INSTITUTO DE PESQUISAS TECNOLÓGICAS DO ESTADO DE SÃO PAULO

VERA LUCIA REIKO YOSHIDA SHIDOMI

Processo de Melhoria da Qualidade dos Dados: Um Estudo de Caso

São Paulo

2004

VERA LUCIA REIKO YOSHIDA SHIDOMI

Processo de Melhoria da Qualidade dos Dados: Um Estudo de Caso

Dissertação apresentada ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT, para obtenção do título de Mestre em Engenharia de Computação.
Área de concentração: Engenharia de Software

Orientadora: Drª Edit Grassiani Lino de Campos

São Paulo

2004

Shidomi, Vera Lúcia Reiko Yoshida
Processo de melhoria da qualidade dos dados: um estudo de caso. / Vera Lúcia Reiko Yoshida Shidomi. São Paulo, 2004.
112p.

Dissertação (Mestrado em Engenharia de Computação) - Instituto de Pesquisas Tecnológicas do Estado de São Paulo. Área de concentração: Engenharia de Software

Orientador: Prof. Dra. Edit Grassiani Lino de Campos

1. Qualidade dos dados 2. Reengenharia de dados 3. Padronização de dados 4. CRM 5. Tese I. Instituto de Pesquisas Tecnológicas do Estado de São Paulo. Centro de Aperfeiçoamento Tecnológico II. Título

CDU 004.052.42(043)
S555p

Dedicatória

À minha mãe com amor e saudades.

Agradecimentos

Agradeço ao meu esposo Carlos Shidomi pela ajuda, dedicação e companheirismo em todos os momentos. À minha orientadora Edit pela paciência, compreensão e incentivo. Aos amigos de trabalho, Paiva, Américo, Daniel e João, que me deram um voto de confiança e permitiram a experiência de conhecer esta área de qualidade de dados. As amigas Carla e Eloisa pela força e ajuda sempre presente. E todos os familiares, amigos e colegas de trabalho que sempre me ajudaram prontamente, especialmente, meu pai Yoshio, meu irmão Paulino, meus Padrinhos Carlos e Luisa, Hiroki e Kátia, Luis e Adriana.

À todos o meu sincero: Muito Obrigada.

RESUMO

A implantação de um projeto para melhoria da qualidade de dados (QD) requer a utilização de uma metodologia específica para essa área. Existem algumas metodologias propostas referentes ao assunto qualidade, mas especificamente para qualidade dos dados com aplicações em ambiente real validando tais metodologias não há praticamente nenhum registro.

Esse trabalho tem como objetivo, utilizando uma metodologia de QD, implementar um caso prático e apresentar os resultados e as dificuldades encontradas em cada fase. Dentre as diversas metodologias levantadas, foi escolhida a TQdM (*Total Quality data Management*), pois ela apresenta com mais detalhes todas as fases de desenvolvimento.

Esta metodologia é aplicada na melhoria da qualidade dos dados de várias fontes de informação referente a clientes de uma instituição financeira visando a geração de um cadastro consolidado. Este processo envolve a limpeza e a padronização dos dados, o enriquecimento das informações com algumas fontes fidedignas, a identificação e eliminação de duplicados, a transformação e disponibilização dos dados.

São documentadas todas as etapas do processo e a partir dos problemas encontrados são sugeridas diversas soluções às dificuldades encontradas, propondo o refinamento da metodologia original com o acréscimo de processos adicionais.

Palavras chaves: qualidade dos dados, reengenharia dos dados, CRM, limpeza de dados, padronização de dados, enriquecimento de dados.

ABSTRACT

The implementation of a project for improving data quality (DQ) requires the use of specific methodology. Some methodologies have been proposed for this issue, but none for data quality applied to a real environment.

The target of this project is to use DQ methodology to implement a case study and presents its results and difficulties encountered in each phase of the process and solutions. Among many methodologies evaluated, TQdM (Total Quality data Management) was chosen because it provides detailed information about all development phases.

The methodology is applied to improve data quality of several client information sources from a financial institution, in order to obtain a consolidated list of unique persons. The process consists of data cleansing and standardization, information enrichment with trustworthy sources, identification and elimination of duplicates, data transformation and delivery.

All process phases are documented and suggestions are given to tackle the problems encountered. Further refinement is recommend with the addition of new processes.

Keys Words: data quality, data reengineering, CRM, data cleasing, data standardization, data enhancement.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas

CRM	Custom Relation Management
DW	Data Warehouse
MIT	Massachusetts Institute of Technology
PDCA	Plan, Do, Check, Act
PDSA	Plan, Do, Study, Act
QD	Qualidade de Dados
TDQM	Total Data Quality Management
TQdM	Total Quality data Management
TQM	Total Quality Management

Sumário

Capítulo 1 - Introdução

1.1.Motivação

O assunto qualidade de dados (QD) já existe há muitos anos e vários autores (Deming[1], Juran[2], Crosby[3], Ishikawa[4]) contribuíram para o desenvolvimento da área, criando e aprimorando processos de melhoria contínua. Não há porém, praticamente nenhum histórico de aplicação de metodologias para projetos de qualidade de dados em ambientes reais de desenvolvimento.

Atualmente, a importância de projetos de QD tem aumentado devido à necessidade das empresas obterem informações íntegras e confiáveis de seus sistemas, principalmente com a implantação de sistemas de relacionamento com o cliente (*Customer Relationship Management*): conhecer o cliente tornou-se fundamental. Além disso, com a crescente expansão de sistemas de suporte à decisão, acessando e replicando várias bases de dados dos sistemas transacionais, e com os diversos processos de fusões de empresas, atingindo diferentes ramos de atividades, surge à necessidade de obter informações íntegras, pois o usuário se torna cada vez mais rígido com relação à qualidade das informações utilizadas.

Fornecer como produto informações divergentes denota o caos, acarretando a perda da credibilidade do sistema sendo utilizado. Frente a seus usuários, o custo de informações de baixa qualidade gera grandes despesas e põe em risco a organização, pois em um ambiente competitivo a qualidade da informação é seu diferencial. É a sua vantagem competitiva. Mas implantar um projeto de QD requer uma metodologia específica para garantir a sua implementação.

Nota-se a necessidade de pesquisas para auxiliar as primeiras implementações de um projeto deste tipo, abordando metodologias de implementação de projetos de QD e apresentando um caso real.

1.2.Objetivo

O objetivo é implementar um projeto de QD utilizando uma metodologia especifica para essa área, apresentando todas as etapas do processo, pontos importantes e sugestões de melhoria em termos de novos processos e procedimentos.

Dentre as várias metodologias, a TQdM[7] (*Total Quality data Management*) foi escolhida por se tratar de uma metodologia que apresenta uma literatura completa com todas as fases de um projeto de QD de forma detalhada. O principal processo da metodologia é a reengenharia, que tem como objetivos: efetuar a limpeza, a padronização, o enriquecimento e a identificação de duplicidades nas informações para obter informações de qualidade.

O estudo de caso tem como objetivo aplicar a metodologia TQdM com o intuito de melhorar a qualidade dos dados das diversas fontes de informação e gerar um cadastro consolidado a ser utilizado em ações estratégicas pelas diversas áreas de negócios de uma instituição financeira. Ao término da implementação do estudo de caso, serão apresentadas soluções de melhorias para os processos existentes e novos processos serão propostos para as próximas implementações.

1.3.Contribuição esperada

Espera-se que este estudo seja de valia para empresas que desejam implantar um projeto na área de QD. Existem algumas referências relatando sobre os conceitos, processos e metodologias aplicáveis a QD, mas nenhuma apresenta um caso prático de um projeto e explica como aplicar a teoria na prática.

Este trabalho, além de documentar todos os processos utilizados no estudo de caso, servirá para que erros e vícios comuns sejam evitados. Espera-se que seja utilizado como uma referência para futuras implementações e para nortear outros projetos de QD, inserindo a área no mercado de forma progressiva.

1.4.Metodologia de Trabalho

Inicialmente foram levantadas e analisadas todas as fontes de informações possíveis, mas a única metodologia encontrada que apresenta todos os procedimentos descritos é a criada por Larry P. English denominada TQdM[7] (*Total Quality data Management)*. Por esse motivo essa metodologia foi escolhida como fundamento principal para a implementação do caso prático de QD.

Paralelamente ao estudo do material encontrado, o caso prático foi implementado em ambiente real de uma grande empresa da área financeira. A equipe formada por um especialista na ferramenta e três analistas de sistemas seniores, sendo um deles a autora deste trabalho, implementaram o projeto.

As informações provenientes dos sistemas transacionais, ambiente *mainframe*, foram transferidas para a plataforma do projeto, ambiente *Sun Solaris*, para serem tratadas pela ferramenta de reengenharia e limpeza dos dados, denominada *Integrity* da empresa Ascential. Os resultados deste projeto foram carregados numa base de dados *Oracle 8i*. Todos os processos foram documentados pela equipe utilizando como ferramentas *Visio, Word, Excel* e *Erwin.*

1.5.Organização do Trabalho

No segundo capítulo, são apresentadas algumas definições e procedimentos para elucidar o assunto sendo tratado e a evolução dos processos de melhoria contínua da qualidade. As metodologias são apresentadas de forma simplificada dando-se destaque à metodologia TQdM[7] que se encontra formalizada.

No terceiro capítulo, o principal processo da metodologia TQdM[7], denominado processo de reengenharia e limpeza de dados, é apresentado em detalhes.

No quarto capítulo são apresentados os objetivos do estudo de caso para o qual, dados os requisitos do problema, utilizou-se à metodologia TQdM[7], bem como o ambiente escolhido para implementação, incluindo a ferramenta de reengenharia e

os procedimentos utilizados. Também, no detalhamento dos processos, são descritos, de forma simplificada, os algoritmos que a ferramenta escolhida utiliza para realizar os processos. São relatados também, todos os problemas encontrados durante a implementação do estudo de caso.

No quinto capítulo, em decorrência da etapa anterior, são apresentadas e justificadas as melhorias sugeridas à metodologia TQdM[7] incluindo uma série de alterações, em particular, inclusão de novos processos, no sentido de sanar alguns dos problemas encontrados no ambiente prático. Apresenta-se um estudo comparativo entre a metodologia original e a proposta, identificando os novos processos e os já existentes. O aspecto “documentação” é aqui enfatizado, pela sua importância.

Concluindo, no sexto capítulo, apresenta-se um resumo do trabalho, as contribuições e as sugestões para futuros trabalhos, tendo como eventual ponto de partida o presente trabalho.

Capítulo 2 - Estado da Arte

2.1. Introdução

Esse capítulo tem como objetivo apresentar as diferentes visões relacionadas à qualidade dos dados, aos principais processos envolvidos na limpeza dos dados e a evolução dos processos de melhoria contínua da qualidade, até chegar nas metodologias existentes no momento.

2.2. Características da Qualidade

A qualidade é muito difícil de ser explicada e mensurada, pois ela não tem uma definição universal; ela varia dependendo do ponto de vista do observador. Por esse motivo à avaliação da qualidade de um produto ou serviço feita sem critérios, pode incorrer em resultados duvidosos e frustrar os seus consumidores.

Do ponto de vista de Wang[20], a qualidade de um produto depende do processo pelo qual o produto foi projetado e produzido, e dessa forma, a qualidade dos dados depende dos processos utilizados durante o projeto e produção para geração dos dados. E para obter a melhor qualidade é necessário primeiro compreender o que significa qualidade e como ela pode ser medida. Wang caracteriza a qualidade de dados como um conceito multidimensional, associando cada estado do mundo real a um estado específico nos sistemas de informação.

As possíveis deficiências de representação que ocorrem durante o projeto do sistema e a produção dos dados são utilizadas para definir as dimensões essenciais de qualidade de dados: Completude (*Complete*), Clareza (*Unambiguous*), Significado (*Meaningful*) e Corretude (*Correct*). Na tabela 2.1 são apresentadas as dimensões da qualidade de dados referentes aos dados, denominadas visões internas, e as relacionadas com a apresentação e utilização dos dados, denominadas visões externas.

Dimensões	
Visão Interna (Projeto, Operação)	**Relacionada aos dados** exatos, seguros, atualizados, completos, consistentes, precisos
	Relacionada ao sistema Confiável
Visão Externa (utilização, valor)	**Relacionada aos dados** Atualizados, relevantes, satisfatório, útil, claro, conciso, detalhado, imparcial, quantitativo, compreensível
	Relacionada ao sistema Atualizado, flexível, formatado, eficiente

Tabela 2. 1 - Visões Internas e Externas ao projeto

Do ponto de vista de Larry English[7], existem vários tipos de qualidade:

- Qualidade da definição: refere-se a dois aspectos que devem ser analisados:
  - Qualidade Inerente: está relacionada a corretude ou precisão dos dados. É o nível em que o dado reflete o objeto do mundo real correspondente. Todo dado é uma abstração ou representação de alguma coisa real.
  - Qualidade Pragmática: está relacionada ao valor do dado preciso na óptica da empresa. Essa qualidade pragmática é o nível de eficiência e eficácia com que o dado possibilita as pessoas com conhecimento possam ir de encontro aos objetivos da empresa. Dados que não são capazes de ajudar a empresa em sua missão não têm qualidade, não importam quão precisos eles sejam.

- Qualidade da arquitetura da informação: é o nível com que as estruturas de dados representam a herança e os relacionamentos a partir dos eventos e objetos do mundo real, estas estruturas devem ser estáveis para possibilitar que novas aplicações a reutilizem e flexíveis para suportar mudanças sem alterações significantes nos esquemas e bancos de dados.

- Satisfatoriedade dos dados: é o nível em que os dados precisos representam as características das entidades ou fatos do mundo real, e possibilitam a satisfação das necessidades dos usuários da informação para executar seus trabalhos de forma eficiente.

- Qualidade da apresentação dos dados: significa que os usuários com conhecimento podem facilmente e rapidamente compreender o significado da informação e aplicá-lo corretamente em seu trabalho.

De uma forma geral[21], os atributos mais citados como características da qualidade dos dados são:

- Precisão: Os dados representam precisamente a realidade;
- Integridade: As estruturas dos dados e seus relacionamentos são mantidos de forma consistente;
- Consistência: Os dados são definidos de forma coerente;
- Completude: Todos os dados necessários estão presentes;
- Validade. Os valores dos dados estão dentro das restrições definidas pelo negócio;
- Disponibilidade: Os dados estão disponíveis quando necessários;
- Acesso: Os dados são facilmente acessados, compreendidos e utilizados.

Analisando esses vários aspectos relacionados com a qualidade da informação, é possível dizer que a qualidade dos dados afeta diretamente a efetividade e a eficiência dos processos do negócio. Nota-se que é muito importante que as informações sejam de qualidade e que possam ser utilizadas de forma assertiva para auxiliar a empresa em assuntos estratégicos.

2.2. Limpeza dos dados

De acordo com o *Gartner Group*[22], os principais processos envolvidos na limpeza dos dados são:

- Elementarização (*Elementizing*): significa separar os dados em componentes, chamados “elementos”;
- Padronização (*Standardizing*): significa aplicar um formato padrão a esses elementos;
- Verificação (*Verifying*): significa examinar os elementos e procurar por erros;
- Identificação (*Matching*): significa detectar elementos idênticos, como por exemplo, endereços ou nomes iguais;
- *Householding*: identifica grupos de elementos que possuem características em comum, como por exemplo pessoas diferentes que residem em um mesmo endereço;
- Documentação (*Documenting*): significa capturar os resultados dos passos anteriores para facilitar futuros exercícios de limpeza de dados.

Para a empresa FirstLogic[11], há três estágios para a melhoria da qualidade dos dados: o primeiro é o processo de limpeza, o segundo é o processo de identificação e o terceiro é o processo de consolidação, conforme apresentado na figura 2.1.

A seguir cada um dos estágios relacionados com a melhoria da qualidade dos dados é apresentado em detalhes:

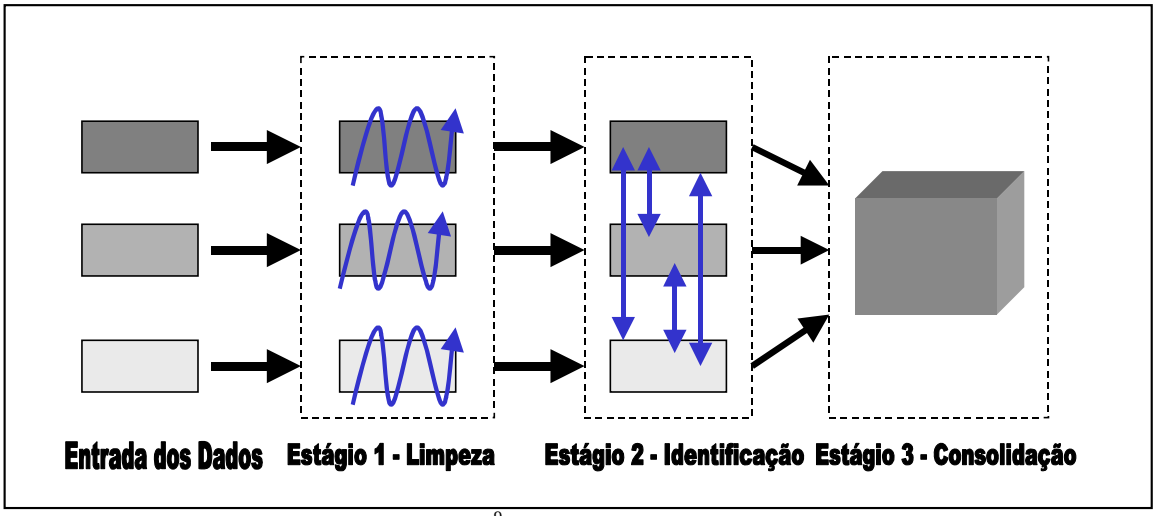

Figura 2. 1 - Estágios da Qualidade de Dados [9]

➢ Estágio 1 - Limpeza (*Cleansing*)

Durante este estágio, os dados são divididos em partes menores (*parsing*) para facilitar a correção, a padronização e a identificação dos dados, pois permite efetuar comparações de componentes individuais de forma mais eficaz do que utilizando o dado inteiro. A figura 2.2 apresenta um exemplo da saída produzida por este processo, neste exemplo o campo Nome foi divido em três campos menores denominados: Primeiro Nome, Nome do Meio, Ultimo Nome, Titulo, e campo Endereço foi separado em cinco campos denominados Endereço, Numero, CEP, Cidade e UF.

Dados Entrada
ANA MARIA SILVA DOUTORA
HOSPITAL SÃO PAULO
RUA PEDRO DE TOLEDO, 50
03434-010 SP SP

Dados Divididos	
Primeiro Nome	ANA
Nome Meio	MARIA
Último Nome	SILVA
Título	DOUTORA
Empresa	HOSPITAL SAO PAULO
Endereço	RUA PEDRO DE TOLEDO
Número	50
CEP	03434010
Cidade	SP
UF	SP

Figura 2. 2 - Exemplo de separação dos dados

Após a divisão dos dados, a próxima fase do processo de limpeza é a correção (*correcting*). A correção tem como objetivo ajustar problemas de discrepâncias de abreviações e formatos na entradas dos dados; ausência de letras causada pela similaridade fonética durante a entrada dos dados por telefone; informações desencontradas devido a mudanças de nome e endereço, entre outras ocorrências. A figura 2.3 apresenta um exemplo da correção do campo Cidade que foi preenchido com a forma abreviada SP e após o processo foi corrigido para SAO PAULO.

Dados Divididos	
Primeiro Nome	ANA
Nome Meio	MARIA
Último Nome	SILVA
Título	DOUTORA
Empresa	HOSPITAL SAO PAULO
Endereço	RUA PEDRO DE TOLEDO
Número	50
CEP	03434010
Cidade	SP
UF	SP

Dados Corrigidos	
Primeiro Nome	ANA
Nome Meio	MARIA
Último Nome	SILVA
Título	DOUTORA
Empresa	HOSPITAL SAO PAULO
Endereço	RUA PEDRO DE TOLEDO
Número	50
CEP	03434010
Cidade	SAO PAULO
UF	SP

Figura 2. 3 - Exemplo de correção dos dados

O passo seguinte é a padronização (*standardization*), que tem como objetivo comparar a representação do dado com uma representação padrão pré-estabelecida, reduzindo assim problemas de abreviações inconsistentes, títulos não utilizados, similaridade e variações de pronúncia, entre outros. A figura 2.4 apresenta alguns resultados deste processo, no exemplo, os campos Titulo e Título do Endereço foram padronizados para DRA e R, respectivamente.

Dados Corrigidos	
Primeiro Nome	ANA
Nome Meio	MARIA
Último Nome	SILVA
Título	DOUTORA
Empresa	HOSPITAL SAO PAULO
Endereço	RUA PEDRO DE TOLEDO
Número	50
CEP	03434010
Cidade	SAO PAULO
UF	SP

Dados Padronizados	
Título	DRA
Primeiro Nome	ANA
Nome Meio	MARIA
Último Nome	SILVA
Empresa	HOSPITAL SAO PAULO
Titulo_End	R
Logradouro	PEDRO DE TOLEDO
Número	50
Complemento	
CEP	03434010
Cidade	SAO PAULO
UF	SP

Figura 2. 4 - Exemplo de padronização dos dados

O passo final desse estágio de limpeza é o enriquecimento, que significa inserir novos dados e completar informações que faltavam. Esse processo pode ser efetuado por outra empresa, ou utilizando outra fonte de informação adicional, ou atualizando manualmente essas informações. A figura 2.5 apresenta um exemplo de enriquecimento utilizando os dados dos Correios, a partir do campo CEP equivalente foi obtido o campo Bairro.

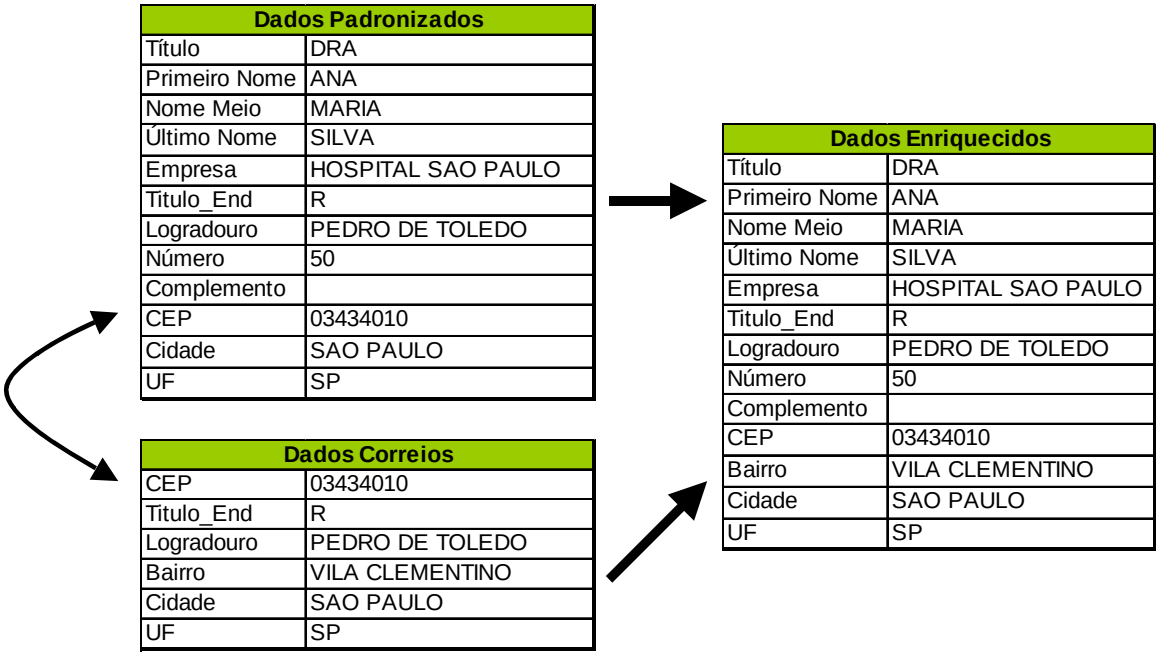

Figura 2. 5 - Exemplo de enriquecimento com Correios

- Estágio 2 - Identificação (*matching)*

O processo de identificação permite detectar dados similares dentre os dados da origem, eliminando duplicidades e consolidando as informações. Conforme apresentado no exemplo da figura 2.6, através deste processo, é possível identificar casos de duplicidades entre fontes de informações distintas mesmo que algumas informações não sejam exatamente equivalentes.

O processo de identificação envolve processos de comparações para identificar dados duplicados. Existem vários tipos de comparações que podem ser utilizadas[11]:

1. Identificação utilizando uma chave (*Key-Code Matching*)

   Esse processo efetua comparações por igualdade utilizando os primeiros caracteres de um ou mais campos. Este método primitivo raramente é praticado porque ele utiliza um pequeno grupo de caracteres, o que pode resultar em falsas duplicatas.

2. Identificação utilizando fonética (*Soundexing*)

   Esse processo detecta similaridade fonéticas das palavras em inglês, como “f” e “ph”. Esses tipos de erros ocorrem muitas vezes devido à entrada dos dados ter sido efetuada por telefone, e que não podem ser padronizados. Contudo, utilizar o som das palavras é inadequado como uma solução isolada porque somente podem ser detectados erros de fonética.

3. Identificação utilizando similaridade (*Similarity Matching ou Fuzzy Matching*)

   Esse processo pode identificar duplicidades computando um grau de confiança entre os dois componentes. Como não são utilizadas comparações por igualdade este processo pode ser utilizado nos casos de problemas de fonética, digitação, entre outros. Ele é considerado o melhor método de identificação, principalmente no caso de dados que não podem ser padronizados, como sobrenomes, nomes de negócios, e números de casas.

4. Identificação utilizando pesos (*Weighted Matching*)

   Esse processo pode ser utilizado em conjunto com a identificação utilizando fonética ou similaridade. Ele permite indicar a importância relativa dos campos atribuindo pesos aos mais importantes e esses valores são computados e utilizados durante o processo de identificação.

➢ Estágio 3 - Consolidação (*Consolidation*)

Depois que os dados duplicados foram identificados no estágio anterior, é possível gerar uma visão consolidada utilizando critérios pré-definidos. Depois de consolidados, os dados são transferidos para um *data warehouse*, *data mart*, ou outro repositório. A figura 2.7 apresenta um exemplo da consolidação dos dados da fonte A e da fonte B, neste caso, o critério adotado durante este processo de consolidação foi sempre obter como resultado o máximo de informações possíveis, por isso apesar da fonte A não possuir a informação Complemento preenchida, no resultado consolidado, este campo foi obtido da fonte B, o mesmo ocorrendo com o campo Titulo que não existia na fonte B, mas existia na fonte A e foi atribuído ao resultado consolidado.

Dados Entrada
ANA MARIA SILVA DOUTORA
HOSPITAL SÃO PAULO
RUA PEDRO DE TOLEDO, 50
03434-010 SP SP

Dados Limpos (FONTE A)	
Título	DRA
Primeiro Nome	ANA
Nome Meio	MARIA
Último Nome	SILVA
Empresa	HOSPITAL SAO PAULO
Titulo_End	R
Logradouro	PEDRO DE TOLEDO
Número	50
Complemento	
CEP	03434010
Bairro	VILA CLEMENTINO
Cidade	SAO PAULO
UF	SP

Dados Limpos (FONTE B)	
Título	
Primeiro Nome	ANA
Nome Meio	M
Último Nome	SILVA
Empresa	HOSPITAL SAO PAULO
Titulo_End	R
Logradouro	PEDRO DE TOLEDO
Número	
Complemento	CONJ 23 BLOCO A
CEP	03434010
Bairro	VILA CLEMENTINO
Cidade	SAO PAULO
UF	SP

Figura 2. 6 - Exemplo do processo de identificação

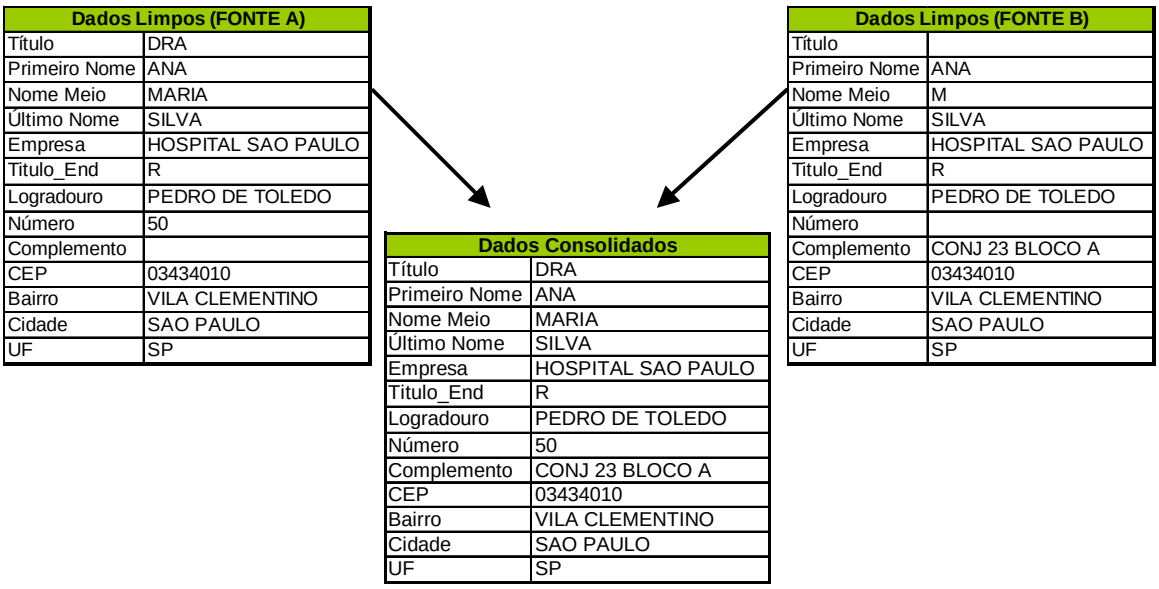

Figura 2. 7 - Exemplo do processo de consolidação

2.3. Ferramentas de QD

As ferramentas de qualidade de dados foram desenvolvidas para três finalidades[19]:

1. Análise da qualidade dos dados

   Estas ferramentas têm como objetivo encontrar anomalias nos dados, analisar a qualidade dos dados, descobrir informações específicas dos dados, suas estruturas e regras de negócio. Essas ferramentas auxiliam no processo contínuo de monitoramento e auditoria dos dados.

2. Identificação dos relacionamentos entre os dados

   Estas ferramentas têm como objetivo identificar ocorrências similares comparando as fontes de informação com fontes fidedignas, como por exemplo utilizando dados dos Correios visando aprimorar as informações e assim aumentar a eficiência no envio de correspondências, denominado “Eficiência do Contato”, ou comparando com outras fontes de informação

visando identificar relacionamentos entre as fontes, denominado "Identificação dos relacionamentos". A maioria das ferramentas de qualidade se enquadram nessa categoria e elas apresentam um rápido retorno do investimento.

3. Reengenharia de Dados

   Estas ferramentas são as mais flexíveis das três categorias, permitindo aos usuários reparar os dados e definir transformações. Possuem capacidade de preenchimento de dados faltantes e ajuste de dados.

De acordo com as categorias apresentadas anteriormente, o grupo *Gartner*[23] de 2002 apresenta, na tabela 2.2, os principais fornecedores de produtos para qualidade de dados e classifica as soluções de cada fornecedor em forte, moderado ou fraco.

Dados 2002				
Fornecedor	**Categoria**			
	Identificação dos relacionamentos entre os dados		**Análise da qualidade dos dados**	**Reengenharia**
	Eficiência do Contato	**Identificação dos relacionamentos**		
Ascential Software	Moderado	Moderado	Forte	Forte
Avellino			Forte	
DataFlux/SAS	Moderado	Moderado	Moderado	
Evoke Software			Forte	
FirstLogic	Forte	Moderado		
Group 1 Software	Forte	Moderado		
Innovative Systems	Forte	Moderado	Forte	Fraco
Trillium Software	Forte	Moderado	Fraco	Fraco

Tabela 2. 2 - Classificação das ferramentas em 2002

Analisando as informações apresentadas e comparando com dados de 1996, na tabela 2.3, é possível notar que os fornecedores estão preocupados em aprimorar as funções de suas ferramentas e expandir em outras finalidades, como por exemplo o fornecedor Innovative Systems que em 1996 somente atuava em ferramentas para identificação dos relacionamentos entre os dados e que em 2002 passou a atuar

também em soluções de análise da qualidade dos dados e de reengenharia, além de aprimorar as antigas funções.

Dados 1996				
Fornecedor	**Categoria**			
	Identificação dos relacionamentos entre os dados		**Análise da qualidade dos dados**	**Reengenharia**
	Eficiência do Contato	**Identificação dos relacionamentos**		
Ascential Software		Moderado	Forte	Forte
Group 1 Software	Forte	Fraco	Fraco	
Innovative Systems	Fraco	Fraco		
Trillium Software	Forte	Fraco		Fraco
i.d. Centric	Forte	Moderado	Fraco	
QDB Solutions			Forte	

Tabela 2. 3 - Classificação das ferramentas em 1996

2.4. Histórico e Estado da Arte

Segundo Larry English[12], a preocupação com a qualidade iniciou-se com a revolução industrial, onde a introdução de processos de produção em massa causaram a ocorrência de grande quantidade de defeitos. Observando esses problemas, os primeiros pensadores, como Deming, Juran, Crosby e Ishikawa, descobriram que esses processos poderiam ser melhorados reduzindo o número de defeitos através de implementação de mecanismos de controle. Somente no ano de 1991 surgiram as primeiras metodologias para melhoria da qualidade voltada aos dados. A seguir são apresentadas as principais contribuições dos estudiosos sobre o assunto que servem de fundamento para as metodologias hoje existentes.

2.4.1. W. Edwards Deming

Dr. W. Edwards Deming é conhecido como o pai do renascimento industrial do Japão no pós-guerra e foi respeitado por muitos como o principal “guru” de qualidade nos Estados Unidos. Sua filosofia pode ser resumida nos famosos “14 Pontos de Deming”[13] e no Ciclo PDCA[15] (*Plan-Do-Check-Act*) apresentados a seguir.

- **14 Pontos de Deming**[13]

Os 14 pontos de Deming são utilizados como princípios fundamentais para o gerenciamento da transformação visando mudanças fundamentais no negócio. Eles são os princípios necessários para mudar a mentalidade das pessoas em “fazer rápido” (*do it fast*) para “fazer corretamente” (*do it right*). A seguir cada um dos 14 pontos é apresentado:

1. Toda empresa precisa ter um plano para o futuro. As metas e as propostas da empresa para melhorar a qualidade da informação precisam estar claras para todos os empregados;
2. Uma nova atitude deve ser assumida para atingir os objetivos desejados. A mudança se faz necessária;
3. O objetivo é eliminar os defeitos na origem, melhorando o processo produtivo e não simplesmente inspecionando o produto final e identificando produtos de baixa qualidade;
4. A aquisição de um produto barato hoje pode gerar gastos extras no futuro com manutenção, por isso, a avaliação da qualidade de um produto não pode ser simplesmente baseada no seu preço;
5. O processo de melhoria deve ser contínuo, um processo sem fim que deve abranger todas as pessoas da empresa. Todos devem se sentir bem dentro desse processo;
6. Todas as pessoas envolvidas no processo devem ser treinadas para que possam atuar de forma correta dentro da sua área. As pessoas são os principais agentes do processo de qualidade e produtividade;
7. O gerenciamento deve ter como objetivo melhorar o processo e não verificar se as quotas numéricas foram atingidas. O objetivo do gerente é liderança e não supervisão;
8. Os profissionais devem estar tranqüilos e confiantes com relação ao seu trabalho, eles não podem estar preocupados com seus empregos, com o gerenciamento, com quotas a serem cumpridas, com os erros cometidos, entre outros;

9. Todos os departamentos da organização devem trabalhar juntos com o mesmo objetivo de obter a satisfação do cliente;
10. “Slogans”, “posters” e qualquer outra forma de divulgação de mensagens não resultam na melhoria da qualidade. Eles geram frustrações e ressentimentos por parte dos empregados, por isso devem ser evitados;
11. Quotas e objetivos numéricos utilizados no gerenciamento da produção devem ser eliminados por serem contraproducentes;
12. Se as pessoas se sentirem satisfeitas e orgulhosas com o seu trabalho, elas farão um bom trabalho;
13. A empresa precisa incentivar processos de auto-estudo, processos onde os profissionais se desenvolvam sozinhos;
14. Para que os 13 pontos descritos anteriormente aconteçam é muito importante que todas as pessoas estejam acompanhando a transformação. A transformação é um trabalho de todos.

- **Ciclo PDCA[15] (*Plan-Do-Check-Act*)**

Walter Shewhart propôs uma técnica para obter a qualidade da informação através de processos de melhorias, conhecido como “Ciclo de Shewhart”: os processos utilizados para a melhoria da qualidade deveriam ser cíclicos com uma etapa de inspeção provendo informações para a etapa de especificação de forma contínua.

Esse método foi adaptado por W. Edwards Deming em 1950, tornando-o conhecido como Ciclo de Deming ou ciclo PDCA. Conforme apresentado na figura 2.8, os passos envolvidos no ciclo PDCA, de forma simplificada, são:

- *PLAN* (Planejar) Projetar e revisar os componentes do processo de negócio para melhorar os resultados;
- *DO* (Executar) Implementar o plano e mensurar seu desempenho;
- *CHECK* (Avaliar) Avaliar os resultados e verificar se estão conforme o esperado;
- *ACT* (Atuar) Decidir e acionar as mudanças necessárias para melhorar o processo.

Posteriormente, Deming substituiu o "*CHECK*" por "*STUDY*" sugerindo uma análise de resultados mais criteriosa. Por isso uma alternativa de abreviação do ciclo de Deming seria PDSA[16] (*Plan, Do, Study, Act*) como apresentado na figura 2.9.

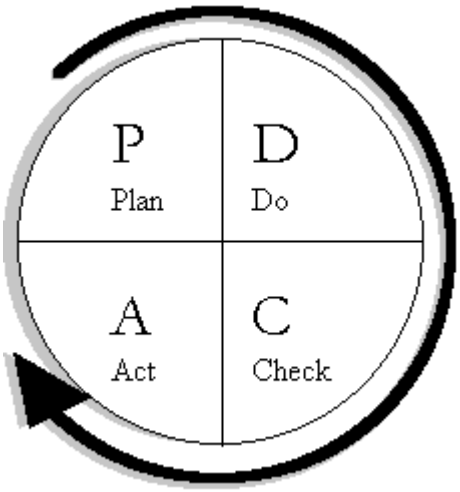

Figura 2. 8 - Ciclo PDCA[15]

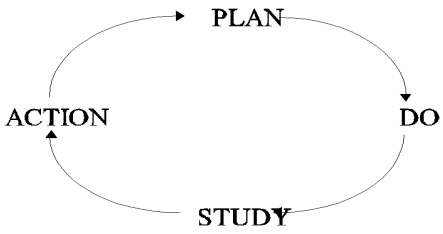

Figura 2. 9 - Ciclo PDSA[16]

Os 14 pontos de Deming e o ciclo PDCA foram utilizados como fundamento para outros pensadores evoluírem suas metodologias relativas à melhoria da qualidade dos dados.

2.4.2. Joseph M. Juran[2]

Juran focou seu trabalho na trilogia[8]: planejamento, controle e melhoria da qualidade. Esses três aspectos deram origem ao "Mapa de planejamento da

qualidade"[2] (*Quality Planning Road Map*). Na tabela 2.4 são apresentados os objetivos e as atividades relacionadas com cada um dos aspectos da trilogia.

Aspecto	Objetivo	Atividade
Planejamento da Qualidade	Prover o apoio operacional visando atender as necessidades dos clientes.	• Identificar quem são os clientes; • Determinar as necessidades desses clientes; • Planejar o que deve ser feito para atender as necessidades; • Desenvolver um produto que corresponda a essas necessidades; • Otimizar as características do produto para atender as necessidades dos clientes e as necessidades técnicas.
Melhoria da Qualidade	Melhorar os processos visando eliminar os erros e o retrabalho, além de aumentar a capacidade do produto estar adequado ao cliente de forma mais eficiente.	• Implementar um processo produtivo que seja capaz de produzir o produto desenvolvido; • Otimizar o processo de produção do produto;
Controle de Qualidade	Manter um nível específico de qualidade dentro do limite aceitável ou ao menos garantir que a qualidade não está ficando pior.	• Garantir que o processo pode produzir o produto em ambiente de produção; • Transferir o processo para produção.

Tabela 2. 4 - Mapa do planejamento da qualidade de Juran[2]

2.4.3. Philip Crosby[3]

Deming e Juran foram os grandes pensadores da revolução da qualidade, mas foi Crosby que popularizou o assunto. Ele conseguiu utilizando uma linguagem simples, disseminar a preocupação com o alto custo resultante da pouca qualidade dos dados. Crosby[10] definiu um programa de qualidade composto por 14 passos apresentados a seguir:

1. Compromisso da gerência: Esclarecer onde a gerência está na qualidade;
2. A equipe de melhoria da qualidade: Executar o programa de melhoria da qualidade;
3. Medidas de qualidade: Prover uma apresentação dos problemas atuais e potencias ocasionados pela não conformidade de forma que permita uma avaliação objetiva e uma ação corretiva;

4. O custo da qualidade: Definir os componentes do custo da qualidade e explicar seu uso como uma ferramenta de gerência;
5. Conscientização da qualidade: Prover um método para avaliar o interesse das pessoas, em toda a companhia, com a conformidade do produto ou serviço e a reputação da qualidade da empresa;
6. Ação corretiva: Prover um método sistemático para resolver para sempre os problemas que são identificados nas etapas precedentes a ação;
7. Planejar zero de defeito: Examinar as várias atividades que devem ser conduzidas durante a preparação do lançamento formal do programa zero de defeitos;
8. Treinamento do supervisor: Definir o tipo de treinamento que os supervisores necessitam a fim de implementar ativamente sua parte no programa de melhoria da qualidade;
9. Dia ZD: Criar um evento que permita que todos os empregados participem apresentando uma experiência pessoal de mudança;
10. Ajuste do objetivo: Assumir promessas e compromissos incentivando o indivíduo a estabelecer seu objetivo de melhoria para si e para seus grupos;
11. Eliminar a causa do erro: Criar um método para que o empregado possa comunicar a gerência às situações críticas que o independem de cumprir o compromisso de melhoraria;
12. Reconhecimento: Apreciar aqueles que participam do processo;
13. Conselhos de qualidade: Aproximar os profissionais de qualidade para uma comunicação do planejamento;
14. Repetir o processo: enfatizar que o programa de melhoria da qualidade nunca termina.

2.4.4. Kaoru Ishikawa[4]

Kaoru Ishikawa desejava que as pessoas pensassem na forma de trabalhar. Ele desejava o envolvimento dos gerentes; ele insistia em que a melhoria da qualidade poderia sempre estar um passo adiante, que a qualidade deveria estar presente no ciclo de vida do produto e não somente durante a produção. Ele expandiu

os quatro passos de Deming no Ciclo PDCA para seis como apresentado na tabela 2.5.

PLAN	1. Determinar objetivos e metas; 2. Determinar métodos para atingir objetivos;
DO	3. Dedicar-se à educação e treinamento; 4. Implementar o trabalho;
CHECK	5. Verificar os efeitos da implementação;
ACT	6. Executar a ação apropriada.

Tabela 2. 5 - Ciclo PDCA de Ishikawa

2.4.5. TQM[5] (*Total Quality Management*)

O gerenciamento da qualidade total (*Total Quality Management*) surgiu no ano 1950 e se tornou popular somente no ano 1980. TQM utiliza como fundamento os "14 Pontos de Deming" e sua premissa é que a produção de bens e serviços é um processo dinâmico que envolve operações e pessoas, que podem estar sujeitas ao processo de melhoria. Ele incorpora os conceitos de qualidade do produto, controle do processo, garantia da qualidade e melhoria do processo[17].

Para uma implementação de TQM com sucesso, uma empresa necessita concentrar-se em 8 elementos chaves:

1) Ética: é imprescindível discernir o certo e o errado em qualquer situação;
2) Integridade: implica em honestidade, moral, valores e outros aspectos considerados importantes pelo cliente;
3) Envolvimento: força a participação integral de todos os membros;
4) Treinamento: é muito importante para que os empregados sejam altamente produtivos e para que possam implementar o TQM dentro de seus departamentos;
5) Grupo de trabalho: propicia melhores soluções para os problemas em menor tempo, pois as pessoas se sentem mais confortáveis trabalhando em grupo;

6) Liderança: é o elemento mais importante do TQM; ela requer que os gerentes consigam guiar seus subordinados; para isso eles precisam estar comprometidos e acreditarem nas práticas diárias do TQM;
7) Reconhecimento: depois de terminado o processo todos necessitam receber elogios e ter o reconhecimento do trabalho efetuado corretamente. Esse reconhecimento pode ser feito de diversas formas, em lugares e tempos diferentes;
8) Comunicação: é imprescindível para que todo o processo dê certo. Ela serve de elo entre o que se deseja e o que foi feito.

2.4.6. TDQM[14] (*Total Data Quality Management*)

O programa de Gerenciamento Total da Qualidade dos Dados criado no MIT (*Massachusetts Institute of Technology*), no ano de 1991 por Dr. Richard Wang em conjunto com Stuart Madnick, tem como objetivo, em longo prazo, a criação de uma teoria de qualidade de dados baseada em disciplinas de ciência da computação, comportamento organizacional, estatística, contabilidade e gerenciamento de qualidade total. Em curto prazo, o objetivo é criar um centro de excelência entre os praticantes de técnicas de qualidade de dados e atuar como um laboratório para métodos eficazes e experiências de projetos.

A metodologia TQDM adota como fundamento o Ciclo de Deming[6] como apresenta a figura 2.10, onde as quatro componentes do programa TQDM são: **definição** (*Define*), **medição** (*Measure*), **análise** (*Analyze*) e **melhoria** (*Improve*). As duas primeiras componentes focam a definição e medição da qualidade dos dados, respectivamente, a análise identifica e calcula o impacto da baixa qualidade de dados e os benefícios de alta qualidade e a melhoria envolve o replanejamento de práticas de negócio e implementação de novas tecnologias para melhorar significamente a qualidade da informação corporativa.

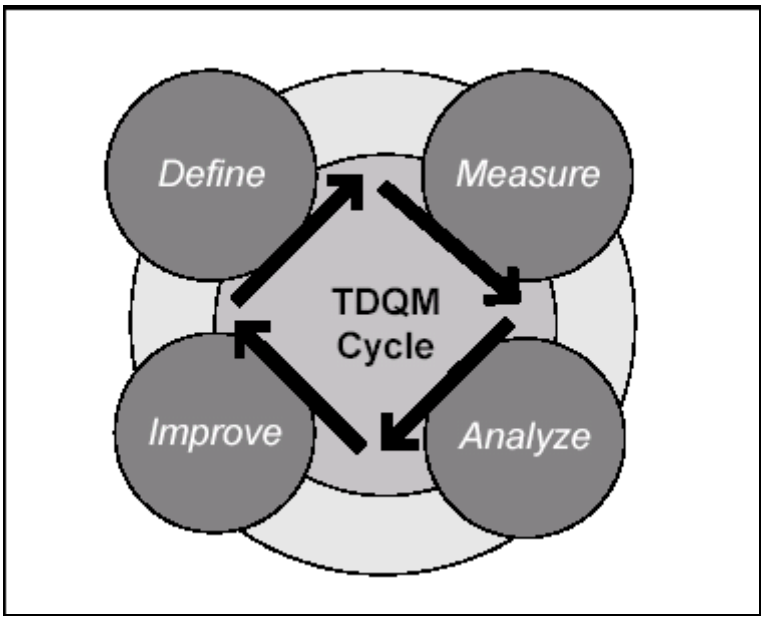

Figura 2. 10 - Ciclo TQDM[6]

2.4.7. TQdM[7] (*Total Quality data Management*)

A metodologia TQdM (*Total Quality data Management*) foi à metodologia adotada para a implementação do estudo de caso por ser a única a apresentar com detalhes todas as etapas de um projeto de qualidade de dados.

Essa metodologia foi desenvolvida por Larry English e considera que um projeto de qualidade de dados está além de um processo de melhoria dos dados ou limpeza dos dados. Considera que todos na empresa estão relacionados pela informação, e por isso, a qualidade da informação é de grande valor para a empresa, e que a satisfação do cliente é o resultado da informação de qualidade, e conseqüentemente, todos devem ser envolvidos e ter responsabilidades no processo de melhoria contínua, procurando assim integrar os princípios e métodos à cultura da empresa.

De uma forma geral, o processo de melhoria da qualidade da informação da metodologia TQdM pode ser representada por um plano de ação, denominado processo 6, e 5 processos gerais ilustrados na figura 2.11 e descritos logo a seguir.

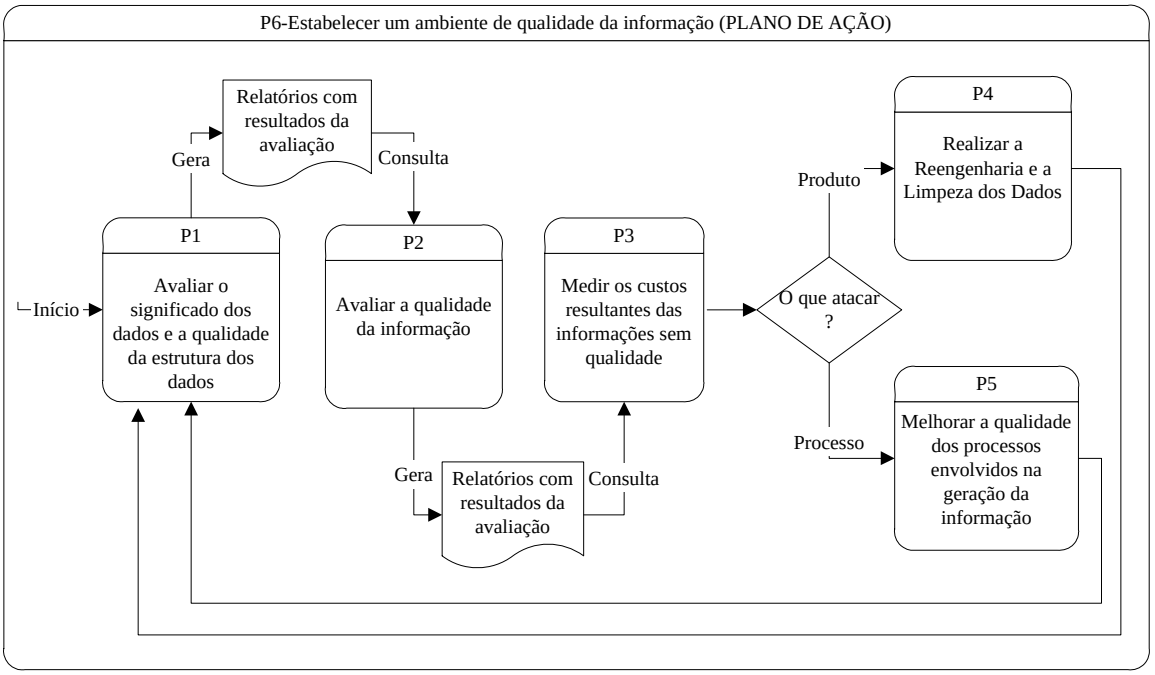

Figura 2. 11 - Metodologia TQdM

Processo 6: Estabelecer o ambiente de qualidade da informação (Plano de Ação)

Este processo representa o sistema, a direção e requisitos culturais para sustentar um ambiente de melhoria de qualidade da informação. Sem essas características culturais, uma iniciativa de qualidade da informação pode alcançar alguns benefícios a curto prazo, mas esses benefícios não são contínuos.

Este processo é o primeiro porque ele é fundamental para a melhoria da qualidade da informação á longo prazo. Este processo representa um plano de ação para implementar os processos de qualidade, servindo de suporte para todos os demais processos. É necessário romper os paradigmas. É necessário compreender o fluxo da informação, identificar quais os usuários da informação e o que eles esperam, educar as pessoas envolvidas para melhorar os processos do dia-a-dia para a própria eficiência e efetividade de suas atividades, incentivar a satisfação do cliente como medida de performance.

Os principais passos desse plano de ação são:

1. Identificar o nível de maturidade da organização em QD dentro de uma matriz similar ao CMM (*Capability Maturity Model*), conforme apresentada na tabela 2.6, para conhecer a situação inicial da empresa;

Tipo de Análise	Estágio 1: Incerteza	Estágio 2: Despertar	Estágio 3: Esclarecimento	Estágio 4: Sabedoria	Estágio 5: Segurança
	Estágio CMM: Inicial	Estágio CMM: Repetitivo	Estágio CMM: Definido	Estágio CMM: Gerenciado	Estágio CMM: Otimizado
Gerenciamento do conhecimento e da atitude	Nenhuma compreensão da qualidade da informação como uma ferramenta de gerenciamento. Tenta culpar a administração dos dados pelos problemas na qualidade da informação.	Reconhece que o gerenciamento da qualidade da informação pode ser de valor, mas não provê dinheiro e nem tempo para fazer isso acontecer.	Enquanto desenvolve programas de melhoria da qualidade aprende mais sobre gerenciamento da qualidade, tornando-o suportável e útil	Compreende o gerenciamento da qualidade da informação. Reconhece a sua função própria de continuar dando ênfase no assunto.	Considera o gerenciamento da qualidade da informação como parte essencial do sistema da companhia.
Situação da qualidade da informação da organização	A qualidade dos dados está escondida nos departamentos de desenvolvimento das aplicações. Os dados auditados provavelmente não fazem parte da organização. Ênfase na correção dos dados ruins.	A função principal da qualidade da informação é "identificar" mas a ênfase continua sendo em corrigir os dados ruins.	Qualidade da informação na organização existe, todos as avaliações estão incorporadas e o gerente tem papel no desenvolvimento de aplicações.	Gerentes da qualidade da informação reportam para o CIO da empresa: situação efetiva e ações preventivas. Envolvimento com áreas de negócio.	Gerente da qualidade da informação é parte do grupo de gerenciamento. Prevenção é o principal foco. Qualidade da informação é pensamento do líder.
Tratamento dos problemas de qualidade da informação	Problemas são combatidos conforme eles ocorrem, não são resolvidos, definições inadequadas, várias brigas e discussões.	Grupos são formados para atacar os principais problemas. Soluções de grande impacto não são solicitadas.	Ações corretivas são comunicadas. Problemas são mostrados abertamente e resolvidos de forma organizada.	Problemas são identificados antecipadamente no seu desenvolvimento. Todas as funções são abertas para sugestões e melhorias.	Exceções na maioria dos casos incomuns, problemas de qualidade da informação são prevenidos.
Custo da qualidade da informação como percentual da receita.	Reportado: Desconhecido	Reportado: 5%	Reportado: 10%	Reportado: 8%	Reportado: 5%
	Atual: 20%	Atual: 18%	Atual: 15%	Atual: 10%	Atual: 5%
Ações para melhoria da qualidade da informação	Nenhuma atividade organizada. Nenhum conhecimento de quais atividades.	Obvio é realizado. Esforço pequeno para motivação	Implementação dos 14 pontos do programa com conhecimento profundo e estabelecendo cada passo.	Continuação do programa dos 14 pontos e começando a otimizar	Melhoria da qualidade da informação é uma atividade normal contínua.
Conclusão da Postura da companhia sobre a qualidade da informação	"Nós não sabemos porque nós temos problemas com a qualidade da informação" "Nós temos problemas com a qualidade da informação?"	"Isto é absolutamente necessário sempre que temos problemas com a qualidade da informação?"	"Através do gerenciamento do compromisso e da melhoria da qualidade da informação nós estamos identificando e resolvendo nossos problemas."	"Prevenir problemas da qualidade da informação é uma parte da rotina de nossa operação"	"Nós sabemos porque nós não temos problemas com qualidade da informação"

Tabela 2. 6 – Modelo de maturidade da gerência da qualidade da informação[7]

2.Planejar os objetivos para o gerenciamento da informação e para melhoria da qualidade da informação visando descrever o possível estado futuro da empresa e que problemas serão resolvidos;
3.Identificar e nomear um líder da qualidade da informação para tomar alguma atitude e iniciar o processo;
4.Conduzir uma pesquisa de satisfação dos clientes com as informações para encontrar as frustrações e barreiras resultantes da informação sem qualidade;
5.Identificar outras transformações, iniciativas de melhorias ou recursos externos de aprendizado. Necessidade de ser capaz de unir essas iniciativas com os processos e aumentar sua probabilidade de sucesso;
6.Selecionar uma área pequena, gerenciável e com alto retorno para conduzir um projeto piloto;
7.Definir um problema do negócio a ser resolvido e as métricas para o sucesso do projeto de melhoria da qualidade da informação;
8.Definir a cadeia de valores da informação e desenvolver um inventário pequeno e seleto de informações críticas;
9.Desenvolver critérios para avaliação da qualidade das informações críticas;
10.Calcular o valor agregado para o cliente, se possível, para estimar a ausência e perda de oportunidades resultantes das informações sem qualidade;
11.Analisar informações relacionadas com as queixas e os problemas dos clientes;
12.Quantificar os custos resultantes dos problemas de qualidade dentro do conjunto de informações críticas do projeto piloto;
13.Identificar e desenvolver afinidade e comunicação pessoal com os responsáveis e proporcionar orientação consciente;
14.Definir informações de controle e regras de qualidade;
15.Definir princípios, processos e objetivos da qualidade da informação;
16.Analisar as barreiras sistemáticas para qualidade da informação e recomendar mudanças;
17.Conduzir uma avaliação da maturidade do gerenciamento da qualidade da informação com a alta gerência e providenciar educação formal;

18. Conduzir um projeto de melhoria da qualidade da informação e quantificar os benefícios alcançados comparando com o estado original;
19. Estabelecer um mecanismo regular de comunicação e educação com a alta administração para sustentar seu envolvimento e compromisso;
20. Continuar melhorando os processos de melhoria da qualidade da informação.

Processo 1: Avaliar o significado dos dados e a qualidade da estrutura dos dados

Não é possível medir a qualidade de um produto sem conhecer se as especificações do produto estão corretas e são o que deveriam ser. Para medir a qualidade da informação em um banco de dados ou fora de um processo, primeiro é necessário analisar a definição dos dados.

O primeiro passo é definir as características essenciais e críticas na definição dos dados e da estrutura dos dados, sendo que, a estrutura dos dados representa os projetos dos modelos de dados e bancos de dados. Esses requisitos mínimos incluem nomes dos dados, definições, conjuntos de valores válidos, e regras de negócios pertinentes.

A seguir, é necessário identificar um grupo de informações importantes para avaliar, um grupo de informações onde a baixa qualidade resulte em altos custos ou conseqüências inaceitáveis. Depois identificar as categorias dos principais envolvidos com a informação desse grupo, como os geradores da informação, pessoas que criam e mantêm a informação; especialistas que utilizam a informação; clientes e envolvidos externos que requisitam as informações.

O próximo passo é executar uma avaliação técnica das definições dos dados para verificar se estão em conformidade com os padrões, se possuem o mínimo exigido. Deve ser efetuada uma avaliação da qualidade da arquitetura da informação, ou seja, o projeto e implementação do banco de dados contra as melhores práticas com essa finalidade. O passo mais importante é medir a satisfação do cliente com a definição da informação com base na opinião dos especialistas que utilizam os dados e os produtores da informação que criam e mantêm os dados.

Processo 2: Avaliar a qualidade da Informação

Este processo verifica e analisa a qualidade da informação da mesma forma como é feito para um produto manufaturado. Há avaliações técnicas, através de análise de amostras, para verificar se os produtos gerados estão em conformidade com as especificações.

Há várias formas para medir o grau de satisfação do cliente de acordo com suas expectativas. O primeiro passo é identificar ou revisar o grupo de informações a serem mensuradas. Depois estabelecer as características da qualidade da informação para serem medidas em um grupo de dados, como completitude, conformidade com as regras de negócio e precisão. A seguir identificar as várias categorias de envolvidos no grupo da informação, como bases de dados que armazenam os dados, todos os processos e aplicações que criam, atualizam, e transformam ou extraem os dados, bem como todos os processos que recebem e utilizam a informação. Identificar quais bases de dados ou processos serão mensurados, identificar as origens de validação dos dados e comparar com a precisão dos dados sendo medidos, extrair uma amostra aleatória dos dados a serem analisados, medir a qualidade da informação e encontrar o melhor caminho para comunicar os resultados aos envolvidos.

Processo 3: Medir os custos resultantes das informações sem qualidade

Um dos mitos da melhoria da qualidade da informação é que os custos da baixa qualidade da informação não podem ser avaliados. Inicialmente é necessário identificar as diretivas do negócio, como aumentar os lucros, a satisfação do cliente ou reduzir os custos, depois analisar os custos da informação, como infra-estrutura para capturar e manter a informação, como o desenvolvimento de aplicações para receber e utilizar as informações como valor agregado, etc. A seguir, determinar os custos da empresa resultantes da ausência e imprecisão dos dados. Identificar os segmentos do cliente e calcular o valor do ciclo de vida do cliente utilizado para medir a ausência ou perda de oportunidade devido à baixa qualidade da informação. Enfim estabelecer o valor da qualidade da informação pela medição das perdas ou

ausências de oportunidades resultantes da não existência de qualidade nos produtos da informação.

Processo 4: Realizar a reengenharia e limpeza dos dados

É o processo para melhorar o produto da informação (sintomas), ou seja, as informações utilizadas pelo usuário final. Este processo é muito importante. Ele serve para transformar os dados defeituosos em dados com um nível aceitável de qualidade. O dado é limpo e retrabalhado.

Inicialmente, identifica-se a origem dos dados que requerem tratamento, a seguir é feita a extração e análise dos dados originais para identificar anomalias e padrões. Depois os dados são padronizados para que possam ser comparados, mesclados, ter suas duplicidades identificadas, etc. Os dados analisados são corrigidos e complementados utilizando processos automáticos ou manuais. O passo seguinte é identificar registros duplicados e consolidar as informações para um único registro utilizando um conjunto de algoritmos disponíveis na ferramenta a ser utilizada. Depois os dados são extraídos e analisados para identificação de alguns padrões de erros que podem ser utilizados para melhorar o processo.

Os demais passos são utilizados para corrigir e enriquecer os dados antes de retornarem para as bases de dados ou data warehouse; são utilizados para mapear os dados limpos para a estrutura de dados destino, sumarizar e derivar os dados, auditar e controlar a extração, transformação e carga dos dados.

Processo 5: Melhorar a qualidade dos processos envolvidos na geração da informação

Este processo verifica os problemas encontrados durante a fase de avaliação da qualidade, analisa os motivos, planeja e implementa processos de melhorias que previnem defeitos, age na causa do problema, ele deveria ser o mais utilizado processo de qualidade da informação, porém devido ao alto custo de implementação existem poucas empresas que realizam esse processo de melhoria. Ele utiliza o ciclo de Shewhart, conhecido como PDCA, para implementar os ciclos de melhoria dos processos.

O primeiro passo deste processo estabelece o problema a ser resolvido, identifica as atividades pertinentes, e estabelece o time de melhoria da qualidade da informação. Depois desenvolve o plano para a melhoria, que inclui identificar as mudanças específicas que precisam ser feitas em um ambiente controlado. Implementa as mudanças corretivas em ambiente controlado, podendo então testar as melhorias. Avalia a efetividade das melhoria das mudanças efetuadas, se a melhoria desejada não foi obtida, então as melhorias são re-implementadas ou re-planejadas. Depois torna efetiva as melhorias e implementações dentro da empresa para transformá-las em um padrão e para produzir resultados de qualidade consistentes.

2.5. Conclusão

Este capítulo apresentou conceitos importantes para compreensão da qualidade de dados, apresentou a evolução dos processos de melhoria contínua até atingir as metodologias existentes para implantação de processos de qualidade de dados. Dentre todas as metodologias apresentadas, somente a TQdM possui detalhamento das fases de desenvolvimento, pois não foi encontrado nenhum registro mais detalhado sobre as demais. O processo de reengenharia e limpeza dos dados da TQdM é o processo que possui a maior importância em um projeto de qualidade de dados, pois é responsável pelo tratamento da informação gerando uma informação de qualidade, e durante a sua implementação, é possível detectar problemas resultantes de falhas de definição na entrada dos dados que poderiam ser ajustadas, por isso esse processo será detalhado no próximo capítulo.

Capítulo 3 – Reengenharia e Limpeza dos Dados (Processo 4)

3.1. Introdução

Este capítulo apresenta a visão de Larry English[7] sobre o processo de reengenharia e limpeza (Processo 4) da metodologia TQdM. Este processo é utilizado como base no estudo de caso e as modificações que se fizerem necessárias ao aplicá-lo são apresentadas nos capítulos 4 e 5.

Na figura 3.1 estão todos os sub-processos utilizados para tratamento da informação, desde o estabelecimento da prioridade das fontes até a disponibilização dos dados e auditoria do processo.

Identifica-se o sub-processo 4.1 como sendo a primeira atividade a ser realizada, que consiste na seleção das fontes mais adequadas para aplicação do processo de qualidade de dados. Selecionadas as fontes, os dados serão extraídos no sub-processo 4.2 e, em seguida, submetidos ao sub-processo 4.3 (padronização), que consiste na definição e aplicação de padrões aos dados, como por exemplo, a equalização das palavras AVENIDA, AVE, AVEN para AV na informação endereço. Padronizada a informação, o sub-processo 4.4 efetuará as correções ou complementações necessárias, como por exemplo, complementação ou correção da informação sexo utilizando-se como referência o nome de uma pessoa. Após a correção, o sub-processo 4.5 consolidará a informação eliminando as possíveis duplicidades. A transformação e o enriquecimento da informação, se necessário, utilizando uma fonte de dados confiável, como por exemplo, dados dos Correios, serão implementados no sub-processo 4.7. Derivações ou sumarizações são obtidas com a aplicação do sub-processo 4.8, e finalmente, os processos de extração, transformação e disponibilização de dados serão auditados e controlados pelo processo 4.9. O processo 4.6 é responsável por identificar e analisar a existência de alguns problemas detectados nos dados durante a implementação dos processos iniciais (P4.2, P4.3, P4.4 e P4.5).

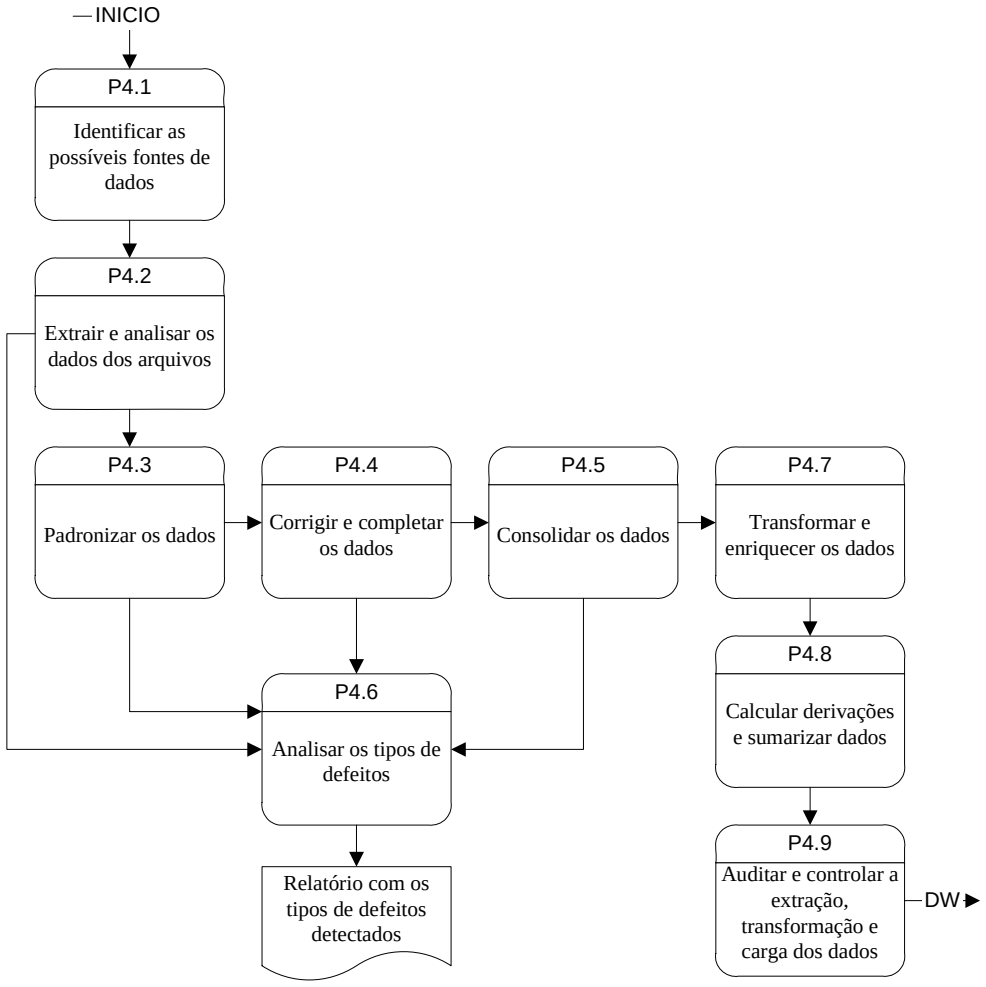

Figura 3. 1 - Metodologia TQdM - Processo 4 – Reengenharia e Limpeza dos Dados[7]

Nos itens a seguir, para cada um dos sub-processos serão apresentados: o seu objetivo, as informações de entrada a serem utilizadas, o detalhamento das atividades envolvidas e as saídas produzidas.

3.2. Identificar as possíveis fontes de dados (P4.1)

- **Objetivo**

O propósito deste processo é criar uma base de referência com as possíveis fontes de dados (bases de dados, arquivos textos, planilhas, entre outros), onde a melhoria da qualidade da informação enriquecerá o desempenho do negócio.

- **Processo**

Este processo identifica quais grupos de informação são confiáveis dentre as várias fontes identificadas e estabelece as regras de seleção e os critérios para sanar possíveis empates na escolha da melhor informação.

As entradas desse processo são:

- Grupos de informações relacionados ao negócio a serem utilizados no agrupamento das referências;
- Inventário de todas as fontes de dados a ser utilizado como referências;
- Cadeia de valores de informação, obtida durante a avaliação dos custos (Processo 3), a ser utilizada como critério na escolha da melhor referência.

As atividades envolvidas durante esse processo são:

- Identificar todas as fontes, operacionais ou não, que contenham todas as informações do negócio da empresa;
- Identificar computadores ou fontes de dados departamentais que possuam informações;
- Levantar as fontes pessoais que possam servir de referência com o principal especialista da informação;
- Documentar os processos e a utilização das fontes analisadas;
- Gerar uma referência de todas as fontes identificadas;
- Eleger o principal registro de referência das fontes utilizando-se dos seguintes critérios:
    - O processo que possui o maior interesse na exatidão de um dado é normalmente o mais confiável. Por exemplo, o departamento de Contas a Receber de uma empresa tem um grande interesse na efetividade do endereço da fatura, pois se este estiver errado, o pagamento poderá não ser efetuado. Logo, o endereço de correspondência dessa fonte de informação seria a mais confiável.

- Os dados que são recentemente atualizados são, geralmente, mais precisos do que os dados mais antigos. Contudo, dados atualizados por pessoas que utilizam os dados são mais precisos do que os dados atualizados por pessoas que não os utilizam, ou simplesmente, utilizam o volume de atualização dos dados como medida de produtividade.
- Conduzir uma avaliação da qualidade da informação sobre as várias fontes para descobrir quais possuem a mais alta qualidade da informação para cada grupo de campos, segundo o autor, é a forma mais efetiva.

- Na ausência de um registro de referência para um grupo de informações, identificar as principais fontes e utilizar os processos 4.2 e 4.3 para determinar a melhor.

➢ **Saídas**

- Lista das principais fontes de dados para cada grupo de informações;
- Lista de redundância das fontes de dados para cada grupo de informações.

3.3. Extrair e analisar os dados dos arquivos fontes (P4.2)

➢ **Objetivo**

O propósito deste processo é extrair os dados dos arquivos fontes e verificar se o conteúdo está condizente com a definição formal do campo. Caso existam informações incompatíveis, efetuar a estruturação de novos campos. Larry English[7] denomina esses dados de “escondidos” (*hidden*).

➢ **Processo**

Este processo extrai os dados das fontes e analisa-os confirmando se estão condizentes com a definição estabelecida. Quando alguma inconformidade é detectada, analisa-se o significado e a maneira como o dado está sendo utilizado.

As entradas desse processo são:

- Lista das principais fontes de dados para cada grupo de informações ;
- Lista de redundância das fontes de dados para cada grupo de informações.

As atividades envolvidas durante esse processo são:

- Extrair as amostras de dados das fontes. A amostra precisa ser o mais representativa possível, ou seja, abranger o maior número possível das anomalias;
- Afinar a granularidade dos dados, reduzindo-os aos fatos. Por exemplo, separar o nome das pessoas em diferentes componentes do campo Nome, como Primeiro Nome, Nome do Meio e Último Nome;
- Analisar o significado da informação com os especialistas das fontes e definir o significado dos novos atributos e tipos de entidades descobertas;
- Documentar a definição, os conjuntos de domínios de valores e as regras de negócio relacionadas com os novos atributos;
- Mapear e documentar os relacionamentos entre os novos atributos e as fontes originais.

Os dados denominados “escondidos” precisam ser encontrados, separados e definidos por três razões:

- Garantir que todos os atributos requeridos para o DW (*Data Warehouse*) foram identificados;
- Compreender o objetivo e a utilização dos dados visando transformá-los e migrá-los corretamente para o DW;
- Identificar os domínios e as regras de negócio da origem dos dados que são utilizados durante a transformação do processo 4.6.

- **Saídas**
  - Identificação de novos atributos que necessitarão de estruturação;
  - Identificação e definição de regras de negócio para cada novo atributo definido;
  - Anomalias de dados a serem corrigidas nos próximos processos.

3.4. Padronizar os dados (P4.3)

- **Objetivo**

O objetivo é estabelecer uma única forma padrão para representação dos dados na empresa, eliminando problemas de formatos inconsistentes e representações dúbias.

- **Processo**

Este processo estabelece uma forma padrão de representação e ajusta os dados para esse padrão. A correção efetuada nos dados, durante esse processo, visa apenas padronizar o conteúdo dos dados e não tem como objetivo complementar, alterar ou corrigir o significado do dado.

As entradas desse processo são:

- Novos atributos identificados com domínios a serem compatibilizados;
- Regras de negócio a serem aplicadas durante a padronização;
- Problemas na qualidade dos dados relacionados aos padrões.

O processo de padronização é aplicado para ajustar os seguintes casos:

- Domínios de valores redundantes

  Este processo visa uniformizar valores de acordo com um padrão pré-estabelecido. Por exemplo, o campo Estado Civil pode ser representado como SOLTEIRO em uma fonte, em outra, como SOL ou SOLT; enfim, o processo deve estabelecer uma única forma para representar o mesmo Estado Civil.

- Formatos inconsistentes

  Este processo visa ajustar formatos inconsistentes como, por exemplo, o campo Código Postal, representado em alguns casos como 02323-089, em outros como 02323900 ou 2323-899; o processo deve determinar uma única formatação.

- Níveis de granularidade distintos

  Este processo estabelece o menor nível de granularidade para os campos, ou seja, decompõe os componentes de um determinado campo em campos distintos visando analisar o significado de cada um. Por exemplo, a informação Endereço, em uma fonte de informação, pode possuir os campos Logradouro e Complemento, e em outra fonte pode possuir os campos Logradouro, Número e Complemento; neste caso, o objetivo é equalizar a informação no mesmo nível de granularidade.

- Várias informações em um mesmo campo

  Uma prática comum a ser corrigida durante esse processo é a presença de vários significados dentro de um mesmo campo. A solução é separar cada significado em um novo atributo, facilitando as futuras consultas e análises. Por exemplo, no caso de um campo Produto contendo as informações SABONETE (HIGIENE PESSOAL) e MACARRAO (COMESTÍVEL), nota-se que o tipo do produto está sendo cadastrado no mesmo campo destinado ao nome do produto. Para resolver este problema, a solução seria incluir um novo campo denominado Tipo do Produto, com domínio HIGIENE PESSOAL e COMESTÍVEL, e eliminar esta informação do campo Produto.

As atividades envolvidas durante a padronização dos dados são:

1. Identificar os atributos comuns a todas as fontes pertinentes;
2. Identificar e selecionar os especialistas da informação que utilizam determinados grupos de dados;
3. Eleger um representante oficial da empresa para cada grupo de dados;
4. Obter um consenso sobre a definição dos dados, as regras do negócio e os valores padrões dos dados;
5. Documentar os resultados obtidos sobre a definição dos dados;

6. Validar a qualidade da definição do dado e estrutura da informação com relação ao critério definido no processo 5 (Qualidade da Especificação do Produto da Informação);
7. Documentar o mapeamento dos dados originais para os dados padronizados;
8. Documentar as regras de transformação dos dados para os tipos de entidades e atributos padronizados.

- **Saídas**
  - Dados definidos e padronizados, incluindo os tipos de entidades, atributos, conjuntos de valores e regras de negócio;
  - Mapeamento dos dados da fonte para os dados padronizados.

3.5. Corrigir e Completar os dados (P4.4)

- **Objetivo**

Este processo tem como objetivo melhorar a qualidade do dado existente, atuando na correção de imprecisões ou valores incorretos, encontrando e capturando valores de dados incompletos. Este processo altera o conteúdo dos dados, diferindo do processo anterior (P4.3), que apenas estabelece um padrão único para o mesmo grupo de dados.

- **Processo**

Este processo procura identificar os valores corretos para cada grupo de dados e aplicá-los nos casos de valores incompletos, incorretos ou aparentemente inapropriados, denominados por Larry English[7] como dados suspeitos.

Dados suspeitos são dados não conhecidos com exatidão, isto é, não “parecem“ certos. Esses dados podem estar fora de uma distribuição normal ou fora dos valores de um domínio, como por exemplo, o campo Data de Nascimento preenchido na maioria das vezes com valor 01/01/1900; dados com valores acima ou abaixo do esperado, como por exemplo, um campo de Salário com conteúdo R$1,00;

dados com valores duplicados quando valores únicos são esperados, como por exemplo, os campos CPF ou CGC.

As entradas para esse processo são:

- As fontes de dados a serem tratadas durante o processo de correção;
- Definição dos dados e documentação das regras de negócio;
- Mapeamento dos dados originais para os tipos de entidades e atributos padronizados;
- Anomalias a serem corrigidas.

É válido considerar que as correções:

- Não poderão ser aplicadas aos dados cujas definições não foram registradas ou se perderam ao longo do tempo;
- Poderão ser realizadas de forma automática utilizando-se ferramentas apropriadas;
- Em alguns casos requerem correções manuais realizadas pelos especialistas, tornando o processo mais dispendioso.

As atividades envolvidas durante a complementação e correção dos dados são:

a. Identificar a ausência dos dados, embora muitas vezes, a ausência do dado não significa simplesmente conteúdo vazio ou nulo; valores *default* podem ser utilizados para indicar a ausência de informação;
b. Identificar os dados incorretos e suspeitos;
c. Determinar a maneira para aplicar o processo de correção:
    - Correção automática, utilizando uma ferramenta para efetuar comparações com uma fonte confiável, como por exemplo, a dos Correios, para complementar ou corrigir as informações;
    - Correção manual, no caso de necessidade de intervenção humana;
    - Correção híbrida (automática e manual).

d. Determinar a ordem de prioridade na escolha de qual conjunto de informações deve ser corrigido primeiro, baseando-se nos custos envolvidos no caso de decisões erradas devido a imprecisões ou ausência de dados;
e. Selecionar as informações a serem corrigidas para cada fonte ;
f. Documentar, em um repositório, o tipo de limpeza a ser adotado para cada tipo de dado;
g. Determinar como manipular um dado incorreto ou suspeito, como por exemplo, rejeitando-o no processo de qualidade e excluindo-o da origem; aceitando-o sem mudanças e sem documentação; aceitando-o sem mudanças mas documentando-o como suspeito, entre outros.
h. Documentar os casos importantes que não serão possíveis corrigir ou aqueles cujos valores serão aproximados;
i. Criar um atributo para codificar os resultados do processo de correção;
j. Levantar e documentar o tempo e os custos envolvidos no processo de limpeza. Esses custos envolvem o tempo despendido para desenvolver rotinas de transformação, custos da ferramenta de limpeza dos dados, tempo gasto para investigar e atualizar valores imprecisos e ausentes, custos envolvidos no processamento e custos dos materiais necessários para validar os dados.

- **Saídas**
  - Dados limpos, corrigidos ou complementados;
  - Dados rejeitados e incorretos.

3.6. Consolidar os dados (P4.5)

- **Objetivo**

  Este processo tem como objetivo gerar dados consolidados sem redundância.

- **Processo**

Este processo analisa os dados para identificar ocorrências de duplicidades, para então consolidá-las. O processo de consolidação deve gerar referências entre o dado consolidado e as informações originais.

As informações de entrada para este processo são todas as fontes de dados para cada entidade.

Os dados padronizados e corrigidos podem ser agora encontrados em uma ou mais fontes de dados. Se a mesma entidade possuir diferentes identificadores em diferentes fontes, esta poderá ser consolidada em uma única ocorrência, durante este processo; a utilização de ferramentas de reengenharia e limpeza poderá facilitar e agilizar a sua execução.

As atividades envolvidas durante esse processo são:

a. Estabelecer os critérios iniciais (campos, tipos de comparação, entre outros fatores) para identificação (*matching*);
b. Verificar a efetividade dos critérios e a saída produzida;
c. Redefinir os critérios até atingir o nível de identificação de duplicados desejado;
d. Validar os critérios de identificação em cada uma das fontes de dados. Depois de validá-los isoladamente, validar o processo de identificação de duplicados utilizando todas as fontes de dados;
e. Marcar e agrupar as ocorrências de acordo com os critérios de identificação comuns;
f. Prepará-los para os futuros processos. Essa organização será utilizada em processos futuros.

- **Saídas**

- Dados identificados, referenciados e consolidados;
- Lista dos dados identificados como duplicados;
- Lista dos dados duplicados suspeitos.

3.7. Transformar e enriquecer os dados (P4.7)

- **Objetivo**

Este processo tem como objetivo preparar os dados para a disponibilização na base de dados destino.

- **Processo**

Este processo é composto por duas etapas. A primeira transforma os dados padronizados para a estrutura de dados destino. A segunda, opcional, combina dados de fontes externas com os dados internos com a finalidade de enriquecê-los.

As informações de entrada utilizadas durante a primeira etapa são:

- Dados corrigidos e consolidados a serem disponibilizados na base de dados destino;
- Mapeamento dos dados originais para os dados padronizados;
- Estrutura dos dados do DW (*Data Warehouse*) ou dados destinos.

As atividades envolvidas durante o processo de transformação dos dados para a estrutura de dados destinos são:

1. Expandir o mapeamento dos dados padronizados para a nova estrutura de dados destino;
2. Definir e implementar as regras de transformação dos dados;
3. Definir e implementar a programação para propagar os dados, baseada no volume de alterações dos dados e no tempo disponível para os processos de extração, transformação e disponibilização dos dados, na complexidade dos processos de consolidação, entre outros;
4. Testar as transformações dos dados para garantir que os processos estão de acordo com a especificação.

Na segunda etapa de enriquecimento, opcional, são utilizadas como informações de entrada, os dados externos de provedores de informação. Existem vários tipos de dados que podem ser utilizados para enriquecer os dados existentes,

como por exemplo, informações postais (Cep, Bairro), dados pessoais (Data de Nascimento, Estado Civil), etc.

Porém, antes de utilizar uma fonte externa para enriquecimento dos dados existentes, é importante conhecer sua definição e semântica, a data em que foram obtidos, sua procedência, o nível de confiança ou nível da qualidade das informações.

A seguir alguns itens que precisam ser verificados durante o processo de formatação e enriquecimento dos dados visando a estrutura dos dados destino:

- Garantir que existe um consenso na definição dos dados da estrutura destino;
- Assegurar que a definição está de acordo com as fontes de informação;
- Garantir que as definições das regras de transformação foram satisfeitas;
- Garantir uma correta e consistente transformação de tipos de dados;
- Assegurar a definição e significado de qualquer dado externo utilizado no processo de enriquecimento;
- Assegurar que quaisquer dados externos utilizados para enriquecer dados operacionais estão sendo identificados corretamente.

➢ **Saídas**

- Dados transformados e enriquecidos;
- Mapeamento dos dados da fonte para os dados padronizados;
- Mapeamento dos dados da fonte para a estrutura destino.

3.8. Calcular derivações e sumarizar dados (P4.8)

➢ **Objetivo**

O objetivo é otimizar o desempenho do DW (*Data Warehouse*), determinando e armazenando os dados derivados para as consultas mais freqüentes

que requerem cálculos complexos, ou seja, balancear os custos de armazenamento de dados com os custos dos cálculos "*online*" de dados derivados.

- **Processo**

Este processo identifica as várias dimensões ou visões dos dados, as fórmulas e regras de negócio utilizadas para calcular os dados derivados com os especialistas do negócio, além de documentar todas as definições e cálculos das derivações e sumarizações.

As informações de entrada utilizadas nesse processo são:

- Dados formatados e enriquecidos;
- Mapeamento dos dados das fontes para a estrutura destino.

As atividades para calcular os dados derivados envolvem:

1. Modelar e identificar os atributos derivados e sumarizados necessários no DW (*Data Warehouse*);
2. Definir as regras e algoritmos para os cálculos;
3. Validar a definição dos dados e regras de cálculo junto aos especialistas no assunto;
4. Implementar as rotinas e especificar os parâmetros na ferramenta para derivação ou sumarização dos dados;
5. Testar as rotinas e certificar-se de que suas execuções estão de acordo com as especificações.

- **Saídas**

- Dados derivados ou sumarizados.

3.9. Auditar e controlar a extração, transformação e carga dos dados (P4.9)

- **Objetivo**

Este processo tem como objetivo assegurar que os dados corretos estão sendo obtidos dos arquivos corretos, transformados de acordo com as especificações e carregados nos campos corretos do banco de dados destino.

- **Processo**

As informações de entrada para esse processo são:

- Dados obtidos em cada uma das fases, desde a extração da fonte de dados original até a carga na base de dados destino;
- Definição dos dados do sistema origem;
- Definição dos dados destino;
- Regras utilizadas durante a transformação, enriquecimento e sumarização dos dados para verificação.

As atividades envolvidas durante o processo de auditoria e controle da extração, transformação e carga de dados incluem:

1. Assegurar que os dados do sistema origem, obtidos no início do processo 4, tenham uma definição clara dos dados, bem como a especificação dos domínios de valores e os critérios a serem utilizados durante o processo;
2. Assegurar que a estrutura de dados destino tenha uma definição clara dos dados, bem como a especificação dos domínios de valores e critérios a serem utilizados durante o processo;
3. Garantir que todas as regras de transformação, sumarização, identificação e consolidação tenham especificações claras;
4. Determinar os critérios de auditoria dos dados e requisitos de controle;
5. Definir os processos para implementar a auditoria e procedimentos de controle ;
6. Desenvolver procedimentos para monitorar e controlar a extração, transformação e carga de dados.

- **Saídas**
  - Relatórios de controle e da auditoria.

3.10. Analisar os tipos de defeitos (P4.6)

- **Objetivo**

O objetivo deste processo é analisar e registrar os problemas identificados nas fontes durante os processos de extração (P4.2), padronização (P4.3), correção e complementação (P4.4) e consolidação dos dados (P4.5). Este processo pode prover subsídios para a melhoria do processo de geração dos dados a ser realizada no processo 5, apresentando os problemas mais freqüentes detectados durante os processos de reengenharia.

- **Processo**

Este processo analisa as saídas de outros processos para compreender os tipos de erros detectados e a freqüência com que esses erros acontecem, visando identificar os custos e os impactos desses erros sobre o negócio.

As informações de entrada para esse processo são:

- Anomalias dos dados;
- Dados rejeitados e incorretos.

As principais atividades envolvidas são:

a. Listar e analisar exemplos dos vários tipos de anomalias encontrados nos processos anteriores;
b. Listar exemplos representativos de cada tipo de defeito;
c. Categorizar os problemas de qualidade da informação e os padrões;
d. Estimar a freqüência de cada problema;
e. Estimar os custos ou impactos relativos aos problemas detectados;
f. Avaliar os tipos de defeitos dos dados e sua importância.

- **Saídas**
  - Lista dos tipos de defeitos identificados.

3.11. Conclusão

Neste capítulo, o processo de reengenharia e limpeza de dados da metodologia TQdM foi apresentado detalhadamente, com todos os sub-processos envolvidos. Dentre todos os processos da metodologia, este foi o principal processo aplicado no estudo de caso, uma vez que o escopo deste é atuar na melhoria da qualidade da informação, e não nos processos que geraram essa informação (Processo 5).

O próximo capítulo apresenta o estudo de caso que aplicará a metodologia TQdM. Serão descritos os objetivos do estudo de caso, as ferramentas utilizadas, o ambiente de implementação, a equipe que implementa o projeto, os resultados obtidos e as lições aprendidas.

Capítulo 4 - Estudo de Caso: Aplicação TQdM na prática

4.1.Introdução

Este capítulo apresenta um estudo de caso aplicando a metodologia TQdM, principalmente o processo 4 de Reengenharia e Limpeza dos dados. Durante este capítulo, o problema encontrado no estudo de caso que motivou a implementação de um projeto de qualidade de dados é apresentado, bem como a equipe e a ferramenta utilizada. Todas as fases que ocorreram durante o seu desenvolvimento, e principalmente, as dificuldades enfrentadas são descritas.

4.2.Apresentação do Problema

Uma grande empresa brasileira da área financeira possui diversos departamentos, sendo que cada departamento mantém o seu próprio cadastro de clientes distribuídos em diversos sistemas de informação. Como cada departamento gerencia o seu próprio cadastro de clientes, cada um armazena somente as informações de clientes que julgam necessários para sua área de atuação.

Como a empresa poderá conhecer todos os seus clientes e os produtos que eles possuem ou poderiam adquirir, se existem diversos cadastros ? E como compatibilizar todas as informações dos diversos departamentos ou mesmo dentro de um departamento, se cada um possui as suas próprias regras para formatos, códigos e tamanhos de informações ?

A solução atual gera um cadastro de clientes a partir de alguns campos chaves que identificam unicamente o cliente, como por exemplo, o campo Cpf ou Cgc. Esse cadastro, armazenado em uma estrutura de banco de dados Oracle 8i, é utilizado pela empresa na área de DW (*Data Warehouse*) para tomada de decisões, não afetando os sistemas legados desenvolvidos para a entrada dos dados.

Porém esta solução não se mostrou eficiente, pois muitas vezes um mesmo número de Cpf ou Cgc pode ser compartilhado por diversas pessoas de uma mesma

família. A falta de um padrão no preenchimento de campos descritivos, como Nome, tornou inviável a utilização deste no processo de identificação. Além disso, a geração de um cadastro de clientes não compatibiliza todas as informações se nenhum processo for criado com esse objetivo. Então, como melhorar a qualidade dos dados e gerar um cadastro de clientes consolidado de qualidade com todas as informações compatibilizadas ? Esse é o objetivo desse estudo de caso que será apresentado a seguir.

4.3.Objetivo

O objetivo desse projeto é melhorar a qualidade dos dados de cada fonte de informação, originada nos diversos departamentos da empresa, e a partir dessas informações de qualidade, identificar unicamente o cliente, gerando um cadastro consolidado utilizando critérios pré-estabelecidos do negócio.

É de extrema importância que a integridade e a compatibilidade das informações consolidadas a partir das fontes origens sejam mantidas, e que nenhuma alteração seja efetuada nos sistemas mantenedores destas informações.

Neste cadastro consolidado de clientes, todas as informações devem ser equalizadas, os dados relevantes devem ser acrescidos e corrigidos, se possível, com a utilização de fontes fidedignas disponíveis, e as duplicidades de informações devem ser eliminadas.

Ao término desse projeto, espera-se obter um modelo padrão para a incorporação de novas fontes de informação ao processo de qualidade e, conseqüentemente, ao cadastro consolidado de clientes.

4.4. Equipe

Para a implementação do estudo de caso que envolveu os processos de avaliação da qualidade (P1 e P2), medição dos custos (P3) e, principalmente, o

processo de reengenharia e limpeza dos dados (P4) foi designada uma equipe com 5 pessoas, conforme apresentada na tabela 4.1.

Perfil	Quantidade
Especialista em QD e na ferramenta adquirida	1
Gerente de Projeto	1
Analista de Sistema Sênior	3

Tabela 4. 1 - Equipe de Trabalho

Durante a fase de avaliação da qualidade (P1 e P2) e medição dos custos (P3) todos os membros da equipe participaram, porém no desenvolvimento do processo de reengenharia (P4), o especialista na ferramenta, por ser o mais experiente em projetos de QD, distribuiu as funções técnicas para os analistas da seguinte forma:

- Análise e definição dos padrões: foi designado um analista responsável pela análise, definição e implementação dos padrões para os quais as fontes deveriam ser convertidas;
- Avaliação da qualidade: foi designado um analista responsável pelas avaliações da qualidade dos dados de todas as fontes de informação;
- Planejamento, implementação e testes de um projeto piloto: foi designado um analista responsável por planejar os sub-processos do processo de reengenharia (P4), implementá-los e testá-los, gerando um modelo para as próximas fontes. A autora participou de perto na geração deste modelo.

Todos os analistas envolvidos já possuíam conhecimento do assunto qualidade de dados e alguns deles já estavam familiarizados com a ferramenta de reengenharia e limpeza adotada.

4.5. Ferramenta

Analisando o objetivo do projeto e verificando a necessidade de melhoria na qualidade dos dados devido aos diferentes formatos, tamanhos e códigos, notou-se a

necessidade de uma solução capaz de reparar os dados e definir transformações. Dentre as soluções apresentadas no capítulo 2, foi escolhido uma ferramenta de reengenharia de dados.

Dentre as ferramentas disponíveis no mercado para essa finalidade, optou-se pela ferramenta Integrity da empresa Ascential. Ela foi escolhida por ser a ferramenta mais completa, conforme apresentado anteriormente no capítulo 2.3, e, principalmente, por possibilitar a customização de regras e padrões de acordo com a necessidade da empresa.

A seguir, foi necessário selecionar a plataforma a ser utilizada. Novamente as considerações do *Gartner Group* foram relevantes nesta decisão.A plataforma mais vendida até o momento da escolha da plataforma no mercado internacional era o ambiente *Mainframe*; além disso, esse era o principal ambiente do cliente.

Porém durante a execução do projeto nessa plataforma, começaram a surgir dificuldades, tais como, indisponibilidade de recursos de memória, de processamento, e principalmente, de área de armazenamento, devido à concorrência com o ambiente de produção, além de dificuldades em manipular arquivos para consultas rápidas nessa plataforma. Por esta razão, a plataforma foi alterada para Sun Solaris por ser um ambiente mais flexível. A nova plataforma requereu nova instalação e configuração da máquina, mas houve um ganho considerável no tempo de execução e verificação dos processos.

O ambiente utilizado para implementação foi uma máquina dedicada para o desenvolvimento e produção do projeto, com sistema operacional SunOS 5.8, 5 processadores, cerca de 1,5 Teras de disco e 10MB de memória. Os softwares instalados foram Integrity 3.11 e SyncSort como utilitário para ordenação.

4.6. Procedimentos

Todos os processos da metodologia TQdM implementados durante o estudo de caso são apresentados a seguir e o processo de Reengenharia e Limpeza dos dados (Processo 4) é apresentado em detalhes em um item à parte.

4.6.1.Processo 6: Estabelecer um ambiente de qualidade da informação (Plano de Ação)

O plano de ação proposto tem como objetivo estabelecer um ambiente de qualidade da informação e propõe diversas iniciativas para garantir que esse ambiente seja obtido. Dentre todas as iniciativas apresentadas no capítulo 2.4.7, a única que foi considerada estratégica para atingir o objetivo do projeto e que foi utilizada durante a implementação do processo de Reengenharia (Processo 4) do estudo de caso foi : “Selecionar uma área pequena, gerenciável e com alto retorno para conduzir um projeto piloto”. Por isso foi selecionada uma fonte para implementar um projeto piloto e validar a solução, antes da implementação do projeto de qualidade envolvendo todas as fontes de informação. Depois do processo ser testado e refinado no projeto piloto, ele foi implementado nas demais fontes.

4.6.2.Processo 1: Avaliar o significado dos dados e a qualidade da estrutura dos dados

Esse processo foi denominado internamente de “Avaliação da Qualidade” e teve como objetivo avaliar, de uma forma global, o ambiente do DW (*Data Warehouse*), as fontes originadores de informação, os dados produzidos e os processos de carga e, principalmente, analisar as informações consolidadas obtidas através do atual processo, que utiliza a informação CPF/CGC como critério de identificação única.

Entrevistas foram realizadas para conhecer o ambiente e efetuar o levantamento das fontes de informação e das especificações detalhadas de cada uma delas, comparativamente à atual estrutura de banco de dados. O escopo da análise abrangeu três fontes de informação, denominadas genericamente de Fonte A, Fonte B e Fonte C, que o cliente julgou serem de extrema importância para o seu negócio e que representam a maior parte das informações relevantes de seus clientes.

4.6.3.Processo 2: Avaliar a qualidade da informação

Após a avaliação da qualidade da definição e da estrutura dos dados, ocorreu a avaliação dos dados; essa fase também foi denominada “Avaliação da Qualidade”.

O objetivo dessa avaliação foi efetuar uma análise nos dados para verificar o seu preenchimento e sua freqüência, além de identificar alguns exemplos de dados preenchidos indevidamente.

Durante as análises, dados sem preenchimento ou com preenchimento utilizando zeros foram considerados vazios e campos com conteúdos incoerentes, devido a problemas de digitação rápida, ou possuindo uma quantidade considerável de dígitos iguais, foram denominados “mascarados”. A tabela 4.2 apresenta um exemplo dos resultados obtidos após a análise do campo Cpf; todos os campos de cada uma das fontes de informação passaram por essa análise.

	Fonte_A	**Fonte_B**	**Fonte_C**
CPF	**Freqüência**	**Freqüência**	**Freqüência**
Vazios	1%	0%	0%
Mascarados	5%	0%	1%

Tabela 4. 2 - Resultados da análise do campo Cpf

A próxima análise teve como objetivo identificar possíveis inconsistências de valores, conforme apresentado na tabela 4.3, onde nomes femininos são classificados com sendo do sexo masculino (Cod_Cliente=2); registros identificados como duplicados por possuírem o mesmo campo Cpf, mas correspondendo a pessoas diferentes (Cod_Cliente=1); registros não identificados como duplicados pois os campos Cpfs eram diferentes (Cod_Cliente=1, 2, 4).

CPF	**Cod_Cliente**	**Nome**	**Endereço**	**Sexo**
16475583901	1	VERA L SANTANA	R TITO 45	F
16475583909	2	VERA LUCIA SANTANA	R TITO 45	M
11111111111	3	VERA L S MORAES	R TITO 45	M
06475583901	4	VERA L SANTANA	R DR TITO 45	F
16475583901	1	ANA L SANTANA	R TITO 45	F
16475583901	5	JOAO P SANTANA	R DR TITO 45	F

Tabela 4. 3 - Exemplos de inconsistências de valores

Os produtos gerados foram relatórios apresentando todos os aspectos analisados, problemas de preenchimento dos campos, e, principalmente, apresentando casos de incoerências e ineficiência no processo atual de identificação. Toda a documentação resultante das análises realizadas foi disponibilizada para conhecimento e tomada de decisão.

4.6.4.Processo 3 : Medir os custos resultantes das informações sem qualidade

O objetivo deste processo é mensurar os custos da falta de qualidade da informação e, conforme descrito na apresentação do problema, a existência de cadastros descentralizados faz com que a identificação única de um cliente seja dificultada; logo conhecer o próprio cliente passa a ser difícil. Conhecer o cliente significa a possibilidade de identificar quais são os produtos mais indicados para serem oferecidos a um determinado cliente segundo seu perfil de risco e capacidade de investimento. Observando o problema, adotou-se como diretriz a redução de custo, pela canalização do esforço correto no sentido de oferecer o produto correto ao cliente correto. Observa-se que a mensuração dos custos é um fator de sigilo da organização e por isso serão abordados apenas os aspectos relativos à análise do custo em função dos números apurados.

A medição do custo sob o ponto de vista adotado passa pela identificação do valor de retorno do custo da implementação da estrutura necessária para executar os processos de qualidade de dados, pela identificação do custo da informação incorreta, que pode vir do custo da mão de obra direta durante o tempo em que esta aplica seu esforço na prospecção de negócios, tendo como alvo o cliente errado ou o custo de postagem de documentos para o cliente errado, e pelo estabelecimento do valor da qualidade da informação que identifica em quanto tempo a implementação de um processo de qualidade de dados possibilita seu retorno sobre o investimento. Esses dados permitem a identificação clara da viabilidade de implementação do processo de qualidade e permitem a tomada da decisão de sua implementação.

4.6.5.Processo 4 : Realizar a reengenharia e limpeza dos dados

O objetivo desse processo é melhorar a qualidade dos dados através de processos de padronização, correção e complementação dos dados, identificação e eliminação de duplicidades, transformação e enriquecimento utilizando fontes de dados fidedignas, e alimentando as bases de dados destinos.

Este processo é o mais importante, por ser o objetivo principal deste trabalho, por isso ele será descrito em detalhes no item 4.7.

4.6.6.Processo 5: Melhorar a qualidade dos processos envolvidos na geração da informação

O objetivo desse processo é melhorar a qualidade dos processos que originaram a informação, fazendo com que os problemas sejam corrigidos na sua origem. Como o objetivo deste projeto de qualidade foi melhorar a qualidade do processo em curso e não nos sistemas de origem das informações, esse processo não foi implementado. Mas futuramente esse processo poderá ser incorporado e será de grande impacto, pois afetará todos os sistemas legados da empresa que geraram as informações com problemas de qualidade.

4.7. Processo 4 – Reengenharia e Limpeza dos dados

O processo de reengenharia e limpeza dos dados é o principal processo deste trabalho por englobar todas as atividades envolvidas para a melhoria da qualidade dos dados e, conseqüentemente, para a geração do cadastro consolidado de clientes. Nos itens a seguir, todas as atividades envolvidas são apresentadas em detalhes.

4.7.1.Identificar as possíveis fontes dos dados (P4.1)

O cliente selecionou inicialmente oito fontes que entrariam no processo para a geração do cadastro único e forneceu os arquivos confiáveis que deveriam ser utilizados para agregar valor às informações de produção. Esses arquivos contêm dados dos Correios e da operadora de telefonia. Destas oito fontes de informação, somente três foram utilizadas devido à falta de tempo para implementação; elas foram selecionadas por serem fundamentais para o negócio da empresa.

Entrevistas com os responsáveis pela informação foram efetuadas para recuperar a definição formal de cada um dos arquivos de dados, obtendo os tipos e conteúdos esperados de cada campo. Com as definições dos arquivos, conforme apresentadas nas tabelas 4.4, 4.5 e 4.6, foi possível criar uma lista de referências (tabela 4.7) que estabelece um relacionamento entre os grupos de campos e as respectivas fontes de informação.

FONTE A	
Nome do Campo	**Tipo/Tamanho/Formato**
Código_cliente	Alfanumérico(08)
CPFCGC	Alfanumérico(14)
Nome	Alfanumérico(70)
Endereço	Alfanumérico(60)
Numero	Alfanumérico(10)
Complemento	Alfanumérico(10)
Cep	Alfanumérico(08)
Bairro	Alfanumérico(25)
Município	Alfanumérico(60)
Estado	Alfanumérico(40)
Sexo	Alfanumérico(01)
Datanascimento	Alfanumérico(08) (Formato "DDMMAA", onde D representa o dia , M o mês e A o ano)
Telefone	Alfanumérico(15)
DataAtualizacao	Alfanumérico(19) (Formato "AAAA-MM-DD HH:mm:SS", onde A representa o ano, M o mês, D o dia, H a hora, M o minuto e S o segundo)

Tabela 4. 4 - Definição dos campos da FONTE A

FONTE B	
Nome do Campo	**Tipo/Tamanho**
Codigo_cliente	Alfanumérico(08)
CPFCGC	Alfanumérico(14)
Nome	Alfanumérico(60)
Endereço	Alfanumérico(100)
Cep5	Alfanumérico(05)
Cep3	Alfanumérico(03)
Bairro	Alfanumérico(15)
Município	Alfanumérico(60)
Estado	Alfanumérico(02)
Sexo	Alfanumérico(01)
Datanascimento	Alfanumérico(08) (Formato "DDMMAA", onde D representa o dia , M o mês e A o ano)
DataAtualizacao	Alfanumérico(19) (Formato "AAAA-MM-DD HH:mm:SS", onde A representa o ano, M o mês, D o dia, H a hora, M o minuto e S o segundo)

Tabela 4. 5 - Definição dos campos da FONTE B

FONTE C	
Nome do Campo	**Tipo/Tamanho**
Codigo_cliente	Alfanumérico(8)
CPFCGC	Alfanumérico(14)
Nome	Alfanumérico(70)
Endereço	Alfanumérico(60)
Numero	Alfanumérico(10)
Complemento	Alfanumérico(10)
Cep	Alfanumérico(8)
Município	Alfanumérico(60)
Estado	Alfanumérico(2)
Telefone	Alfanumérico(15)
DataAtualizacao	Alfanumérico(19) (Formato "DDMMAA", onde D representa o dia , M o mês e A o ano)

Tabela 4. 6 - Definição dos campos da FONTE C

A lista de referência apresentada na tabela 4.7, além de associar as fontes de informação com os grupos de informações, também estabelece, a partir de regras do negócio, o critério de desempate a ser utilizado na escolha da melhor informação através de uma numeração, por exemplo, onde quanto menor o número, mais importante é a fonte. Por exemplo, durante a escolha da melhor informação considerando o campo CPF, a informação contida na fonte A é mais confiável do que

a informação contida na fonte B, por esta razão foram atribuídos os números 1 e 2 respectivamente, e por sua vez, a fonte B é mais confiável do que a fonte C, por esta razão foi atribuído o número 3 para esta fonte C. Porém considerando o campo Nome o mesmo não ocorre; neste caso a informação mais confiável é a informação contida na fonte B.

LISTA DE REFERÊNCIAS						
Nome do Campo	**Fonte A**		**Fonte B**		**Fonte C**	
Codigo_cliente	Alfanumérico(08)		Alfanumérico(08)		Alfanumérico(08)	
CPFCGC	Alfanumérico(14)	1	Alfanumérico(14)	2	Alfanumérico(14)	3
Nome	Alfanumérico(70)	2	Alfanumérico(60)	1	Alfanumérico(70)	3
Endereço	Alfanumérico(60) Número Alfanumérico(10) Complemento Alfanumérico(10)	2	Alfanumérico(100)	1	Alfanumérico(60) Número Alfanumérico(10) Complemento Alfanumérico(10)	3
CEP	Alfanumérico(8)	2	CEP5 Alfanumérico(05) CEP3 Alfanumérico(03)	1	Alfanumérico(08)	3
Bairro	Alfanumérico(25)	2	Alfanumérico(15)	1	-	3
Município	Alfanumérico(60)	2	Alfanumérico(60)	1	Alfanumérico(60)	3
Estado	Alfanumérico(40)	2	Alfanumérico(02)	1	Alfanumérico(02)	3
Sexo	Alfanumérico(01)	1	Alfanumérico(01)	2	-	
DataNascimento	Alfanumérico(08)	1	Alfanumérico(08)	2	-	
Telefone	Alfanumérico(15)	2	-	3	Alfanumérico(15)	1
DataAtualização	Alfanumérico(19)		Alfanumérico(19)		Alfanumérico(19)	

Tabela 4. 7 - Lista de referências de campos

Além dos *layouts* das fontes de dados, foram levantados também os *layouts* dos arquivos a serem utilizados como fontes fidedignas para enriquecimento, no caso foram os arquivos dos Correios e da operadora de telefonia, conforme as tabelas 4.8 e 4.9.

CORREIOS	
Nome do Campo	**Tipo/Tamanho**
CEP	Alfanumérico(08)
Endereço	Alfanumérico(100)
FaixaInicial	Alfanumérico(05)
FaixaFinal	Alfanumérico(05)
Bairro	Alfanumérico(20)
Localização	Alfanumérico(60)
Estado	Alfanumérico(02)

Tabela 4. 8 - Definição dos campos dos Correios

Operadora_Telefonica	
Nome do Campo	**Tipo/Tamanho**
DDD	Alfanumérico(4)
Cidade	Alfanumérico(60)
Prefixo	Alfanumérico(4)

Tabela 4. 9 - Definição dos campos da operadora telefônica

4.7.2.Extrair e analisar os dados dos arquivos fontes (P4.2)

No processo de “Avaliação dos Dados” (P1 e P2), a análise foi efetuada utilizando somente uma amostra dos arquivos; nesta fase todas as informações de uma mesma data de referência foram consideradas na análise.

Esta análise é denominada de “investigação” pela ferramenta de reengenharia utilizada, e tem como objetivo avaliar a qualidade da informação de cada um dos campos dos arquivos fontes.

A tabela 4.10 apresenta um exemplo dos resultados obtidos durante a análise do arquivo FONTE A, onde foram analisados, para cada um dos campos, o fator de preenchimento, a quantidade de informações inválidas e válidas (considerando somente as informações preenchidas), sendo que foram consideradas informações inválidas aquelas que possuíam informações fora do domínio padrão. Esse processo de investigação foi efetuado para todas as fontes de informação do projeto.

	Preenchidos		% Válidos		% Inválidos		Brancos	
Campo	**Qtd**	**%**	**Qtd**	**%**	**Qtd**	**%**	**Qtd**	**%**
Codigo_Cliente	5.000.000	100.00	5.000.000	100.00	-	0.00	-	0.00
Nome	4.999.500	99.99	4.999.500	100.00	-	0.00	500	0.01
Endereço	4.977.000	99.54	4.970.530	99.87	6.470	0.13	23.000	0.46
Numero	4.999.500	99.99	4.999.500	100.00	-	0.00	500	0.01
Complemento	1.587.500	31.75	1.586.706	99.95	794	0.05	3.412.500	68.25
Cep	4.985.000	99.70	4.985.000	100.00	-	0.00	15.000	0.30
Bairro	4.925.000	98.50	4.920.075	99.90	4.925	0.10	75.000	1.50
Município	175.000	3.50	173.740	99.28	1.260	0.72	4.825.000	96.50
Estado	2.872.500	57.45	2.843.775	99.00	28.725	1.00	2.127.500	42.55
Sexo	2.450.000	49.00	2.448.530	99.94	1.470	0.06	2.550.000	51.00
DataNascimento	2.550.000	51.00	2.547.450	99.90	2.550	0.10	2.450.000	49.00
Telefone	350.000	7.00	245.000	70.00	105.000	30.00	4.650.000	93.00
DataAtualização	750.000	15.00	750.000	100.00	-	0.00	4.250.000	85.00

Tabela 4. 10 - Resultados obtidos da análise da FONTE A

Após a análise campo a campo de cada um dos arquivos fontes, uma análise do conteúdo dos dados foi efetuada, gerando relatórios e gráficos apresentando os dez valores mais freqüentes e os dez menos freqüentes para cada um dos campos. Todas as análises efetuadas tiveram como objetivo avaliar a qualidade da informação; lembrando que, as especificações técnicas e do negócio foram fornecidas pelo cliente para possibilitar a análise completa das informações.

Ao término deste processo foi possível identificar a necessidade da criação de novos atributos para equalizar a granularidade do campo Endereço, além da necessidade de ajustes nos domínios e formatos dos campos Estado, Data de Nascimento, Telefone e Sexo para adequar a um único formato padrão.

4.7.3. Padronizar os dados (P4.3)

Este processo tem como objetivo estabelecer uma forma única para representação dos dados, envolvendo o seu conteúdo e o seu formato de saída. Para ser possível à definição e a implementação dos padrões na ferramenta de reengenharia (Integrity) foi necessário compreender como ela funciona.

A figura 4.1, apresenta as atividades envolvidas na definição dos padrões de classificação, de conversão e de formatação dos campos utilizados pelo processo na ferramenta. Inicialmente todos os campos comuns em todas as fontes que necessitam de padronização são selecionados (processo 1 - figura 4.1). Após algumas reuniões com os responsáveis pelas informações, são estabelecidas as formas padronizadas dos campos, objetivando eliminar os possíveis domínios redundantes (processo 2 – figura 4.1). Se necessário, alguns novos atributos poderão ser gerados para eliminar casos de várias informações em um mesmo campo. Os *layouts* padrões de saída são estabelecidos durante esse processo, visando eliminar alguns formatos de campos inconsistentes e nível de granularidade distintos (processo 3 – figura 4.1).

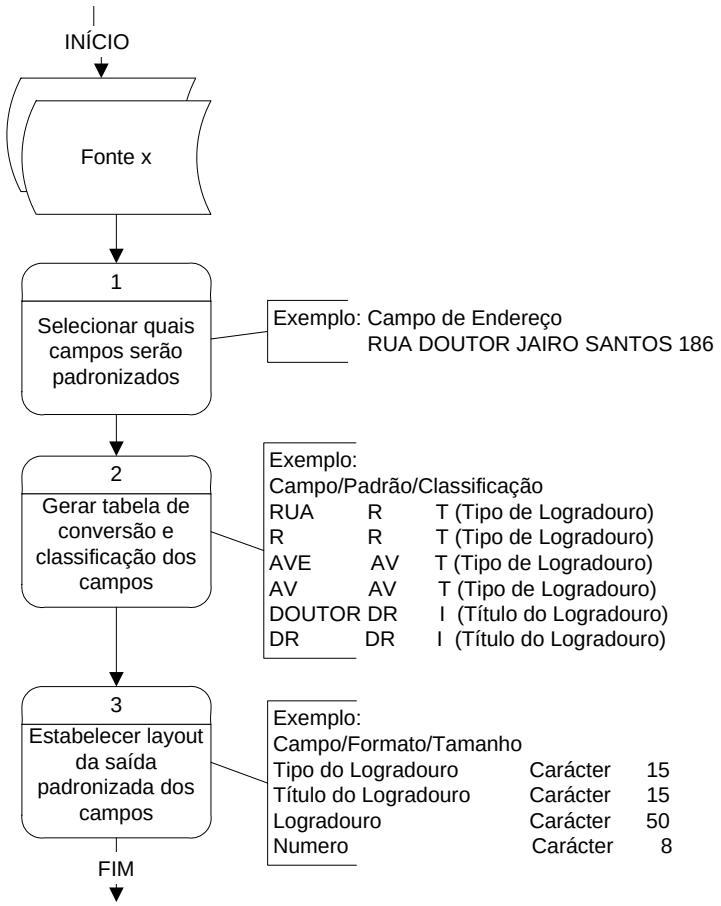

Figura 4. 1 - Atividades envolvidas na definição dos padrões

Após a definição de todos os padrões a serem utilizados, a próxima fase é implementar esses padrões na ferramenta. Para isso é necessário que uma regra seja criada para cada tipo de campo. Regra é a denominação de um recurso da ferramenta Integrity que possibilita a obtenção dos padrões definidos.

A figura 4.2 apresenta o funcionamento das regras. A primeira etapa consiste na classificação do conteúdo do campo em padrões (*tokens*) utilizando uma tabela de classificações da própria ferramenta (tabela 4.11) ou realizando algumas customizações, se necessário, como por exemplo, Tipo de Logradouro e Título de Logradouro (processo 1 – figura 4.2). Após a conversão em *tokens*, a próxima etapa é responsável por converter os domínios dos campos utilizando tabelas de conversões definidas anteriormente, como por exemplo, DOUTOR para DR (processo 2 – figura 4.2). A seguir, os dados classificados e convertidos são formatados para o *layout* de saída (processo 3 – figura 4.2) gerando assim o arquivo com os dados padronizados.

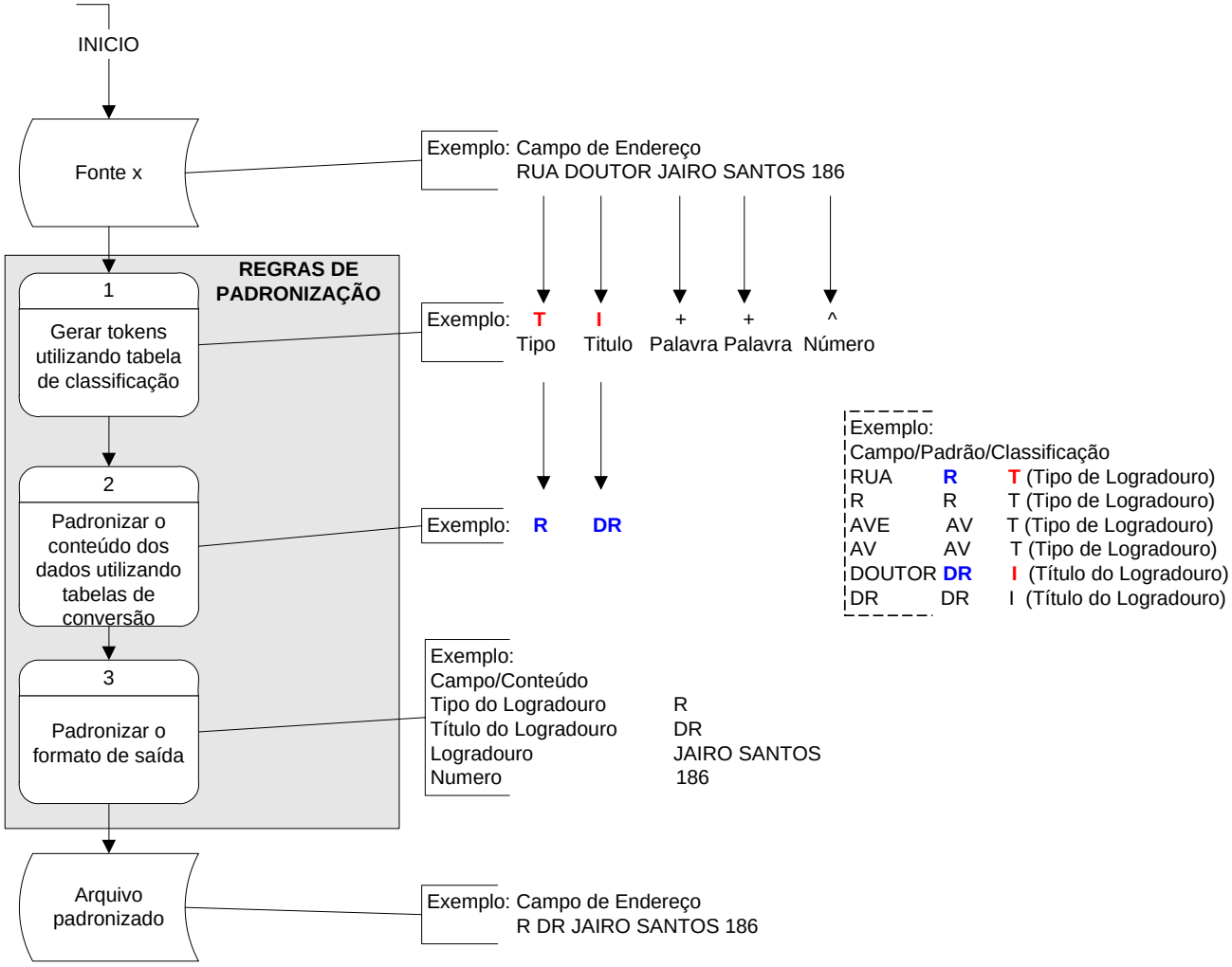

Figura 4. 2 - Funcionamento das regras de padronização

Classificação	Descrição
^	Valores numéricos. Exemplo: 9
?	Uma ou mais palavras desconhecidas consecutivas. Exemplo: ANA MARIA
+	Uma única palavra alfabética Exemplo: ANA
&	Um único padrão de qualquer tipo Exemplo: XXX
>	Números e Letras consecutivos Exemplo: 999SAO
<	Letras e Números consecutivos Exemplo: SAO999
@	Misto (Letras, Números) Exemplo: 999SAO999

Tabela 4. 11 - Tabela de padrões da ferramenta Integrity

Após a análise de quais campos necessitavam de padrões, a partir dos resultados obtidos no processo anterior (P4.2), as regras para os campos Nome, Endereço, Telefone, Data, Cidade e UF foram criadas.

O objetivo das regras de Endereço e Nome, além de compatibilizar algumas grafias, é ajustar os diferentes níveis de granularidade existentes entre as fontes de informação auxiliando os futuros processos de comparações. Já as demais regras foram criadas com o objetivo de ajustar os domínios redundantes e formatos inconsistentes. Alguns exemplos de padronização utilizando todas as regras definidas são apresentados na tabela 4.12.

Regra	Entrada	Saída Padronizada
Endereço	AVENIDA PAULISTA 45 10 ANDAR	Tipo: AV Titulo: Logradouro: PAULISTA Número: 45 Complemento: 10 AND Endereço: AV PAULISTA 45 10 AND
	AVE DOUTOR ARNALDO 345 APTO 45	Tipo: AV Titulo: DR Logradouro: ARNALDO Número: 345 Complemento: AP 45 Endereço: AV DR ARNALDO 345 AP 45
Telefone	(011) 3434-5555	DDD: 011 Número: 34345555 Ramal:
	5587-3333 RAMAL 34	DDD: Número: 55873333 Ramal: 34
Data	121003	Data: 12102003
Cidade	SP	Cidade: SAO PAULO
	SA PAULO	Cidade: SAO PAULO
UF	SAO PAULO	UF: SP
	S PAULO	UF: SP

Tabela 4. 12 - Exemplos de padronização

A tabela 4.13 apresenta a regra definida para o campo Nome. É importante ressaltar que, essa regra produz duas informações novas: Sexo e Tipo de Pessoa, que foram determinados através de pesquisas em tabelas de conversões. Essas tabelas possuem a maior parte dos nomes de pessoas e o sexo relacionado, como por exemplo, ANA determinando Sexo FEMININO. No caso da classificação do Tipo de

Pessoa, esta é obtida através de palavras chaves existentes no nome que determinam pessoa jurídica, por exemplo, LIMITADA, SA, entre outras.

Entrada	Saída Padronizada
PROFESSORA MARIA SIQUEIRA	Prefixo: PROFA Primeiro_Nome: MARIA Nome_Meio: Último_Nome: SIQUEIRA Sufixo: Nome: PROFA MARIA SIQUEIRA Sexo: F Tipo_Pessoa:
JOAO MANOEL MARTINS FILHO	Prefixo: Primeiro_Nome: JOAO Nome_Meio: MANOEL Último_Nome: MARTINS Sufixo: FL Nome: JOAO MANOEL MARTINS FL Sexo: M Tipo_Pessoa:
EMPRESA SANTOS LIMITADA	Prefixo: Primeiro_Nome: Nome_Meio: Último_Nome: Sufixo: Nome: EMPRESA SANTOS LTDA Sexo: Tipo_Pessoa: J

Tabela 4. 13 – Exemplos de padronização utilizando regra de Nome

A próxima etapa foi responsável por identificar, para cada uma das fontes de dados, quais seriam as regras aplicadas em cada campo, produzindo uma relação Regra x Campo conforme apresentada na tabela 4.14.

Ao término desse processo, todos os campos que necessitavam ser equalizados quanto a padrões (conteúdos e formatos), encontravam-se padronizados e prontos para a próxima fase de reengenharia, conforme tabela 4.15.

Fonte	Campo	Regra
FONTE A	Nome	Nome
	Endereco + Numero + Complemento	Endereco
	Municipio	Cidade
	Estado	UF
	DataNascimento	Data
	Telefone	Telefone
	DataAtualização	Data
FONTE B	Nome	Nome
	Endereco	Endereco
	Municipio	Cidade
	Estado	UF
	DataNascimento	Data
	DataAtualizacao	Data
FONTE C	Nome	Nome
	Endereco + Numero + Complemento	Endereco
	Municipio	Cidade
	Estado	UF
	Telefone	Telefone
	DataAtualizacao	Data
CORREIOS	Endereco	Nome
	Localizacao	Cidade
	Estado	UF
TELEFONICA	DDD + Telefone	Telefone

Tabela 4. 14 - Associação Campos x Regras

	Preenchidos		Padronizados	
Campo	**Qtd**	**%**	**Qtd**	**%**
Codigo_Cliente	5.000.000	100.00	-	0.00
Nome	4.999.500	99.99	4.999.500	99.99
Endereço	4.977.000	99.54	4.977.000	99.54
Numero	4.999.500	99.99	4.999.500	99.99
Complemento	1.587.500	31.75	1.587.500	31.75
Cep	4.985.000	99.70	-	0.00
Bairro	4.925.000	98.50	-	0.00
Município	175.000	3.50	175.000	3.50
Estado	2.872.500	57.45	2.872.500	57.45
Sexo	2.450.000	49.00	2.450.000	49.00
DataNascimento	2.550.000	51.00	2.550.000	51.00
Telefone	350.000	7.00	350.000	7.00
DataAtualização	750.000	15.00	-	0.00

Tabela 4. 15 – Resultados após padronização

4.7.4.Corrigir e completar os dados (P4.4)

Este processo tem como objetivo atuar na correção dos dados, e quando possível, na complementação dos dados. A primeira atividade envolvida foi à identificação de quais campos necessitavam de correções, isto foi feito utilizando-se os relatórios obtidos durante a extração e análise dos dados (Processo 4.2). Após várias reuniões com os especialistas das fontes dos dados decidiu-se que o campo Sexo seria corrigido e o campo Tipo de Pessoa seria um campo a ser implementado. A seguir, foi definido que os ajustes dos campos seriam efetuados de forma automática utilizando a ferramenta de reengenharia. A próxima fase foi identificar quais seriam os critérios de correção e complementação a serem aplicados para cada um dos campos.

O critério adotado para o campo Sexo e exemplificado na tabela 4.16, foi que o conteúdo do campo Sexo ser gerado pela regra de padronização, definida no processo anterior (P4.3). Caso a regra não conseguisse classificar o nome como sendo um nome feminino ou masculino, ou seja, o conteúdo do campo Nome não fosse encontrado na tabela de conversão durante a padronização, por exemplo, YURI, DARCI, VALDECI, etc, o campo a ser utilizado seria o de entrada.

Nome	Campo Sexo (Entrada)	Campo Sexo (Regra)	Campo Sexo (Corrigido)
ANA MARIA DA SILVA	M	F	**F**
YURI MARTINS	F		**F**

Tabela 4. 16 - Exemplo de correção

Para o campo Tipo de Pessoa, o processo utilizado possuiu como objetivo a complementação, pois a regra para Nome utilizada durante a padronização somente conseguiu identificar os casos de nomes de pessoas do tipo “jurídica”, nos outros casos, o campo não foi determinado. Por essa razão, foi definido que todos os demais casos determinariam pessoas do tipo “física”, como apresentado na tabela 4.17 para os nomes ANA e YURI.

Nome	Campo Tipo de Pessoa (Entrada)	Campo Tipo de Pessoa (Regra)	Campo Tipo de Pessoa (Complementado)
ANA MARIA DA SILVA			**F**
YURI MARTINS			**F**
EMPRESA XPTO LTDA		J	**J**

Tabela 4. 17 - Exemplos de complementação

Ao final deste processo, o campo Sexo foi corrigido e o campo Tipo de Pessoa foi complementado, conforme apresentado na tabela 4.18. Lembrando que, o conteúdo do campo Sexo pode pertencer a um domínio válido, como por exemplo, FEMININO, e mesmo assim, ter sido corrigido através deste processo para MASCULINO, pois não estava condizente com o conteúdo do campo Nome (MAURO) .

		Preenchidos		% Válidos		% Inválidos		Corrigidos / Complementados	
Campo	**Total_Regs**	**Qtd**	**%**	**Qtd**	**%**	**Qtd**	**%**	**Qtd**	**%**
Sexo	5.000.000	2.450.000	49.00	2.448.530	99.94	1.470	0.06	3.000.000	60.00
TipoPessoa	5.000.000	-	0.00	-	0.00	-	0.00	2.250.000	45.00

Tabela 4. 18 – Resultados após correção e complementação

4.7.5.Consolidar os dados (P4.5)

Este processo teve como objetivo gerar dados consolidados sem redundância. Visando identificar os casos de duplicidades, foram analisados e discutidos os relatórios gerados durante o processo de extração e análise dos dados (Processo 4.2). Definiu-se que todos os critérios de identificação a serem adotados neste projeto seriam conservadores, ou seja, no caso de qualquer dúvida ou incerteza, os dados não seriam caracterizados como duplicados, pois as conseqüências de alguma identificação indevida seriam piores do que os ganhos com a eliminação de dados duplicados.

Para implementar corretamente os critérios estabelecidos para a consolidação dos dados na ferramenta de reengenharia foi necessário conhecer e compreender as

variáveis utilizadas durante o processo de consolidação ou identificação (*matching*). As variáveis requeridas pela ferramenta são apresentadas a seguir:

- **Tipo de Identificação**

  O processo de identificação pode ser utilizado para identificar as duplicidades existentes dentro de um mesmo arquivo (*UNDUP*), bem como, pode ser utilizado para identificar informações equivalentes entre dois arquivos (*MATCH*).

- **Bloco**

  Após a seleção do tipo de identificação, a ferramenta requer pelo menos uma definição de bloco. Um bloco representa um grupo de informações que possuem um ou mais campos idênticos. Definir um bloco significa determinar quais os grupos de informações com possibilidade de duplicidade. A ferramenta implementa a utilização de blocos para reduzir a quantidade de registros envolvidos durante o processo de comparação, e assim, reduz o tempo de pesquisa e torna o processo mais eficiente. Por isso durante o processo de escolha dos campos é importante considerar que, quanto maior o grupo gerado, pior será o desempenho devido ao número de comparações necessárias durante o processo de identificação, além disso, a ferramenta possui limitação para o tamanho do bloco.

- **Tipo de comparação**

  Após o agrupamento dos dados, é necessário definir quais campos serão utilizados na comparação e qual o tipo de comparação a ser utilizada para cada um dos campos escolhidos. A ferramenta possui diversos tipos de comparações, como por exemplo, comparação por caracteres (*CHAR*), comparação numérica (*NUMERIC*), comparação alfanumérica (*UNCERT*), etc. Cada tipo de comparação possui parâmetros específicos a serem definidos. A tabela 4.19 apresenta os parâmetros requeridos pela comparação alfanumérica denominada *UNCERT*.

Parâmetro	Descrição	Valores Possíveis
m-prob	Probabilidade de o campo ser equivalente dado que os pares de registros foram identificados	0.9 (valor default)
u-prob	Probabilidade de o campo ser equivalente dado que os pares de registros não foram identificados	0.01 (valor default)
Param1	Determina o grau exigido para que a comparação seja considerada equivalente. Quanto maior o valor neste campo, mais exigente será a comparação aproximada.	900–Dados Idênticos 850–Dados estão tão próximos que podem ser considerados equivalentes 800-Dados são provavelmente os mesmos 750–Dados são provavelmente diferentes 700–Dados são certamente diferentes

Tabela 4. 19 - Parâmetros requeridos pelo *UNCERT*

Depois de conhecer as variáveis necessárias para implementar o processo de consolidação, foi possível definir corretamente os critérios a serem utilizados. Um dos critérios estabelecidos foi o de denominar “ocorrências duplicadas” aquelas cujo campo CPF possuíssem valores equivalentes e as que possuíssem o campo Nome contendo valores similares ou aproximados. Implementando este critério na ferramenta, foram obtidos os valores apresentados na tabela 4.20, onde o tipo de identificação definido foi *UNDUP*, pois o processo foi criado com o objetivo de identificar duplicatas na mesma fonte de dados. Como se determinou que o CPF deveria ser idêntico, ele foi eleito como o campo a ser utilizado para geração dos blocos, e dentro dos blocos foram realizadas comparações alfanuméricas (*UNCERT*) no campo Nome, considerado que a probabilidade do conteúdo deste campo estar correto era de 50% (m-prob 0.5) e que o conteúdo provavelmente possuiria variação na escrita (Param1 700) entre as ocorrências.

Variável	Valor
Tipo de Identificação	UNDUP
Bloco	CPF
Tipo de comparação aproximada Parâmetros	UNCERT Campo NOME m-prob 0.5 u-prob 0.01 Param1 700

Tabela 4. 20 – Parâmetros definidos para critério 1

Definidos os valores das variáveis, o processo de consolidação foi implementado e testes foram realizados até atingir o nível de refinamento desejado para este critério. Após a execução do processo de identificação, as ocorrências receberam uma nova numeração (Dup) e para aquelas identificadas como duplicadas, um mesmo número é atribuído, conforme apresentado na tabela 4.21.

Código Cliente	CPF	Nome	Endereço	Data Nasc	Dup
1	25421212449	PAULO SERGIO GONZAGA	AV PAULISTA 50	12121960	**1**
2	25421212449	PAULO S GONZAGA	AL SANTOS 34	01011900	**1**
3	25421212488	PAULO SERGIO GONZAGA	PAULISTA 50	12121960	**2**
4	11111111111	PAULO S GONZAGA	AV PAULISTA 50	12121960	**3**
5	25421212567	PAULO S GONZAGA	BRIG TOBIAS 2	01011900	**4**

Tabela 4. 21 – Resultados da consolidação utilizando critério 1

Além da geração de um novo arquivo identificado, ao término do processo, um relatório é gerado apresentando as ocorrências de duplicidades, conforme a figura 4.3. Neste relatório, os campos utilizados durante as comparações são apresentados. Uma nota é determinada para uma ocorrência denominada principal (XA), no caso, 0.32, e as demais ocorrências são comparadas com esta. De acordo com a proximidade das demais ocorrências com a principal, uma nota é atribuída. Uma nota mínima pode ser definida para que casos aparentemente duvidosos não sejam considerados como duplicados.

Analisando os resultados obtidos no primeiro critério definido, notou-se que este era insuficiente, pois haviam algumas ocorrências com CPFs diferentes que caracterizavam duplicidades. Por isso, um segundo critério foi definido, considerando os casos de ocorrências com CPFs diferentes, mas com os campos Nome e Endereços similares ou aproximados.

Este critério foi implementado na ferramenta utilizando os parâmetros apresentados na tabela 4.22. Analisando cada um dos parâmetros temos que, o tipo de identificação foi mantido (*UNDUP*) por se tratar do mesmo processo, e como se determinou que o CPF poderia ser diferente, foi necessário definir outro campo para compor o bloco, definiu-se então que, o primeiro caractere do campo Nome seria utilizado na geração dos grupos.

```
Results for match: CRITERI1
Passes completed : 1
Page: 1

Results from pass: 1
     Match cutoff:    0.00   Clerical cutoff:    0.00
TYPE     WGT FLAGS   RECNUM   NOME

[XA]    0.32            1 PAULO SERGIO GONZAGA
[DA]    0.00            2 PAULO S GONZAGA


************************************************
*
*  OUTPUT STATISTICS FOR REPORT GENERATOR: CRITERI1
*  PASSES COMPLETED: 1
*
*       2  Records written to file: c:\Projetos\DEMO\Data
*
*       1  REGULAR  records written
*       0  CLERICAL duplicates written
*       1  DUPLICATE 'A' records written
*
************************************************

************************************************
*
*  STATISTICS FOR MATCH: CRITERI1
*  PASSES COMPLETED: 1
*
*       5  Total A records
*
* Totals for pass: 1
*
* NOTE: REGULAR records are passed to next pass as residual
*
*       5  A records read
*       4  Blocks processed
*       0  Overflow blocks
*       2  Max A block size
*     1.3  Average A block size
*       1  REGULAR records
*       0  Clerical duplicates
*       1  A duplicates
*       0  EXACT A duplicates
*       4  A residuals (including REGULAR records)
*
*
* Totals for all passes
*
*       4  Blocks processed
*       0  Overflow blocks
*       1  A duplicates
*       4  A residuals (including REGULAR records)
```

Figura 4. 3 - Relatório de Consolidação - Critério 1

Os campos Nome e Endereço foram utilizados durante as comparações alfanuméricas (*UNCERT*), considerando-se que, a probabilidade do conteúdo estar correto era de 50% (m-prob 0.5) e a probabilidade de que os conteúdos dos campos estariam com a escrita diferente (Param1 700).

Variável	Valor
Tipo de Identificação	UNDUP
Bloco	NOME1 (Primeiro caractere do campo Nome)
Tipo de comparação aproximada Parâmetros	UNCERT Campo NOME m-prob 0.5 u-prob 0.01 Param1 800 UNCERT Campo ENDEREÇO m-prob 0.8 u-prob 0.01 Param1 700

Tabela 4. 22 - Parâmetros definidos para critério 2

Após a execução e o refinamento deste segundo critério foi gerado um arquivo contendo as ocorrências identificadas como duplicadas, conforme tabela 4.23, além de um relatório apresentando as notas atribuídas para cada ocorrência (figura 4.4).

Código Cliente	CPF	Nome	Endereço	Data Nasc	Dup
1	25421212449	PAULO SERGIO GONZAGA	AV PAULISTA 50	12121960	**1**
2	25421212449	PAULO S GONZAGA	AL SANTOS 34	01011900	**2**
3	25421212488	PAULO SERGIO GONZAGA	PAULISTA 50	12121960	**1**
4	11111111111	PAULO S GONZAGA	AV PAULISTA 50	12121960	**1**
5	25421212567	PAULO S GONZAGA	BRIG TOBIAS 2	01011900	**3**

Tabela 4. 23 - Resultados da consolidação utilizando critério 2

Novamente, após a análise dos resultados produzidos, os resultados ainda não foram satisfatórios e foi necessário mais um critério para garantir que a maioria das ocorrências de duplicidade estavam sendo consideradas. Foi gerado um terceiro critério para abranger os casos em que os cincos primeiros dígitos do campo Cpf eram idênticos (Bloco) e o campo Nome era similar ou aproximado. Os parâmetros utilizados após o processo de refinamento e testes estão apresentados na tabela 4. 24.

```
Results for match: CRITER2
Passes completed : 1
Page: 1

Results from pass: 1
     Match cutoff:    0.00   Clerical cutoff:    0.00
TYPE     WGT FLAGS   RECNUM   NOME                END

[XA]    1.32            1 PAULO SERGIO GONZAGA  AV PAULISTA 50
[DA]    0.48            3 PAULO SERGIO GONZAGA  PAULISTA 50
[DA]    1.00            4 PAULO S GONZAGA       AV PAULISTA 50


 ************************************************
 *
 *  OUTPUT STATISTICS FOR REPORT GENERATOR: CRITER2
 *  PASSES COMPLETED: 1
 *
 *       3  Records written to file: c:\Projetos\DEMO\Data\CRITER2
 *
 *       1  REGULAR  records written
 *       0  CLERICAL duplicates written
 *       2  DUPLICATE 'A' records written
 *
 ************************************************
 ************************************************
 *
 *  STATISTICS FOR MATCH: CRITER2
 *  PASSES COMPLETED: 1
 *
 *       5  Total A records
 *
 * Totals for pass: 1
 *
 * NOTE: REGULAR records are passed to next pass as residuals
 *
 *       5  A records read
 *       1  Blocks processed
 *       0  Overflow blocks
 *       5  Max A block size
 *     5.0  Average A block size
 *       1  REGULAR records
 *       0  Clerical duplicates
 *       2  A duplicates
 *       0  EXACT A duplicates
 *       3  A residuals (including REGULAR records)
 *
 * Totals for all passes
*       1  Blocks processed
 *       0  Overflow blocks
 *       2  A duplicates
 *       3  A residuals (including REGULAR records)
```

Figura 4. 4 - Relatório de consolidação – Critério 2

Variável	Valor
Tipo de Identificação	UNDUP
Bloco	CPF5 (Campo CPF com 5 posição)
Tipo de comparação aproximada Parâmetros	UNCERT Campo NOME m-prob 0.5 u-prob 0.01 Param1 800

Tabela 4. 24 - Parâmetros utilizados no critério 3

Após a execução do terceiro critério, um arquivo com as ocorrências duplicadas (tabela 4.25) e um relatório (figura 4.5) apresentando as notas atribuídas para cada ocorrência foram gerados.

Cód_Cliente	CPF	Nome	Endereço	DataNasc	Dup
1	25421212449	PAULO SERGIO GONZAGA	AV PAULISTA 50	12121960	**1**
2	25421212449	PAULO S GONZAGA	AL SANTOS 34	01011900	**1**
3	25421212488	PAULO SERGIO GONZAGA	PAULISTA 50	12121960	**1**
4	11111111111	PAULO S GONZAGA	AV PAULISTA 50	12121960	**2**
5	25421212567	PAULO S GONZAGA	BRIG TOBIAS 2	01011900	**1**

Tabela 4. 25 - Resultados da consolidação utilizando critério 3

```
Results for match: CRITER3
Passes completed : 1
Page: 1

Results from pass: 1
     Match cutoff:    0.00   Clerical cutoff:    0.00
TYPE    WGT FLAGS   RECNUM   NOME

[XA]   0.32            1 PAULO SERGIO GONZAGA
[DA]   0.00            2 PAULO S GONZAGA
[DA]   0.32   X         3 PAULO SERGIO GONZAGA
[DA]   0.00            5 PAULO S GONZAGA

************************************************
*
*  OUTPUT STATISTICS FOR REPORT GENERATOR: CRITER3
*  PASSES COMPLETED: 1
*
*       4  Records written to file: c:\Projetos\DEMO\Data
*
*       1  REGULAR  records written
*       0  CLERICAL duplicates written
*       3  DUPLICATE 'A' records written
*
************************************************
************************************************
*
*  STATISTICS FOR MATCH: CRITER3
*  PASSES COMPLETED: 1
*
*       5  Total A records
*
* Totals for pass: 1
*
* NOTE: REGULAR records are passed to next pass as residual
*
*       5  A records read
*       2  Blocks processed
*       0  Overflow blocks
*       4  Max A block size
*     2.5  Average A block size
*       1  REGULAR records
*       0  Clerical duplicates
*       3  A duplicates
*       1  EXACT A duplicates
*       2  A residuals (including REGULAR records)
* Totals for all passes
*       2  Blocks processed
*       0  Overflow blocks
*       3  A duplicates
*       2  A residuals (including REGULAR records)
```

Figura 4. 5 - Relatório gerado critério 3

Após refinar cada um dos três critérios individualmente, o processo seguinte foi gerar um processo de identificação que englobasse todos os casos desejados. Para isso foi necessário utilizar os três critérios em conjunto, sendo que o resultado obtido seria a soma de cada um dos critérios executados separadamente. Dessa forma os parâmetros finais utilizados para realizar esse processo são apresentados na tabela 4.26 e o arquivo com todos os registros devidamente identificados são apresentados na tabela 4.27.

Variável	**Valor**
Tipo de Identificação	UNDUP
Bloco 1	**CPF**
Tipo de comparação aproximada Parâmetros	UNCERT Campo NOME m-prob 0.5 u-prob 0.01 Param1 700
Bloco 2	**NOME1 (Campo nome com 1 posição)**
Tipo de comparação aproximada Parâmetros	UNCERT Campo NOME m-prob 0.5 u-prob 0.01 Param1 800 UNCERT Campo ENDEREÇO m-prob 0.8 u-prob 0.01 Param1 700
Bloco 3	**CPF5 (Campo CPF com 5 posição)**
Tipo de comparação aproximada Parâmetros	UNCERT Campo NOME m-prob 0.5 u-prob 0.01 Param1 800

Tabela 4. 26 - Parâmetros finais

Código Cliente	**CPF**	**Nome**	**Endereço**	**Data Nasc**	**Dup**
1	25421212449	PAULO SERGIO GONZAGA	AV PAULISTA 50	12121960	**1**
2	25421212449	PAULO S GONZAGA	AL SANTOS 34	01011900	**1**
3	25421212488	PAULO SERGIO GONZAGA	PAULISTA 50	12121960	**1**
4	11111111111	PAULO S GONZAGA	AV PAULISTA 50	12121960	**1**
5	25421212567	PAULO S GONZAGA	BRIG TOBIAS 2	01011900	**1**

Tabela 4. 27 - Resultados da consolidação utilizando todos os critérios

O relatório (figura 4.6) gerado apresenta as ocorrências identificadas como duplicadas em cada um dos três blocos definidos e as notas atribuídas, conforme explicado anteriormente. Através da análise deste relatório é possível verificar se os critérios definidos estão realmente identificando ocorrências duplicadas ou gerando falsos duplicados. Falsos duplicados são aquelas ocorrências que atendem as exigências impostas através dos parâmetros definidos, mas não são casos de duplicados. O refinamento dos critérios demanda uma grande quantidade de tempo até atingir uma solução satisfatória, como no caso apresentado.

```
Results for match: UNDUP
Passes completed : 3
Page: 1

Results from pass: 1
     Match cutoff:    0.00   Clerical cutoff:    0.00
TYPE     WGT FLAGS   RECNUM   NOME                              END

[XA]    0.32             1 PAULO SERGIO GONZAGA                 AV PAULISTA 50
[DA]    0.00             2 PAULO S GONZAGA                   AL SANTOS 34
Results from pass: 2
     Match cutoff:    0.00   Clerical cutoff:    0.00
[XA]    0.64             1 PAULO SERGIO GONZAGA                 AV PAULISTA 50
[DA]    0.45             3 PAULO SERGIO GONZAGA                 PAULISTA 50
[DA]    0.32             4 PAULO S GONZAGA                   AV PAULISTA 50
Results from pass: 3
     Match cutoff:    0.00   Clerical cutoff:    0.00
[XA]    0.32             1 PAULO SERGIO GONZAGA                 AV PAULISTA 50
[DA]    0.00             5 PAULO S GONZAGA                   BRIG TOBIAS 2

 ************************************************
 *  OUTPUT STATISTICS FOR REPORT GENERATOR: UNDUP
 *  PASSES COMPLETED: 3
 *
 *       7  Records written to file: c:\Projetos\DEMO\Data\UNDUP.RPT
 *
 *       3  REGULAR  records written
 *       0  CLERICAL duplicates written
 *       4  DUPLICATE 'A' records written
 *
 ************************************************
 ************************************************
 *
 *  STATISTICS FOR MATCH: UNDUP
 *  PASSES COMPLETED: 3
 *
 *       5  Total A records
 * Totals for pass: 1
 *
 * NOTE: REGULAR records are passed to next pass as residuals
 *       5  A records read
 *       4  Blocks processed
 *       0  Overflow blocks
 *       2  Max A block size
 *     1.3  Average A block size
 *       1  REGULAR records
 *       0  Clerical duplicates
 *       1  A duplicates
 *       0  EXACT A duplicates
 *       4  A residuals (including REGULAR records)
 *
```

```
* Totals for pass: 2
*
* NOTE: REGULAR records are passed to next pass as residuals
*
*        4  A records read
*        1  Blocks processed
*        0  Overflow blocks
*        4  Max A block size
*      4.0  Average A block size
*        1  REGULAR records
*        0  Clerical duplicates
*        2  A duplicates
*        0  EXACT A duplicates
*        2  A residuals (including REGULAR records)
*
* Totals for pass: 3
*
* NOTE: REGULAR records are passed to next pass as residuals
*
*        2  A records read
*        1  Blocks processed
*        0  Overflow blocks
*        2  Max A block size
*      2.0  Average A block size
*        1  REGULAR records
*        0  Clerical duplicates
*        1  A duplicates
*        0  EXACT A duplicates
*        1  A residuals (including REGULAR records)
*
*
* Totals for all passes
*
*        6  Blocks processed
*        0  Overflow blocks
*        4  A duplicates
*        1  A residuals (including REGULAR records)
```

Figura 4. 6 - Relatório final

Ao término deste processo, as ocorrências duplicadas foram identificadas e receberam um novo código, mas para manter a compatibilidade com os sistemas legados, foi gerado também um arquivo de referências (tabela 4.28) contendo os códigos que identificam unicamente a ocorrência na fonte de dados original, como por exemplo, o campo Código do Cliente, e os novos códigos gerados durante o processo de identificação (Dup).

Dup	Cod_Cliente
1	1
1	2
1	3
1	4
1	5

Tabela 4. 28 - Arquivo com referências

É possível avaliar o ganho do processo analisando a quantidade de registros duplicados por fonte. A ferramenta permite gerar, além de um arquivo contendo as ocorrências identificadas como duplicadas, um grupo de informações denominadas suspeitas, que representam ocorrências identificadas com uma baixa confiabilidade e que requerem um tratamento especial e manual. A tabela 4.29 apresenta um exemplo dos resultados finais obtidos ao término do processo de consolidação.

Quantidade		
Total de Registros	5.000.000	**%**
Duplicados	600.000	12.00
Suspeitos	50.000	1.00

Tabela 4. 29 - Resultados após consolidação

4.7.6.Transformar e Enriquecer os dados (P4.7)

O objetivo deste processo pode ser dividido em duas etapas: transformar as informações para adequá-las à estrutura de dados destino e depois enriquecer os dados utilizando uma fonte de dados confiável.

A primeira etapa deste processo é ajustar as informações à estrutura de dados destino, para isso, após a geração do arquivo identificado no processo anterior (P4.6), é necessário realizar um processo de "sobrevivência". Este processo tem como objetivo gerar uma única ocorrência para cada grupo de ocorrências identificadas como duplicadas pelo processo de consolidação utilizando critérios específicos. Após várias reuniões com os especialistas na informação, os critérios utilizados na escolha da melhor informação foram definidos conforme a tabela 4.30.

Campo	**Critério**
Cpf	Selecionar o Cpf que possuir a data de atualização mais recente
Nome	Selecionar o campo Nome que possuir o conteúdo de maior comprimento
Endereço	Selecionar o campo Endereço que possuir o conteúdo de maior comprimento
DataNascimento	Selecionar a Data de Nascimento mais freqüente

Tabela 4. 30 - Critérios para consolidação

O resultado obtido utilizando os critérios determinados está apresentado na tabela 4.31, considerando que, para cada um dos campos deste arquivo foi aplicada uma forma distinta de escolha da melhor informação.

Cod_Novo	CPF	Nome	Endereço	Data Nasc
1	25421212449	PAULO SERGIO GONZAGA	AV PAULISTA 50	12121960

Tabela 4. 31 - Resultado da consolidação

Aplicando este processo de sobrevivência das informações para cada uma das fontes, o problema foi resolvido de maneira isolada, pois ainda restava solucionar o problema das duplicidades existentes entre as fontes. Para isso novamente o processo de identificação foi utilizado e outros critérios de consolidação foram definidos. Porém os critérios a serem aplicados na escolha da melhor informação já foram definidos no processo 4.1 e são apresentados de forma simplificada na tabela 4.32. Durante este processo, a ordem de prioridade de escolha da melhor informação para cada um dos campos é determinada por um número, como por exemplo, no caso do campo CPFCGC; se o campo existir na Fonte A, o valor contido nesta fonte sempre será considerado o melhor, caso contrário, será o conteúdo da Fonte B e a seguir da Fonte C, diferindo do comportamento do campo Nome, onde o conteúdo da Fonte B é o mais importante, seguido da Fonte B e C.

Nome do Campo	Fonte A	Fonte B	Fonte C
CPFCGC	1	2	3
Nome	2	1	3
Endereço	2	1	3
CEP	2	1	3
Bairro	2	1	3
Município	2	1	3
Estado	2	1	3
Sexo	1	2	-
DataNascimento	1	2	-
Telefone	2	3	1

Tabela 4. 32 - Prioridade entre as fontes de informação

Ao término deste processo, todas as informações foram adequadas à estrutura de dados destino. O próximo processo foi efetuar o enriquecimento. O processo de

enriquecimento utiliza os mesmos princípios empregados durante o processo de consolidação (tipo de comparação, blocos, comparações, etc) apresentados anteriormente no processo 4.6. O enriquecimento foi aplicado aos campos Endereço e Telefone utilizando dados dos Correios e da operadora de telefonia. Antes de utilizar essas informações, elas foram devidamente analisadas e padronizadas para viabilizar os processo de identificação e enriquecimento das informações.

Nos campos de endereço, bairro, cidade e estado, o enriquecimento foi efetuado utilizando como bloco o campo Cep e comparação aproximada do Endereço. No caso de identificar o endereço correspondente na base dos Correios, os campos Bairro, Cidade e Estado foram substituídos com as informações obtidas nos Correios, conforme figura 4.7.

Já o campo Telefone foi enriquecido utilizando informações da operadora de telefonia. A melhoria de qualidade deste campo ocorreu principalmente no prefixo, pois foi possível ajustá-lo de acordo com a estruturação ocorrida nacionalmente. Além disso foi possível identificar o DDD correto para esse telefone de acordo com o prefixo e Cidade/Estado do endereço.

SAÍDA PADRONIZADA DOS DADOS DOS CORREIOS							DADOS NÃO PADRONIZADOS	
Tipo	Titulo	Logradouro	Numero	Complemento	Cidade	UF	CEP	Bairro
R		SANTOS			SAO PAULO	SP	02422180	PAULISTA
AV		BRASIL			SAO PAULO	SP	03030001	JD EUROPA
AV	DR	ARNALDO			SAO PAULO	SP	03434001	JARDINS

SAÍDA PADRONIZADA DOS DADOS DE ENTRADA							DADOS NÃO PADRONIZADOS	
Tipo	Titulo	Logradouro	Numero	Complemento	Cidade	UF	CEP	Bairro
AV		ARNALDO	456	APTO 44	SAO PAULO		03434001	
R		SANTOS	22				02422180	

SAÍDA PADRONIZADA E ENRIQUECIDA COM CORREIO							DADOS NÃO PADRONIZADOS	
Tipo	Titulo	Logradouro	Numero	Complemento	Cidade	UF	CEP	Bairro
AV	DR	ARNALDO	456	APTO 44	SAO PAULO	SP	03434001	JARDINS
R		SANTOS	22				02422180	

Figura 4. 7 - Enriquecimento utilizando dados dos Correios

Depois de efetuar o processo de enriquecimento nas fontes de informação foi possível mensurar os benefícios deste processo, conforme apresentado na tabela 4.33, principalmente analisando o campo Município que, antes deste processo, estava

preenchido em somente 3.5% dos casos, e após o enriquecimento, em 60% dos casos, estava preenchido.

		Preenchidos		% Válidos		% Inválidos		Enriquecidos	
Campo	**Total_Regs**	**Qtd**	**%**	**Qtd**	**%**	**Qtd**	**%**	**Qtd**	**%**
Endereço	5.000.000	4.977.000	99.54	4.970.530	99.87	6.470	0.13	2.986.200	60.00
Bairro	5.000.000	4.925.000	98.50	4.920.075	99.90	4.925	0.10	2.905.750	59.00
Município	5.000.000	175.000	3.50	173.740	99.28	1.260	0.72	2.986.200	60.00
Estado	5.000.000	2.872.500	57.45	2.843.775	99.00	28.725	1.00	2.986.200	60.00
Telefone	5.000.000	350.000	7.00	245.000	70.00	105.000	30.00	120.000	2.40

Tabela 4. 33 - Resultados após enriquecimentos

Ao término deste processo as informações se encontravam adequadas à estrutura de dados destino e os campos Endereço, Bairro, Município, Estado e Prefixo e DDD do Telefone enriquecidos com dados dos Correios e da operadora de telefonia.

4.7.7.Calcular derivações e sumarizar dados (P4.8)

O objetivo deste processo é gerar dados derivados ou sumarizados para otimizar o desempenho durante consultas aos dados relacionados com valores. Este processo não foi implementado no estudo de caso, pois até o momento não foi necessário realizar nenhuma sumarização ou derivação, mas pode ser que ainda haja a necessidade de incluir esses processos.

4.7.8.Auditar e controlar a extração, transformação e carga dos dados (P4.9)

O objetivo deste processo é garantir que as informações corretas estão sendo extraídas, transformadas e carregadas no seu destino. Para isso, todos os processos envolvidos no projeto de qualidade de dados (padronizações, enriquecimentos, identificação, eliminação de duplicados, consolidação e carga), após terem sido

implementados e validados no ambiente de desenvolvimento, foram cadastrados em uma ferramenta de controle de execução de programas já existente utilizada na empresa. Além disso, durante os 3 primeiros meses, todos os históricos de execução (*logs*) dos processos de qualidade, desde a geração do arquivo origem dos dados até a carga na base de dados do DW (*DataWarehouse*), foram verificados e analisados, e os resultados produzidos foram enviados para os especialistas dos dados para verificação e validação.

4.7.9.Analisar os tipos de defeitos (P4.6)

O objetivo deste processo é registrar e analisar os principais problemas de qualidade detectados durante o processo de reengenharia (P4). Para fornecer subsídios para futuras análises e, conseqüentemente, melhorias nos processos, durante a implementação do projeto, foram criados alguns mecanismos (campos e arquivos) indicando se as informações originais foram modificadas (padronizadas, corrigidas, complementadas, enriquecidas, etc) para que o processo de qualidade pudesse ser acompanhado e continuamente melhorado. Através desses mecanismos as origens dos problemas mais freqüentes podem ser encontradas e possivelmente acertadas para que os erros sejam minimizados.

4.8. Problemas Encontrados

Durante a implementação do estudo de caso utilizando a metodologia TQdM foram detectados alguns pontos críticos apresentados a seguir, porém as principais dificuldades encontradas durante a implementação ocorreram durante o processo de Reengenharia e Limpeza dos dados (Processo 4), pois este processo foi implementado para cada uma das fontes de informação envolvidas na geração do cadastro unificado.

4.8.1.Avaliação da qualidade dos dados

O processo de extração e análise dos dados das fontes de informação (P4.2) foi realizado com o objetivo isolado de analisar a qualidade dos dados e produzir relatórios estatísticos, quando poderia ter sido executado sob uma visão mais ampla e atender às necessidades das etapas seguintes do projeto, como os processos de padronização, correção, complementação, consolidação e enriquecimento.

Por esse motivo, durante as fases seguintes, sentiu-se a necessidade de relatórios específicos para cada fase e foi preciso repetir esse processo de análise ou investigação nas fontes de informação novamente gerando retrabalho.

4.8.2. Estudo de Viabilidade

A metodologia TQdM sugere a utilização de um plano de ação (P6) para a implementação de um projeto e este plano inclui a realização de um projeto piloto em uma área estratégica e gerenciável para validar e testar a solução a ser utilizada. Conforme apresentada na figura 4.8, durante a implementação do estudo de caso, logo após o processo de medição dos custos resultantes da falta de qualidade dos dados (P3), foi decidido que as informações de baixa qualidade (Produto) seriam ajustadas através de um processo de reengenharia e limpeza dos dados. Antes de implementar este processo, notou-se a falta de um estudo de viabilidade para verificar se a ferramenta a ser utilizada durante a implementação do projeto atendia às necessidades e expectativas, e se seria compatível com o volume de informações a serem tratadas durante este processo.

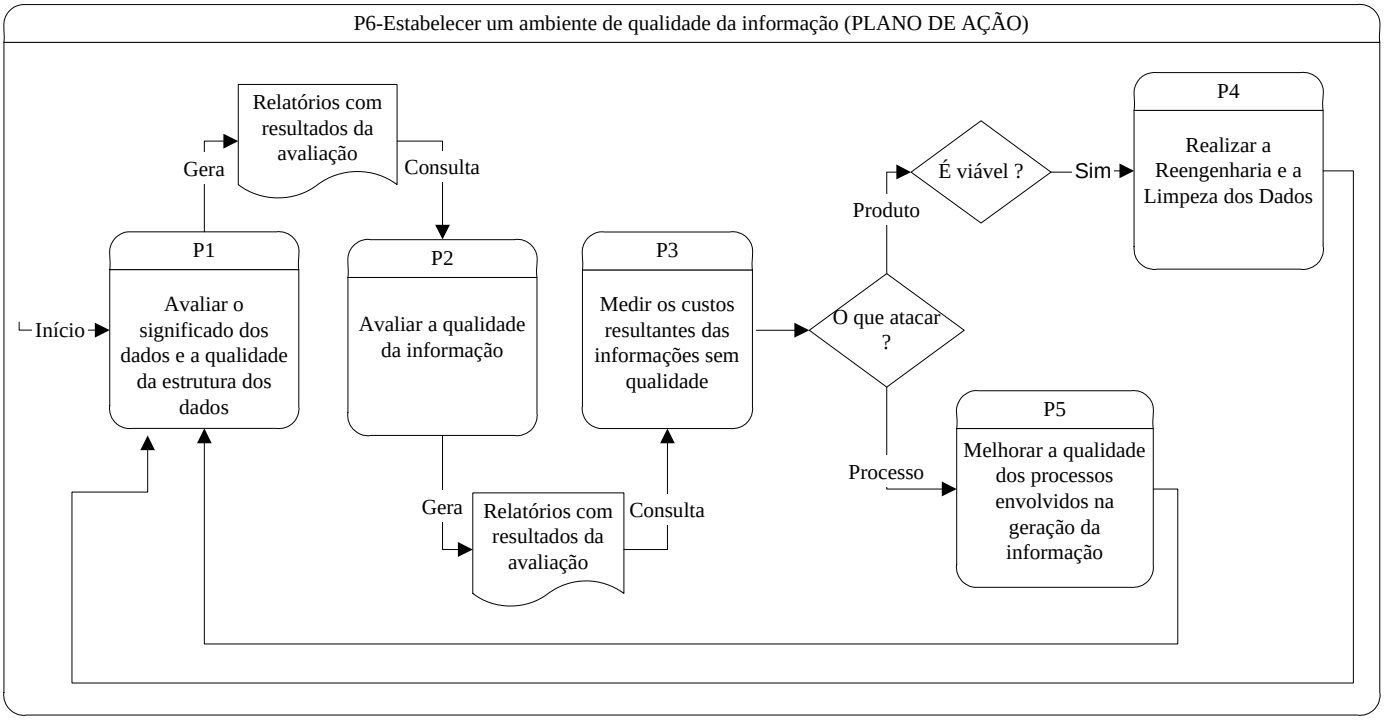

Figura 4. 8- Metodologia TQdM aplicada ao caso prático[7]

4.8.3. Regras ou padrões

As regras ou os padrões foram criados com o objetivo principal de equalizar as diversas informações entre as diferentes fontes de dados, por isso nos processos posteriores notou-se a necessidade de ajustes nos padrões para aumentar a efetividade dos processos de consolidação e enriquecimento. Novamente, ocorreu o mesmo problema apresentado no caso da avaliação da qualidade dos dados. Quando alguma falha na definição e implementação dos padrões era notada, já não havia mais tempo hábil para refazer processos anteriores. Para suprir a falta de padronização de alguns campos, como por exemplo, os campos Apartamento e Casa, um trabalho redobrado de análise foi gerado nas fases de identificação e detecção de duplicados.

4.8.4.Paralelismo de atividades

O estudo de caso apresentado implementa várias vezes o processo de reengenharia e limpeza de dados (processo 4), uma vez para cada uma das fontes de informação; por este motivo, durante a fase de implementação as atividades envolvidas foram executadas em paralelo visando otimizar o tempo despendido; por exemplo, um analista foi responsável pelo processo de reengenharia da Fonte A, outro, da Fonte B e outro, da Fonte C. Embora o tempo de implementação dos processos de melhoria da qualidade das fontes tenha sido reduzido e otimizado, o resultado não foi satisfatório. O principal problema foi à integração dos módulos; foram realizados vários testes, mas nenhum teste em conjunto, e quando foi possível a sua realização, detectaram-se vários problemas, como por exemplo, *layouts* discrepantes de arquivos que deveriam ser equivalentes, implementações a princípio equivalentes estavam diferentes, regras de padronização em versões diferentes, enfim, várias diferenças e conseqüente perda de informação.

4.8.5.Controle dos Processos

A metodologia apresenta todos os processos envolvidos em um projeto de qualidade de dados e como devem ser implementados, porém, para torná-la prática seria importante à criação de um esquema ou um diagrama de processos mais detalhado para representar e documentar exatamente como os processos devem ser implementados. Porém diagramas de processos isolados não seriam suficientes para controlar todos os processos.

Notou-se a necessidade de mecanismos para documentar a análise e a execução dos processos de identificação e detecção de duplicados, principalmente porque, durante o processo de refinamento dos critérios, vários testes foram efetuados, mas não houve compartilhamento com toda a equipe, gerando retrabalho e perda do conhecimento adquirido.

4.8.6.Níveis mínimos de qualidade

Durante a implementação deste estudo de caso foi possível notar que é muito importante a definição de um nível mínimo de qualidade aceitável para uma determinada fonte de informação, pois no decorrer do projeto, caso este nível mínimo não tenha sido apresentado e conhecido por todos os envolvidos, todo o trabalho para a qualidade dos dados pode ser comprometido se em alguma fonte de informação esta qualidade mínima não for atingida.

4.8.7. Projeto Modelo

Ao término da implementação da metodologia TQdM, sentiu-se a necessidade de divulgar o conhecimento adquirido, com a finalidade de que todo o trabalho e tempo investidos não ficasse perdido e esquecido pelos demais departamentos da empresa. Notou-se que seria importante a geração de algum projeto modelo para que as futuras implementações pudessem utilizá-lo para implementar a metodologia TQdM.

4.9. Conclusão

Neste capítulo o estudo de caso foi apresentado, todos os processos utilizados durante a implementação da metodologia TQdM foram detalhados e as dificuldades encontradas durante a execução do projeto foram apresentadas. No próximo capítulo serão apresentadas algumas propostas de melhorias na metodologia utilizada visando sanar os problemas detectados durante a sua implementação.

Capítulo 5 - Soluções Propostas

5.1. Introdução

Este capítulo apresenta algumas sugestões de melhorias para a metodologia TQdM visando minimizar os problemas identificados durante o estudo de caso. Alguns destes problemas ocorrem devido a uma avaliação incompleta da qualidade dos dados (P4.2), à ausência de uma fase para verificar a viabilidade da solução técnica proposta, às insuficientes definições de regras ou padrões (P4.3), à ausência de mecanismos para controle dos processos, à ausência de definições dos níveis mínimos de qualidade a serem adotados, e principalmente, à ausência da implementação de um projeto piloto para validar todo a solução definida e para servir de modelo para futuros processos de qualidade de dados.

As sugestões de melhorias, que são apresentadas em detalhes a seguir, envolvem a criação de novos processos, a divisão e inversão de processos já existentes e a utilização de formulários padrões para documentar os resultados obtidos.

5.2. Visão geral

A metodologia TQdM, apresentada na figura 5.1, é composta por cinco processos responsáveis por avaliar o significado dos dados e a qualidade da estrutura dos dados (P1), avaliar a qualidade da informação (P2), medir os custos das informações sem qualidade (P3), realizar a reengenharia e a limpeza dos dados (P4) e melhorar a qualidade dos processos envolvidos na geração da informação (P5). Além de um sexto processo, denominado de plano de ação (P6), responsável por estabelecer um ambiente de qualidade da informação.

Logo após a execução do processo responsável pela medição dos custos (P3) é possível identificar qual atitude deve ser tomada para melhorar a qualidade dos dados. Há dois caminhos possíveis: realizar a reengenharia e limpeza dos dados ou ajustar os processos envolvidos na geração da informação. Antes de iniciar a

implementação de um desses processos, sentiu-se a necessidade de um novo processo responsável por validar a solução técnica adotada, por isso um novo processo denominado Estudo de viabilidade técnica (P7) foi implementado.

O objetivo deste processo é validar a efetividade da solução técnica escolhida. Esta solução pode ser implementada para melhorar a qualidade dos dados, realizada pelo processo de reengenharia ou limpeza de dados (Processo 4 – figura 5.1), ou para melhorar a qualidade do processo que originou as informações de baixa qualidade (Processo 5 – figura 5.1).

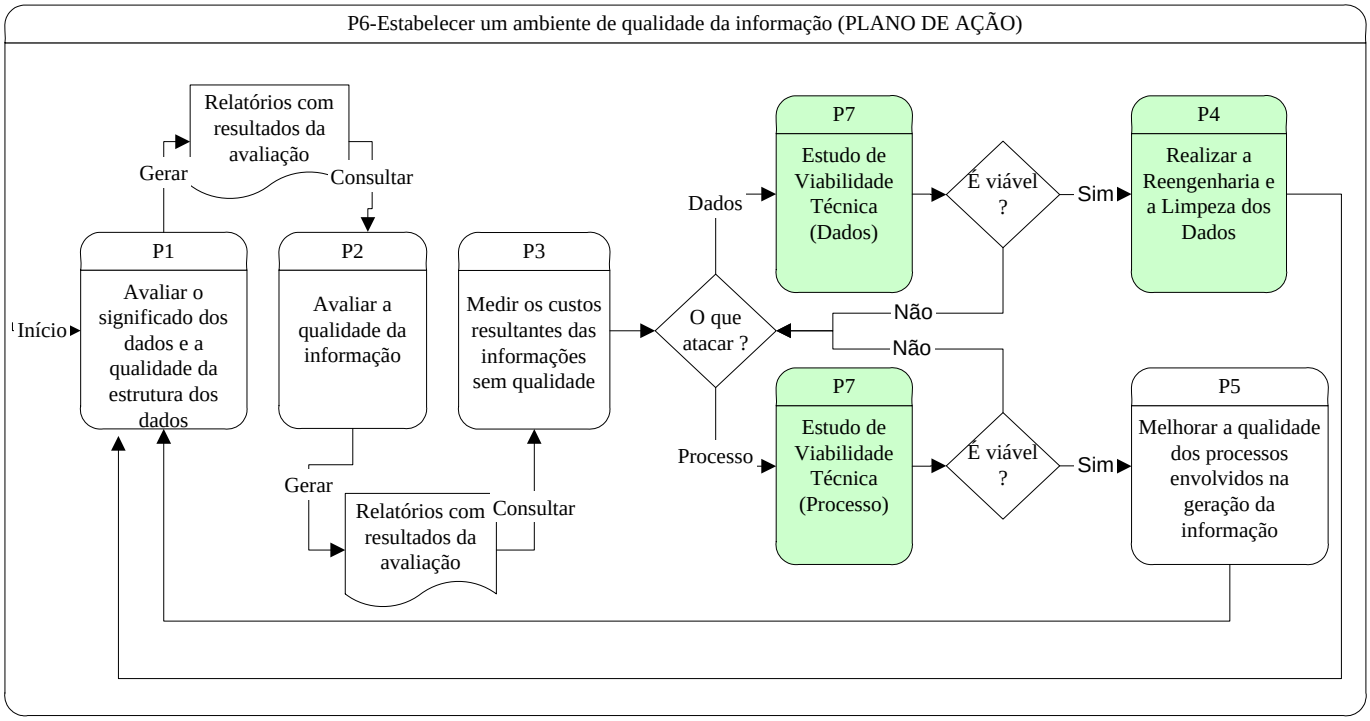

Figura 5. 1 - Sugestão de novo processo na metodologia TQdM

As atividades realizadas durante este processo envolvem a instalação e configuração, em ambiente real, da ferramenta e dos utilitários necessários, como por exemplo, um programa para ordenação dos dados. Após a instalação, é possível implementar uma demonstração da solução com todas as principais funcionalidades da ferramenta escolhida.

No estudo de caso, uma vez instalada a ferramenta, um processo foi construído e executado utilizando volumes de dados reais para validar a ferramenta de reengenharia. Ao término da implementação deste processo foi possível concluir

que a ferramenta era adequada ao volume de informações que deveriam ser tratadas, além de ser eficiente e eficaz durante os processos de padronização, enriquecimento e identificação de informações duplicadas. Após a validação da solução técnica, iniciou-se o processo de reengenharia e limpeza dos dados que é descrito a seguir.

5.3. Processo de Reengenharia e Limpeza dos dados (P4)

Durante o estudo de caso, o processo de reengenharia e limpeza dos dados foi implementado visando melhorar a qualidade dos dados. Os problemas de qualidade identificados nos processos anteriores (P1 e P2) são tratados nesta fase.

O processo de reengenharia e limpeza dos dados da metodologia TQdM implementado no estudo de caso, apresentado anteriormente na figura 3.1, é composto por nove sub-processos responsáveis por identificar as possíveis fontes de dados (P4.1), extrair e analisar os dados das fontes (P4.2), padronizar os dados (P4.3), corrigir e complementar os dados (P4.4), consolidar os dados (P4.5), transformar e enriquecer os dados (P4.7), calcular derivações ou sumarizações dos dados (P4.8), auditar e controlar a extração, transformação e carga dos dados (P4.9) e analisar os tipos de defeito (P4.6).

A figura 5.2 apresenta a metodologia TQdM com algumas modificações em virtude das sugestões de melhorias visando solucionar os problemas identificados durante o estudo de caso. O processo responsável pela padronização dos dados (P4.3) é representado pelos processos na cor rosa (P4.3.1 e P4.3.3). O processo responsável pela transformação e enriquecimento dos dados (P4.7) é representado pelos processos na cor azul (P4.7.1 e P4.7.2), sendo que, na metodologia original este processo encontra-se logo após o processo de consolidação dos dados (P4.5). O processo na cor verde (P4.3.2) representa um novo processo responsável por agrupar as fontes de informação com as mesmas características e planejar uma solução padrão para cada grupo, este novo processo foi incorporado ao processo de padronização. A seguir cada um dos sub-processos é apresentado com as devidas modificações.

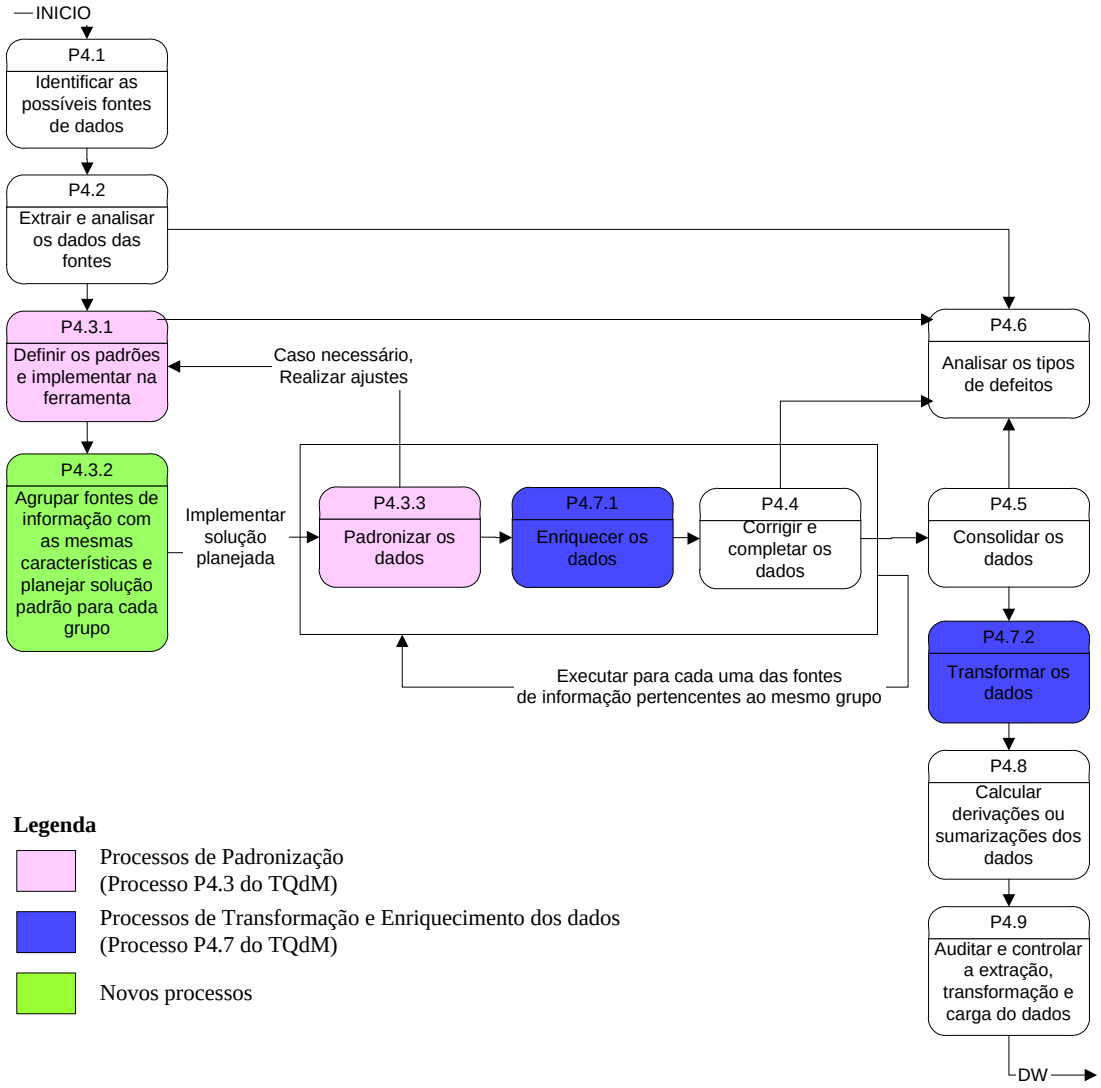

Figura 5. 2 - Sugestões de melhorias no processo de reengenharia e limpeza dos dados(P4)

5.3.1 – Identificar as possíveis fontes de dados (P4.1)

O principal problema identificado durante a implementação deste processo foi à ausência de formulários específicos para realizar a documentação das fontes de informação, e assim, garantir a quantidade de informações mínimas necessárias de cada fonte.

Como sugestão para solucionar este problema, depois de efetuar o levantamento completo, os resultados gerados podem ser armazenados em

formulários específicos e disponibilizados em local de fácil acesso para a equipe toda. Para cada fonte de informação avaliada pode ser preenchido um formulário dedicado, como exemplificado na tabela 5.1, contendo as seguintes informações:

- Nome da fonte que foi avaliada;
- Nome do departamento responsável pela fonte;
- Nome do sistema que originou as informações;
- Nome do sistema que receberá as informações;
- Data em que o levantamento foi efetuado;
- Nome do responsável pela fonte;
- Tipo da fonte; por exemplo, arquivo seqüencial, planilha, tabela de um banco de dados;
- Freqüência com que as informações são atualizadas; como por exemplo atualização diária, mensal, quinzenal, anual.
- Para cada um dos campos desta fonte, informar o nome do campo, uma breve descrição da função deste campo, o tipo do campo (numérico, alfanumérico, data), o tamanho do campo, a máscara utilizada durante o preenchimento das informações (DDMMAAAA, XXXXX, 99999), o domínio de valores que este campo pode conter (“F” ou “M”, 1 ou 0).

DETALHAMENTO DAS FONTES DE INFORMAÇÃO					
Nome Fonte: FONTE_A			**Data:** 01/06/2003		
Departamento: Financeiro			**Responsável:** MARIA SANTOS		
Sistema Origem: Sistema EA			**Tipo Fonte:** Arquivo seqüencial		
Sistema Destino: Sistema EO			**Freqüência Atualização:** Mensal		
Campo	**Descrição**	**Tipo**	**Tamanho**	**Máscara**	**Domínio**
Nome	Nome Cliente	Texto	60	-	-
Endereço	Endereço de Correspondência	Texto	100	-	-
Nascimento	Endereço de Correspondência	Texto	100	-	-

Tabela 5. 1- Exemplo de formulário preenchido para detalhamento das fontes

Após realizar o levantamento das definições de cada uma das fontes de informação, a próxima etapa deste processo é identificar, para cada um dos campos, a ordem de importância das informações dadas as diferentes fontes. A tabela 5.2

apresenta um exemplo de como os resultados desta fase podem ser representados. Lembrando que, os seguintes itens precisam ser verificados:

- Todos os campos analisados estão representados pela primeira coluna deste formulário;
- Todos os nomes das fontes de dados analisadas estão representadas pela primeira linha desta tabela (Fonte_A, Fonte_B, Fonte_n);
- Os diferentes formatos dos campos em cada uma das fontes de informação estão apresentados lado a lado para futura compatibilização;
- A ordem de escolha da melhor informação entre as diferentes fontes é identificada pela numeração ao lado de cada coluna para uma determinada fonte. Quanto menor o número atribuído maior a importância desta fonte para o negócio da empresa.

LISTA DE REFÊNCIAS						
Campo	**Fonte_A**	**nro**	**Fonte_B**	**Nro**	**Fonte_n**	**Nro**
Nome	Texto(60)	1	Texto(30)	3	Texto(40)	2
Endereço	Texto(100)	2	Texto(120)	1	Texto(40)	3
…						

Tabela 5. 2 – Exemplo de formulário preenchido da lista de referências

Ao término deste processo, todas as fontes analisadas se encontram devidamente documentadas permitindo que os resultados obtidos nesta fase possam ser utilizados pelos próximos processos.

5.3.2 – Extrair e analisar os dados dos arquivos (P4.2)

Como ocorreu no processo anterior, o principal problema deste processo foi à falta de formulários padrões para documentar os resultados obtidos durante a análise dos dados e auxiliar a tomada de decisão dos processos seguintes.

A tabela 5.3 apresenta um exemplo de como as análises efetuadas e implementadas na ferramenta de reengenharia podem ser documentadas. As seguintes informações precisam ser levantadas:

- Nome da fonte de informação;
- Os procedimentos implementados na ferramenta são descritos na primeira coluna da tabela:
- O tipo de análise utilizado para avaliação da qualidade (discreta, concatenada, utilizando padrões, etc);
- Nome do campo analisado;
- Máscara utilizada durante a análise, podendo ser caractere, responsável pela análise do conteúdo do campo, ou análise do tipo de dado que analisa o formato do campo ou alguma máscara específica.

Nome da fonte: FONTE_A

Nome Processo	Tipo	Campo	Máscara
ARQ1I001	Discreta	Nome	CCCCCCCCCCCCC
ARQ1I002	Discreta	CPF	CCCCCCCCCCC
ARQ1I003	Discreta	CPF	TTTTTTTTTTT
…			

Tabela 5. 3 – Exemplo de formulário preenchido na avaliação dos dados

Após a execução de todos os processos de análise dos dados, todos os resultados são consolidados e relatórios são gerados para cada fonte de dados. Estes relatórios precisam informar, principalmente, a freqüência de todos os campos, o fator de preenchimento de cada campo, os valores mais freqüentes de cada campo, os formatos mais utilizados durante o preenchimento dos campos, entre outros. Além disso, uma análise comparativa avaliando o fator de preenchido dos campos precisa ser realizada produzindo um relatório geral sobre a qualidade dos campos entre as diversas fontes, conforme apresentado na tabela 5.4. Visando auxiliar os processos de consolidação, seria importante apresentar também o valor de maior e menor freqüência identificados para cada um dos campos. O resultado deste processo será útil durante a implementação dos processos de identificação e consolidação.

Nome Campo	Fonte_A	Fonte_B	Fonte_C	Fonte_n
Nome	90%	87%	45%	n%
CPF	99%	98.5%	50%	n%
…				

Tabela 5. 4 – Exemplo de relatório gera com Campo x Fonte x Fator de preenchimento

5.3.3 – Padronizar os dados (P4.3)

Durante a implementação deste processo em todas as fontes de informação, notou-se que a atividade responsável por definir padrões e implementá-los na ferramenta estava sendo executada de forma repetitiva e ineficiente. Se houvesse uma fase responsável por analisar todas as fontes e gerar padrões genéricos, o processo de definição e implementação dos padrões seria realizado uma única vez atendendo à todas as fontes e somente o processo de padronização seria aplicado em todas as fontes.

A cada implementação de uma nova fonte, notou-se uma similaridade dos processos de padronização, enriquecimento, correção e complementação dos dados entre as fontes. Se estas semelhanças fossem detectadas antes, poderia ter sido planejada uma solução padrão para fontes de um mesmo grupo reduzindo assim o tempo gasto com o planejamento de uma nova solução a cada fonte.

Visando solucionar estes problemas, uma sugestão é dividir o processo de padronização da metodologia TQdM em três partes: a primeira responsável por definir os padrões utilizados e implementá-los na ferramenta (P4.3.1); a segunda responsável por agrupar as fontes de informação com as mesmas características e planejar uma solução padrão para cada grupo (P4.3.2) e a terceira responsável por padronizar os dados (P4.3.3). A seguir cada um destes processos são apresentados em detalhes.

- **Definir os padrões e implementá-los na ferramenta (P4.3.1)**

Este processo tem como objetivo estabelecer uma forma padrão de representação dos dados, para isso se faz necessário estabelecer os padrões ou os formatos desejados para cada campo a ser tratado, e a seguir, implementar estes padrões na ferramenta de reengenharia gerando regras específicas.

Inicialmente, é possível utilizar, caso existam, as regras padrões (*standard*) que a ferramenta possui e aplicá-las às fontes de informação. Os resultados gerados são apresentados aos especialistas do negócio para avaliação e compreensão sobre os conceitos e os processos envolvidos durante a padronização. Após analisar os resultados obtidos, eles estarão aptos para sugerir alguma melhoria nos padrões atuais e, talvez, seja identificada a necessidade de novas regras. Lembrando que, a principal função da padronização é equalizar cada grupo de dados para um único padrão, independente da fonte de onde eles provêem.

Após concluir as alterações e as novas implementações solicitadas, as regras de padronização são aplicadas em todas as fontes de informação, utilizando dados e volumes reais, em seguida, algumas amostras representativas são extraídas para os especialistas validarem as regras definidas. Caso alguma incoerência seja encontrada, os ajustes necessários são efetuados e, novamente, as regras são aplicadas em todas as fontes até atingir o nível de qualidade desejado para a padronização. A etapa seguinte é documentar, de forma detalhada, todas as regras e os padrões definidos e empregados na padronização.

O primeiro formulário preenchido é responsável por especificar as entradas e saídas da regra, ou seja, todos os parâmetros esperados na entrada da regra e todos os campos padronizados gerados, sendo que, para cada um dos campos é informado o tamanho e o formado de saída, conforme exemplificado na tabela 5.5.

Nome Regra	BRNOME
Objetivo	Padronizar os campos nomes
Parâmetros de Entrada	NOME
Saída Padronizada	TITULO Alfanumérico (15) PRIMEIRO_NOME Alfanumérico (25) NOME_MEIO Alfanumérico (50) ULTIMO_NOME Alfanumérico (25) SUFIXO Alfanumérico (15)

Tabela 5. 5 - Exemplo de formulário das entradas e saídas das regras

A fase seguinte é reponsável por documentar os padrões utilizados durante a implementação de cada uma das regras definidas. Para cada campo padronizado é apresentado a relação dos padrões utilizados, conforme apresentado na tabela 5.6. Além disso um documento é gerado para cada regra explicando em detalhes o seu funcionamento e, os tratamentos especiais adotados, como por exemplo, endereçamento da cidade de Brasília que possui regras de composição de logradouros especiais utilizando blocos, lotes, quadras, áreas especiais, etc.

Nome Regra	BRNOME
Campo padronizado	TITULO
Entrada	**Saída Padronizada**
DOUTOR	DR
DOUTORA	DRA
PROFESSOR	PROF
PROFESOR	PROF
PROFESSORA	PROFA
…	
Campo padronizado	SUFIXO
Entrada	**Saída Padronizada**
JUNIOR	JR
NETO	NT
FILHO	FL
…	

Tabela 5. 6 - Exemplo detalhamento das regras

Ao término deste processo, as regras específicas para cada tipo de campo estão prontas para serem utilizadas no processo de padronização (P4.3.3) para todas as fontes de informação. Este processo somente necessita ser reexecutado se durante a padronização dos dados algum problema for identificado que requeira alguma correção.

- **Agrupar as fontes de informação com as mesmas características e planejar uma solução padrão para cada grupo (P4.3.2)**

Este novo processo tem como objetivo agrupar as fontes de informação com as mesmas características e planejar para cada grupo uma solução padrão de

melhoria da qualidade dos dados envolvendo os processos de padronização, enriquecimento, correções e complementações dos dados. Ao agrupar as fontes é importante analisar os campos existentes e preenchidos de cada fonte, verificar se o nível de atualização, de duplicidade e de qualidade das informações são similares, enfim, considerar todos os resultados obtidos durante a fase de análise e de avaliação da qualidade (P4.2).

A fase seguinte é planejar uma solução padrão para cada um dos grupos identificados, esta solução envolve principalmente a implementação do processo de padronização (P4.3.3) utilizando as regras definidas anteriormente (P4.3.1) e as atividades relacionadas com o enriquecimento (P4.7.1), ou seja, quais campos serão enriquecidos, quais fontes fidedignas serão utilizadas, como será o processo de comparação a ser implementado, ou seja, todos os critérios utilizados no processo de enriquecimento são definidos durante esta fase. Esta solução padrão também define quais campos serão corrigidos ou complementados (P4.4), quais os critérios serão utilizados e de que forma serão implementados (automática ou manual).

Após o planejamento da solução padrão, uma fonte de informação de cada grupo é selecionada para implementar a solução definida na ferramenta de reengenharia e validá-la. Os resultados produzidos são apresentados aos responsáveis para análise e, caso haja necessidade, a solução é refinada e implementada até atingir os resultados desejados. Talvez haja necessidade de ajustar as regras definidas no processo 4.3.1. Ao término deste processo será possível obter uma solução padrão validada para a ser aplicada em todas as fontes de informação, além de obter todos os critérios, conforme exemplificado na tabela 5.7, a serem utilizados no enriquecimento, na correção e na complementação dos dados garantindo um nível mínimo de qualidade em todas as fontes.

Solução Padrão 1			
Critério	**Campo**	**Tipo Processo**	**Descrição**
CRIT1	Sexo	Correção	Se a padronização do campo Nome retornar um campo Sexo preenchido, este valor substituirá o valor do campo Sexo da entrada; Caso contrário, será mantido o valor de entrada.
CRIT2	CEP	Enriquecimento	Se os campos Endereço, Cidade e UF forem encontrados no arquivo dos Correios, então o campo CEP encontrado substituirá o valor do campo CEP da entrada; Caso contrário, será mantido o valor de entrada.
CRIT3	Bairro (NOVO)	Complementação	Se os campos Endereço, Cidade e UF forem encontrados no arquivo dos Correios, então o campo Bairro encontrado será incorporado como um campo novo; Caso contrário, será atribuído o valor nulo.

Tabela 5. 7- Exemplo de documentação dos critérios utilizados na solução padrão

- **Padronizar os dados (P4.3.3)**

O objetivo deste processo é realizar a padronização dos dados utilizando as regras definidas anteriormente. O processo consiste em verificar, para cada um dos campos, se existe alguma regra padrão específica para o tipo do campo, se existir, um processo responsável pela padronização precisa ser implementado e executado na ferramenta de reengenharia e documentado informando para cada campo a regra de padronização aplicada, conforme apresentado na tabela 5.8.

Nome Fonte	FONTE_A
Campo	**Regra**
Nome	BRNOME
Endereço	BRENDER
Telefone	BRTEL
…	

Tabela 5. 8 - Exemplo de formulário gerado pela padronização dos dados

Ao término deste processo, todos os dados que necessitavam de padronização, já se encontram devidamente equalizados em um único padrão e aptos a serem utilizados nos processos seguintes.

5.3.4 – Corrigir e completar os dados (P4.4)

O objetivo deste processo é implementar na ferramenta de reengenharia os processos responsáveis por corrigir e complementar os dados utilizando os critérios, previamente, definidos na solução padrão e estabelecidos no processo 4.3.2. A documentação produzida por este processo relaciona o nome do processo implementado na ferramenta com o critério utilizado.

5.3.5 – Consolidar os dados (P4.5)

Este processo tem como objetivo identificar casos de duplicidades em uma mesma fonte de informação ou entre fontes distintas. Durante o estudo de caso, este processo de consolidação ficou comprometido, pois necessitou de informações complementares para auxiliar o processo de identificação que somente seriam obtidas após o processo de enriquecimento (P4.7.1),. Para evitar que estes problemas ocorram novamente, o processo de transformação e enriquecimento dos dados (P4.7) poderia ser dividido e o processo de enriquecimento poderia ser implementado antes do processo de consolidação, conforme apresentado na figura 5.2.

Desta forma, além das informações estarem padronizadas, corrigidas e complementadas, elas estariam enriquecidas com fontes fidedignas, proporcionando assim mais informações a serem utilizadas durante o processo de identificação de duplicados. Os critérios utilizados durante este processo e implementados na ferramenta de reengenharia seriam documentados conforme apresentado na tabela 5.11.

Processo	UNDUP	
Bloco	CPF	
Campo	**Tipo de Comparação**	**Parâmetros**
Primeiro_Nome	UNCERT	m-prob 0.9 u-prob 0.01 Param1 900
Nome_Meio	UNCERT	m-prob 0.9 u-prob 0.01 Param1 850
Ultimo_Nome	UNCERT	m-prob 0.9 u-prob 0.01 Param1 900
Sufixo	CHAR	m-prob 0.9 u-prob 0.01 Param1 800
Nota Corte	Nenhuma	

Tabela 5. 9 - Exemplo de critérios utilizados na consolidação

Após a execução deste processo, todas as ocorrências de duplicidades são identificadas, conforme tabela 5.10, e continuam existindo no arquivo consolidado, pois ainda não foi realizado nenhum processo para retirá-las deste arquivo.

CPF	**Primeiro_Nome**	**Nome_Meio**	**Ultimo_Nome**	**Sufixo**	**Dup**
2345678901	ANA	MARIA	SANTOS		1
3345678901	ANA	LUCIA	SANTOS		2
2345678901	ANA	MARIA	SANTOS		1
3345678901	ANA	L	SANTOS		2
72343212	JARBAS	SANTOS	SILVA	JR	3

Tabela 5. 10 - Exemplo dos dados após processo de consolidação

Além disso, ao término do processo, todos os parâmetros utilizados e resultados obtidos se encontram devidamente documentados e disponíveis para consultas.

5.3.6 – Transformar e Enriquecer os dados (P4.7)

- **Enriquecer os dados (P4.7.1)**

O objetivo deste processo é implementar a solução padrão de enriquecimento, definida anteriormente no processo 4.3.2, utilizando a ferramenta de reengenharia.

Todos os critérios definidos na ferramenta são documentados, conforme exemplificado na tabela 5.11, para servirem de entrada e base para a próxima fonte a ser tratada e todos os problemas enfrentados são documentados para manter um histórico sobre como foram obtidos os parâmetros utilizados, como apresentado na tabela 5.12.

Fonte Fidedigna	Correios	
Critério	CRIT2	
Processo	MATCH2	
Bloco	UF + Cidade + Endereco5 (5 posições do endereço)	
Campo	**Tipo de Comparação**	**Parâmetros**
Endereço	UNCERT	m-prob 0.9 u-prob 0.01 Param1 800
Nota Corte	Nenhuma	
Enriquecimento	CEP	

Tabela 5. 11 - Exemplo do detalhamento do enriquecimento

Fonte Fidedigna	Receita Federal
Problemas	**Soluções**
Conteúdo do campo CPF idêntico, mas Nomes diferentes: ANA MARIA DOS SANTOS LUCIA MARIA DOS SANTOS	Reduzir 50 pontos quando a primeira letra do Primeiro Nome for diferente (Parâmetro DISAGREE –50)
…	

Tabela 5. 12 - Exemplo de soluções adotadas no enriquecimento

Ao término deste processo, todas as informações se encontram padronizadas e enriquecidas com a utilização de fontes fidedignas e estão prontas para serem corrigidas ou complementadas pelo próximo processo.

- **Transformar os dados (P4.7.2)**

O objetivo deste processo é adequar os dados de acordo com o formato de entrada esperado no banco de dados do DW (*Data Warehouse*). A primeira transformação, denominada de sobrevivência pela ferramenta de reengenharia, é responsável por eliminar os registros identificados como duplicados no processo anterior (P4.5). Este processo utiliza principalmente a lista de referências obtida no processo 4.1 para escolher a melhor informação dentro de cada grupo, além de outros critérios a serem utilizados no caso de empate, como por exemplo, selecionar a informação mais atual ou a informação que apresentar o maior conteúdo, etc. Ao término desta transformação, todas as duplicidades foram eliminadas, conforme apresenta a tabela 5.13.

CPF	Primeiro_Nome	Nome_Meio	Ultimo_Nome	Sufixo	Dup
2345678901	ANA	MARIA	SANTOS		1
3345678901	ANA	LUCIA	SANTOS		2
72343212	JARBAS	SANTOS	SILVA	JR	3

Tabela 5. 13 - Exemplo dos dados após processo de sobrevivência

A próxima fase é responsável por transformar os dados para adequá-los ao formato esperado na plataforma destino. Por exemplo, se o formato esperado para o campo CPF for alfanumérico com onze posições, então o campo CPF precisa ser formatado com zeros à esquerda, conforme apresentado na tabela 5.14, para ajustá-lo ao formato esperado no destino.

CPF	Primeiro_Nome	Nome_Meio	Ultimo_Nome	Sufixo	Dup
02345678901	ANA	MARIA	SANTOS		1
03345678901	ANA	LUCIA	SANTOS		2
00072343212	JARBAS	SANTOS	SILVA	JR	3

Tabela 5. 14 - Exemplo dos dados após transformação

Após os dados serem transformados, eles estão prontos para as próximas fases responsáveis por sumarizar ou derivar algum campo (P4.8), caso necessário, auditar e controlar a extração, transformação e carga dos dados (P4.9), analisar os

tipos de defeitos identificados nos processos anteriores (P4.6) e, finalmente, realizar a atualização dos dados no banco de dados do ambiente DW.

5.3. Conclusão

Neste capítulo foram apresentadas algumas sugestões de melhorias à metodologia TQdM visando sanar as dificuldades enfrentadas durante a implementação do estudo de caso. As sugestões envolvem a criação de novos processos e sub-processos visando suprir a falta de alguns processos importantes para tomada de decisão (P7-Estudo de Viabilidade Técnica) e agilizar a implementação das fontes (P4.3.2-Agrupar as fontes de informação com as mesmas características e planejar uma solução padrão para cada grupo); a divisão de processos já existentes com o objetivo de separar atividades distintas (P4.3-Padronizar os Dados e 4.7-Transformar os dados e Enriquecê-los) e a inversão de alguns processos visando agregar todas as informações relevantes antes de realizar o processo de consolidação (P4.7.1-Enriquecer os dados). E principalmente, apresentar algumas sugestões de formulários a serem preenchidos em cada um dos principais processos que requerem documentação especial.

Capítulo 6 - Conclusão

6.1 - Resumo

Para este trabalho inicialmente levantou-se as características relacionadas com a qualidade de dados, os principais processos envolvidos na limpeza dos dados e a evolução dos processos de melhoria da qualidade até atingir as metodologias existentes no momento, como por exemplo, a TQM (*Total Quality Management*), TDQM (*Total Data Quality Management*) e TQdM (*Total Quality data Management*). Porém dentre todas as metodologias, notou-se que, somente a TQdM, desenvolvida por Larry English, possuia detalhamento das fases de desenvolvimento para implementação de um projeto de qualidade de dados, por isso ela foi escolhida para ser aplicada no estudo de caso.

A metodologia TQdM é composta por 6 processos. O primeiro processo é responsável por avaliar o significado dos dados e a qualidade da estrutura dos dados (P1), o processo seguinte é responsável por avaliar a qualidade da informação (P2), sendo que, os resultados dessas análises são empregados durante o processo de medição dos custos resultantes das informações sem qualidade (P3). Ao término deste processo, duas medidas podem ser tomadas: realizar a reengenharia e a limpeza dos dados visando melhorar a qualidade dos dados (P4) ou melhorar a qualidade dos dados envolvidos na geração da informação (P5). Além desses 5 processos, existe um processo, denominado plano de ação, responsável por estabelecer um ambiente de qualidade da informação (P6). Dentre todos esses processos, o processo de reengenharia e limpeza dos dados é o processo mais importante, pois ele é responsável pelo tratamento da informação gerando uma informação de qualidade e, durante a sua implementação, foi possível detectar alguns problemas a serem ajustados na entrada dos dados.

O processo de reengenharia e limpeza dos dados (P4) tem como objetivo realizar o tratamento das informações desde o estabelecimento da prioridade das fontes até a carga dos dados e auditoria do processo. Após conhecer todos os

processos envolvidos na metodologia TQdM, eles foram utilizados para implementar um estudo de caso.

O estudo de caso ocorreu em uma grande empresa da área financeira que possuía diversos cadastros de clientes descentralizados e objetivava melhorar a qualidade dos dados desses cadastros isolados e gerar um cadastro consolidado. A principal premissa foi que nenhuma mudança deveria ser efetuada na entrada dos dados já existentes e que o novo cadastro gerado deveria possuir um relacionamento com as fontes de informações.

Iniciou-se o estudo de caso pelo processo de avaliação da qualidade da definição e do conteúdo dos dados de uma amostra (P1 e P2), onde os resultados obtidos foram registrados e enviados para a fase de avaliação dos custos relacionados com a baixa qualidade dos dados (P3). Nesta fase decidiu-se por prosseguir o processo de melhoria da qualidade dos dados implementando o processo de reengenharia e limpeza dos dados (P4). Este foi o principal processo do estudo de caso. Nele as fontes de informações foram novamente analisadas (P4.1 e P4.2), mas desta vez considerando o volume real de informações. Após a avaliação dos dados, o processo seguinte foi a definição dos padrões a serem adotados para equalizar todos os dados de um mesmo grupo, denominado de processo de padronização (P4.3), e a construção de regras para sustentar a implementação deste procedimento. Com as informações padronizadas foi possível corrigir e complementar os dados faltantes (P4.4) e a seguir realizar a consolidação dos dados (P4.5). Depois, o processo de transformação e enriquecimento dos dados (P4.7) utilizando as informações dos Correios e da operadora de telefonia foi aplicado. A fase final foi auditar e controlar a extração, transformação e carga dos dados (P4.9) no ambiente de DW (*DataWarehouse*). Durante a implementação dos processos foram inseridos alguns mecanismos para detectar possíveis problemas de qualidade nos dados visando auxiliar o processo de análise dos defeitos (P4.6) a ser realizado no futuro.

Durante a implementação dos processos da metodologia TQdM, alguns problemas foram identificados, como por exemplo, a ausência de fases responsáveis pelo estudo de viabilidade técnica da ferramenta a ser utilizada durante o processo de reengenharia e limpeza dos dados, a ausência de uma fase de prototipação para implementar e validar a solução desejada e a ausência de formas padrões para

realizar a documentação dos processos. Visando minimizar os problemas identificados, algumas sugestões de melhorias na metodologia foram apresentadas envolvendo a criação de novos processos, a divisão e a inversão de processos já existentes e a utilização de formulários padrões para documentar os resultados encontrados.

6.2 – Resultados Obtidos e Contribuições

Este estudo aplica-se à empresas que desejam implantar um projeto na área de QD. Implantações de projetos nesta área necessitam de conhecimento teórico e prático sobre o assunto porém, observa-se que, apesar de alguns trabalhos relatarem sobre conceitos, processos e metodologias aplicáveis a QD, não foi encontrado nenhum trabalho que apresentasse um caso prático de um projeto ou que explicasse como aplicar a teoria na prática.

Este trabalho apresenta os conceitos principais utilizados em um projeto de QD, aplicados em um estudo de caso prático, onde é possível observar a implementação da metodologia, passo-a-passo, em cada um dos seus processos. O estudo de caso mostra como diversos problemas que surgiram durante o andamento do projeto foram solucionados pela equipe de trabalho, como por exemplo, a dificuldade de aliar o conhecimento específico da metodologia e da ferramenta para solucionar um problema real. Apresenta sugestões de como a metodologia e a aplicação prática poderiam ser conduzidas para minimizar os problemas no andamento dos projetos QD. E apresenta também, um conjunto de formulários criados para implementação do estudo de caso, cuja utilização possibilita maior controle e documentação de todo o processo de QD.

6.3 – Futuros Trabalhos

Este trabalho abordou a implementação da metodologia TQdM (*Total Quality data Management*) visando a melhoria da qualidade dos dados (P4), ou seja, atuou na melhoria do produto final. Para futuros trabalhos sugere-se a implementação e a

avaliação de um estudo de caso envolvendo a melhoria dos processos que geram as informações de baixa qualidade (P5) fechando assim, o ciclo de melhoria da qualidade dos dados.

Os processos de melhoria da qualidade são cíclicos, sempre podem ser melhorados; o que é possível fazer é refinar o processo a cada iteração. Para auxiliar este processo seria interessante e importante criar mecanismos para avaliar se a qualidade dos dados do processo atual é satisfatória. Seria importante criar formas de extração de amostras inteligentes para avaliar constantemente a qualidade dos dados e manter sempre o nível de qualidade desejado.

Outro ramo que pode ser explorado é continuar as pesquisas de outras metodologias voltadas para projetos de qualidade dos dados, como por exemplo a TDQM (Total Data Quality Management), e avaliar a sua efetividade em um projeto real indicando os pontos fracos e fortes desta metodologia. Além disso, seria de grande valia a realização de trabalhos comparativos envolvendo as diversas ferramentas existentes e avaliar em que casos elas podem ser utilizadas, ou seja, de acordo com o objetivo e escopo do projeto apontar qual seria a categoria de ferramenta adequada e dentre todas qual a melhor ferramenta a ser utilizada.

Referências Bibliográficas

[1] SKYMARK. Management Resources. Thinkers. In: **Dr. W. Edwards Deming**. Disponível em: http://www.skymark.com/resources/leaders/deming.asp [01/10/2002 23:57]

[2] HO, S & FUNG, C. **Juran´s Message**. TQM Guru´s Idea. TQMEX Model, 1998. Disponível em: http://www.hkbu.edu.hk/~samho/tqm/tqmex/juran.htm [01/10/2002 00:05]

[3] SKYMARK. Management Resources. Thinkers. In: **Philip Crosby: The fun uncle of Qualiy Revolution**. Disponível em: http://www.skymark.com/resources/leaders/crosby.asp [01/10/2002 23:51]

[4] SKYMARK. Management Resources. Thinkers. In**: Kaoru Ishikawa: One Step Further**. Disponível em: http://www.skymark.com/resources/leaders/ishikawa.asp [01/10/2002 24:01]

[5] PADHI, N. **The Eight Elements of TQM.** In: SIX SIGMA. Disponível em: http://www.isixsigma.com/library/content/c021230a.asp [01/05/2003 20:08]

[6] WANG, R. **Total Data Quality Management Cycle.** In: Raising the Bar for Data Quality in the New Millennium. p2. 2000. Disponível em: http://www.niss.org/affiliates/dqworkshop/presentations/wang-presentation.pdf [08/04/2003 23:13]

[7] ENGLISH, L. **Improving Data Warehouse and Business Information Quality**. New York: Wiley, 1999.

[8] BRECKER ASSOCIATES. **Quality-Based Problem-Solving / Process Improvement**. Disponível em: http://www.brecker.com/quality.htm [10/05/2003 21:45]

[9] FIRSTLOGIC. **Customer Data Quality – Building the Foundation for a One-to-one Customer Relationship. A White Paper**. Disponível em: http://www.firstlogic.com/pdfs/db_oldWhitepaper.pdf [08/10/2002 21:12]

[10] CROSBY, P. B.: **Quality is free**, p. 149-222, New York: Peguin Group, 1979.

[11] FIRSTLOGIC. **Which data is most important to your company.** In: Customer Data Quality – Building the Foundation for a One-to-one Customer Relationship. A White Paper. p.2. Disponível em: http://www.firstlogic.com/pdfs/db_oldWhitepaper.pdf [08/10/2002 21:12]

[12] ENGLISH, L. **10 years of Information Quality Advances: What Next ?** In: DM Review, Fevereiro 2001. Disponível em http://www.dmreview.com/master.cfm?NavID=55&EdID=3009 [01/10/2002 23:35]

[13] LEADERSHIP INSTITUTE. **What are Dr. Demings 14 Points?** In: Who is Dr. W. Edwards Deming? Disponível em: http://www.lii.net/deming.html [01/10/2002 22:10]

[14] MASSACHUSETTTS INSTITUTE OF TECHNOLOGY. **The MIT Total Quality Management Program**. Disponível em: http://web.mit.edu/tdqm/www/about.shtml [01/10/2002 22:10]

[15] HCI. **PDCA CYCLE - From problem-faced to problem-solved**. In: Articles, papers and tools. Disponível em: http://www.hci.com.au/hcisite2/toolkit/pdcacycl.htm#From%20problem-faced%20to%20problem-solved [08/04/2003 23:13]

[16] HO, S & FUNG, C. **Deming´s Message**. TQM Guru´s Idea. TQMEX Model, 1998. Disponível em: http://www.hkbu.edu.hk/~samho/tqm/tqmex/deming.htm [20/05/2003 23:34]

[17] SCHLENKER, J. **Total Quality Management – Na Overview**. In: HRZone. Overview for TQM. Disponivel em: http://www.hrzone.com/topics/tqm.html#overview [07/04/2003 23:17]

[18] ST. NORBERT COLLEGE. **Introduction to TQM in POM**. In: Production / Operations Management. Disponível em: http://www.snc.edu/socsci/chair/333/333_tqm1.htm [23/05/2003 23:34]

[19] O´NEILL, P. **It´s a Dirty Job: Cleaning Data in the Warehouse. Database and Data Warehousing Software Worldwide. Market Analysis**. In: GartnerGroup, 12 Janeiro, 1998. Disponível em: http://www4.gartner.com/UnrecognizedUserHomePage.jsp Note Number: DBDW-WW–DP–9801 [08/10/2002 21:06]

[20] WAND,Y & WANG, R. Y.: **Anchoring Data Quality Dimensions in Ontological Foundations**. In: Communications of the ACM, 39 (1996), No. 11, p.86-95.

[21] ECKERSON, W. W.: **Data Quality and the Bottom Line – Achieving Business Success through a Commitment to High Quality Data**. In: TDWI Report Series-The Data Warehousing Institute. Disponível em: http://www.c4dq.com/files/TDWI-DQReport.pdf [13/11/2002 10:03]

[22] NELSON, D.S. & SINGHAL, R. & JANOWSKI, W. & FREY, N.: **Customer Data Quality and Integration: The Foundation of Successful CRM**. Disponível em: www.gartner.com [10/07/2002 11:36]

[23] GARTNER: Tutorial: **Data Warehouse: Selecting the BI and ETL Products That Are Right for You!** In: Gartner Symposium ITXPO 2002. Disponivel em: http://symposium.gartner.com/docs/symposium/itxpo_orlando_2002/documentation/sym12_05c.pdf [08/08/2002 23:13]