UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

VINÍCIUS DE FREITAS SOARES

IDENTIFICAÇÃO ÚNICA DE PACIENTES EM FONTES DE DADOS DISTRIBUÍDAS E HETEROGÊNEAS

VITÓRIA
2009

Vinícius de Freitas Soares

IDENTIFICAÇÃO ÚNICA DE PACIENTES EM FONTES DE DADOS DISTRIBUÍDAS E HETEROGÊNEAS

Dissertação submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo como requisito parcial para obtenção do grau de Mestre em Informática.

Orientador:
Prof. Dr. Alvaro Cesar Pereira Barbosa

Coorientador
Prof. Dr. Saulo Bortolon

Universidade Federal do Espírito Santo - UFES
Centro Tecnológico
Departamento de Informática

VITÓRIA - ES
2009

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

Soares, Vinícius de Freitas, 1968-
S676i Identificação única de pacientes em fontes de dados distribuídas e heterogêneas / Vinícius de Freitas Soares. – 2009.
152 f. : il.

Orientador: Alvaro Cesar Pereira Barbosa.
Co-Orientador: Saulo Bortolon.
Dissertação (mestrado) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Banco de dados. 2. Organização da informação. 3. Sistemas de recuperação da informação - Saúde pública. 4. Pacientes - Identificação. I. Barbosa, Alvaro Cesar Pereira. II. Bortolon, Saulo. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 004

Identificação Única de Pacientes em Fontes de Dados Distribuídas e Heterogêneas

Vinícius de Freitas Soares

Dissertação submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo como requisito parcial para obtenção do grau de Mestre em Informática.

Aprovada em 25/08/2009 por:

Prof. Dr. Alvaro Cesar Pereira Barbosa - DI/UFES

Prof. Dr. Saulo Bortolon - DI/UFES

Prof. Dr. Anilton Salles Garcia - DI/UFES

Profª. Drª. Cláudia Medina Coeli - IESC/UFRJ

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

Vitória-ES, agosto de 2009

Dedico este trabalho à mulher da minha vida,
Maria Tereza e ao homem da minha vida,
Meu filho, Caio.

AGRADECIMENTOS

A Deus, por ter me feito acreditar Nele e por ter acreditado em mim.

Aos meus pais por ter me dado a vida, o meu caráter, a minha formação, e por tantas outras coisas que se eu fosse citar daria uma outra dissertação.

Às minhas irmãs pelo apoio e, sobretudo, por serem minhas irmãs. Sempre verei minha irmã Sandra com cinco metros de altura e Mary, para mim, é exemplo de tudo.

Ao professor Alvaro Cesar Pereira Barbosa, que me aceitou no Mestrado e me aturou este tempo todo, apesar das brigas. Para mim, o professor Alvaro é mais que um amigo ou orientador, ele é aquilo que eu quero ser quando crescer.

Aos professores Saulo Bortolon, Ricardo Falbo, Sérgio de Freitas e Berilhes por me exigirem e por terem me feito enxergar o quanto eu conseguiria aprender.

Aos colegas do Prodest, em especial ao Glauber (ex-Prodest), Tiago, Breno, Sylvia, Suzana, Wideraldo, Denise, Fernando Pêgo, Olga, Ulisses, Marco Aurélio, Fábio Modenese, Rosângela, Giancarlo e Sérgio que não só torceram como me ajudaram a desenvolver este trabalho.

Às seguintes instituições do Governo do Estado do Espírito Santo, na figura dos seus gestores: Instituto de Tecnologia da Informação e Comunicação do Estado do Espírito Santo (Prodest), Secretaria de Estado da Saúde (SESA), Secretaria de Estado de Gestão e Recursos Humanos (SEGER) e Departamento Estadual de Trânsito (DETRAN).

Aos meus colegas do Mestrado, com destaque para os meus companheiros de caminhada: Carlos, Ramon, Patrícia e Camilo Carvalho.

Ao meu mentor espiritual, Padre José Pedro Luchi, prova de que fé e sabedoria podem andar juntas em qualquer nível cultural ou científico.

Ao meu primo Carlos Eduardo e ao meu quase irmão, Marcus De Angelis.

RESUMO

No decorrer de sua vida, um paciente é atendido por várias instituições de saúde e é submetido a uma série de procedimentos. A quantidade de informações armazenadas sobre esse paciente é crescente, tanto em volume quanto em diversidade. Existem ainda diferentes identificações para um mesmo paciente, gerando alto custo com duplicação de procedimentos e colaborando com a imprecisão dos diagnósticos e tratamentos. Nesse sentido, o presente trabalho utiliza técnicas de *Record Linkage* e geração de MPI (*Master Patient Index*), combinadas com as especificações do perfil de integração PIX (*Patient Identifier Cross-Referencing*), para estabelecer uma identificação única de pacientes em diferentes sistemas de informação em saúde, que contenham fontes de dados heterogêneas e distribuídas. Com a utilização desses conceitos e tecnologias, foi especificado um projeto e desenvolvido um protótipo de um IHE (*Integrating the Healthcare Enterprise*)/PIX. Experimentos foram realizados em três cenários com dados reais.

Palavras-Chave: Banco de Dados, Armazenamento e Recuperação de Informação, Integração de Dados, Sistemas de Informação em Saúde, Identificação de Pacientes, IHE/PIX, *Record Linkage*.

ABSTRACT

Through its lifetime, patient's healthcare information is generated in several health care institutions and submitted to a series of procedures. The amount of stored information is incremental by nature, both in volume and diversity. There are also differences in the patient identification, bringing high expenses with duplicated procedures and contributing to inaccuracy in diagnosis and treatments. Therefore, the present work uses Record Linkage techniques during the creation of a MPI (Master Patient Index), combined with the specification of a PIX (Patient Identifier Cross-Referencing) integration profile, in order to establish a unique patient identification in different healthcare information systems, which include homogeneous and distributed data sources. Using those concepts and technologies, a project has been specified and it gave birth to an IHE (Integrating the Healthcare Enterprise) /PIX prototype. Experiments have been carried out in three sceneries with real data.

Keywords: Databases, Information Storage and Retrieval, Data Integration, Health Information Systems, Patient Identification, IHE/PIX, Record Linkage.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

ACC - *American College of Cardiology*

ADT - *Admission, Discharge, Transfer*

ANS - Agência Nacional de Saúde Suplementar

ANSI - *American National Standards Institute*

APAC-SIA - Autorização para Procedimentos de Alto Custo/Complexidade para o Sistema de Informações Ambulatoriais do SUS

ASP - *Active Server Pages*

ATNA - *Audit Trail and Node Authentication*

CA - *Computer Associates*

CAPS - *Composite Application Platform Suite*

CASE - *Computer-Aided Software Engineering*

CBR - *Case Base Reasoning*

CCR - *Continuity of Care Records*

CDA - *Clinical Document Architecture*

CDC - *Centers for Disease Control and Prevention*

CEF - Caixa Econômica Federal

CEM - Código de Ética Médica

CEN - *Comité Européen de Normalisation* (*European Committee for Standardization*)

CEP - Comitê de Ética em Pesquisa

CFM - Conselho Federal de Medicina

CID - Classificação Internacional de Doenças

CIH - Comunicação de Internação Hospitalar

CIS - *Clinical Information System*

CNH - Carteira Nacional de Habilitação

CNES - Cadastro Nacional de Estabelecimentos de Saúde

CNS - Cartão Nacional de Saúde

CoDIMS - *Configurable Data Integration Middleware System*

COM - *Component Object Model*

Conasems - Conselho Nacional de Secretarias Municipais de Saúde

CORBA - *Common Object Request Broker Architecture*

CPB - Código Penal Brasileiro

CPF - Cadastro de Pessoas Físicas

CT - *Computed Tomography*

CT - *Consistent Time*

DATASUS - Departamento de Informática do SUS

DETRAN - Departamento Estadual de Trânsito

DHF - Diagrama Hierárquico de Funções

DICOM - *Digital Imaging Communications in Medicine*

DLL - *Dynamic Link Libraries*

ECG - *Electrocardiogram*

EHR - *Electronic Healthcare Record*

EM - *Expectation-Maximization*

EPR - *Eletronic Pacients Record*

ER - *Entity Resolution*

ESB - *Enterprise Service Bus*

EUA - *Enterprise User Authentication*

FAPES - Fundação de Apoio à Ciência e Tecnologia do Espírito Santo

Febrl - *Freely Extensible Biomedical Record Linkage*

FMUSP - Faculdade de Medicina da Universidade de São Paulo

FRIL - *Fine-grained Record Integration and Llinkage*

GM - Gabinete do Ministro

GNU - *General Public License*

HC - Hospital das Clínicas

HIMSS - *Health Information Management Systems Society*

HIS - *Hospital Information System*

HL7 - *Health Level Seven*

HOI - *Health Outcomes Institute*

IARC - *International Agency for Research on Cancer*

ICD - *International Classification of Diseases*

IDC - *International Data Corporation*

IDE - *Integrated Development Environment*

IESC - Instituto de Estudos em Saúde Coletiva

IFES - Instituto Federal do Espírito Santo

IHE - *Integrating the Healthcare Enterprise*

IIS - *Internet Information Services*

IMS - Instituto de Medicina Social

ISO - *International Organization for Standardization*

ITI - *IHE IT Infrastructure*

IUPAC - *International Union of Pure and Applied Chemistry*

J2EE - *Java 2 Enterprise Edition*

LANL - *Los Alamos National Labs*

LIS - *Laboratory Information System*

LOINC - *Logical Observation Identifiers Names and Codes*

MACDP - *Metropolitan Atlanta Congenital Defects Program*

MDC - Modelo de Dados Comum

MPI - *Message Passing Inteface*

MPI - *Master Patient Index*

MR - *Magnetic Resonance*

MS - Ministério da Saúde

MVC - *Model-view-controller*

NDC - *National Drug Codes*

NHII - *National Health Information Infrastructure*

NIS - Número de Identificação Social

NPCR - *National Program of Cancer Registries*

OHF - *Open Healthcare Framework*

OMG - *Object Management Group*

OpenEHR - *Open Electronic Health Records*

PACS - *Picture Archiving and Communication System*

PAM - *Patient Administration Management*

PDMS - *Peer-to-Peer Data Management Systems*

PDQ - *Patient Demographics Query*

PEP - Prontuário Eletrônico do Paciente

PHI - *Protected Healthcare Information*

PIDS - *Person Identification Service*

PIX - *Patient Identifier Cross Referencing*

PMC - Proporção Mínima de Concordância

PNIIS - Política Nacional de Informação e Informática em Saúde

PPA - Plano Plurianual

PSA - *Patient Synchronized Application*

PSF - Programa Saúde da Família

PWP - *Personnel White Pages*

RCP - *Rich Client Platform*

RES - Registro Eletrônico de Saúde

RFD - *Retrieve Form for Data Capture*

RID - *Retrieve Information for Display*

RIS - *Radiology Information System*

RSNA - *Radiological Society of North America*

RUTE - Rede Universitária de Telemedicina

SALI - *Software for Automated Linkage in Italy*

SAMHSA - *Substance Abuse and Mental Health Administration*

SAP - *Systeme, Anwendungen und Produkte in der Datenverarbeitung*

SCNS - Sistema do Cartão Nacional de Saúde

SDO - *Standards Developing Organization*

SEDU - Secretaria da Educação

SEGER - Secretaria de Estado de Gestão e Recursos Humanos

SESA - Secretaria da Saúde

SESP - Secretaria de Segurança Pública e Defesa Social

SGBD - Sistema Gerenciador de Banco de Dados

SIAB - Sistema de Informação da Atenção Básica

SIARHES - Sistema Integrado de Administração de Recursos Humanos do Estado do Espírito Santo

SIB - Sistema de Informações de Beneficiários

SIEPI - Sistema de Informações Epidemiológicas

SIH - Sistema de Internações Hospitalares

SIH-SUS - Sistema de Informações Hospitalares do SUS

SIM - Sistema de Informação sobre Mortalidade

Sinan - Sistema de Informação de Agravos de Notificação

Sinasc - Sistema de Informação sobre Nascidos Vivos

SIS - Sistema(s) de Informação em Saúde

SNOMED - *Systemized Nomenclature of Medicine*

SO - Sistema Operacional

SOAP - *Simple Object Access Protocol*

SOP - Solicitação de Propostas

SUS - Sistema Único de Saúde

TISS - Troca de Informações em Saúde Suplementar

UERJ - Universidade do Estado do Rio de Janeiro

UFES - Universidade Federal do Espírito Santo

UFRJ - Universidade Federal do Rio de Janeiro

UMDNS - *Universal Medical Device Nomenclature System*

UML - *Unified Modeling Language*

UMLS - *Unified Medical Language System*

XCA - *Cross-Comunity Access*

XDM - *Cross-Enterprise Document Media Interchange*

XDR - *Cross-Enterprise Document Reliable Interchange*

XDS - *Cross Enterprise Document Sharing*

XDS-SD - *Cross-Enterprise Sharing of Scanned Documents*

XML - *Extensible Markup Language*

SUMÁRIO

1. Introdução

Neste Capítulo, estão descritos o contexto no qual este trabalho está inserido, sua motivação e justificativa, seu objetivo, a metodologia e as contribuições esperadas. No final do Capítulo, é apresentada a organização deste trabalho.

1.1. Contexto

Atualmente, 92% das informações produzidas no mundo são criadas em formato digital. Os meios analógicos, como o papel, abrigam apenas 8% dos dados [Favaro e Vieira 2008]. Esses dados digitais que, segundo o IDC – International Data Corporation[1], estavam em 2007 na casa dos 281 exabytes (281 bilhões de gigabytes), estão armazenados em diversas fontes de dados. Frequentemente, essas fontes estão dispersas geograficamente (distribuídas), podendo estar interligadas por uma rede. Geralmente, são desenvolvidas de maneira independente, em plataformas de hardware e software distintas, possuindo diferentes modelos de dados, formas de consulta e protocolos de gerenciamento de transações (heterogêneas) [Özsu e Valduriez 2001]. É comum também que as fontes de dados usem estruturas e terminologias diferentes (heterogeneidades estrutural e semântica, respectivamente) [Kim e Seo 1991] e [Pitoura, Bukhres e Elmagarmid 1995].

Com o crescimento da distribuição e da heterogeneidade dos dados, aumenta a necessidade de se ter uma visão integrada desses dados. Uma demanda atual é identificar registros correspondentes a uma mesma entidade de interesse (indivíduos, empresas, regiões geográficas ou famílias) [Gu et al. 2003]. Essa tarefa é trivial nos casos em que cada fonte de dados possui um identificador unívoco: um campo obrigatório, não duplicado, e usado para identificar cada um dos registros, como, por exemplo, o CPF. Esse tipo de associação é definido como método determinístico [Coeli et al. 2006].

No entanto, há bases de dados onde não há um identificador unívoco comum a duas fontes. Nesses casos, são utilizados campos não unívocos, presentes nas duas fontes de dados, que permitam indicar que, provavelmente, um par de registros se refira a uma mesma entidade de interesse. Esse método é definido como probabilístico [Coeli et al. 2006].

[1] http://www.idc.com/

Os sistemas de saúde, especificamente, são ricos em fontes de dados que não podem ser combinadas pelo método determinístico, em especial, na identificação de um mesmo paciente em diversas bases de dados. É comum que um mesmo paciente seja identificado de diferentes maneiras pelos diversos prestadores de serviço de saúde que o atenderão no decorrer de sua vida. Como solução para este problema de integração, é fundamental a identificação de cada paciente de maneira única. A ausência dessa identificação pode causar efeitos catastróficos.

Nos Estados Unidos, cerca de 98.000 pessoas morrem por causa de negligência médica quando estão hospitalizados: 13% desse número se deve a irregularidades ocorridas em cirurgias e 67% a falhas na transfusão de sangue. Essas falhas podem ser causadas por identificação errônea do paciente [JCI 2005] e [Mettler, Fitterer e Rohner 2007]. Estudos na Inglaterra e na Austrália revelam que entre 12% e 16% de todos os pacientes estão expostos a um "evento adverso" [Mettler, Fitterer e Rohner 2007]. As principais razões para que esses eventos ocorram com frequência são a insuficiência de comunicação, a falta de trabalho em equipe e a verificação inadequada da identificação do paciente [Chassin e Becher 2002].

No Brasil, a preocupação com a identificação precisa dos pacientes e a integração das bases de dados de saúde se dá por ações governamentais. O Ministério da Saúde, através do Departamento de Informação e Informática do SUS (Sistema Único de Saúde), estabeleceu uma Política Nacional de Informação e Informática em Saúde [MS 2004]. Essa política possui como uma de suas diretrizes:

> *"Estabelecer sistema de identificação unívoca de usuários, profissionais e estabelecimentos de saúde que seja progressivamente adotado, aprimorando o processo de integração dos sistemas de informação de saúde e viabilizando o registro eletrônico de saúde. O Cartão Nacional de Saúde - que identifica univocamente usuários e profissionais - e o Cadastro Nacional de Estabelecimentos de Saúde – que identifica univocamente os estabelecimentos – são o passo inicial na construção deste novo paradigma."*

Essa necessidade de identificação única de pacientes também é seriamente discutida por instituições como a ANS – Agência Nacional de Saúde Suplementar[2], Ministério do Planejamento[3],

[2] http://www.ans.gov.br/portalv4/site/home/default.asp

[3] http://www.planejamento.gov.br

CONASEMS - Conselho Nacional de Secretarias Municipais de Saúde[4] e SAS - Secretaria de Atenção à Saúde[5], órgão pertencente ao Ministério da Saúde [CONIP 2006].

A identificação única de pacientes é necessária tanto para a integração de instituições diferentes, quanto para sistemas diferentes pertencentes a uma mesma instituição. A rigor, cada prestador de serviço pode ter seu próprio sistema, com sua maneira peculiar de identificar seus pacientes. Muitas vezes, esses pacientes são identificados com um nível de imprecisão muito grande. As instituições de saúde resistem em alterar seus sistemas, mesmo que em nome de uma integração ou de uma unificação, devido o custo que isto representa. Portanto, soluções de integração de dados devem ser flexíveis a ponto de permitir uma interoperabilidade entre os sistemas, sem que esses sofram alterações. As diversas fontes de dados poderiam apenas fornecer dados para permitir a identificação única do paciente, como pode ser observado na Figura 1.

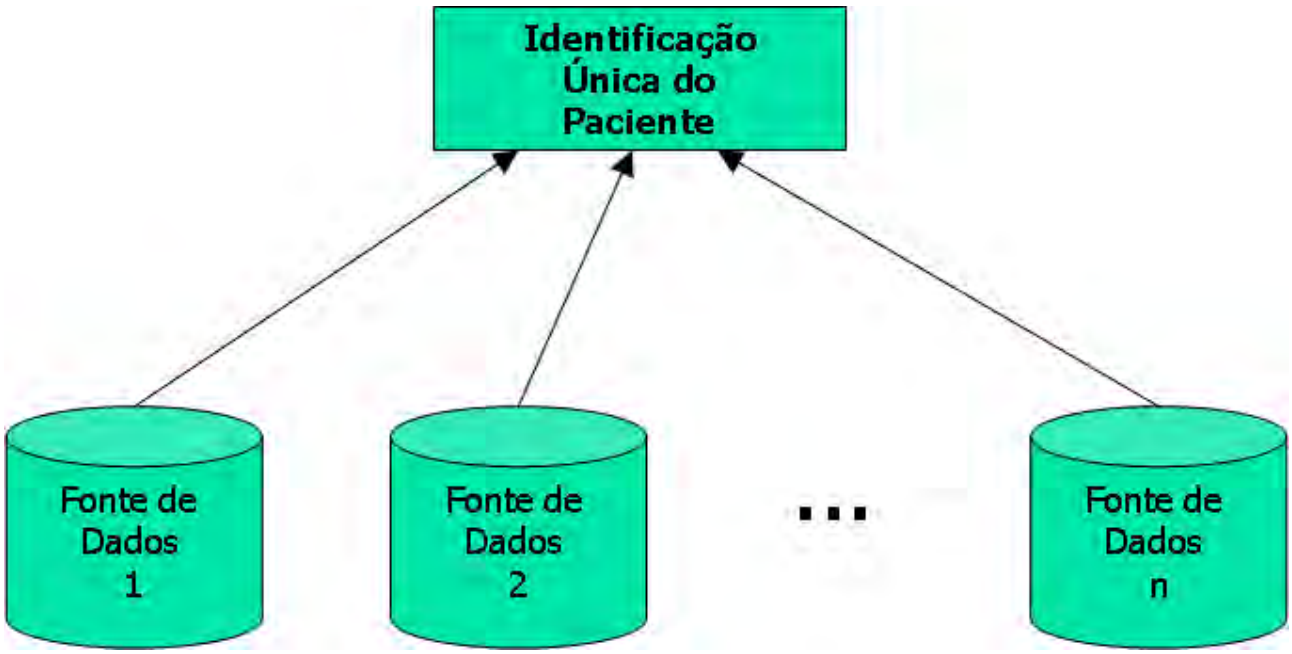

Figura 1: Fornecimento de informações pelas fontes de dados para permitir a identificação única do paciente.

É importante, também, destacar que um mesmo prestador de serviço de saúde pode conter mais de um tipo de sistema: sistemas de informações hospitalares (HIS – *Hospital Information System*), prontuário eletrônico de pacientes (EPR – *Eletronic Pacients Record*), sistema de informações de radiologia (RIS – *Radiology Information System*), sistema de comunicação e arquivamento de imagens (PACS - *Picture Archiving and Communication System*) e tantos outros. Além disso, os

[4] http://www.conasems.org.br/cgi-bin/pagesvr.dll/Get?id_sec=2

[5] http://www.saude.gov.br/sas

dados dos pacientes podem estar armazenados nas mais diferentes formas: dados cadastrais estruturados, laudos e diagnósticos escritos em texto livre, exames no formato de imagens e vídeos, etc. Também vale lembrar que a integração de todo este universo de dados, em função do paciente, deve respeitar a normas legais e a preceitos éticos, que são discutidos no Capítulo 3.

Já existem padrões para a integração de dados médicos e, mais especificamente, para estabelecer a identificação única de pacientes. O padrão IHE (*Integrating the Healthcare Enterprise*) [IHE 2009], reconhecido em diversas partes do mundo, sobretudo na Europa e América do Norte, define Perfis de Integração, que especificam as interações entre várias aplicações de saúde. O IHE é uma iniciativa que recomenda o uso de padrões existentes, tais como HL7 (*Health Level Seven*) [HL7], DICOM [DICOM] e outros, além de definir novos padrões de integração [Stolba e Schanner 2007].

Dentre os perfis de integração do IHE, encontra-se o PIX (*Patient Identifier Cross-Referencing*) [ACC, HIMSS e RSNA 2008], responsável pela integração de dados de pacientes, através da geração do MPI (*Master Patient Index*). O perfil de integração IHE/PIX promove a referência cruzada entre os diversos identificadores de pacientes, usando um índice único gerado, o MPI. Dessa forma, é possível identificar o mesmo paciente em diversos sistemas de uma organização de saúde ou entre organizações.

1.2. Motivação

A heterogeneidade e distribuição dos dados dos pacientes produzem uma fragmentação do conhecimento sobre sua saúde. Neste caso, um determinado paciente tem seus dados em diversas fontes (clínicas, consultórios, laboratórios, hospitais), o que pode gerar:

- A não identificação imediata do paciente ou o desconhecimento da existência de seus diversos prontuários em situações de emergência;
- A solicitação de exames em duplicidade, com aumento de custos;
- Erros de diagnóstico devido à falta de dados atuais e do histórico do paciente.

Como já dito, o uso de padrões é fundamental para a identificação única do paciente e, consequentemente, para a elaboração de um prontuário eletrônico único. Na literatura, são encontradas diversas soluções para integração de dados de pacientes e geração de uma identificação única para ele. Para a geração do MPI, pode-se empregar procedimentos de *Record Linkage*, que

são técnicas de integração entre sistemas existentes, combinando dados de dois bancos de dados ou mais, verificando a sobreposição dos mesmos [Rötzch 2006]. As soluções de PIX mais citadas na literatura e as implementações de *Record Linkage* que possuem maior destaque são apresentadas e discutidas no Capítulo 2 deste trabalho. Contudo, não basta identificar o mesmo paciente em fontes de dados diferentes. É necessário criar mecanismos que permitam, a partir da identificação única, retornar às fontes originais, a fim de consultar não apenas os dados demográficos do paciente como seu histórico clínico.

Propõe-se, nesta dissertação, desenvolver um protótipo de implementação que permita a identificação única de pacientes, cujos dados estejam em fontes de dados heterogêneas, localizadas nas diversas organizações de saúde. Esta implementação deve reunir características do padrão IHE/PIX e técnicas de *Record Linkage*. Para suprir esta necessidade, devem-se atender alguns requisitos:

- Com o intuito de seguir padrões da área de informática em saúde, a solução deve atender pelo menos parte das especificações do perfil de integração IHE/PIX, cuja última revisão foi realizada em dezembro de 2008 [ACC, HIMSS e RSNA 2008];
- Deve permitir que sejam consultadas as fontes de dados originais, sem que essas sejam alteradas e sem a geração de arquivos intermediários;
- Não utilização de arquiteturas proprietárias, com camadas de hardware ou software dependentes de um único fornecedor.

1.3. Objetivos

O objetivo geral deste trabalho é utilizar as técnicas de *Record Linkage* e geração de MPI, combinadas com parte das especificações do perfil de integração IHE/PIX, para estabelecer uma identificação única de pacientes em diferentes sistemas de informação em saúde, que contenham fontes de dados heterogêneas e distribuídas.

Destacam-se como objetivos específicos: (*i*) análise das características de arquiteturas implementadas de IHE/PIX e *Record Linkage*; (*ii*) especificação e modelagem do projeto de um IHE/PIX, usando técnicas de *Record Linkage*; (*iii*) desenvolvimento de um protótipo que permita a combinação probabilística de fontes de dados distintas, com a possibilidade de retorno aos dados originais.

1.4. Justificativa

A razão de esta pesquisa ter sido proposta é o fato de haver poucas soluções na literatura para identificação única de pacientes, com retorno às fontes de dados originais que, por sinal, estão em constante atualização. É mais comum encontrar soluções que exigem que os dados sejam extraídos para que sejam combinados dentro de um formato específico. Em outros casos, são encontradas soluções de identificação de pacientes e geração de MPI que não seguem padrão algum, muito menos o padrão IHE/PIX.

Há ainda soluções comerciais que implementam o perfil de integração IHE/PIX. Contudo, exigem plataformas de software proprietárias ou, até mesmo, a aquisição de um hardware específico.

A solução proposta permite consulta às fontes de dados originais, após a geração do MPI; obedece a parte das especificações do perfil PIX, como pode ser observado no Capítulo 5; e permite carga incremental do MPI, identificando novos registros nas fontes de dados, associados a um mesmo paciente.

1.5. Metodologia de Desenvolvimento do Trabalho

Para a realização deste trabalho, inicialmente foi necessário um estudo sobre integração de dados heterogêneos e distribuídos.

A partir de um estudo dirigido, que teve como tema a Integração de Dados em Sistemas de Saúde, foi estudado o padrão IHE. Como já descrito, um dos perfis de integração do IHE é o PIX, que integra dados de pacientes, gerando um índice mestre. Após o estudo desses padrões, foram definidas duas linhas de pesquisa. A primeira, foi um estudo sobre a geração do MPI, especialmente, sobre a técnica de *Record Linkage*, detalhada no Capítulo 4. A segunda linha de pesquisa diz respeito aos PIX já implementados tanto em nível acadêmico quanto em nível comercial ou governamental. Como implementações de perfis PIX ainda são raras, também foram analisados os projetos clássicos que utilizam técnicas de *Record Linkage* e geração de MPI.

Uma vez compreendidas as técnicas de geração de MPI, através do estudo de *Record Linkage* e do padrão PIX, foi especificado um projeto e desenvolvido um protótipo que permite que sejam identificados pacientes em fontes de dados heterogêneas e distribuídas, permitindo acesso a essas fontes de dados originais a partir do índice gerado. Além disso, foram desenvolvidos recursos

como: (*i*) **carga incremental do índice** de acordo com as atualizações das fontes; (*ii*) **promoção de par duvidoso em par combinado**: a partir da exibição dos índices de confiança dos pares duvidosos e dos dados das suas fontes originais, o usuário decide se é ou não o mesmo paciente; (*iii*) **configuração dos parâmetros** de todas as fases do *Record Linkage*.

Por fim, no Estudo de Caso, é testado o protótipo em três cenários, com destaque para o terceiro, que relaciona os pacientes com catarata com aqueles que possuem glaucoma, com base nos cadastros da Secretaria da Saúde do Estado do Espírito Santo.

1.6. Contribuições

Este trabalho faz parte do projeto de pesquisa "Software Livre e Interoperabilidade em Saúde", financiado pela FAPES[6] (Fundação de Apoio a Ciência e Tecnologia do Estado do Espírito Santo), outorga 032/2007. O objetivo desse projeto de pesquisa é explorar a infraestrutura da RUTE (Rede Universitária de Telemedicina) para pesquisar, desenvolver, testar e implantar soluções de software livre para PACS e desenvolver soluções para integrar o prontuário tradicional, baseado em papel, com sistemas de imagens, através de soluções de digitalização de baixo custo.

Um artigo sobre este trabalho foi submetido, aceito e apresentado no CBIS'2008 (XI Congresso Brasileiro de Informática em Saúde) [Soares, Barbosa e Costa 2008].

Além disso, são esperadas as seguintes contribuições com este trabalho:

- Consolidação dos conceitos sobre integração de dados, em especial sobre a técnica de *Record Linkage* e geração de MPI.
- Resumo sobre padrões de Informática em Saúde, com destaque para o IHE e seu perfil de integração PIX.
- Pesquisa e documentação da arquitetura de implementações PIX e *Record Linkage*, seja de instituições acadêmicas, governamentais ou pertencentes ao mercado. Vários destes softwares foram desenvolvidos por instituições públicas ou acadêmicas e encontram-se disponíveis em modalidade de *open source*. Outros são soluções comerciais, com cópias de avaliação disponíveis.
- Formalização dos passos para geração do MPI, através da técnica de *Record Linkage*.

[6] http://www.fapes.es.gov.br

- Concepção e modelagem de uma sistema de informação que une as especificações do IHE/PIX com as técnicas de *Record Linkage*.
- Desenvolvimento de um protótipo que integra fontes de dados heterogêneas de pacientes, gerando um índice único.

Os estudos desenvolvidos aqui são, na visão do autor, um passo importante para implantação no Estado do Espírito Santo de um prontuário único de pacientes na rede pública estadual. Essa ação está planejada desde setembro de 2003, quando foi elaborado o Relatório de Gestão da Secretaria de Estado da Saúde de 2002 [SESA 2003]. O prontuário único informatizado envolverá 11 hospitais da rede pública estadual, além dos centros de referência e da rede conveniada.

1.7. Organização da Dissertação

Além desta Introdução, este trabalho está dividido nos seguintes capítulos:

- **Capítulo 2:** São apresentados e discutidos trabalhos com soluções implementadas de perfis de integração PIX ou projetos clássicos de *Record Linkage* e geração de MPI.
- **Capítulo 3:** São apresentados os principais fundamentos e tecnologias envolvidas neste trabalho.
- **Capítulo 4:** São descritas, detalhadamente, todas as etapas do método de *Record Linkage*, com a respectiva geração do MPI.
- **Capítulo 5:** Apresenta a concepção, modelagem e implementação de um protótipo que une as especificações do IHE/PIX com as técnicas de *Record Linkage*, com o objetivo de integrar fontes de dados heterogêneas de pacientes, gerando um índice único.
- **Capítulo 6:** Descreve o estudo de caso, com a utilização de três cenários que combinam fontes de dados heterogêneas.
- **Capítulo 7:** Apresenta uma síntese do trabalho realizado, suas principais contribuições e a indicação de alguns trabalhos futuros para a continuidade desta pesquisa.

Finalmente, são listadas as referências bibliográficas utilizadas.

2. Trabalhos Relacionados

Neste Capítulo, são apresentadas arquiteturas de soluções PIX já implementadas. Contudo, como essas implementações ainda são raras, também foram analisados os projetos clássicos que utilizam técnicas de *Record Linkage* e geração de MPI. Também foram estudadas soluções de combinação de fontes de dados que utilizam técnicas específicas ou que são destaque na área de informática em saúde.

2.1. Introdução

O padrão IHE está organizado em domínios operacionais e clínicos. Em cada um desses domínios, usuários com experiência clínica ou operacional identificam prioridades de integração e compartilhamento de informações. Os domínios também são úteis para que empresas desenvolvedoras de sistemas de saúde forneçam soluções baseadas em padrões [IHE 2009].

Um desses domínios é o *IHE IT Infrastructure* (ITI). Ele trata da implementação de soluções, baseadas em padrões de interoperabilidade, que buscam melhorar o compartilhamento de informações, o fluxo de trabalho e a assistência ao paciente.

O domínio *IHE IT Infrastructure* iniciou-se em 2003 e foi lançado pelo *Health Information Management Systems Society* (HIMSS)[7]. Em 2008, o GIP-DMP (*Groupement d'Intérêt Public pour le Dossier Médical Personnel*)[8] juntou-se ao HIMSS para ajudar a patrocinar o ITI na Europa. Um dos perfis de integração pertencentes a este domínio é o PIX, que foi proposto e homologado na revisão 1.1 do *IT Infrastructure Technical Framework* [ACC, HIMSS e RSNA 2004], em 30 de julho de 2004. O PIX sofreu sua primeira revisão em 15 de agosto de 2005, na versão 2.0 do *IT Infrastructure Technical Framework* [ACC, HIMSS e RSNA 2005]. A Revisão mais recente do PIX ocorreu em 12 de dezembro de 2008, com a versão 5.0 do *IHE IT Infrastructure Techinical Framework* [ACC, HIMSS e RSNA 2008].

Embora, em sua essência, a especificação do padrão IHE/PIX tenha permanecido a mesma, este perfil de integração é muito recente e ainda demanda revisões que garantam sua estabilidade. As implementações clássicas de IHE/PIX encontradas na literatura são abordadas na Seção 2.2.

[7] http://www.himss.org/ASP/index.asp

[8] http://www.d-m-p.org/index.php?option=com_content&task=view&id=283&Itemid=303

Na Seção 2.3, é traçado um histórico sobre a técnica de *Record Linkage* e são apresentadas as soluções mais referenciadas na literatura. Estudos como o realizado por Elmagarmid, Ipeirotis e Verykios (2007) apresentam um *survey* sobre soluções clássicas de *Record Linkage*. Soluções de *Record Linkage* menos referenciadas na literatura também são apresentadas, devido a proximidade com a proposta deste trabalho.

Na Seção 2.4, são analisadas alternativas de combinação de fontes de dados que utilizam técnicas próprias desenvolvidas para este fim. Também são analisadas, nessa seção, soluções de identificação única de pessoas específicas da área de saúde, sejam do meio acadêmico-científico, do meio governamental ou do mercado.

Ao final deste Capítulo, é elaborado um quadro-resumo, que possui como objetivo mostrar um estudo comparativo entre as soluções estudadas, que se aproximam do que é proposto.

2.2. Implementações IHE/PIX

Poucas são as soluções IHE/PIX encontradas na literatura. Nesta Seção, são descritas três arquiteturas de implementações PIX, sendo duas delas, soluções comerciais.

A importância de descrever essas arquiteturas é ressaltar o caráter não proprietário deste trabalho. Nas arquiteturas citadas, são observadas infraestruturas proprietárias, como barramentos, *bridges* ou camadas de hardware e software dependentes de um único fabricante.

2.2.1. LENUS

O LENUS [Krechel e Hartbauer 2008] e [LENUS] é um gerador de índice mestre de pacientes (MPI), desenvolvido pela *SER Solutions Deutschland GmbH*[9], que está associado a um registro único de pacientes e a um barramento de serviços de assistência médica. Esses serviços de assistência são compreendidos por três funcionalidades:

- Admissão do paciente, com a sua consequente inclusão no MPI;
- Transferência de pacientes para os diversos departamentos de um hospital e
- Dispensa do paciente.

Essas três funcionalidades são normalmente resumidas na sigla ADT (*Admission – Discharge –*

[9] http://www.ser.de/ww/de/pub/solutions.cfm

Transfer).

O LENUS parte do princípio de que alguns aspectos da arquitetura de sistemas heterogêneos não podem ser resolvidos pela simples adoção de um MPI. Então, ele combina um registro de pacientes centralizado (LENUS EPR) com o componente MPI para garantir mais funcionalidade. Com isso, ele torna possível o armazenamento de conteúdos de diferentes sistemas em um repositório central. Além disso, ele agrega, nesse repositório, dados DICOM e HL7 provenientes de sistemas especialistas.

Baseado nos casos de uso descritos no cenário IHE/PIX, descritos na Seção 5.1, o LENUS possui as seguintes funcionalidades para que o MPI seja gerado:

- Ligação (*Linking*) da identificação de pacientes de diferentes sistemas e organizações em um índice de paciente central.
- Disponibilização de algoritmos inteligentes de busca, que combinam dados demográficos enviados por aplicações cliente com dados referentes ao índice central gerado.
- Emparelhamento (*Matching*) semi-automático de dados, através de parâmetros configuráveis.
- Emparelhamento manual de pacientes incertos. Pacientes cuja probabilidade de existência em mais de uma fonte de dados está em uma faixa de dúvida.
- Bloqueio (*Lock)* e desbloqueio (*unlock)* de pacientes.
- Funções sofisticadas de busca integradas a um cliente MPI.
- Fusão manual de pacientes.
- Inclusão de dados de pacientes novos.
- Atualização de dados de pacientes.
- Cancelamento de dados de pacientes.
- Fusão de dados de pacientes.
- Notificação a todos os sistemas interligados ao LENUS no caso de inclusão de novos pacientes.

A SER, empresa desenvolvedora do LENUS, adota o conceito de barramento para prover a comunicação entre diversos sistemas de saúde tanto com o LENUS MPI, gerador do identificador único de pacientes do LENUS, quanto com o LENUS EPR, seu prontuário eletrônico. O barramento de serviços corporativos (ESB – *Enterprise Service Bus*) pode ser mais bem compreendido, observando-se a Figura 2. Nessa arquitetura de barramento, o ESB é chamado de HSB (*Healthcare*

Service Bus), ou Barramento de Serviços de Saúde, já que adota conceitos do domínio de saúde.

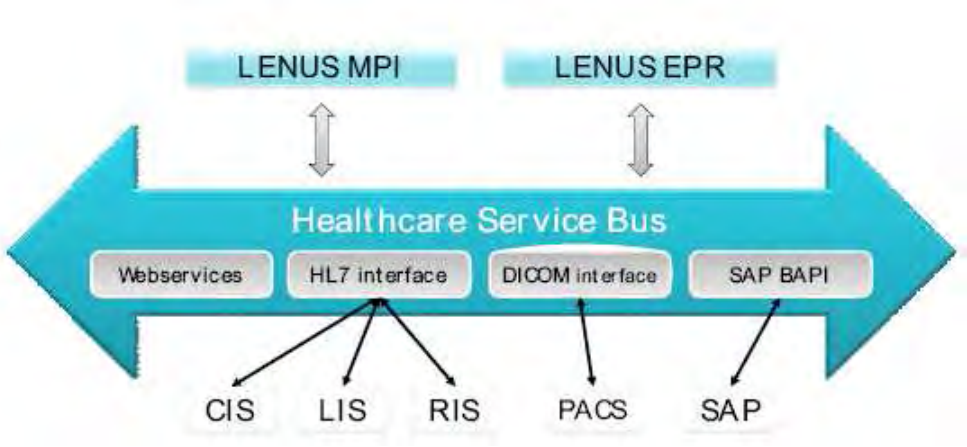

Figura 2: *Healthcare Service Bus* utilizado pelo LENUS [Krechel e Hartbauer 2008].

A Figura 2 também mostra os adaptadores que ligam o LENUS aos sistemas da área de saúde:

- **HL7 interface:** São adaptadores que seguem o padrão HL7. Muitos sistemas de saúde, como HIS, RIS, sistemas de informações de laboratórios (LIS – *Laboratory Information System*) e sistemas de informações clínicos (CIS – *Clinical Information System*) já geram e compartilham dados no padrão HL7.
- **DICOM interface:** O DICOM é um padrão fortemente usado no domínio da radiologia. Adaptadores DICOM são usados para troca de dados de imagens médicas de pacientes através do barramento.
- **SAP BAPI:** Para áreas administrativas na Alemanha e Áustria, foco comercial do LENUS, conexões com sistemas SAP[10] são muito importantes. Portanto, foi desenvolvido o adaptador SAP-BAPI/RFC para que sistemas desenvolvidos com tecnologia SAP pudessem ser conectados ao barramento.
- Todos os outros sistemas podem ser conectados através de um adaptador XML/SOAP genérico, que permite interface com arquivos ou bancos de dados.

O LENUS MPI recebe mensagens filtradas e metadados através do *Healthcare Service Bus* e se comunica com os ambientes clínicos heterogêneos via interface com os adaptadores. Ou seja, toda comunicação do LENUS MPI com os sistemas de saúde é feita através do *Healthcare Service Bus*. O demais componentes da arquitetura do LENUS *Master Patient Index* estão descritos na Figura 3.

[10] http://www12.sap.com/index.epx

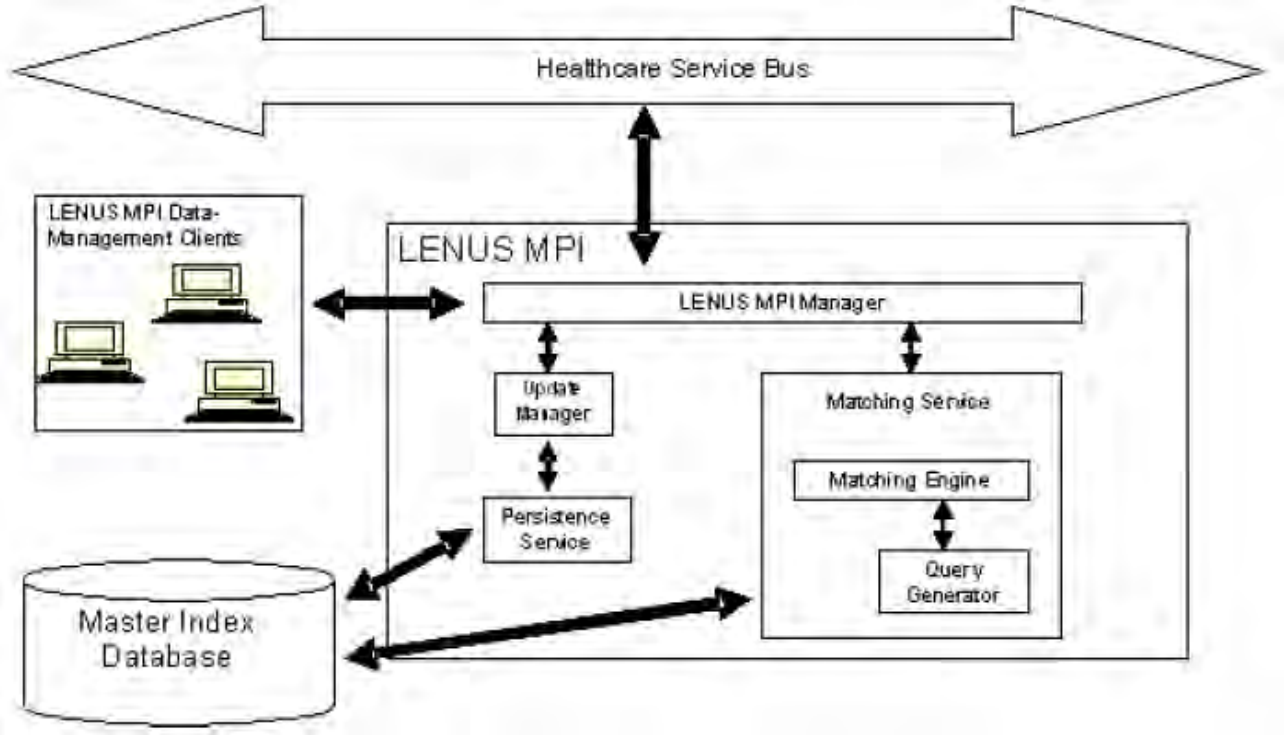

Figura 3: Arquitetura do LENUS MPI [Krechel e Hartbauer 2008].

Quando um sistema conectado transfere dados para o índice, o LENUS *MPI Manager* envia uma requisição para o *Matching Service*. O componente *Matching Service*, automaticamente, gera uma *query* para o *Master Index Database* e executa uma comparação *fuzzy* dos dados recuperados do banco com os dados recebidos.

O componente *Matching Engine* usa versões estendidas de métodos de busca associativa (*associative search*) e Raciocínio Baseado em Casos (CBR - *Case Base Reasoning*) para combinar partes estruturadas e desestruturadas dos dados recebidos.

Dependendo dos resultados comparados, uma combinação automatizada é executada ou há necessidade de intervenção humana para decidir entre as combinações candidatas. Para um paciente novo, um identificador único é gerado. Também é adicionada uma chave de referência para o sistema de origem dos dados. Todas as transações executadas no *Master Index Database* que são reportadas aos sistemas dependentes ocorrem através do *Healthcare Service Bus*. O componente *Update Manager* determina quais os sistemas devem ser informados sobre a execução de uma determinada transação.

O LENUS MPI faz o papel do *Patient Identifier Cross-reference Manager* (*PIX Manager*) no perfil IHE/PIX e suporta as transações *Patient Identity Feed*, *PIX Query* e *PIX Update Notification*, vistos na Seção 3.4.

2.2.2. Projeto ARTEMIS

O projeto ARTEMIS [Christophilopoulos 2005], [Eichelberg, Aden e Thoben 2005] e [ARTEMIS], fundado pela União Europeia, em 1º de janeiro de 2004, tem por objetivo fornecer interoperabilidade entre sistemas de informações clínicos de diferentes organizações, baseado em serviços de *Semantic Web Services* [McIlraith, Son e Zeng 2001] e ontologias de domínio. O projeto ARTEMIS conta com a parceria da Turquia, Alemanha, Grécia e Reino Unido.

Uma organização da área de saúde pode juntar-se à rede ponto-a-ponto (P2P) ARTEMIS e publicar serviços eletrônicos, tais como: fornecimento de acesso ao prontuário eletrônico de pacientes (mediante uma autorização adequada), acesso à admissão de pacientes e acesso ao sistema de laboratório. Com a rede ARTEMIS, serviços poderão ser invocados dinamicamente. Por exemplo, dados de pacientes poderão ser traduzidos e mapeados entre diferentes organizações de saúde. No ARTEMIS, grupos de organizações participantes estão acoplados a um *SuperPeer*, que estão ligados por sua vez a uma grande rede ponto-a-ponto.

A disponibilidade de acesso às informações clínicas em toda a organização e, possivelmente, até mesmo entre países, exige um serviço eficiente para localizar registros clínicos relativos a um determinado paciente. O acesso aos registros clínicos encontrados deve contar com alto índice de confidencialidade. Dentro da rede ARTEMIS, registros clínicos podem ser localizados através do *Patient Identification Protocol (PID Protocol)*, um protocolo criptográfico escalável, que permite identificar e localizar instituições que têm registros clínicos disponíveis para um determinado paciente. Quando um solicitante identificado localiza uma ou mais instituições que possuem registros clínicos do paciente desejado, o acesso ao registro é negociado pelo *PID Protocol*.

O ARTEMIS obedece a diversos padrões. Dentre eles:

- **GEHR/openEHR:** iniciativa que começou em 1992, como um projeto de pesquisa da União Europeia, que tem por finalidade promover a interoperabilidade entre os Registros Eletrônicos de Saúde (EHR – *Electronic Healthcare Record*). Atualmente, esse projeto é mantido pela fundação *open*EHR[11].
- **EHRcom:** Padrão *Electronic Healthcare Record Communication* (CEN/TC 251), desenvolvido pela CEN[12] (*European Committee for Standardization*).

[11] http://www.openehr.org/home.html

[12] http://www.cen.eu/cenorm/homepage.htm

- **ISO:** Padrões do Comitê Técnico de Informática em Saúde da ISO (*ISO Technical Committee on Health Informatics*[13]). Em especial àqueles padrões que versam sobre EHR.
- **MML:** *Medical Markup Language*. Padrão desenvolvido pelo Grupo de Pesquisa em Registro Eletrônico de Pacientes do Ministério de Saúde e Bem-Estar do Japão (*Electronic Health Record Research Group of the Japanese Ministry of Health and Welfare*).
- Padrões já citados: HL7, IHE e DICOM.

A integração entre o projeto ARTEMIS e o padrão IHE se dá pela integração entre os perfis RID (*Retrieve Information for Display*) e PIX. O perfil de integração RID permite acesso, somente leitura, aos dados clínicos dos pacientes. Ele suporta acesso a documentos nos mais diversos formatos: PDF, JPG, CDA, etc. Ele suporta também acesso a dados essenciais à vida do paciente, tais como, alergias, medicamentos atuais e resumo dos relatórios [ACC, HIMSS e RSNA 2008] e [ACC, HIMSS e RSNA 2008a].

Com a integração desses dois perfis e a adoção de alguns componentes do perfil PIX em cada nó da rede ARTEMIS, seria possível consultar um índice geral, gerido por um componente *PIX Manager* centralizado, e obter a localização dos pontos que contém informações sobre um determinado paciente. Contudo, devido ao dinamismo da informação, optou-se pela distribuição do índice de localização dos pacientes entre os pontos. Portanto, foi criado mais um componente, chamado *Patient Indentification Server* (PID Server) em cada nó da rede ARTEMIS. Esse componente, descrito na Figura 4, implementa o mapeamento da identificação do candidato semelhante ao perfil IHE PIX. Ele também é alimentado por informações de admissão, atualização e fusão de pacientes e mensagens ADT, conforme a seta 1 da Figura 4. Dependendo do tipo da mensagem ADT, um novo registro de paciente é criado, sua identificação é atribuída, juntamente com chaves de *hash* e algumas variáveis (seta 2, Figura 4). Essas variáveis representam os dados demográficos dos pacientes, úteis à sua identificação.

[13] http://www.iso.org/iso/standards_development/technical_committees.htm

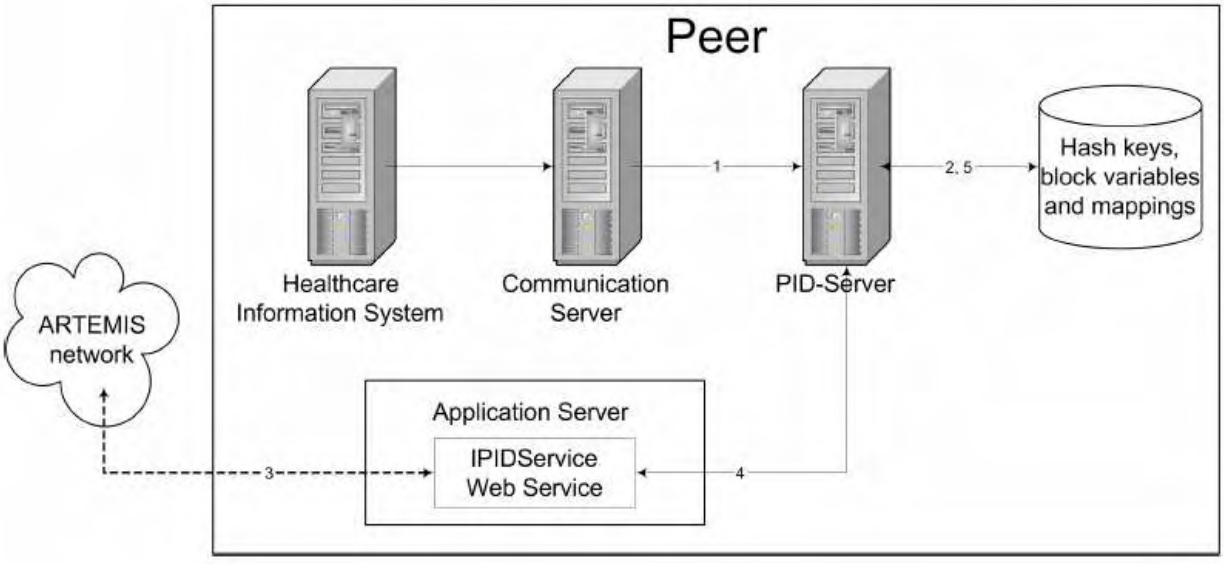

Figura 4: Componente PID Server, presente em cada nó da rede ARTEMIS [Christophilopoulos 2005].

Uma consulta é disparada de um *peer* ao outro, através de uma *PID Protocol query* (representada pela seta 3 da Figura 4). A responsabilidade do *Web Service IPIDService* é selecionar, a partir do bloco de variáveis, os pacientes que estão possivelmente relacionados a uma consulta (tarefa representada pelas setas 4 e 5 da Figura 4).

2.2.3. OHF

O *Eclipse Open Healthcare Framework* (OHF Eclipse)[14] [Renly 2007] e [Smith et al. 2006] é um projeto desenvolvido em Eclipse[15], com o propósito de disseminar a tecnologia de informática em saúde. O projeto é composto por ferramentas e *frameworks* que enfatizam a utilização de padrões existentes e emergentes, tais como IHE, HL7 e DICOM. O principal objetivo do projeto é incentivar a interoperabilidade de plataformas *open source*.

O OHF Eclipse, ou simplesmente OHF, é desenvolvido pela *Eclipse Fundation*[16], corporação sem fins lucrativos, que tem como objetivo incentivar a criação, evolução, promoção e suporte de soluções na Plataforma Eclipse. Essa fundação também tem como finalidade fornecer serviços, produtos e capacitação para a comunidade *open source*.

[14] http://www.eclipse.org/ohf/

[15] http://www.eclipse.org/

[16] http://www.eclipse.org/org

O projeto OHF implementa *frameworks* extensíveis e ferramentas para suportar padrões de informática em saúde, encapsuladas em *plugins*. Há *plugins* desenvolvidos que obedecem às especificações da *IHE IT Infrastructure Techinical Framework*. Mais especificamente, há *plugins* desenvolvidos que suportam transações definidas para o perfil de integração IHE/PIX.

Em linhas gerais, a arquitetura do OHF é composta por aplicações clientes, *gateways* e aplicações servidoras. No lado cliente, a forma convencional de usar o OHF é criar uma aplicação Eclipse RCP (*Rich Client Platform*)[17]. Essa aplicação segue o padrão de arquitetura de software MVC (*Model-view-controller*) [Freeman e Freeman 2007]. Na camada *Model*, são usados componentes OHF, conforme ilustrado na Figura 5.

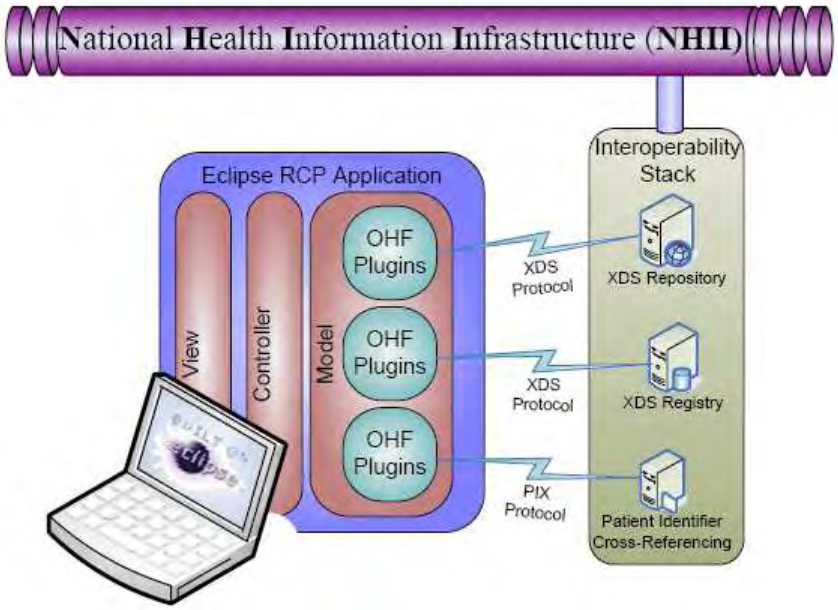

Figura 5: Componentes OHF embutidos em uma aplicação Eclipse RPC [Smith et al. 2006].

Muitas aplicações de saúde existentes atualmente foram desenvolvidas em ambiente .NET[18] ou em alguma outra linguagem de programação própria para plataformas cliente/servidor ou web. Poucas aplicações deste domínio utilizam ambiente Java e dificilmente empregam soluções Eclipse RCP. A proposta do projeto OHF para resolver esse problema é usar, no lado servidor, o Eclipse e o Apache Axis[19] para disponibilizar as funcionalidades dos componentes OHF via *WebServices*. Essa solução é chamada Ponte OHF (OHF *Bridge*)[20]. O OHF *Bridge* usa um componente em tempo de execução (*runtime*), chamado "*OSGI on Server*", que embute os *plugins* e expõe um subconjunto de

[17] http://www.eclipse.org/community/rcp.php

[18] http://msdn.microsoft.com/netframework/

[19] http://ws.apache.org/axis/

[20] http://www.eclipse.org/ohf/components/bridge/

suas funcionalidades como *WebServices*. Usando o OHF *Bridge*, soluções que suportam ambiente SOAP, como, por exemplo, aquelas desenvolvidas em PHP ou em .NET, podem usar componentes OHF. Essa arquitetura pode ser observada na Figura 6.

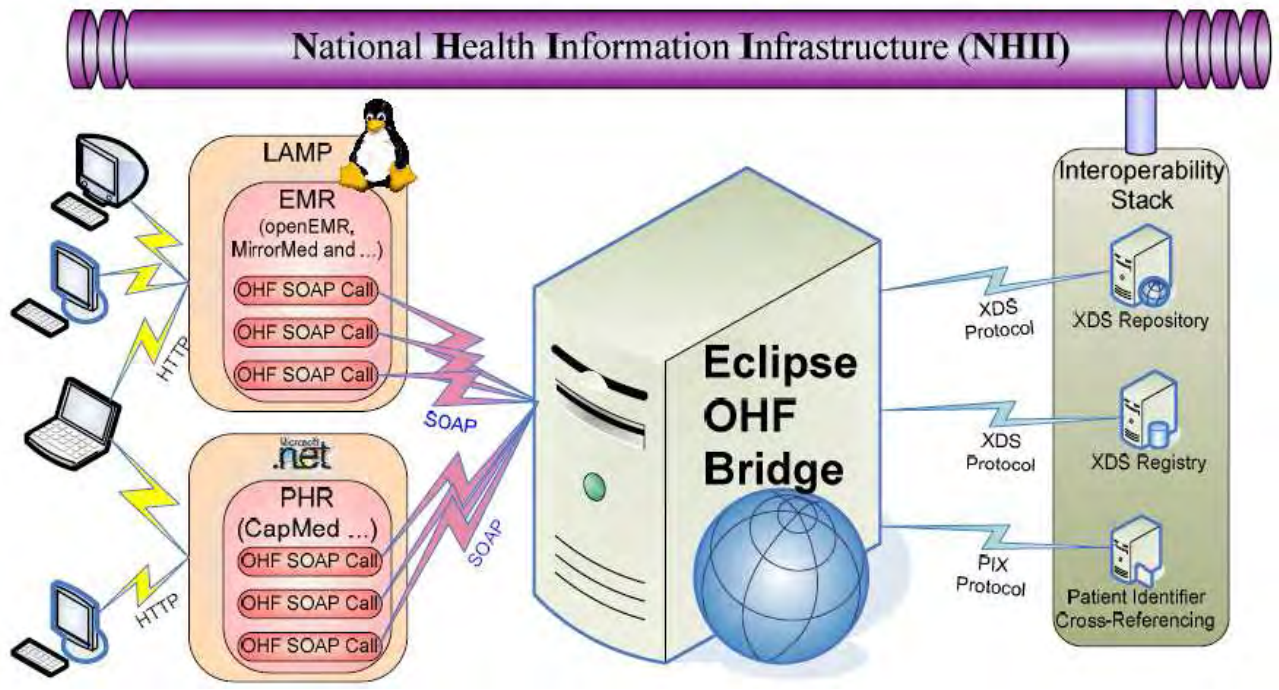

Figura 6: Componente OHF *Bridge* sendo consumido por aplicações que suportam SOAP [Smith et al. 2006].

O *National Health Information Infrastructure* (NHII)[21], presente nas Figuras 5 e 6, é uma iniciativa do Departamento de Saúde e Serviços Humanos do Governo Norte-Americando (*US Department of Health and Human Services*) de fornecer uma rede baseada em conhecimento, que permita a interoperabilidade entre sistemas de saúde. Esta rede tem por objetivo melhorar a tomada de decisão, tornando a informação de saúde disponível quando e onde ela for necessária.

2.3. Soluções de *Record Linkage*

O termo *Record Linkage* refere-se à tarefa de encontrar entidades de interesse iguais em duas ou mais fontes de dados. Esta técnica é utilizada quando não há entre conjuntos de dados diferentes uma chave única que as relacione, sendo necessário escolher campos que, uma vez comparados, indiquem o quanto é provável que um par de registros se refira a mesma entidade.

[21] http://aspe.hhs.gov/sp/NHII/

A primeira referência que cita o termo *Record Linkage* é encontrada no trabalho do Dr. Halbert Dunn, chefe do *U.S. National Office of Vital Statistics* (Dunn 1946), que descreveu, de forma metafórica, que cada pessoa possui um livro da vida (*Book of Life*) e o *Record Linkage* é o nome dado ao processo de montar as páginas deste livro em um único volume.

Newcombe et al. (1959) é um dos pioneiros no desenvolvimento de uma metodologia que permite combinar automaticamente registros de fontes de dados diferentes, encontrando indivíduos iguais nessas fontes. Newcombe e Kennedy (1962), ao publicar o artigo com o título *Record Linkage- Making Maximum Use of the Discrimination Power of Identifying Information*, associaram o método de combinação de registros ao termo *Record Linkage*.

Fellegi e Sunter (1969) estenderam os conceitos originais de Newcombe e deram tratamento matemático e formal ao processo de *Record Linkage* [Jaro 1989]. A técnica desenvolvida por eles permite a combinação probabilística de registros, tornado-se a mais aceita na literatura especializada.

Neste trabalho, o termo *Record Linkage* é utilizado de forma mais geral. Muitas vezes, o simples uso do termo *Record Linkage* se refere à técnica desenvolvida por Fellegi e Sunter (1969) de combinação probabilística de registros. Em outros casos, esse termo é utilizado para se referir a um software que emprega essa técnica. O uso da expressão *Record Linkage* de uma forma mais livre é adotada por vários autores e torna o texto menos prolixo.

Também é comum o uso do termo aproximação de registros, que se refere ao processo de combinar ou vincular registros que se relacionam com a mesma entidade ou evento em um ou mais conjunto de dados [Churches, Christen, Lim e Zhu 2002]. Usa-se o termo **deduplicação (*deduplication*)**, quando se quer buscar a mesma entidade em apenas um conjunto de dados e, como já visto, o termo ***Record Linkage*** quando a busca pela entidade ocorre em dois ou mais conjuntos de dados.

O detalhamento do método de *Record Linkage*, com a descrição de cada uma de suas etapas, é visto no Capítulo 4.

A seguir são analisadas as soluções de *Record Linkage* mais citadas na literatura e aquelas que mais se aproximam do projeto especificado neste trabalho.

2.3.1. Febrl

O *Febrl* (*Freely Extensible Biomedical Record Linkage*) [Christen 2008] e [Christen 2008a] vem sendo desenvolvido desde 2002, como parte de um projeto de pesquisa colaborativo entre a Universidade Nacional Australiana (*Australian National University*)[22] em Camberra e o Departamento de Saúde de Nova Gales do Sul (*New South Wales Department of Health*)[23] em Sydney, ambas instituições australianas. O objetivo desse projeto é desenvolver novas técnicas que melhore cada uma das fases do processo de *Record Linkage* entre bases de dados do domínio saúde.

Ao contrário dos softwares comerciais de *Record Linkage*, que se apresentam como caixas-preta (*black-box*), o projeto *Febrl* disponibiliza toda a documentação e o código-fonte. Com isso, as técnicas pesquisadas encontram-se disponíveis para a comunidade científica ou para instituições que tenham a necessidade de integrar bases de dados.

O sistema *Febrl* foi escrito na linguagem de programação *Python*[24]. Sua interface gráfica está baseada na biblioteca PyGTK[25] e no Glade toolkit[26]. A linguagem de programação *Phyton* foi escolhida por também possuir seu código fonte aberto e por contar com uma capacidade excelente para manipulação de *strings* (cadeia de caracteres). Além disso, possui outras características importantes, tais como: orientação a objetos; estruturas de dados do tipo conjunto, listas e arrays associativos (*associative arrays*); e facilidade de acesso a bancos de dados.

O *Febrl* utiliza várias técnicas de aproximação de registros para detectar registros duplicados [Martinhago 2006] e [Christen e Churches 2005]. Nele, os atributos utilizados para aproximação de registros podem ser categorizados em cinco classes: nome, endereços, datas e horas, atributos categóricos (tais como sexo ou nacionalidade) e quantidades escalares (tais como altura e peso). Recentemente, foi incorporado a essas categorias o número de telefone. Martinhago (2006) fez uma adaptação de algumas dessas classes para os padrões brasileiros.

Dois grandes passos devem ser seguidos para detectar registros duplicados no *Febrl*:

[22] http://www.anu.edu.au/index.php

[23] http://www.health.nsw.gov.au/

[24] http://www.python.org/

[25] http://www.pygtk.org

[26] http://glade.gnome.org

- **Limpeza e Padronização (*Data Cleaning and Standardisation*):** Neste passo, os dados são limpos e padronizados, conforme as classes citadas anteriormente. Então é gerado um novo conjunto de dados (*data set*), que é utilizado para o processo de deduplicação ou *Record Linkage*.
- **Aproximação de Registros:** Este passo é a busca por duplicação de registros, propriamente dita. Essa fase é composta por três etapas: Indexação (*Blocking*), Comparação e Classificação. Essas etapas são clássicas na busca de registros duplicados e são detalhadamente exploradas no Capítulo 4.

O *Febrl* também suporta paralelismo [Christen, Churches e Hegland 2004], empregando técnicas de computação paralela e de alto desempenho, tais como *clusters*, servidores multiprocessados e supercomputadores.

O paralelismo ocorre de forma transparente para o usuário. O padrão de troca de mensagens utilizado é o *Message Passing Inteface* (MPI)[27], biblioteca baseada em programação paralela e troca de mensagens. Para que a tecnologia MPI seja utilizada no ambiente Febrl, é necessária a instalação do módulo *Pypar*[28], que permite que scripts da linguagem *Python* sejam executados em paralelo. Contudo, segundo Martinhago (2006), o *Febrl* não consegue realizar a etapa de classificação de forma paralela. Cada processador fica esperando o processador responsável pelo bloco anterior concluir sua execução, para então, poder executar seu bloco.

2.3.2. RecLink

O RecLink [Camargo e Coeli 2000], [Camargo e Coeli 2006] e [EPSJV 2008] é um sistema brasileiro desenvolvido com a finalidade de relacionar bases de dados, fundamentado na técnica de relacionamento probabilístico de registros. O sistema foi escrito na linguagem de programação C++, utilizando o ambiente de programação Borland[29] C++ Builder[30]. Este software foi desenvolvido, inicialmente, por profissionais do Departamento de Planejamento e Administração em Saúde do Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro (UERJ)[31] e do

[27] http://www.mpi-forum.org/

[28] http://datamining.anu.edu.au/~ole/pypar/

[29] http://www.borland.com/

[30] http://www.codegear.com/tabid/123/Default.aspx

[31] http://www.uerj.br/

Departamento de Medicina Preventiva da Faculdade de Medicina e Núcleo de Estudos de Saúde Coletiva, Universidade Federal do Rio de Janeiro (UFRJ)[32].

O software consiste em uma interface com bancos de dados flexíveis que permite ao usuário designar, de modo interativo, as regras de associação entre duas tabelas. O programa foi avaliado a partir de dados coletados para um estudo de viabilidade da implantação de sistema de vigilância do *diabetes mellitus* na população idosa residente em uma determinada área da cidade do Rio de Janeiro. As fontes de dados empregadas neste estudo foram: o Sistema de Informação sobre Mortalidade (SIM), o Sistema de Informações Hospitalares do SUS (SIH-SUS) e estatísticas ambulatoriais provenientes de unidades de saúde que registraram casos de *diabetes mellitus* na área estudada. O gerenciador de banco de dados utilizado nessa primeira experiência foi o Visual dBASE[33]. Neste mesmo estudo, outros dados ambulatoriais foram utilizados.

Atualmente, o RecLink está em sua 3ª versão. Como nas versões anteriores, a mais recente também é desenvolvida na linguagem C++. Contudo, essa versão, utiliza o ambiente de programação *Borland Development Studio* e apresenta as seguintes melhorias:

- **Uso de assistentes:** Sequências de telas que orientam o usuário passo a passo na execução de uma determinada rotina.
- **Combinação:** Na nova versão do programa, as rotinas para combinação de arquivos, seleção de pares verdadeiros e geração de arquivos reduzidos para os passos de blocagem foram combinados em um grande painel de controle, facilitando a configuração por parte do usuário. Esse painel de controle, combinando várias configurações do RecLink pode ser observado na Figura 7.
- **Deduplicação:** A técnica de deduplicação foi incorporada ao *software*, permitindo que sejam identificados registros duplicados internamente em uma base de dados.

[32] http://www.ufrj.br/

[33] http://www.dbase.com/Index.asp

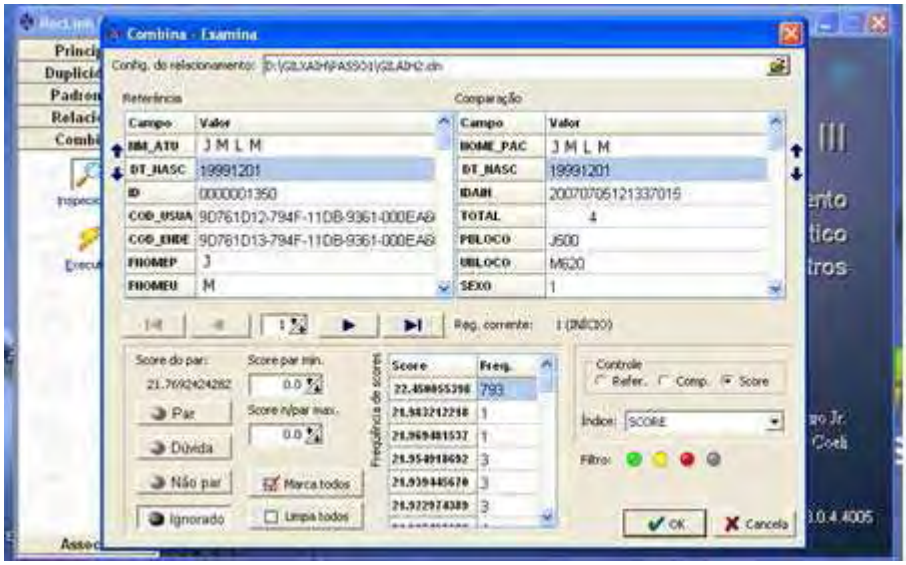

Figura 7: Painel de controle para configurar a combinação de registros no RecLink [Camargo e Coeli 2006].

Hoje em dia, o foco do RecLink é integrar bases de dados em saúde, tanto na área epidemiológica, quanto na área administrativa. Recentemente, o RecLink foi utilizado para analisar o perfil da população cadastrada no PSF (Programa Saúde da Família) em três unidades de saúde do município de Juiz de Fora/MG [Coeli et al. 2006]. A população coberta pelas unidades de saúde de Progresso, Santa Rita e Parque Guarani corresponde a 15.790 habitantes cadastrados no PSF em 2006.

2.3.3. FRIL

O *Fine-grained Record Integration and Llinkage* (FRIL) [Jurczyk et al. 2008] é uma ferramenta de *Record Linkage*, desenvolvida pela Universidade de Emory (*Emory University*)[34], localizada no estado da Geórgia - Estados Unidos. Essa ferramenta de código aberto (*open source*) tem como proposta associar técnicas tradicionais de *Record Linkage* com um rico e configurável conjunto de parâmetros. O FRIL oferece como vantagens:

- Funções avançadas para combinar tanto esquemas quanto dados. No caso dos atributos das fontes de dados, há a possibilidade de fundir dois atributos em um único, dividir partes de um único atributo ou substituir atributos por expressões regulares.
- Métodos de pesquisa customizáveis pelo usuário.

[34] http://www.emory.edu/home/index.html

- Suporte a sistemas cuja CPU possui mais de um núcleo (*multi-core systems*), de forma transparente para o usuário.
- Análise dinâmica dos parâmetros que foram configurados.

O FRIL é uma ferramenta baseada em Java, que possui uma interface gráfica que auxilia os usuários a compreender os efeitos causados pelos parâmetros escolhidos. A arquitetura geral do FRIL pode ser observada na Figura 8.

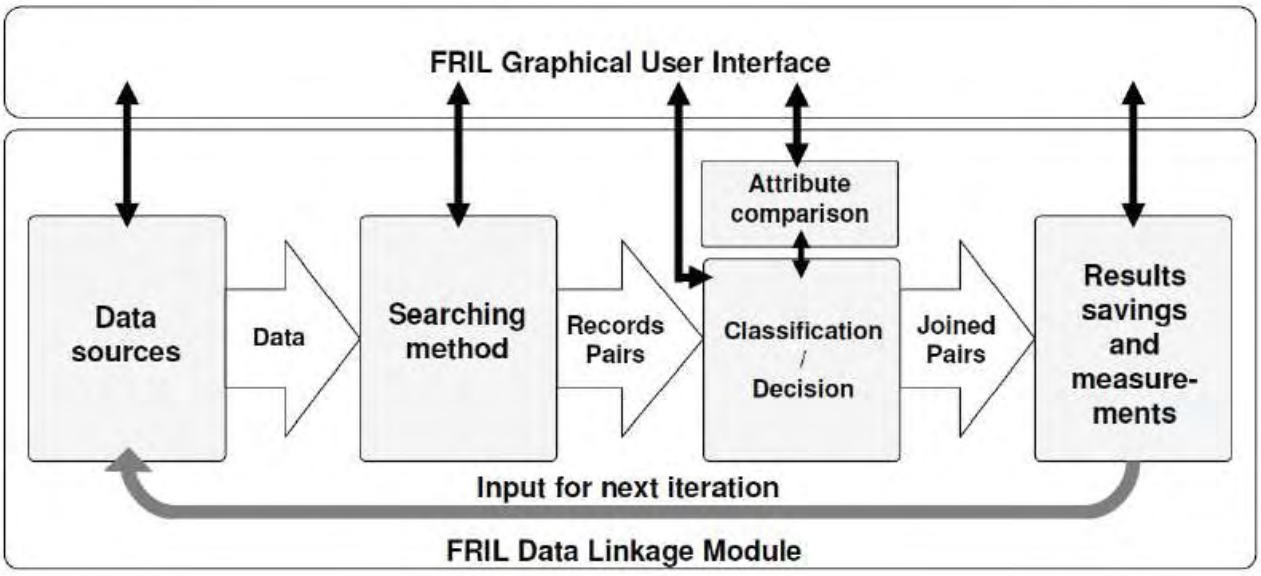

Figura 8: Arquitetura do FRIL [Jurczyk et al. 2008].

Observe que o usuário participa de todas as fases do processo de *Record Linkage*, seja configurando parâmetros seja analisando resultados. O FRIL conta ainda com uma retroalimentação (*Input for next iteration*), que permite com que resultados obtidos em um experimento sejam utilizados para o próximo.

A pesquisa realizada por Jurczyk et al. (2008) teve como objetivo obter possíveis ligações entre duas fontes de dados:

- A primeira fonte de dados diz respeito a casos de nascimento de crianças com defeitos congênitos, filhos de moradores de 5 condados centrais da cidade de Atlanta entre os anos de 1997 e 2006. Foram registrados mais de 12.700 desses casos pela MACDP (*Metropolitan Atlanta Congenital Defects Program*)[35]. O MACDP tem como propósito detectar e pesquisar casos de problemas congênitos ocorridos no nascimento. Esse programa foi criado

[35] http://www.cdc.gov/ncbddd/bd/macdp.htm

em 1967 pelas entidades: CDC (*Centers for Disease Control and Prevention*)[36], Universidade de Emory e Instituto de Saúde Mental da Geórgia (*Georgia Mental Health Institute*)[37].

- A outra fonte de dados é a base de certidões de nascimento e possui 1,25 milhões de registros de crianças nascidas no estado da Geórgia, também entre os anos de 1997 e 2006. A entidade detentora desses dados é o Departamento de Recursos Humanos da Geórgia (*Georgia Department of Human Resources*)[38].

Foram realizados quatro experimentos com as fontes de dados selecionadas. Os resultados dos três primeiros experimentos apresentaram uma eficácia superior a 95%. O quarto experimento contou com muitos falsos positivos, uma vez que foram levados em consideração os nomes de recém-nascidos desconhecidos, caracterizados com “Baby Smith” (ou “Smith B”).

Os experimentos demonstraram ainda uma diminuição significativa de tempo para execução do processo de *Record Linkage*, quando comparado com o processo manual. Inclusive, devido ao volume dos dados, a comparação manual pura e simples não foi possível, tendo sido usados softwares estatísticos adaptados para combinação de registros.

2.3.4. BigMatch

O programa BigMatch [Yancey 2004] e [Yancey 2007] é um *Record Linkage* desenvolvido pelo departamento de censo dos Estados Unidos (*U.S. Bureau of Census*)[39], cujo objetivo é extrair combinações plausíveis de fontes de dados de grande volume. Ele permite que sejam configurados diferentes critérios de blocagem. Esse software também é utilizado para descobrir duplicações em um arquivo único.

O BigMatch foi projetado para tirar vantagem da grande capacidade de memória dos computadores atuais. Em sua configuração padrão, o usuário é capaz de especificar as variáveis de blocagem, os campos que serão comparados e os parâmetros de combinação. Todos esses conceitos são abordados no Capítulo 4.

[36] http://www.cdc.gov/

[37] http://www.gmhcn.org/

[38] http://dhr.georgia.gov/portal/site/DHR/

[39] http://www.census.gov/

Contudo, as variáveis de blocagem, que determinam os blocos que serão comparados entre si, devem ser estabelecidas antes da execução do *Record Linkage* propriamente dito. Nesse caso, os dois arquivos que serão comparados são pré-ordenados, de acordo com os parâmetros de configuração. Após esse processo, é que os pares podem ser comparados. Se o usuário desejar outro critério de blocagem para combinação de registros, os arquivos devem ser ordenados novamente pelo novo critério, para depois ocorrer o processo de *Record Linkage*.

A justificativa para não fazer a ordenação dos arquivos pelo critério de blocagem durante a etapa de comparação de campos é, justamente, o fato do BigMatch manipular grandes fontes de dados. Segundo Yancey (2004), para arquivos muito grandes, o processo de ordenação pode demorar exageradamente. Mesmo para máquinas rápidas, ordenar 300 milhões de registros pode levar cerca de 12 horas. Outro fator que deve ser levado em consideração no processo de ordenação é o espaço em disco. Para que um arquivo seja classificado, é necessária uma área em disco de, aproximadamente, três vezes o seu tamanho original. O tempo de processamento também se reduz bastante, com o isolamento do processo de ordenação.

O processo de *Record Linkage* executado pelo BigMatch se mostra mais eficiente quando as operações são executadas para um arquivo **A** muito grande e um arquivo **B** de tamanho moderado. Entende-se por arquivo moderado, aquele que cabe totalmente na memória principal. Esse é, inclusive, o propósito original do programa BigMatch.

Outra característica peculiar do BigMatch é que cada bloco possui suas variáveis de blocagem, campos de comparação e parâmetros definidos pelo usuário. Para a ordenação de cada bloco, o BigMatch utiliza uma versão do *Quicksort* otimizada para *strings*, desenvolvida por Bentley e Sedgewick (1997).

O programa BigMatch é escrito na linguagem de programação C. Ele pode ser executado nas plataformas Windows, UNIX e VAX. Nas Workstations UNIX do *U.S. Bureau of Census*, são processadas 100 milhões de comparações de pares por segundo. Nesse órgão, há estações de trabalho com 32 gigabytes de memória principal, sendo que 100 milhões de registros do Censo Norte-Americano cabem em quatro gigabytes de memória.

Atualmente, o software BigMatch roda nas plataformas: *Compaq Alphas*, *Sun workstations* e Windows PCs. Em um processador Intel® Itanium®[40], utilizando, aproximadamente, 1,5 gigabytes

[40] http://www.intel.com/products/processor/itanium/index.htm

de memória RAM, o BigMatch é capaz de processar cerca de 300.000 pares por segundo [Yancey 2007].

Winkler e Yancey (2006) estão pesquisando versões paralelas da tecnologia BigMatch, com o objetivo de aumentar significativamente a velocidade de comparações de pares em relação ao BigMatch monoprocessado.

O BigMatch será maciçamente utilizado para o Censo de 2010 dos Estados Unidos. Um grande problema é a potencial duplicação de registros oficiais. O programa BigMatch identificará prováveis duplicações no registros de pessoas, utilizando dados do Censo de 2000 [Ikeda e Porter 2007].

O código-fonte, escrito em C, do programa BigMatch pode ser obtido através de contato com o *U.S. Bureau of Census*. Já o manual deste software pode ser encontrado no endereço http://www.census.gov/srd/papers/pdf/rrc2007-01.pdf.

2.4. Outros Trabalhos Relacionados

Além das soluções vistas anteriormente, existem trabalhos que utilizam outros conceitos e técnicas para identificar de maneira única dados de uma mesma entidade ou indivíduo em uma ou mais fontes de dados. Destacam-se, aqui, as soluções: **D-Dupe** e **SecondString**.

O D-Dupe [Bilgic et al. 2006] e [D-Dupe] é uma ferramenta que foi desenvolvida pelo Departamento de Ciência da Computação da Universidade de Maryland (*Department of Computer Science, University of Maryland*)[41], que combina algoritmos de mineração de dados (*data mining*) para Resolução de Entidade (*Entity Resolution* – ER) com uma visualização em rede de tarefas específicas.

A técnica de *Entity Resolution* [Benjelloun et al. 2005] localiza e combina registros que representam a mesma entidade no mundo real. O D-Dupe fornece aos usuários uma visão em rede, na qual exibe um contexto de colaboração para potenciais duplicatas (*Collaboration Context Network*). Esse contexto de colaboração revela para um par de potenciais duplicatas as suas relações de vizinhança. Essa visualização permite aos usuários identificar rapidamente com quem as prováveis duplicatas se relacionam. Isso facilita enormemente a decisão de se é ou não uma

[41] http://www.cs.umd.edu/

duplicata no mundo real. A Figura 9 mostra um exemplo de utilização do D-Dupe, que tem como objetivo identificar autores iguais em uma fonte de dados, com base na similaridade de seus nomes em suas relações de vizinhança. Usando esses critérios, o software estima que o autor George G. Robertson (indicado com o número 1) tem 95% de chance de ser a mesma pessoa indicada com o número 2: George C. Robertson. O percentual de similaridade entre os autores pode ser observado no painel de controle (*Control Panel*), localizado na parte superior direita da Figura 9.

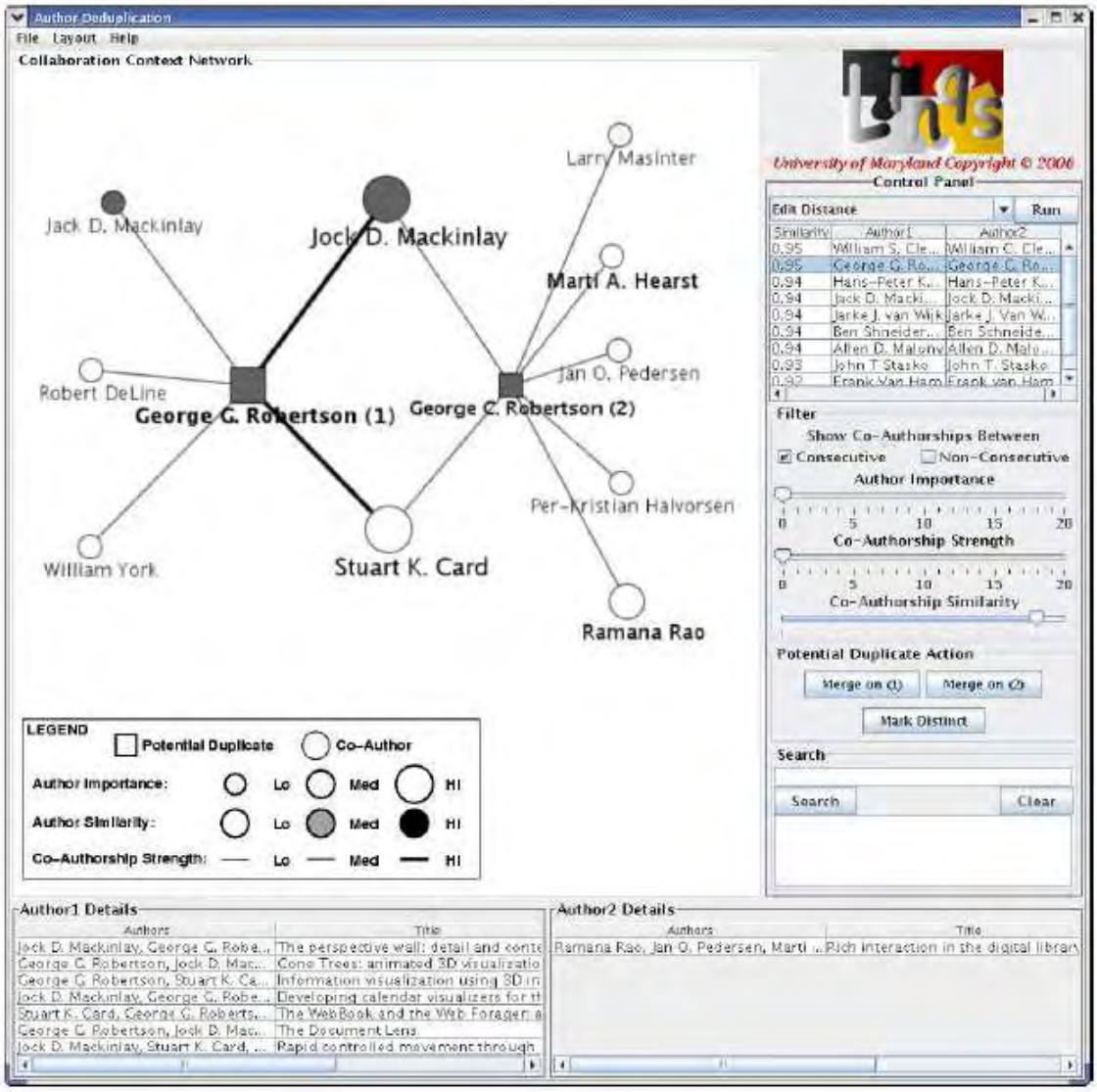

Figura 9: Rede de Contexto de Colaboração (*Collaboration Context Network*) de dois autores no D-Dupe [Bilgic et al. 2006].

O D-Dupe foi desenvolvido na linguagem de programação C#, no IDE (*Integrated Development Environment*) MS Visual Studio.Net 2005. O software é suportado pelos Sistemas Operacionais Windows XP ou Windows 2000.

O SecondString [Cohen, Ravikumar e Fienberg 2003] é uma ferramenta *open source* desenvolvida em Java, usando técnicas de *Record Linkage*. O SecondString foi desenvolvido pela Universidade de Carnegie Mellon (*Carnegie Mellon University*)[42] e comporta várias técnicas de cálculo de distância e similaridade de *strings*. Um destaque na sua arquitetura é o objeto *distance function learner*, que tem a capacidade de "aprender", à medida que processa a comparação de pares de *strings*.

2.4.1. Soluções na Área de Saúde

Ainda na área científica, existem diversos institutos de pesquisa que criaram ferramentas que combinam bases de dados através de métodos probabilísticos. Muitos desses centros de pesquisa estão vinculados à área de saúde e alguns deles pertencem a organismos governamentais.

O relacionamento entre dados e registros de fontes diferentes ganha uma especial importância no setor da saúde por possuir objetivos tais como: ajudar a melhorar as políticas públicas, detectar fraudes, reduzir custos e descobrir reações adversas de drogas [Christen 2008].

Neste sentido, as soluções pesquisadas que possuem maior destaque são:

a. Link Plus [SAUDE-SC 2004] e [Link Plus]**:** Este software é um *Record Linkage* probabilístico, desenvolvido pela Divisão de Controle e Prevenção do Câncer, um dos Centros para Controle e Prevenção de Doenças do Departamento de Saúde e Serviços Humanos dos Estados Unidos (*Division of Cancer Prevention and Control*[43], *Department of Health and Human Services*[44], *Centers for Disease Control and Prevention*[45]). O seu desenvolvimento também contou com o apoio do Programa Nacional de Registros de Câncer daquele país (NPCR - *National Program of Cancer Registries*)[46]. O Link Plus é uma aplicação desenvolvida para ambiente Microsoft® Windows®, que pode ser executada em dois modos: (*i*) para detectar registros duplicados de câncer em um banco de dados, ou (*ii*) para relacionar registros de câncer de um arquivo com arquivos externos. Embora originalmente concebido para ser utilizado para registros de câncer, esse software pode ser utilizado com qualquer tipo de dados em formato delimitado ou largura fixa.

[42] http://www.cmu.edu/index.shtml

[43] http://www.cdc.gov/cancer/

[44] http://www.hhs.gov/

[45] http://www.cdc.gov

[46] http://www.cdc.gov/cancer/npcr/

Atualmente, o Link Plus roda no Sistema Operacional Windows XP ou Vista e é necessária a instalação da versão mais recente do *web browser* Internet Explorer[47]. A linguagem de programação utilizada foi o Microsoft Visual Basic[48], versão 6, e muitas DLLs (*Dynamic Link Libraries*) incorporadas ao programa foram escritas em C. Os planos para evolução do Link Plus envolve: um estudo de usabilidade para tornar a interface mais amigável e eficiente; o desenvolvimento de métodos de combinação para telefones e endereços; conversão para .Net.

No Brasil, o Link Plus foi utilizado para avaliar o impacto nas taxas de incidência de tuberculose com a exclusão de registros indevidamente repetidos no sistema de notificação. Nesse estudo, foram analisados dados do Sistema de Informação de Agravos de Notificação (Sinan) do Ministério da Saúde, referentes ao período de 2000 a 2004. Os registros repetidos que foram identificados, foram classificados em seis categorias mutuamente exclusivas: falta de dados, duplicidade verdadeira, recidiva, reingresso, transferências entre unidade de saúde e inconclusiva. As categorias determinaram a remoção, vinculação ou permanência de cada indivíduo duplicado na base [Bierrenbach et al. 2007].

b. SALI [Dal Maso, Braga e Franceschi 2001]**:** Pesquisadores do Serviço de Epidemiologia do Centro de Referência do Câncer de Aviano na Itália (*Centro di Riferimento Oncologico, Aviano, Italia*)[49], e da Agência Internacional para Pesquisa do Câncer de Lyon na França (IARC - *International Agency for Research on Cancer*)[50] desenvolveram o software SALI - *Software for Automated Linkage in Italy*, com o objetivo de relacionar bases de dados de AIDS e de câncer da Itália. Essas bases de dados possuem cerca de 100 mil registros cada. O software foi desenvolvido em CA-Clipper e as bases de dados estão no formato DBF. O SALI possui um algoritmo específico que trata erros de soletração em italiano e em outras línguas latinas. O processo de *Record Linkage* do SALI possui uma grande sensibilidade, ou seja, ele é capaz de descobrir as combinações corretas entre as bases de dados. Por outro lado, o software captura um volume exagerado de falsos positivos, (sua especificidade é de cerca de 70%), exigindo que haja uma inspeção manual para descarte de pares incorretos.

[47] http://www.microsoft.com/brasil/windows/internet-explorer/default.aspx

[48] http://msdn.microsoft.com/en-us/vbasic/default.aspx

[49] http://www.cro.sanita.fvg.it/

[50] http://www.iarc.fr/

c. OpenEMed [OpenEMed]**:** É um conjunto de componentes distribuídos, utilizados por serviços de saúde, construído de acordo com as especificações da OMG™ (*Object Management Group*)[51] e do HL7. A OMG™ é um organismo internacional, sem fins lucrativos, de adesão aberta, responsável pelo desenvolvimento e disseminação de padrões para integração corporativa. Dentre as especificações de serviços normatizadas pela OMG™, está o Serviço de Identificação de Pessoas (PIDS - *Person Identification Service*) [OMG 2001]. O PIDS é usado por sistemas de informação hospitalares, quando há a necessidade de identificar os pacientes para acesso aos seus dados clínicos, administrativos e financeiros. Isso ocorre porque, em muitos casos, um mesmo hospital pode possuir vários sistemas heterogêneos, que foram desenvolvidos ou comprados em épocas distintas, utilizando linguagens e ambientes operacionais diferentes [Cardoso e Sabbatini 1999].

O PIDS, bem como os demais serviços que compõem o OpenEMed, foi desenvolvido em Java, na plataforma J2EE. Inicialmente, o projeto OpenEMed foi construído pela LANL (*Los Alamos National Labs*)[52]. Atualmente, o OpenEMed é um software *open source*, cujos fontes e documentação encontram-se disponíveis em http://sourceforge.net/projects/openmed.

d. PIDS do InCor [Fiales et al. 2001]**:** O Complexo do Hospital das Clínicas (HC)[53] da Faculdade de Medicina da Universidade de São Paulo (FMUSP)[54] ocupa uma área total de 352 mil metros quadrados com cerca de 2.200 leitos distribuídos entre os seus seis institutos especializados, dois hospitais auxiliares, uma divisão de reabilitação e um hospital associado [HC-FMUSP]. A exemplo do OpenEMed, o InCor, um dos institutos especializados do Hospital das Clínicas de São Paulo, desenvolveu um PIDS que integra todo o Complexo HC.

A necessidade da construção de um sistema onde os pacientes atendidos em uma das instituições participantes do complexo pudessem ser encontrados nas demais instituições, sem a utilização de um sistema centralizado, levou à construção de uma estrutura de federação. Nessa federação, um novo paciente, cadastrado em qualquer domínio, pode ser encontrado pelos demais sistemas. Para que a identificação do paciente fosse possível, foi necessário estabelecer uma estrutura de correlação. Nessa arquitetura, diferentes identificadores, de um mesmo paciente, são correlacionados em um domínio especial, chamado domínio de correlação. Esse

[51] http://www.omg.org/

[52] http://www.lanl.gov/

[53] http://www.hcnet.usp.br/

[54] http://www.fm.usp.br/

domínio tem a função singular de correlacionar tais identificadores nos demais domínios. Na Figura 10, é mostrado o esquema de federação do PIDS adotado no Hospital das Clínicas de São Paulo.

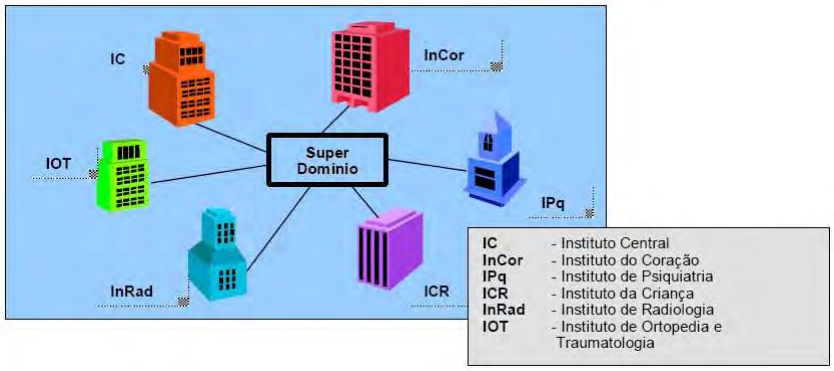

Figura 10: Esquema de federação do PIDS no Complexo HC [Fiales et al. 2001].

Nesse esquema de federação do PIDS, cada instituto do HC é um domínio a ser correlacionado. O *Super Domínio* é a entidade que representa o domínio de correlação. Nele está a base de dados que possui as informações básicas de identificação de um paciente (nome do paciente, nome da mãe, data de nascimento, local de nascimento e sexo, domínio do paciente - ex.: iot.usp.br, incor.usp.br, etc. -, identificador do paciente no domínio cadastrado, identificador do paciente no *Super Domínio*). Os Institutos são entidades que possuem seu próprio sistema de identificação de pacientes implementado. Além disso, os Institutos possuem acesso ao sistema de identificação de pacientes do *Super Domínio*. Com isso, é possível a identificação de pacientes a partir de qualquer site da federação.

Para implementação do PIDS, foi utilizada a linguagem Java. A implementação do padrão CORBA[55] ficou a cargo da ferramenta Inprise Visibroker®[56]. O ambiente de desenvolvimento utilizado foi o JBuilder®[57] da Borland.

2.4.2. Soluções Comerciais

Além das iniciativas governamentais e acadêmicas para combinar registros de fontes de dados heterogêneas, há várias soluções comerciais que implementam *Record Linkage*, IHE/PIX e geração de MPI.

[55] http://www.corba.org/

[56] http://www.borland.com/us/products/visibroker/index.html

[57] http://www.borland.com/br/products/jbuilder/

Dentre os softwares de *Record Linkage* comerciais mais citados na literatura, destacam-se o LinkageWiz [LinkageWiz] e o The Link King© [Campbell 2005]. O LinkageWiz foi desenvolvido pela LinkageWiz Software e pode ser utilizado tanto para combinação de bases de dados diferentes, quanto no processo de deduplicação. O LinkageWiz é utilizado por empresas, agências governamentais, universidades e outras organizações nos Estados Unidos, Canadá, Reino Unido, Austrália e França. O software possui uma versão gratuita que combina arquivos de até 2 mil registros. Acima desse patamar, o software é comercializado em função do volume de registros que os arquivos que serão combinados possuem.

The Link King© é um software desenvolvido pela MEDSTAT, originalmente para um projeto de integração de bases de dados da Administração de Serviços de Abuso de Substâncias e Saúde Mental (SAMHSA - *Substance Abuse and Mental Health Administration*)[58]. Embora seja de domínio público, para que The King Link© seja executado, é necessário que seja instalada uma versão atualizada do SAS/AF®[59], que é um gerador de aplicativos comercial. The King Link© possui algumas características interessantes, como a identificação de apelidos (*nickname identification*). Esse recurso faz com que nomes como "Bill", "William", "Billy", "Will" etc. sejam tratados de forma equivalente.

As grandes corporações de software também possuem soluções de mercado para combinação probabilística de registros e geração de MPI.

A Sun Microsystems[60] possui uma solução para implementação do IHE/PIX e geração de MPI, chamada *IHE Compliant Master Patient Index* [Sun], que é baseada no produto *Sun Java*™ *Composite Application Platform Suite (CAPS)*[61]. A solução de MPI da Siemens Medical[62] é o produto *Soarian® Integrated Care (Soarian IC)* [Siemens 2007].

A Microsoft apresenta um gerador de MPI integrado a uma solução chamada *Microsoft Connected Health Framework Architecture and Design Blueprint* [Microsoft 2009]. A Figura 11 mostra a geração de um índice central, em um ambiente de federação. Esse índice central é responsável pela tradução entre os diferentes identificadores que podem ser utilizados pelos nós

[58] http://www.samhsa.gov/

[59] http://www.sas.com/software/

[60] http://www.sun.com/

[61] http://www.sun.com/software/javaenterprisesystem/javacaps/

[62] http://www.medical.siemens.com

participantes. Dessa forma, a partir de cada fonte de dados, é possível consultar dados do paciente contidos em toda a federação. O nó central também pode conter os metadados relevantes, associados a cada nó: descrição do dado, natureza do dado, tipo, tamanho, etc. Isso permite filtrar os pedidos e enviá-los apenas para os nós que possuam itens relevantes, como, por exemplo, resultados laboratoriais, em vez do histórico médico completo do paciente.

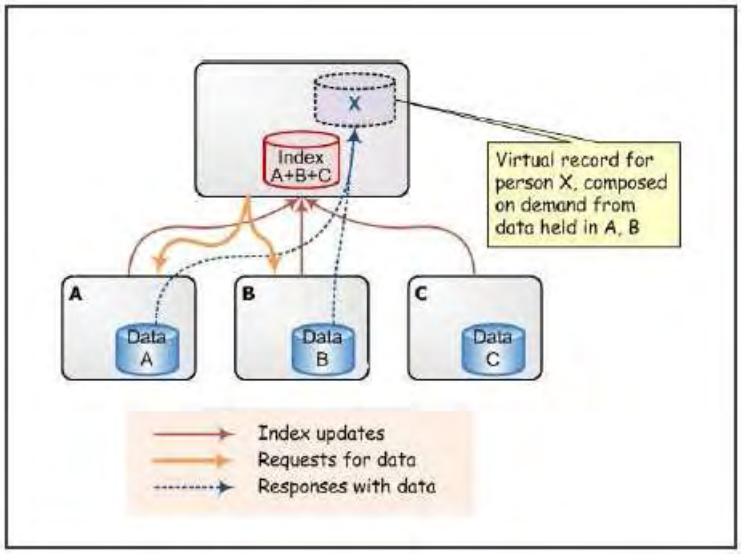

Figura 11: Modelo de federação com índice central, proposto pela Microsoft [Microsoft 2009].

2.5. Conclusão

Neste Capítulo, apresentou-se, resumidamente, as soluções de IHE/PIX, *Record Linkage* e Geração de MPI existentes na literatura especializada ou propostas pelo mercado.

Houve muita dificuldade de encontrar perfis de integração IHE/PIX implementados. É possível concluir que essas implementações ainda sejam incipientes. Um argumento que reforça essa conclusão é o fato de que nos resultados dos testes de produtos realizados no *IHE Asia-Pacific 2008 Connectathon*[63], principal evento para consolidação do padrão IHE no mercado, houve apenas um produto que implementa o ator *PIX Manager* (Seção 3.4).

Quanto aos geradores de MPI, as soluções comerciais respondem mais rapidamente do que a academia, sobretudo com produtos que tem como objetivo integrar uma única instituição de saúde.

[63] http://www.ihe.net.au/Results-2008.htm

As soluções de *Record Linkage*, por sua vez, são diversificadas e encontram-se em um nível de maturidade bastante razoável. Contudo, essas soluções ainda são usadas de forma isolada e para objetivos específicos. Soluções que integrem grandes bases relacionadas à saúde pública ainda são um campo a ser explorado.

A partir das soluções vistas, é possível estabelecer um quadro-resumo (Tabela 1) com destaque para a disponibilidade e ambiente operacional de cada software.

Tabela 1: Quadro-resumo das soluções de IHE/PIX e *Record Linkage*.

Solução	Abordagem	Disponível na Web				Linguagem de Programação	Sistema Operacional
		Fontes	Software*	Manual	Tutorial		
Lenus (SER)	IHE/PIX	Não	Não	Não	Não	Não identificada	Todos os SOs podem ser conectados
Artemis (União Europeia)	IHE/PIX	Não	Não	Não	Não	Java	Não especificado
OHF (IBM)	IHE/PIX	Sim	Não	Sim	Sim	Java	Windows, Linux e Mac
Febrl (Universidade Nacional Australiana)	*Record Linkage*	Sim	Sim	Sim	Sim	Python	Os principais
RecLink (IESC/UFRJ)	*Record Linkage*	Sim	Sim	Sim	Sim	C++	Windows
FRIL (Universidade de Emory)	*Record Linkage*	Sim	Sim	Sim	Sim	Java	Os principais
BigMatch (U.S. Bureau of Census)	*Record Linkage*	Enviado por e-mail	Não	Sim	Não	C	UNIX, VAX e Windows
D-Dupe (Universidade de Maryland)	*Entity Resolution*	Não	Sim	Não	Em vídeo	C#	Windows
SecondString (Univesidade de Carnegie Mellon)	*Record Linkage*	Sim	Não	Não	Não	Java	Independente de Plataforma
Link Plus (CDC)	*Record Linkage*	Não	Sim	Não	No próprio software	Visual Basic 6	Windows
SALI (IARC)	*Record Linkage*	Não	Não	Não	Não	CA-Clipper	MS-DOS e Windows
LinkageWiz (LinkageWiz Software)	*Record Linkage*	Não	Sim (com limitações)	Sim	Sim	Visual Basic 6	Windows
The Link King© (MEDSTAT)	*Record Linkage*	Não	Sim	Sim	Sim	SAS/AF®	Windows
OpenEMed (LANL)	PIDS	Sim	Sim	Sim	Não	Java	Independente de Plataforma
PIDS do InCor	PIDS	Não	Não	Não	Não	Java	Windows

* Software para instalação

Foram também investigadas as funcionalidades de cada solução e observou-se que a estratégia utilizada neste trabalho une a possibilidade das fontes de dados pertencerem a qualquer domínio ou formato com a capacidade de consulta às fontes originais a partir do índice único gerado, o que é raro nas implementações pesquisadas. Além disso, este trabalho não usa arquiteturas proprietárias, como barramentos, *bridges* ou camadas de hardware e software dependentes de um único fabricante.

3. Referencial Teórico

Neste Capítulo, são analisados alguns conceitos e tecnologias abordados ou empregados neste trabalho. Também são apresentadas ações atuais para adoção de uma identificação única de usuários de saúde, sobretudo, do governo brasileiro. Alguns padrões para área de Informática em Saúde também são descritos aqui.

3.1. Estratégias e Políticas para Identificação de Pacientes

O primeiro aspecto que deve ser observado quando se pensa na integração de dados de saúde é o que integrar. Nada adianta promover a identificação única de pacientes se seus dados encontram-se armazenados de maneira inadequada. Portanto, é necessário que os dados dos pacientes estejam em meio eletrônico e minimamente estruturados para que possam ser recuperados e combinados com outras bases de dados.

Existem dois instrumentos que contribuem para a padronização dos dados dos pacientes. O primeiro é denominado Prontuário Eletrônico do Paciente (PEP), ou EPR – *Eletronic Pacients Record*. O segundo é o Registro Eletrônico de Saúde (RES), ou EHR – *Electronic Healthcare Record*.

Segundo Massad et al. (2003), o PEP é uma estrutura eletrônica para manutenção de informações sobre o estado de saúde e o cuidado recebido por uma pessoa durante toda a sua vida. Os dados armazenados no PEP possuem vários formatos, estão em locais distintos e foram incluídos em épocas diferentes, por diversos profissionais de saúde. As principais vantagens dos prontuários eletrônicos vêm (*i*) do avanço no processo de tomada de decisão, uma vez que permite rápido acesso aos dados do paciente; (*ii*) da melhoria dos tratamentos e do atendimento aos pacientes; (*iii*) da redução de custos, com a otimização dos recursos.

No Brasil, a implantação do PEP teve início em 1999, com a *Recomendação Final do Comitê de Padronização de Registros Clínicos sobre a SOP 001 – Versão 1.0* [MS 1999]. Em 2003, foi realizada a 11ª Conferência Nacional de Saúde. Nessa ocasião o Ministério da Saúde estruturou a "Política Nacional de Informação e Informática em Saúde" (PNIIS). Essa política, dentre outras coisas, estabelece a geração automática dos registros eletrônicos para os SIS (Sistemas de Informação em Saúde) nacionais. Atualmente, o PNIIS está em sua versão 2.0 [MS 2004], que foi

deliberada na 12ª. Conferência Nacional de Saúde e PPA do Ministério da Saúde, ocorrida em março de 2004.

Para Peter Waegemann (1996) e Massad et al. (2003), a construção de um prontuário eletrônico é um processo evolutivo de 5 níveis:

- **Nível 1 - Registro Médico Automático:** o formato do prontuário é em papel, apesar do fato de que aproximadamente 50% das informações tenham sido geradas por computadores. Desta forma, papel e registro eletrônico coexistem.
- **Nível 2 – Sistema de Registro Médico Computadorizado:** muito semelhante ao nível 1, exceto pelo fato de que incorpora imagens capturadas via *scanners*. Em geral, esse tipo de sistema é departamentalizado, com pouca integração.
- **Nível 3 – Registro Médico Eletrônico:** diferentemente do nível anterior, requer que o sistema esteja implantado na instituição toda e contenha elementos como integração com sistema de gerenciamento da prática, sistemas especialistas como alertas clínicos e programas de educação ao paciente. Nesse nível, os requisitos de confidencialidade, segurança e proteção dos dados são atendidos.
- **Nível 4 – Sistema de registro eletrônico do paciente:** o escopo de informação presente é maior do que o suposto registro médico. As informações constantes vão além das paredes da instituição que está atendendo o paciente. Assim, esse nível requer que a identificação do paciente seja única e feita a nível nacional.
- **Nível 5 – Registro eletrônico de saúde:** inclui uma rede de fornecedores e locais, tendo o paciente como centro. A informação não é baseada somente nas necessidades do serviço de saúde; é baseada na saúde e doença do indivíduo e da comunidade.

O segundo instrumento que contribui para a padronização dos dados dos pacientes é, exatamente, o Registro Eletrônico de Saúde, ou RES, que é o nível final na escala evolutiva do PEP, proposto por Waegemann (1996).

Segundo o Conselho Federal de Medicina (CFM)[64], a definição de RES é

> *"O registro longitudinal da vida de uma pessoa, não-somente dos eventos relacionados à doença, mas também de informações de saúde, tais como hábitos alimentares, prática desportiva e atividades de lazer. Todos os eventos relacionados*

[64] http://www.portalmedico.org.br/novoportal/index5.asp

> *à saúde da pessoa devem estar registrados neste prontuário, do nascimento à morte, agregados em torno de um identificador único. A informação deve estar representada de tal modo que a troca entre instituições e a recuperação de dados seja feita de forma transparente para aqueles que estiverem acessando a informação. Acima de tudo, o RES deve atender aos requisitos essenciais de integridade, autenticidade, disponibilidade e privacidade da informação.* [CFM 2002].

Pereira (1995) relata que, na Inglaterra a localização das pessoas é feita através do "Registro Central" do Serviço Nacional de Saúde, que cobre todo o país. Trata-se de um registro populacional dentro do sistema de saúde, ao qual outras bases de dados estão relacionadas, tais como a de mortalidade, fatores de risco ou práticas de saúde. O autor salienta ainda que nos países do norte da Europa, cada cidadão recebe um número que o acompanha do nascimento à morte. Esse processo de identificação é particularmente útil em investigações epidemiológicas, pois as perdas são reduzidas ao mínimo [Oliveira 2007]. Na Áustria, quando um paciente é admitido em um hospital ou é consultado por um médico, ele é identificado pelo seu *e-Card*, que é um cartão inteligente que identifica o paciente em todo o país. Quando o paciente é identificado, os direitos de acesso do médico que o está assistindo são verificados. Então, a partir daí, o médico tem acesso a documentos clínicos do paciente [Stolba e Schanner 2007].

No Brasil, a principal iniciativa para identificação única do paciente em nível nacional é o Cartão Nacional de Saúde (CNS), também conhecido como Cartão SUS [MS-CNS]. Há também outras iniciativas, visando à integração de bases de dados de saúde, como o Cadastramento Nacional de Usuários do Sistema Único de Saúde [DATASUS 2009], que é o primeiro passo para a implantação do Cartão SUS em todo território nacional, e a construção do Sistema de Informações Epidemiológicas – SIEPI, a partir de grandes bancos de dados nacionais.

3.1.1. Cartão Nacional de Saúde

Gestores estaduais e municipais planejam e executam várias ações que utilizam a tecnologia da informação com a finalidade de atender necessidades da área de saúde. Contudo, o surgimento de sistemas de informações locais, com diversidade de padrões e protocolos, pode prejudicar o desenvolvimento de um RES nacional. O Cartão SUS é uma ação do governo federal que vem atender à necessidade de implementação de um padrão de informações único, utilizado por todos os usuários de saúde do Brasil.

O Cartão Nacional de Saúde é um instrumento que possibilita a vinculação dos procedimentos executados no âmbito do Sistema Único de Saúde (SUS) ao usuário, ao profissional que os realizou e também à unidade de saúde onde foram realizados [CNS]. O CNS foi instituído pela portaria N.º 1.560/GM, de 29 de agosto de 2002. As portarias 1.589/GM de 3 de setembro de 2002 e 1.740/GM de 2 de outubro de 2002 revisam ou complementam a portaria original. Já a portaria SIS/SE nº 39, publicada em 19 de abril de 2001, trata da operacionalização do processo de cadastramento nacional e traz os termos de adesão, municipal e estadual, bem como o manual de preenchimento do formulário.

Além de haver um cartão magnético (físico) que permite a identificação do usuário de saúde, o qual pode ser observado na Figura 12, há o sistema do Cartão Nacional de Saúde (SCNS), que possui os seguintes objetivos [CNS]:

- Construção de uma base de dados de histórico clínico;
- Imediata identificação do usuário, com agilização no atendimento;
- Ampliação e melhoria de acesso da população a medicamentos;
- Possibilidade de revisão do processo de compra de medicamentos;
- Integração de sistemas de informação;
- Acompanhamento dos fluxos assistenciais, ou seja, acompanhamento do processo de referência e contra-referência dos pacientes;
- Revisão dos critérios de financiamento e racionalização dos custos;
- Acompanhamento, controle, avaliação e auditoria do sistema e serviços de saúde;
- Gestão e avaliação de recursos humanos.

O tripé do SCNS são três cadastros que identificam univocamente os usuários do SUS: 1) o de usuários do SUS, com a geração de um número único de identificação, de âmbito nacional; 2) o de unidades de saúde; 3) e o de profissionais que executam procedimentos no sistema [Oliveira 2007], [CFM 2002] e [MS-CNS].

Figura 12: Cartão de usuários do SUS.

Um objetivo do projeto Cartão Nacional de Saúde é promover a integração entre os sistemas de informação utilizados no âmbito do Sistema Único de Saúde, sejam eles sistemas de base nacional ou sistemas de uso local. Para que tal objetivo seja viabilizado, o sistema demanda a definição de um conjunto de padrões de representação e troca de informação. A padronização compreende não apenas os aspectos de hardware e software (que devem obrigatoriamente ser abertos), mas, também, os aspectos de representação, transmissão, acesso e armazenamento da informação em saúde [CNS].

Quanto à arquitetura, o CNS está baseado em cinco componentes principais:

- Os cartões de identificação dos usuários e profissionais;
- Os terminais de atendimento e os equipamentos para armazenamento e tratamento da base de dados (servidores);
- Os softwares;
- A rede de comunicação;
- Os aspectos de segurança.

Os cartões de identificação dos usuários e dos profissionais são lidos por um terminal especialmente desenvolvido para o projeto. Esses cartões utilizam a tecnologia de tarja magnética exclusiva para leitura e refletem a preocupação com diversas variáveis, dentre elas durabilidade, custo e controle de uso. Os cartões são instrumentos de identificação e não de armazenamento de informações. O cartão do profissional traz embutida uma senha para acesso ao sistema.

Há diversos desafios para a efetiva implantação do Cartão SUS no Brasil. Dentre esses desafios, destacam-se a dificuldade de identificação de recém-nascidos e pessoas que não possuem documentos de identificação oficiais. Há também dificuldade de integração das bases de dados dos prestadores privados de serviços de saúde. Outros desafios relacionam-se à negociação do

Ministério da Saúde com estados, municípios e entidades profissionais; ao gerenciamento de contratos com terceiros; a falta de capacitação na área de informática; a insuficiência de recursos humanos; a baixa qualidade das linhas telefônicas e as definições quanto a padrões de informática e políticas de acesso. O grande número de desafios conceituais, operacionais e de gestão na implantação do Cartão SUS é decorrente do grau de inovação, magnitude e complexidade do projeto [Cunha 2002].

3.1.2. Cadastramento Nacional de Usuários do Sistema Único de Saúde

O cadastramento consiste no processo por meio do qual são identificados os usuários do Sistema Único de Saúde e seus domicílios de residência.

Com o cadastro, é possível a emissão do Cartão Nacional de Saúde para os usuários e a vinculação de cada usuário ao domicílio de residência, permitindo uma maior eficiência na realização das ações de natureza individual e coletiva desenvolvidas nas áreas de abrangência dos serviços de saúde.

O Cadastramento permite ainda a construção de um banco de dados para diagnóstico, avaliação, planejamento e programação das ações de saúde. A realização de um cadastramento domiciliar de base nacional, aliado à possibilidade de manutenção dessa base cadastral atualizada, pode permitir aos gestores do SUS a construção de políticas sociais integradas e intersetoriais (educação, trabalho, assistência social, tributos etc.) nos diversos níveis de governo [DATASUS 2009].

Além de ser o primeiro passo para a implantação do Cartão Nacional de Saúde em todo o território nacional, o Cadastro Nacional de Usuários é uma ferramenta importante para a consolidação do Sistema Único de Saúde (SUS), facilitando a gestão do sistema e contribuindo para o aumento da eficiência no atendimento direto ao usuário.

Há uma série de produtos (*softwares*) desenvolvidos que contribuem para a implantação do Cadastro dos usuários do SUS. A Tabela 2 apresenta um quadro com os softwares disponíveis atualmente, com suas respectivas descrições [DATASUS 2009a].

Tabela 2: Produtos desenvolvidos para suportar a implantação do Cadastramento Nacional de Usuários do SUS.

Software	Descrição
CadSUS	Aplicativo de cadastro e manutenção de usuários do Sistema Único de Saúde e seus domicílios de residência.
CadSUS Simplificado	Cadastro de usuários em Unidades de Saúde. Sem informações de domicílio, mas com obrigatoriedade de motivo de cadastramento e de Número Provisório.
CADWEB	Cadastramento via Internet dos usuários do Sistema Único de Saúde – SUS. É mais uma ferramenta da metodologia de implantação do Cartão Nacional de Saúde em todo o território nacional.
Crítica	Aplicativo recomendado para Municípios ou Unidades de Saúde que já possuem sistema de informação e desejam apenas enviar o conteúdo de suas bases.
Centralizador	Entrada de dados cadastrais captados a nível municipal por aplicativos externos ao sistema CADSUS, para obtenção do número do Cartão Nacional de Saúde.
Corretor	Aplicativo utilizado para correção de registros recusados pela Caixa Econômica Federal (CEF)[65]. Permite que haja uma correção descentralizada, mesmo nos municípios que não centralizaram as informações. Segundo a Portaria N.º 1.560/GM, que instituiu o Cartão SUS, o número individual de identificação gerado tem por base o Número de Identificação Social – NIS, administrado pela Caixa Econômica Federal, acrescido de 4 (quatro) dígitos de uso exclusivo da saúde.
Transmissor	Aplicativo que possibilita o envio e a recepção de arquivos nos aplicativos centralizadores de informações.
BNL (Base Nacional Leve)	Mais do que uma Base de Dados, a BNL é um projeto que envolve soluções de negócios e tecnologia que permitirão ao DATASUS[66] prover estados, municípios e outros sistemas internos do próprio DATASUS de informações necessárias para controle e gestão da saúde.
Data Mart	Aplicativo de geração de base de consulta e extração de dados, utilizando o TabWinSql[67].

3.1.3. Sistema de Informações Epidemiológicas

O objetivo da construção do Sistema de Informações Epidemiológicas (SIEPI) é o de produzir informações epidemiológicas para a tomada de decisão por parte dos gestores do SUS e demais atores da saúde suplementar [MS 2008]. Também é objetivo do SIEPI conhecer a situação de saúde da população beneficiária de planos privados, fornecendo subsídios para que a ANS, o Ministério da Saúde e demais atores do setor de saúde desenvolvam e fomentem políticas públicas voltadas às ações de proteção e promoção à saúde da população [MS 2007a].

Além desses objetivos gerais, o SIEPI possui os seguintes objetivos específicos [MS 2008]:

- Subsidiar a avaliação e regulação do sistema de saúde suplementar;

[65] http://www.caixa.gov.br/

[66] http://w3.datasus.gov.br/datasus/datasus.php

[67] http://www.datasus.gov.br/tabwin

- Fomentar práticas de saúde orientadas para as necessidades epidemiológicas da população beneficiária;
- Completar o perfil epidemiológico do complexo quadro sanitário brasileiro;
- Tornar comparáveis e integrar os subsistemas público e privado do SUS;
- Subsidiar o planejamento e programação das ações e serviços de saúde na perspectiva de melhorar a qualidade da atenção à população beneficiária dos planos e seguros de saúde;
- Disseminar informações epidemiológicas às operadoras, prestadores de serviços, gestores públicos e população em geral.

Para que esse conjunto de objetivos seja cumprido, o SIEPI capta as informações de saúde dos beneficiários e relaciona as diversas bases de dados com a utilização de técnicas de *Record Linkage*. As bases de dados nacionais utilizadas são:

- SIH (Sistema de Internações Hospitalares),
- SIM (Sistema de Informação sobre Mortalidade),
- Sinasc (Sistema de Informação sobre Nascidos Vivos),
- CIH (Comunicação de Internação Hospitalar) e
- SIB (Sistema de Informações de Beneficiários) da própria ANS.

A Figura 13 mostra o esquema geral de integração das bases de dados promovida pelo SIEPI.

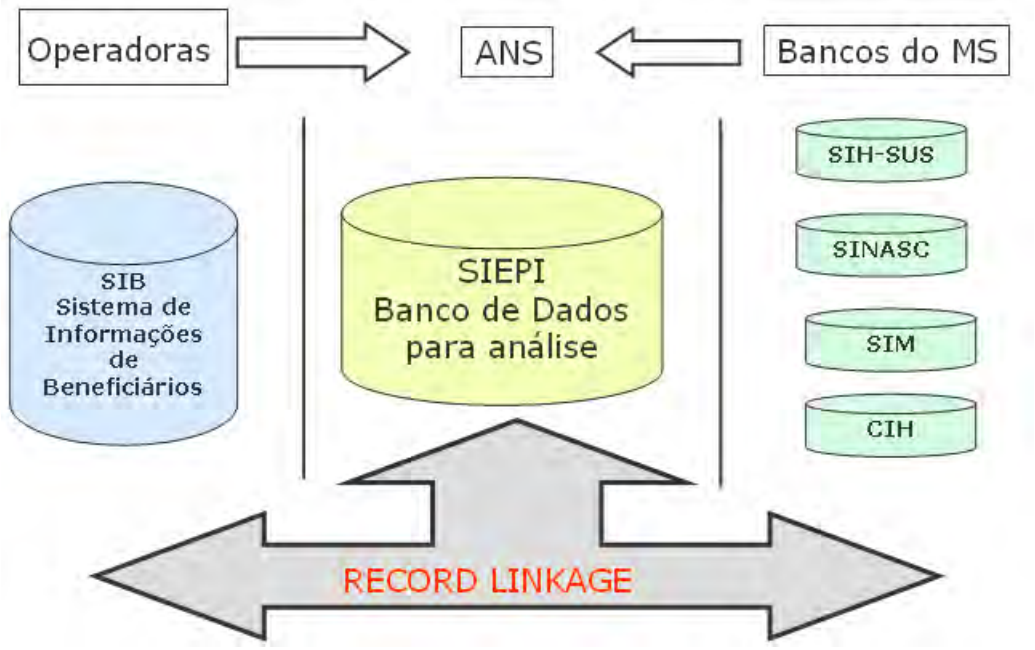

Figura 13: Esquema das bases de dados integradas pelo SIEPI [Rötzch 2006].

Atualmente, apenas as informações sobre ressarcimento (SIH e SIB) e sobre mortalidade (SIM) de beneficiários estão disponíveis no SIEPI, no sítio da ANS, no endereço http://www.ans.gov.br/portal/site/informacoesss/siepi/default.asp.

No Brasil, há ainda diversos registros epidemiológicos e administrativos do Sistema Único de Saúde que cobrem eventos distintos e necessitam de integração, a fim de evitar duplicidade de dados, coleta de dados desnecessários e sobrecarga dos profissionais de saúde. Além das bases de dados nacionais cobertas pelo SIEPI, destacam-se: o Sinan; as APAC-SIA (Autorizações para Procedimentos de Alto Custo/Complexidade para o Sistema de Informações Ambulatoriais do SUS) [Silva et al. 2006]; o CNES (Cadastro Nacional de Estabelecimentos de Saúde) e o SIAB (Sistema de Informação da Atenção Básica). Estes dois últimos especialmente importantes na análise de subnotificação de registros de nascidos vivos em sistemas de informação em saúde [Drumond, Machado e França 2008].

3.2. Aspectos Legais na Identificação de Pacientes

Consultar o histórico clínico do paciente, a partir dos prontuários médicos das diversas instituições de saúde que o atendeu, não é uma tarefa simples nem do ponto de vista organizacional nem do tecnológico. Há também, aspectos legais ou éticos que podem dificultar ainda mais esta tarefa.

Segundo a Consulta Nº. 450/97 sobre o Parecer do CFM Nº. 38/97, que versa sobre a "Legalidade de se Manter Arquivo Eletrônico no Consultório", "*o prontuário médico pertence ao paciente tendo este o direito de solicitá-lo e sendo o médico em seu consultório ou a instituição de assistência médica os fiéis guardiões de tão importante documentação*".

O Código Penal Brasileiro [BRASIL 1940] prevê em seus artigos 153, 154 e 325, respectivamente, crime contra divulgação de segredos, violação do segredo profissional e violação de sigilo funcional.

A Constituição da República Federativa do Brasil [BRASIL 1988], no seu artigo 5º, inciso X, afirma que *"são invioláveis a intimidade, a vida privada, a honra e a imagem das pessoas, assegurado o direito a indenização pelo dano material ou moral decorrente de sua violação"*.

O Código de Ética Médica (CEM) [CFM 1988], em seu artigo 102, diz que é vedado ao médico: *"Revelar fato de que tenha conhecimento em virtude do exercício de sua profissão, salvo por justa*

causa, dever legal ou autorização expressa do paciente”, proibindo de fazê-lo mesmo se for de conhecimento público ou se o paciente tiver falecido. Mesmo quando do depoimento como testemunha, a proibição persiste. Inclusive, todo o Capítulo IX do CEM se refere ao segredo médico.

Todo esse arcabouço jurídico pode ser interpretado como obstáculo para o compartilhamento das informações médicas de um paciente, mesmo que para outro médico [Soares, Barbosa e Costa 2008]. Essa limitação ético-jurídica pode ser vista em dois exemplos:

- Proibição da colocação do diagnóstico codificado (CID) ou do tempo de doença no preenchimento das guias da TISS (Resolução do CFM nº 1.819/2007 de 17 de maio de 2007).
- Artigo 205 da Lei nº 8.112, de 11 de dezembro de 1990, que institui o Regime Jurídico dos Servidores Públicos Civis da União, das autarquias, inclusive daquelas em regime especial, e das fundações públicas federais. Esse artigo fala, textualmente: *“O atestado e o laudo da junta médica não se referirão ao nome ou natureza da doença, salvo quando se tratar de lesões produzidas por acidente em serviço, doença profissional ou qualquer das doenças especificadas no art. 186, § 1o.”*.

Recentemente, várias resoluções do CFM vêm endossando a implantação tanto do PEP quanto do RES nacional. Contudo, é necessário que também haja um esforço jurídico para viabilizar a integração das bases de dados dos usuários de saúde, desde que sejam cumpridos os requisitos essenciais de integridade, autenticidade, disponibilidade e privacidade da informação. Essa iniciativa deve começar pela saúde pública, uma vez que há um grande esforço para construção de um cadastro dos usuários do SUS, conforme já descrito neste trabalho. Afinal, toda a população brasileira é usuária da Rede Pública de Saúde, não apenas a população de baixa renda. Basta observar as campanhas de vacinação.

Outros aspectos legais devem ser observados no domínio da Informática em Saúde, como é o caso da segurança da informação a ser armazenada e trafegada, o tempo de guarda dos documentos eletrônicos e a validade do prontuário eletrônico como prova legal [Francisco Jr. et al. 2008].

3.3. Padrões na Área de Informática em Saúde

Segundo Massad et al. (2003), atualmente, a informática mudou substancialmente os mecanismos de armazenamento, processamento e transmissão de dados. Entretanto, independente da mídia utilizada, para que a representação e troca de informações possam ocorrer de forma compreensível entre dois ou mais usuários de um sistema de comunicação, são necessárias duas condições básicas: a) definição de um vocabulário comum para representação e registro de conceitos; b) que a comunicação ocorra segundo um conjunto de regras compartilhadas pelos usuários.

Essa visão exposta por Massad et al. (2003) deixa clara a necessidade de adoção de padrões para a comunicação entre sistemas. Sobretudo, entre sistemas de saúde, uma vez que, nesses casos, há vocabulário e conceitos bem característicos e há diversas regras protocoladas que devem ser seguidas em cada procedimento médico. Portanto, o uso de padrões na área de informática em saúde contribui para a integração de dados médicos. Especificamente, para a identificação única de pacientes, o uso de padrões é fundamental, já que, cada vez mais, sistemas que vêm sendo desenvolvidos ou comercializados utilizam padrões para armazenamento e troca de informações em saúde.

O problema é a diversidade de padrões existentes: CCR (*Continuity of Care Records*), CDA (*Clinical Document Architecture*), CEN EN 13606 EHRcom, DICOM, HL7, HOI (*Health Outcomes Institute*), ICD (*International Classification of Diseases*), IUPAC (*International Union of Pure and Applied Chemistry*), LOINC (*Logical Observation Identifiers Names and Codes*), NDC (*National Drug Codes*), OpenEHR, SNOMED (*Systemized Nomenclature of Medicine*), TISS (Troca de Informações em Saúde Suplementar), UMDNS (*Universal Medical Device Nomenclature System*), UMLS (*Unified Medical Language System*). Procurou-se, neste trabalho, destacar aqueles padrões mais consolidados na área de informática em saúde e que possam contribuir de forma significativa para a identificação única de pacientes.

3.3.1. HL7 – *Health Level Seven*

O *Health Level Seven*, HL7 [HL7], é uma organização sem fins lucrativos, fundada nos Estados Unidos, em 1987, que desenvolve padrões internacionais de saúde. Trata-se de uma SDO (*Standards Developing Organization*), ou seja, uma entidade que produz padrões, acreditada pela

ANSI (*American National Standards Institute*)[68]. A missão do HL7 é "*prover sistemas e padrões relacionados para a troca, integração, compartilhamento e recuperação de informação eletrônica na saúde, para apoio da prática médica e administrativa, permitindo um maior controle dos serviços de saúde*". Vários países, incluindo o Brasil[69], são filiados à organização HL7.

O termo HL7 também se refere a um conjunto de padrões específicos para dados clínicos e administrativos, que possibilita, além de outras funcionalidades, checagem de segurança, identificação de usuários, checagem de disponibilidades, mecanismos de negociação de trocas e intercâmbio de informações. O objetivo principal desses padrões é tornar mais simples a implementação de interfaces entre aplicações de diferentes empresas. Nos Estados Unidos, o HL7 está incorporado em mais de 93% dos sistemas de informações hospitalares [Chaudry et al. 2006].

3.3.2. DICOM – *Digital Imaging Communications in Medicine*

DICOM (*Digital Imaging Communications in Medicine* - Comunicação de Imagens Digitais em Medicina) é um conjunto de normas para tratamento, armazenamento e transmissão de imagens médicas num formato eletrônico [DICOM]. O DICOM define um protocolo que permite que imagens médicas geradas por diferentes dispositivos possam se integrar aos sistemas de informação da área de saúde e, em especial, aos PACS. O padrão DICOM não é apenas um formato de arquivo. Trata-se de um conjunto de regras que permite que imagens médicas e informações associadas a elas sejam trocadas entre equipamentos de diagnóstico geradores de imagens, computadores e hospitais.

O modelo DICOM está estruturado em quatro níveis:

1. Pacientes;
2. Estudos;
3. Séries e equipamentos;
4. Imagens, ondas e relatórios estruturados.

[68] http://www.ansi.org/

[69] http://www.hl7brasil.org.br/

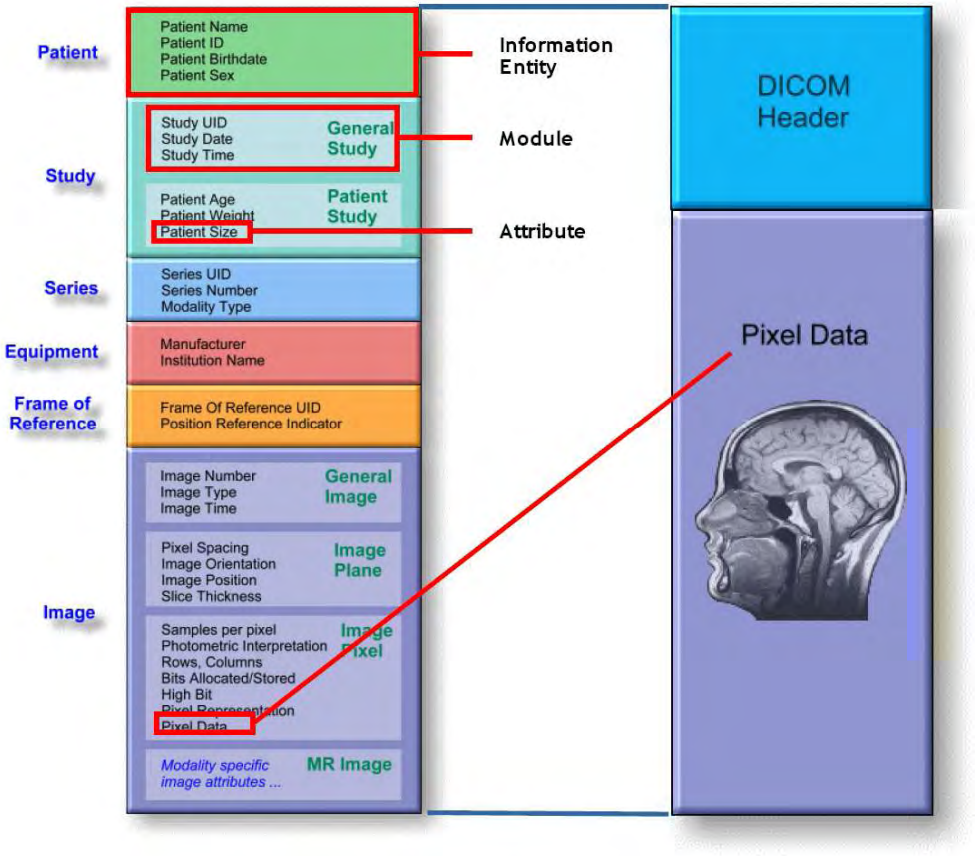

Figura 14: Estrutura do DICOM [Martins 2008].

Para cada paciente, podem existir vários estudos, que são modalidades de exames suportados pelo padrão DICOM, como tomografia computadorizada (CT – *Computed Tomography*), eletrocardiogramas (ECG - *Electrocardiogram*) e ressonância magnética (MR - *Magnetic Resonance*). Cada um desses estudos pode conter informações sobre dispositivos ou uma série de exames. O formato de um arquivo DICOM pode ser visto na Figura 14.

3.3.3. Padrão TISS – Troca de Informações em Saúde Suplementar

A ANS estabeleceu padronização obrigatória para troca de informações em saúde entre operadoras de planos de saúde e prestadores de serviços. Essa padronização, denominada padrão TISS, tem por objetivo atingir a compatibilidade e a interoperabilidade funcional e semântica entre os diversos sistemas independentes, com a finalidade de avaliar a assistência à saúde (caráter clínico, epidemiológico ou administrativo) e seus resultados, orientando o planejamento do setor [MS 2007]. Outro objetivo do padrão TISS é facilitar o ressarcimento ao SUS, previsto no artigo 32 da Lei nº 9.656/98, que estipula que as operadoras dos planos de saúde cujos beneficiários foram atendidos em hospitais da rede pública de saúde e nos hospitais contratados ou conveniados com o

SUS, devem ressarcir essas despesas aos cofres públicos. O padrão TISS foi instituído pela Resolução Normativa ANS (RN) nº 153/2007:

> *"Art. 1º - A presente Resolução estabelece padrão obrigatório para troca de informações em saúde suplementar (TISS) entre operadoras de plano privado de assistência à saúde e prestadores de serviços de saúde sobre os eventos de saúde realizados em beneficiários de plano privado de assistência à saúde, e mecanismos de proteção à informação em saúde suplementar"*.

As principais categorias de padrões na área de Informática em Saúde são padrão de comunicação, de vocabulário, de conteúdo e estrutura e de privacidade, confidencialidade e segurança [MS 2007]. O padrão TISS trata cada uma dessas quatro categorias da seguinte forma:

1. **Padrão de comunicação:** é adotada a linguagem de marcação XML/Schema;
2. **Padrão de vocabulário:** adota-se, atualmente, o padrão o CID 10 para descrição dos diagnósticos do paciente;
3. **Padrão de conteúdo e estrutura:** são os padrões definidos nas guias e demonstrativos;
4. **Padrão de privacidade, confidencialidade e segurança:** foram adotadas as normas editadas pelo Conselho Federal de Medicina.

O padrão TISS é composto por guias e demonstrativos de pagamento, mensagens eletrônicas e estrutura SCHEMA-XML. As guias definidas no padrão TISS são: Guia de Consulta, Guia de Serviços Profissionais / Serviço Auxiliar Diagnóstico e Terapia, Guia de Solicitação de Internação, Guia de Resumo de Internação, Guia de Honorário Individual Legenda, Guia de Outras Despesas, Guia de Tratamento Odontológico – Solicitação, Guia de Tratamento Odontológico – Cobrança, Demonstrativos de retorno, Demonstrativo de Pagamento, Demonstrativo de Análise de Conta Médica e Guia de Tratamento Odontológico – Demonstrativo de Pagamento.

Além de uma série de normas e padrões, com a adoção do padrão TISS, a ANS disponibiliza softwares que têm como objetivo fomentar a implantação desse novo padrão de troca de informações em saúde suplementar. Duas dessas ferramentas encontram-se desenvolvidas e disponíveis para download:

- **AplicaTISS:** Software desenvolvido na linguagem de programação Delphi[70], com banco de dados Interbase[71], para a plataforma Windows 2000. Ele possui funcionalidades básicas para a realização da troca de informações e administração do negócio em saúde suplementar, tais como: cadastro de beneficiários/pacientes, contratos prestados/operadora, contratos coletivo/individual, cadastro de eventos assistenciais, controle de rede de prestadores, controle de autorizações, controle de reembolso, ficha financeira, valoração de guias, relatórios e indicadores. Futuramente, as entidades poderão realizar alterações no aplicativo, com a finalidade de adequá-lo às suas necessidades de negócio. Contudo, as possíveis alterações no código-fonte deverão ser comunicadas à Agência Nacional de Saúde Suplementar, conforme descrito na licença de uso.
- **TissNET:** Software livre desenvolvido na linguagem de programação Java que tem como objetivo o gerenciamento de filas de mensagens eletrônicas trocadas entre as operadoras de planos de saúde e os prestadores de serviços. Esse software utiliza canal seguro através de uma porta TCP dedicada, que pode ser escolhida e configurada para a certificação digital das entidades.

Há ainda uma série de soluções comerciais que dão suporte à utilização do padrão TISS, como os softwares: Central TISS da Centralx[72] e Facilitiss[73], e o serviço *on-line* TISSXML da FIT Inovação e Tecnologia[74].

3.3.4. IHE – *Integrating the Healthcare Enterprise*

O padrão IHE é uma iniciativa de profissionais de saúde e da indústria para estimular a integração de sistemas de informações que suportam instituições de saúde [Stolba e Schanner 2007]. O IHE promove o uso coordenado de padrões estabelecidos, como o DICOM e o HL7.

O IHE é fortemente suportado pela indústria, sobretudo na Europa e nos Estados Unidos. Mais de 300 organizações já desenvolveram soluções adequadas ao padrão IHE [IHE 2009a] e a instituição internacional que define o padrão IHE é composta por 253 membros [IHE 2009b]. O IHE está organizado em nove domínios clínicos e operacionais. Cada domínio é subdividido em

[70] http://www.codegear.com/products/delphi/win32

[71] http://www.codegear.com/products/interbase

[72] http://www.tiss.med.br/

[73] https://www.facilitiss.com.br/

[74] http://www.tissxml.com.br/

perfis de integração, que são casos de uso que descrevem cenários selecionados do mundo real. Os sistemas de informação envolvidos são atores ("*Actor*") e as interações entre os atores são definidas através de transações ("*Transactions*") [Dogac, Bicer e Okcan 2006].

Os domínios do IHE são: Patologia Anatômica (*Anatomic Pathology*), Cuidados com a Visão (*Eye Care*), Infraestrutura de TI (*IT Infrastructure*), Laboratório (*Laboratory*), Coordenação do Atendimento ao Paciente (*Patient Care Coordination*), Dispositivos de Cuidado ao Paciente (*Patient Care Devices*), Qualidade, Pesquisa e Saúde Pública (*Quality, Research and Public Health*), Radiação Oncológica (*Radiation Oncology*) e Radiologia (*Radiology*).

Como mencionado no Capítulo 2, o domínio *IHE IT Infrastructure* (ITI) trata da implementação de soluções baseadas em padrões de interoperabilidade, que buscam melhorar o compartilhamento de informações, o fluxo de trabalho e a assistência ao paciente. Este domínio é composto pelos seguintes perfis de integração [IHE 2009]:

- *Consistent Time* (CT): assegura que os *clocks* e *time stamps* dos computadores em uma rede estejam bem sincronizados (erro médio menor que 1 segundo).
- *Audit Trail and Node Authentication* (ATNA): autentica sistemas usando certificação digital e envia eventos PHI (*Protected Healthcare Information*) auditados para um repositório, com o objetivo de facilitar a implantação de políticas de confidencialidade.
- *Retrieve Information for Display* (RID): fornece um acesso somente leitura (*read-only*), através de browser, a informações clínicas do paciente, tais como alergias ou resultados laboratoriais, localizadas externamente à aplicação do usuário.
- *Enterprise User Authentication* (EUA): estabelece um único acesso do usuário (*single sign-on*) que é utilizado em todos os dispositivos e softwares.
- *Patient Identifier Cross Referencing* (PIX): promove a referência cruzada entre a identificação do paciente para múltimplos domínios.
- *Patient Synchronized Application* (PSA): permite a seleção de um paciente em um aplicativo e, a partir dessa seleção, visualizar seus dados em diversas aplicações independentes. Evita que o usuário busque informações de um mesmo paciente, diversas vezes, em diversas aplicações.
- *Patient Demographics Query* (PDQ): permite consultas a um servidor central com informações sobre pacientes e a recuperação de dados demográficos e informações sobre visitas de um determinado paciente.

- *Patient Administration Management* (PAM): permite saber a situação atual do paciente. A localização e quadro clínico do paciente são obtidos através de mensagens ADT (*Admission, Discharge, Transfer*).
- *Cross Enterprise Document Sharing* (XDS): registra e compartilha documentos relativos a registros eletrônicos de saúde entre instituições de saúde. Estas instituições abrangem desde consultórios médicos até grandes hospitais.
- *Personnel White Pages* (PWP): Fornece informações sobre os trabalhadores da área de saúde.
- *Cross-Enterprise Document Media Interchange* (XDM): transfere documentos XDS e metadados, usando CDs, memórias USB ou anexos de e-mails.
- *Cross-Enterprise Document Reliable Interchange* (XDR): permite troca de documentos de saúde entre instituições, usando uma rede ponto-a-ponto.
- *Cross-Enterprise Sharing of Scanned Documents* (XDS-SD): define como são integradas as informações clínicas obtidas através de sistemas legados: papéis, filmes, documentos digitalizados, etc.
- *Retrieve Form for Data Capture* (RFD): extrai dados de uma aplicação, com a finalidade de satisfazer os requisitos de um sistema externo.
- *Cross-Comunity Access* (XCA): permite a consulta e recuperação de dados de saúde pertencentes a outras comunidades. Cada comunidade pode ser um domínio de afinidade XDS (*XDS Affinity Domains*), que define quais documentos podem ser usados por outros perfis XDS ou por outras comunidades. A estrutura interna de cada comunidade é transparente.

3.4. Perfil de Integração IHE/PIX

O perfil de integração *Patient Identifier Cross-referencing* (PIX) é destinado a organizações de saúde de diferentes tamanhos (hospitais, clínicas, consultórios médicos, etc.). Esse perfil suporta a referência cruzada de identificadores de pacientes de múltiplos domínios independentes, através das seguintes interações:

- A transmissão da informação da identidade do paciente para a formação do índice único.
- A capacidade de acessar as fontes de dados originais, através da lista de identificadores referenciados, seja para fazer uma consulta aos dados do paciente, seja para notificar atualizações ocorridas nas outras fontes.

Os atores envolvidos no PIX, bem como as transações entre eles são descritos na Figura 15.

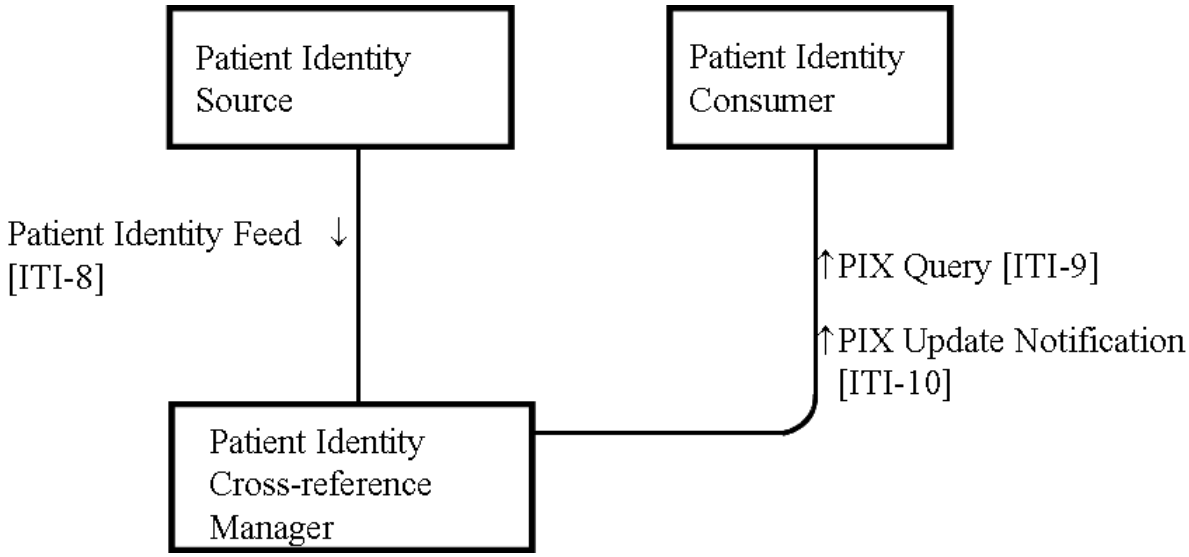

Figura 15: Diagrama com atores e transações do *Patient Identifier Cross-referencing* (PIX) [ACC, HIMSS e RSNA 2008].

Os *Patient Identity Source* são as fontes de dados relativas aos pacientes pertencentes a cada uma das instituições de saúde participantes (hospitais, clínicas, laboratórios, etc.). Este ator é responsável por determinar identidades de pacientes contidas em seus respectivos domínios. Ele notifica ao *Patient Identity Cross-reference Manager* (*PIX Manager*) todos os eventos relacionados à identificação do paciente (criação de campos, fusão de atributos, etc.). Sua forma de comunicação com o PIX Manager é através de envios de cargas de identificação de pacientes (*Patient Identity Feed*).

O *Patient Identifier Consumer* solicita informações sobre identificadores de pacientes em outros domínios. Este ator requisita e recebe dados de identificação de pacientes através do *PIX Manager*, utilizando para isso consultas PIX (*PIX Query*). Opcionalmente, o *Patient Identifier Consumer* recebe notificação de atualização de dados (*PIX Update Notification*).

O *Patient Identity Cross-reference Manager* (*PIX Manager*) recebe informações de identificação de pacientes dos atores *Patient Identity Source*, gera o MPI e gerencia a referência cruzada de identificadores entre os domínios.

Como qualquer perfil de integração do padrão IHE, aqui não há definições de políticas ou de algoritmos que devem ser utilizados. O PIX define a interoperabilidade necessária para troca de informações entre os atores. Fica a cargo do desenvolvedor escolher as políticas e algoritmos mais

adequados às organizações que farão parte do domínio PIX (*Patient Identifier Cross-reference Domain*).

A Figura 16 ilustra o domínio PIX, detalhando o fluxo do processo que envolve um *PIX Manager* e vários *Patient Identifier Domains*, representados pelas letras A, B e C.

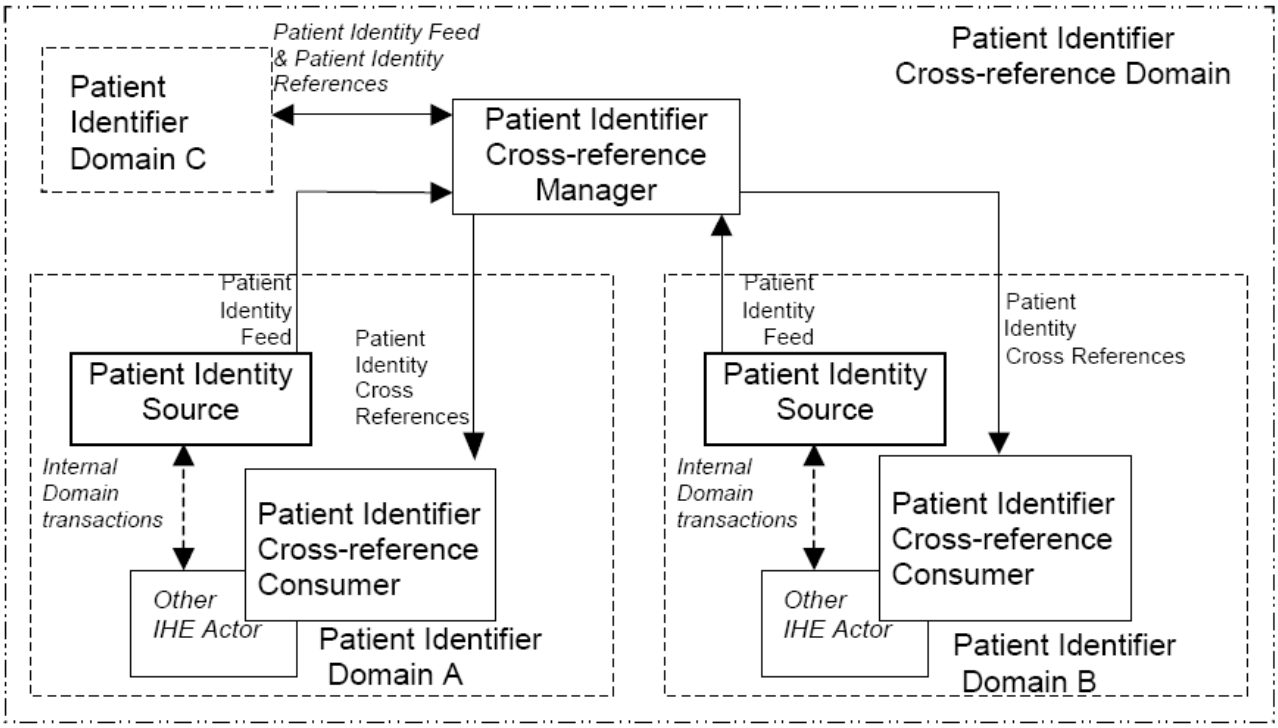

Figura 16: Fluxo do processo do *Patient Identifier Cross-referencing* (PIX) [ACC, HIMSS e RSNA 2008].

Um *Patient Identifier Domain* é definido como um sistema único ou um conjunto de sistemas interconectados que compartilham o mesmo esquema de identificação (mesmo identificador e mesmo processo de identificação de um paciente). Além disso, todos os *Patient Identifier Domain* que integram o domínio PIX devem possuir autoridade para publicar identificadores de pacientes e serem reconhecidos como autores desses dados. Em resumo, cada *Patient Identifier Domain* é uma fonte de dados que irá fornecer ao *PIX Manager* dados para identificação única de seus pacientes.

O *PIX Identifier Cross-reference Manager* (*PIX Manager*), como já mencionado, exerce, nesse cenário, o papel de criar e manter o MPI e de fornecer referência cruzada entre as diversas fontes de dados. O *PIX Manager* não é responsável por melhorar a qualidade dos dados fornecidos pelas fontes (*Patient Identifier Domain*). A qualidade da informação de identificação do paciente e de seus dados demográficos é de responsabilidade de cada fonte de dados.

O perfil de integração PIX interage com vários outros perfis. Ele interage com o CT, na medida em que cada *Patient Identifier Domain* precisa estar sincronizada com o *PIX Manager*. Deve haver uma interação com o perfil PSA, uma vez que é necessária uma consulta ao MPI para ser possível visualizar dados de pacientes em diversas aplicações independentes. A integração com o XDS se dá pela necessidade de troca de documentos entre os diversos *Patient Identifier Domain*. Como se trata de documentos relativos a pacientes, é necessário o uso do PIX e, sobretudo do MPI, para identificar de forma unívoca esses pacientes. De forma análoga, também há integração entre PIX e XDM. Para que a consulta a dados demográficos dos pacientes seja possível, é necessária uma integração entre PIX e PDQ. Por fim, para que seja possível localizar e recuperar dados clínicos de pacientes em diversas bases de dados, com um único acesso do usuário, é necessário combinar o uso dos perfis RID, EUA e PIX.

3.5. Alternativas para Integração de Dados

Em muitos casos, não basta integrar os dados de dois sistemas. É necessária a integração dos modelos de dados ou dos esquemas de sistemas heterogêneos. Nos sistemas de saúde, essa situação ocorre no caso da semântica dos dados fazer diferença, ou seja, quando se pretende buscar informações de um paciente associadas a um contexto. Por exemplo, quando for preciso buscar, a partir de um MPI, em fontes de dados distintas, somente exames médicos associados a uma doença de um determinado paciente. Nesse caso, o formato e a semântica do dado devem ser considerados.

Há questões clássicas de integração de dados heterogêneos, como a heterogeneidade semântica, que ocorre quando, em diferentes fontes de dados, há discordância sobre o significado, interpretação ou uso pretendido do mesmo dado ou de dados relacionados [Shet e Larson 1990].

A heterogeneidade semântica entre modelos de dados distintos pode gerar alguns tipos de conflitos [Özsu e Valduriez 2001] e [Pitoura, Bukhres e Elmagarmid 1995]:

1. **Conflito de identidade:** ocorre quando o mesmo conceito é representado por diferentes objetos em diferentes bancos de dados. Por exemplo, quando dois pacientes têm diferentes identificadores em diferentes hospitais. Este conflito é resolvido com a geração do MPI.

2. **Conflito de esquema:** ocorre quando os esquemas que representam o mesmo conceito não são idênticos. Os conflitos de esquema são subdivididos em:
   a. **Conflito de nome:** há homônimos e sinônimos. Os homônimos ocorrem quando o mesmo nome é usado por diferentes conceitos. Por exemplo, o termo “usuário” em um

hospital pode significar pacientes, enquanto que em outro pode representar os usuários de um sistema. Já os sinônimos ocorrem quando o mesmo conceito é descrito por diferentes nomes. Por exemplo, um hospital usa o termo admissão para inclusão do paciente no sistema, enquanto outro usa o termo entrada para o mesmo procedimento.

b. **Conflito estrutural:** ocorre quando o mesmo conceito é representado por diferentes construtores do modelo de dados ou, os construtores usados têm diferentes estruturas ou diferentes comportamentos. Por exemplo, em um hospital, o número de vezes que um paciente foi internado é dado por um atributo do paciente, enquanto que em outro hospital ele deve ser calculado por um método, quando necessário.

3. **Conflito semântico:** ocorre quando o mesmo conceito é interpretado de forma diferente em bancos de dados distintos. Por exemplo, em um HIS, o termo “reserva” é usado para reservar uma sala no centro cirúrgico, enquanto que em outro HIS significa uma área restrita a um tipo específico de profissional.

4. **Conflito de dados:** ocorre quando os valores dos dados de mesmo conceito são diferentes em diferentes bancos de dados. Por exemplo, o mesmo paciente aparece tendo diferentes dados biométricos em diferentes sistemas de saúde.

As alternativas para integrar dados, resolvendo os conflitos expostos, foram divididas em cinco abordagens, descritas a seguir. Ao final, é feita uma síntese sobre integração de dados e esquemas.

3.5.1. Integração Manual

Nesta forma de integração, os usuários interagem com todos os sistemas de informação e, manualmente, integram os dados selecionados [Ziegler e Dittrich 2004]. Portanto, os usuários têm que trabalhar com diferentes interfaces e linguagens de consulta. Além de ter que conhecer os detalhes sobre a localização dos dados, sua representação lógica e sua semântica.

Uma forma de amenizar o esforço causado por essa abordagem é a utilização de uma interface padrão, como *web browser*, por exemplo.

Também podem ser construídas aplicações de integração que acessam várias fontes de dados e retornam resultados integrados para o usuário. Entretanto, com o crescimento dos sistemas de informação que compõem o ambiente ou com a incorporação de um novo sistema ao cenário de integração, é necessário novo esforço de desenvolvimento.

3.5.2. Centralizado

No mundo corporativo, o mais comum é a utilização de grandes repositórios globais. Esses repositórios, que concentram todos os dados da organização é um exemplo típico da abordagem centralizada. É também comum a catalogoção de serviços disponíveis, o que facilita a consulta e o gerenciamento dos dados [Melo et al. 2005].

Essa forma de integração não é muito prática, principalmente, quando envolve várias organizações. Uma alternativa é gerar um *data warehouse*, com todos os dados das bases locais e com uma atualização periódica. Mesmo assim, há muita dificuldade em manter uma base de dados centralizada atualizada quando há o envolvimento de várias organizações. Há, também, a perda da autonomia dos sistemas locais que participam da comunidade.

3.5.3. Distribuído com Replicação de Metadados

Essa abordagem, proposta por Melo et al. (2005), permite manter os dados localmente em cada site, mesmo os semi-estruturados e os não estruturados, como arquivos .doc, .ppt, .pdf, etc. Nessa forma de integração, todos os metadados compartilhados no ambiente são replicados em cada site, permitindo uma visão global dos dados. Os metadados funcionam como ponteiros, que permitem consultas locais.

A principal vantagem, em relação à abordagem centralizada, é o fato dos metadados exigirem muito menos espaço de armazenamento do que o conteúdo em si. Permitindo, inclusive, menor tráfego na rede. Como principal desvantagem, há a necessidade de um mecanismo de controle mais complexo para a replicação dos metadados.

3.5.4. Ponto a Ponto

Nessa abordagem, cada membro da comunidade funciona como fornecedor e requerente de recursos para a comunidade como um todo. Recursos podem ser entendidos como conteúdos ou serviços. Conteúdos podem ser os dados propriamente ou esquemas de metadados. Pode haver serviços como busca, replicação e mapeamento [Melo et al. 2005].

Podem-se integrar dados utilizando uma abordagem ponto-a-ponto, através da criação de um Sistemas de Gerenciamento de Dados Ponto-a-Ponto (PDMS - *Peer-to-Peer Data Management Systems*). Essa técnica permite que cada ponto tenha funcionalidades de SGBDs e autonomia local,

podendo, inclusive, entrar na comunidade e sair dela a qualquer momento. O compartilhamento dos dados é descentralizado, ou seja, cada ponto se comunica com os outros para executar consultas e transações, embora, as implementações clássicas tratem somente consultas. O processamento e o armazenamento dos dados são distribuídos em cada um dos pontos. Paradoxalmente, há PDMSs que utilizam *peers* de maior poder computacional, responsáveis pelo compartilhamento e gerenciamento de recursos, chamados *super-peers*. Considerando as características dos PDMSs, como volatilidade, escalabilidade, autonomia no gerenciamento de dados, utilização de esquema local e compartilhamento de dados entre os nós, [Sung et al. 2005], propôs uma arquitetura de referência para cada um dos pontos, como pode ser observado na Figura 17.

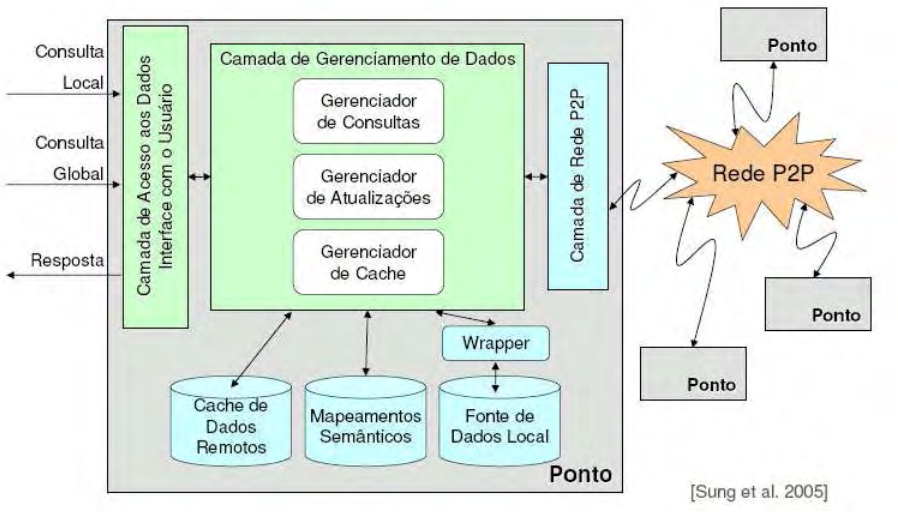

Figura 17: Arquitetura genérica de um ponto em um PDMS, proposta por [Sung et al. 2005].

A principal vantagem dessa forma de integração é a autonomia local associada ao acesso global. Já a sua principal desvantagem é a possibilidade de serem obtidas respostas incompletas para uma consulta. Ou seja, o conjunto resultante de uma consulta pode não ser formado pelo conteúdo de informação da totalidade dos nós da rede. Pode haver nós que nem estavam conectados no momento da consulta.

3.5.5. *Middleware* de Integração

Usando essa alternativa, o usuário não envia consultas diretamente para as fontes de dados. Não é necessário que o usuário conheça detalhes sobre as fontes de dados, nem que interaja com cada uma delas individualmente [Barbosa 2001]. Para isso, é necessário o uso de padrões de interoperabilidade e *wrappers*.

Os padrões de interoperabilidade têm como principal objetivo a integração de aplicações distribuídas através da tecnologia de objetos. De forma transparente, objetos fazem requisições e recebem respostas de outros objetos. Cada padrão de interoperabilidade define um subconjunto de mecanismos para viabilizar a construção e a comunicação dos componentes dos objetos. Entretanto, os objetos se comunicam sem conhecer como os componentes dos outros objetos foram implementados.

Wrappers são componentes de software que traduzem uma consulta para a linguagem de consulta específica de cada fonte de dados [Melo et al. 2005]. As fontes de dados executam essas consultas e enviam as respostas de volta para os *wrappers*. Portanto, os wrappers convertem dados vindos de fontes heterogêneas e distribuídas para um esquema de dados comum.

Historicamente, existem duas abordagens para sistemas de integração de dados heterogêneos, utilizando *middleware* de integração:

a. Mediadores: que são um tipo de *middleware* que ligam fontes de dados heterogêneas através de programas de aplicação [Wiederhold e Genesereth 1995].

b. SGBDH: que é uma camada de software que possui o objetivo de viabilizar a comunicação entre sistemas de bancos de dados autônomos, heterogêneos e distribuídos, usando recursos de SGBDs, sem a necessidade de alterar os sistemas de bancos de dados existentes. Como os esquemas dos diferentes bancos de dados componentes podem estar expressos em diferentes modelos, há a necessidade da criação de um Modelo de Dados Comum (MDC), descrito em um esquema global. O MDC é o modelo de dados usado pelo sistema integrador [Barbosa 2001].

Com o passar do tempo, houve uma convergência entre as soluções que utilizam *middleware* de integração. As soluções que utilizam mediadores cada vez mais implementam recursos de SGBDs, enquanto as soluções baseadas em SGBDH usam recursos de programação.

3.5.6. Síntese sobre Integração de Dados

Tradicionalmente, instituições possuem dados armazenados localmente que são manipulados por aplicações legadas. A integração de dados heterogêneos e distribuídos deve exigir dessas instituições o menor esforço possível.

Para que essa premissa seja atingida, podem-se utilizar alternativas como integração manual de dados e aplicações, centralização dos dados, replicação e distribuição dos metadados de cada modelo, uso de um modelo conceitual global, redes ponto-a-ponto e uma série de outras técnicas. Com o surgimento de novas necessidades de integração de dados, há também o surgimento de novas técnicas. Um exemplo disso é a consulta a textos livres em vários bancos de dados heterogêneos de forma eficiente. Para suprir essa necessidade, usa-se a geração do índice invertido (*Inverted Index*), que consiste na atualização de um único e centralizado índice composto por palavras (*index term*) que estão contidas em um ou mais documentos. O índice invertido deve estar associado a uma máquina de busca (*search engine*) que encontra as palavras pesquisadas [Zobel e Moffat 2006].

O fato é que integrar dados não é uma tarefa trivial. Há três décadas ou mais são propostas soluções de integração de banco de dados e ainda não há uma solução estilo "bala de prata" (*silver bullet*), que resolva todos os problemas [Ziegler e Dittrich 2004].

Especificamente, neste trabalho, a procura de dados de um mesmo paciente em diferentes fontes de dados heterogêneas e distribuídas será feita com a utilização de técnica de *Record Linkage*, descrita a seguir.

4. *Record Linkage*

A necessidade de implantação de um RES nacional ou o simples interesse de integrar bases de dados de duas instituições de saúde diferentes demanda o uso técnicas de integração de dados. Em geral, o objetivo dessa integração é identificar pacientes iguais em fontes de dados diferentes.

A tarefa de identificar indivíduos iguais em fontes de dados heterogêneas, seja no domínio da saúde, seja em qualquer outro domínio, é trivial nos casos em que os registros de cada fonte possuem um campo comum que permita a identificação de cada registro de forma unívoca, como, por exemplo, o CPF. No entanto, em bases de dados de saúde, é comum não encontrarmos um campo dessa natureza. Diante deste problema, a integração de bases de dados pode ser realizada através de dois métodos [Coeli et al. 2006]:

- **Método determinístico:** É necessária a existência de um identificador unívoco nas fontes de dados a serem relacionadas, como o CPF ou o número do Cartão Nacional de Saúde. Este identificador unívoco é formado por um campo ou por um conjunto de campos obrigatório, não duplicado e usado para identificar cada uma das tuplas de uma tabela. Este método emprega a comparação exata entre campos e é, comumente, de simples entendimento e implementação. Contudo, em alguns casos, este método envolve decisões subjetivas, que pode tornar a tarefa laboriosa e consumir muito tempo [Romero 2008].

- **Método probabilístico:** Baseia-se na utilização conjunta de campos presentes nas duas fontes de dados, que possuem o objetivo de identificar o quanto é provável que um par de registros se refira a um mesmo indivíduo [Coeli et al. 2006].

Como descrito no Capítulo 2, o método probabilístico desenvolvido por Fellegi e Sunter (1969) é referenciado, neste trabalho, como *Record Linkage*, cujo esquema geral foi formalizado por Christen (2008) e é apresentado na Figura 18. Neste esquema, podem ser observadas as etapas desse método e seus relacionamentos; as entradas, representadas pelas fontes de dados; e saídas: pares combinados, não combinados e duvidosos, classificados após atribuição de pesos ou escores. Cada etapa é descrita a seguir.

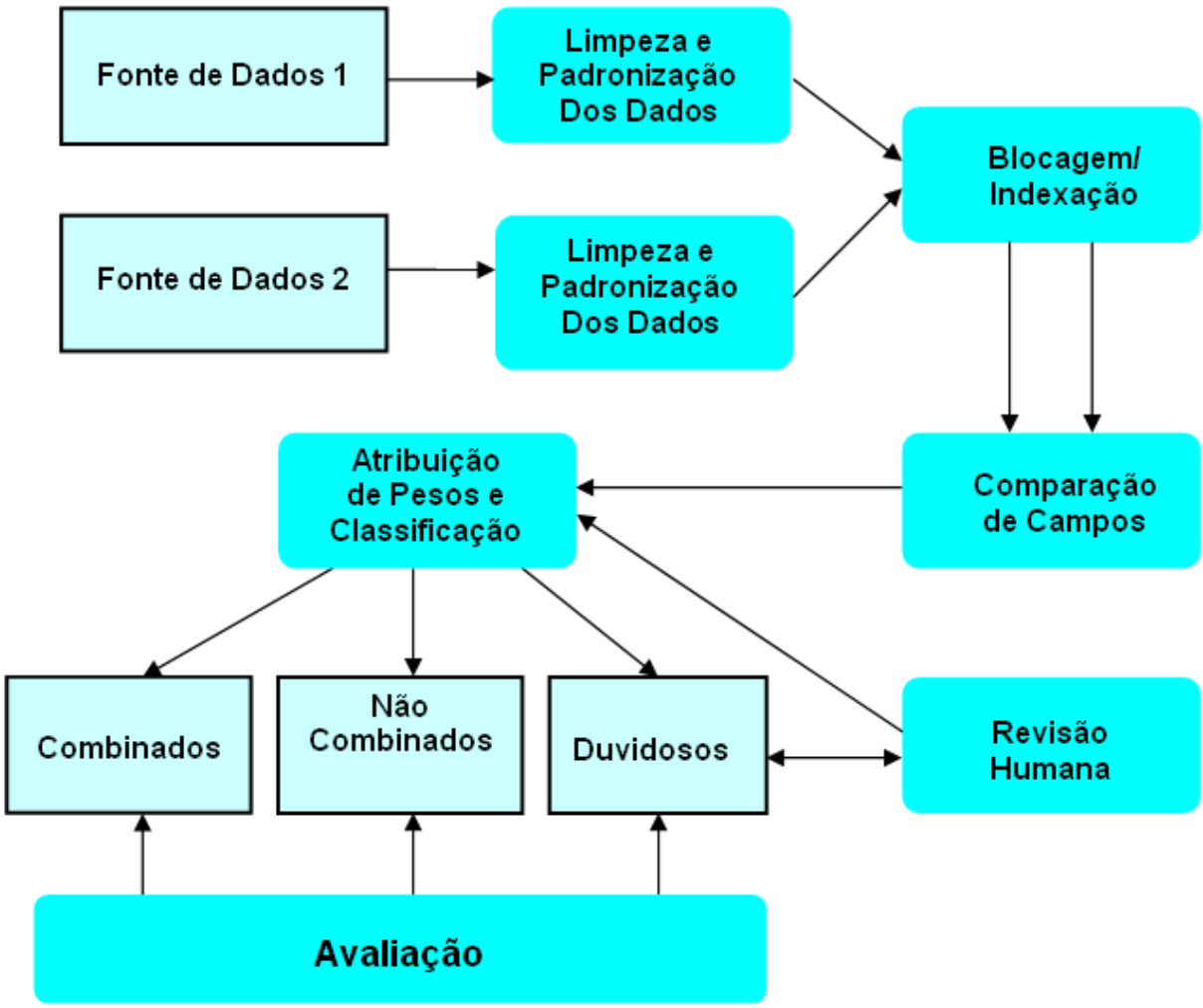

Figura 18: Esquema Geral do Método de *Record Linkage*. Adaptado de Christen (2008).

4.1. Escolha e Obtenção das Fontes de Dados

Antes de ser iniciado o *Record Linkage* propriamente dito, é necessário decidir quais fontes de dados integrar. A obtenção das fontes de dados, em geral, faz parte de um projeto maior. Como toda fase de um projeto, há necessidade de estimativa de tempo e custo para obtenção desses dados. O nível de formalização e privacidade para conseguir os dados também deve ser levado em consideração. Obtenção de dados de saúde, por exemplo, exige um comprometimento extremo com privacidade e segurança. A Resolução 196/96 do Conselho Nacional de Saúde – Ministério da Saúde, que se fundamenta em declarações e diretrizes como o Código de Nuremberg (1947), a Declaração dos Direitos do Homem (1948) e a Declaração de Helsinque (1964 e suas versões posteriores de 1975, 1983 e 1989), trata o uso de dados de pacientes como **pesquisa envolvendo seres humanos** (II-2). Assim, é exigido que qualquer protocolo de pesquisa nesta área seja submetido ao Comitê de Ética em Pesquisa (CEP) da instituição fonte para poder ser executado.

Outro fator que deve ser levado em consideração é a maneira como é feito o acesso: diretamente no servidor de banco de dados do sistema que se pretende consultar ou se os dados estão replicados em um site ou em uma mídia externa. A etapa de obtenção dos dados para geração do índice único não deve incluir ou alterar nenhum dado em sua origem.

4.2. Limpeza e Padronização dos Dados

As fontes de dados que são utilizadas no processo de *Record Linkage* devem ser estruturadas, como tabelas em um banco de dados ou, no máximo, semi-estruturadas, como arquivos em formato XML. O relacionamento probabilístico de fontes de dados não estruturadas, como documentos gerados em um editor de texto ou páginas web não é foco deste trabalho.

Segundo [Rahm e Do 2000], [Christen e Churches 2005] e [Oliveira 2007], a maioria das fontes de dados que se pretende relacionar contém codificações e formatos diferentes entre si. Além disso, possuem dados “sujos”, incompletos ou ultrapassados. O objetivo da limpeza e padronização é converter dados brutos originários das fontes de dados em dados bem definidos, resolvendo as inconsistências na forma como os campos estão representados ou codificados. Em resumo, deve haver uma padronização dos campos que fazem parte da etapa de comparação.

Para viabilizar a integração das fontes de dados diferentes, é preciso que elas possuam definições compatíveis para os seus campos [Oliveira 2007]. Por exemplo, o campo data de nascimento em uma fonte de dados pode ter seu formato de armazenamento diferente do da outra fonte de dados. Contudo, os dois campos possuem o mesmo significado. É necessária que haja uma conversão de formato para que eles possam ser comparados.

Outra questão é o fato de uma fonte de dados possuir uma chave composta, ou seja, uma chave primária composta por mais de um atributo. Neste caso, não há necessidade de conversão. Contudo, há necessidade de armazenar todos os campos que fazem parte da chave no MPI.

Já que o intuito da padronização é minimizar a ocorrência de erros de compatibilidade durante a etapa de **Comparação de Campos**, é necessária a criação de uma série de funções que padronize e elimine incompatibilidades nos campos que serão comparados, fazendo com que ambos possuam o mesmo formato. Essas funções devem agir, sobretudo, em campos com formatos livres, como, por exemplo, naqueles que armazenam nomes de pessoas. A Tabela 8, apresentada na Seção 5.2, mostra

uma lista de funções de compatibilidade de campos, com seus respectivos parâmetros de entrada e descrição da funcionalidade.

É comum também nesta etapa de **Limpeza e Padronização dos Dados** a geração de arquivos ou estruturas dados intermediárias, com a finalidade de aumentar o desempenho do *Record Linkage* como um todo.

4.3. Blocagem

No momento da comparação dos campos de cada fonte de dados, é possível que seja feita uma comparação “um para um”. Neste tipo de comparação, todos os registros da primeira fonte de dados são comparados com os da segunda [Oliveira 2007]. Neste caso, observa-se um desempenho $O(mn)$ ou $O(n^2)$.

Para reduzir a grande quantidade de comparações de potenciais pares de registros, utiliza-se a técnica de indexação ou filtragem, conhecida por **Blocagem** (*blocking*). Nessa técnica, um determinado campo, chamado de chave de blocagem (*blocking key*) é usado para dividir as fontes de dados em blocos. Pode ser usada, também, a combinação de vários campos ou parte de um campo para formar a chave de blocagem [Christen 2008] e [Oliveira 2007]. As fontes de dados são logicamente divididas em blocos mutuamente exclusivos. Os registros de determinado bloco possuem o mesmo valor da chave de blocagem escolhida. Nas duas fontes de dados, só serão comparados registros que estiverem contidos no mesmo bloco.

Embora possa haver grande aumento de desempenho no *Record Linkage* com a adoção da **Blocagem**, essa técnica apresenta riscos. Por exemplo, se for usada como chave de blocagem a primeira letra do primeiro nome do paciente nas duas fontes de dados, o paciente “Elton”, escrito sem “H” e com “H” (Helton) estarão em blocos diferentes e seus registros não serão comparados, mesmo que o restante dos campos coincida integralmente.

Segundo Jaro (1989), a chave de blocagem escolhida deve permitir a divisão das fontes de dados no maior número de blocos possível e, ao mesmo tempo, ser sujeita à baixa probabilidade de erros de registro. Uma **Blocagem** a partir do campo “sexo”, por exemplo, dividiria cada fonte de dados em apenas dois blocos, trazendo pouco ganho de desempenho na etapa de **Comparação de Campos** [Camargo e Coeli 2000]. Já a utilização do último nome como chave de blocagem, permitiria uma divisão em muitos blocos, mas, o problema da classificação de registros do mesmo

indivíduo em blocos diferentes se agravaria. Por exemplo, a omissão do último nome do marido no nome da esposa pode provocar distorções na comparação deste campo.

Jaro (1989) recomenda ainda que seja realizada a **Blocagem** em vários passos, com o intuito de diminuir o erro da classificação incorreta de registros. Ou seja, emprega-se determinada chave de blocagem e procede-se a comparação dos registros. Os registros não combinados na primeira etapa seriam novamente blocados, com o emprego de nova chave de registro. O que deve ser avaliado é se há ganho de desempenho com sucessivos passos de **Blocagem** e **Comparação de Campos**.

Baxter et al. 2003 propõe e analisa vários métodos de blocagem em seu estudo chamado "*A Comparison of Fast Blocking Methods for Record Linkage*".

4.4. Comparação de Campos

Utilizando-se ou não a **Blocagem**, campos comuns às duas fontes de dados precisam ser escolhidos para serem comparados. Esses campos, em seu conjunto, devem garantir uma alta probabilidade de identificar de maneira única cada registro. Na grande maioria das vezes, são escolhidos campos com conteúdo alfanumérico.

Quando são comparados conteúdos exatos de campos alfanuméricos, como primeiro nome, último nome ou endereço, podem ser desprezadas várias distorções: erros de digitação, variação ortográfica de nomes próprios, variações fonéticas e de pronúncia, etc.

Outro tipo de comparação que deve ser observado é aquele que ocorre entre duas datas. Por exemplo: em um par avaliado, um registro da primeira fonte de dados apresenta data de nascimento igual a 03/04/1970, enquanto na segunda fonte, a data é igual 27/10/1964. Não há problema algum em dizer que essas datas são diferentes. Contudo, se a segunda fonte de dados apresentasse a data de nascimento igual a 30/04/1970, embora diferente da primeira fonte, existiria a dúvida de se houve ou não erro de digitação.

Segundo Winkler (2004), em uma base de dados que possui uma taxa de erro de 5% para cada uma das variáveis (primeiro nome, último nome, ano, mês e dia de nascimento), se fossem considerados apenas os pares que concordam caractere a caractere, haveria uma perda de mais de 20% das combinações.

Portanto, o algoritmo de **Comparação de Campos** deve levar em consideração essas possíveis distorções. Segundo Oliveira (2007), as comparações de *strings* podem ser feitas com a utilização de dois tipos de códigos:

- **Códigos fonéticos:** Identificam campos com a mesma pronúncia.
- **Códigos ortográficos:** Refletem variações na escrita. Esses algoritmos utilizam métricas de similaridade de *strings* para determinar se dois valores são semelhantes o bastante para serem iguais.

Neste trabalho, o algoritmo fonético utilizado para nomes próprios é o *Soundex*. Alguns outros algoritmos fonéticos específicos para a Língua Portuguesa e, sobretudo, para sua pronúncia brasileira, foram analisados. Contudo, como o domínio estudado é a área de saúde e o Brasil é um país de várias origens, não é razoável a limitação de pronúncias de nomes e sobrenomes a apenas um idioma. Portanto, foi escolhido o *Soundex*, por ser um algoritmo mais genérico, com aceitação em várias partes do mundo. O algoritmo ortográfico escolhido para comparação de datas é a Distância *Levensthein*. As datas são convertidas para o formato "AAAAMMDD" (4 caracteres para ano, 2 para mês e 2 para dia) na etapa de **Limpeza e Padronização dos Dados**.

Existe uma enorme quantidade de algoritmos que calcula a similaridade entre *strings*. Um excelente trabalho é o Tutorial *Record Linkage: Similarity Measures and Algorithms* [Koudas et al. 2006], que faz um estudo minucioso sobre este tema. Não é objetivo deste trabalho fazer um estudo comparativo entre os algoritmos de aproximação e similaridade entre *strings*.

4.4.1. *Soundex*

O *Soundex* é um índice para codificação de nomes, que se preocupa mais com o som do que com a forma de como um nome está escrito. Ele foi usado inicialmente para codificar sobrenomes (*surnames*) pela Administração de Arquivos e Registros Nacionais dos Estados Unidos (*National Archives and Records Administration*)[75]. Nomes que possuem o mesmo som, mas estão escritos de forma diferente, têm o mesmo código [Soundex].

O método foi criado e patenteado em 1918, por Margaret O'Dell e Robert C. Russel [Knuth 1973]. O departamento de censo dos Estados Unidos (*U.S. Bureau of Census*) codificou todos os

[75] http://www.archives.gov/

registros relativos aos censos de 1880 a 1920 [Branting 2003] e [Oliveira 2007], utilizando *Soundex*.

O algoritmo *Soundex* produz um código padrão, composto pela primeira letra da palavra a ser codificada, seguida por três dígitos numéricos. Os dígitos variam de 0 a 6. Os seis números significativos representam classes fonéticas dos sons da fala humana: bilabial, labiodental, dental, alveolar, velar e glotal [Knuth 1973], [Bhagat e Hovy 2007] e [Oliveira 2007]. O esquema de codificação do *Soundex*, com o relacionamento entre letras e números, está descrito na Tabela 3. A seguir, estão descritas as regras para geração do código *soundex*.

Segundo Schaback e Li (2007), o *Soundex* mapeia 87% das *strings* que possuem erros ortográficos, gerando o mesmo código fonético para essas *strings*. Há casos em que nomes com sons diferentes podem gerar o mesmo código e nomes semelhantes podem não produzir o mesmo *soundex*, que é o que ocorre quando nomes com sons idênticos começam com letras diferentes [Oliveira 2007].

Tabela 3: Codificação fonética do *Soundex*.

Valor a ser atribuído à Letra	Letras
0	A, E, I, O, U, Y, H, W
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Fonte: Zobel e Dart (1996).

Para a codificação do *soundex*, são aplicadas as seguintes regras, baseadas no algoritmo de Zobel e Dart (1996):

a. O código é composto da letra inicial da *string* mais três dígitos, atribuídos conforme a Tabela 3. As demais consoantes são ignoradas;

b. Se o código gerado for menor que quatro caracteres, zeros serão acrescidos.

c. As vogais A, E, I, O, U e as letras Y, W e H, bem como os demais caracteres que não são letras não serão considerados.

d. Repetições adjacentes são ignoradas e é tratada apenas a primeira letra. Essas repetições ocorrem nos seguintes casos: consoantes duplas ou consoantes seguidas pertencentes ao mesmo grupo de código; consoante imediatamente após a letra inicial que pertença ao mesmo grupo de código da letra inicial; consoantes do mesmo grupo de código separadas por W ou H.

Seguindo as regras apresentadas, o nome **Vinícius**, por exemplo, poderia ser escrito com diversas grafias: **Venicios**, **Venicius**, **Vinicios**, **Vinicius**, que mesmo assim, teria codificação *Soundex* igual a **V522**.

Em conformidade com a Tabela 3 e as regras anteriormente descritas, a Figura 19 mostra um código-fonte escrito em VBScript[76], com a implementação da função *Soundex*.

```
'
'Algoritmo SOUNDEX.
'
Function Soundex(pTexto)      'As String
'
Dim resultado                 'As String
Dim texto                     'AS String
Dim i                         'As Long
Dim ivalorSoundex             'As Integer
Dim valorPrimeiraLetra        'As Integer
'
' soundex é case-insensitive
texto = UCase(pTexto)
' a primeira letra é copiada no resultado
resultado = Left(texto, 1)
valorPrimeiraLetra = ValorSoundex(resultado)
'
For i = 2 To Len(texto)
    ivalorSoundex = valorSoundex(Mid(texto, i, 1))
    If ivalorSoundex <> 0 And valorPrimeiraLetra <> ivalorSoundex Then
       resultado = resultado & ivalorSoundex
    End If
    valorPrimeiraLetra = ivalorSoundex
    ivalorSoundex = 0
Next
Soundex = Mid(resultado, 1, 4)
If Len(resultado) < 4 Then
   Soundex = Soundex & String(4 - Len(resultado), "0")
End If
End Function
'------------------------------
'Tabela com valores SOUNDEX.
'
Function valorSoundex(pCaracter) 'As Integer
'
Select Case pCaracter
       Case "B", "F", "P", "V"
            valorSoundex = "1"
       Case "C", "G", "J", "K", "Q", "S", "X", "Z"
            valorSoundex = "2"
       Case "D", "T"
            valorSoundex = "3"
       Case "L"
            valorSoundex = "4"
       Case "M", "N"
            valorSoundex = "5"
       Case "R"
            valorSoundex = "6"
End Select
End Function
```

Figura 19: Função *Soundex* escrita em VBScript.

[76] http://msdn.microsoft.com/en-us/library/t0aew7h6.aspx

4.4.2. Distância *Levenshtein*

A distância de edição, também chamada de "*combinação de strings com K diferenças*" ou distância *Levenshtein* foi concebida pelo cientista russo Vladimir Levenshtein em 1965. Levenshtein publicou os trabalhos [Levenshtein 1965] e [Levenshtein 1966], descrevendo a técnica que é hoje a mais conhecida e aplicada para medir similaridade entre *strings* [Navarro 2001]. O objetivo dessa técnica é ter o menor número de operações para tornar duas *strings* iguais. As operações permitidas são: inserções (*insertions*), remoções (*deletions*) e substituições (*substitutions*) de caracteres. Todas as operações têm custo (ou peso) igual a um. O algoritmo *Levenshtein* retorna um valor que vai de 0 (quando as duas *strings* são iguais) ao tamanho da maior *string* (quando as duas *strings* são totalmente diferentes). Formalmente, a distância *Levenshtein* é dada pela seguinte expressão (1) [Navarro 2001]:

$$0 \leqslant d(x, y) \leqslant \max(|x|, |y|) \qquad (1)$$

O algoritmo para calcular a distância de edição usa uma matriz $(n + 1) \times (m + 1)$, onde n e m é o número de caracteres de cada uma das duas *strings*. O objetivo desse algoritmo é transformar o segmento inicial *X1..i* no segmento *Y1..j* utilizando um mínimo de *Ci,j* operações. Ao final da execução, o elemento no canto inferior direito da matriz, *Cm,n*, contém a distância de edição entre *X* e *Y* [Navarro 2001] e [Oliveira 2007].

A Figura 20 ilustra uma matriz utilizada para calcular a distância de edição entre as datas 03/04/1970 e 27/10/1964, citadas no início da Seção 4.4. Ambas as datas estão no formato "AAAAMMDD". Como as datas são bem diferentes entre si, o escore é 6, quando o máximo é 8. As células que compõem a diagonal principal indicam o caminho para o resultado final.

		1	**9**	**7**	**0**	**0**	**4**	**0**	**3**
	0	1	2	3	4	5	6	7	8
1	1	**0**	1	2	3	4	5	6	7
9	2	1	**0**	1	2	3	4	5	6
6	3	2	1	**1**	2	3	4	5	6
4	4	3	2	2	**2**	3	3	4	5
1	5	4	3	3	3	**3**	4	4	5
0	6	5	4	4	3	3	**4**	4	5
2	7	6	5	5	4	4	4	**5**	5
7	8	7	6	5	5	5	5	5	**6**

Figura 20: Matriz gerada para o cálculo de distância de edição entre as datas 03/04/1970 e 27/10/1964.

A Figura 21 ilustra uma matriz utilizada para calcular a distância de edição entre as datas 03/04/1970 e 30/04/1970, também citadas no início desta Seção. Observe que essas datas são bem próximas e pode ter havido um erro de digitação, quando da entrada de dados. O escore neste caso é 2, para o máximo de 8. Assim como na Figura 20, as células da diagonal principal indicam o caminho para o resultado final.

		1	**9**	**7**	**0**	**0**	**4**	**0**	**3**
	0	1	2	3	4	5	6	7	8
1	1	**0**	1	2	3	4	5	6	7
9	2	1	**0**	1	2	3	4	5	6
7	3	2	1	**0**	1	2	3	4	5
0	4	3	2	1	**0**	1	2	3	4
0	5	4	3	2	1	**0**	1	2	3
4	6	5	4	3	2	1	**0**	1	2
3	7	6	5	4	3	2	1	**1**	1
0	8	7	6	5	4	3	2	1	**2**

Figura 21: Matriz gerada para o cálculo de distância de edição entre as datas 03/04/1970 e 30/04/1970.

Oliveira (2007) apresenta, na Figura 22, o algoritmo que calcula a distância de *Levenshtein*, aplicado nas *strings X* e *Y*. A *string X* possui comprimento *n*, enquanto a *string Y* tem comprimento *m*.

```
int DistanciaLevenshtein (char X[1..n], char Y[1..m])
   // C é uma matriz com n+1 linhas e m+1 colunas
   declare int C[0..n, 0..m]
   // i e j são usados nas iterações de X e Y
   declare int i, j, cost

   for i from 0 to n
       C[i, 0] := i
   for j from 0 to m
       C[0, j] := j

   for i from 1 to n
       for j from 1 to m
           if X[i] = Y[j] then cost := 0
                          else cost := 1
           C[i, j] := minimum(
                       C[i-1, j  ] + 1,     // deleção
                       C[i  , j-1] + 1,     // inserção
                       C[i-1, j-1] + cost   // substituição
                             )
    return C[n,m]
```

Figura 22: Algoritmo que calcula a distância de *Levenshtein* [Oliveira 2007].

A Figura 23 apresenta a implementação do algoritmo de distância de *Levenshtein* em VBScript, utilizada neste trabalho. Nessa implementação, há dois destaques:

- O primeiro, evidenciado pelo número 1, mostra uma adaptação do código-fonte, uma vez que a linguagem de programação VBScript não possui a função *min( )* ou *minimum( )* implementada internamente.
- O segundo destaque, evidenciado pelo número 2, mostra a criação de um *Fator*, que será utilizado na etapa de **Classificação e Atribuição de Pesos**. Como a distância de *Levenshtein*, neste trabalho, é utilizada para cálculo de distância entre datas no formato "AAAAMMDD", calcula-se o *Fator* com a seguinte expressão (2):

$$Fator = 1 - \left(\frac{distância\ de\ Levenshtein}{8} \right) \quad (2)$$

Então, quando a distância de *Levenshtein* é igual a 0 (zero), o *Fator* é igual a 1, ou seja, as duas *strings* são 100% iguais. Quando a distância de edição é igual a 8, o *Fator* é igual a 0, significando que as datas são completamente diferentes. O *Fator*, portanto, calcula o percentual de semelhança entre as duas datas.

```
'
'Algoritmo Levenshtein
'
Function Levenshtein( a, b )
'
      Dim i             'As Integer
      Dim j             'As Integer
      Dim cost          'As Integer
      Dim d             'As Integer
      Dim min1          'As Integer
      Dim min2          'As Integer
      Dim min3          'As Integer
      Dim Fator         'As Double

      If Len( a ) = 0 Then
         Levenshtein = Len( b )
         Exit Function
      End If

      If Len( b ) = 0 Then
         Levenshtein = Len( a )
         Exit Function
      End If

      ReDim d( Len( a ), Len( b ) )

      For i = 0 To Len( a )
          d( i, 0 ) = i
      Next

      For j = 0 To Len( b )
          d( 0, j ) = j
      Next

      For i = 1 To Len( a )
          For j = 1 To Len( b )
              If Mid( a, i, 1 ) = Mid( b, j, 1 ) Then
                 cost = 0
              Else
                 cost = 1
              End If
              '
              min1 = ( d( i - 1, j ) + 1 )           'deleção
              min2 = ( d( i, j - 1 ) + 1 )           'inserção
              min3 = ( d( i - 1, j - 1 ) + cost )    'substituição
              '
              If min1 <= min2 And min1 <= min3 Then
                 d( i, j ) = min1
              ElseIf min2 <= min1 And min2 <= min3 Then
                     d( i, j ) = min2
                  Else
                     d( i, j ) = min3
                  End If
          Next
      Next
      Fator = 1 - (d( Len( a ), Len( b ) )/8)
      Levenshtein = Fator
End Function
```

1

2

Figura 23: Função *Levenshtein* escrita em VBScript.

4.5. Atribuição de Pesos e Classificação

Esta etapa ocorre concomitantemente com a **Comparação de Campos**. Cada par, composto pelos registros de cada uma das fontes de dados, possui um conjunto de seus campos que são comparados. Para cada comparação de campos (ou parte de campos) é calculado um escore. Quando os campos são iguais ou possuem uma situação de concordância aceitável, este escore contribui positivamente para classificar este par como combinado (*macth*). Caso contrário, o escore contribui negativamente, pesando para que este par seja classificado com não-combinado (*non-match*). O escore final, que classifica os pares como combinados, não combinados ou duvidosos (*possible links*), é a soma desses escores parciais. A todo este ciclo, alguns autores [Camargo e Coeli 2000], chamam de pareamento.

Para que seja mais bem compreendida a etapa de pareamento, é necessário que alguns conceitos sejam abordados. Esses conceitos foram inicialmente propostos por Newcombe et al. (1959) e desenvolvidos posteriormente por Fellegi & Sunter (1969):

- **Escore limiar:** são limites que classificam os pares em três categorias: combinados, não-combinados e duvidosos. Ou seja, os pares que apresentam escores totais maiores ou iguais a um valor predeterminado (**limiar superior** ou ***upper threshold***) são classificados como verdadeiros (ou combinados). Já aqueles pares que apresentarem um escore total abaixo do **limiar inferior** (ou ***lower threshold***) são considerados falsos. Os pares cujos escores totais estiverem entre os dois limiares são classificados como duvidosos e devem ser revisados manualmente.

- **Escore total:** também conhecido como escore final. São os escores de cada par, obtidos a partir da soma dos escores ponderados de cada campo. Esses escores ponderados são calculados durante a etapa de **Comparação de Campos**. Cada campo comparado contribui de modo diferenciado para o escore total do par [Camargo e Coeli 2000].

Quando um par de campos proveniente de registros de fontes de dados diferentes é comparado, podem ocorrer quatro possibilidades:

1. Se os campos concordarem entre os dois registros e se tratar de um par verdadeiro, é dito que o par é um **verdadeiro positivo** (***true match***) ou, simplesmente, verdadeiro;
2. Se os campos concordarem entre os dois registros e for um par falso, então este par é um **falso positivo** (***false match***);

3. Se os campos discordarem entre os dois registros e for um par falso, então o par é um **verdadeiro negativo** (***true non-match***) ou, simplesmente, falso;
4. Se os campos discordarem entre os dois registros e for um par verdadeiro, então o par é um **falso negativo** (***false non-match***).

Para melhor ilustrar as possibilidades ao se comparar os campos na etapa de pareamento, foi produzida a Figura 24.

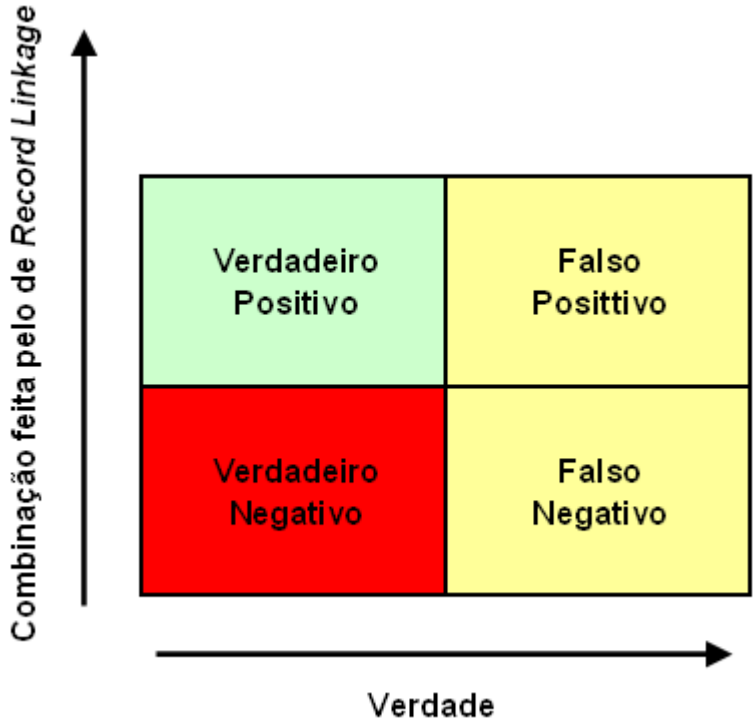

Figura 24: Possibilidades ao se comparar campos no Método *Record Linkage*.

Para cada par de campos ***i***, é definido m_i como a probabilidade do par ser verdadeiro. Essa probabilidade é chamada de **sensibilidade** (ou ***sensitivity***). Já u_i é a probabilidade do par ser um falso positivo. Essa probabilidade é expressa pela fórmula (**1 – especificidade**, ou ***1 - specificity***).

De forma análoga, ($1 - m_i$) representa a probabilidade de o par ser um falso negativo. Ao passo que ($1 - u_i$) representa a probabilidade de o par ser falso. Esta probabilidade é chamada de **especificidade** (ou ***specificity***).

A Tabela 4 mostra um quadro-resumo dos conceitos de probabilidade descritos anteriormente.

Tabela 4: Conceitos de probabilidade para o par de campos comparado.

Definição	Fórmula	Conceito
Sensibilidade	m_i	Probabilidade de o par ser verdadeiro.
Especificidade	$(1 - u_i)$	Probabilidade de o par ser falso
(1 – Sensibildade)	$(1 - m_i)$	Probabilidade de o par ser falso negativo.
(1 – Especificidade)	u_i	Probabilidade de o par ser falso positivo.

Segundo Camargo e Coeli (2000), os valores de m_i e u_i, assim como os valores dos limiares superiores e inferiores podem ser estimados. Fellegi e Sunter (1969) e Jaro (1989) propõem uma metodologia para a estimativa desses parâmetros. Também podem ser utilizados valores depurados a partir de um histórico de execução do método de *Record Linkage* várias vezes.

Herzog et al. (2007) fez um estudo sobre o uso de algoritmos EM (*Expectation-Maximization*) [Moon 1996] para estimar os parâmetros de pareamento: sensibilidade e especificidade para comparação entre campos. Neste trabalho, m_i e u_i foram estimados com base no estudo realizado por Camargo e Coeli (2000) e seguem, basicamente, a Tabela 5.

Tabela 5: Parâmetros de sensibilidade e especificidade seguidos neste trabalho.

Tipo de Campo	Sensibilidade (m_i)	1 – Especificidade (u_i)
Nome (ou parte do nome)	**90%**	**5%**
Data	**90%**	**10%**
Sexo	**95%**	**50%**

Os autores encontraram esses números baseados no poder discriminatório dos campos e na probabilidade de terem seus conteúdos registrados corretamente. Por exemplo: o campo sexo mostra baixo poder discriminatório (há apenas duas possibilidades de preenchimento), mas o seu registro é, em geral, feito de forma correta. Já o campo “último nome”, apesar de possuir um bom poder discriminatório (poder de identificar um indivíduo), está mais sujeito a erros de registro.

Com base nas probabilidades estudadas, são construídos dois fatores de ponderação: um, para a situação de concordância e outro, para a situação de discordância [Camargo e Coeli 2000]. Um determinado campo do primeiro registro é comparado com um do segundo registro, se eles concordarem, aplica-se o **fator de ponderação de concordância** (3):

$$wc_i = \log_2\left(\frac{m_i}{u_i}\right) \qquad (3)$$

Caso contrário, aplica-se o **fator de ponderação de discordância** (4):

$$wd_i = \log_2\left[\frac{(1-m_i)}{(1-u_i)}\right] \qquad (4)$$

O **escore** total (E_t), representado pela expressão (5), de determinado par é obtido a partir da soma dos fatores de ponderação atribuídos após a comparação de cada campo.

$$E_t = \sum_{i=1}^{n} wx_i \qquad (5)$$

Nessa expressão (5), *n* é o número de campos que são comparados de cada registro e *x* é o fator de ponderação de concordância ou discordância.

Como, no caso abordado nesta dissertação, $\boldsymbol{m_i}$ é sempre maior que $\boldsymbol{u_i}$, o **fator de ponderação de concordância** gera um número positivo, contribuindo positivamente para o **escore total**. Já o **fator de ponderação de discordância** gera um número negativo, contribuindo para diminuir o valor do **escore total** [Jaro 1989].

Jaro (1989) ressalta que nem sempre é fácil decidir sobre a concordância ou discordância entre dois campos de determinado par. Como visto nos cálculos de datas na etapa de **Comparação de Campos**, o simples fato de uma data ser diferente da outra não significa que essa diferença tem que contribuir para a diminuição drástica do **escore total**. Nesses casos, Jaro (1989) propõe a adoção de uma **proporção mínima de concordância**. Portanto, quando a discordância entre os campos é pequena, é atribuído um valor que contribui positivamente com o **escore total**. Porém, essa contribuição é menor do que aquela utilizada no caso da concordância ser exata. Por exemplo: Se o *Fator* da distância de *Levenshtein*, visto na expressão (2) for maior ou igual a 0,75 e menor que 1, é aplicado este *Fator* ao **fator de ponderação de concordância.** Se este *Fator* for menor que 0,75, os campos são tratados como diferentes, sendo aplicado o **fator de ponderação de discordância**.

A seguir, é apresentado um exemplo de aplicação dos conceitos de **Classificação e Atribuição de Pesos**.

4.5.1. Aplicação de Conceitos

Dadas duas fontes de dados: **A** e **B**, e um planejamento da comparação de campos, realizado de acordo com a Tabela 6:

Tabela 6: Planejamento da comparação de campos.

Campo	Tipo de Campo	Algoritmo	Sensibilidade (m_i)	1 – Especificidade (u_i)	PMC*
C1	Primeiro Nome	*Soundex*	90%	5%	-
C2	Último Nome	*Soundex*	90%	5%	-
C3	Sexo	Comparação Exata	95%	50%	-
C4	Data de Nascimento	*Levenshtein*	90%	10%	75%

* PMC: Proporção Mínima de Concordância

Foi produzido o código genérico descrito na Tabela 7. No exemplo hipotético, não foi utilizada blocagem. Ou seja, são comparados todos os registros da fonte de dados **A** com todos os da fonte de dados **B**.

Tabela 7: Código exemplo para comparação de campos, classificação e atribuição de pesos.

Código Genérico	Fator de Ponderação de Concordância	Fator de Ponderação de Discordância
`Et = 0`	-	-

`Se Soundex(A.C1) = Soundex(B.C1)`
`  Et = Et + log2(90/5)`
`Senão`
`  Et = Et + log2(10/95)`
	4,170	-3,248

`Se Soundex(A.C2) = Soundex(B.C2)`
`  Et = Et + log2(90/5)`
`Senão`
`  Et = Et + log2(10/95)`
	4,170	-3,248

`Se A.C3 = B.C3`
`  Et = Et + log2(95/50)`
`Senão`
`  Et = Et + log2(5/50)`
	0,926	-3,322

`Se Fator(A.C4,B.C4) >= 0.75`
`  Et = Et + Fator(A.C4,B.C4)*log2(90/10)`
`Senão`
`  Et = Et + log2(10/90)`
	3,170	-3,170

Neste exemplo, o maior **escore total** (E_t) possível é 12,436 e o menor é -12,988. Com base nesses números, é possível estabelecer os limiares superior e inferior. Por exemplo, poderiam ser

considerados verdadeiros ou pares combinados aqueles que tivessem **escore total** acima de 10. Os pares duvidosos poderiam ser aqueles que tivessem obtido escores positivos abaixo de 10. Já os falsos, poderiam ser os pares que tivessem **escore total** negativo.

Os limiares definidos podem ser refeitos a partir da **Revisão Humana e Avaliação dos Resultados**, que será vista na próxima Seção.

4.6. Revisão Humana e Avaliação dos Resultados

Após a classificação dos pares comparados, ainda há a necessidade de execução de dois processos:

- A revisão humana dos pares classificados como duvidosos e
- A avaliação do método de *Record Linkage* utilizado, com o objetivo de aprimorá-lo.

Após a execução da parte automatizada do *Record Linkage*, é gerado um gráfico semelhante ao apresentado na Figura 25, chamado de gráfico dos escores [Grannis 2008].

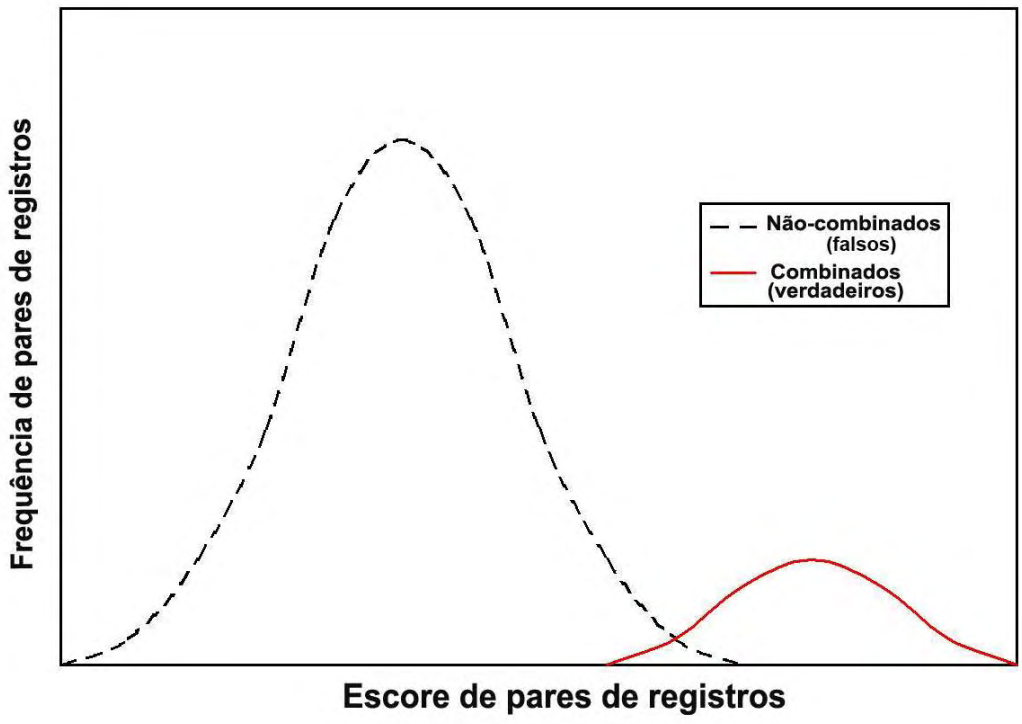

Figura 25: Gráfico dos Escores [Grannis 2008].

A abscissa do gráfico representa os escores obtidos pelos pares. O eixo das ordenadas representa a frequência dos pares de registros, ou seja, o número de pares de registros comparados que

obtiveram o mesmo escore. Com isso, a linha pontilhada representa os pares comparados que não apresentaram um escore suficiente para serem verdadeiros. Já a linha sólida, com amplitude menor, representa os pares que obtiveram um escore suficiente para serem considerados verdadeiros ou duvidosos.

Se a interseção entre os gráficos com pares verdadeiros e falsos for ampliada, um gráfico semelhante à Figura 26 é obtido.

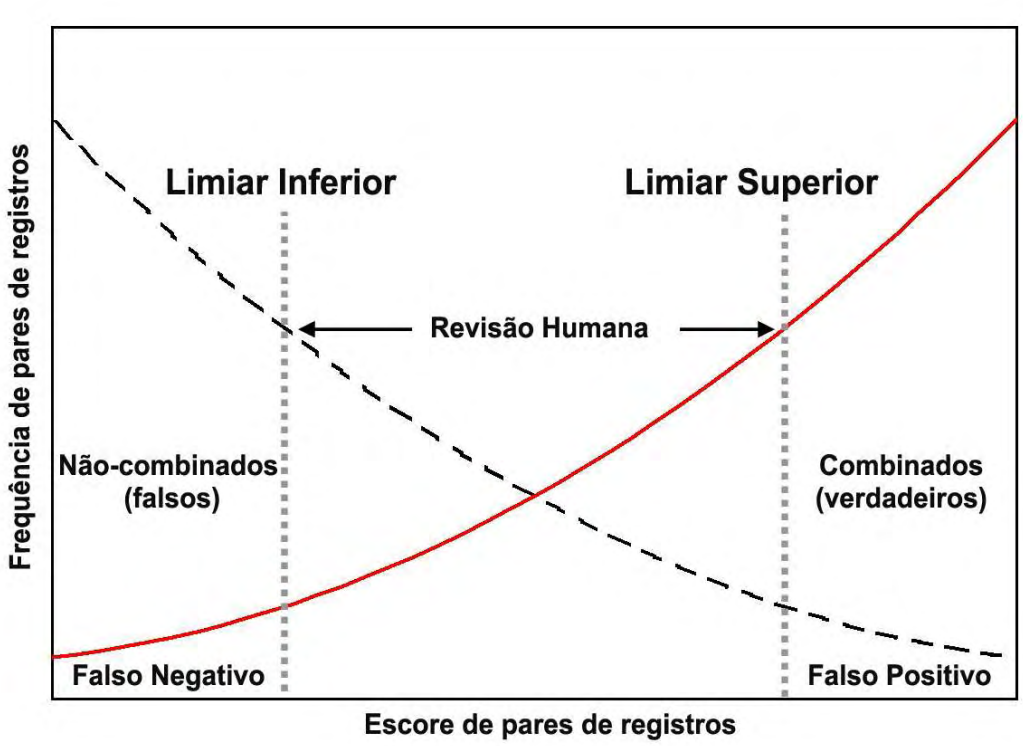

Figura 26: Ampliação da interseção entre os gráficos da Figura 25 [Grannis 2008].

Este trecho de gráfico apresentado na Figura 26 é bem rico em detalhes. Aqui são observados os limiares estabelecidos, tanto o superior quanto o inferior. Os pares contidos entre os dois limiares necessitarão de **Revisão Humana**. Já os pares falsos que estiverem à direita do **limiar superior** serão considerados verdadeiros e comporão o universo de pares falsos positivos. Da mesma forma, os pares verdadeiros que estiverem à esquerda do **limiar inferior** serão considerados falsos, compondo o grupo de pares falsos negativos.

Nem sempre é possível fazer uma revisão manual de toda a base. Em casos de fontes de dados grandes, essa comparação torna-se impraticável. Mesmo se fosse possível este tipo de comparação, a probabilidade de falha seria muito grande, por mais criteriosos que fossem os conferentes.

Às vezes é impraticável conferir até mesmo os duvidosos. Se o número de pares entre os dois limiares estiver na casa dos milhares, a revisão além de onerosa e longa, também estará sujeita a falhas. A **Revisão Manual** é recomendada quando os duvidosos estiverem na casa das centenas. É necessário também que os conferentes tenham à disposição um ferramental de software para que eles possam tomar a decisão de transformar um duvidoso em positivo, gerando um MPI para ele, ou em negativo, descartando-o.

A eficácia de um sistema de *Record Linkage* pode ser medida com uma adaptação dos conceitos de *Recall* e *Precision*, provenientes da Recuperação da Informação (*Information Retrieval*) [Baeza-Yates e Ribeiro-Neto 1999]. Quanto menor for o número de falsos positivos e falsos negativos encontrados, maior será a eficácia do método utilizado.

Nesse contexto, o ***Recall*** representa o índice de recuperação ou retorno. Ele é a proporção entre o número de documentos retornados e o número de documentos que realmente deveriam ser retornados. A expressão (6) mostra o cálculo do ***Recall***, originalmente concebido:

$$Recall = \frac{número\ de\ documentos\ relevantes\ recuperados}{número\ de\ documentos\ relevantes\ no\ sistema} \qquad (6)$$

Adaptando o conceito de ***Recall*** para o método de *Record Linkage*, obtém-se a expressão (7):

$$Recall = \frac{número\ de\ pares\ verdadeiros}{número\ de\ pares\ verdadeiros + número\ de\ falsos\ negativos} \qquad (7)$$

O parâmetro ***Precision*** representa a proporção de documentos recuperados que são relevantes para a pesquisa. Ele é a taxa de documentos úteis entre o total de documentos recuperados. A expressão (8) mostra o cálculo do ***Precision***, originalmente concebido:

$$Precision = \frac{número\ de\ documentos\ relevantes\ recuperados}{número\ de\ documentos\ recuperados} \qquad (8)$$

Adaptando o conceito de ***Precision*** para o método de *Record Linkage*, obtém-se a expressão (9):

$$Precision = \frac{número\ de\ pares\ verdadeiros}{número\ de\ pares\ verdadeiros + número\ de\ falsos\ positivos} \qquad (9)$$

Segundo Branting (2003), deve haver um equilíbrio entre o *Recall* e o *Precision*. O autor completa que o melhor algoritmo é aquele que otimiza o equilíbrio entre os dois parâmetros. O *Recall* deve ser o mais alto possível, sem penalizar o *Precision*.

Para expressar o balanceamento entre o *Recall* e o *Precision*, Rijsbergen (1979) combinou esses dois parâmetros e formulou uma medida de desempenho geral *F*, baseada na média harmônica entre o *Recall* e o *Precision*, representada pela expressão (10):

$$F = \frac{2 \ PR}{P + R} \qquad (10)$$

Winkler (2004) conclui que relacionar bases de dados usando o método probabilístico ainda é uma questão em aberto.

4.7. Geração do MPI

Os pares combinados e os duvidosos que, após uma revisão manual, forem promovidos a combinados irão compor o *Master Patient Index* (MPI).

O MPI é composto por um identificador único gerado, associado aos identificadores de cada fonte de dados. Seu objetivo é relacionar esse identificador-mestre às múltiplas fontes de dados ou a vários sistemas de informação. Com isso, é possível garantir uma identificação única de pacientes em fontes de dados heterogêneas e distribuídas, tema deste trabalho. Essa identificação única permitirá a consulta às fontes de dados originais.

5. Projeto e Implementação

A partir do estudo detalhado do processo de *Record Linkage*, apresentado no Capítulo anterior, e da definição dos requisitos de um IHE/PIX clássico [ACC, HIMSS e RSNA 2008a], neste Capítulo, é apresentada a modelagem da solução proposta. Esta modelagem inclui: casos de uso padrões do perfil de integração PIX, diagrama de sequência, diagrama de pacotes, diagramas de classes, projeto lógico e físico do banco de dados, diagrama hierárquico de funções e desenho das páginas web. Ao final, são apresentados os ambientes de hardware e software utilizados neste projeto.

A solução proposta tem como base o perfil de integração PIX do padrão IHE. Contudo, esse perfil de integração, assim como qualquer perfil de integração do padrão IHE, não define políticas, técnicas ou algoritmos que devem ser utilizados. A especificação do PIX tem a função de definir a interoperabilidade necessária para troca de informações entre os atores.

Neste trabalho, não só foi desenvolvido um PIX, baseado nos requisitos definidos pelo padrão IHE, como também foram desenvolvidas funcionalidades que agregam valor ao processo de identificação de pares existentes em diversas fontes gerando um índice único. Destacam-se como valores agregados à definição do PIX convencional: (*i*) o retorno às fontes originais de forma on-line, (*ii*) a atualização incremental do índice gerado, (*iii*) a flexibilidade na configuração das etapas que compõem o *Record Linkage* e (*iv*) a combinação das fontes de dados com tipos e formatos diferentes.

5.1. Requisitos para Desenvolvimento de um IHE/PIX

O perfil de integração PIX, descrito na Seção 3.4, possui casos de uso padrões definidos por documentos técnicos oficiais homologados pela instituição IHE [ACC, HIMSS e RSNA 2008a]. Esses casos de uso são baseados na arquitetura do PIX, também descrito na Seção 3.4.

O caso de uso ***Patient Identity Feed*** (Alimentação da Identidade do Paciente), apresentado na Figura 27, possui como escopo o fornecimento de informações sobre o paciente, incluindo a confirmação de dados demográficos, após a identidade de o paciente ter sido estabelecida, modificada ou combinada com outros dados.

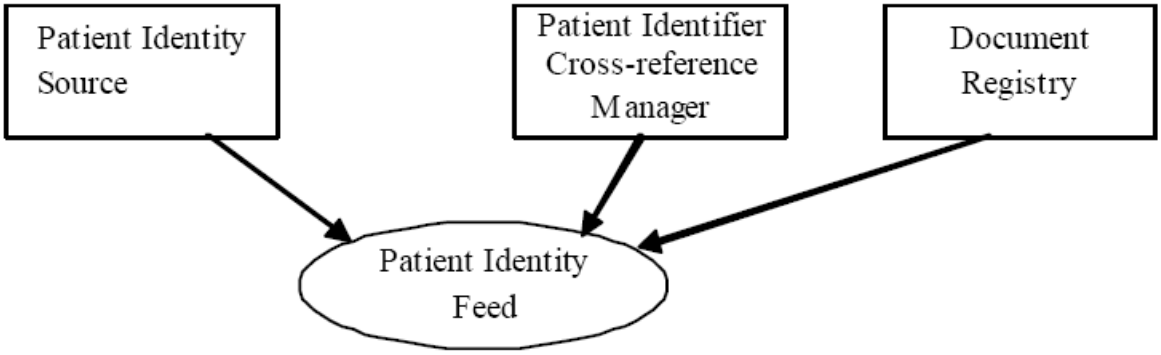

Figura 27: Caso de uso *Patient Identity Feed* [ACC, HIMSS e RSNA 2008a].

- **Ator:** *Patient Identity Source* (Fonte da identidade do paciente).
- **Papel:** Envia uma notificação para o ator *Patient Identifier Cross-reference* (*PIX Manager*) e para o ator *Document Registry* **(**ator pertencente ao perfil de integração XDS) toda vez que há uma inclusão, alteração ou combinação nos dados utilizados para identificação do paciente.
- **Ator:** *Patient Identifier Cross-reference* (Referência cruzada para identificadores de pacientes).
- **Papel:** Baseado nas informações fornecidas por cada fonte de dados, através dos atores *Patient Identity Source*, ele gerencia a referência cruzada entre os identificadores de pacientes nessas fontes.
- **Ator:** *Document Registry* (Registro do documento).
- **Papel:** Este ator não tem função no PIX. Ele é utilizado no perfil XDS. O *Document Registry* usa identificadores de pacientes fornecidos pelo *Patient Identity Source* para se assegurar de que metadados de documentos XDS estejam associados com pacientes conhecidos. Ele também atualiza a identificação de pacientes em metadados de documentos, através do monitoramento de operações de mudança de identidade, como o agrupamento de dados de dois pacientes, classificando como um único.

O caso de uso ***PIX Query*** (Consultas PIX), apresentado na Figura 28, recebe uma requisição do ator *Patient Identifier Cross-reference Consumer* com a identificação de um paciente conhecida pelo cliente (consumidor). A requisição é recebida pelo *Patient Identifier Cross-reference Manager*, que processa a requisição e retorna uma resposta, em formato de lista, com todos os identificadores daquele paciente, caso haja algum.

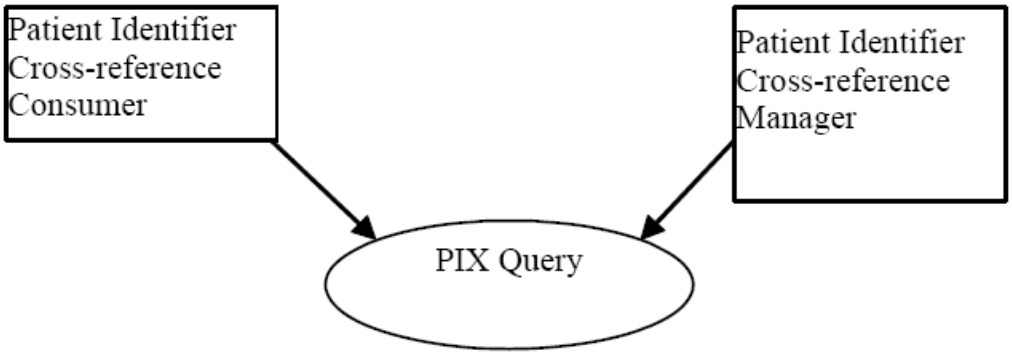

Figura 28: Caso de uso *PIX Query* [ACC, HIMSS e RSNA 2008a].

- **Ator:** *Patient Identifier Cross-reference Consumer* (Consumidor da referência cruzada de identificador de paciente).
- **Papel:** Consulta o *Patient Identifier Cross-reference Manager* (*PIX Manager*) com o objetivo de obter uma lista de identificadores de pacientes, se houver.
- **Ator:** *Patient Identifier Cross-reference Manager* (*PIX Manager*).
- **Papel:** Gerencia a referência cruzada de identificadores de pacientes. Mediante uma requisição recebida, retorna uma lista de identificadores correspondentes ao paciente, se houver.

O caso de uso ***PIX Update Notification*** (Notificação de Atualização PIX), apresentado na Figura 29, possui o seguinte escopo: o ator *Patient Identifier Cross-reference Manager* fornece notificações de atualizações aos diversos *Patient Identifier Cross-reference Consumers* associados. O *PIX Manager* possui uma lista, que pode ser configurada, com a relação dos *Patient Identifier Cross-reference Consumers* interessados em receber cada tipo de notificação. Essa transação usa a mensagem genérica "*Update Person Information*" do protocolo HL7 para comunicar esse tipo de informação [HL7 2003].

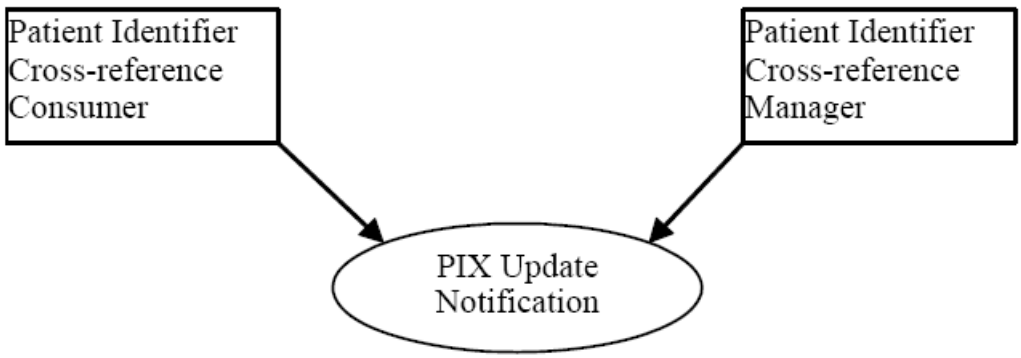

Figura 29: Caso de uso *PIX Update Notification* [ACC, HIMSS e RSNA 2008a].

- **Ator:** *Patient Identifier Cross-reference Manager* (*PIX Manager*).
- **Papel:** O *PIX Manager* serve a um conjunto bem definido de Domínios de Identificação de Pacientes (*Patient Identification Domains*), gerenciando a referência cruzada de identificadores de pacientes neste domínio. Ele fornece uma lista de "aliases" de identificação de pacientes (*patient ID aliases*), via notificações de atualização, para uma lista configurada de *Patient Identifier Cross-reference Consumer* interessados.
- **Ator:** *Patient Identifier Cross-reference Consumer*.
- **Papel:** Recebe notificações com mudanças nos *patient ID aliases*, vindas do ator *Patient Identifier Cross-refernce Manager*. Tipicamente, o *Patient Identifier Cross-reference Consumer* usa essas notificações para manter links de informações sobre pacientes em diferentes domínios.

Há diagramas de sequência específicos e formais para o perfil de integração PIX [ACC, HIMSS e RSNA 2008] e [ACC, HIMSS e RSNA 2008a]. Contudo, eles não representam de maneira clara cada caso de uso visto anteriormente. Por esse motivo, foi elaborado um diagrama de sequência que retrata os requisitos descritos nesses casos de uso e baseado nos diagramas de sequência encontrados nos documentos oficiais (Figura 30). Para uma melhor compreensão do diagrama, foram utilizados dois *Patient Identity Sources*: Domínios A e B.

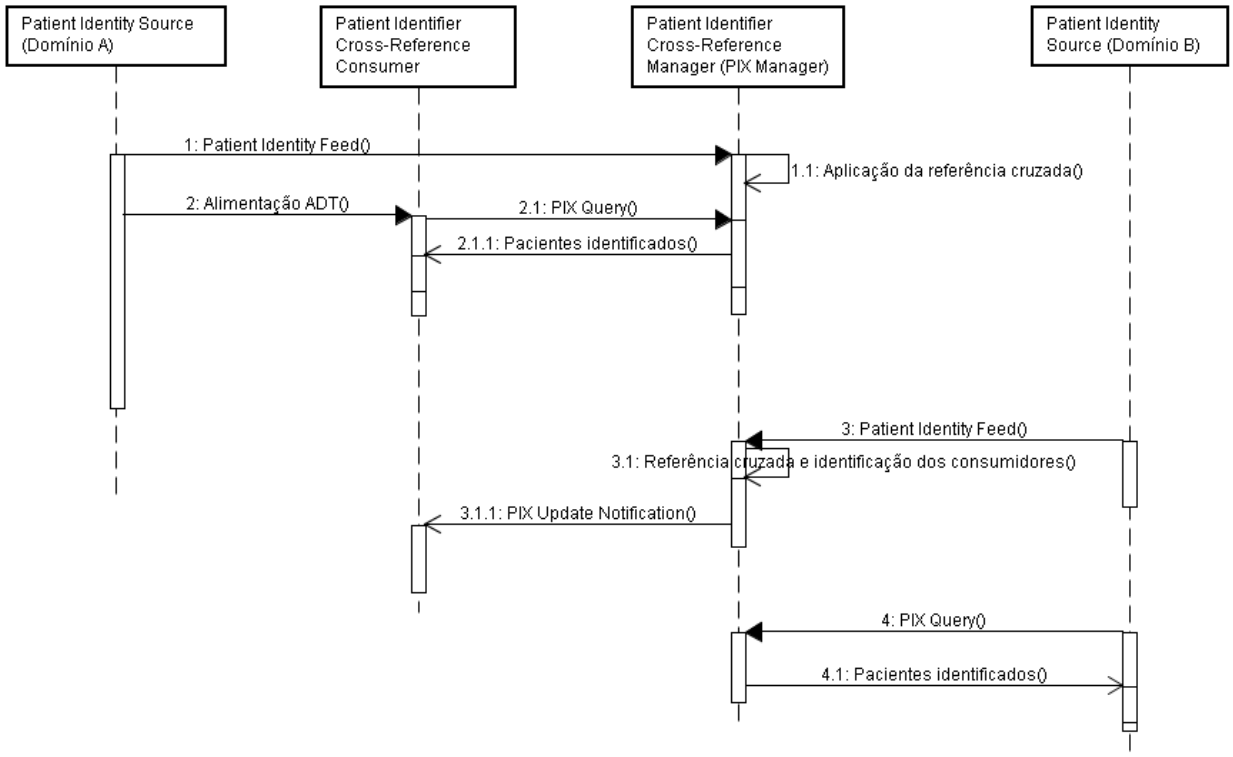

Figura 30: Diagrama de sequência do perfil de integração PIX.

O evento 1 começa com o disparo de uma mensagem *Patient Identity Feed* para o *PIX Manager*, que aplica uma referência cruzada entre os dados recebidos e os dados que integram o índice dos pacientes cadastrados. O caso de uso *PIX Query* pode tanto ser atendido pelo evento 2 quanto pelo evento 4. No primeiro caso, o Domínio A dispara uma alimentação ADT para um *Patient Identificer Cross-Reference Consumer*, que nesse caso, pode ser um sistema de informação específico (RIS, por exemplo). Esse sistema de informação necessita saber se o paciente que sofreu uma transação ADT (entrou, saiu ou foi transferido) já existe no índice geral. Então, ele dispara um *PIX Query* para o *PIX Manager*, que retorna uma lista de pacientes identificados, se houver. Da mesma forma, uma fonte de dados específica (Domínio B) também pode enviar um *PIX Query* diretamente para o *PIX Manager*, obtendo também uma lista de pacientes identificados (evento 4). Por fim, no evento 3, um *Patient Identity Source* (Domínio B, no diagrama) pode disparar um *Patient Identity Feed* para o *PIX Manager*, que faz a referência cruzada entre os pacientes, atualiza o índice, identifica os consumidores interessados nesta atualização e os notifica (mensagem *PIX Update Notification*).

Outra forma de ilustrar os casos de uso descritos é através do painel apresentado na Figura 31 [Henderson e Bao 2005]. Observe que a ligação dos dados dos pacientes entre os diversos domínios é feita pelas notificações enviadas aos domínios interessados, mais especificamente os domínios **C** e **D**. Estes domínios usam estas informações para manter referências cruzadas entre pacientes de outros domínios.

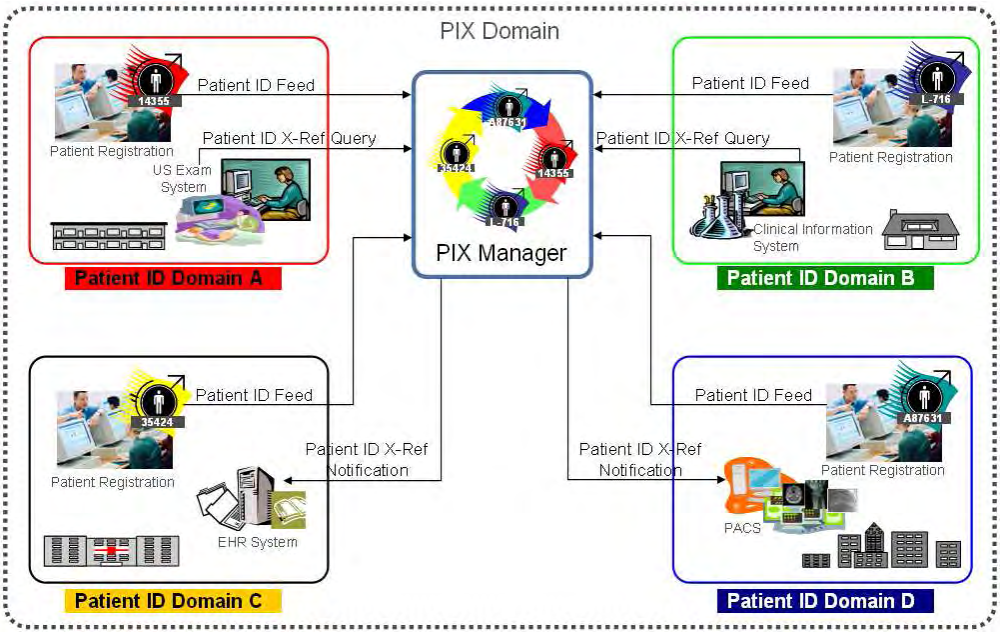

Figura 31: Painel com os casos de uso do perfil PIX [Henderson e Bao 2005].

Contudo, as diversas notificações para manter uma referência cruzada em cada domínio podem ser substituídas pela adoção de um MPI, solução adotada neste trabalho e ilustrada na Figura 32 [Henderson e Bao 2005]. Neste novo painel, é possível observar que o MPI é o instrumento que permite o gerenciamento da referência cruzada entre os diversos identificadores de pacientes.

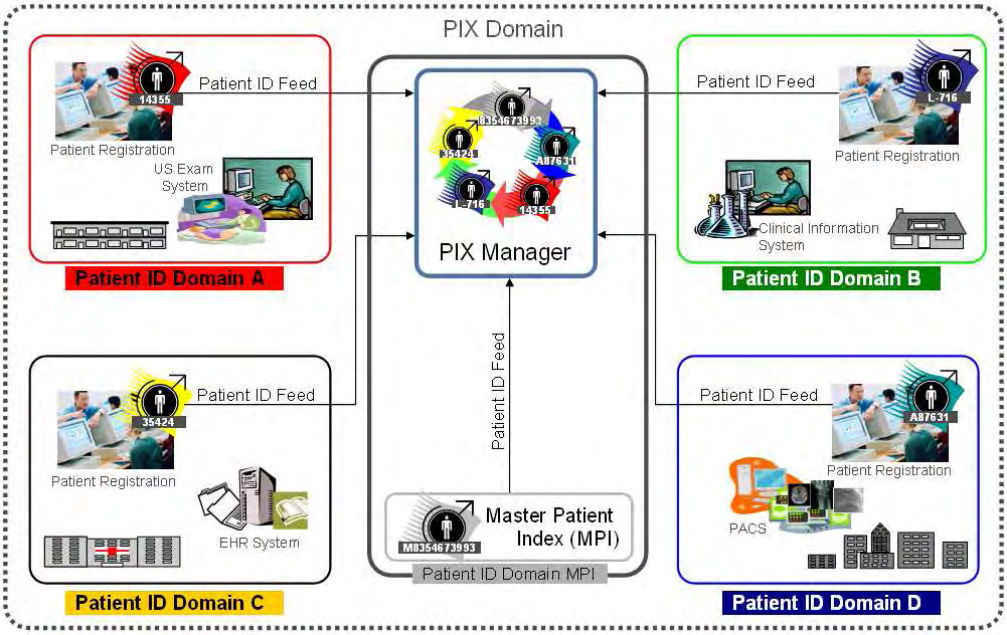

Figura 32: Perfil PIX usando MPI [Henderson e Bao 2005].

5.2. Projeto Conceitual

Com base nos requisitos para desenvolvimento de um IHE/PIX, com o uso de um MPI, e das especificações do método de *Record Linkage*, foi elaborado um projeto conceitual cujo Diagrama de Pacotes é apresentado na Figura 33.

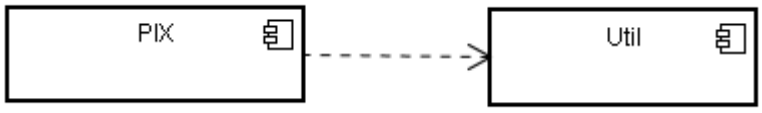

Figura 33: Diagrama de Pacotes da solução proposta.

Este diagrama é composto por dois pacotes: o ***PIX***, que mostra todas as etapas do processo de *Record Linkage* e a respectiva geração do índice único de cada linkage, chamado de MPI ou metaíndice; e o pacote ***Util***, que fornece classes e métodos que fazem o tratamento dos campos que serão comparados, além de conter os algoritmos *Soundex* e *Levenshtein*.

O diagrama de classes, apresentado na Figura 34, representa as classes utilizadas para realização do *Record Linkage* entre duas fontes de dados configuráveis. A abordagem utilizada nesta solução permite que sejam combinadas diversas fontes de dados heterogêneas duas a duas.

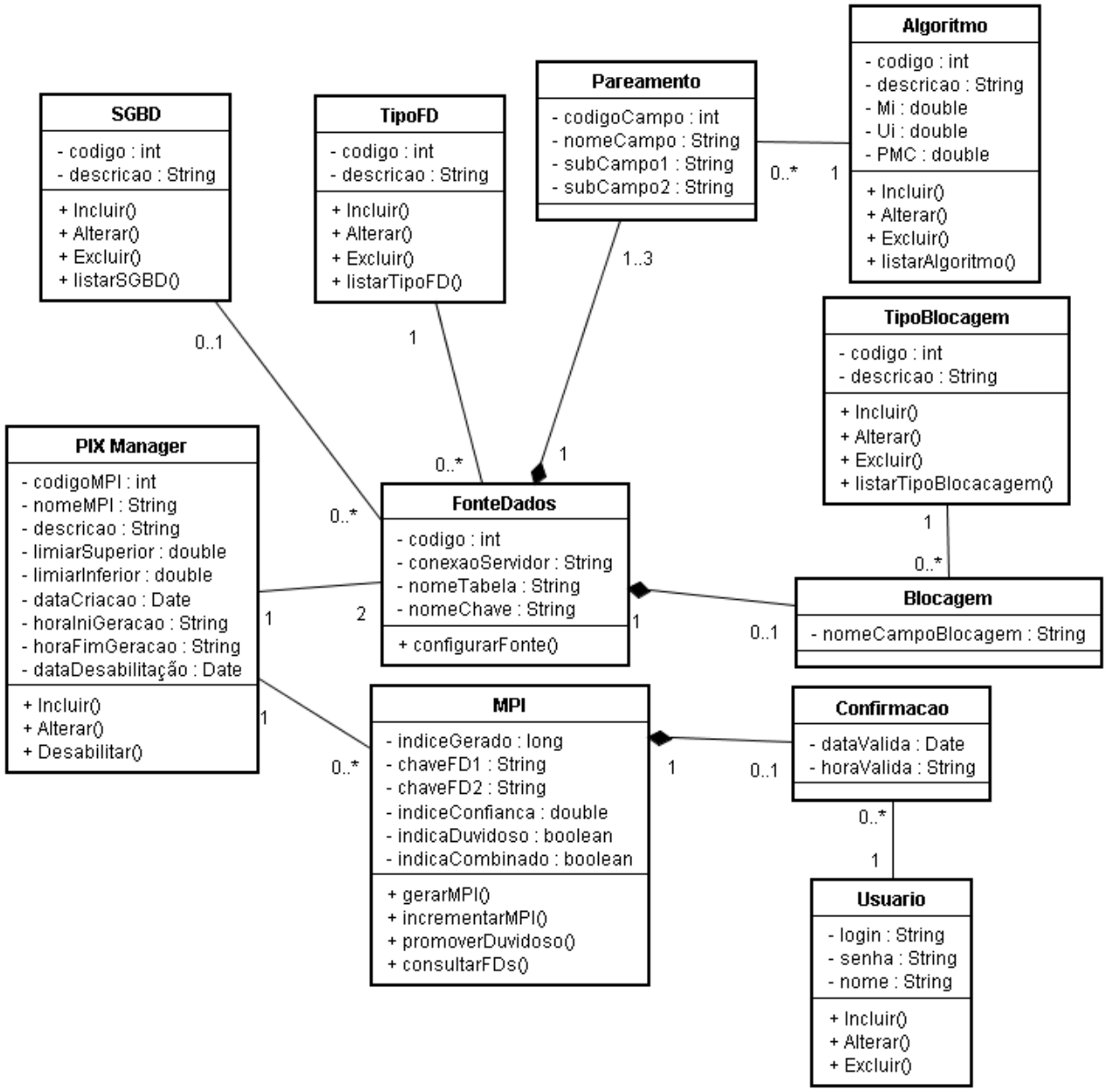

Figura 34: Diagrama de Classes do pacote *PIX*.

Cada combinação, envolvendo duas fontes distintas, é descrita na classe *PIX Manager*. A configuração de cada fonte de dados é feita através do método *configurarFonte( )* da classe *FonteDados*. Para isso, é necessário informar:

- Qual o tipo de fonte de dados: relacional ou XML;
- A *string* de conexão à fonte de dados;

- Em caso de fontes de dados relacionais, qual o SGBD utilizado;
- Em caso de fontes de dados XML, qual o esquema;
- Qual a tabela que se deseja combinar e qual a chave desta tabela.

É possível também configurar a blocagem, escolhendo o campo da tabela e o tipo de blocagem: todo o campo, primeiro nome (representa o primeiro *token* da *string*) e último nome (último *token* da *string*).

Através da classe *Pareamento*, é possível configurar quais os campos a serem comparados. Para os campos do tipo *String*, é possível subdividir a comparação em primeiro nome e último nome. O algoritmo utilizado também pode ser escolhido. Para os campos em formato texto, pode-se escolher entre comparação exata ou *Soundex*. Para os campos do tipo *Date*, é possível escolher a comparação exata ou *Levenshtein*. Os algoritmos possuem $\boldsymbol{m_i}$, $\boldsymbol{u_i}$ e PMC (Proporção Mínima de Concordância) fixos, baseados no estudo de [Camargo e Coeli 2000].

Uma vez configuradas as fontes de dados que serão comparadas, é possível gerar o MPI. O método *gerarMPI( )* da classe *MPI* é responsável por essa tarefa. Este método utiliza uma série de funções para manipulação e comparação de *strings*. Essas funções são métodos de classes, contidas no pacote ***Util***, cujo diagrama de classes é mostrado na Figura 35.

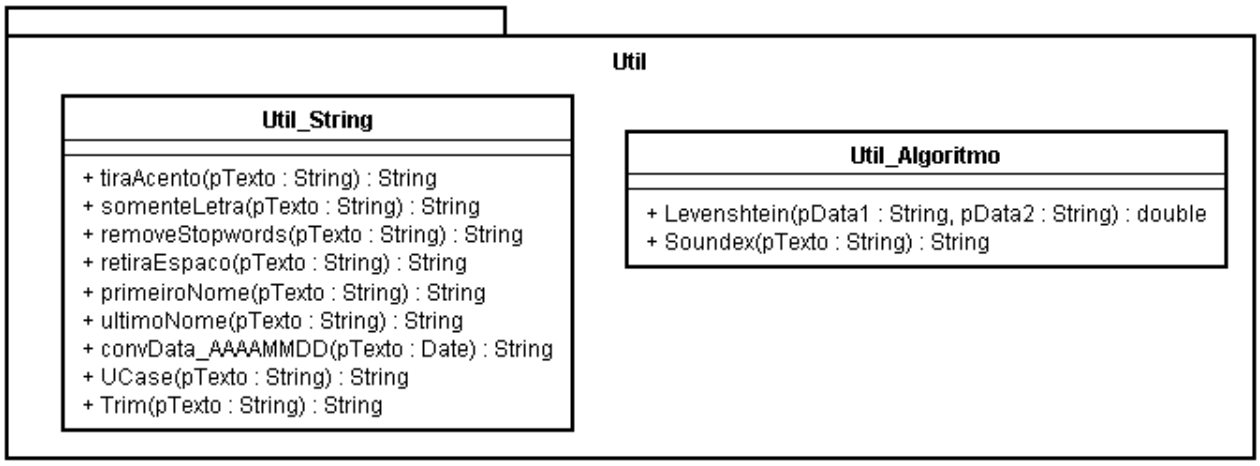

Figura 35: Diagrama de Classes do pacote *Util*.

A Tabela 8 mostra a funcionalidade de cada método de manipulação de *strings*, com os seus respectivos parâmetros de entrada. Esses métodos são funções que permitem compatibilidade entre os campos que serão comparados.

Tabela 8: Funções úteis na padronização de campos.

Função	Parâmetros de entrada	Funcionalidade
tiraAcento(String pTexto)	**pTexto**[77]: Campo do tipo string.	Remove acentuação dos nomes.
somenteLetra(String pTexto)	**pTexto**: Campo do tipo string.	Substitui caracteres especiais e números por espaços em branco.
removeStopwords(String pTexto)	**pTexto**: Campo do tipo string.	Remove preposições comuns a nomes próprios da Língua Portuguesa: de, da, do, das, dos.
retiraEspaco(String pTexto)	**pTexto**: Campo do tipo string.	Retira espaços em branco de uma *string*.
primeiroNome(String pTexto)	**pTexto**: Campo do tipo string.	Retorna o primeiro nome de um nome próprio. Retorna o conteúdo encontrado que vai do primeiro caracter preenchido da *string* até o caracter anterior ao primeiro espaço em branco. Primeiro *token* de uma *string*.
ultimoNome(String pTexto)	**pTexto**: Campo do tipo string.	Retorna o último nome de um nome próprio. Retorna o conteúdo encontrado que vai do caracter preenchido após o último espaço em branco até o final da *string*. Último *token* de uma *string*.
convData_AAAAMMDD(Date pData)	**pData**: Campo do tipo date.	Converte o formato original do campo data para o formato "AAAAMMDD".
UCase(String pTexto)	**pTexto**: Campo do tipo string.	Converte todos os caracteres de uma *string* para maiúsculo. Muitas linguagens de programação já possuem esta função embutida.
Trim(String pTexto)	**pTexto**: Campo do tipo string.	Remove os espaços em branco à direita e à esquerda da *string*. Muitas linguagens de programação já possuem esta função embutida.

O termo *Stopwords*, que aparece na descrição do método *removeStopwords( )*, é comum na Recuperação de Informações (*Information Retrieval*) [Baeza-Yates e Ribeiro-Neto 1999] e representa palavras com alta frequência numa coleção de documentos que não contribuem para a diferenciação de um documento do outro. Essas palavras são geralmente artigos, preposições, conjunções, pronomes, contrações e demais classes de palavras auxiliares [Martha et al. 2004] e [Borsato et al. 2006], além de alguns advérbios e verbos de ligação. Especificamente neste trabalho, *Stopwords* são preposições comuns a nomes próprios da Língua Portuguesa: de, da, do, das, dos.

Ainda na etapa de limpeza e padronização do processo de *Record Linkage*, é gerada uma tabela intermediária com os campos a serem comparados de cada fonte de dados. Nessas tabelas, os campos já estão tratados, apresentando o resultado das formatações e dos algoritmos. No

[77] Em geral, campos textuais de formato livre: Nome, Endereço, etc.

experimento descrito como primeiro cenário do Estudo de Caso (Capítulo 6), o uso dessa tabela intermediária permitiu que o número de pares comparados por segundo crescesse em até sete vezes.

Outros recursos e estratégias para combinação de fontes de dados são utilizados neste trabalho. Após análise, é possível promover um par duvidoso em um par combinado. Isto é feito através do método *promoverDuvidoso( )* da classe *MPI*. Ao realizar esta operação, é armazenado um pequeno *log*, indicando quem fez e quando foi feito.

Na classe *MPI*, há o método *incrementarMPI( )*. Como as fontes de dados estão sendo atualizadas constantemente, este método permite que haja uma nova comparação entre os pares não-combinados, com o objetivo de encontrar novos combinados e, com isso, atualizar o MPI. Ao contrário de um IHE/PIX convencional, a solução proposta não permite uma atualização do índice único a cada atualização em uma das fontes de dados. Em outras palavras, a transação *PIX Update Notification* não foi implementada neste trabalho. A atualização do MPI ocorre de forma incremental e é disparada pelo usuário.

Em alguns casos, existe a necessidade de se retornar às fontes de dados originais, a partir do MPI gerado, com a finalidade de consultar todas as informações relativas a um mesmo paciente. Isso permite uma visão integrada dos dados que estão armazenados em diferentes fontes e pertencentes a um mesmo paciente. Essa função é realizada pelo método *consultarFDs( )*.

As classes *SGBD*, *TipoFD*, *Algoritmo*, *TipoBlocagem* e *Usuario* contam com os métodos clássicos de *Incluir( )*, *Alterar( )* e *Excluir( )* objetos. Sendo que as quatro primeiras classes citadas contam ainda com métodos que permitem listagem das respectivas descrições. Esses métodos são utilizados para alimentar elementos gráficos do tipo *listbox*.

A classe *PIX Manager* possui métodos para incluir e alterar objetos. Contudo, não possui o método *Excluir( )*. Uma vez criado um ambiente para combinação entre duas fontes de dados, ele não poderá mais ser removido. Este ambiente poderá somente ser desabilitado, através do método *Desabilitar( )*.

5.3. Projeto Lógico

Ainda com base nos requisitos para desenvolvimento de um IHE/PIX, foi desenvolvido um banco de dados, cujo diagrama de tabelas [Elmasri e Navathe 2005] é apresentado na Figura 36.

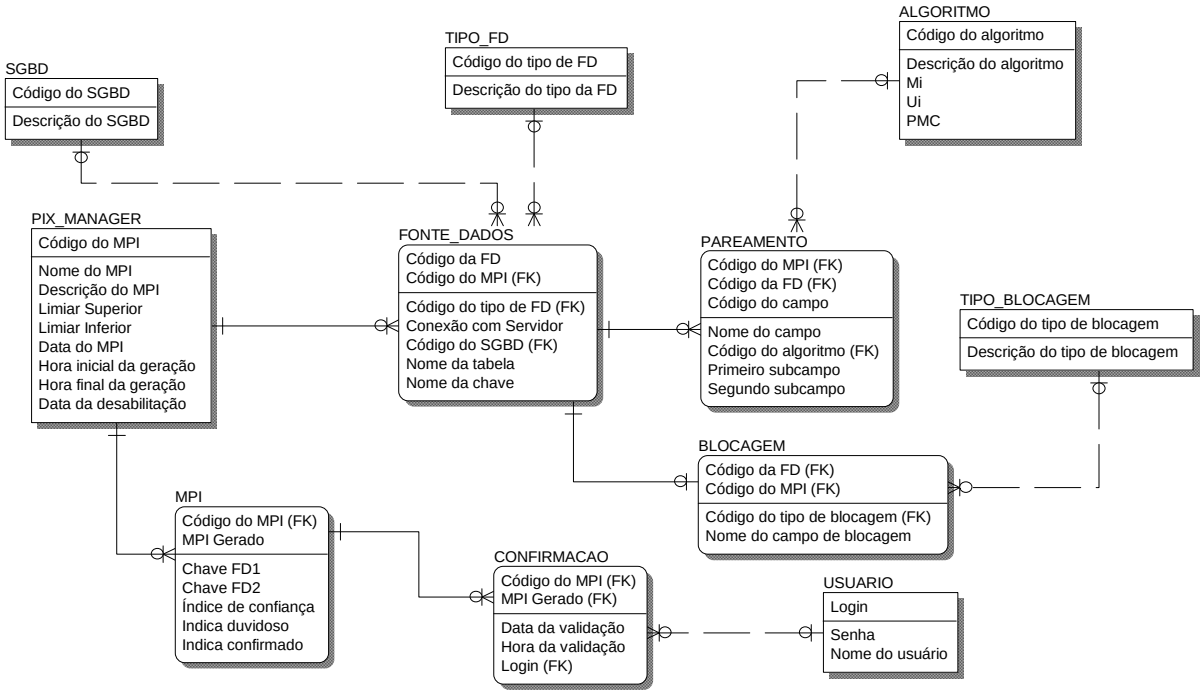

Figura 36: Diagrama de tabelas da solução implementada.

A Tabela 9 apresenta a descrição de cada uma das tabelas apresentadas na Figura 36.

Tabela 9: Descrição das tabelas.

Tabela	Descrição
ALGORITMO	Algoritmo utilizado para comparação de campos.
BLOCAGEM	Blocagem utilizada para comparação das duas fontes de dados.
CONFIRMACAO	Promoção de um par duvidoso em um par combinado.
FONTE_DADOS	Dados sobre cada fonte de dados.
MPI	MPI Gerado. Apresenta cada par combinado ou duvidoso com seu respectivo índice de confiança. Também é possível que todas as comparações sejam armazenadas.
PAREAMENTO	Parâmetros para comparação de campos no processo de *Record Linkage*.
PIX_MANAGER	Armazena os dados sobre cada combinação entre duas fontes de dados.
SGBD	Identificação do SGBD
TIPO_BLOCAGEM	Tipo de blocagem utilizada.
TIPO_FD	Tipo da fonte de dados a ser combinada.
USUARIO	Cadastro de usuários.

5.4. Projeto Físico

A partir do projeto lógico, foi desenvolvido um projeto físico cuja implementação foi feita na versão Express 2008 do Microsoft SQL Server[78]. O projeto físico pode ser observado na Figura 37 e o dicionário de dados na Tabela 10. Neste dicionário, é descrita cada tabela com seus respectivos

[78] http://www.microsoft.com/brasil/servidores/sql/default.mspx

atributos e indicação de campo chave. Há também, na Tabela 10, o nome físico e a descrição de cada campo. Os tipos e tamanhos de cada campo são apresentados no próprio diagrama, na Figura 37.

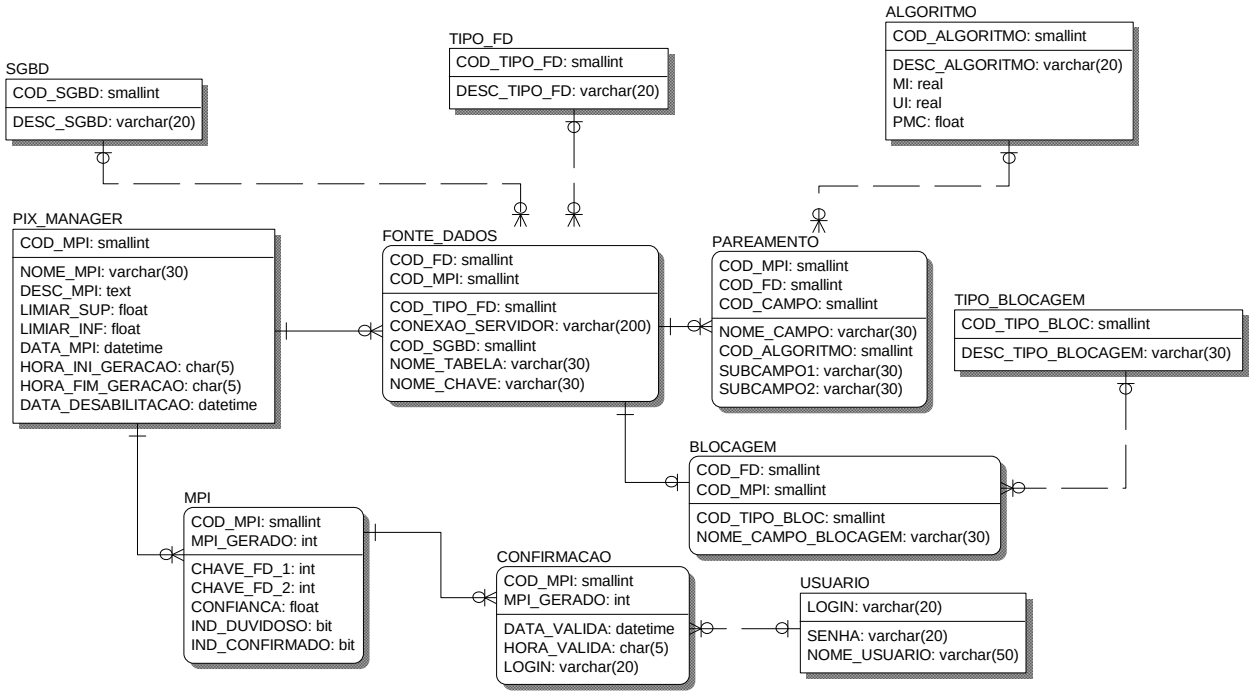

Figura 37: Projeto físico da solução implementada.

Tabela 10: Dicionário de dados.

Tabela	Atributo	Nome do campo	Descrição
ALGORITMO	Código do algoritmo	COD_ALGORITMO	Código do algoritmo utilizado para comparar campos.
	Descrição do algoritmo	DESC_ALGORITMO	Nome do algoritmo.
	Mi	MI	Sensibilidade atribuída ao algoritmo.
	Ui	UI	(1 - Especificidade) atribuída ao algoritmo.
	PMC	PMC	Proporção mínima de concordância do campo.
BLOCAGEM	Código da FD	COD_FD	Código de cada fonte de dados de um MPI.
	Código do MPI	COD_MPI	Código do MPI. Cada combinação entre duas fontes de dados.
	Código do tipo de blocagem	COD_TIPO_BLOC	Código do tipo de blocagem.
	Nome do campo de blocagem	NOME_CAMPO_BLOCAGEM	Nome do campo utilizado para fazer a blocagem.
CONFIRMACAO	Código do MPI	COD_MPI	Código do MPI. Cada combinação entre duas fontes de dados.
	MPI Gerado	MPI_GERADO	Código que representa o MPI gerado.
	Data da validação	DATA_VALIDA	Data da confirmação manual feita pelo usuário, promovendo um par duvidoso em combinado.
	Hora da validação	HORA_VALIDA	Hora da confirmação manual feita pelo usuário, promovendo um par duvidoso em combinado.
	Login	LOGIN	*Login* do usuário que fez a confirmação manual.

Tabela 10: Dicionário de dados - Continuação.

Tabela	Atributo	Nome do campo	Descrição
FONTE_DADOS	Código da FD	COD_FD	Código de cada fonte de dados de um MPI.
	Código do MPI	COD_MPI	Código do MPI. Cada combinação entre duas fontes de dados.
	Código do tipo de FD	COD_TIPO_FD	Código do tipo da fonte de dados.
	Conexão com Servidor	CONEXAO_SERVIDOR	*String* de conexão com o servidor de banco de dados.
	Código do SGBD	COD_SGBD	Código do SGBD.
	Nome da tabela	NOME_TABELA	Nome da tabela do banco de dados que irá participar do *Record Linkage*.
	Nome da chave	NOME_CHAVE	Nome da chave primária da tabela utilizada no *Record Linkage*.
MPI	Código do MPI	COD_MPI	Código do MPI. Cada combinação entre duas fontes de dados.
	MPI Gerado	MPI_GERADO	Código que representa o MPI gerado.
	Chave FD1	CHAVE_FD_1	Conteúdo da chave da primeira fonte de dados.
	Chave FD2	CHAVE_FD_2	Conteúdo da chave da segunda fonte de dados.
	Índice de confiança	CONFIANCA	Índice de confiança ou escore total do par combinado.
	Indica duvidoso	IND_DUVIDOSO	Indicação de que é um par duvidoso.
	Indica confirmado	IND_CONFIRMADO	Indicação de que é um par combinado.
PAREAMENTO	Código do MPI	COD_MPI	Código do MPI. Cada combinação entre duas fontes de dados.
	Código da FD	COD_FD	Código de cada fonte de dados de um MPI.
	Código do campo	COD_CAMPO	Código do campo que deve ser comparado.
	Nome do campo	NOME_CAMPO	Nome do campo que deve ser comparado.
	Código do algoritmo	COD_ALGORITMO	Código do algoritmo que será utilizado para comparar campos.
	Primeiro subcampo	SUBCAMPO1	Primeira subdivisão do campo que é comparado. Exemplo: primeiro nome de um indivíduo.
	Segundo subcampo	SUBCAMPO2	Segunda subdivisão do campo que é comparado. Exemplo: último nome de um indivíduo.
PIX_MANAGER	Código do MPI	COD_MPI	Código do MPI. Cada combinação entre duas fontes de dados.
	Nome do MPI	NOME_MPI	Nome do MPI.
	Descrição do MPI	DESC_MPI	Texto descrevendo cada combinação entre duas fontes de dados. Justificativa de porque integrar as duas fontes de dados.
	Limiar Superior	LIMIAR_SUP	Quando o escore total de um par de registros comparados for maior ou igual ao **Limiar Superior**, significa que é um par combinado.
	Limiar Inferior	LIMIAR_INF	Quando o escore total de um par de registros comparados for menor ou igual ao **Limiar Inferior**, significa que é um par não combinado. Quando o escore total estiver entre os limiares, será um par duvidoso.
	Data do MPI	DATA_MPI	Data da geração do MPI.
	Hora inicial da geração	HORA_INI_GERACAO	Hora do início de geração do MPI.
	Hora final da geração	HORA_FIM_GERACAO	Hora do término de geração do MPI.
	Data da desabilitação	DATA_DESABILITACAO	Data de desabilitação do MPI.
SGBD	Código do SGBD	COD_SGBD	Código do SGBD.
	Descrição do SGBD	DESC_SGBD	Nome do SGBD.
TIPO_BLOCAGEM	Código do tipo de blocagem	COD_TIPO_BLOC	Código do tipo de blocagem.
	Descrição do tipo de blocagem	DESC_TIPO_BLOCAGEM	Descrição do tipo de blocagem.
TIPO_FD	Código do tipo de FD	COD_TIPO_FD	Código do tipo da fonte de dados.
	Descrição do tipo da FD	DESC_TIPO_FD	Descrição do tipo da fonte de dados.
USUARIO	Login	LOGIN	*Login* do usuário que fez a confirmação manual.
	Senha	SENHA	Senha do usuário criptografada.
	Nome do usuário	NOME_USUARIO	Nome do usuário do sistema.

5.5. Implementação

A partir das especificações descritas anteriormente, foi desenvolvido um protótipo que apresenta as seguintes funcionalidades:

- Páginas que permitem testar as funções *Soundex* e distância de *Levenshtein*;
- *Record Linkage* propriamente dito, com a respectiva geração do índice único (MPI);
- Promoção de pares duvidosos em pares combinados, a partir de intervenção humana;
- Carga incremental do MPI;
- Consulta às fontes de dados originais a partir do MPI gerado.

5.5.1. Interface

As funcionalidades desenvolvidas são páginas web, dispostas no diagrama hierárquico de funções (DHF) [Martin e McClure 1991], apresentado na Figura 38. O menu principal (Figura 39) leva a cada uma das funcionalidades do protótipo.

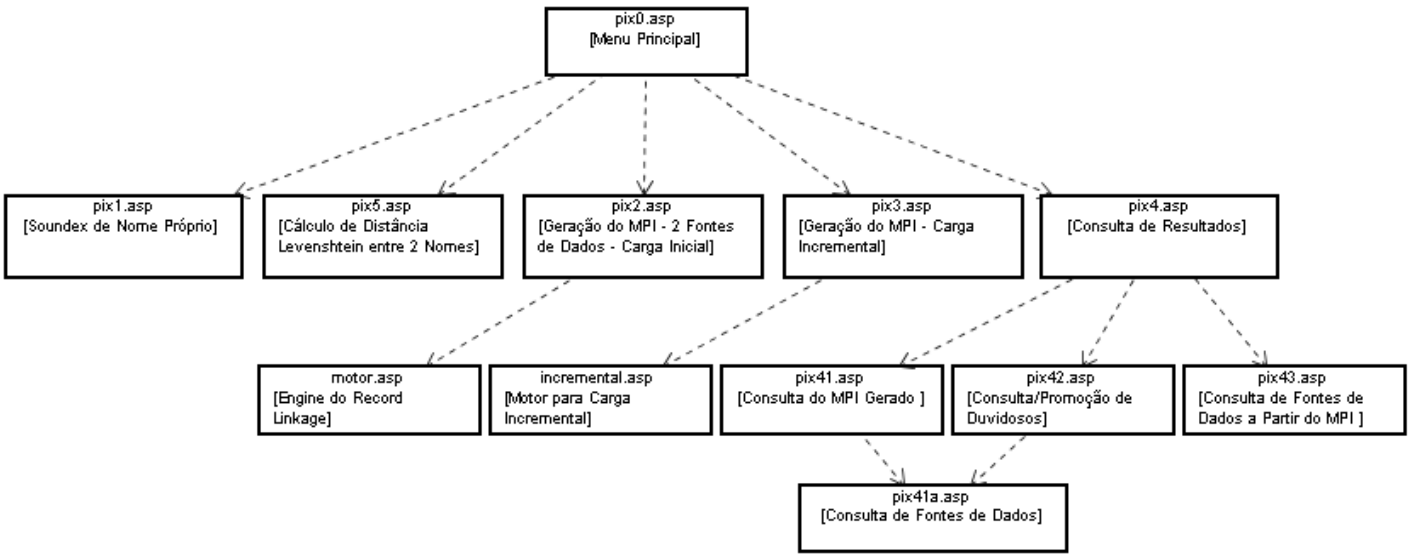

Figura 38: DHF do protótipo desenvolvido.

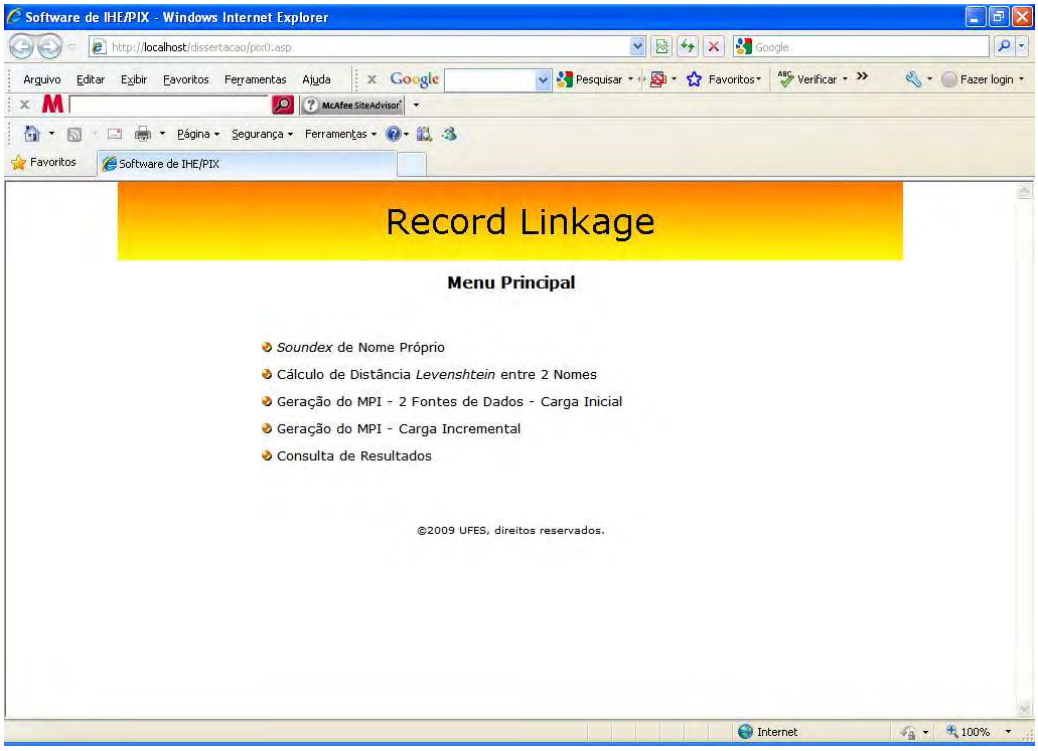

Figura 39: Menu principal do protótipo.

A função que permite promover pares duvidosos em combinados está agrupada no menu "Consulta de Resultados" (Figura 40), por se tratar de um dos resultados obtidos com o processo de *Record Linkage*.

Figura 40: Menu "Consulta de Resultados".

A Figura 41 mostra uma página que permite testar a função *Soundex*. Nessa página, o usuário informa qualquer nome próprio, que é trabalhado, sofrendo as seguintes ações: (*i*) todos os caracteres são convertidos para maiúsculo; (*ii*) são retirados os acentos; (*iii*) são retirados caracteres que não são letras; (*iv*) são eliminados caracteres em branco à esquerda e à direita do nome informado e (*v*) são removidos os *Stopwords*. Por fim, é aplicado o algoritmo *Soundex* no texto resultante.

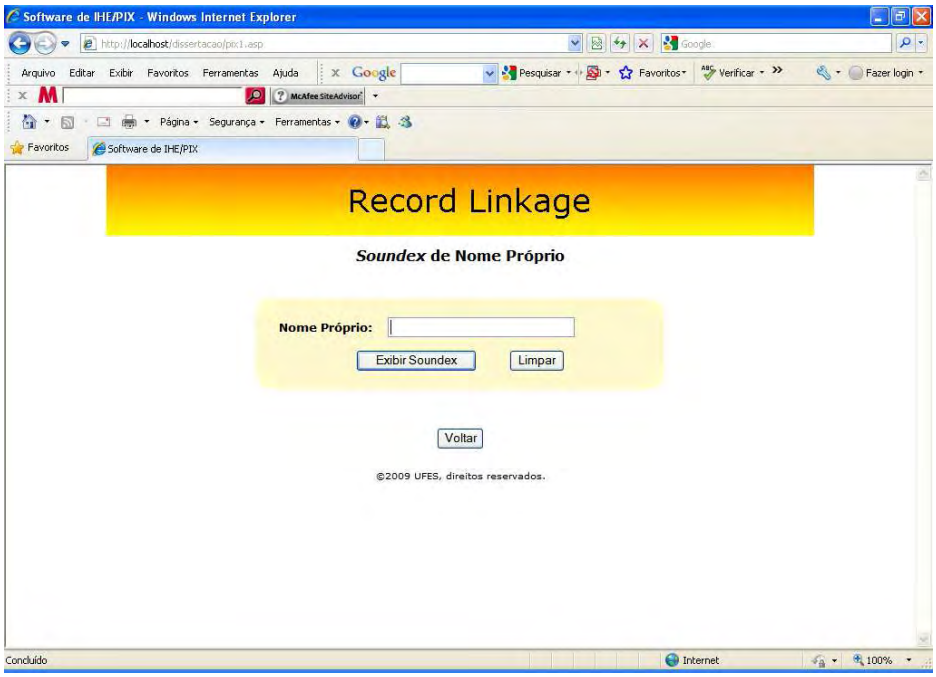

Figura 41: Página com a função *Soundex*.

A Figura 42 apresenta uma página que calcula a distância de *Levenshtein* entre dois nomes informados. O usuário pode informar tanto nomes quanto sequências numéricas, como datas. O resultado é apresentado no formato da expressão (2), vista na Subseção 4.4.2 deste trabalho. Quando os dois nomes informados são absolutamente iguais o *Fator* é igual a 1, quando são completamente diferentes o *Fator* é igual a 0, e quando há semelhança entre dois nomes não idênticos, o *Fator* apresentado é um número maior que zero e menor que 1.

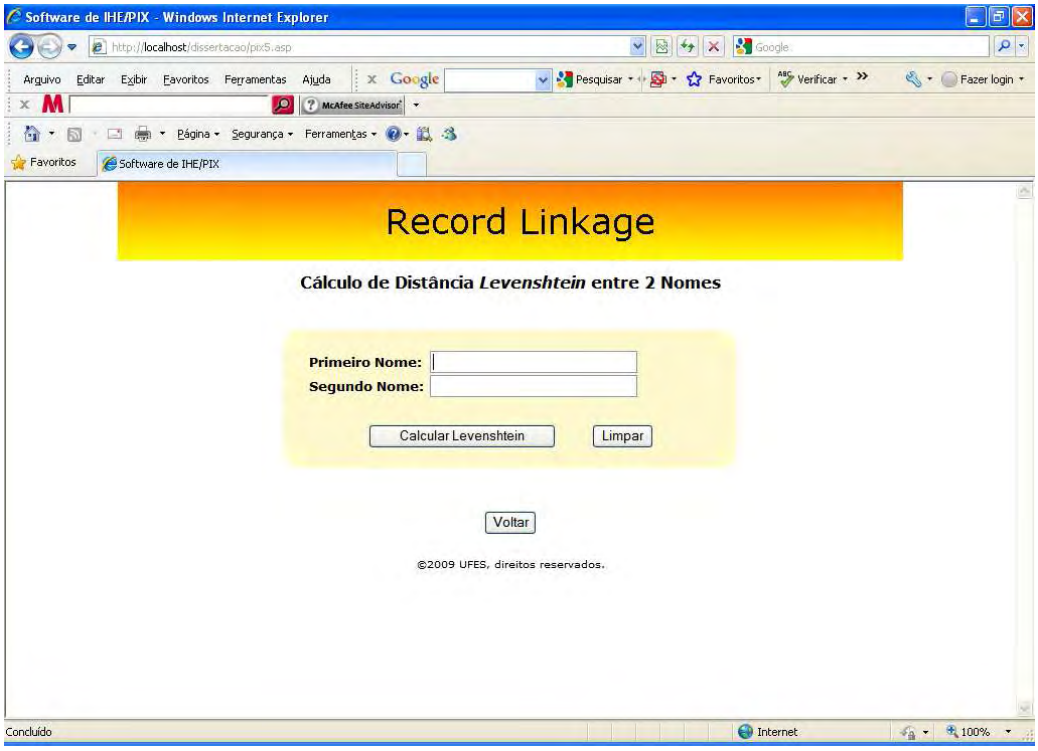

Figura 42: Página com a função *Levenshtein*.

O objetivo principal do protótipo gerado corresponde a execução do método de *Record Linkage* e a geração de um índice único entre duas fontes de dados heterogêneas e distribuídas. Para isso, é necessário que o software desenvolvido seja o mais configurável e flexível possível.

A Figura 43 mostra um exemplo de configuração de duas fontes de dados que são comparadas. Nela, é mostrada a configuração utilizada no experimento descrito no primeiro cenário do Capítulo 6. Inicialmente, o usuário informa um nome para o MPI, completando com uma breve descrição. Em seguida, o usuário informa os limiares superior e inferior. Para cada fonte de dados, é informado o tipo de fonte de dados, a *string* de conexão e a identificação do banco de dados ou esquema XML. Também são informadas a tabela (no caso de banco de dados) ou *tag* (no caso de XML), chave e blocagem. Finalmente, são informados dados sobre o pareamento entre os campos, indicando, para cada campo, seu nome, o algoritmo utilizado e sua subdivisão, se houver.

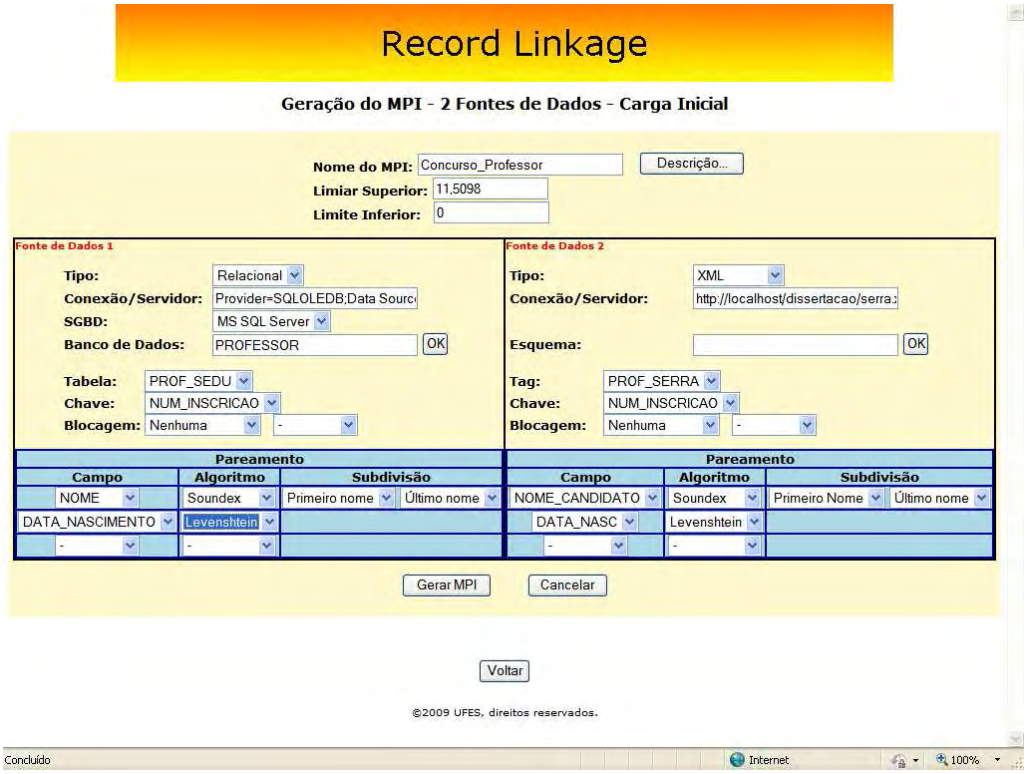

Figura 43: Exemplo de configuração de duas fontes de dados.

Outro recurso do protótipo é a carga incremental de um MPI, cuja interface para execução é apresentada na Figura 44.

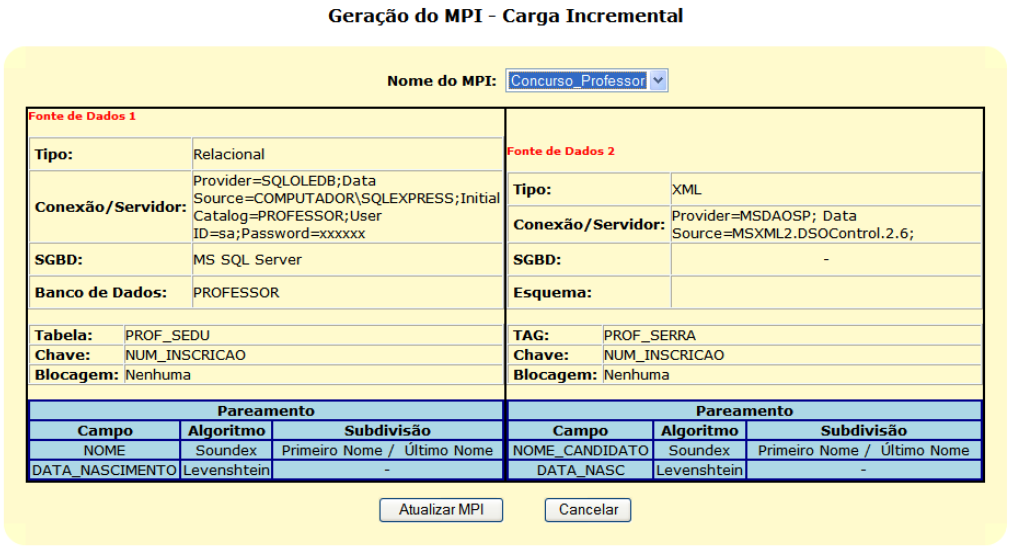

Figura 44: Interface para carga incremental do MPI.

Outra funcionalidade desenvolvida é a possibilidade de promoção de um par duvidoso em combinado. O usuário pode consultar os pares duvidosos entre duas fontes de dados combinadas, ou seja, aqueles que estão entre os limiares superior e inferior, conforme mostrado na Figura 45.

Figura 45: Relação de pares duvidosos.

Após análise de cada um dos pares, inclusive com a consulta aos dados das fontes originais, o usuário decide se um ou mais pares duvidosos podem ser promovidos a combinados. Para isso, ele seleciona os pares que desejar, clicando na caixa de verificação (*checkbox*), localizada à direita da lista, e pressiona o botão “Atualizar MPI”. Com esse procedimento, os pares duvidosos selecionados passarão a ser considerados combinados. São registrados para cada par promovido o login de quem fez a operação, e a data e hora de sua realização.

Não houve preocupação com implementação de procedimentos de segurança da informação no desenvolvimento deste protótipo. Essa tarefa deve fazer parte da relação de trabalhos futuros.

Por fim, é possível, com o protótipo gerado, consultar as fontes de dados originais. Este procedimento pode ser feito de três maneiras diferentes.

A primeira maneira é a partir da consulta à relação dos pares combinados de um determinado MPI (Figura 46).

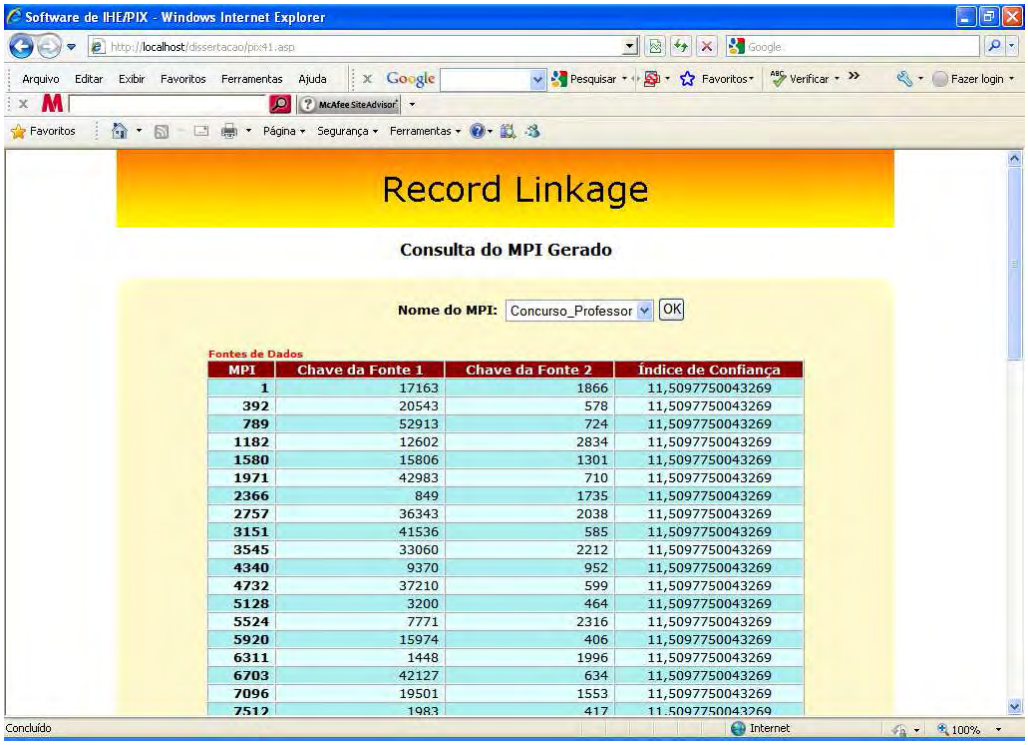

Figura 46: Consulta aos pares combinados de um MPI.

Cada código de MPI gerado é um link que permite a consulta às fontes de dados originais. Os campos de cada fonte de dados são exibidos, com seus respectivos conteúdos. Os dados de cada fonte são dispostos lado a lado, como visto na Figura 47.

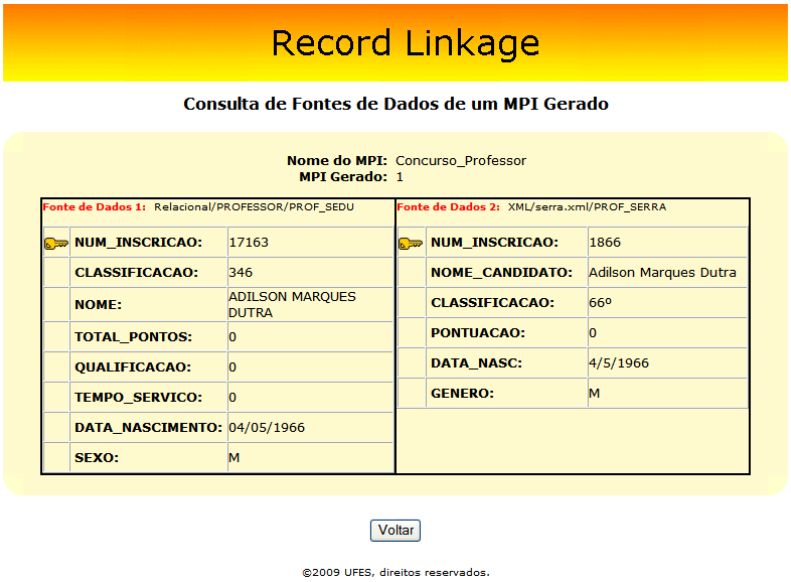

Figura 47: Consultas às fontes de dados originais de um MPI gerado.

Outra maneira de consultar as fontes originais é a partir da consulta/promoção de pares duvidosos (Figura 45). Da mesma forma que na consulta aos pares combinados, o código dos MPIs

gerados para os pares duvidosos também são links que possibilitam consultas às fontes de dados originais.

Uma terceira forma de consulta é a escolha de um MPI (par de fontes de dados), juntamente com um código de MPI gerado conhecido (Figura 48). A exemplo das duas maneiras anteriores, os dados originais daquele par combinado é exibido.

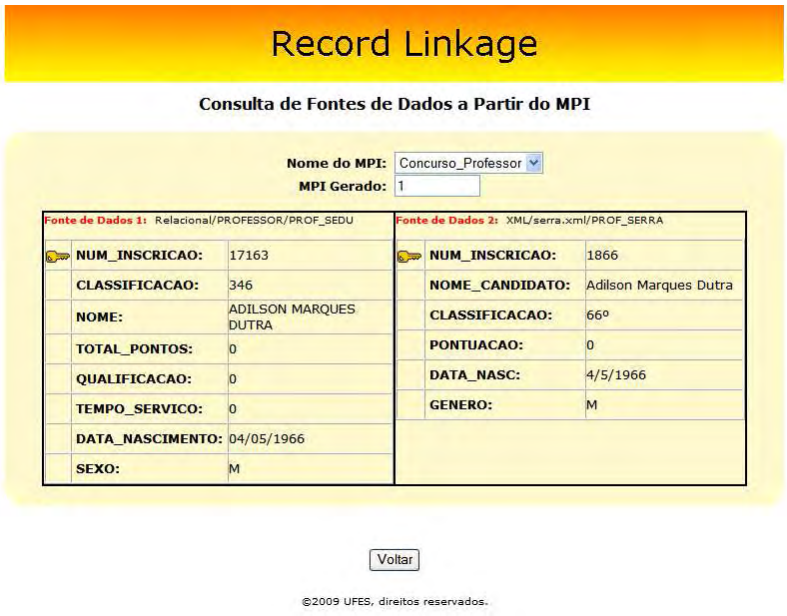

Figura 48: Consultas às fontes de dados originais a partir de um código conhecido.

5.5.2. Ambiente

O protótipo foi desenvolvido em ambiente web, na linguagem de programação ASP (*Active Server Pages*)[79], em servidores IIS (*Internet Information Services*)[80], em Sistemas Operacionais Windows (XP e Server 2003). O SGBD utilizado foi o MS-SQL Server. Para integrar fontes de dados no formato XML, foi utilizado o componente XMLDOM da Microsoft. Artefatos próprios da modelagem UML foram produzidos com a utilização do software JUDE[81] Community 5.5. A modelagem de dados foi feita na ferramenta CASE CA ERwin® Data Modeler.

[79] http://msdn.microsoft.com/en-us/library/aa286483.aspx

[80] http://www.iis.net/

[81] http://jude.change-vision.com/jude-web/index.html

Os testes foram realizados utilizando recursos computacionais do Prodest (Instituto de Tecnologia da Informação e Comunicação do Estado do Espírito Santo)[82], o DataCenter do Governo do Estado. A razão da escolha desse ambiente vem da necessidade de realização de testes com cenários de combinação de fontes de dados de grande volume e de alta confidencialidade. Algumas bases pertencentes ao Governo do Estado do Espírito Santo que comporão tais cenários não podem ser extraídas ou replicadas para fora desse DataCenter.

Os resultados obtidos com esses testes são descritos em três cenários, detalhados a seguir.

[82] http://www.prodest.es.gov.br/

6. Estudo de Caso

Com o objetivo de consolidar os conceitos estudados e validar a eficácia do protótipo desenvolvido, foram realizados experimentos com utilização de três cenários.

O objetivo principal do primeiro cenário é testar o protótipo desenvolvido, já que ele possui um número relativamente pequeno de combinações: cerca de 30 mil. Este cenário utiliza resultados de dois concursos públicos que possuem requisitos para admissão de professores bastante semelhantes, havendo, portanto, muita chance de encontrar um grande número de candidatos coincidentes. Com poucas combinações, é possível executar as funcionalidades do protótipo muitas vezes, o que facilita a depuração de erros. Além disso, o pouco volume de dados possibilita também a conferência manual dos pares combinados e a identificação de pares verdadeiros, permitindo identificar falsos positivos e falsos negativos.

O segundo cenário tem por objetivo verificar a eficácia do processo de *Record Linkage* desenvolvido. Assim, indivíduos coincidentes são procurados em duas bases de dados governamentais que possuem identificação unívoca: o CPF. Com isso, é possível comparar o resultado obtido com a junção entre as tabelas das duas bases de dados com a aplicação do método probabilístico. Com a identificação de falsos positivos e falsos negativos através de métodos computacionais, é possível calcular com precisão a eficácia do processo.

O objetivo do terceiro cenário é identificar pacientes coincidentes entre uma relação de pessoas com catarata com uma relação de indivíduos que sofrem de glaucoma. Este cenário, portanto, combina duas fontes de dados do domínio saúde, foco deste trabalho.

6.1. Primeiro Cenário

Para este cenário, são utilizados resultados de dois processos seletivos para contratação de professores de designação temporária da disciplina de História, disponíveis na Web. Devido à semelhança dos requisitos para admissão dos professores e pelo fato da contratação se dar para um mesmo município, havia muita chance de encontrar o mesmo indivíduo nas duas listas.

A primeira lista é referente ao "*Resultado do Processo Seletivo Simplificado do Magistério - Edital Nº 0045/2008*", para o cargo de professor de História, promovido pelo município da Serra/ES, que está disponível em

http://app.serra.es.gov.br/concursos/downloads/resultado_historia_045_2008.pdf. A lista com o resultado é composta pelos seguintes campos: número de ordem, número de inscrição, nome do candidato, classificação, pontuação e data de nascimento. Esta lista foi armazenada em um arquivo XML, com 79 ocorrências. A Figura 49 mostra como o documento **serra.xml** está estruturado.

Figura 49: Documento XML com o *"Resultado do Processo Seletivo Simplificado do Magistério - Edital Nº 0045/2008"*.

A segunda lista se refere ao "*Processo Seletivo de Professor para Designação Temporária - 2008/2009*", para o cargo de professor de História no município da Serra/ES, promovido pelo Governo do Estado do Espírito Santo, disponível no site: http://www.sedu.es.gov.br/download/dt_sedu/70_2.pdf. A lista com o resultado é composta pelos seguintes campos: ordem de classificação, nome, inscrição, total de pontos, pontos da qualificação, pontos para tempo de serviço, data de nascimento. Essa lista foi armazenada em uma tabela no ambiente MS-SQL Server 2008 Express, contendo 390 linhas. A Figura 50 descreve a tabela PROF_SEDU projetada em um banco de dados MS-SQL Server.

Nome da Coluna	Tipo de Dados	Permitir Nulos
CLASSIFICACAO	smallint	☑
NOME	nvarchar(50)	☑
NUM_INSCRICAO	int	☑
TOTAL_PONTOS	real	☑
QUALIFICACAO	real	☑
TEMPO_SERVICO	real	☑
DATA_NASCIMENTO	datetime	☑
SEXO	nvarchar(1)	☑

Figura 50: Tabela em MS-SQL Server do "*Processo Seletivo de Professor para Designação Temporária - 2008/2009*".

O objetivo de usar fontes de dados distintas é, exatamente, testar a flexibilidade do protótipo desenvolvido. Neste cenário, mesmo com a mudança na forma de acesso às fontes de dados originais, não houve prejuízo à execução do processo de *Record Linkage*.

Para identificar o mesmo indivíduo nas duas fontes de dados, foram comparados os seguintes conteúdos dos candidatos: primeiro nome, último nome e data de nascimento. Devido ao pouco universo de campos comparados, só foram considerados pares combinados aqueles que obtiveram índice de confiança máximo. Foram considerados duvidosos aqueles pares que obtiveram índice de confiança positivo. Os pares que apresentaram índice de confiança negativo foram considerados não combinados. Não foi utilizada blocagem, devido ao baixo número de comparações: 30.810.

A Tabela 11 apresenta os parâmetros de configuração utilizados na comparação de campos. Os valores de m_i, u_i e PMC são baseados no estudo de Camargo e Coeli (2000), conforme descrito na Tabela 5, Capítulo 4.

Tabela 11: Parâmetros de configuração para comparação de campos no 1º Cenário.

Tipo de Campo	Algoritmo	Sensibilidade (m_i)	1 – Especificidade (u_i)	PMC*
Primeiro Nome	*Soundex*	90%	5%	-
Último Nome	*Soundex*	90%	5%	-
Data de Nascimento	Levenshtein	90%	10%	75%

*PMC: Proporção Mínima de Concordância

Com base nos parâmetros da Tabela 11 e com a utilização da expressão 5 (Seção 4.5), o maior índice de confiança possível é 11,5098 e o menor é -9,6658.

Após a execução do *Record Linkage*, foi obtido o resultado apresentado na Tabela 12.

Tabela 12: Resultado do *Record Linkage* do 1º Cenário.

Índice de Confiança	Número de ocorrências
11,5098	66
5,1699	1
4,0919	2
3,2994	21
-2,2479	585
-3,3259	1
-3,7222	83
-4,1184	865
-9,6658	29186

Pode-se observar na Tabela 12 que foram encontrados 66 pares combinados (índice de confiança máximo), 24 pares duvidosos (índices de confiança positivos abaixo do índice máximo) e 30.720 pares não combinados (índice de confiança negativo).

Como forma de validação do método, foi realizada conferência manual, comparando as 79 ocorrências da primeira fonte de dados com as 390 linhas da segunda, chegando-se às seguintes conclusões:

- Não foram encontrados falsos positivos: os 66 pares combinados obtidos com a aplicação do método são verdadeiros;
- O par combinado com pontuação 5,1699 se refere a duas pessoas diferentes que possuem o mesmo *soundex* para o primeiro nome: *Luciana* e *Luciano*. Esses indivíduos também possuem o mesmo último nome: *Oliveira*. O uso do campo sexo contribuiria para baixar em muito o índice de confiança deste par;
- Foram encontrados dois falsos negativos, com índice de confiança de 4,0919. Esses pares são de duas professoras que omitiram o último nome em uma das duas fontes de dados. Provavelmente, o nome de casada. A ocorrência desses dois pares falsos negativos contribuiu para que o *Recall* ficasse na casa dos 97%. Caso esses dois falsos negativos sejam promovidos a pares combinados, o número de pares verdadeiros subirá para 68, ou seja, para o número correto de professores comuns contidos nas duas fontes de dados.
- Não foi encontrada coincidência alguma nos demais pares.

Aplicando as fórmulas de *Precision*, *Recall* e da medida de desempenho geral *F*, citadas na Seção 4.6, foram obtidos os resultados a seguir:

Recall = 97,05% *Precision* = 100% *F* = 98,5%

Foi gerado um índice único contendo, inicialmente, os 66 pares combinados. Foram acrescidos, posteriormente, os 2 falsos negativos, com a utilização da funcionalidade “Consulta/Promoção de Pares Duvidosos”, totalizando os esperados 68 pares.

6.2. Segundo Cenário

O segundo cenário consiste na busca do mesmo indivíduo em duas bases de dados governamentais de grande volume: (*i*) cadastro de servidores do Poder Executivo do Estado do Espírito Santo e (*ii*) cadastro do condutores habilitados do Estado do Espírito Santo. O objetivo de combinar essas duas bases de dados é a medição mais precisa da eficácia do método utilizado. O número de pares combinados encontrados é comparado com o relacionamento entre as duas tabelas dos dois bancos de dados. Como há um relacionamento unívoco entre estas duas tabelas, com a utilização do CPF, o *Precision* e o *Recall* do algoritmo podem ser medidas com exatidão.

Considerando os recursos computacionais necessários, em especial, o tempo de processamento, não foram utilizadas as duas bases de dados em sua totalidade. Mesmo com o uso de blocagem, o processamento da combinação integral entre as duas bases geraria um processamento na casa de dezenas de bilhões de pares comparados. A Tabela 13 mostra o volume de registros em função da primeira letra do nome do indivíduo em cada uma das fontes de dados escolhidas.

Tabela 13: Número de registros em função da primeira letra do nome.

Primeira letra	Sistema de Recursos Humanos	Sistema de Habilitação	Número de Combinações
A	18.321	115.704	2.119.812.984
B	1.794	11.262	20.204.028
C	9.161	49.997	458.022.517
D	6.592	37.024	244.062.208
E	11.774	66.843	787.009.482
F	5.051	37.295	188.377.045
G	5.829	40.942	238.650.918
H	2.422	14.439	34.971.258
I	4.480	18.441	82.615.680
J	13.588	120.545	1.637.965.460
K	1.983	7.326	14.527.458
L	11.597	57.382	665.459.054
M	26.828	89.978	2.413.929.784
N	4.583	18.417	84.405.111
O	1.763	12.252	21.600.276
P	3.655	27.056	98.889.680
Q	69	240	16.560
R	11.831	69.948	827.554.788
S	8.384	39.363	330.019.392
T	3.380	14.934	50.476.920
U	285	2.252	641.820
V	5.065	31.785	160.991.025
W	2.871	27.719	79.581.249
X	10	75	750
Y	263	701	184.363
Z	1.324	3.357	4.444.668
Total	**162.903**	**915.277**	**10.564.414.478**

Em função da análise dos dados anteriores, foi escolhida a letra “B” como primeira letra do nome do indivíduo a ser comparado. A razão dessa escolha vem do número de comparações, que ficou na casa de 20 milhões de registros. Este número representa um meio termo entre um número de comparações inexpressivo, como é o caso da letra “X” e um número de comparações que demandaria recursos computacionais e tempo bastante preciosos, como é o caso da letra “M”, que exige 2,4 bilhões de comparações.

A primeira fonte de dados é a relação de servidores do Poder Executivo do Estado do Espírito Santo que começam com a letra “B”. Essa fonte de dados é proveniente do Sistema Integrado de Administração de Recursos Humanos do Estado do Espírito Santo (SIARHES) e seu acesso foi obtido com a autorização da Secretaria de Estado de Gestão e Recursos Humanos (SEGER)[83]. A base de dados do SIARHES está em um SGBD Oracle 10g. Mas, por razões de segurança e desempenho, não pôde haver um acesso direto a essa base. Foi feita uma replicação de dados para um banco de dados em MS-SQL Server 2000 em um servidor do Prodest. Essa tabela possui 1.794 registros, foi gerada em 7 de julho de 2009 e conta com os campos descritos na Tabela 14.

Tabela 14: Campos da primeira fonte de dados do 2º Cenário.

Nome do campo	Descrição	Tipo de dados
NUMFUNC	Número funcional	numeric(8,0)
NOME	Nome do servidor	varchar(60)
SEXO	Sexo do servidor	char(1)
DTNASC	Data de nascimento do servidor	datetime
MAE	Nome da mãe do servidor	varchar(60)
CPF	CPF do servidor	numeric(11,0)

Os dados dessa fonte possuem uma qualidade bastante razoável. Há 4 CPFs nulos, 41 CPFs com conteúdo igual a “11111111111” (onze algarismos um), e 2 indivíduos com CPFs iguais. Existem ainda 55 registros com o nome da mãe nulo e 2 registros com o nome da mãe igual a “MAE”. Os CPFs nulos ou com conteúdo formado apenas por algarismos “1” não fizeram parte do relacionamento unívoco entre as duas fontes de dados, uma vez que a segunda fonte de dados não possui CPFs repetidos ou nulos. Para o nome da mãe nulo ou com conteúdo igual a “MAE” foi aplicado o **fator de ponderação de discordância** (expressão 4, Seção 4.5).

A segunda fonte de dados é a relação de condutores (motoristas habilitados) do Estado do Espírito Santo que começam com a letra “B”. Essa fonte de dados é proveniente do Sistema de Habilitação e seu acesso foi autorizado pelo Departamento Estadual de Trânsito (DETRAN)[84]. A

[83] http://www.seger.es.gov.br/

[84] http://www.detran.es.gov.br/

base de dados do Sistema de Habilitação está em um cluster de servidores MS-SQL Server 2000. Pelas mesmas razões expostas em relação ao SIARHES, os dados foram replicados para o mesmo banco de dados criado para hospedar o conteúdo da primeira fonte de dados. A tabela criada possui 11.262 registros, foi gerada em 7 de julho de 2009 e conta com os campos descritos na Tabela 15.

Tabela 15: Campos da segunda fonte de dados do 2º Cenário.

Nome do campo	Descrição	Tipo de dados
NumRegistroCNH	Número da CNH	numeric(11,0)
Nome_Condutor	Nome do condutor	varchar(60)
Cod_Sexo	Sexo do condutor	char(1)
Data_Nascimento	Data de nascimento do condutor	datetime
Nom_Mae	Nome da mãe do condutor	varchar(60)
Num_CPF	CPF do condutor	numeric(11,0)

A qualidade dos dados desta fonte é excelente. Não há a presença de CPFs nulos ou duplicados. Existem 35 registros cujo nome da mãe está nulo. A exemplo da primeira fonte de dados, também foi aplicado o **fator de ponderação de discordância** nestes casos.

Para identificar o mesmo indivíduo nas duas fontes de dados, foram comparados os seguintes conteúdos: Primeiro nome, último nome, data de nascimento, sexo, primeiro nome da mãe e último nome da mãe. A Tabela 16 apresenta os parâmetros de configuração utilizados na comparação de campos deste cenário.

Tabela 16: Parâmetros de configuração para comparação de campos no 2º Cenário.

Tipo de Campo	Algoritmo	Sensibilidade (m_i)	1 – Especificidade (u_i)	PMC*
Primeiro Nome	*Soundex*	90%	5%	-
Último Nome	*Soundex*	90%	5%	-
Data de Nascimento	Levenshtein	90%	10%	75%
Sexo	Comparação exata	95%	50%	-
Primeiro nome da mãe	*Soundex*	90%	5%	-
Último nome da mãe	*Soundex*	90%	5%	-

*PMC: Proporção Mínima de Concordância

Para este cenário, não foi utilizada blocagem, uma vez que a massa de teste já é um subconjunto dos dados das duas fontes (bloco).

Foi então aplicado o relacionamento puro e simples entre as duas tabelas. Aplicou-se uma junção com utilização da cláusula “where” (produto cartesiano), como pode ser observado na Figura 51.

```
SELECT NUMFUNC, NumRegistroCNH,NOME, Nome_condutor, CPF
FROM   FUNCIONARIOS, CONDUTOR
WHERE  FUNCIONARIOS.CPF = CONDUTOR.Num_CPF
ORDER  BY NOME
```

Figura 51: Instrução SQL para relacionamento das duas tabelas.

O resultado foi que 931 condutores e servidores possuem o mesmo CPF. Esses indivíduos foram considerados os **pares verdadeiros positivos**, que é um dado fundamental para medição da eficácia do algoritmo.

Por razões de desempenho e disponibilidade de recursos, o processo de *Record Linkage* foi realizado em três noites, seguindo o planejamento da Tabela 17.

Tabela 17: Planejamento para execução do *Record Linkage* no 2º Cenário.

Blocos	Nº Funcionários	Nº Condutor	Nº Combinações
NOME < 'BE'	142	11.262	1.599.204
NOME > 'BE' e NOME < 'BI'	750	11.262	8.446.500
NOME > 'BI'	902	11.262	10.158.324
Total	**1.794**	**11.262**	**20.204.028**

Com base nos parâmetros da Tabela 16 e com a utilização da expressão 5 (Seção 4.5), o maior índice de confiança possível neste cenário é 20,7756 e o menor é -19,4835. Na tabela MPI, só foram gravados índices de confiança positivos.

Após a execução do *Record Linkage* neste cenário, foi obtido o resultado apresentado na Tabela 18.

Tabela 18: Resultado do *Record Linkage* do 2º Cenário.

Índice de Confiança	Número de ocorrências
20,7756	503
20,3794	6
19,9831	189
16,5277	11
16,1315	2
15,7352	31
14,4358	521
13,3578	71
12,9615	68
12,5653	627
10,1878	267
9,1098	5
8,7136	27
8,3174	323
7,0179	14.357
5,9399	27
5,5437	477
5,1474	4.824
2,7700	7.782
1,6920	13
1,2958	15
1,2958	310
0,8995	182
0,8995	1
0,8995	3.374
Índices Negativos	**20.170.015**

A partir desses resultados e analisando o conteúdo dos dados, conclui-se que:

1. Foram considerados pares combinados aqueles que obtiveram índices de confiança maiores ou iguais a 15,7352 (6 primeiras linhas em destaque na Tabela 18).
2. Os pares que obtiveram escore total menor ou igual a 14,4358 foram considerados não combinados, por possuírem datas de nascimento diferentes em mais de 25% do Fator de Distância de *Levenshtein* (expressão 2, Subseção 4.4.2).
3. 503 pares combinados obtiveram índice de confiança máximo (20,7756), ou seja, todos os campos comparados estavam iguais.
4. 6 pares combinados obtiveram escore total igual a 20,3794. Sua diferença para o índice de confiança máximo foi um dígito diferente na data de nascimento.
5. 189 pares combinados obtiveram escore total igual a 19,9831. Sua diferença para o índice de confiança máximo foi a inversão do mês com o dia na data de nascimento, respeitando-se o limite de 75% da semelhança entre as datas.
6. 11 pares combinados obtiveram escore total igual a 16,5277. A diferença desse escore para o índice de confiança máximo foi a troca de sexo.
7. 2 pares combinados obtiveram escore total igual a 16,1315. A diferença desse escore para o índice de confiança máximo foi a troca de sexo associada à diferença de um dígito na data de nascimento.
8. 31 pares combinados obtiveram escore total igual a 15,7352. A diferença desse escore para o índice de confiança máximo foi a troca de sexo associada à inversão do mês com o dia na data de nascimento, respeitando-se o limite de 75% da semelhança entre as datas.
9. A soma das 6 primeiras linhas da Tabela 18 é igual a 742.
10. Dos 742 pares considerados combinados pelo processo de *Record Linkage* adotado, 57 são falsos positivos. Ou seja, somente 685 são verdadeiramente pares combinados pelo critério de relacionamento das duas tabelas através do CPF. Portanto, obtivemos o seguinte *Precision*:

$$Precision = \frac{número\ de\ pares\ verdadeiros}{número\ de\ pares\ verdadeiros + número\ de\ falsos\ positivos}$$

$$Precision = 685 \div (685+57) \quad \Rightarrow \quad Precision = 92{,}31\%$$

11. Também foram encontrados 246 falsos negativos. Ou seja, 246 pares que estão combinados abaixo do escore total de 15,7352, mas que possuem CPFs iguais. Com isso, o *Recall* deste cenário ficou assim calculado:

$$Recall = \frac{número\ \ de\ \ pares\ verdadeiros}{número\ \ de\ \ pares\ verdadeiros\ +\ número\ de\ falsos\ negativos}$$

$$Recall = 685 \div (685+246) \quad \Rightarrow \quad Precision = 73{,}57\%$$

12. A medida de desempenho geral *F* ficou em 81,9%.

Há ainda algumas conclusões a serem tiradas a respeito da qualidade dos dados utilizados:

1. Existem 4 indivíduos que possuem o mesmo CPF nas duas fontes de dados que obtiveram índices de confiança negativos. (*i*) Em um dos casos, o nome da mãe no cadastro de condutores estava vazio; (*ii*) em outra situação, os últimos nomes tanto da mãe quanto do indivíduo haviam sido suprimidos; (*iii*) houve também um caso de ausência do nome da mãe no cadastro de servidores; (*iv*) por fim, a grafia de um sobrenome (do indivíduo e da mãe) estavam tão diferentes que geravam códigos *soundex* totalmente distintos em relação ao cadastro de condutores.
2. De forma inversa à situação anterior, houve 3 pares combinados que obtiveram índice de confiança máximo que não foram relacionados quando da junção dos CPFs. Estes casos foram: (*i*) CPF nulo no cadastro de funcionários; (*ii*) CPF inválido em condutor (faltava um dígito no meio); (*iii*) um mesmo condutor com 2 CPFs diferentes.

O desempenho inferior em relação ao primeiro cenário se dá pelo fato de haver algumas inconsistências nas duas fontes de dados deste cenário. Inclusive, este estudo é um aliado poderoso para tratar essas inconsistências nos respectivos cadastros.

6.3. Terceiro Cenário

O terceiro cenário é um caso típico e real da aplicação da identificação única de pacientes em fontes de dados distintas. Este cenário consiste na busca do mesmo indivíduo em dois cadastros diferentes, relacionados a duas doenças que afetam a visão: catarata e glaucoma.

Segundo Almeida (2006), o que há em comum entre a catarata e o glaucoma é a idade. Ambas são doenças relacionadas ao processo de envelhecimento. Assim sendo, é muito comum que essas doenças estejam presentes num mesmo paciente.

A catarata é a opacificação do cristalino. A sua causa mais frequente é o próprio envelhecimento: são as chamadas *Cataratas Senis*, que podem surgir já a partir dos 40 anos de idade.

De uma maneira simplificada, o glaucoma é uma elevação da pressão no interior do olho, o que acarreta atrofia do nervo óptico. A hipertensão ocular é nociva aos filamentos nervosos que formam o nervo óptico, que é o elemento condutor dos estímulos visuais para o cérebro. Com o tempo, o nervo óptico vai se atrofiando e comprometendo a visão, de maneira irreversível. Existem dezenas de tipos de glaucoma. O mais frequente é o *Glaucoma Crônico de Ângulo Aberto*. Esse tipo de glaucoma tem uma incidência de 1 a 2% na população em geral. Este índice aumenta com a idade, podendo atingir 6 a 7% após os 70 anos de idade.

Pela sua frequência, a catarata e o glaucoma são, ao lado da retinopatia diabética, as principais causas de cegueira no Brasil e no mundo. A incidência dessas doenças tende a crescer com o aumento da expectativa de vida da população.

Diante de um paciente com catarata e glaucoma com indicação cirúrgica, a maioria dos oftalmologistas (61,1%) prefere indicar a cirurgia combinada. Destes, 72% não usam incisão combinada, enquanto 28% utilizam esta técnica [Santhiago et al. 2009].

O Ministério da Saúde, através da Portaria nº 339 de 08 de Maio de 2002, determina que as secretarias de saúde dos estados, do Distrito Federal e dos municípios em Gestão Plena do Sistema Municipal enviem àquele órgão um cadastro dos procedimentos oftalmológicos executados, incluindo procedimentos envolvendo catarata e glaucoma.

Portanto, o objetivo deste cenário é a combinação dos cadastros referentes a essas duas doenças: glaucoma e catarata, mantidos pela Secretaria de Saúde do estado do Espírito Santo (SESA) [85]. Esses dados foram obtidos, mediante aprovação pelo Comitê de Ética em Pesquisa da SESA do Protocolo de Pesquisa 24/2009, em reunião ordinária realizada em 30 de junho de 2009.

[85] http://www.saude.es.gov.br/

A primeira fonte de dados deste cenário é o cadastro de pacientes com catarata, composto por 31.867 registros, cadastrados desde 19 de abril de 1995 até 2 de julho de 2009. Os campos que compõem este cadastro estão explicitados na Tabela 19, juntamente com os seus respectivos percentuais de utilização.

Tabela 19: Campos que compõem o cadastro de pacientes com catarata.

Nome do campo	Descrição	Quantidade	% de utilização
CODIGO_PAC	Código do Paciente	31.867	100
NOME_PAC	Nome do paciente	31.867	100
DTNAS_PAC	Data de nascimento do paciente	31.643	99,30
LOCAL_NASC	Local de nascimento do paciente	164	0,51
DTCAD_PAC	Data de cadastramento do paciente	31.867	100
SEXO_PAC	Sexo do paciente	421	1,32
ESCOLA_PAC	Escolaridade do paciente	31.863	99,99
EST_CIVIL	Estado civil do paciente	31.864	99,99
CONJUG_PAC	Nome do cônjuge do paciente	1	0,003
PAI_PAC	Nome do pai do paciente	1	0,003
MAE_PAC	Nome da mãe do paciente	204	0,64
ENDER_PAC	Endereço do paciente	24.386	76,52
COMPL_PAC	Complemento do endereço do paciente	729	2,29
BAIRRO_PAC	Bairro onde o paciente reside	23.004	72,19
CEP_PAC	CEP da residência do paciente	14.630	45,91
CIDADE_PAC	Cidade onde o paciente reside	30.870	96,87
ESTADO_PAC	Estado onde o paciente reside	31.435	98,64
TELEFO_PAC	Telefone do paciente	11.215	35,19
CIC_PAC	CPF do paciente	3.815	11,97
DOCUM_PAC	Documento do paciente	19	0,06
OBSERV_PAC	Observação sobre o paciente	12	0,04

Os campos nome do cônjuge, nome do pai, documento do paciente e observações sobre o paciente (com fundo em destaque mais claro), devido ao baixíssimo número de registros com este conteúdo, nitidamente não estão sendo utilizados. Local de nascimento, sexo e nome da mãe também são campos muito pouco preenchidos, não sendo úteis no processo de *Record Linkage*. Os campos escolaridade e estado civil, embora com um alto percentual de utilização, possuem problemas quanto ao seu conteúdo: 31.853 registros possuem escolaridade igual a zero (“0”) e 31.860 registros possuem estado civil igual a “S”. O CPF do paciente, que poderia ser utilizado como identificador único dessa fonte de dados, está preenchido em menos de 12% dos registros. Há ainda dados referentes à localização do paciente: endereço com complemento, bairro, CEP, cidade, estado e telefone (com fundo em destaque mais escuro), que não devem ser utilizados para comparação com a segunda fonte de dados, já que podem ter sofrido alterações com o passar do tempo.

A segunda fonte de dados deste cenário é o cadastro de pacientes com glaucoma, composto por 4.755 registros, cadastrados desde 9 de outubro de 2003 até 2 de julho de 2009. Os campos que compõem esta fonte de dados estão explicitados na Tabela 20, juntamente com os seus respectivos percentuais de utilização.

Tabela 20: Campos que compõem o cadastro de pacientes com glaucoma.

Nome do campo	Descrição	Quantidade	% de utilização
CODIGO_PAC	Código do Paciente	4.755	100
NOME_PAC	Nome do paciente	4.755	100
DTNAS_PAC	Data de nascimento do paciente	4.728	99,43
DTCAD_PAC	Data de cadastramento do paciente	4.755	100
SEXO_PAC	Sexo do paciente	43	0,90
ENDER_PAC	Endereço do paciente	4.653	97,85
COMPL_PAC	Complemento do endereço do paciente	173	3,64
BAIRRO_PAC	Bairro onde o paciente reside	4.475	94,11
CEP_PAC	CEP da residência do paciente	2.325	48,90
CIDADE_PAC	Cidade onde o paciente reside	4.741	99,71
ESTADO_PAC	Estado onde o paciente reside	4.754	99,98
TELEFO_PAC	Telefone do paciente	3.720	78,23
CIC_PAC	CPF do paciente	1.697	35,69
DOCUM_PAC	Documento do paciente	66	1,39
REQUERENTE	Dados sobre o requerente	4	0,08
DOCUM_REQ		1	0,02
TELEFO_REQ		1	0,02
OBSERV_PAC	Observação sobre o paciente	-	0,00

A exemplo da primeira fonte de dados deste cenário, aqui também há campos que não estão sendo utilizados, devido ao baixo percentual de utilização (fundo mais claro, em destaque na Tabela 20): documento do paciente, dados sobre o requerente e observações sobre o paciente. Nessa fonte, há também dados referentes a localização do paciente (fundo mais escuro, em destaque na Tabela 20), que não são comparados por serem mutáveis. O campo CPF é utilizado em mais de 35% dos registros, mas este percentual é insuficiente para identificação única do paciente. O campo sexo só está preenchido em 43 registros.

Em resumo, os únicos campos contidos nas duas fontes de dados que podem ser comparados são o nome do paciente e sua data de nascimento. Mesmo assim, há algumas inconsistências nestes campos em ambas as fontes. Foram, portanto, executados alguns procedimentos, visando à limpeza e padronização dos dados (Seção 4.2).

Na fonte de dados referente aos pacientes com catarata, dos 31.867 registros, havia 31.643 com datas de nascimento válidas. Somente esses registros foram considerados. Quanto ao campo nome, existem 6 registros com nomes começados com espaços em branco, que foram suprimidos. Havia ainda um nome começado com 0 (zero) no lugar da letra “O”. Essa correção foi feita manualmente.

Contudo, o grande problema na primeira fonte de dados é a **deduplicação** (registros com nomes e datas de nascimento idênticos). Foram encontradas 1.218 casos de duplicatas, 91 casos de triplicatras, 29 situações nas quais 4 registros apresentam nomes e datas de nascimento idênticos, 9 casos com 5 registros com nomes e datas de nascimento iguais, 3 casos com 6 registros com nomes e datas de nascimento iguais , 1 caso com 7 e 1 com 8. Foram removidas todas as deduplicações.

Foi considerado apenas o registro mais antigo de cada conjunto repetido de nome e data de nascimento. Ou seja, o registro com menor código do paciente. Ao final da etapa de limpeza e padronização dos dados, foi gerada uma tabela em um banco dados MS-SQL Server 2008 Express com 30.092 registros, cuja estrutura é apresentada na Tabela 21.

Tabela 21: Campos da primeira fonte de dados do 3º Cenário (pacientes com catarata).

Nome do campo	Descrição	Tipo de dados
CODIGO_PAC	Código do Paciente	int
NOME_PAC	Nome do paciente	nvarchar(32)
DTNAS_PAC	Data de nascimento do paciente	datetime

Na fonte de dados referente aos pacientes com glaucoma, dos 4.755 registros, existem 4.728 com datas de nascimento válidas. Somente esses registros foram considerados. Não existem nomes iniciados com espaço em branco ou 0 (zero). Foram encontrados 31 casos de deduplicação (nomes e datas de nascimento idênticos). Só foram encontradas duplicatas. Não foram encontradas triplicatas ou deduplicações com grau maior. Essas deduplicações foram removidas, considerando apenas o registro com menor código do paciente, semelhante ao que foi feito com a primeira fonte de dados. Ao final dessa etapa de limpeza e padronização de dados, foi gerada uma tabela em MS-SQL Server 2008 Express contendo 4.697 registros, com a mesma estrutura da primeira fonte de dados, descrita na Tabela 21.

Os conteúdos comparados entre as duas fontes de dados foram o primeiro nome do paciente, o último nome do paciente e sua data de nascimento. A Tabela 22 apresenta os parâmetros de configuração utilizados na comparação de campos deste cenário.

Tabela 22: Parâmetros de configuração para comparação de campos no 3º Cenário.

Tipo de Campo	Algoritmo	Sensibilidade (m_i)	1 – Especificidade (u_i)	PMC*
Primeiro Nome	*Soundex*	90%	5%	-
Último Nome	*Soundex*	90%	5%	-
Data de Nascimento	Levenshtein	90%	10%	75%

*PMC: Proporção Mínima de Concordância

Considerando que: (*i*) a comparação entre as duas fontes de dados se deu apenas com três campos, (*ii*) um dos campos é o *soundex* (Subseção 4.4.1) do primeiro nome e (*iii*) o *soundex* usa, obrigatoriamente, a primeira letra para formar sua codificação, foi utilizada blocagem com a primeira letra do primeiro nome do paciente. Com isso, o número de comparações entre as duas fontes de dados caiu de 141.342.124 (30.092 registros da primeira fonte multiplicado por 4.697 registros da segunda) para 12.205.469, como pode ser observado na Tabela 23.

Tabela 23: Número de comparações do 3°Cenário com blocagem da 1ª letra do primeiro nome do paciente.

Primeira letra	Pacientes com catarata	Pacientes com glaucoma	Número de Combinações
A	4.569	621	2.837.349
B	377	76	28.652
C	1.309	230	301.070
D	1.403	189	265.167
E	1.887	275	518.925
F	746	80	59.680
G	1.091	199	217.109
H	585	67	39.195
I	1.211	212	256.732
J	3.565	543	1.935.795
K	17	7	119
L	1.753	260	455.780
M	4.839	915	4.427.685
N	1.102	169	186.238
O	971	121	117.491
P	531	87	46.197
Q	8	2	16
R	1.057	172	181.804
S	996	169	168.324
T	516	80	41.280
U	54	5	270
V	734	111	81.474
W	335	45	15.075
X	1	-	-
Y	22	4	88
Z	413	58	23.954
Total	**30.092**	**4.697**	**12.205.469**

Com base nos parâmetros da Tabela 22 e com a utilização da expressão 5 (Seção 4.5), o maior índice de confiança possível neste cenário é 11,5098 e o menor é -9,6658. Na tabela MPI, só foram gravados índices de confiança positivos.

Após a execução do *Record Linkage* para este cenário, foi obtido o resultado apresentado na Tabela 24.

Tabela 24: Resultado do *Record Linkage* do 3º Cenário.

Índice de Confiança	Número de ocorrências
11,5098	771
11,1135	155
10,7173	1.208
5,1699	53.144
4,0919	225
3,6957	5.541
3,2994	61.136
Índices negativos	**12.083.289**

Ao contrário do 2° cenário, aqui não há uma maneira precisa de detectar falsos positivos e falsos negativos, devido ao fato de não existir um identificador unívoco. Além disso, o conjunto de campos comparados neste cenário é bastante inferior ao anterior. Por essas razões, foram considerados pares combinados somente aqueles que obtiveram índice de confiança máximo. Embora, em uma rápida conferência visual, tenham sido detectados casos de falsos positivos, mesmo neste patamar.

Para contornar esse problema, deve-se considerar, em princípio, que o par com índice igual a 11,5098 representa o mesmo indivíduo. Ao se detectar visualmente que se trata de indivíduos distintos, no momento de uma consulta, deverá ser desenvolvido um recurso que possibilite a retirada desses indivíduos diferentes da posição de igualdade.

Os pares com índices de confiança iguais a 11,1135 são considerados duvidosos. Os 155 pares que compõem essa faixa devem ser conferidos, analisando os outros dados do cadastro e, quando for o caso, serão promovidos de duvidosos a combinados. A distinção entre um par com índice de confiança máximo e aquele que obteve um índice de confiança igual a 11,1135 é apenas um dígito diferente na data de nascimento.

Os pares com índices de confiança inferiores a 11 não devem ser considerados. Nestes casos, já são observadas variações grosseiras, sobretudo, nas datas de nascimento. É possível que haja pares combinados nessas faixas, mas, somente, nos casos em que houver erros de digitação.

Após as considerações descritas, conclui-se que 771 pacientes pertencem às duas fontes de dados e há dúvida em relação a 155 pacientes. Essa massa de dados serve como referência para ações combinadas de um mesmo paciente que sofre tanto de glaucoma quanto de catarata. Isso representa economia de gastos, menor tempo para o tratamento ou execução de procedimento em um mesmo paciente e, muito provavelmente, uma ação de combate a cegueira.

7. Conclusões

O uso de métodos probabilísticos para integrar fontes de dados heterogêneas e distribuídas pode ser bastante útil para encontrar a mesma entidade, sobretudo a mesma pessoa, em bases diferentes. Essa demanda ocorre, especialmente, no domínio saúde, onde a inexistência de um identificador único como o CPF ou o Cartão SUS é bastante comum. No contexto da área de saúde, podem-se ter indivíduos cuja identificação não está limitada a uma documentação padronizada, sobretudo, nos atendimentos de emergências. Grupos como recém-nascidos, estrangeiros, detentos, moradores de rua ou indígenas podem não possuir qualquer tipo de documentação. Nesses casos, é fundamental identificá-los através de dados demográficos, como nome, data de nascimento, nome da mãe, sexo, naturalidade e nacionalidade.

Este trabalho apresenta um estudo sobre soluções que permitem identificar pares existentes em diversas fontes de dados não dotadas de uma identificação unívoca que distingue cada registro, gerando um índice único. Ele permite ainda retorno às fontes de dados originais de forma *on-line* e atualização incremental do índice gerado. Outro ponto forte deste trabalho é sua flexibilidade: trata-se de uma solução que permite configurar as etapas que compõem o *Record Linkage* e combinar fontes de dados com tipos e formatos diferentes.

Para o desenvolvimento deste trabalho, foram analisadas as soluções hoje existentes, capazes de atender a necessidade identificada. Nessas soluções, são observadas a flexibilidade, a natureza não proprietária e adoção de padrões, em especial, a adoção do padrão IHE/PIX, próprio para este fim.

Foi estudado o método *Record Linkage* propriamente dito. Embora este método seja consagrado pela literatura e alvo de muito esforço científico, não há um material que encerre o assunto sobre o *Record Linkage*. O leque de opções no uso dessa técnica é bastante extenso, sobretudo em relação à blocagem, determinação de limiares, especificidade e sensibilidade.

Após esses estudos e baseado nas especificações do padrão IHE/PIX [ACC, HIMSS e RSNA 2008], foi modelada e projetada uma solução para identificação única de pacientes em duas fontes de dados distintas. A partir da modelagem, foi gerado um protótipo, que foi testado em três cenários.

No primeiro cenário, foram combinadas duas fontes de dados heterogêneas (uma relacional e a outra XML), usando dados provenientes de resultados de dois concursos públicos. Neste cenário, foi obtida uma eficácia de 98,5%.

No segundo cenário, foram utilizadas duas bases de dados governamentais que possuem como identificador unívoco o CPF. O objetivo do uso deste cenário é avaliar a eficácia do método utilizado em grandes bases de dados, já que são conhecidos os pares combinados que deveriam ser identificados antes da execução do método. Neste cenário, a eficácia do método ficou em 81,9%.

No último cenário, foram utilizados dados reais de pacientes. Foram combinadas as bases de dados de pacientes com glaucoma e catarata, tratados na Rede Pública do Estado do Espírito Santo. Neste cenário, foram encontrados 771 pacientes que pertencem às duas bases de dados, ou seja, 16,41% dos pacientes com glaucoma também sofrem de catarata e 2,56% dos pacientes com catarata também tem glaucoma. Este dado é especialmente útil, uma vez que 61,1% dos oftalmologistas preferem a cirurgia combinada nos casos em que os pacientes possuem as duas doenças com indicação de cirurgia. Isso representa diminuição de custo, de risco cirúrgico e, muito provavelmente, prevenção à cegueira, uma vez que glaucoma e catarata, juntamente com a retinopatia diabética, são as maiores causas de cegueira no Brasil e no mundo.

Não houve conferência manual no terceiro cenário devido ao volume dos dados.

7.1. Contribuições e Publicações

Como contribuições deste trabalho destacam-se:

- Resumo das questões legais e de ações governamentais que envolvem a identificação de pacientes em diferentes fontes de dados, sobretudo no Brasil;
- Consolidação dos conceitos formais e práticos do método de *Record Linkage*, com a explicação detalhada de cada etapa, chegando ao nível dos códigos-fonte dos algoritmos utilizados;
- Definição dos requisitos do perfil de integração IHE/PIX;
- Modelagem conceitual, lógica e física de uma solução de PIX, utilizando o método de *Record Linkage* e geração de MPI. Nesta solução, não há a proposição de arquiteturas de softwares ou hardwares proprietários;

- Implementação de um protótipo, baseado nos requisitos e modelagem anteriormente descritos, testado em três cenários com dados reais;
- Contribui para o desenvolvimento do projeto “Software Livre e Interoperabilidade em Saúde”, financiado pela FAPES (Fundação de Apoio a Ciência e Tecnologia do Estado do Espírito Santo), outorga 032/2007.
- A partir deste projeto, torna-se possível a identificação única de pacientes em diversas fontes de dados heterogêneas, localizadas em instituições diversas, com adoção de métodos e padrões consolidados pela literatura, sem utilização de arquiteturas proprietárias.

Publicação relacionada a este trabalho:

- Identificação Única de Pacientes em Fontes de Dados Distribuídas e Heterogêneas, CBIS'2008, Campos do Jordão [Soares, Barbosa e Costa 2008].

7.2. Dificuldades Encontradas

Dentre as dificuldades encontradas na execução deste trabalho destacam-se:

- **Compreensão dos conceitos para desenvolvimento de um *Record Linkage*:** As informações encontradas na literatura estão muito dispersas, dificultando a absorção dos conhecimentos necessários para desenvolvimento de cada uma das fases do método. Pelas mesmas razões, também houve muita dificuldade no desenvolvimento da configuração flexível de cada etapa do método de *Record Linkage*.
- **Obtenção das fontes de dados para composição dos cenários e teste do protótipo:** A autorização para uso das bases de dados do segundo cenário contou com um alto grau de dificuldade. Além de necessitar do aval da alta cúpula de cada órgão gestor da informação, foi determinado que os dados fossem manipulados apenas nas dependências do Prodest.
- Para utilização dos dados do terceiro cenário, a restrição foi ainda maior. Manipulação de dados de pacientes é considerada pela Resolução 196/96 do Conselho Nacional de Saúde - Ministério da Saúde como **pesquisa envolvendo seres humanos** (II-2). Portanto, foi necessária a elaboração de um protocolo de pesquisa nos padrões da citada resolução. Este protocolo de pesquisa exige a assinatura de um termo de compromisso por parte dos pesquisadores envolvidos na pesquisa (orientado, orientador e coorientador). Também é necessária uma declaração de apoio institucional por parte da instituição de pesquisa. Neste

caso do coordenador do Programa de Pós-Graduação em Informática - Universidade Federal do Espírito Santo (UFES). Foi também necessária uma autorização do uso dos dados por parte do gestor da informação, neste caso, do Subsecretário de Assuntos de Gestão Hospitalar. Feito isso, a documentação foi submetida ao CEP (Comitê de Ética em Pesquisa) para avaliação. Após parecer de um relator e da avaliação do CEP, os dados foram liberados.

7.3. Trabalhos Futuros

Como trabalhos futuros destacam-se:

- Conclusão da implementação da carga incremental para geração do MPI. Neste caso em particular, é necessária a criação de uma federação de dados [Özsu e Valduriez 2001], que pode absorver os conceitos do *Patient Identifier Cross-reference Domain*, explicitados no Capítulo 3. Neste ambiente de federação, cada vez que for executada a carga incremental de um MPI, são formados novos pares combinados e duvidosos, a partir das fontes de dados originais, que estão em constante atualização. Uma sugestão para composição inicial deste ambiente de federação é o prontuário médico de dois hospitais informatizados que compõem a rede pública estadual do Espírito Santo. Uma outra alternativa de trabalho futuro, ainda neste sentido, é a implementação clássica da transação *PIX Update Notification*, detalhada no Capítulo 5. Neste caso, mensagens ADT alimentariam o MPI.
- União das tecnologias de *Record Linkage* com *Middleware* de Integração [Barbosa 2001], com a finalidade de resolver conflitos semânticos entre metadados, possibilitando uma escolha automática de campos que serão comparados. Neste trabalho futuro, é necessária uma integração entre a fase de Comparação de Campos do método *Record Linkage* com o Componente Gerência de Metadados do CoDIMS [Barbosa 2001] e [Silvestre 2005], que é um ambiente flexível e configurável que possibilita gerar sistemas para integração de dados heterogêneos e distribuídos.
- Incorporação de uma camada de segurança da informação ao protótipo desenvolvido. Implantação de *login* de acesso; habilitação, associando cada usuário a uma ou mais tarefas; alimentação de um *log*, detalhando quem fez o que e quando.
- Estudo sobre paralelismo do método *Record Linkage*, sobretudo na etapa de comparação de campos.

- Conversão para software livre e orientado a objetos. Neste sentido, está sendo desenvolvido um trabalho de conclusão de curso intitulado "*Engine* em Java Fundamentada no Método *Record Linkage* para Integração de Fontes de Dados Heterogêneas", pelo IFES (Instituto Federal do Espírito Santo)[86], que tem como objetivo desenvolver uma camada de software configurável e flexível contendo todos as etapas do processo de *Record Linkage*. Este projeto teve sua proposta aceita e encontra-se em fase de desenvolvimento. O estudo de caso utilizado terá como base dados do programa Bolsa Família [Romero 2008].

- Aprimoramento do protótipo desenvolvido, incorporando algumas funcionalidades, tais como: (*i*) combinação com outros tipos de fontes de dados, como, por exemplo, dados provenientes de SGBDs orientados a objetos; (*ii*) possibilidade de chave composta para geração do MPI; (*iii*) possibilidade de alteração de localização das fontes de dados sem perda do histórico e com a manutenção da funcionalidade de carga incremental; (*iv*) prover recurso de exclusão de falso positivo do MPI, ou seja, excluir um índice gerado quando ele não representar o mesmo indivíduo.

- Desenvolvimento de testes, visando o aumento da eficácia do método utilizado. Estes testes devem envolver: (*i*) substituição do algoritmo *Soundex* pela distância de *Levenshtein* para comparação de nomes; (*ii*) tratamento de campos com conteúdos nulos ou vazios; (*iii*) adoção de novas técnicas de blocagem; e (*iv*) cálculos mais precisos para sensibilidade e especificidade na etapa de comparação entre campos.

- Em função de alguns resultados obtidos, sobretudo no estudo do segundo cenário, será desenvolvido um projeto que buscará inconsistências nos cadastros de servidores públicos estaduais, de condutores de veículos e de proprietários de veículos. Deverão ser buscadas deduplicações, identificadores como CPF nulos ou duplicados. Também será feito um batimento entre o cadastro de condutores de veículos e o cadastro de óbitos, fornecido pelo Ministério da Previdência Social[87], através do Sisobnet, Sistema Informatizado de Óbito[88]. O objetivo desse batimento é descobrir motoristas que já morreram, evitando que o Estado continue tendo gastos com sua regularização. Para este novo projeto, serão disponibilizados recursos de hardware próprios, uma vez que processos de *Record Linkage* dessa natureza consomem várias horas de processamento.

[86] http://www.ifes.edu.br/

[87] http://www.previdencia.gov.br/

[88] http://www.dataprev.gov.br/sisobi/

- Desenvolvimento do projeto de integração de bases de dados criminais, promovido pela Secretaria de Segurança Pública e Defesa Social[89] (SESP) do Estado do Espírito Santo.

[89] http://www.sesp.es.gov.br/sitesesp/index.jsp

REFERÊNCIAS

[ACC, HIMSS e RSNA 2004] ACC; HIMSS; RSNA. IT Infrastructure Technical Framework. Volume 1 (ITI TF-1). Integration Profiles, Revision 1.1, Final Text Version, July 2004. Disponível em http://www.ihe.net/Technical_Framework/upload/ihe_iti_tf_1.1_vol1_FT.pdf. Acessado em março de 2009.

[ACC, HIMSS e RSNA 2005] ACC; HIMSS; RSNA. IT Infrastructure Technical Framework. Volume 1 (ITI TF-1). Integration Profiles, Revision 2.0, Final Text Version, August 2005. Disponível em http://www.ihe.net/Technical_Framework/upload/ihe_iti_tf_2.0_vol1_FT_2005-08-15.pdf. Acessado em agosto de 2008.

[ACC, HIMSS e RSNA 2008] ACC; HIMSS; RSNA. IT Infrastructure Technical Framework. Volume 2 (ITI TF-1). Integration Profiles, Revision 5.0, Final Text Version, December 2008. http://www.ihe.net/Technical_Framework/upload/IHE_ITI_TF_5-0_Vol1_FT_2008-12-12.pdf Acessado em junho de 2009.

[ACC, HIMSS e RSNA 2008a] ACC; HIMSS; RSNA. IT Infrastructure Technical Framework. Volume 2 (ITI TF-2). Transactions, Revision 5.0, Final Text Version, December 2008. http://www.ihe.net/Technical_Framework/upload/IHE_ITI_TF_5-0_Vol2_FT_2008-12-12.pdf Acessado em junho de 2009.

[Almeida 2006] ALMEIDA, H. G. Catarata e Glaucoma. Entrevista concedida à revista SAÚDE em 27 de julho de 2006. Uberlândia, MG. Disponível em http://www.iobh.com.br/arquivos/entrevista_Uberlandia.doc. Acessado em junho de 2009.

[ARTEMIS] Artemis Project Home Page. Disponível em http://www.srdc.metu.edu.tr/webpage/projects/artemis/. Acessado em março de 2009.

[Baeza-Yates e Ribeiro-Neto 1999] BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. *Modern Information Retrieval*. New York: ACM Press/Addison-Wesley, 1999.

[Barbosa 2001] BARBOSA, A. C. P. *Middleware para Integração de Dados Heterogêneos Baseado em Composição de Frameworks*. Tese (Doutorado) – PUC-Rio, 2001. Disponíıvel em http://codims.lprm.inf.ufes.br/publicacoes/tese_alvaro.pdf. Acessado em fevereiro de 2009.

[Baxter, Christen e Churches 2003] BAXTER, R.; CHRISTEN, P.; CHURCHES, T. A comparison of fast blocking methods for record linkage. In *'ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation'*, Washington DC, pp. 25-27. 2003.

[Benjelloun et al. 2005] BENJELLOUN, O.; GARCIA-MOLINA, H.; SU, Q.; WIDOM, J. Swoosh: A generic approach to entity resolution. *Technical report*, Stanford University, 2005.

[Bentley e Sedgewick 1997] BENTLEY, J.L.; SEDGEWICK, R.A. Fast Algorithms for Searching and Sorting Strings. *Proceedings of the Eighth ACM-SIAM Symposium on Discrete Algorithms, 360-36*, 1997.

[Bhagat e Hovy 2007] BHAGAT, R.; HOVY, E. Phonetic Models for Generating Spelling Variants, In *Proceedings International Joint Conference of Artificial Intelligence (IJCAI)*. Hyderabad, India. 2007.

[Bierrenbach et al. 2007] BIERRENBACH, A. L.; STEVENS, A. P.; GOMES, A. B. F.; NORONHA, E. F.; GLATT, R.; CARVALHO, C. N.; OLIVEIRA JR., J. R; SOUZA, M. F. M. Efeito da remoção de notificações repetidas sobre a incidência da tuberculose no Brasil. *Revista Saúde Pública*. 2007; 41(Supl. 1): 67-76.

[Bilgic et al. 2006] BILGIC, M.; LICAMELE, L.; GETOOR, L.; SHNEIDERMAN, B. D-Dupe: an interactive too for entity resolution in social networks. *IEEE Symposium on Visual Analytics Science and Technology 2006* (Baltimore, MD), IEEE Computer Society; Silver Spring, MD, 2006; 43-50.

[Borsato et al. 2006] BORSATO, G. G.; SCALABRIN, E. E.; DIAS, J. S.; ENEMBRECK, F. Recuperação de Informação em Situações de Urgência-Emergência no Atendimento Pré-Hospitalar. In: *RESI - Revista Eletrônica de Sistemas de Informação*. Edição 9, Nº3 2006.

[Branting 2003] BRANTING, L. K. A comparative evaluation of name-matching algorithms. In: *International Conference on Artificial Intelligence and Law (ICAIL)*, pages 224–232. ACM Press, 2003.

[BRASIL 1940] BRASIL. Decreto-Lei n° 2.484 de 07 de Dezembro de 1940. Código Penal. Brasília, DF: Senado Federal.

[BRASIL 1988] BRASIL. Constituição da República Federativa do Brasil. Brasília, DF: Senado Federal.

[Camargo e Coeli 2000] CAMARGO JR., K. R.; COELI, C. M. *Reclink*: aplicativo para o relacionamento de bases de dados, implementando o método "*probabilistic record linkage*". *Cadernos Saúde Pública 2000*; 16: 439-47, 2000.

[Camargo e Coeli 2006] CAMARGO JR., K. R.; COELI, C. M. *RecLink* 3: Nova Versão do Programa que implementa a Técnica de Associação Probabilística de Registros (*Probabilistic Record Linkage*). *Cadernos Saúde Coletiva*, Rio de Janeiro, 14 (2): 399 - 404, 2006.

[Campbell 2005] CAMPBELL, K. M. Rule Your Data with The Link King (a SAS/AF application for record linkage and unduplication). *SUGI 30*, 2005.

[Cardoso e Sabbatini 1999] CARDOSO, S. H.; SABBATINI, R. M. E. Novo Padrão Facilitará Intercâmbio de Dados sobre Pacientes. *Revista Informática Médica*, Volume 2, Número 1, Janeiro/Fevereiro 1999.

[CFM 1988] Conselho Federal de Medicina. Código de Ética Médica. 1988.

[CFM 2002] Conselho Federal de Medicina. Câmera Técnica de Informática em Saúde. Processo Consulta CFM n. 1.401/2002. *Prontuário eletrônico*. Brasília, 2002. Disponível em: http://www.portalmedico.org.br/pareceres/cfm/2002/30_2002.htm. Acessado em junho de 2009.

[Chassin e Becher 2002] CHASSIN, M. R.; BECHER, E. C. The Wrong Patient. In: *Annals of Internal Medicine*, 136, p. 826-833, 2002.

[Chaudry et al. 2006] CHAUDRY, N. G.; ILYAS, S; SHAHZAD, A; SALEEM, A.; RASHID, M.; CHAUDHRY, T. A. An Open Source Health Care Management System for Pakistan, *Paper ID-ICOST2006-10*, 2006.

[Christen 2008] CHRISTEN, P. Febrl: a freely available record linkage system with a graphical user interface. *Conferences in Research and Practice in Information Technology Series; Vol. 327 archive. Proceedings of the second Australasian workshop on Health data and knowledge management - Volume 80*, 2008.

[Christen 2008a] CHRISTEN, P. Febrl – An open source data cleaning, deduplication and record linkage system with a graphical user interface. In: *ACM International Conference on Knowledge Discovery and Data Mining, Las Vegas (2008) 1065–1068*, 2008.

[Christen e Churches 2005] CHRISTEN, P.; CHURCHES, T. A probabilistic deduplication, record linkage and geocoding system. *Proceedings of the ARC Health Data Mining workshop*, pp. 109-116. The Australian National University, Canberra, AU. 2005

[Christen, Churches e Hegland 2004] CHRISTEN, P.; CHURCHES, T.; HEGLAND, M. Febrl – a parallel open source data linkage system. In: *PAKDD, Springer LNAI 3056*, pages 638–647, Sydney, 2004.

[Christophilopoulos 2005] CHRISTOPHILOPOULOS, E. ARTEMIS (IST-1-002103-STP): A Semantic Web Service-based P2P Infrastructure for the Interoperability of Medical Information Systems. *InnoFire Medical Cooperation Network Newsletter*, September 2005.

[Churches, Christen, Lim e Zhu 2002] CHURCHES, T.; CHRISTEN, P.; LIM, K.; ZHU, J. X. Preparation of name and address data for record linkage using hidden Markov models. *BMC - Biomed Central Medical Informatics and Decision Making, 2(9),* 2002. Disponível em http://www.biomedcentral.com/1472-6947/2/9/. Acessado em março de 2009.

[CNS] Portal da Saúde. Cartão Nacional de Saúde. Disponível em http://portal.saude.gov.br/portal/saude/Gestor/area.cfm?id_area=944#. Acessado em junho de 2009.

[Coeli et al. 2006] COELI, C. M. et al. *Uso Integrado de Bases de Dados Avaliação em Saúde*. Projeto de Pesquisa, CNPq, 2006. Disponível em http://www.nates.ufjf.br/novo/encontro/Xaps/pdf/palestras/usointegrado.pdf. Acessado em março de 2009.

[Cohen, Ravikumar e Fienberg 2003] COHEN, W.; RAVIKUMAR, P.; FIENBERG, S. A comparison of string metrics for matching names and records. *KDD-2003 Workshop on Data Cleaning and Object Consolidation*.

[CONIP 2006] DUTRA, R.; BRETAS, N.; FERREIRA, S.; GAMARSKI, R. Integração de informações: Identificação Única de Pacientes, Profissionais e Estabelecimentos de Saúde. *Congresso de Inovação da Gestão Pública (CONIP Saúde 2006)*, 2006. Disponível em http://ww2.conip.com.br/saude2006/programacao.php. Acessado em fevereiro de 2009.

[Cunha 2002] CUNHA, R. E. Cartão Nacional de Saúde – os desafios da concepção e implantação de um sistema nacional de captura de informações de atendimento em saúde. *Revista Ciência e Saúde Coletiva*, Rio de Janeiro, 2002, v. 7, n. 4, p. 869-878. ISSN 1413-8123.

[Dal Maso, Braga e Franceschi 2001] DAL MASO, L.; BRAGA C.; FRANCESCHI S. Methodology Used for "Software for Automated Linkage in Italy" (SALI). *Journal of Biomedical Informatics*, 34:387-395, 2001.

[DATASUS 2009] Portal de Cadastros Nacionais. Disponível em http://cartaonet.datasus.gov.br/. Acessado em junho de 2009.

[DATASUS 2009a] Portal de Cadastros Nacionais. Produtos Desenvolvidos. Disponível em http://cartaonet.datasus.gov.br/Produtos.html#cadsus. Acessado em junho de 2009.

[D-Dupe] D-Dupe: A Novel Tool for Interactive Data Deduplication and Integration. Disponível em http://www.cs.umd.edu/projects/linqs/ddupe/. Acessado em junho de 2009.

[DICOM] NEMA. DICOM - Digital Imaging and Communications in Medicine. Disponível em http://dicom.nema.org/. Acessado em fevereiro de 2009.

[Dogac, Bicer e Okcan 2006] DOGAC, A.; BICER, V.; OKCAN, A. Collaborative Business Process Support in IHE XDS through ebXML Business Processes, In *proc. of International Conference on Data Engineering (ICDE2006)*, Atlanta, USA, April 2006.

[Drumond, Machado e França 2008] DRUMOND, E. F.; MACHADO, C. J.; FRANÇA, E. Subnotificação de nascidos vivos: procedimentos de mensuração a partir do Sistema de Informação Hospitalar. *Revista de Saúde Pública 2008*; 42 (1): 55-63, 2008.

[Dunn 1946] DUNN, H. L. Record linkage. *American Journal of Public Health*, Washington, D.C, v. 36 n. 12, p. 1412-1416, Dec., 1946.

[Eichelberg, Aden e Thoben 2005] EICHELBERG, M.; ADEN, T.; THOBEN, W. A Distributed Patient Identification Protocol based on Control Numbers with Semantic Annotation. In: *Journal on Semantic Web and Information Systems*, Editor: Sheth, A., Lytras, M. D., Vol. 1, No. 4, pages 4-43, 2005.

[Elmagarmid, Ipeirotis e Verykios 2007] ELMAGARMID, A. K.; IPEIROTIS P. G.; VERYKIOS V. S. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19(1):1-16*, 2007.

[Elmasri e Navathe 2005] ELMASRI, R. E.; NAVATHE, S. B. *Sistemas de banco de dados*. Rio de Janeiro: Addison-Wesley, 2005.

[EPSJV 2008] Escola Politécnica de Saúde Joaquim Venâncio (Organização). *Pesquisa: "uso integrado de base de dados na avaliação em saúde": material didático (tutorial)*. Rio de Janeiro, RJ, 2008.

[Favaro e Vieira 2008] FAVARO, T.; VIEIRA, V. Não será por falta de memória – Quadro: Conteúdo maior que a memória. In.: *Revista VEJA*, 2058ª ed. Editora Abril, pág. 98 a 99. 30 de abril de 2008.

[Fellegi e Sunter 1969] FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. *Journal of American Statistical Association*, 66(1):1183-1210.

[Fiales et al. 2001] FIALES, V. R.; NARDON, F. B.; FURUIE, S. S. Construção de um Serviço de Identificação de Pacientes. *Revista Eletrônicade Iniciação Científica*, Agosto, 2001.

[Francisco Jr. et al. 2008] FRANCISCO JR., G. O.; GOTTBERG, H.; MANCINI, F.; LEDERMAN, H. M.; PISA, I. T. Validade do Prontuário Médico Eletrônico como Prova Jurídica. XI Congresso Brasileiro de Informática em Saúde (CBIS'2008), 2008.

[Freeman e Freeman 2007] FREEMAN, E.; FREEMAN, E. *Use a cabeça! Padrões de Projeto (Design Patterns)*. 2. ed. Rio de Janeiro, RJ: Alta Books, 2007.

[Grannis 2008] GRANNIS, S. An Overview of Patient Matching. *Website*: http://www.docstoc.com. Disponível em http://www.docstoc.com/docs/2359888/An-Overview-of-Patient-Matching. Acessado em junho de 2009.

[Gu et al. 2003] GU, L.; BAXTER, R.; VICKERS, D.; RAINSFORD, C. Record linkage: Current practice and future directions. *Tech. Rep. 03/83, CSIRO Mathematical and Information Sciences*, 2003.

[HC-FMUSP] Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo – Quem Somos. Disponível em http://www.hcnet.usp.br/instituicao/quem_somos.htm. Acessado em junho de 2009.

[Henderson e Bao 2005] HENDERSON, M; E BAO, Y. Patient Identifier Cross-Referencing for MPI (PIX). *IHE IT Infrastructure Technical Committee. IHE Technical Webinar 2005*.

[Herzog et al. 2007] HERZOG, T. N.; SCHEUREN, F. J.; WINKLER, W. E. *Data Quality and Record Linkage Techniques*. Springer, July 2007.

[HL7] Health Level Seven. HL7 Version 2.3 Specification. 1997. Disponível em http://www.hl7.org/. Acessado em agosto de 2008.

[HL7 2003] HL7. HL7 - Application Protocol for Electronic Data Exchange in Healthcare Environments Version 2.5, June 2003.

[IHE 2009] Integrating the Healthcare Enterprise. Disponível em http://www.ihe.net/. Acessado em junho de 2009.

[IHE 2009a] Integrating the Healthcare Enterprise. IHE Connectathon Results Browser. Disponível em http://sumo.irisa.fr/con_result/. Acessado em junho de 2009.

[IHE 2009b] Integrating the Healthcare Enterprise. Member Organizations. Disponível em http://www.ihe.net/governance/member_organizations.cfm. Acessado em junho de 2009.

[Ikeda e Porter 2007] IKEDA, M ; PORTER, E. Initial Results from a Nationwide BigMatch Matching of 2000 Census Data. *Census Bureau, Statistical Research Division*, 2007.

[Jaro 1989] JARO, M. A. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84:414-420, 1989.

[JCI 2005] Joint Commission International Center for Patient Safety (Eds.). Technology in Patient Safety - Using Identification Bands to Reduce Patient Identification Errors. In: *Joint Commission Perspectives on Patient Safety*, 5, p. 1–10, 2005.

[Jurczyk et al. 2008] JURCZYK, P. et al. FRIL: A Tool for Comparative Record Linkage. *American Medical Informatics Association (AMIA) 2008 Annual Symposium*, 2008.

[Kim e Seo 1991] KIM, W.; SEO, J. Classifying schematic and data heterogeneity in multidatabase systems, *IEEE Computer Society Press*, v. 24, n. 12, p. 12-18, 1991. ISSN 0018-9162.

[Knuth 1973] KNUTH, D. *The Art of Computer Programming - Volume 3: Sorting and Searching*. Addison-Wesley Publishing Company, 1973.

[Koudas et al. 2006] KOUDAS, N.; SARAWAGI S.; SRIVASTAVA D. Record linkage: similarity measures and algorithms. *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, June 27-29, 2006, Chicago, IL, USA

[Krechel e Hartbauer 2008] KRECHEL, D.; HARTBAUER, M. The LENUS Master Patient Index: Combining Hospital Content Management with a Healthcare Service Bus, *21st IEEE International Symposium on Computer-Based Medical Systems, 2008. CBMS '08*, p. 170-172, 2008. ISSN 1063-7125.

[LENUS] Ihr Spezialist für digitale Dokumentenlogistik. Disponível em http://www.lenus.info/ww/de/pub/healthcare.cfm. Acessado em março de 2009.

[Levenshtein 1965] LEVENSHTEIN, V. Binary codes capable of correcting spurious insertions and deletions of ones. Probl. Inf. Transmission 1, 8-17.

[Levenshtein 1966] LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions and reversals. Sov. Phys. Dokl. 10, 8, 707-710. Original in Russian in Dokl. Akad. Nauk SSSR 163, 4, 845-848, 1965.

[Link Plus] Cancer - Registry Plus™ Link Plus - NPCR. Disponível em http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm. Acessado em junho de 2009.

[LinkageWiz] LinkageWiz Inc. Record linkage software. Version 5.0. Disponível em http://www.linkagewiz.com/. Acessado em junho de 2009.

[Martha et al. 2004] MARTHA, A. S.; BARRA, P. S. C.; CAMPOS, J. R. Recuperação de Informações em Textos Livres de Prontuários do Paciente. In: *Congresso Brasileiro de Informática Médica*, Ribeirão Preto, 2004.

[Martin e McClure 1991] MARTIN, J.; MCCLURE, C. *Técnicas Estruturadas e CASE*. São Paulo: Makron Books, 1991.

[Martinhago 2006] MARTINHAGO, C. Z. *Customização Em Ambientes de Qualidade de Dados*. Dissertação (Mestrado) - UFPR, 2006. Disponível em

http://dspace.c3sl.ufpr.br/dspace/bitstream/1884/4797/1/dissertacao_adriana.pdf. Acessado em março de 2009.

[Martins 2008] Martins, A. C. Norma DICOM. *Copyright © Siemens AG 2008. All rights reserved.* Disponível em http://www.urbana.fm/~antocm/files/docs/ucp.pt/Advanced-Dicom-2008-v2.pdf. Acessado em agosto de 2009.

[Massad et al. 2003] MASSAD, E. et al. *Prontuário eletrônico do paciente: definições e conceitos*. In: MARIN, Heimar de Fátima et al. O Prontuário eletrônico do paciente na assistência, informação e conhecimento médico. São Paulo: H. de F. Marin, 2003, p. 1-20. Capítulo 1. OPAS/OMS, 2003.

[McIlraith, Son e Zeng 2001] MCILRAITH, S. A.; SON, T. C.; ZENG, H. Semantic Web Services, *IEEE Intelligent Systems*, 16(2), 46-53., 2001.

[Melo et al. 2005] MELO, R. N.; MOURA, S. L.; SILVA, F. J. C.; SIQUEIRA, S. W. M. Integrating Repositories of Learning Objects Using Web-Services and Ontologies. *International Journal of Web Services Practices*, Seoul, v. 1, n. 1-2, p. 57-72, 2005.

[Mettler, Fitterer e Rohner 2007] METTLER, T.; FITTERER, R.; ROHNER, P. Strategies for a Systematical Patient Identification. *European Conference on eHealth (ECEH 2007)*, p. 23–30.

[Microsoft 2009] Microsoft Corporation. Connected Health Framework Architecture and Design Blueprint, A Stable Foundation for Agile Health and Social Care, Part 3 - Technical Framework. *Knowledge Driven Health - Second Edition Published March 2009*. Disponível em http://www.microsoft.com/industry/healthcare/technology/healthframework.mspx. Acessado em junho de 2009.

[Moon 1996] MOON, T. K. The expectation-maximation algorithm. *IEEE Signal ProcessingMagazine*, pages 47–70, Nov. 1996.

[MS 1999] Ministério da Saúde. Departamento de Informática do SUS (DATASUS). Recomendação Final do Comitê de Padronização de Registros Clínicos sobre a Solicitação de Proposta SOP 001/99 - Versão 1.0. Brasília: Ministério da Saúde, 1999.

[MS 2004] Ministério da Saúde. Política Nacional de Informação e Informática em Saúde - Proposta Versão 2.0. Brasília: Ministério da Saúde, 2004.

[MS 2007] Ministério da Saúde. Agência Nacional de Saúde Suplementar. Padrão TISS – Troca de Informações em Saúde Suplementar. Versão 3.00, atualizada em maio de 2007. Disponível em http://www.ans.gov.br/portal/site/_hotsite_tiss/pdf/texto_completo.pdf. Acessado em junho de 2009.

[MS 2007a] Ministério da Saúde. Agência Nacional de Saúde Suplementar. Padrão A integração com o SUS, na perspectiva da DIDES. Disponível em http://www.ans.gov.br/portal/site/Biblioteca/encontro_ans_ops_natal.asp. Acessado em junho de 2009.

[MS 2008] Ministério da Saúde. Agência Nacional de Saúde Suplementar. Caderno de Informação de Ressarcimento e Integração com o SUS. Rio de Janeiro: Ministério da Saúde, 2008.

[MS-CNS] Ministério da Saúde. Secretaria de Gestão de Investimentos em Saúde. *Cartão Nacional de Saúde: instrumento para um novo modelo de gestão em saúde.* Brasília: Ministério da Saúde. Disponível em: http://dtr2001.saude.gov.br/cartao. Acessado em junho de 2009.

[Navarro 2001] NAVARRO, G. A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, v. 33, n. 1, p. 31-88, mar. 2001.

[Newcombe et al. 1959] NEWCOMBE, H. B.; KENNEDY, J. M.; AXFORD, S. J.; JAMES, A. P. Automatic Linkage of Vital Records. *Science,* 130, 954-959. 1959.

[Newcombe e Kennedy 1962] NEWCOMBE, H. B.; KENNEDY, J. M. Record Linkage making Maximum Use of the Discrimination Power of Identifying Information. *Commun ACM* 1962; 5: 563-6.

[Oliveira 2007] OLIVEIRA, I. C. *Desenvolvimento e Aplicação de um Modelo para Relacionar Diferentes Sistemas de Informação na Área da Saúde*. Tese (Doutorado) - Universidade Federal de Santa Catarina, 2007.

[OMG 2001] Object Management Group, Inc. Person Identification Service (PIDS) Specification, Version 1.1, April 2001. Disponível em http://www.omg.org/technology/documents/formal/person_identification_service.htm. Acessado em junho de 2009.

[OpenEMed] OpenEMed. Disponível em http://www.openemed.org/. Acessado em junho de 2009.

[Özsu e Valduriez 2001] ÖZSU, M. T.; VALDURIEZ, P. *Princípios de sistemas de bancos de dados distribuídos*. 2. ed. Rio de Janeiro, RJ: Campus, 2001. Tradução da 2. ed. americana, Vandenberg D. de Souza.

[Pereira 1995] PEREIRA, M. G. *Epidemiologia: teoria e prática*. Rio de Janeiro: Guanabara Koogan S.A., 1995.

[Pitoura, Bukhres e Elmagarmid 1995] PITOURA, E.; BUKHRES, O.; ELMAGARMID, A. Object orientation in multidatabase systems. *ACM Comput. Surv.*, ACM Press, v. 27, n. 2, p. 141-195, 1995. ISSN 0360-0300.

[Rahm e Do 2000] RAHM, E; DO, H. H. Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23(4):1-11, 2000.

[Renly 2007] RENLY, S, R. Patient Identifier Cross-reference Consumer Architecture & API Documentation, Version 0.2.0. *OHF - Open Healthcare Framework*, December 2007. Disponível em http://wiki.eclipse.org/images/c/c4/OHF_Client_Patient_Identifier_Cross-reference_Consumer_beta.pdf. Acessado em março de 2009.

[Rijsbergen 1979] BAEZA-YATES, C. *Information Retrieval*. London: Butterworths, 2nd Edition, 1979.

[Romero 2008] ROMERO, J. A. R. *Utilizando o Relacionamento de Bases de Dados para Avaliação de Políticas Públicas: Uma Aplicação para o Programa Bolsa Família*. Tese (Doutorado) - Universidade Federal de Minas Gerais, 2008.

[Rötzch 2006] RÖTZCH, J. M. Relacionamento Nominal de Banco de Dados - Record Linkage. *ANS - Agência Nacional de Saúde Suplementar - Ministério da Saúde*, 2006. Disponível em http://www.ans.gov.br/portal/upload/informacoesss/Relacionamento_banco_dados.ppt. Acessado em fevereiro de 2009.

[Santhiago et al. 2009] SANTHIAGO, M. R; GOMES, B. A. F.; GAFFREE, F. F. P.; VARANDAS, V. S; COSTA FILHO, A. A. Tendências evolutivas dos cirurgiões de catarata presentes no IV Congresso Brasileiro de Catarata e Cirurgia Refrativa. *Revista Brasileira Oftalmologia*. 2009; 68 (1): 13-7

[SAUDE-SC 2004] Secretaria de Estado da Saúde do Estado de Santa Catarina. *Link Plus - Guia simplificado do Usuário - Versão 1.0 de 4 de janeiro de 2004*. Texto original do *User's Guide do Link Plus*, traduzido e modificado pelo DASIS/SVS/MS. Disponível em http://www.saude.sc.gov.br/download/LinkPlus/. Acessado em junho de 2009.

[Schaback e Li 2007] SCHABACK, J. E; LI, F. Multi-level feature extraction for spelling correction. In: *International Joint Conference on Artificial Intelligence (IJCAI), Workshop on Analytics for Noisy Unstructured Text Data*, pages 79–86, Hyderabad, India.

[SESA 2003] Secretaria de Estado da Saúde do Estado do Espírito Santo. Relatório de Gestão - 2002. Vitória: *Secretaria de Estado da Saúde*, setembro de 2003.

[Shet e Larson 1990] SHET, A. P.; LARSON, J. A. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases. *ACM Computing Surveys*, Vol. 22, No. 3, pp. 183-236, September 1990.

[Siemens 2007] SIEMENS Medical. RHÖN-KLINIKUM AG and Siemens Medical Solutions introduce intra-facility electronic health record (WebEPA) - One of Europe's biggest e-health projects is launched. *Bad Neustadt a. d. S. / Erlangen*, 18 September 2007. Disponível em http://www.medical.siemens.com/siemens/sv_SE/rg_marcom_FBAs/files/Press_Releases/2007/Word/Rhoen_EPA_e.doc. Acessado em junho de 2009.

[Silva et al. 2006] SILVA, J. P. L.; TRAVASSOS, C.; VASCONCELLOS, M. M.; CAMPOS, L. M. Revisão Sistemática sobre Encadeamento ou *Linkage* de Bases de Dados Secundários para uso em Pesquisa em Saúde no Brasil. *Cadernos Saúde Coletiva*, Rio de Janeiro, 14 (2): 197 - 224, 2006.

[Silvestre 2005] SILVESTRE, L. J. *Uma Abordagem Baseada em Ontologias para a Gerência de Metadados do CoDIMS*. Dissertação (Mestrado) - UFES, 2005. Disponível em http://codims.lprm.inf.ufes.br/publicacoes/dissertacaoLeonardo.pdf. Acessado em julho de 2009.

[Smith et al. 2006] SMITH, E. et al. Eclipse Open Healthcare Framework. Eclipse *Summit Europe 2006, Server-Side Eclipse Symposium*, Germany, October 2006. Disponível em http://www.eclipsecon.org/summiteurope2006/presentations/ESE2006_OHF_Bridge.pdf. Acessado em março de 2009.

[Soares, Barbosa e Costa 2008] SOARES, V. F.; BARBOSA, A. C. P.; COSTA, R. G. Identificação Única de Pacientes em Fontes de Dados Distribuídas e Heterogêneas. *XI Congresso Brasileiro de Informática em Saúde (CBIS'2008)*, 2008.

[Soundex] National Archives and Records Administration - Soundex System. Disponível em http://www.archives.gov/genealogy/census/soundex.html. Acessado em junho de 2009.

[Stolba e Schanner 2007] STOLBA, N.; SCHANNER, A. eHealth Integrator - Clinical Data Integration in Lower Austria. *Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED 2007)*, Plymouth, England, July 2007.

[Sun] Sun in the Healthcare Industry: IHE. IHE Compliant Master Patient Index based on Sun Java™ Composite Application Platform Suite (CAPS). Disponível em http://www.sun.com/solutions/landing/industry/healthcare/ihe_mpi/ihe_javacaps.pdf. Acessado em junho de 2009.

[Sung et al. 2005] SUNG, L. G. A.; AHMED, N.; BLANCO, R.; LI, H; SOLIMAN, M. A.; HADALLER, D. A Survey of Data Management in Peer-to-Peer Systems. *Technical report, School of Computer Science, University of Waterloo*, 2005. Disponível em http://www.cs.uwaterloo.ca/research/tr/2006/CS-2006-18.pdf. Acessado em junho de 2009.

[Waegemann 1996] WAEGEMANN, C.P. The five levels of electronic health records. *MD Computing*, v.13, n. 3, 1996.

[Wiederhold e Genesereth 1995] WIEDERHOLD, G.; GENESERETH, M. The basis for mediation. In *Proc. 3rd International Conference on Cooperative Information Systems (COOPIS95)*, 1995.

[Winkler 2004] WINKLER, W. E. Methods for evaluating and creating data quality. *Information Systems*, v.29 n.7, p.531-550, October 2004.

[Winkler e Yancey 2006] WINKER, W. E.; YANCEY, W. E. Parallel BigMatch. *technical report*, to appear.

[Yancey 2004] YANCEY, W.E. Bigmatch: A Program for Large-Scale Record Linkage. In: *Proceedings of the Survey Research Methods Section, American Statistical Association*, Session 100: Record Linkage and Unduplication Methodology and Results, 2004.

[Yancey 2007] YANCEY, W.E. Bigmatch: BigMatch: A Program for Extracting Probable Matches from a Large File for Record Linkage. *Research Report RR 2007-01,* Statistical Research Division, US Bureau of the Census, March 2007.

[Ziegler e Dittrich 2004] ZIEGLER, P.; DITTRICH, K. R. Three Decades of Data Integration - All Problems Solved? *18th IFIP World Computer Congress (WCC 2004)*, Volume 12, Building the Information Society, Kluwer 156(), 3-12.

[Zobel e Dart 1996] ZOBEL, J.; DART, P. Phonetic string matching: lessons from information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, p.166-172, August 18-22, 1996, Zurich, Switzerland.

[Zobel e Moffat 2006] ZOBEL, J.; MOFFAT, A. Inverted files for text search engines, *ACM Computing Surveys*, article 6, Vol. 38 No.2. 2006.