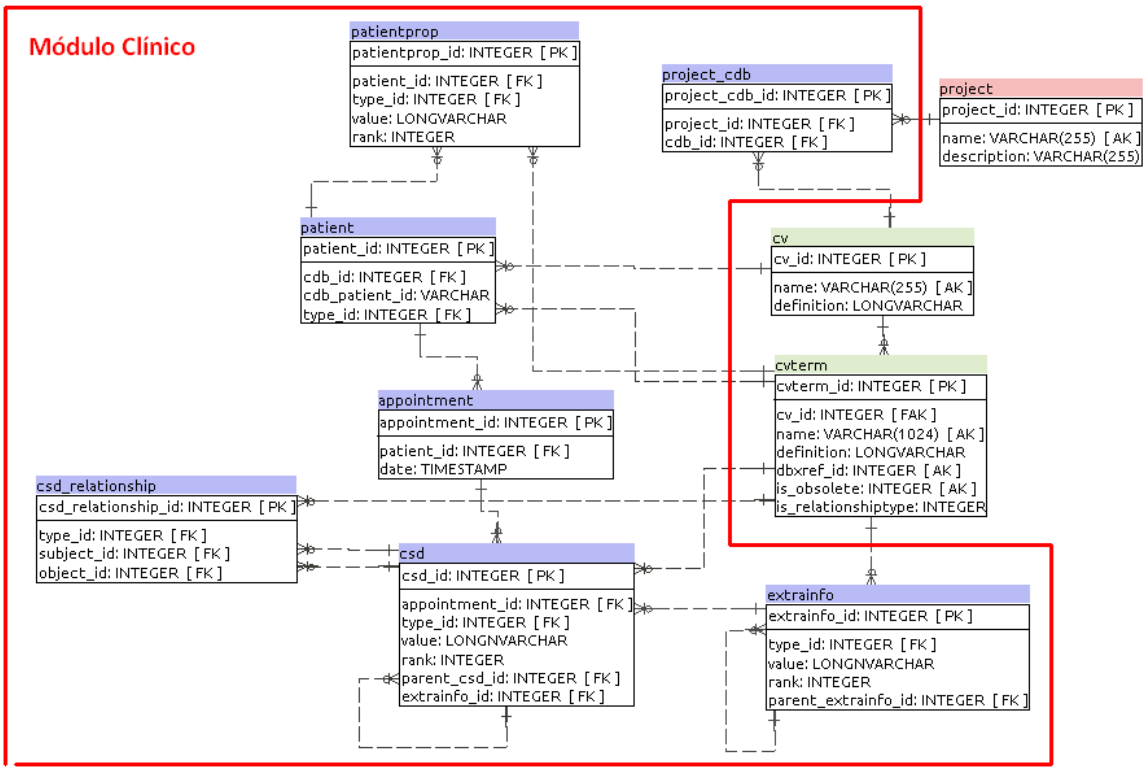

FIGURA 10. MÓDULO CLÍNICO

O Módulo Clínico é composto por 7 tabelas: *patient, patientprop, appointment, project_cdb, csd, csd_relationship e extrainfo.* A tabela *patient* é autoexplicativa, onde são armazenados os dados relacionados aos pacientes. Chado já possui uma tabela *project*, onde um projeto define um contexto no qual são agrupados um conjunto de informações relacionadas, por exemplo à conjunto de experimentos em um estudo. Cada paciente pertence a um banco clínico (clinical database – cdb) e cada banco clínico pode estar vinculado a vários projetos. Para isso foi criada a tabela *project_cdb*.

As informações clínicas ou sócio-demográficas que não se alteram com o tempo ou cujo um registro temporal não precise ser necessariamente mantido, tais como nome, RG, CPF, sexo, são armazenadas na tabela *patientprop.* Para as informações em que é necessário ter uma data associada é utilizada a tabela *appointment* para agregar essas informações em um único momento temporal (obtidas, por exemplo, em uma nova consulta clínica ou retorno do paciente).

Os tipos de informação clínica e sócio demográfica tais como idade, peso, tamanho do tumor, tipo do tumor, são representados em uma ontologia que é armazenada no módulo *Controlled Vocabulary*, mais especificadamente nas tabelas *cv* e *cvterm*.

A tabela *csd* (*clinico-social data*) é o local onde a maioria da informação é armazenada. Esta tabela foi projetada para representar, de uma forma flexível, qualquer tipo de informação clínica ou sócio demográfica relacionada a um paciente. A semântica da informação clínica é dada pela coluna *type_id,* que é uma chave estrangeira da coluna *cvterm_id* da tabela *cvterm* que, por sua vez, armazena os termos de uma ontologia que representa os tipos de informação clínica e sócio demográfica. A coluna *value* representa o conteúdo da informação. A coluna *rank* é utilizada quando é necessário armazenar um mesmo tipo de informação para o mesmo paciente. Para cada instância da informação, a coluna *rank* recebe um novo valor. Outra coluna importante é a *parent_csd* que representa um auto relacionamento. Esta coluna é utilizada quando determinada informação clínica está relacionada com outra informação do mesmo paciente.

A tabela *extrainfo* é responsável pelo armazenamento da informação que é independente do paciente. Normalmente estes tipos de dados estão em tabelas que são referenciadas por chaves estrangeiras na tabela paciente ou em qualquer tabela dependente do paciente, por exemplo, informações sobre cidades, medicamentos, hospitais, procedimentos, etc. São o tipo de informação que existe independentemente de um paciente.

A tabela *csd_relationship* é utilizada quando é necessário representar relacionamentos complexos entre dados clínicos ou sócio-demográficos. Nessa tabela é possível relacionar duas informações clínicas, usando as colunas *subject_id* e *object_id* que são chave estrangeira da tabela *csd por meio* de um relacionamento que é dado pela coluna *type_id*, que por sua vez é chave estrangeira da tabela *cvterm*.

5.3. ONTOLOGIA DE REFERÊNCIA

A escolha da ontologia de referência é um passo importante para definição da plataforma de integração. Essa ontologia não só é utilizada como base para integração das fontes de dados, atuando como um esquema global, como também é por meio dela que o usuário define as consultas. Ferramentas de análise também a utilizam como base para consultar o banco de dados integrado. Por definir a semântica dos dados gerenciados, a escolha da ontologia de referência irá delimitar a abrangência da plataforma de integração.

Devido à grande abrangência da área de medicina translacional, decidimos concentrar os esforços no domínio da oncogenômica, mais especificadamente em câncer de cabeça e pescoço. Na escolha da ontologia de referência, foi realizado um estudo bibliográfico das ontologias utilizadas para integração de dados clínicos e ontologia do domínio da oncologia e foram analisadas em profundidade as seguintes ontologias: SNOMED, ACGT Master Ontology e Translational Medicine Ontology.

ESCOLHA DA ONTOLOGIA

SNOMED é uma ontologia generalista da área da medicina, compreendendo conceitos até de outros domínios. A *Translational Medicine Ontology* (TMO) é uma ontologia menor, contendo conceitos mais gerais da área da medicina translacional mas possuindo mapeamentos para outras ontologias mais específicas. A ACGT Master Ontology é uma ontologia do domínio do câncer, que também compreende alguns aspectos clínicos.

A enorme quantidade de conceitos e relacionamentos da SNOMED (mais de 300.000 conceitos e 1.000.000 de relacionamentos) permite uma grande cobertura do domínio médico e até domínios mais gerais. Porém essa grande quantidade também dificulta o processo de mapeamento manual e diminui a acurácia em uma ferramenta de combinação automatizada de conceitos. A restrição quanto à licença de utilização também contribuiu para que não utilizássemos a SNOMED como ontologia de referência.

Foi escolhido utilizar a versão da TMO que possui mapeamento com a ACGT Master Ontology. Dessa forma são representados conceitos mais gerais da área de medicina translacional além de uma maior especificidade no domínio da oncologia permitindo a

representação com maior granularidade da informação nesta área. Por meio da TMO é possível estender a plataforma para outros domínios da medicina translacional. Para isso é necessário realizar o mapeamento da TMO para a ontologia que descreve o domínio de interesse ou utilizar uma das 60 ontologias já mapeadas para TMO.

5.4. Pipeline de Migração dos Dados

Foi desenvolvido um *pipeline* para migração de dados de uma fonte de informação para o banco de dados integrado. O *pipeline* consiste de 5 passos (Figura 11):

Passo 1 – Criar o Esquema Conceitual do Banco de Dados Clínico

Este passo consiste na criação de um modelo para descrever a estrutura do banco de dados clínico (CDB – *Clinical DataBase*) que armazena os dados originais. Esse modelo representa de um modo único o esquema da fonte de dados clínico.

Passo 2 – Armazenar o Esquema Conceitual do Banco Clínico no Banco de Dados Integrado

O esquema conceitual do banco de dados clínico é estruturado em uma hierarquia simples, podendo ser representado da mesma forma que uma ontologia. Existem várias maneiras de armazenar uma ontologia no Chado. Isso dependerá da linguagem utilizada para representação do conhecimento. A linguagem mais comum é a OWL (*Web Ontology Language).* No domínio da biomedicina, ontologias no formato OBO (*Open Biomedical Ontology Format*) também são bastante comuns. Uma maneira simples é utilizar *scripts* em Perl fornecidos pelo grupo GMOD. O esquema conceitual do banco de dados clínico é armazenado, assim como outras ontologias de domínio e a própria ontologia de referência, principalmente nas tabelas *cv* e *cvterm* do módulo *Controlled Vocabulary.*

Passo 3 – Armazenar os Dados no Módulo Clínico

Para migrar os dados armazenados no banco de dados clínico legado para o Módulo Clínico no Chado, é necessário planejar um processo de ETL (*Extraction, Transformation and Load*). Neste passo, é importante manter uma correta "tipagem" da informação de acordo com o modelo do banco clínico armazenado no módulo *Controlled Vocabulary.* Em outras palavras,

é necessário corretamente relacionar a informação armazenada no Módulo Clínico com o respectivo termo no esquema conceitual do banco clínico.

PASSO 4 – MAPEAR O ESQUEMA CONCEITUAL DO BANCO CLÍNICO COM A ONTOLOGIA DE REFERÊNCIA

Neste passo é realizado o mapeamaento entre os conceitos que compõe o esquema conceitual do banco clínico sendo integrado com conceitos da ontologia de referência. Esse mapeamento deve representar um *match* exato, de forma que o conceito na ontologia de referência possa representar univocamente o conceito no banco clínico integrado.

PASSO 5 – CRIAR A CAMADA DE COMPATIBILIDADE COM O BANCO CLÍNICO

O próximo passo é a criação da camada de compatibilidade com o banco de dados clínico. A camada de compatibilidade consiste em um conjunto de visões que representam a estrutura do banco de dados clínico por meio do Chado. A vantagem de criar uma Camada de Compatibilidade (do inglês *Bridge Layer*) é para facilitar a consulta e a adaptação de ferramentas computacionais legadas, ou seja, que foram projetadas originalmente para o banco clínico, para trabalhar corretamente e sem modificações pelo intermédio do Chado.

A metodologia de migração proposta aqui pode ser utilizada para adaptar bancos de dados legados para o framework proposto. Essa metodologia pode ser aplicada em dados armazenados em bancos de dados relacionais, arquivos de valores separados por vírgulas e *dumps* de bancos de dados em SQL. A integração ocorre quando os modelos das fontes clínicas são mapeados para uma ontologia de referência comum.

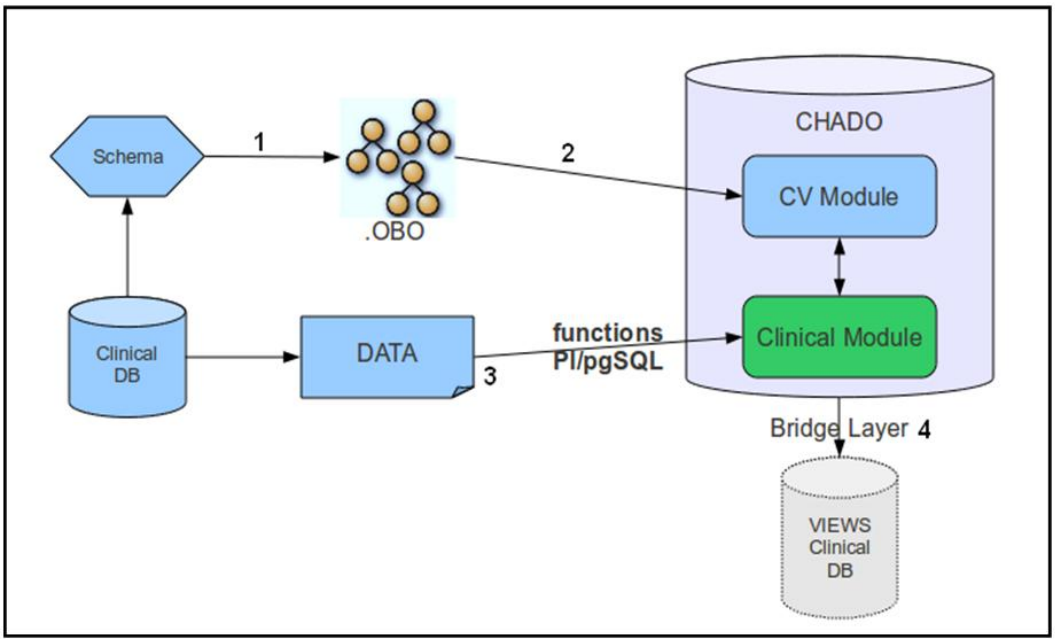

FIGURA 11. METODOLOGIA PARA MIGRAÇÃO DE BANCOS DE DADOS CLÍNICOS

5.5. MÓDULO DE GERENCIAMENTO DE USUÁRIOS E PROJETOS

O acesso aos bancos de dados clínicos e aos experimentos de *microarrays*, assim como às ferramentas de análises, são feitas pelos projetos (Figura 12). A hierarquia de acesso do usuários é definida da seguinte forma: no âmbito do sistema existem dois tipos de usuários: administrador e usuário comum. O administrador é o único responsável por importar novos bancos clínicos, realizar o mapeamento com a ontologia de referência e criar as camadas de compatibilidade. O acesso do usuário comum a essas informações depende de seu papel dentro de um projeto. No âmbito de um projeto existem três papeis diferentes: coordenador, analista e membro. O membro só pode visualizar as informações, sem alterá-las. O analista além de visualizar, pode gerenciar os pacientes, os experimentos de *microarrays* e as amostras biológicas. Por fim, o coordenador possui as mesmas permissões do analista e também pode gerenciar os demais membros do projeto, adicionando novos, mudando papéis e removendo outros.

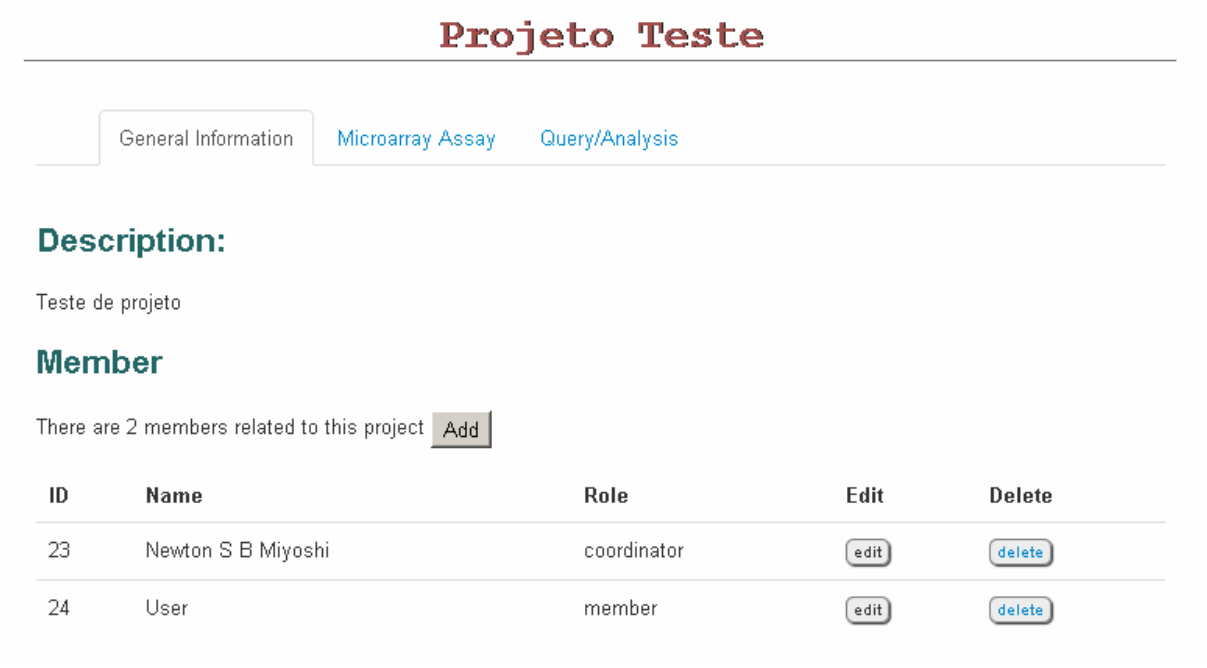

FIGURA 12. INTERFACE PARA GERENCIAMENTO DE PROJETO

5.6. MÓDULO DE GERENCIAMENTO DE AMOSTRAS BIOLÓGICAS E MICROARRAYS

Chado possui um módulo para representação de experimentos de *microarray* e amostras biológicas denominado MAGE *Module*. É formado por um conjunto de tabelas e visões para representar uma plataforma de *microarray*, algoritmos e medidas de análise de expressão gênica assim como os resultados de um determinado experimento. Porém não existe uma ferramenta computacional para gerenciamento desses experimentos e do material biológico provenientes de pacientes que utilize deste conjunto de tabelas.

Na plataforma IPTrans foi estendido o módulo MAGE do Chado e desenvolvido o Módulo de Gerenciamento de *Microarrays* e Amostras Biológicas para gerenciar plataformas e experimentos de *microarray* (Figura 13) assim como material biológico utilizado nesses experimento fazendo a ligação com os pacientes presentes no Módulo Clínico desenvolvido.

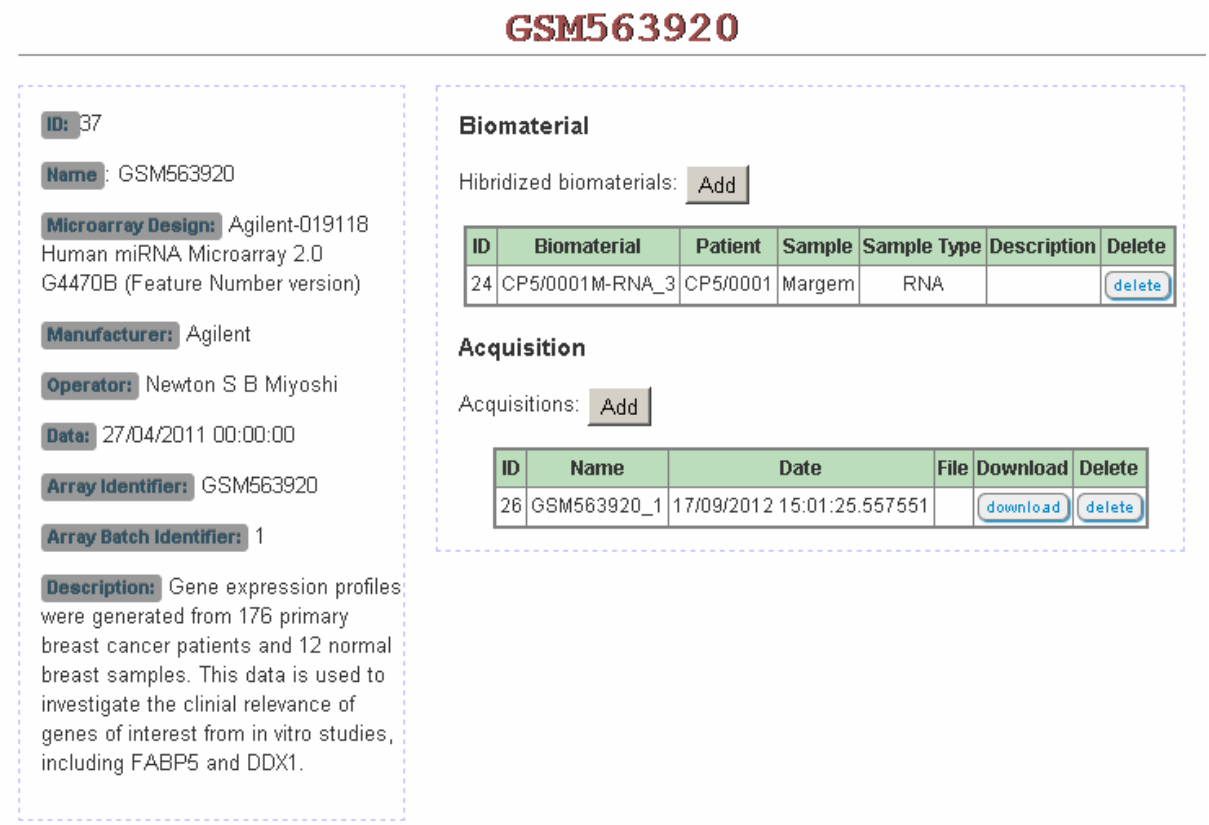

FIGURA 13. INTERFACE PARA GERENCIAMENTO DE UM EXPERIMENTO DE MICROARRAY

Um dos maiores bancos de dados públicos de experimentos de *microarray* é o GEO (EDGAR R; M; AE., 2002) (*Gene Expression Omnibus*) do NCBI. Por meio dele é possível obter tanto o design das plataformas de *microarray* quanto milhares de experimentos que utilizam essas plataformas. O GEO possui dois padrões de arquivos principais: SOFT (*Simple Omnibus Format in Text*)(NCBI, 2012) e MINiML (MIAME *Notation in Markup Language*)(NCBI, 2012). Tanto o SOFT quanto MINiML representam o mesmo conjunto de informações que variam desde o organismo em estudo, protocolos utilizados, o tipo do estudo, até a quantidade de elementos analisados, fabricante e a tecnologia utilizada. A diferença entre os formatos está na sintaxe do arquivo: SOFT define um arquivo tabular, com um cabeçalho seguindo um determinado padrão de caracteres que é facilmente analisado por scripts em linguagem de programação tais como Perl, Python e Java. O formato MINiML é o SOFT traduzido em XML. Possui o mesmo conjunto de característica, porém, por ser feito em XML, pode ser analisado e editado por diversos programas desenvolvidos para trabalhar com XML. O GEO também permite o download de dados

brutos dos experimento de *microarray* assim como de arquivos no formato do software utilizado que dependerá da plataforma e do fabricante do *microarray,* por exemplo, é comum a utilização do software Agilent Feature Extraction para *microarrays* da Agilent.

O Módulo de Gerenciamento de *Microarrays* e Amostras Biológicas permite a importação de plataformas e experimentos de *microarrays* nos formatos SOFT do GEO, e no formato específico do software Feature Extraction da Agilent. Sua arquitetura foi projetada para permitir que outros formatos de arquivo sejam importados. O fluxo de dados e os principais componente do módulo estão demonstrados na Figura 14.

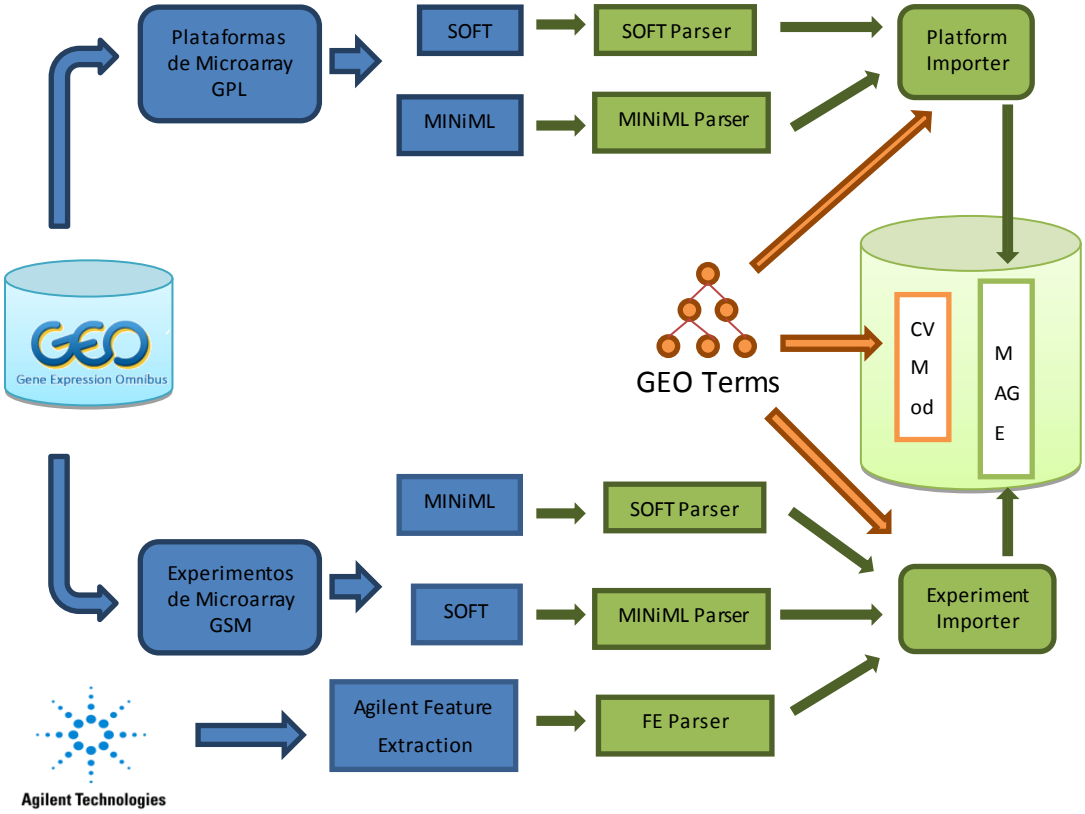

FIGURA 14. PIPELINE PARA IMPORTAÇÃO DE PLATAFORMAS E EXPERIMENTOS DE MICROARRAYS

Os principais componentes são: *Platform Importer, Experiment Importer* e o vocabulário controlado denominado *GEO Terms.* Este vocabulário controlado foi criado a partir de conceitos importantes do GEO e é utilizado para anotar semanticamente os

experimentos e as plataformas de *microarray* facilitando a sua importação e exportação para o GEO. *Platform Importer* é o componente responsável por realizar a importação das plataformas de *microarray* para o módulo MAGE. A importação das informações de um experimento de *microarray* é realizada pelo componente *Experiment Importer*. Ambos utilizam *parsers* para importar arquivos em diferentes formatos tais como SOFT e o *Feature Extraction* da Agilent. Cada parser implementa uma interface de objeto em comum sendo facilmente estendido para outros tipos de formato de arquivos.

Foi construída uma interface para cadastro, busca e gerenciamento das amostras biológicas (Figura 15). Também foi desenvolvido um pequeno vocabulário controlado para realizar anotações das amostras biológicas. Esse vocabulário foi construído a partir do banco de amostras biológicas do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto. Este módulo foi projetado para se adaptar, caso novos termos sejam adicionados ou retirados conforme a necessidade de se anotar informações adicionais. Essa adaptação ocorre tanto no nível do banco de dados quanto no nível da interface do usuário, de forma automática.

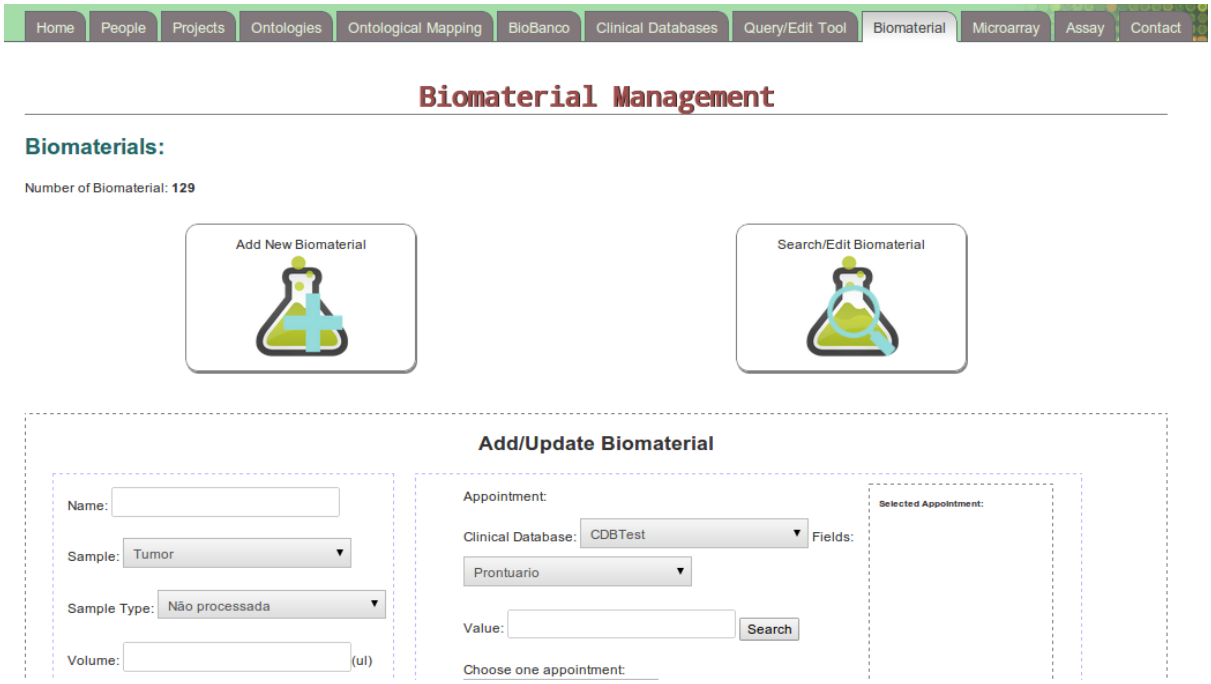

FIGURA 15. INTERFACE PARA GERENCIAMENTO DE AMOSTRAS BIOLÓGICAS

5.7. MÓDULO DE CONSULTA DE DADOS CLÍNICOS E BIOMOLECULARES

O Módulo de Consulta foi implementado no IPTrans permitindo três tipos gerais de consultas: consulta aos dados clínicos, aos dados biomoleculares e integrando dados clínicos e biomoleculares. A consulta aos dados clínicos pode ser feita de duas formas diferentes: (i) consulta simples que é realizada apenas para uma das fontes de informação; e (ii) a consulta integrada que utiliza da ontologia de referência para consultar todas as fontes de informação representadas. A vantagem da consulta simples é poder utilizar todas as possibilidades de termos para realizar a busca, porém buscando especificadamente em apenas uma fonte. Na busca integrada será consulta é feita em termos da ontologia de referência que representa o denominador comum entre as fontes. A vantagem é poder buscar em todas as fontes de informação integradas. A Figura 16 mostra a interface web criada para realizar a busca no conjunto de pacientes.

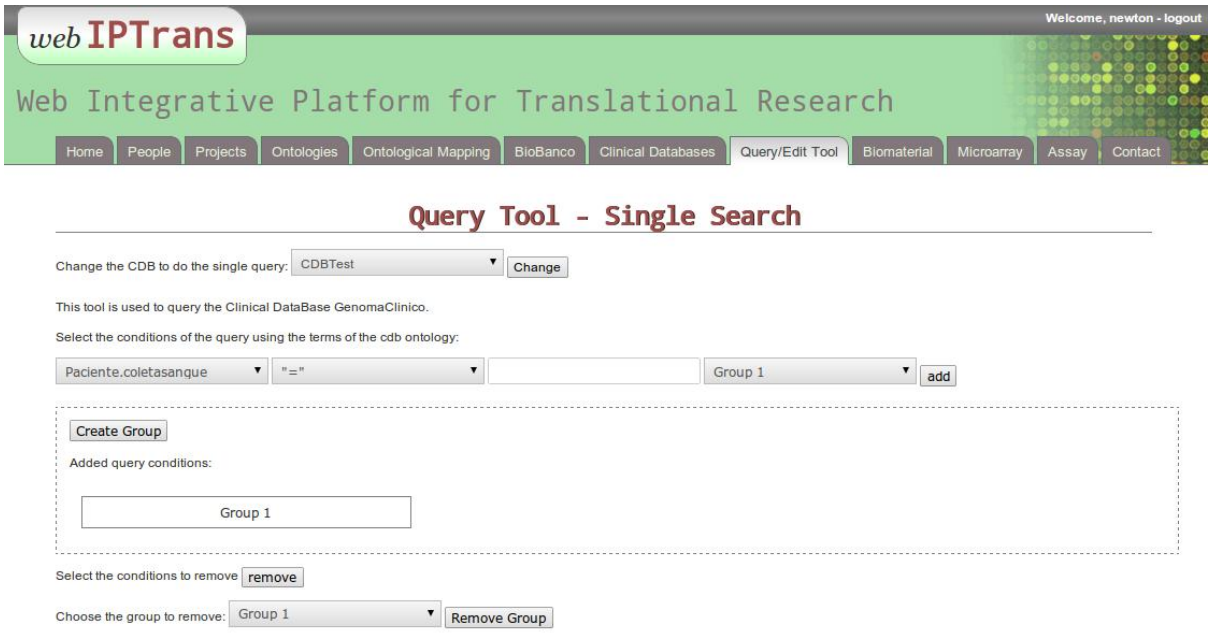

FIGURA 16. APLICATIVO WEB

A busca é realizada criando grupos de filtros. Por exemplo, para buscar tanto pacientes fumantes, do sexo masculino, acima de 50 anos quanto mulheres ex-fumantes cria-se dois grupos utilizando dos termos do esquema conceitual do banco de dados clínico como demonstrado na Figura 17.

Group 1

(105472) Paciente.sexo = M

AND

(105479) Paciente.idade > 50

AND

(105497) Paciente.fuma > Sim, ainda fuma

OR Group 2

(105472) Paciente.sexo = F

AND

(105497) Paciente.fuma = Somente no passado

FIGURA 17. EXEMPLO DE BUSCA UTILIZANDO-SE GRUPOS DE FILTROS

A consulta aos dados biomoleculares foi implementada para os experimentos de *microarray* (Figura 18). A partir da seleção de um experimento de *microarray* é possível definir uma lista de genes de interesse e verificar qual o nível de expressão desses genes. Também é possível definir um limiar e verificar quais genes, dentre todos os analisados, ou a partir de uma lista de interesse, tiveram nível de expressão maior que o limiar definido.

GSM563920_1

Details

ID: 26 **Name:** GSM563920_1 **Date:** 17/09/2012 15:01:25.557551 **Assay:**

Quantifications

Normalized signal intensity | value | Enter Gene List | Filter

Total: **41359**

ID	Spot ID	Feature	Normalized signal intensity
41818	A_23_P105747(APPL2)	85567	1.24
41819	A_23_P105753(PWP1)	85569	1.46
41820	A_23_P105794(EPSTI1)	85571	5.98
41821	A_23_P105803(FGF9)	85573	1.43
41822	A_23_P105815	85574	2.07
41823	A_23_P105833(BIVM)	85576	0.56
41824	A_23_P105856(SLC46A3)	85578	1.04
41825	A_23_P105862(FRY)	85580	0.31
41826	A_23_P105873(TRPC4)	85582	0.6
41827	A_23_P105886(C1QTNF9B)	85584	0.21

Page 1 of 4136 | Next Page

FIGURA 18. CONSULTA DO NÍVEL DE EXPRESSÃO GÊNICA EM UM EXPERIMENTO DE MICROARRAY

A consulta integrando dados clínicos e biomoleculares pode ser realizada de duas formas diferentes. A primeira abordagem consiste em selecionar um conjunto de experimentos de *microarray*, possivelmente definir uma lista de genes de interesse e um limiar para o nível de expressão gênica, e o conjunto de dados clínicos de interesse (Figura 19).

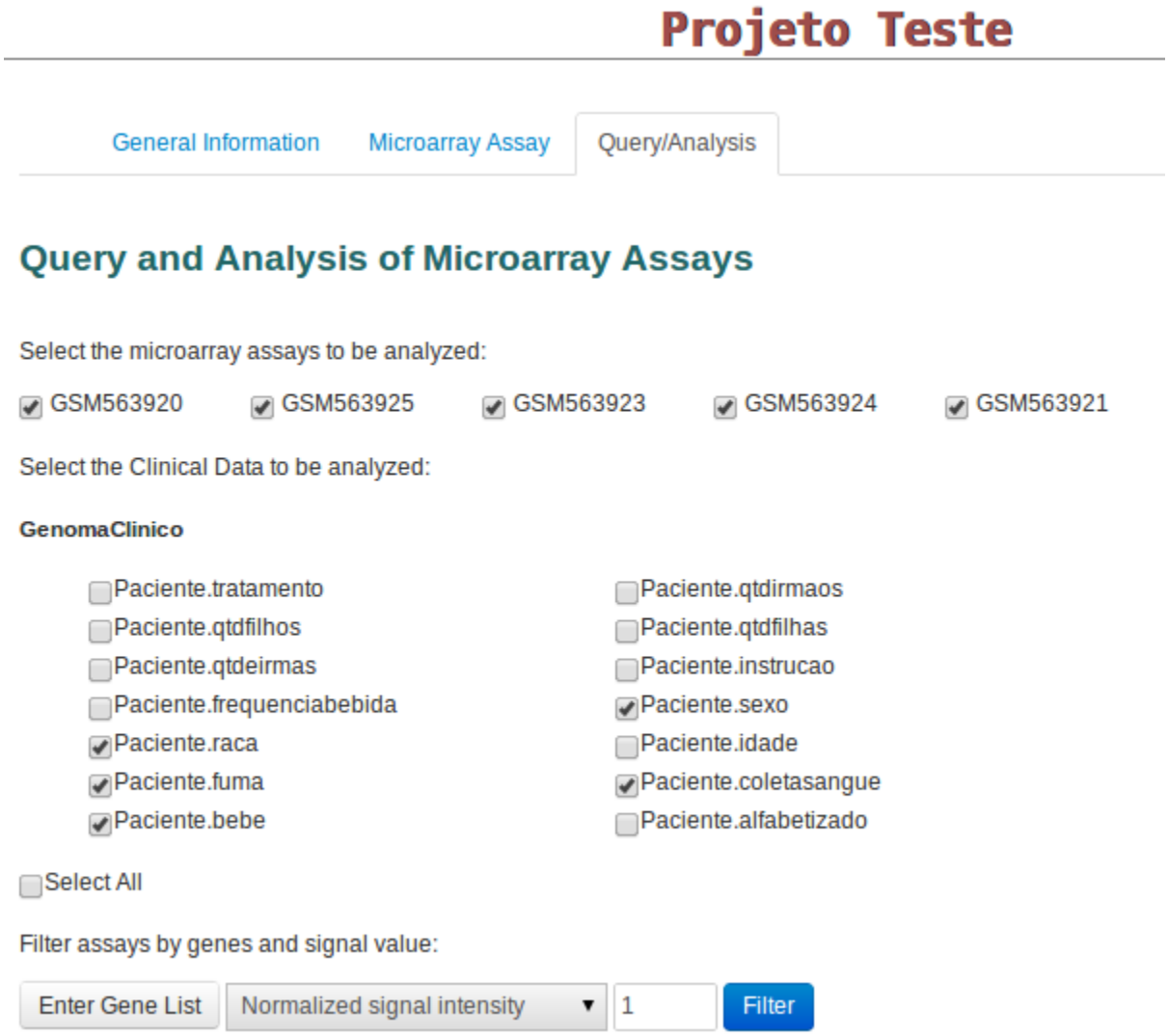

FIGURA 19. INTERFACE PARA CONSULTA DE DADOS CLÍNICOS E BIOMOLECULARES

A partir dessa entrada o sistema irá selecionar os experimentos cujos genes de interesse tiveram um nível de expressão maior que o limiar e retornar a lista dos pacientes com os dados clínicos selecionados e um conjunto de gráficos mostrando a distribuição dos valores desses dados clínicos (Figura 20).

Clinical Analysis

Patients Count: 5 Appointments Count: 9 Clinical Count: 9

Delete	View/Edit	Patient ID	Appointment ID	Paciente.fuma	Paciente.bebe	Paciente.frequenciabebida	Paciente.coletasangue	Paciente.sexo	Paciente.raca
del	edit	8903	13782	Sim, ainda fuma	So no passado	Ambos	Sim	M	Branco
del	edit	8903	13783	Sim, ainda fuma	So no passado	Ambos	Sim	M	Branco
del	edit	8903	13784	Sim, ainda fuma	So no passado	Ambos	Sim	M	Branco
del	edit	8904	13785	Sim, ainda fuma	So no passado	Ambos	Sim	M	Mulato
del	edit	8905	13786	Sim, ainda fuma	Sim, ainda bebe	Ambos	Sim	M	Branco
del	edit	8906	13787	Sim, ainda fuma	Sim, ainda bebe	Entre as Refeicoes	Sim	M	Mulato
del	edit	8906	13788	Sim, ainda fuma	Sim, ainda bebe	Entre as Refeicoes	Sim	M	Mulato
del	edit	8906	13789	Sim, ainda fuma	Sim, ainda bebe	Entre as Refeicoes	Sim	M	Mulato
del	edit	8909	13804	Sim, ainda fuma	Sim, ainda bebe	Ambos	Sim	M	Branco

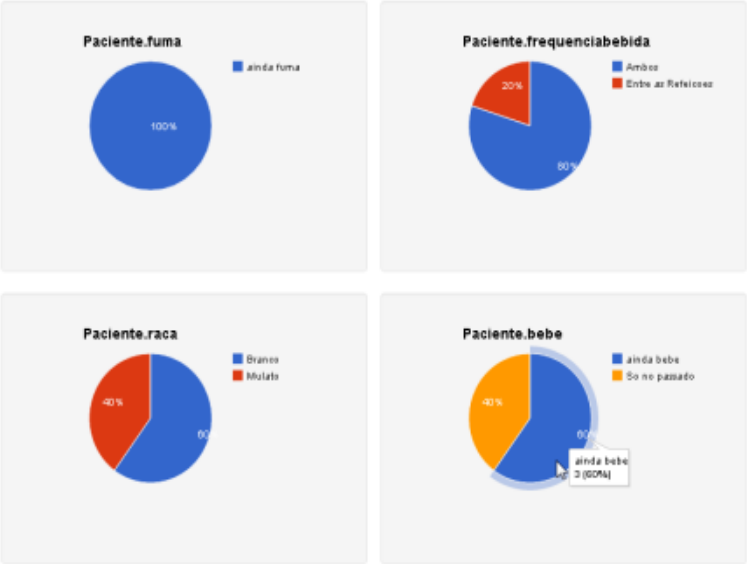

FIGURA 20. RESULTADO DA CONSULTA INTEGRANDO DADOS CLÍNICOS E BIOMOLECULARES A PARTIR DE INFORMAÇÕES DE MICROARRAY

Esse mesmo tipo de consulta integrando dados clínicos e biomoleculares pode ser realizada da maneira inversa. A partir da definição de um conjunto de informações clínicas de interesse, de maneira análoga à demonstrada na Figura 17 e da seleção de um grupo de pacientes, o sistema retorna ao usuário uma lista dos genes que tem um maior nível de expressão. Estas análises só serão possíveis se todos os paciente fizerem parte de um mesmo estudo, ou seja, se a plataforma de *microarray* e os protocolos utilizados forem os mesmos.

Query Biomolecular Data

Patient appointments query: 9

Query microarray assays: 5

GSM563923 - GSM563920 - GSM563925 - GSM563921 - GSM563924 -

Analysis

Gene Name	Mean	Variance	Stand Dev	Min	Max	Box-Plot
APPL2	1.378	0.0995699999999995	0.315547143862846	1.1	1.81	
PWP1	1.186	0.298582222222222	0.546426776633633	0.56	2.1	
EPSTI1	1.61	3.6144	1.90115754213058	0.37	5.98	
FGF9	1.384	1.20923	1.09964994430046	0.27	2.56	
BIVM	0.89	0.2323	0.481975103091435	0.56	1.72	
FRY	1.1676	0.257293999999999	0.507241559811496	0.31	2.36	
TRPC4	1.131	0.161987777777778	0.40247705248595	0.6	1.68	

FIGURA 21. RESULTADO DA CONSULTA INTEGRANDO DADOS CLÍNICOS E BIOMOLECULARES A PARTIR DE INFORMAÇÕES CLÍNICAS

5.8. CONSIDERAÇÕES FINAIS

Neste capítulo apresentamos a ferramenta IPTrans (*Integrative Platform for Translational Research*). Esta ferramenta é a implementação do *framework* conceitual apresentado no capítulo 4 para o domínio da medicina translacional, mais especificamente para oncologia. Foram implementados os módulos de integração de esquema, de importação de dados, de consulta integrada e a camada de compatibilidade. O Chado foi utilizado como base para os dados biológicos e propusemos um novo módulo clínico para representação das informações dos pacientes que segue o modelo EAV. Através deste modelo podemos integrar dados não conhecidos *a priori* devido à sua flexibilidade. Propusemos a utilização da TMO com mapeamento para ACGT-MO como ontologia de referência, pois, apesar da sua simplicidade comparada à SNOMED, ela abrange conceitos gerais da área de medicina translacional e também conceitos específicos da área de oncologia. O módulo de integração

de esquemas foi implementado de forma a permitir o mapeamento manual entre conceitos dos esquemas das fontes de dados com a ontologia de referência. A utilização da ontologia de referência como um esquema mediator é importante pois, ao adicionar novas fontes de dados, não é necessário realizar o mapeamento para todas as outras fontes, é necessário mapear apenas para o esquema mediator. A menor quantidade de conceitos da TMO também auxilia no processo de mapeamento manual. O módulo de integração de esquemas também foi projetado para permitir uma futura implementação do processo de combinação automática ou semi-automática dos esquemas. Além das funcionalidades principais propostas no *framework* conceitual também foram desenvolvidos módulos adicionais para gerenciamento de usuários, projetos, microarrays e amostras biológicas.

6. VALIDAÇÃO DO FRAMEWORK

6.1. CONSIDERAÇÕES INICIAIS

Neste capítulo é discutido a aplicação da ferramenta em um conjunto de dados reais buscando validar o projeto do *framework* e a plataforma computacional desenvolvida a partir deste. Esses conjuntos de dados são de pacientes com câncer de cabeça e pescoço provenientes do Hospital das Clínicas de Ribeirão Preto e do Instituto Ludwig de São Paulo.

6.2. PROJETO "ONCOGENÔMICA APLICADA À TERAPIA DE CABEÇA E PESCOÇO"

O projeto "Oncogenômica Aplicada à Terapia de Carcinoma de Cabeça e Pescoço" da Rede GENOPROT (CNPq) tem como objetivo realizar uma pesquisa conjunta focada na análise de mecanismos genéticos e epigenéticos responsáveis pela regulação do transcriptoma e secretoma de carcinomas de cabeça e pescoço, buscando identificar biomarcadores para diagnóstico, prognóstico e que poderiam ser utilizados por alvos terapêuticos.

Por meio deste projeto foram obtidas duas fontes de dados clínicos diferentes. A primeira consiste no banco de dados relacional do Projeto Genoma Clínico do Instituto Ludwig com apoio da FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo). Esse banco de dados possui 20 tabelas com algumas possuindo até 120 colunas (Figura 22). O sistema gerenciador do banco de dados relacional é o MySQL (ORACLE, 2010). Nesse banco estão cadastrados 423 pacientes.

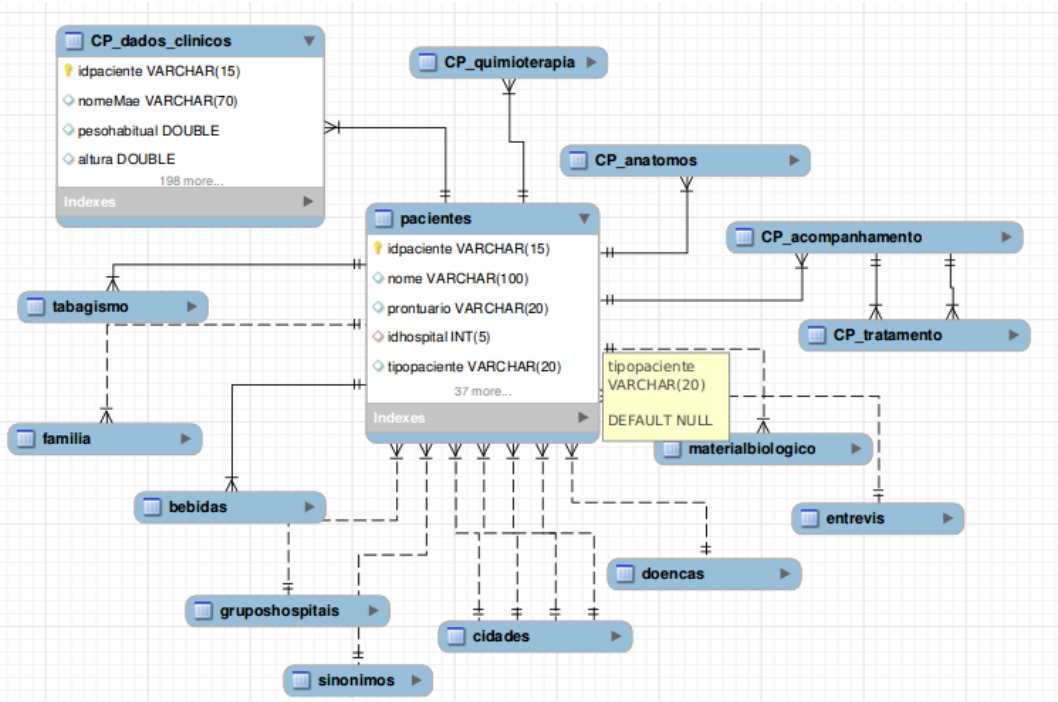

FIGURA 22. PARTE DO BANCO DE DADOS DO PROJETO GENOMA CLÍNICO

A segunda fonte de informação é uma planilha exportada a partir do banco de dados do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HC-FMRP). Essa planilha contém 40 colunas e cada linha representa um atendimento de um paciente. No total são 16384 atendimentos (Figura 23).

Instituto	Tp.Atend	Tipo de Atendimento	Convênio	Dt.	Dt.
CA	EF	ENFERMARIA	SUS	21/	24/0
CA	EF	ENFERMARIA	SUS	17/	18/0
CA	AB	AMBULATORIO	SUS	04/	04/0
CA	EF	ENFERMARIA	SUS	26/	27/0
CA	EF	ENFERMARIA	SUS	13/	18/0
UE	PS	PRONTO SOCORRO	SUS	26/	26/0
CA	AB	AMBULATORIO	SUS	20/	20/0
CA	EF	ENFERMARIA	SUS	19/	21/0
CA	EF	ENFERMARIA	SUS	29/	02/1
CA	EF	ENFERMARIA	SUS	07/	10/1
CA	EF	ENFERMARIA	SUS	05/	06/1
CA	EF	ENFERMARIA	SUS	12/	15/0
CA	EF	ENFERMARIA	SUS	07/	13/0
CA	EF	ENFERMARIA	SUS	16/	29/0

FIGURA 23. PARTE DA PLANILHA COM DADOS DE PACIENTE DO HC-FMRP

O primeiro passo da importação consiste em obter a representação do esquema conceitual das fontes de informação utilizando a linguagem de representação de

conhecimento OBO-Format. Primeiramente são fornecidas informações para acesso às fontes de dados. O IPTrans automaticamente gera o conjunto de termos que compõe o esquema conceitual e, utilizando-se da ferramenta OBO-Edit (DAY-RICHTER et al., 2007), são criados os modelos conceituais (Figura 24).

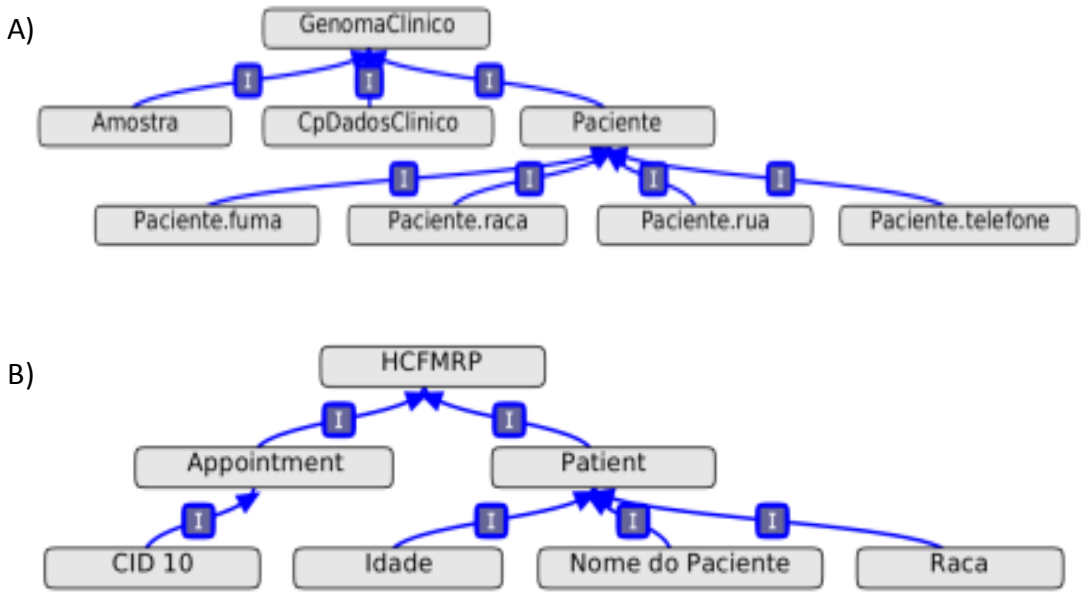

FIGURA 24. ESQUEMA CONCEITUAL DAS FONTES DO PROJETO GENOMA CLINICO (A) E DO HC-FMRP (B)

Durante o passo 2, descrito na seção 5.4 a carga do esquema conceitual representado no formato OBO foi realizada usando *scripts* em Perl provenientes do grupo GMOD sendo realizada automaticamente no IPTrans durante a sequência de importação de um banco clínico.

A Figura 25 mostra um exemplo de uma paciente no banco de dados clínico (CDB) e como essa informação é armazenada, após o carregamento do esquema conceitual do banco de dados clínicos, no módulo *Controlled Vocabulary*. O primeiro passo é representar a tabela paciente e suas respectivas colunas tais como 'idade', 'altura' e 'peso' para o esquema conceitual e então armazená-lo no módulo *Controlled Vocabulary*. Isto pode ser feito especificamente na tabela *cv*, que representa as ontologias armazenadas, e na tabela *cv_term* que representa os termos que compõe essas ontologias.

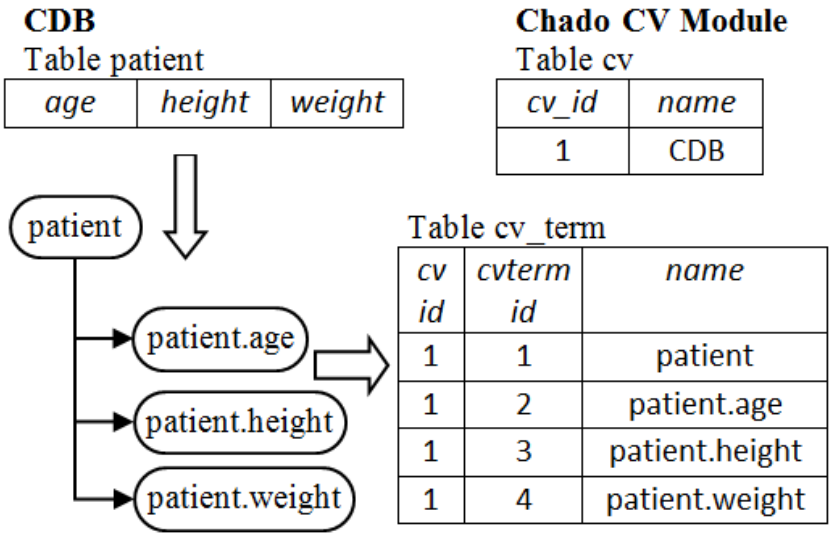

FIGURA 25 REPRESENTAÇÃO DA TABELA *PATIENT* DO BANCO CLÍNICO POR MEIO DO SCHEMA CONCEITUALD O BANCO CLÍNICO

No passo 3, um processo ETL (*Extraction, Transformation and Load*) foi realizado, consistindo na extração dos dados do banco de dados clínico, transformação da informação quando necessário e a carga desses dados no Módulo Clínico. Esse processo é semi-automático, onde o usuário precisa selecionar qual entre as tabelas é a tabela paciente para o caso de um SGBD relacional. No caso de uma planilha, é necessário definir quais as colunas são informações invariantes no tempo e quais estão relacionadas a um evento clínico. Feita essa pré-configuração a plataforma realiza a importação automática dos dados para o Módulo Clínico.

A Figura 26 ilustra como a informação original do banco de dado clínico pode ser armazenada no Módulo Clínico. Neste exemplo, os dados extraídos são idade, altura e peso de um paciente. Primeiro um registro na tabela *patient* do Módulo Clínico é criado e esse registro recebe um identificador interno (neste exemplo, seria o id "9"), nessa tabela também é armazenado o identificador do paciente no banco clínico. É criado um registro na tabela *appointment*, representando uma consulta do paciente. Então os dados clínicos e demográficos desse paciente são armazenados na tabela *csd* do Módulo Clínico, que segue o modelo EAV. O tipo do dado é determinado por meio da coluna *type_id* que é chave estrangeira da coluna *cvterm_id* da tabela *cvterm* onde são armazenados os termos do modelo conceitual do banco de dados clínico. Cada informação que é armazenada na tabela

csd é "tipada", ou seja, é semanticamente representada por um termo na ontologia do banco clínico armazenada no módulo *Controlled Vocabulary.*

CDB

Table patient

age	*height*	*weight*
11	*150*	*46*

Chado CV Module

Table cv_term

cv id	*cvterm id*	*name*
1	2	patient.age
1	3	patient.height
1	4	patient.weight

Chado Clinical Module

Table csd

csd id	*patient id*	*type id*	*value*	*rank*	*parent csd*	*extrainfo id*
1	*9*	*2*	*11*	*0*	*null*	*null*
2	*9*	*3*	*150*	*0*	*null*	*null*
3	*9*	*4*	*46*	*0*	*null*	*null*

FIGURA 26 REPRESENTAÇÃO DA INFORMAÇÃO PROVENIENTE DO BANCO DE DADOS CLÍNICO

Para representar parte da informação do banco de dados clínico foram utilizadas as colunas *rank* e *parent_csd*. A coluna *rank* recebe um valor sequencial e foi utilizada quando é armazenado o mesmo tipo de informação (mesmo *type_id*) para o mesmo paciente. Um caso seria quando é necessário representar os medicamentos usados por um paciente (suponha Carboplatina, Decadron e Furosemida). Uma maneira de realizar isso é criar três registros na tabela *csd* para o mesmo paciente e com o mesmo *type_id (cvterm* que referencia o termo "medicamento"). Dessa forma cada medicamento é marcado por um valor diferente de *rank*.

A coluna *parent_csd* é um auto relacionamento. Foi utilizada para representar a dosagem dos medicamentos. Considerando Carboplatina, um dos medicamentos que foram mencionados anteriormente, para relacionar a dosagem com o medicamento correto é

utilizada a coluna *parent_csd*. A Figura 27 ilustra a estrutura e o conteúdo da tabela *csd* com o exemplo discutido.

Chado CV Module
Table cv_term

cv id	*cvterm id*	*name*
1	10	drug
1	11	drug.dosage

Chado Clinical Module
Table csd

csd id	*appointment id*	*type id*	*value*	*rank*	*parent csd*	*extrainfo id*
5	*9*	*10*	*Carboplatin*	*0*	*null*	*null*
6	*9*	*10*	*Decadron*	*M1*	*null*	*null*
7	*9*	*10*	*Furosemide*	*2*	*null*	*null*
8	*9*	*11*	*150 mg*	*0*	*5*	*null*

FIGURA 27 EXEMPLO DA UTILIZAÇÃO DA COLUNA *PARENT_CSD*

Após a importação das fontes de dados clínicos foi realizado o mapeamento para a ontologia de referência composta pela Translational Medicine Ontology e a ACGT Master Ontology. Apesar da grande cobertura da ontologia (aproximadamente 2100 conceitos) no domínio da medicina translacional e oncogenômica, nem todos os termos das fontes puderam ser mapeados (Figura 28).

Genoma Clínico	TMO + ACTG MO	HCFMRP
Paciente.nome		Nome do Paciente
Paciente.raca	race	Raca
Paciente.sexo	sex	Sexo
Paciente.idade	age in years	Idade
Doenca.descricao	Diagnostic Result	CID10
		Convenio
Paciente.fuma	Smoker status	
Paciente.bebe		
CpDadosClinicos.peso_habitual	weight in kg	
CpDadosClinicos.altura	height in cm	

FIGURA 28. MAPEAMENTO ENTRE O AS FONTES DE DADOS CLÍNICAS E A ONTOLOGIA DE REFERÊNCIA

Finalmente, a Camada de Compatibilidade para o banco de dados clínico é construída. A construção da camada de compatibilidade consiste em realizar um processo

de pivoteamento (*pivoting*), ou seja, transformar os dados no formato EAV (*row* modelled) para o formato colunar tradicional. Esse processo pode ser feito a partir da criação de uma visão constituída de uma série de junções do tipo *FULL OUTER* com a própria tabela que segue o modelo EAV, neste caso é a tabela *csd*. Cada junção deve ser realizada para um determinado atributo que se queira transformar no formato colunar. Esse processo pode ocasionar uma grande perda de desempenho no banco de dados, devido à operação de junção ser muito custosa. Para solucionar esse problema pode-se materializar essas visões que são construídas. A ferramenta web legada construída para o banco de dados clínico poderia ser adaptada ao IPTrans por meio da Camada de Compatibilidade. O mesmo pode ser feito para outras aplicações .

A Camada de Compatibilidade poderia ser construída para outros bancos de dados utilizando partes da informação armazenadas no módulo clínico. Dessa forma outras ferramentas e aplicativos construídos para esses bancos de dados poderiam ser utilizadas sem a necessidade de alterá-los.

6.3. CONSIDERAÇÕES FINAIS

Neste capítulo apresentamos a validação do *framework* desenvolvido a partir da integração de dados de duas fontes de dados clínicos heterogêneos. Demonstramos como os dados foram importados para a tabela que segue o modelo EAV. Identificamos que o processo de importação é custoso devido ao pivoteamento das informações, porém isto é feito somente uma vez para cada uma das fontes importadas, não sendo um passo crítico para a utilização da ferramenta. Foi realizado também o mapeamento manual dos esquemas conceituais das fontes de dados para a ontologia de referência. Durante o mapeamento identificamos que houve conceitos que não puderam ser mapeados ou que não existiam em ambas as fontes. Esta diferença pode ser resolvida a partir da extenção da ontologia de referência com conceitos que não são abrangidos, se forem informações relevantes para a consulta integrada.

7. DISCUSSÃO E CONCLUSÕES

Transformar o conhecimento gerado pela ciência em um benefício real para melhora da saúde humana é um dos principais objetivos da pesquisa translacional. Para que isso aconteça, uma infraestrutura computacional é necessária permitindo o armazenamento, gerenciamento, integração e análise de ambas informações clínicas e biológicas.

O projeto descrito nesta dissertação tem como objetivo dar um passo em direção a esta infraestrutura, propondo *um framework que* permite a representação de informação clínica, sócio-demográfica e biológica em uma base de dados integrada, apoiada por um ambiente ontológico de maneira flexível e robusta. Esse *framework* foi implementado em uma plataforma computacional, denominada IPTrans, que possui uma arquitetura de quatro camadas: camada de dados, camada semântica, camada de aplicação e camada de interface de usuário.

O modelo de banco de dados biológico Chado foi estendido com a criação do Módulo Clínico que permite a representação de informação clínica e sócio demográfica de maneira genérica. O real benefício de adotar um modelo genérico para representação da informação torna-se concreto, com o surgimento de diversas aplicações e ferramentas de análise que são construídas e mantidas pela comunidade que adota esse modelo. Além disso, facilita a integração de aplicações e de troca de dados entre grupos de pesquisa e também entre pesquisadores que não adotam o Chado e talvez passem a utilizá-lo depois da extensão proposta.

A adoção do Chado como o modelo básico de banco de dados biológico permite a reutilização das ferramentas existentes construída pelo grupo GMOD ou adaptados usando a camada de compatibilidade para análise e visualização de dados moleculares. Com a proposta do Módulo Clínico, esta solução torna-se uma plataforma robusta para área de medicina translacional pois permite a representação de dados clínicos e sócio-demográficos, que antes não poderiam ser representados somente no modelo original do Chado.

Com o uso de uma abordagem ontológica, por meio da construção da camada semântica e do uso de uma ontologia de referência, é possível gerenciar e integrar tipos de

dados altamente heterogêneos, tais como dados clínicos e sócio-demográficos. A ontologia de referência serve como uma estrutura conceitual, possibilitando o mapeamento da informação clínica a partir de diferentes fontes para um esquema global. A utilização da TMO com mapeamento para a ACGT Master Ontology permite tanto a representação de conceitos gerais da medicina translacional, quanto conceitos específicos do domínio da oncologia e permite a possível extensão para outras áreas de pesquisa.

O uso desta plataforma em um conjunto de dados reais, demonstrou a viabilidade da proposta de integração, destacando suas características de flexibilidade e robustez. Por meio do *framework* de integração e da plataforma computacional desenvolvida é dado um novo passo para cumprir o *gap* tecnológico que existe entre a bancada e a prática clínica, permitindo a reutilização de ferramentas de bioinformática e também permitindo uma maneira flexível de integrar diferentes fontes de dados clínicos e sócio-demográficos.

7.1. TRABALHOS FUTUROS

Neste trabalho foi salientado a questão da representação e da integração de dados clínicos e biomoleculares. Muito precisa ser feito para suprir as necessidades computacionais da pesquisa translacional. Os trabalhos futuros foram divididos em três aspectos: do ponto de vista da integração de dados, do ponto de vista biológico e do ponto de vista clínico.

No aspecto da integração será necessário implementar o módulo de resolução de entidades que permita a identificação de uma mesma entidade em diferentes bancos de dados que são integrados. Será necessário também a implementação de algoritmos de combinação de esquemas automático ou semi-automático. Uma possível solução seria a utilização da plataforma OpenII (SELIGMAN et al., 2010) que é um sistema de código aberto e possui alguns dos principais algoritmos de combinação de esquemas. Seria possível integrar o OpenII na ferramenta IPTrans permitindo o mapeamento automático dos modelos dos bancos clínicos com a ontologia de referência.

Na parte biológica seria interessante estender a parte de *microarray* para permitir o *parsing* de outros formatos de arquivos e consequentemente a importação de outros tipos de experimentos. Um outro aspecto seria integrar a ferramenta com dados de sequenciadores de nova geração. Um outro ponto seria agregar ao IPTrans outras ferramentas de bioinformática que usam o Chado como base de dados.

No aspecto da informação clínica, seria importante estender o modulo de importação de dados para permitir dados que estejam em padrões de informação em saúde como HL7, TISS, TUSS. Também seria importante oferecer um maior suporte a imagens médicas utilizando padrão DICOM, por exemplo. Um outro aspecto seria um módulo de segurança para implementar algoritmos de anonimização e outros mecanismo de segurança como acesso via SSL (*Security-Service-Layer*).

8. REFERÊNCIAS BIBLIOGRÁFICAS

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature genetics**, v. 25, n. 1, p. 25-9, maio. 2000.

BERNSTEIN, P. A.; HAAS, L. M. Information integration in the enterprise. **Communications of the ACM**, v. 51, n. 9, p. 72, 1 set. 2008.

BODENREIDER, O.; STEVENS, R. Bio-ontologies: current trends and future directions. **Briefings in bioinformatics**, v. 7, n. 3, p. 256-74, set. 2006.

BRAZHNIK, O.; JONES, J. F. Anatomy of data integration. **Journal of biomedical informatics**, v. 40, n. 3, p. 252-69, jun. 2007.

CASANOVA, M. A. et al. Database Conceptual Schema Matching. **Computer**, v. 40, n. 10, p. 102-104, 1 out. 2007.

CHEN, R. S. et al. Exploring performance issues for a clinical database organized using an entity-attribute-value representation. **Journal of the American Medical Informatics Association : JAMIA**, v. 7, n. 5, p. 475-87, 2000.

CONSORTIUM, T. G. O. The Gene Ontology in 2010: extensions and refinements. **Nucleic acids research**, v. 38, n. Database issue, p. D331-5, jan. 2010.

DAY-RICHTER, J. et al. OBO-Edit: An ontology editor for biologists. **Bioinformatics (Oxford, England)**, v. 23, n. 16, p. 2198-200, ago. 2007.

DEVAKI. **Torque's type map**. Disponível em: <http://nextobjects.sourceforge.net/torque-type-map.html>. Acesso em: 13 fev. 2012.

EDGAR R; M, D.; AE., L. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. **NUCLEIC ACIDS RESEARCH**, v. 30, n. 1, p. 207-210, 2002.

GRUBER, T. R. A Translation Approach to Portable Ontology Specifications. **Knowledge Acquisition**, v. 5, n. April, p. 199-220, 1993.

HARRIS, P. A. et al. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. **Journal of Biomedical Informatics**, v. 42, n. 2, p. 377-381, abr. 2009.

HUBBARD, T. et al. The Ensembl genome database project. **Nucleic acids research**, v. 30, n. 1, p. 38-41, 2002.

I2B2. **i2b2: Informatics for Integrating Biology & the Bedside**. Disponível em: <https://www.i2b2.org/index.html>. Acesso em: 21 fev. 2011.

IHTSDO. **SNOMED CT ® User Guide January 2012 International Release (US English)**. [s.l: s.n.]. p. 2002-2012

JOHN, A. S. **Catalyst 5.8 The Perl MVC Framework**. 1. ed. [s.l.] Packt Publishing, 2009. p. 244

KASPRZYK, A. et al. EnsMart: a generic system for fast and flexible access to biological data. **Genome research**, v. 14, n. 1, p. 160-9, jan. 2004.

LENZERINI, M. **Data integration: a theoretical perspective**Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02. **Anais**...New York, New York, USA: ACM Press, 3 jun. 2002Disponível em: <http://dl.acm.org/citation.cfm?id=543613.543644>. Acesso em: 24 out. 2011

LOWE, H. J. et al. STRIDE--An integrated standards-based translational research informatics platform. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, v. 2009, p. 391-5, jan. 2009.

LUCIANO, J. S. et al. The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. **Journal of biomedical semantics**, v. 2 Suppl 2, n. Suppl 2, p. S1, jan. 2011.

MARTIN, L. et al. **Ontology based integration of distributed and heterogeneous data sources in ACGT**First International Conference on Health Informatics - HEALTHINF. **Anais**...Madeira - Portugal: 2008Disponível em: <http://apps.isiknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=1D3kH23ImE3mlPgOLJF&page=5&doc=41&colname=WOS>. Acesso em: 14 jul. 2010

MENDIS, M. et al. Integration of Hive and cell software in the i2b2 architecture. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 1048, jan. 2007.

MUNGALL, C. J.; EMMERT, D. B.; THE FLYBASE CONSORTIUM. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. **Bioinformatics (Oxford, England)**, v. 23, n. 13, p. i337-46, jul. 2007.

MURPHY, S. N. et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 548-52, jan. 2007.

MURPHY, S. N. et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). **Journal of the American Medical Informatics Association : JAMIA**, v. 17, n. 2, p. 124-30, 1 jan. 2010.

NCBI. **GEO SOFT**. Disponível em: <http://www.ncbi.nlm.nih.gov/geo/info/soft2.html>. Acesso em: 1 out. 2012.

NCBI. **MINiML (MIAME Notation in Markup Language)**. Disponível em: <http://www.ncbi.nlm.nih.gov/geo/info/MINiML.html>. Acesso em: 1 nov. 2012.

NIH. **NIH Roadmap National Centers for Biomedical Computing**. Disponível em: <http://www.ncbcs.org/summary.html>. Acesso em: 9 nov. 2012.

O'CONNOR, B. D. et al. GMODWeb: a web framework for the Generic Model Organism Database. **Genome biology**, v. 9, n. 6, p. R102, jan. 2008.

ORACLE. **MySQL :: The world's most popular open source database**. Disponível em: <http://www.mysql.com/>. Acesso em: 23 nov. 2010.

RUBIN, D. L.; SHAH, N. H.; NOY, N. F. Biomedical ontologies: a functional perspective. **Briefings in bioinformatics**, v. 9, n. 1, p. 75-90, jan. 2008.

RUSSELL, S.; MEADOWS, L. A.; RUSSELL, R. R. **Microarray Technology in Practice**. 1. ed. Cambridge: Elsevier, 2009.

SELIGMAN, L. et al. **OpenII: an open source information integration toolkit** Proceedings of the 2010 international conference on Management of data - SIGMOD '10. **Anais**...New York, New York, USA: ACM Press, 6 jun. 2010Disponível em: <http://dl.acm.org/citation.cfm?id=1807167.1807285>. Acesso em: 19 abr. 2012

STEIN, L. D.; THIERRY-MIEG, J. AceDB: a genome database management system. **Computing in Science & Engineering**, v. 1, n. 3, p. 44-52, 1999.

VIANGTEERAVAT, T. et al. Slim-prim: a biomedical informatics database to promote translational research. **Perspectives in health information management / AHIMA, American Health Information Management Association**, v. 6, p. 6, jan. 2009.

VIANGTEERAVAT, T. et al. Biomedical Informatics Unit (BMIU): Slim-prim system bridges the gap between laboratory discovery and practice. **Clinical and translational science**, v. 2, n. 3, p. 238-41, jun. 2009.

VIANGTEERAVAT, T. et al. Protected Research Information Management Environment (PRIME) provides a secure open source data management option for clinical and scientific research. **BMC Bioinformatics**, v. 12, n. Suppl 7, p. A8, 2011.

VIDAL, V. et al. **An Ontology-Based Framework for Geographic Data Integration**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. v. 5833p. 337-346

W3C. **Semantic Web Health Care and Life Sciences (HCLS) Interest Group**. Disponível em: <http://www.w3.org/blog/hcls>. Acesso em: 25 mar. 2010.

W3C. **Translational Medicine Task Force - HCLC Semantic Web**. Disponível em: <http://www.w3.org/wiki/HCLSIG/PharmaOntology>. Acesso em: 22 nov. 2012.

WOOLF, S. H. The meaning of translational research and why it matters. **JAMA: the journal of the American Medical Association**, v. 299, n. 2, p. 211-3, 2008.