PROCESSO PARA A AVALIAÇÃO DE QUALIDADE DE DADOS EM SISTEMAS DE INFORMAÇÃO UTILIZANDO A METODOLOGIA TDQM

Trabalho de Conclusão de Curso

Engenharia da Computação

Gízia Dielle Leite Lama
Orientadora: Profª. Maria Lencastre Pinheiro de Menezes e Cruz

Universidade de Pernambuco
Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação

GÍZIA DIELLE LEITE LAMA

PROCESSO PARA A AVALIAÇÃO DE QUALIDADE DE DADOS EM SISTEMAS DE INFORMAÇÃO UTILIZANDO A METODOLOGIA TDQM

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, outubro de 2014.

Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação
Projeto de Final de Curso

ESCOLA POLITÉCNICA
DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 12 de 12 de 2014, às 8:00 horas, reuniu-se para deliberar a defesa da monografia de conclusão de curso do discente GIZIA DIELLE LEITE LAMA, orientado pelo professor Maria Lencastre Pinheiro de Menezes Cruz, sob título Processo para a Avaliação de Qualidade de Dados em Sistemas de Informação Utilizando a Metodologia TDQM, a banca composta pelos professores:

Maria Lencastre Pinheiro de Menezes Cruz

João Murilo Azevedo

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

☒ Aprovada ☐ Aprovada com Restrições* ☐ Reprovada

e foi-lhe atribuída nota: 9,5 (nove e meio)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O discente terá 7 dias para entrega da versão final da monografia a contar da data deste documento.

Maria Lencastre

MARIA LENCASTRE PINHEIRO DE MENEZES CRUZ

João Murilo D'Azevedo

JOÃO MURILO AZEVEDO

* Este documento deverá ser encadernado juntamente com a monografia em versão final.

Tempora mutantur et nos mutamur in illis.

- Provérbio latino

Agradecimentos

Àqueles que dedicaram seu tempo de forma a ajudar, seja assumindo o papel de orientadores ou apenas de ouvintes, meus sinceros agradecimentos.

Resumo

A informação é o recurso mais valioso de uma organização; por isso, ela deve ser relevante, confiável, e estar disponível a tempo de proporcionar decisões corretas. Os dados, como matéria-prima da informação, precisam de uma adaptação a um conjunto de regras e técnicas que ajudem a manter a sua corretude e viabilizem sua interpretação da maneira eficaz; esse esforço propociona uma economia considerável de recursos monetários e de tempo. Para realizar a implementação das técnicas e o reconhecimento da má administração dos dados, é necessário um procedimento conciso, adaptável às diferentes situações e complexidade de um determinado problema. Muitos dos métodos e ferramentas existentes no mercado não são usados por necessitarem de um alto investimento de recursos, além disso, a implementação da qualidade geralmente apenas abrange uma parte do problema. Este trabalho busca orientar os esforços envolvidos para a obtenção de um dado de qualidade, na medida que apresenta etapas para que o objetivo seja alcançado. Para isto foi desenvolvido um processo que utiliza o conjunto de regras e características da metodologia *Total Data Quality Management* – combinado com recursos de administração de dados já consolidados. O processo foi aplicado de forma simplificada a um estudo de caso real, de grande porte, que proporcionou uma avaliação inicial da proposta deste trabalho. Como consequência da realização do trabalho, ficou clara a necessidade de integração entre os esforços realizados na área de qualidade de dados precisa ser reconhecida e trabalhada para que esta prática seja acessível aos sistemas e usuários finais.

Abstract

Information is the most valuable resource of an organization and to serve this purpose must be provided reliably, relevant and available in time to provide correct decisions. The data, as raw material of information, need to be adapted to a set of rules and techniques that enable its interpretation in the most effective way possible. This effort provides quick decisions that in the context of today's world, results in considerable monetary resources and time savings. To carry out the implementation of the techniques and the recognition of data mismanagement, a concise procedure is necessary, adaptable to different situations and complexity of the given problem. Many of the methods and tools available on the market suffer rejection because they require a large investment of resources and the quality is generally implemented covering only part of the problem. This work seeks to guide the effort involved reducing the resources expended in obtaining a sufficient data quality, as presents steps to the goal intended to be reached. For this goal a process is developed using the set of rules and features of Total Data Quality Management methodology - combined with data management features already consolidated. The need to integrate the efforts on data quality field needs to be recognized and worked for this practice to be accessible to systems and end users.

Sumário

Índice de Figuras

Índice de Tabelas

Tabela de Símbolos e Siglas

CNAE – Classificação Nacional de Atividades Econômicas

CPF – Cadastro de Pessoas Físicas

ECA – Evento, Condição e Ação

ETL – Do inglês, *Extract, Transform and Load*

IAIDQ – Do inglês, *International Association for Information and Data Quality*

IBM – Do inglês*, International Business Machines*

KDD – Do inglês, *Knowledge Discovery in Databases*

MIT – Do inglês, *Massachusetts Institute of Technology*

NAICS – Do inglês, *North American Industry Classification System*

QD – Qualidade de dados

RN – Regras de Negócio

SGBDs – Sistemas de Gerenciamento de Banco de Dados

SIC – Do inglês, *Standard Internacional Classification*

SISOBI – Sistema de Controle de Óbitos

SRC – Sistema de Registro Civil

SUB – Sistema Único de Benefícios

TDQM – Do inglês, *Total Data Quality Management*

TI – Tecnologia da informação

TQM – Do inglês, *Total Quality Management*

XML – Do inglês, *EXtensible Markup Language*

Capítulo 1

Introdução

Neste capítulo estão descritos a motivação para a pesquisa, os objetivos deste trabalho, assim como a descrição da estrutura do documento.

1.1 Motivação e caracterização do problema

A cada ano, dezenas de bilhões de dólares são perdidos mundialmente devido à falta de qualidade de dados nos sistemas de informação [2,33,41,44]. Tal prejuízo está relacionado à duplicação de informações, erros ortográficos, dados errados, incompletos e inacessíveis, além do conjunto de eventos consequentes – desaparecimento de clientes, de credibilidade – que geram uma cadeia contínua e imensurável de novas perdas financeiras, de produtividade e de tempo.

Segundo Eckerson [13], a informação é considerada a unidade monetária de maior valor na nova economia. Portanto, os dados são essenciais ao capital intelectual e a qualidade com que são apresentados influencia diretamente na capacidade de absorção e resposta das organizações. Como um tema inicialmente proposto em 1972 por Ivanov [18], qualidade de dados é definido por McGilvray [26] como a adaptabilidade da informação em relação ao seu uso, de maneira que este uso seja o mais eficiente e preciso possível para o usuário.

De acordo com a metodologia *Total Quality Management Data* (TDQM) [27] desenvolvida pelo *Massachussets Institute of Technology* (MIT), para que um dado seja considerado de qualidade é necessária a garantia de características divididas pelas seguintes dimensões: precisão, veracidade, relevância e pontualidade. A aplicação das dimensões deve ser adaptada considerando as singularidades do sistema alvo. Portanto, é essencial a criação de diretrizes que ao investigar o sistema, proporcionem a correta avaliação dos dados durante o processo de implementação de qualidade. Desta forma, ao adaptar a conceptualização de um modelo geral e amplamente utilizado como o TDQM, deve-se considerar os diferentes tipos de

complexidade dos sistemas, a quantidade de informação utilizada, a maneira como o dado é apresentado e principalmente atender às necessidades do usuário com êxito [5].

Muitas organizações optam por não aplicar políticas que garantam qualidade de dados devido ao alto custo de recursos envolvidos. De fato, para alguns sistemas, atingir uma qualidade de dados ótima pode não compensar os esforços aplicados. No entanto, este trabalho busca orientar os esforços envolvidos facilitando a aplicação dos recursos despendidos na obtenção de um dado de qualidade. Outro fator comum é que quando existe a má qualidade dos dados, algumas medidas são realizadas apenas em favor dos dados já existentes. Ou seja, ao utilizar apenas ferramentas de correção dos dados, a causa do problema – inexistência de padrões e regras por exemplo – pode não ser abrangida. Logo, será necessária novamente a correção assim constituindo um ciclo tedioso, caro e evitável.

O objetivo deste trabalho é propor um processo que auxilie na análise da qualidade de dados e sugerir correções necessárias através de um processo com quatro etapas utilizando a metodologia TDQM. O *Total Data Quality Management* fornece um escopo para o estudo dos dados a partir da avaliação da existência de características – como precisão, segurança, completude – requisitos necessários a um dado de qualidade. Esta metodologia será utilizada como parâmetro ao desenvolvimento de ações à medida que precisam ser garantidas e implementadas ao tratamento do dado. As etapas do processo serão desenvolvidas a fim de contemplar desde a constatação da necessidade de implementação de qualidade de dados até o ciclo pós implementação, onde é necessária a continuidade da qualidade gerada.

1.2 Objetivos

1.2.1 Objetivo Principal

Este trabalho tem como objetivo principal oferecer um processo que implemente técnicas e estratégias de qualidade de dados, de acordo com a metodologia TDQM, e que oriente formas para que a qualidade seja preservada.

1.2.2 Objetivos secundários

Para alcançar tal finalidade e resultados significativos neste trabalho, as seguintes metas foram estabelecidas:

- investigar modelos padrões de qualidade de dados;
- investigar tecnologias consolidadas que realizem a manipulação de dados;
- analisar e selecionar técnicas de tratamento de qualidade preventiva e reativa;
- definição das etapas na criação de um processo de qualidade de dados utilizando TDQM.
- desenvolver o processo; e,
- aplicar o processo a um exemplo real e realização de uma análise dos resultados alcançados.

1.3 Estrutura do documento

Este documento está dividido em 5 capítulos, cujos respectivos resumos estão apresentados abaixo:

- **Capítulo 1: introdução**

  Neste capítulo estão inseridos os textos introdutórios deste trabalho: a motivação e a caracterização do problema, os objetivos e a descrição da organização do documento.

- **Capítulo 2: formulação teórica**

  Neste capítulo é fornecido o embasamento teórico do trabalho. São descritos os conceitos de informação, dados, conhecimento, de qualidade dados, metodologia *Total Quality Management*, engenharia de requisitos e suporte à administração de dados.

- **Capítulo 3: processo proposto**

Neste capítulo apresenta a principal contribuição do trabalho: a proposta de um processo para esquematizar a implantação de qualidade de dados e suas etapas.

- **Capítulo 4: aplicação prática e resultados**

  Neste capítulo é detalhada a aplicação do processo proposto em um sistema real de grande porte; Alguns conceitos são considerados, incluindo: técnicas de qualidade de entrada, regras de padrões de dados, escalabilidade do sistema e a contínua garantia da qualidade, confiabilidade das informações e segurança de dados.

- **Capítulo 5: conclusões e trabalhos futuros**

  Neste capítulo são apresentadas as conclusões do trabalho, suas contribuições e possíveis trabalhos futuros.

Capítulo 2

Formulação teórica

A finalidade deste capítulo é fornecer as informações necessárias para o entendimento da proposta desta monografia. A primeira seção define conceitos sobre dados, informação e conhecimento, a segunda seção é composta por conceitos e definições sobre qualidade de dados, a terceira seção descreve a metodologia *Total Quality Management* com ênfase no projeto TDQM, por fim, a última seção trata da abrangência interdisciplinar do suporte à qualidade de dados: armazenamento e integração de dados, mineração de dados, engenharia de requisitos e governança de dados.

2.1 O dado, a informação e o conhecimento

O dado é definido como uma sequência de símbolos quantificados ou quantificáveis, podendo ser tratado como uma entidade matemática e descrito através de representações formais [33]. Os dados podem ser considerados ainda, como uma lista ordenada de fatos brutos que representam eventos ocorrentes num ambiente organizacional ou físico, antes de serem organizados de forma que possam ser compreendidos [20].

A informação, segundo McGee e Prusak [25], são dados coletados, organizados e ordenados de forma a ter um significado e contexto específico, cuja missão é informar. A informação é fundamental no apoio às estratégias e processos na tomada de decisão, bem como nas operações organizacionais [5, 23].

O conhecimento é definido como uma mistura de experiências adquiridas ao longo da vida: valores, agregação de novas experiências e informação contextual de forma a permitir uma avaliação [11]. Tanto as origens como a aplicação do conhecimento estão na mente dos conhecedores. Ou seja, não pode ser descrito: o que se descreve é a informação.

Os conceitos de dados, informação e conhecimento são citados conforme a Tabela 1.

Tabela 1. Dados, informação e conhecimento

Dado	Informação	Conhecimento
Simples observações sobre o estado do mundo.	Dados dotados de relevância e propósito.	Informação valiosa da mente humana. Inclui reflexão, síntese e contexto.
➢ Facilmente estruturado ➢ Facilmente obtido através de máquinas ➢ Frequentemente quantificado ➢ Facilmente previsível	➢ Requer unidade de análise ➢ Exige consenso em relação ao significado ➢ Exige necessariamente a mediação humana	➢ Difícil estruturação ➢ Difícil captura por máquinas ➢ Frequentemente implícito

Fonte: extraído de [9]

Para este trabalho será considerada a ênfase no estudo sobre dados como matéria-prima da informação e, consequentemente, do conhecimento. Para facilitar o entendimento da representação dos dados é necessário categorizá-los.

2.1.1 Categoria de dados

Categorias de dados são agrupamentos de dados que possuem atributos em comum [26]. Ainda, segundo McGilvray, categorias são úteis no gerenciamento, pois diferentes tipos de dados devem ser manipulados de acordo com a sua classificação. Portanto, ao entender a relação e dependência entre as diferentes categorias é possível otimizar os esforços no alcance da qualidade de dados. As 6 categorias descritas em [26] são: *master,* transacional, referência, metadados, históricos e temporários – apresentadas na Tabela 2.

Tabela 2. Categoria de dados x Definição

Categoria de dados	Definição
Master	Descreve pessoas, lugares e objetos envolvidos em um sistema. Esses tipos de dados são usados, geralmente, em múltiplos processos e sistemas, por isso, sua padronização e sincronização é crítica no sucesso da integralização de um sistema.
Transacional	Descreve uma transação ou um evento interno ou externo. Esses tipos de dados são agrupados, tipicamente, em registros transacionais que incluem associação com dados *master* e de referência.
Referência	Descreve um conjunto de valores ou esquemas de classificação que são referenciados por sistemas, banco de dados, processos e relatórios do sistemas. Dados de referência são padronizados facilitando o compartilhamento e *feedback* da informação. São relacionados com dados transacionais e *master.*
Metadado	Representa o dado sobre dados. Busca caracterizar e descrever outros dados tornando sua manipulação mais fácil. Metadados podem ser associados a qualquer categoria de dados.
Histórico	Descreve um fato em determinado tempo que não deve ser alterado exceto pra corrigi-lo.
Temporário	Descreve um dado que é mantido na memória do sistema para acelerar seu processamento.

Fonte: traduzido e parcialmente adaptado de [26]

A importância na categorização é dada pela clareza proporcionada com relação à especificação do dado. Esse aspecto auxilia na compreensão do contexto e na

definição de estratégia de qualidade de dados a ser aplicada. Por exemplo, num projeto, focado em melhoria de qualidade em dados *master* pode-se ter como causa dos problemas de qualidade uma falha nos dados de referência, já que ambos podem assumir relações entre si. Ter o conhecimento das categorias de dados permite reduzir o tempo na descoberta de falhas e na escolhas de estratégias [21].

2.1.2 Especificação de dados

Na especificação de um dado, o processo de padronização é muito importante para a garantia de sua qualidade. Desta forma, esta característica deve existir nas etapas iniciais da implementação de um sistema e devem persistir, através de documentação, para possibilitar a sua continuidade. A coleção de padrões, regras e guias que definem como nomear e definir um dado e estabelecer valores válidos, além de especificar regras de negócio, é chamada de padronização de dados. O *International Association for Information and Data quality* [17], ou (IAIDQ), provê um glossário e referências dos termos que remetem à qualidade de dados. A seguir, estão destacados os conceitos de padronização mais relevantes ao projeto desta monografia:

- Convenção de nomes para tabelas e campos – um determinado dado deverá seguir padrões de acordo com o seu recipiente. Por exemplo, se um dado é relacionado a um campo nome, a este e seus semelhantes deverão ser atribuídos prefixos do tipo NM: NM_Primeiro e NM_Sobrenome.
- Definições e convenções de dados na formulação de regras de negócio – cada campo deverá ter um documento padrão que descreva, ao menos, o mínimo de informações relativas a ele como: nome do campo, descrição, exemplo de conteúdo, obrigatoriedade e o valor *default.*
- Estabelecimento, documentação e atualização de listas de valores válidos – deve haver um consenso sobre os valores que podem ser aceitos por um determinado campo. Em caso de mudanças, a documentação deverá conter o processo que identifique como estas podem ser executadas e quais outros processos podem estar envolvidos.
- Valores de referência, classificação e categorização mais aceitos para determinada atividade desenvolvida na organização. Por exemplo, como

referência para a produção de estatísticas sobre a atividade econômica e identificação desta nos cadastros e registros, no Brasil é utilizado o CNAE – Classificação Nacional de Atividades Econômicas, derivado do SIC – *Standard Industrial Classification System.* Já nos Estados Unidos, México e Canadá é utilizado o NAICS – *North American Industry Classification System*, também derivado do SIC. O uso de um padrão que define o segmento de atuação de uma organização permite um alto nível de comparação entre estatísticas de negócio, mas para elevar a comparação a nível mundial deve-se adotar o padrão mais aceito, ou o vigente com o sistema alvo.

Em adição às especificações de dados ainda temos o conceito de modelos de dados – representação visual da estrutura de um dado, ou ainda, a especificação de como o dado deve ser representado num banco de dados. Os termos entidade e atributo, conceitos centrais da modelagem, definem um conjunto de coisas cujas instâncias são unicamente identificáveis e a definição de uma característica, qualidade ou propriedade de uma entidade respectivamente. A forma com que o dado se apresenta ainda pode ser classificada como [26]:

- Conceitual: o modelo se refere a uma visão compreensiva da organização.
- Lógico: específico a uma tecnologia de gerência de dados – XML (*eXtensible Markup Language*), por exemplo.
- Físico: descreve a tecnologia em detalhes – partições utilizadas, tamanho de tabelas, por exemplo.

Por fim, em termos de especificação, têm-se as regras de negócio, ou o conjunto de princípios que descreve interações de negócio e estabelece regras para as ações resultando na integridade do dado. Para Ross [31], regras servem como guias para a condução de uma ação e provêm critérios na tomada de decisões. A seguir, na Tabela 3, são exemplificadas duas regras de negócio e como elas influenciam a qualidade de dados.

Tabela 3. Regras de negócio e sua influência em qualidade de dados

Tipo de regra de negócio	Exemplo de regra de negócio	Ação	Checagem de qualidade de dados
Restrição	Um cliente não pode ultrapassar R$ 2.000,00 em compras na fatura do cartão de crédito.	Um serviço será executado para determinar se o total do valor gasto pelo cliente excedeu R$ 2.000,00, se sim, o cartão será temporariamente desativado.	A regra é violada se: Total_Fatura >= "2.000" e Cartao_Status = "Ativo" .
Inferência	Um cliente deve ser considerado especial se já está com a mesma empresa há 10 anos.	Um serviço será designado para checar a data inicial do cadastro e a data atual.	O tipo do cliente, no caso especial, poderá ser inferido se: o número total de anos for > = 10.

Fonte: extraído de [31]

2.2 Qualidade de dados

A origem de pesquisa sobre qualidade de dados é proveniente da estatística. Segundo Batini e Scannapieca [3], os estatísticos foram os pioneiros na descoberta de problemas com a qualidade de dados. Estudos relacionados a dados estatísticos duplicados foram realizados em 1960 e somente na década de 1990, estudiosos de computação consideraram o problema. Nessa época surgiram os esforços em definir, medir e melhorar a qualidade em processos organizacionais que envolvam dados.

O conceito *Fitness for use*, amplamente aceito como definição de qualidade, refere-se à adaptabilidade da informação, como produto do dado, ao usuário final. Essa definição enfatiza a importância de se considerar a visão do consumidor de qualidade, pois, ele que irá julgar se o produto é ou não apto para uso [10,11,37,41]. Estendendo esse conceito, McGilvray [26], afirma que a qualidade não envolve

somente a confiabilidade do dado, além disso, para essa informação ou dado ajudar a administrar e gerenciar negócios, tomar decisões eficazes, dar suporte ao cliente é necessário que as informações estejam disponíveis na hora, lugar certo e para as pessoas certas.

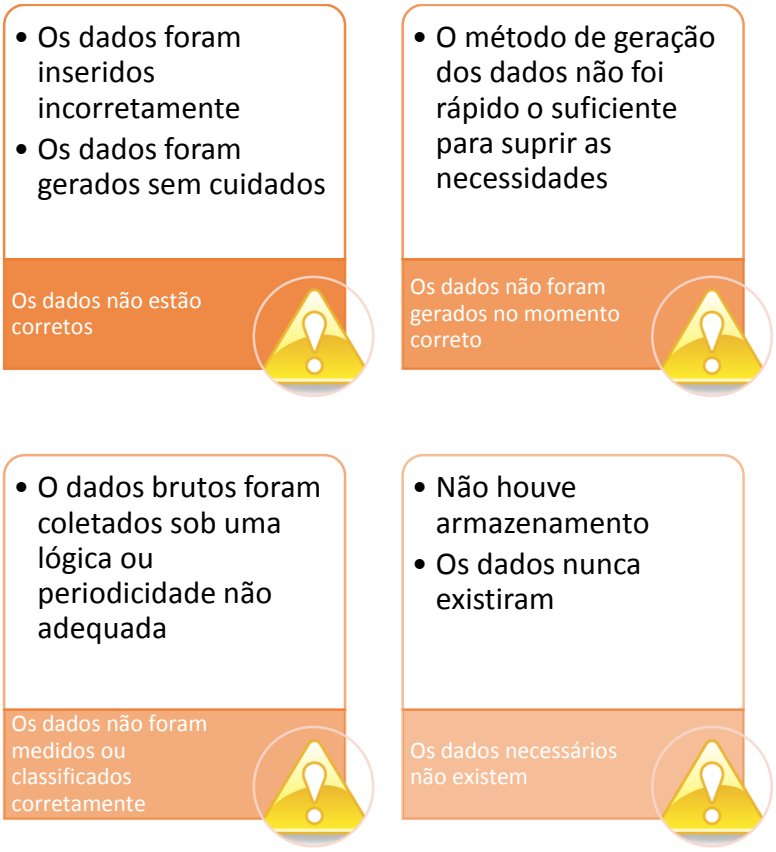
A falta de qualidade de dados custa bilhões de dólares para as diversas organizações mundiais [16, 36]. O estudo mais recente, realizado em 2012 pela *InsightSquared* [42], realiza esforços na medição dos custos e estastísticas sobre a má gestão dos dados através de pesquisa, conforme descrito:

- O custo do dado sem qualidade nos Estados Unidos excede 600 bilhões de dólares anualmente.
- O prejuízo causado pela falta de qualidade de dados nos maiores supermercados do Reino Unido custará 1 bilhão de dólares nos próximos cinco anos.
- Dados fracos, sem qualidade, é citado pela maioria das empresas como a razão principal de aumentar o custos de projeto.
- Boas práticas de qualidade de dados aumentam os lucros em 66%.

A dificuldade em definir estatísticas sobre a falta de qualidade de dados também está inserida num contexto subjetivo – em como estimar as oportunidades perdidas provenientes do mau gerenciamento de dados. Como afirma Drescher [12], a má qualidade de dados significa a possibilidade da existência de informação imprecisa, incompleta, redundante e até mesmo fictícia. Os problemas acarretados incluem a diminuição da confiança do cliente, perda de oportunidade de negócio e tomadas de decisão equivocadas ocasionadas pela imprecisão e a falta de completude dos dados.

Ainda em relação às consequências da falta de qualidade de dados, Turban, Wetherbe e Mclean [37], realizaram uma pesquisa descobrindo a origem e a forma dos problemas. O resultado da pesquisa é encontrado na Figura 1.

Figura 1. Pesquisa sobre a origem dos problemas de qualidade de dados



Fonte: extraído e adaptado de [37]

A partir desse estudo foi possível sintetizar as causas e propor quatro características essenciais a qualidade de dados conforme [26]:

- Qualidade intrínseca: relativa aos dados precisos, objetivos e confiáveis.
- Qualidade de acesso: referindo-se aos dados acessíveis e com segurança em seu acesso.
- Qualidade de contexto: relativo aos dados importantes, com valor agregado, convenientes, ou, na medida certa.
- Qualidade de representação: onde os são dados interpretados de maneira concisa e consistente.

Assumindo as características propostas e o surgimento e evolução de modelos e ferramentas de referência em qualidade, é possível tomar boas decisões, a partir da precisão da informação. Além disso, os benefícios são também monetários,

quantificáveis – relativos a economia de recursos como o tempo –, e intangíveis, como, por exemplo, a satisfação de um cliente.

2.3 Metodologia *Total Quality Management*

A produção de qualidade consiste numa estratégia de administração orientada a criar consciência da qualidade em todos os processos organizacionais. Com esta premissa, o *Total Quality Management* (TQM), desenvolvido para todas as escalas de uma organização [10], dá origem a algumas metodologias que seguem a idéia base de planejamento e continuidade do projeto com foco na qualidade em processos em geral. No entanto, direcionando o foco para qualidade de dados, o *Massachussets Institute of Technology* criou o *Total Data Quality Management* [27] em 1995. Os esforços contínuos do grupo de pesquisa do TDQM e MITIQ têm buscado refinar a metodologia e/ou especializá-la, isto, para solidificar a disciplina de qualidade de dados, divulgar o impacto de pesquisas e promover colaborações entre universidades, indústria, governo e organizações. Por ser a base formal para o gerenciamento de qualidade de dados mais utilizada e de contínuo desenvolvimento, esta será usada no desenvolvimento do processo proposto neste projeto de monografia.

O TDQM pode ser descrito como um conjunto de teorias sólidas utilizadas para prover métodos práticos que melhorem a qualidade de dados [22, 34]. O conjunto da metodologia abrange as quatro categorias de qualidade: intrínseca, contextual, representacional e de acessibilidade, originando 16 características. Ao longo do processo de pesquisa na consolidação desta metodologia, mais de 100 atributos de qualidade de dados foram identificados, e servem hoje de referência para a implementação de qualidade em sistemas computacionais. A seguir, a Figura 2, detalha o esquema desta metodologia.

Figura 2. Categorias e atributos do TDQM

Qualidade de dados			
Categoria Intrínseca	Categoria Contextual	Categoria Representacional	Categoria de Acessibilidade
• Precisão • Credibilidade • Imparcialidade • Reputação	• Valor agregado • Relevância • Atualização • Completude • Quantidade	• Interpretação • Entendimento • Consistência • Representação concisa	• Acesso • Segurança • Manipulação

Fonte: extraído de [6]

Já no ciclo de administração da qualidade de dados do TDQM, o foco é a produção de dados com qualidade adaptados às necessidades do sistema. O processo engloba [6, 24] :

- Definição – a definição abrange a indentificação das características dos dados em dois níveis: agente decisor e analista de TI. Utiliza as quatro categorias na avaliação – intrínseca, contextual, representacional e de acessibilidade (Figura 2).
- Medição – a medição busca especificar cada categoria em divisões. É possível utilizar estudos de pesquisa para registrar dados estatísticos como: taxa de erro e dispersão.
- Análise – interpreta os resultados da medição. É necessário considerar os diferentes usuários e seu envolvimento com os dados; o conhecimento possuído por cada um revela diferentes aspectos e níveis de conhecimento; mede a qualidade dos dados considerando os atributos de qualidade propondo ações corretivas.
- Ação – desenvolve ações de melhoria da qualidade de dados. Visando mudar os dados diretamente ou o processo que produz o dado.

O *framework* do TDQM é extensível e adaptável de acordo com o processo descrito. O ponto essencial é considerar o dado como um produto manufaturado pela maioria das organizações [40].

2.4 Suporte à qualidade de dados

A qualidade de dados é obtida através da união de esforços de diferentes disciplinas que contribuem com tradições teóricas e metodológicas. Portanto, é importante relacionar as linhas de pesquisas que participam neste objetivo em comum; entre elas tem-se técnicas de armazenamento e integração de dados, engenharia de requisitos, mineração de dados e governança de dados. A contribuição de cada uma dessas linhas de pesquisa para a qualidade de dados é detalhada nas subseções a seguir.

2.4.1 Armazenamento e integração de dados

A qualidade dos dados presentes em uma base de armazenamento de dados é fundamental para o sucesso das atividades de uma organização. Um banco de dados pode ser conceitualizado como uma coleção de dados inter-relacionados, representando informações sobre um domínio específico – precisa fornecer informações relevantes e de confiança para a organização [26,45]. Bancos de dados são operados através de sistemas de gerenciamento de banco de dados (SGBDs) – representados por *softwares* que possuem recursos capazes de manipular as informações do banco de dados e interagir com os usuários.Os SGBDs provêem formas de ajuda a garantir a qualidade dos dados, atuando na consistência e integridade destes, à medida que, por exemplo, operam de acordo com especificações mesmo diante de tentativas deliberadas de agir de maneira diferente. É essencial uma política de provimento de qualidade na integração de dados entre diferentes bancos, devido à necessidade de compartilhamento e manipulação dos dados entre várias aplicações. Esta relação, muitas vezes ameaça a integridade do banco devido à ausência de padrões e regras para a manipulação dos dados à medida que estão expostos a vários usuários.

Mais específico, o *data warehouse* (armazém de dados) foca na análise de grandes volumes de dados históricos, possibilitando a análise de eventos passados, oferecendo suporte à tomada de decisão e à previsão de eventos futuros [24]. Utilizado para armazenar informações relativas às atividades de uma organização em bancos de dados, um *data warehouse* é construído, geralmente, a partir de processos ETL [28]: *extract, transform* e *load.* Segundo Eckerson e White, o ETL é uma

ferramenta que desempenha um papel crítico na criação de um *data warehouse*, a base do *Business Intelligence*. Ferramentas ETL agem como um funil, integrando dados heterogêneos de diferentes bases, transformando-os em um formato consistente e significativo, armazenado-os no *data warehouse*. Seu funcionamento visa simplificar e garantir ao processo a precisão e integridade dos dados através de benefícios como a manutenção de códigos mais rápida e fácil [1, 2], a escalabilidade entre servidores e a segurança – obtida através da modularização de processos.

2.4.2 Mineração de dados

A mineração de dados é uma disciplina relevante como uma alternativa ao suporte à avaliação de qualidade de dados. Mineração de dados, ou *data mining*, é o processo de análise de conjuntos de dados, que tem por objetivo a descoberta de padrões interessantes e que possam representar informações úteis. O valor agregado aos dados está na capacidade de extração da informação de mais alto nível, antigindo um maior grau de confiança, ou seja, útil ao suporte de decisões.

Na mineração de dados são definidas as tarefas e os algorítimos que serão utilizados de acordo com os objetivos do estudo, a fim de se obter uma resposta para o problema [40,41]. As tarefas possíveis de um algorítmo de extração de padrões podem ser agrupadas em atividades preditivas e descritivas.

Em geral, um processo de descoberta de conhecimento consiste em uma iteração das seguintes etapas [43]:

- Preparação: é o passo onde os dados são preparados para serem apresentados às técnicas de *data mining*. Os dados são selecionados (quais os dados que são importantes), purificados (retirar inconsistências e incompletude dos dados) e pré-processados (reapresentá-los de uma maneira adequada para o *data mining*). Este passo é realizado sob a supervisão e conhecimento de um especialista, pois o mesmo é capaz de definir quais dados são importantes, assim como o que fazer com os dados antes de utilizá-los no *data mining*.

- Mineração de dados: onde os dados preparados são processados. O principal objetivo desse passo é transformar os dados de uma maneira que permita a identificação mais fácil de informações importantes.
- Análise de Dados: o resultado da mineração é avaliado, visando determinar se algum conhecimento adicional foi descoberto, assim como definir a importância dos fatos gerados.

Data mining [14] é uma das ferramentas mais utilizadas para extração de conhecimento através de bancos de dados (*Knowledge Discovery in Databases* - KDD), tanto no meio comercial quanto no meio científico.

2.4.3 Engenharia de requisitos

A Engenharia de Requisitos baseia-se no estudo e aplicação de uma abordagem sistemática, disciplinada e quantificável para o desenvolvimento, operação e manutenção de *software*. Isto, a partir da criação de princípios sólidos de engenharia que tornem o *software* mais confiável, econômico e eficiente. Segundo Kotonya e Sommerville [19], um processo de engenharia de requisitos deve contemplar, tipicamente, as atividades a seguir:

- Elicitação – descreve a atividade da descoberta dos requisitos do sistema. Para definir o problema a ser resolvido, características, desempenho do sistema e outros recursos relacionados, os desenvolvedores trabalham junto com os clientes e os usuários nesta etapa.
- Análise – esta atividade descreve as necessidades de usuários e clientes de acordo com as definições dos requisitos de software. Serve para para analisar de maneira detalhada e solucionar possíveis conflitos de requisitos entre os envolvidos com o sistema.
- Documentação – a documentação deve abranger os requisitos com a finalidade de servir de base para o restante do processo de desenvolvimento.
- Validação – para esta atividade é necessário que a especificação dos requisitos do software seja validada com os requisitos do sistema.

- Gerenciamento – esta atividade deve atender, ao longo do processo de desenvolvimento, a manutenção da evolução dos requisitos.

É através da engenharia de requisitos que se tem acesso a dados essenciais que vão garantir a corretude dos dados e a sua evolução, uma vez que a engenharia de requisitos faz o levantamento de regras de negócio.

Análoga ao objetivo desta monografia, a disposição de metas da análise de requisitos contempla etapas que serão utilizadas na construção do processo de implementação de qualidade de dados – mais precisamente a etapa de diagnóstico dos problemas do dados. As metas são demonstradas a seguir [19]:

- Entendimento do domínio da aplicação: entendimento geral da área na qual o sistema será aplicado.
- Entendimento do problema: compreensão dos detalhes do problema específico a ser resolvido com o auxílio do contexto do sistema a ser desenvolvido.
- Entendimento do negócio: entender como o sistema irá afetar a organização e como contribuirá para que os objetivos do negócio e os objetivos gerais da organização sejam atingidos.
- Entendimento das necessidades e das restrições dos interessados: entender as demandas de apoio para a realização do trabalho de cada um dos interessados no sistema, entender os processos de trabalho a serem apoiados pelo sistema e o papel de eventuais sistemas existentes na execução e condução dos processos de trabalho. Consideram-se interessados no sistema, todas as pessoas que são afetadas pelo sistema de alguma maneira, dentre elas clientes, usuários finais e gerentes de departamentos onde o sistema será instalado.

A rastreabilidade, alocada na etapa de gerenciamento da engenharia de requisitos, é a atividade que permite a compreensão do relacionamento que existe com e entre requisitos de software, projeto e implementação [51]. Esses relacionamentos auxiliam o projetista a mostrar quais elementos do projeto satisfazem os requisitos à medida que podem fornecer uma base mais eficaz para a garantia da qualidade do sistema, a gerência das mudanças e a manutenção do software.

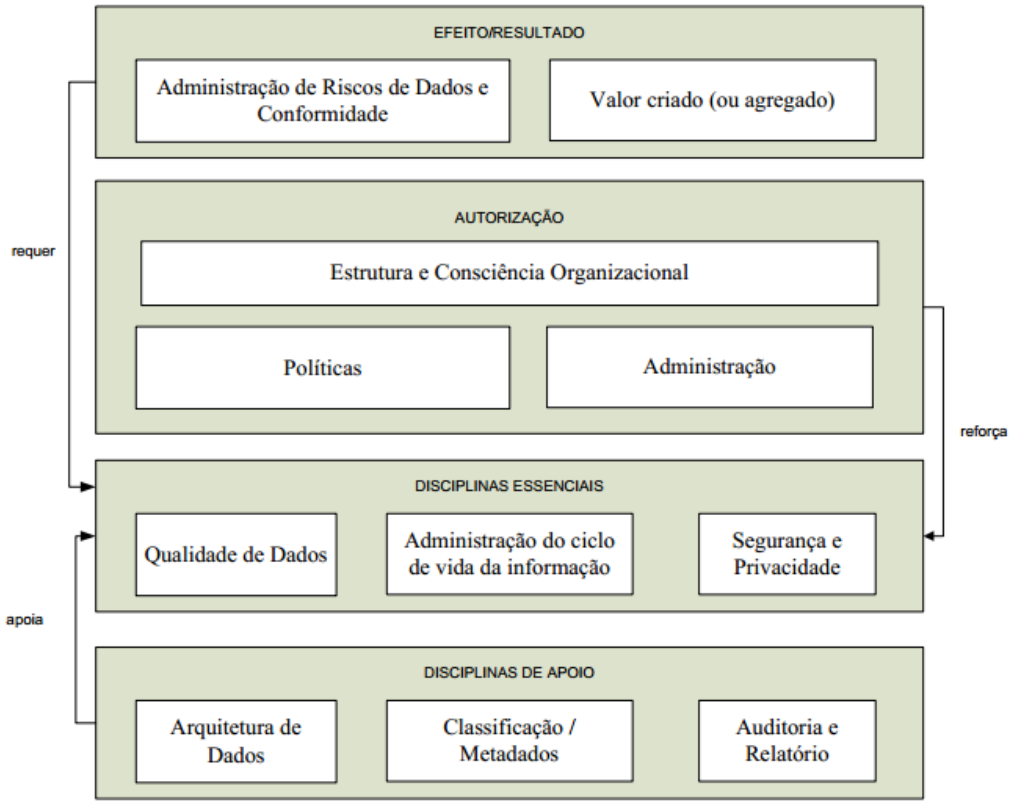
2.4.4 Governança de dados

Governança de Dados pode ser definida como a gestão conjunta de políticas, processos, pessoas e tecnologias, que visa estruturar e administrar os ativos da informação. Com essa estratégia, busca-se aprimorar a eficiência operacional e promover a rentabilidade do negócio a partir do suporte à tomada de decisão.

A governança de dados proporciona a estrutura e processos para a sustentabilidade da qualidade de dados [26]. À medida que busca estabelecer uma estratégia de gerenciamento da informação, a governança de dados envolve regras de governo para os dados – como indicadores de desempenho, prioridades e responsabilidades – e essa abordagem é necessária no provimento da qualidade.

A implementação de qualidade de dados é um projeto contínuo que mesmo após terminada sua execução, ou seja, a correção de falhas e inclusão de políticas de qualidade deve manter a manutenção de sua garantia. É importante lembrar que existem várias correntes de Governança de Dados no cenário atual, no entanto, para este trabalho são considerados os esforços da IBM neste tópico e a utilização de conceitos do *framework* de governança de dados da mesma empresa. O framework é apresentado na Figura 3.

Figura 3. Framework de governança de dados da IBM



Fonte: extraído de [32]

Os quatro grupos existentes no *framework*: Efeito/Resultado, Autorização, Disciplinas Essenciais e Disciplinas de apoio são compostos de 11 disciplinas essenciais [32] No grupo Efeito/Resultado temos:

- Administração de riscos de dados e Conformidade: define a metodologia pela qual os riscos são identificados, qualificados, quantificados, evitados, aceitos ou mitigados.
- Valor criado: processo no qual ativos relativos a dados são qualificados e quantificados para permitir ao negócio maximizar o valor criado.

Do grupo Autorização, fazem parte as disciplinas:

- Estrutura e consciência Organizacional: descreve o nível de respondabilidade entre negócio e TI nos diferentes níveis da administração.
- Políticas: articulação por escrito do comportamento organizacional desejado em relação aos dados obtidos e produzidos.

- Administração: disciplina de controle de qualidade definida para assegurar o cuidado do dado de forma a aumentar seu valor como ativo.

Do grupo Disciplinas essenciais, fazem parte as disciplinas:

- Qualidade de dados: define métodos para mensurar, aprimorar e certificar a qualidade e integridade da produção, teste e arquivamento de dados.
- Administração do ciclo de vida da informação: política sistemática com enfoque na coleção, utilização, retenção e deleção da informação.
- Segurança e Privacidade: descrição de políticas, práticas e controles utilizados pela organização para mitigar risco e proteger seu ativo, o dado.

Do grupo Disciplinas de apoio, fazem parte as disciplinas:

- Arquitetura de dados: tratamento do desenho arquitetural de dados estruturados e não-estruturados utilizados nos sistemas e aplicações.
- Classificação / Metadados: refere-se aos métodos e ferramentas utilizadas para criação de definições semânticas comuns para termos de negócio e TI.
- Auditoria e Relatório: definição dos processos organizacionais para monitoramento e mensuração do valor dos dados, riscos e eficácia da governança.

Para o objetivo deste trabalho, os conceitos de governança de dados serão aplicados em todo o processo da construção do processo devido à similaridade dos objetivos.

Com as disciplinas de suporte apresentadas neste capítulo espera-se:

- contemplar a rastreabilidade de informações, através de requisitos corretamente levantados;
- especificar bem os dados, facilitando a aplicação da qualidade; e,
- realizar a documentação contemplando as etapas do processo facilitando a gerência e fornecendo meios que garantam a qualidade de dados.

Capítulo 3

Processo Proposto

Este capítulo apresenta o processo proposto para suporte à qualidade de dados. Este inclui conceitos da metodologia TDQM [25] relacionando-os com outros recursos que oferecem estratégias para otenção de qualidade como: governança de dados e o Framework IBM [28] (mais especificamente a qualidade de dados). De uma forma complementar o capítulo também descreve as técnicas de processamento de dados propostas para identificar e tratar a má qualidade dos dados (como por exemplo, higienização e mineração dos dados), assim como fazer a sua documentação.

3.1 Visão geral

Em sua visão geral, o processo é composto de 4 etapas descritas a seguir:

- Etapa 1 – Diagnóstico. Esta etapa orienta a avaliação da qualidade nos dados de um determinado sistema; ela inclui métodos de pesquisa a serem usados para a coleta da informação; analisa aspectos relacionados com o impacto da falta de qualidade sobre o negócio, e o custo *versus o* benefício da aplicação de técnicas para sua melhoria. As informações resultantes desta etapa definem o motivo e o objeto (parte do sistema) sobre o qual vai ser implementada a qualidade .
- Etapa 2 – Análise de requisitos do problema. Esta etapa contempla o mapeamento das necessidades, através da coleta e avaliação dos dados envolvidos no processo (objeto identificado na Etapa 1); os dados são especificados e verificados quanto à sua qualidade.
- Etapa 3 – Desenvolvimento de ações. Esta etapa promove o uso de técnicas para o ajuste dos dados como: regras de negócio (geralmente definidas a nível de requisitos), higienização dos dados, testes de escalabilidade e mineração de dados.

- Etapa 4 – Suporte pós-produção. Nesta etapa final, o objetivo é garantir a continuidade da manutenção da qualidade no sistema através de medidas como: treinamento e verificação e a conclusão da documentação dos passos realizados neste processo.

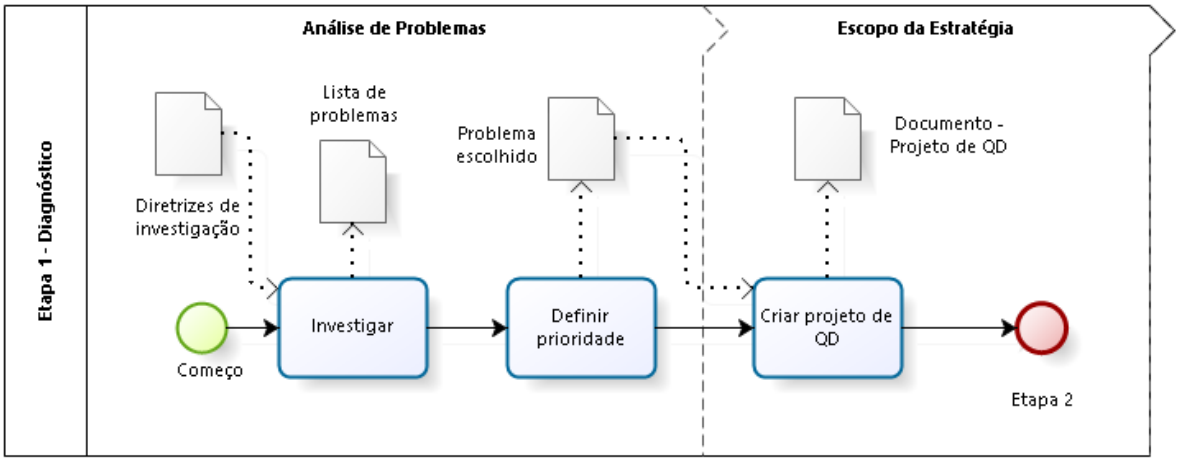
Enquanto que as Etapas 1 e 2 são direcionadas apenas a sistemas que já existem, a Etapa 3 e 4 podem ser aplicadas tanto num processo reativo, no qual o sistema já existe, ou num proativo, ou seja, presentes na criação e concepção de um sistema. Ao longo de cada etapa será apresentada a modelagem correspondente, para isso será utilizado o software Bizagi Suite [35].

3.2 Etapa 1: diagnóstico

Em qualquer tipo de organização, por mais que seja difícil manter o padrão de qualidade em todos os seus sistemas de informação, a implementação de qualidade deve ser priorizada, uma vez que sua ausência pode ocasionar grandes prejuízos. A referência a essa importância deve fazer parte da argumentação, direcionada à organização, na justificativa dos esforços incluídos no processo.

Para estabelecer a justificativa do esforço, tempo e custo a ser despendido, pode-se começar por escolher um ou mais problemas críticos. Observa-se que nesta etapa inicial vários aspectos técnicos ainda são desconhecidos, geralmente tem-se somente suas consequências; sabe-se que por exemplo, uma determinada empresa de telefonia é obrigada a pagar valores altos em multas por cobrança duplicada – neste caso tem-se um problema típico e altamente provável de falta de qualidade relacionada a dados. Logo, a partir da consequência pode-se rastrear o motivo, o objeto que necessita da qualidade. As informações relevantes a serem coletadas sobre o problema tornam-se perceptíveis através da análise das consequências da falta de qualidade de dados; essa coleta deve ser realizada no domínio da organização, conforme processo descrito na Figura 4, que é subdividido em dois outros sub-processos (Análise de Problemas e Definição do Escopo da Estratégia) detalhados a seguir.

Figura 4. Modelagem da Etapa 1: diagnóstico



3.2.1 Análise de problemas

Neste passo é necessário estabelecer quais são os problemas que podem ter ligação com a ausência de qualidade de dados observada na organização. Para isto é preciso conhecimento e habilidade na percepção das falhas existentes.

Algumas diretrizes de investigação podem orientar na identificação dos problemas e falhas críticos em um determinado caso de negócio; geralmente são considerados fatos como:

- vendas e oportunidades perdidas,
- custos desnecessários,
- baixa produtividade, e
- catástrofes

Formas complementares, que ajudam no levantamento de problemas, podem ser aplicadas internamente à empresa – através de pesquisa entre as pessoas envolvidas com o sistema, observação direta do sistema, e inspeção física; ou a partir de levantamentos externos – como, por exemplo, entre consumidores.

Um exemplo de percepção de falhas de custo desnecessário é dado a seguir. Em um determinado banco são destinados R$ 2 milhões para correspondências

anualmente; no entanto, 30% das cartas enviadas não chegam ao destinatário corretamente. O problema, neste caso, pode estar no cadastro de clientes, incompleto e suscetível falhas e provavelmente estará afetando outra tarefa que dependem deste cadastro. Sabendo-se da subjetividade dos diversos problemas que poderão ser encontrados, é preciso priorizar o foco na relação da informação com as categorias da qualidade de dados: intrínseca, contextual, representacional e de acessibilidade.

Após realizado o levantamento dos problemas críticos, é preciso priorizar o problema a ser tratado e identificar seus componentes: dados, pessoas, organizações, processos e tecnologias envolvidas. Esta atividade de identificação é necessária na medida que: uma vez que vários problemas podem ser identificados, deve ser possível determinar o impacto de cada um e por fim, realizar a priorização dos mesmos. É importante salientar que alguns sistemas já podem possuir uma documentação o que facilita muito na identificação dos componentes envolvidos.

3.2.2 Definição do escopo da estratégia para o problema escolhido

Nesta etapa é definida e documentado o escopo do problema escolhido, isto é, o documento relativo ao projeto de qualidade de dados do sistema; o seu objetivo é facilitar a criação da estratégia de ação. O modelo do documento é apresentado na Figura 5.

Figura 5. Modelo de documento do projeto de qualidade de dados (Projeto de QD)

<table>
<tr><td colspan="2">Nome do projeto:
Data:
Responsável:</td><td colspan="2"><nome do projeto>
<data do projeto>
<responsável pelo projeto></td></tr>
<tr><td colspan="2">Histórico de revisão:
Data:
Responsável:
Mudanças:
Técnicas utilizadas:</td><td colspan="2">
<data da revisão após etapa 3>
<responsável pela revisão após etapa 3>
<mudanças realizadas após etapa 3>
<técnicas utilizadas a partir da etapa 3></td></tr>
<tr><td colspan="2">Recursos:</td><td colspan="2"><listagem do responsável pelo projeto, membros do time do projeto e departamentos associados></td></tr>
<tr><td>Falhas</td><td>Processos</td><td>Prioridade</td><td>Comentários</td></tr>
<tr><td><descrição da falha encontrada></td><td><processos relacionados com a falha></td><td><prioridade atribuída ao problema></td><td><observações></td></tr>
</table>

Escopo do projeto:	<problema escolhido> <prioridade> <consequências à falta de qualidade de dados> <benefícios esperados>
Condições do projeto:	<riscos observados associados ao projeto e ou dependências> <estimativa de custo> <estimativa de tempo>
Treinamento:	<observações relativas ao treinamento, abrangência e métodos utilizados (Etapa 4)>

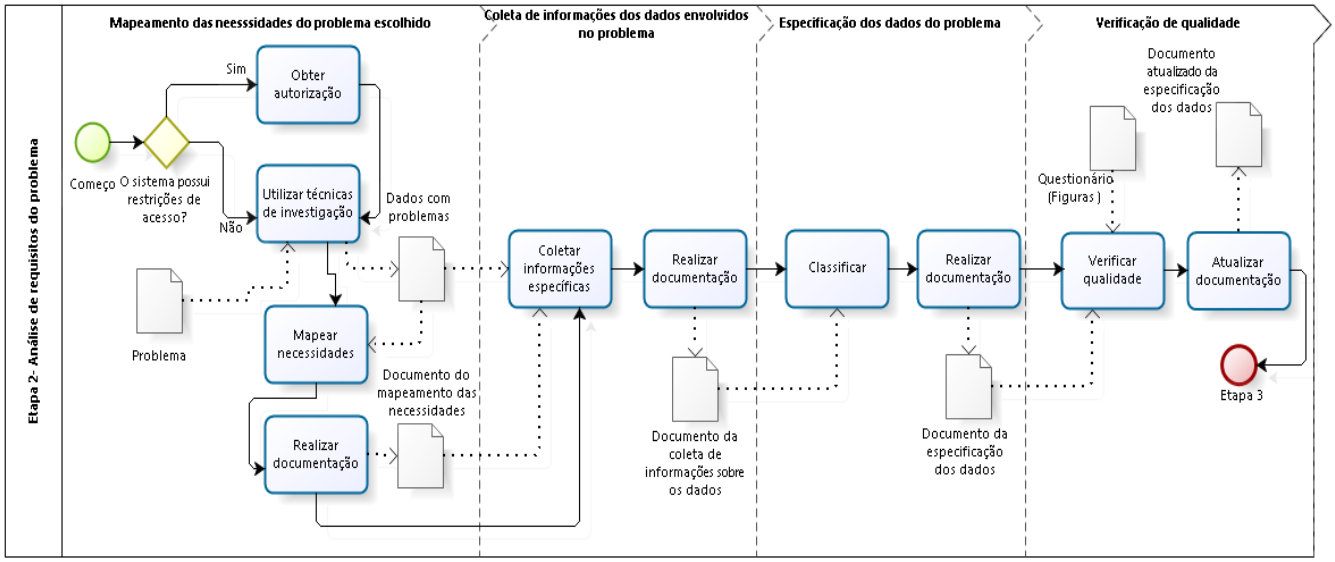
Fonte: parcialmente adaptado de [26]

O objetivo da Etapa 1 como um todo, além de incluir a escolha do problema – que irá influenciar na estratégia de ação – é ter o escopo de abrangência bem definido.

3.3 Etapa 2: análise de requisitos do problema

Esta etapa refere-se à análise de requisitos como estratégia de investigação a partir da informação adquirida na etapa 1, para a aplicação de conceitos de qualidade da metologia TDQM. A estrutura da Etapa 2 é descrita através de quatro sub-processos conforme apresentado na Figura 6. São eles: Mapeamento das necessidades do problema escolhido, Coleta das informações dos dados envolvidos no problema, Especificação dos Dados do Problema, e Verificação da qualidade.

Figura 6. Modelagem da Etapa 2: análise de requisitos do problema



3.3.1 Mapeamento das necessidades

Esta etapa visa investigar melhor o problema para que seja possível tomar conhecimento da sua origem. Nessa investigação, prováveis riscos ao projeto, como restrições de segurança de acesso, deverão ser solucionados/eliminados; para esta investigação podem ser utilizadas técnicas como pesquisas, entrevistas, avaliação da documentação e uso do sistema. Outro artifício para o estudo do problema, que possibilita a estimativa de estatísticas a partir de uma varredura nas tabelas do sistema, são ferramentas de qualidade de dados; empresas como Oracle, ASSESSO e HP são algumas que dispõem de ferramentas que realizam estas buscas e são bem aceitas no cenário atual; entre essas ferramentas estão:

- HP *Quality Center Enterprise*: calcula o risco e os esforços de todos os requisitos do sistema e realiza o gerenciamento de qualidade [30].
- ASSESSO *DataCare*: a versão *Batch* é voltada para a construção de processos que tratam diretamente os arquivos de dados, sem a necessidade de interação com as aplicações da empresa. Já a versão Transacional disponibiliza as funções do *DataCare* para os sistemas da empresa que necessitam validar e tratar dados e, desta forma, garantir a

qualidade da informação no momento da sua captação, seja em ambiente *batch* ou *on-line*, *in-house* ou via internet [7].

- *Oracle Enterprise Data Quality*: provê, dentre outras funcionalidades, a padronização, análise e higienização dos dados [29].

Como resultado da investigação é esperado um mapeamento das necessidades contemplando as seguintes características do problema:

- Problema: título do problema específico que deve ser corrigido no que se refere à qualidade de dados.
- Características: origem dos dados associados; se existirem, normas internas e políticas relacionadas a estes dados; e a informação obtida dos dados.
- Dependências: processos relacionados ao problema, armazenamento, organizações, e informações produzidas pelo problema.

O modelo do documento descrito pode ser encontrado na Figura 7. O nível de detalhe nesta etapa deve ser proporcional à grandeza do problema.

Figura 7. Modelo de documento do mapeamento de necessidades

Problema:	<título do problema específico>
Data:	<data do mapeamento>
Responsável:	<responsável pelo mapeamento>
Características:	
Origem:	<repositório de origem dos dados>
Normas internas:	<existência de regras ou padrões aplicados aos dados>
Informação obtida:	<qual a informação obtida através dos dados>
Dependências:	
Sistemas:	<sistema(s) que utilizam os dados>
Processos:	<processos que dependem dos dados do problema>
Armazenamento:	<repositórios que utilizam os dados do problema>
Aplicações:	<aplicações que utilizam os dados do problema>
Organizações:	<outras organizações que utilizam os dados do problema>

Fonte: parcialmente adaptado de [26]

3.3.2 Coleta de informações sobre os dados envolvidos no problema

Como consequência do mapeamento das necessidades, é preciso coletar os dados envolvidos no processo no qual o problema está inserido e é importante à medida que verifica o que o dado deve significar e do que está, no momento atual, significando. Esta coleta tem o objetivo de reunir informações específicas sobre o dado como:

- Cabeçalho: nome atribuído (campo), banco de dados onde se encontra armazenado, tabela e aplicação relacionada.
- Categoria: definição do tipo de dado de acordo entre: *máster,* transacional, referência, metadado, histórico ou temporário. Dependendo da categoria o dado deverá conter informações obrigatórias ou não. Por exemplo, para dados de referência são esperados o conjunto de valores aceitos, padronizados e as tabelas de referência.
- Descrição: deve conter a função que o dado exerce no sistema.

O modelo do documento descrito pode ser encontrado na Figura 8.

Figura 8. Modelo da coleta de informações sobre os dados envolvidos no problema

Repositório: Data: Responsável: Nome: Tabela(s):	<local onde estão armazenados os dados do problema> <data da coleta> <responsável pela coleta> <campo atribuído ao dado> <tabela(s) associada(s) ao dado>
Categoria do dado:	<definição entre *máster,* transacional, referência, metadado, histórico ou temporário>
Descrição do dado:	<contêm a informação que o dado exerce no sistema>

Fonte: parcialmente adaptado de [26]

3.3.3 Especificação dos dados do problema

A coleta das informações nesta fase permite a especificação dos dados do problema. Esta tarefa é base para a inserção no contexto de qualidade, devendo abranger (ver Figura 9):

- Padrões de dados: deve ser realizada a verificação dos dados para saber se atendem a padrões como: convenção de nomes, normas de entrada (tratamento de pontuação, letras maiúsculas, abreviações aceitas, entre outros) e valores de referência, classificação e categorização (CNAE, SIC e NAICS por exemplo).
- Modelos de dados: deve ser realizada a verificação dos dados em relação ao tipo de chave (primária, estrangeira), cardinalidade (quantas instâncias devem ser relacionadas a outra instância) e nulidade(obrigatória ou não).
- Manipulação (permissões) de dados: devem ser verificadas quais são as operações relacionadas aos dados (inserção, atualização, deleção e seleção). É necessário também contemplar os relacionamentos com outras tabelas e repositórios de dados.

Figura 9. Modelo da coleta de informações sobre os dados envolvidos no problema

Repositório: Data: Responsável:	<local onde estão armazenados os dados do problema> <data da especificação> <responsável pela especificação>
Padrão de dado: Existência de padronização? Descrição	 <presença ou ausência da existência de padronização> <Descrição das regras utilizadas>
Modelo do dado: Tipo de chave: Cardinalidade: Nulidade:	 <especificação entre chave primária e estrangeira> <quantas instâncias estão relacionadas a outras instâncias> <resposta sobre a obrigatoriedade do dado>
Manipulação dado: Operações realizadas: Relacionamento:	 <operações realizadas entre inserção, atualização, deleção e seleção> <existência de relacionamento entre outras tabelas e repositórios >
Verificação de qualidade: Ausência de características:	 <categorizar o dado de acordo com o esquema das Figuras 10 e 11>

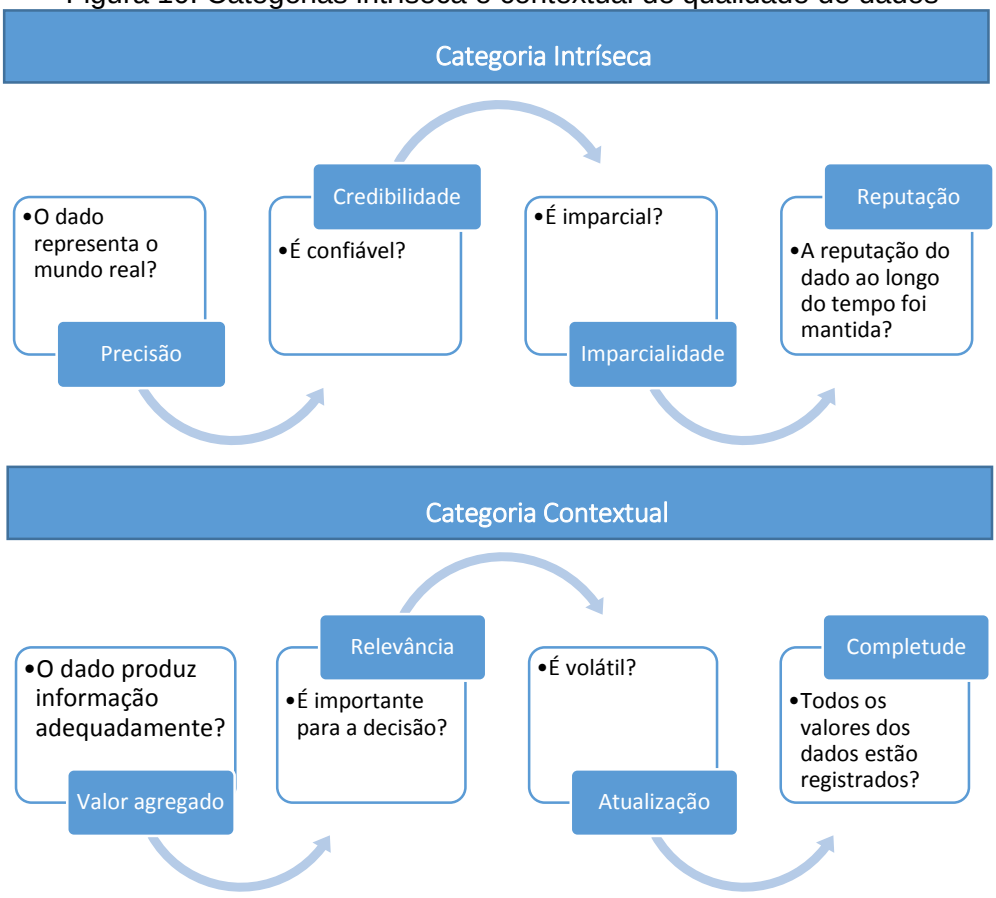
Fonte: parcialmente adaptado de [26]

3.3.4 Verificação de qualidade

Após ser realizado o mapeamento das necessidades (item 3.3.1), é possível inserir as dimensões e categorias de qualidade de dados da metodologia *Total Quality Management Data* [27]. Neste passo deverão ser verificadas as propriedades de acordo com as categorias (ver Figura 10 e 11):

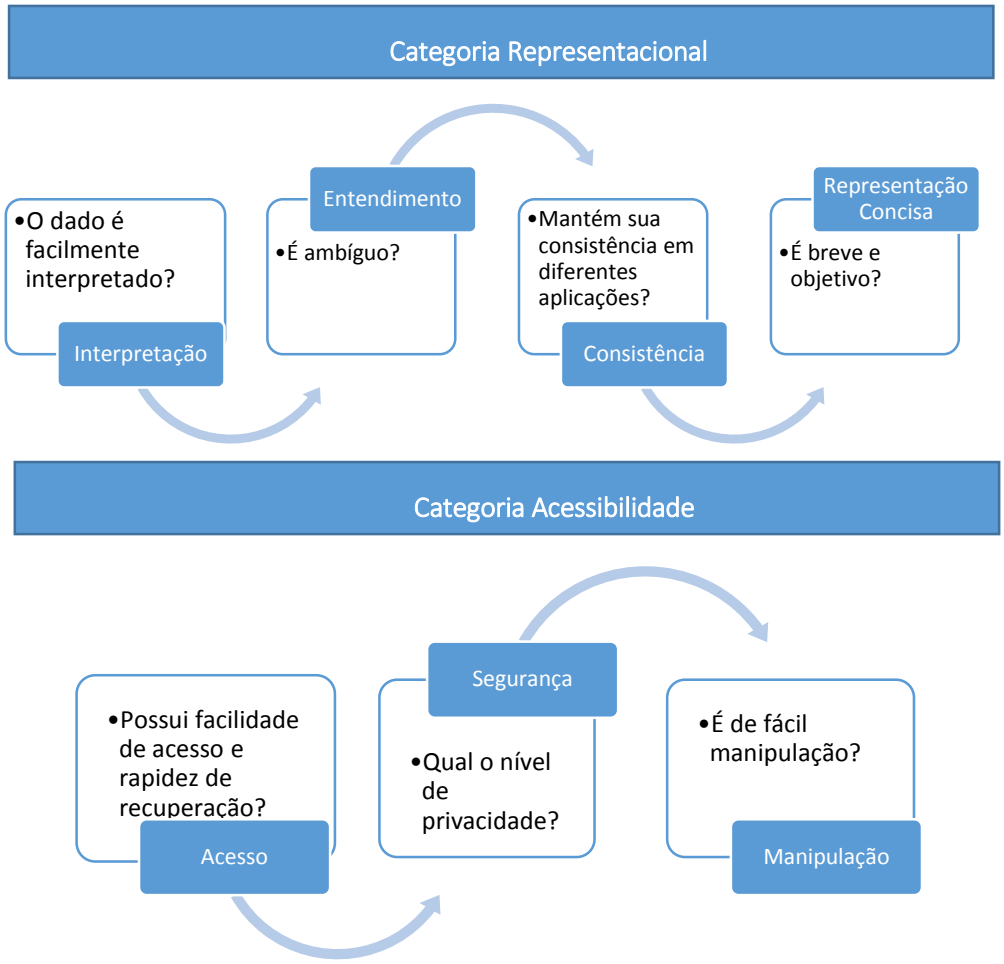
- Intrínseca – propriedades relacionadas com a precisão, credibilidade, imparcialidade e reputação do dado;
- Contextual - propriedades relacionadas com valor agregado, relevância, atualização, e completude;
- Representacional - propriedades relacionadas com interpretação, consistência e representação concisa;
- acessibilidade de qualidade de dados - propriedades relacionadas com acesso, segurança e manipulação.

Figura 10. Categorias intríseca e contextual de qualidade de dados



Fonte: adaptado do Total Data Quality Management

Figura 11. Categorias representacional e acessibilidade de qualidade de dados



Fonte: adaptado do Total Data Quality Management

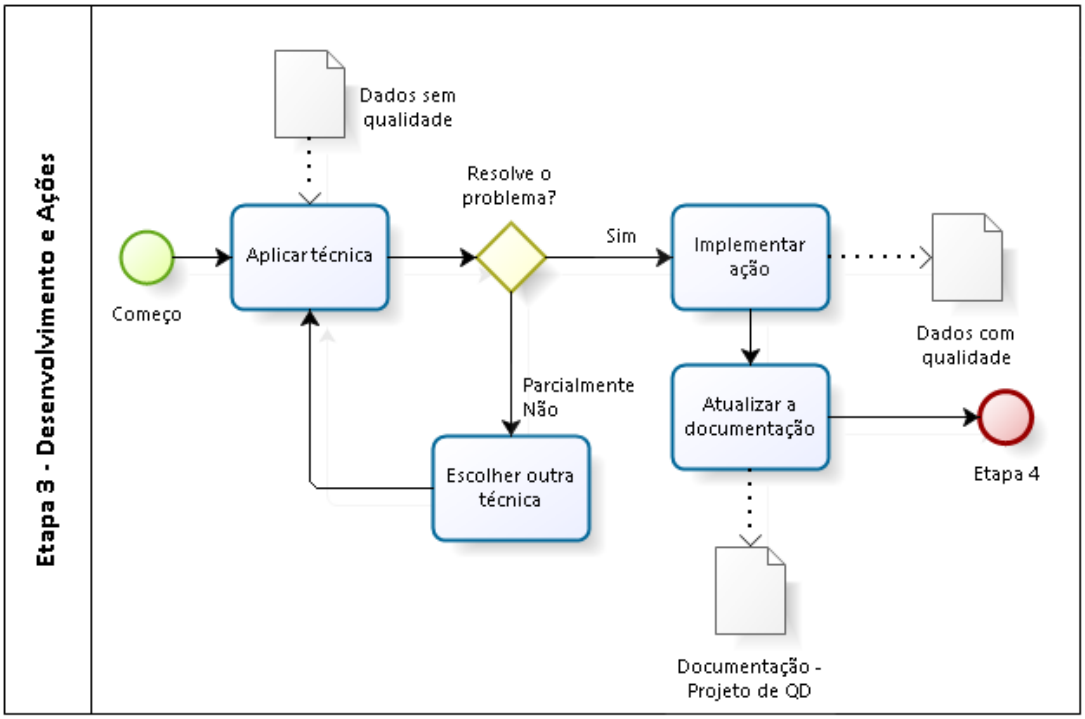
Como explicado no Capítulo 2, as categorias e o que elas representam ao dado – são atributos que ajudam a compor um dado de qualidade. Portanto, ao analisar um determinado dado, a partir do seu mapeamento e investigação, é possível descobrir quais atributos estão falhos ou não existem, proporcionando sua classificação como sendo ou não de qualidade. Após este passo é essencial atualizar a documentação da coleta de informações sobre os dados envolvidos no problema, incluindo finalmente, se o dado atende às dimensões de qualidade; caso negativo, são determinadas quais dimensões precisam ser garantidas através de correção.

3.4 Etapa 3: desenvolvimento e ações

Nesta etapa são desenvolvidas ações visando melhorar a qualidade dos dados. A partir da coleta e classificação, realizadas nas etapas 1 e 2, é possível relacionar

técnicas para obtenção da qualidade preenchendo os requisitos não atendidos e, por fim, ajustando os dados de origem de acordo com o problema. Algumas técnicas são básicas para implementar correção em estruturas de armazenamento de dados como: checagem de valores válidos em listas e a determinação da exclusividade de valores de uma determinada tabela. Porém, para uma abordagem mais completa é necessário o estudo de outras técnicas mencionadas nesta etapa. A estrutura da Etapa 3 é descrita através do fluxo mostrado na Figura 12.

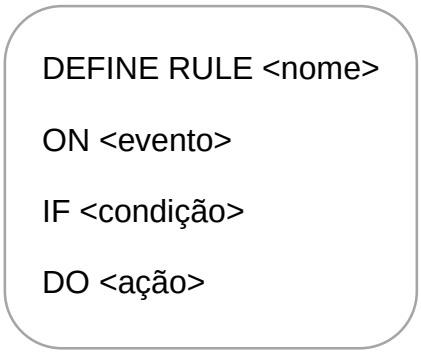
Figura 12. Modelagem da Etapa 3: desenvolvimento e ações



3.4.1 Regras de negócio

As regras de negócio (RN) têm o objetivo de determinar a forma com que o negócio será executado a partir de regras que causem, previnam ou sugiram o acontecimento de ações. Elas podem ser aplicadas para garantir dimensões de qualidade de dados a partir de regras de evento, condição e ação (ECA), genericamente ilustrada na Figura 13.

Figura 13. Regra ECA



Fonte: extraído de [39]

Algumas regras de negócio podem ser aplicadas genericamente para prover qualidade, sendo ajustadas de acordo com o problema (num cenário reativo) e na implementação e criação de um sistema (cenário proativo). Regras de negócio devem ser utilizadas para estabelecer padrões para ações que resultem na integridade e no comportamento do dado. Tipos de regras mais comuns são listados a seguir [4, 8, 24, 26]:

- RN1 – Cada evento deve estar associado a apenas a uma tabela.
- RN2 – A informação referente a um banco de dados deve estar associada ao usuário que acessou o sistema; desta forma, é possível o uso de bancos de dados executando aplicações diferentes pela mesma ferramenta, mas com usuários diferentes.
- RN3 – As tabelas disponibilizadas ao usuário deverão ter relacionamento direto com a tabela do evento (1:N e N:N).
- RN4 – Cada campo da tabela deverá ser documentado com nome, descrição, caso de uso, obrigatoriedade de sua existência e valor padrão.
- RN5 – A classificação e categorização do sistema deverá seguir valores de referência do padrão brasileiro CNAE.
- RN6 – Todos os tipos de conta deverão seguir a nomenclatura CN_. Exemplo CN_Corrente, CN_Investimento.
- RN7 – O sistema deverá aceitar apenas as abreviações previamentes definidas.

- RN8 – Os nomes para os campos da tabela deverão ser precisos e de fácil compreensão.
- RN9 – Números não serão aceitos nos campos de tipo nome.

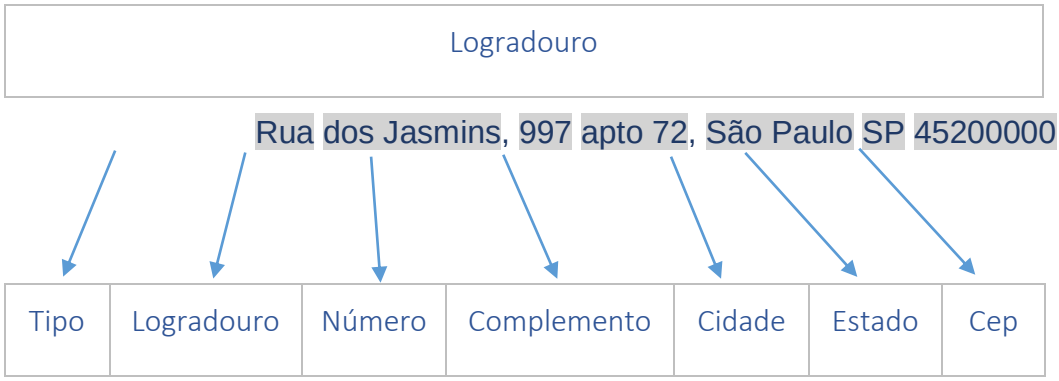
3.4.2 Higienização de dados

A higienização de dados é o processo de limpeza de uma base de dados. Neste procedimento é realizada a detecção e remoção de erros e inconsistências. O processo de limpeza envolve um conjunto de atividades na análise, transformação, conversão, verificação, restrição de valores, fusão de dados, entre outros. A aplicação das atividades está definida e exemplificada a seguir, e inclui técnicas como: *parsing*, análise de digito verificador, verificação de registros, e uso de entradas pré-definidas.

3.4.2.1 *Parsing*

Este método realiza a verificação de uma determinada variável para identificar ocorrências que podem ser tratadas separadamente como na Figura 14.

Figura 14. Realização do *parsing* na variável endereço



Fonte: adaptado de [36]

3.4.2.2 Dígito verificador

Dígito verificador ou algarismo de controle é uma técnica utilizada para verificar a autenticidade de um valor numérico. Ele visa prevenir erros na entrada dos dados e consiste de um ou mais algarismos acrescentados ao valor original; seu valor é calculado através de um determinado algoritmo. As rotinas mais tradicionais de algoritmo são módulos 10 e 11; outra bastante utilizada ainda é o módulo 10, que é

calculado conforme esquematizado na Figura 15, que apresenta um exemplo de conta corrente.

Figura 15. Realização do algoritmo módulo 10

1	4	8	7	2	1
*1	*2	*1	*2	*1	*2
1	8	8	14	2	1
1	4	8	5	2	1

1 + 4 + 8 + 5 + 2 + 1 = (21÷10) = 2, resto 1. Dígito verificador = (10 – 1) = 9

No esquema da Figura 15, para a conta corrente de número 148721 foi obtido o dígito verificador 9. O cálculo é realizado selecionando cada dígito do número, a partir do menos significativo para o mais significativo. Em seguida, é multiplicado, na ordem, por 2, depois 1, e assim sucessivamente. Após, será realizado o somatório dos dígitos das multiplicações. Este, será dividido por 10 e se o resto (módulo 10) for diferente de zero, o dígito será 10 subtraído deste valor. Se o resto (módulo 10) for zero (0) o dígito verificador também será zero (0).

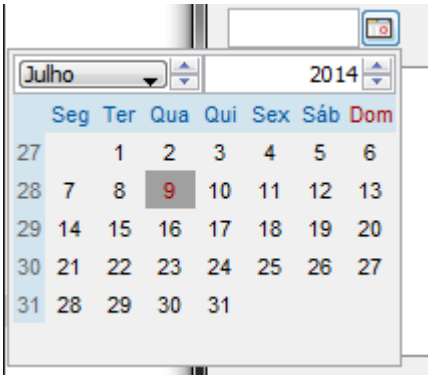
3.4.2.3 Verificação de registro

Os registros informados em cadastro podem ser verificados diretamente no cadastro original certificado. Essa ação faz com que o sistema só aceite a manipulação do dado se ele for validado antes. Permite também que dados verídicos ou precisos sejam extraídos do banco de dados consultado. Por exemplo, para a classificação de registro de pessoa física ou jurídica, pode ser utilizado o sistema da Receita Federal na verificação do número informado e ainda, sua situação como: ativo, pendente ou cancelado. Outros exemplos de sistemas para verificação incluem dados de falecimentos (fornecidos por cartórios) e dados de endereço (fornecidos pelo Diretório Nacional de Endereços).

3.4.2.4 Entradas pré-definidas

Para valores que podem ser pré-estabelecidos, esta regra deve ser aplicada no sistema. Sua aplicação faz com que, possíveis erros de digitação ou desconhecimento do campo, produzam dados inconsistentes. Nos campos que recebem a entrada de datas, no exemplo da Figura 16, é possível aplicar a higienização definindo valores aceitáveis como base para dias, meses e anos. É importante também, se aplicável, a verificação de exceções, como o ano bissexto. Desta forma, 29 de fevereiro pode ser válido.

Figura 16. Higienização do campo data



3.4.3 Testes de escalabilidade

Os sistemas de informação muitas vezes são projetados para um determinado número de usuários. Com o passar do tempo, esses sistemas ultrapassam a capacidade antes estabelecida e apresentam falhas ou um processamento e capacidade abaixo do requerido. Então, é importante testar o sistema em situações extremas, assim mensurando o limite de seu desempenho.

Para aplicações destinadas a sistemas integrados e distribuídos a preocupação deve incluir a segurança dos dados. Políticas de segurança como criptografia por exemplo, consomem mais recursos do sistema alterando seu desempenho.

3.4.4 Mineração de dados

A mineração permite o pré-processamento de grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências

temporais a partir de relacionamentos sistemáticos entre variáveis, detectando assim, novos subconjuntos de dados [15]. Esta técnica pode ser utilizada na limpeza, integração, transformação, redução e discretização de dados. Pode ser utilizada, ainda, como uma ferramenta para reprodução do valor agregado ao dado, auxiliando o suporte à decisão.

A seguir, possíveis problemas de qualidade que podem ser resolvidos por mineração:

- Dados incompletos: realizar o processamento de dados na descoberta de valores e atributos faltantes. Ex: nome = ””.
- Dados ruidosos: realizar o processamento de dados para encontrar valores e atributos contendo erros. Ex: salário = “-54”.
- Dados inconsistentes: realizar o processamento de dados para encontrar valores e atributos contendo discrepâncias. Ex: data de nascimento= “29/05/1990”, idade = “38”.

Algumas soluções de mineração em qualidade de dados:

- Resolver redundâncias: definição de um intervalo para um determinado atributo e de medidas de correção para registros com ocorrência fora do intervalo escolhido.
- Resolver dados faltantes: substituição por valores “*null* ”; preenchimento com medidas estatísticas (média e moda); utilizar modelos preditivos para sugestão no preenchimento.
- Resolver dados ruidosos: clusterização na detecção e remoção de *outliers*; detecção de valores suspeitos; suavização através do ajuste de dados.
- Previsão: estimativa de vendas através da previsão de cargas ao servidor ou tempo de inatividade.
- Agrupamento: separando entidades de um sistema a partir da análise e previsão de afinidades entre atributos.

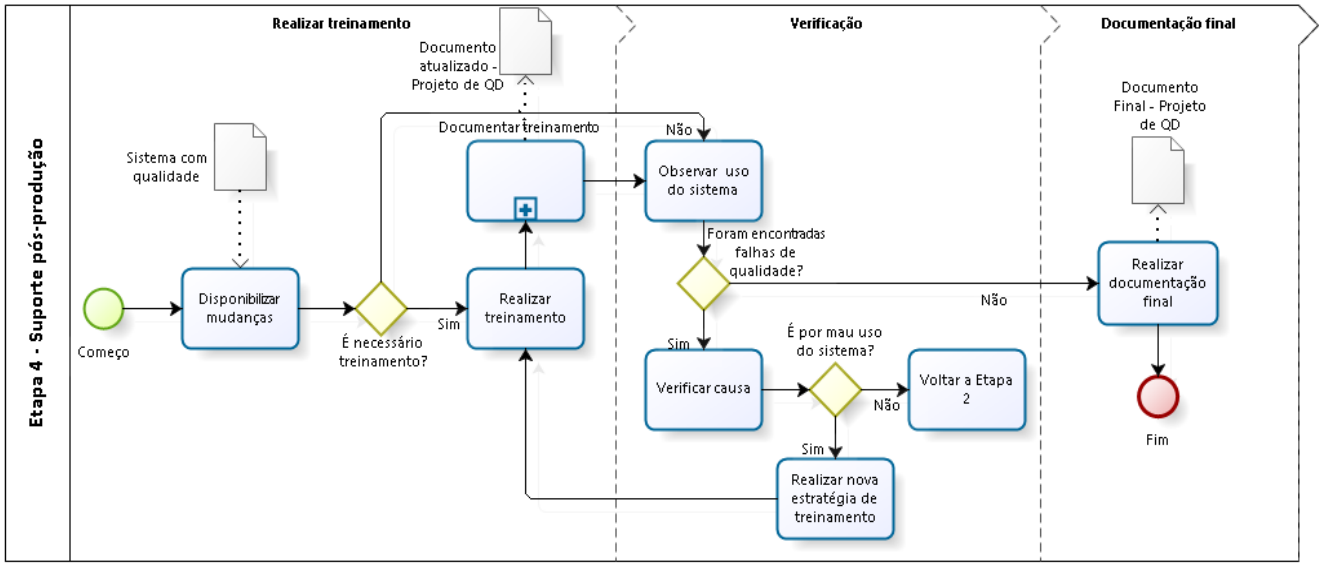
Técnicas de mineração, no entanto, não são indicadas para sistemas de dados extremamente falhos em qualidade, uma vez que estes não provêm indicadores aceitáveis para a previsão.

As técnicas apresentadas nesta etapa devem ser adaptadas de acordo com a necessidade do problema. Mesmo considerando a unicidade dos sistemas, as dimensões de qualidade são mantidas para os dados, logo, as técnicas apresentadas são estáveis na implementação de qualidade. Em sistemas complexos e integrados, por exemplo, pode-se utilizar recursos como a modelagem e simulação da estratégia de qualidade antes de sua aplicação no sistema real.

3.5 Etapa 4: suporte pós produção

Esta última etapa tem o objetivo de, após o diagnóstico, análise de requisitos e desenvolvimento de ações, demonstrar técnicas e estratégias que garantam a continuidade da qualidade no sistema. A estrutura da Etapa 4 é descrita através do fluxo mostrado na Figura 17.

Figura 17. Modelagem da Etapa 4: suporte pós produção



3.5.1 Realizar treinamento

Os problemas com a qualidade dos dados são causados por diversos fatores, começando pela entrada de dados manual, propensa a erros e de funcionários em vários departamentos, cada um com regras e métodos próprios.

Uma das medidas obrigatórias para a garantia da qualidade de dados é o treinamento. Ao realizar mudanças no sistema ou condicioná-lo a fim de garantir a qualidade de dados, é importante aplicar as mudanças – não somente a partir da documentação - mas com atividades de apoio ao desenvolvimento de habilidades para a garantia qualidade mínima do sistema. A implementação de qualidade de dados visa diminuir a possibilidade dos dados não atenderem às dimensões de qualidade; no entanto, algumas funcionalidades podem não receber esta cobertura, por isso o condicionamento a um bom uso do sistema é essencial.

Independente do papel do usuário dos dados (captação, manutenção e uso) o padrão de qualidade escolhido e implantado deve atender a todas as camadas de usuários. O processo de treinamento deverá ser adaptado à organização envolvida. A criação de um ambiente virtual para experimentação do sistema e apresentação da documentação, por níveis de usuário, é uma alternativa que satisfaz e dá apoio a esse processo.

3.5.2 Verificação

Após a implementação de políticas de qualidade é importante verificar a consistência das mudanças geradas. Conflitos entre tecnologias e sistemas agregados ao projeto, treinamento falho de usuários e divergências entre o que foi implementado em comparação ao que era necessário, podem ameaçar a integridade dos dados no sistema. Para este passo, um acompanhamento temporal deve ser realizado através de folhas de verificação – que deverão estar contidas na documentação final; esta é uma alternativa na verificação, como mostrado na Figura 18.

Figura 18. Modelo de folha de verificação

Verificação de qualidade

Período : 20/10/2014 a 27/10/2014 (semana 3)

Responsável:

Funcionalidade: cadastro de clientes

Falhas	Ocorrências	Percentual/Total	Relação com a semana 2
Cadastro duplicado	10	0.04%	Redução
Cadastro incompleto	32	0.98%	Estável
Cadastro desatualizado	50	1.32%	Redução

3.5.3 Documentação final

Para qualquer implementação de qualidade de dados, é essencial desenvolver uma documentação das atividades desenvolvidas em cada etapa da execução deste processo. Esta ação deve ser aplicada nas etapas de *design* da base de dados, iteração das aplicações e acessibilidade. Seu objetivo principal é auxiliar na produção de alta qualidade, reuso de técnicas e dados e prevenir potenciais problemas na pós-produção como: conflitos de definições e normas, redundância e dificuldades no compartilhamento de dados entre aplicações.

Uma documentação de qualidade deverá conter:

- Análise dos problemas (item 3.2.1)
- Escopo da estratégia (item 3.2.2)
- Mapeamento das necessidades do problema (item 3.3.1)
- Coleta de informações sobre os dados envolvidos no processo (item 3.3.2)
- Especificação dos dados do problema (item 3.3.3)
- Técnicas utilizadas na implementação de qualidade (etapa 3)
- Documentos de estratégia de treinamento (item 3.5.1)

- Técnicas de verificação (item 3.5.2)

Baseando-se na aplicação deste processo, várias direções podem ser escolhidas. As etapas aqui apresentadas são um recurso para a análise e medições na construção da qualidade de dados.

Neste capítulo, estruturou-se a qualidade de dados envolvendo processos distintos entre seu desenvolvimento e uso. No próximo capítulo, será aplicado o processo aqui desenvolvido tornando mais visível sua aplicação em um sistema.

Capítulo 4

Aplicação do processo e resultados

Este capítulo descreve a aplicação do processo a um sistema real realizada para avaliar o processo proposto, assim como os resultados obtidos. Dentre os casos analisados, inseridos no Apêndice B, o caso escolhido foi um sistema de grande porte, de uma organização pública, integrado a outros sistemas públicos e privados; a função deste sistema é manter a manipulação e gerência do cadastro de óbitos no território brasileiro – Sistema de Controle de Óbitos (SISOBI), desta forma provendo informações à organização pública sobre o pagamento de benefícios – Sistema Único de Benefícios (SUB). A implementação do processo nesta etapa visa sugerir, a partir de um sistema sem qualidade dados, estratégias na qualificação do sistema e dos dados.

O principal objetivo do Sistema de Controle de Óbitos é cancelar ou suspender benefícios em virtude do óbito dos segurados, por meio de cruzamentos com a base de dados do Sistema Único de Benefícios. Entretanto, o SISOBI vêm apontando falhas nesse processo, à medida que terceiros se aproveitam do benefício de uma pessoa que já faleceu.

Os dados do SISOBI são abastecidos pelos cartórios de registro civis. Estes, constituem a fonte primária das informações de óbitos ocorridos em sua área de competência. Tendo em vista que o objetivo principal do Sistema de Controle de Óbitos é dar maior agilidade e segurança aos procedimentos de cancelamento de benefícios, a entrada de dados no sistema é um processo crítico. Informações de identificação como nome e números de documentos pessoais proporcionam a localização de benefícios de pessoas falecidas. Outras informações influenciam os valores pagos, como a data do óbito, que define a data de cessação do benefício previdenciário. Dessa forma, erros, falhas, omissões e atrasos, intencionais ou não, podem gerar prejuízos ao dinheiro público.

Para o Sistema de Controle de Óbitos são avaliados se seus dados correspondem à realidade, considerando que o principal objetivo deste sistema é

cancelar ou suspender benefícios por meio de cruzamentos com a base de dados do Sistema Único de Benefícios.Também, é feita uma análise das perdas financeiras, bem como a avaliação da qualidade de dados no contexto da organização. Para cada etapa apresentada no processo, Capítulo 3, será feita a analogia com aplicação do processo a um exemplo real, aqui dividida em Aplicação por Etapa.

4.1 Aplicação da Etapa 1: diagnóstico

Considerando o caso escolhido, para a Aplicação da Etapa 1 será iniciada a Análise de problemas (item 3.2.1).

A identificação do problema no sistema foi percebida através da investigação por entrevista com os servidores da organização, desta forma obteve-se relatos sobre falhas. Não foi necessária uma investigação mais complexa neste momento inicial pois, a partir do método utilizado, foi possível obter a constatação de uma falha grave: problemas no cancelamento de benefícios previdenciários em razão de óbito do segurado têm ocasionado pagamentos indevidos.O problema, para esta organização desde o lançamento do sistema, acarreta despesas financeiras de cerca de R$ 2,3 bilhões aos cofres públicos [38], prejuízo estimado a partir de uma auditoria do órgão regulador desta organização - e de credibilidade – à medida que o prejuízo financeiro atinge a organização, estimula também a exploração desta falha: pessoas de má-fé, com acesso à manipulação de dados do sistema tentam obter benefícios para terceiros. Ao longo da ocorrência de casos o sistema torna-se depreciado e, consequentemente, a organização envolvida terá responsabilidades jurídicas e fiscais, ocasionando mais prejuízos.

Para uma organização, grande e pública, como a estudada neste capítulo, é preciso considerar que estas duas características estão associadas, de maneira substancial, a falhas causadas por problemas de comunicação e conhecimento no uso do sistema. Outra importante consequência de uma organização grande e pública é a ação deliberada das pessoas envolvidas, que por ventura, possuem algum tipo de conflito com a organização.

Após a coleta de informações por entrevista, utilizou-se a observação direta e inspeção física: os resultados esperados do sistema foram analisados e foi possível

perceber discrepâncias entre por exemplo, no cadastro de nomes, a não utilização de padrões para entrada de dados. A partir desta constatação, tem-se um problema e sua relação com a falta de qualidade aos dados de entrada. Por intuição, se a ausência de Regras de Negócio é perceptível num simples teste de utilização da consulta de cadastros então, deve-se atentar no mínimo ao longo do uso do processo, para dados imprecisos. Na utilização do sistema também foi possível perceber os sistemas integrados diretamente a ele: Cadastro Nacional de Informações Sociais e Sistema Único de Benefícios.

Escolhido o problema e sabendo da sua justificativa para implementar qualidade de dados – evitar custos desnecessários – o próximo passo é definir o escopo da estratégia. Para isso, será utilizado o modelo de documento do projeto de qualidade de dados (Figura 5) e preenchidas as informações conhecidas até o momento, descritas na Figura 19.

Figura 19. Documento do projeto de qualidade de dados para o SISOBI

<table>
<tr><td colspan="2">Nome do projeto:
Data:
Responsável:</td><td colspan="2">ImQob – Implementação de qualidade no cadastro de óbitos
09/10/2014
<responsável pelo projeto></td></tr>
<tr><td colspan="2">Histórico de revisão
Data:
Responsável:
Mudanças:
Técnicas utilizadas:</td><td colspan="2">
<data da revisão após etapa 3>
<responsável pela revisão após etapa 3>
<mudanças realizadas após etapa 3>
<técnicas utilizadas a partir da etapa 3></td></tr>
<tr><td>Falhas</td><td>Processos</td><td>Prioridade</td><td>Comentários</td></tr>
<tr><td>Dados imprecisos, desatualizados.</td><td>Cadastro, atualização.</td><td>Alta</td><td>Relatos sobre a fragilidade do sistema entre funcionários.</td></tr>
<tr><td colspan="2">Escopo do projeto:

Prioridade:</td><td colspan="2">Implementação de qualidade de dados no sistema de cadastro de óbitos.
Alta</td></tr>
</table>

Consequências dos problemas atuais: **Benefícios esperados:**	Custos desnecessários estimados em R$2,3 bilhões. Perda da credibilidade do sistema Perda da credibilidade da organização Reverter as falhas encontradas; garantir integridade da informação.
Condições do projeto: **Observada dependência com os sistemas:**	Cadastro Nacional de Informações Sociais Sistema Único de Benefícios
Treinamento:	<observações relativas ao treinamento, abrangência e métodos utilizados (Etapa 4)>

O SISOBI apresenta uma documentação não uniforme, isto prejudica a integração e continuidade do sistema. No campo de escopo do projeto tem-se a constatação inicial da causa do problema. Porém, durante a investigação das próximas etapas, é possível especializar o objetivo ou ainda descobrir novos problemas menores, diretamente relacionados ao objeto de estudo. Neste caso, os problemas dependentes deverão ser adicionados observando o controle de revisão.

Para a próxima etapa, a análise de requisitos será utilizada como estratégia de investigação a partir da informação adquirida na Etapa 1.

4.2 Aplicação da Etapa 2: análise de requisitos do problema

Para ser feito o aprofundamento do problema é necessária a investigação dos próprios dados, bem como de suas estruturas de armazenamento. Para isso foi obtida a autorização de acesso. Casos como o aqui apresentado requerem discrição acerca das informações do sistema que possam comprometer tanto os usuários quanto a organização se expostas.

Para a investigação do problema verificou-se a consistência dos dados de cadastro. Isto foi possível devido a coleta de estastísticas, método mais apropriado devido ao grande número de entradas. Para tal, foi utilizado um subsistema para uso gerencial e estatístico dos dados de óbitos. Na coleta, foi selecionado o período de um mês. Os resultados são demonstrados na Tabela 4.

Tabela 4. Dados estatísticos sobre a ocorrência de entradas com falhas

DADO	SITUAÇÃO	QUANT.	%TOTAL
Total de registros		**87.670**	**100**
Data nascimento	Inválida	102	0,12
Data nascimento	Zerada	1.259	1,44
Data óbito	Menor que 24/04/1986	108	0,12
Nome falecido	Informação inválida	415	0,47
Nome mãe	Informação inválida	72	0,08
UF nascimento	Inválida	5.419	6,18
Número da identidade	Preenchido com UF em branco	5.069	5,78
Número da identidade	Preenchido com Órgão Emissor de Identidade em branco	6.975	7,96
Município	Inválido	22.290	25,42
CPF	Inválido	37.218	42,45

Fonte: adaptado de [8]

O registro “Data óbito” deve conter datas posteriores a 24 de abril de 1986 devido a uma regra de negócio entre os sistemas de óbito em geral. Para o mês de amostragem foi possível perceber que mais de 50% dos cadastros apresentavam uma ou mais falhas. As mais comuns no entanto, em relação ao CPF (Cadastro de Pessoas Físicas)

inválido, 42,45% das falhas encontradas, e Municípios inválidos, 25,42%. Cada campo com falha torna o cadastro inválido. Com o cadastro inválido, não é possível cancelar o benefício no Sistema Único de Benefícios, pois para qualquer discrepância entre os dados o sistema não pode associar o cancelamento do benefício à pessoa que faleceu porque entende que são duas pessoas diferentes. Logo, para o sucesso da operação de cancelamento, todos os resultados têm que ser igual entre os cadastros. A partir desta limitação é possível também estimar o custo mínimo da ausência de qualidade de dados nesse mês de aferição como mostra a Figura 20:

Figura 20. Estimativa do prejuízo no SUB

Em posse das informações da Tabela 4 é possível, a partir do conhecimento da verificação de qualidade (item 3.3.4), dar continuidade a análise de requisitos. Para isso, será utilizado, adaptado e preenchido o modelo de documento do mapeamento de necessidades, demonstrado na Figura 21.

Figura 21. Documento do mapeamento de necessidades

Problema: Data: Responsável:	Ausência de mecanismos de padronização dos dados 05/10/2014 <responsável pelo mapeamento>
Características: Origem: Normas internas: Informação obtida:	 BD_Obitos N.I. 01: Data óbito superior a 24 abril 1986. Cadastros com dados inválidos, nulos, duplicados e sem obedecer à N.I. 01 representam mais de 50% dos dados do sistema.
Dependências: Armazenamento: Aplicações: Organizações: Sistemas:	 BD_Beneficiados <Sistema de Pagamento de Benefícios> Sistema de Controle de Óbitos e Sistema de Pagamento Benefícios Instituto de Benefícios, Cartórios de Registro Civil Cadastro Nacional de Informações Sociais, Sistema Único de Benefícios

Foram resumidas nas características da Figura 21 os resultados inconsistentes à função do sistema encontrados na Tabela 4. Já nas dependências foi investigado a relação entre os dados e seus usuários. Para esta aplicação, será realizada a suspensão dos sistemas: Cadastro Nacional de Informações Sociais e Sistema Único de Benefícios. O motivo oferecido é que será necessária a alteração do sistema de cadastro. Para um caso de implementação de qualidade apenas nos dados, tal intervenção pode não ser necessária.

A partir do conhecimento mais específico, do banco e dos dados que ele armazena, foi feita a análise das tabelas que armazenam os dados fornecidos no cadastro. O Sistema único de Benefícios – sistema integrado e dependente do Sistema de Controle de Óbitos, possui Regras de Negócio relacionadas às entradas dos dados bem como utiliza certas informações para atualizar a situação do benefício entre mantido e suspenso. Para isso, é preciso ajustar os dados que ambos utilizam na checagem para um padrão único. Neste caso, o padrão escolhido para referência de ajuste no SISOBI será o do SUB.

Figura 22. Modelo de coleta de dados do SISOBI

Repositório:	BD_Obitos
Data:	06/10/2014
Responsável:	<responsável pela coleta>
Tabela(s):	Cadastro_Obitos

Campo	Categoria	Descrição
CO_CPF	*Master*	Número de CPF do falecido. Ex:89976532310
CO_Data_nasc	Referência	Data de nascimento do falecido. Ex:21/10/1950
CO_Data_obito	Referência	Data de óbito do falecido. Ex:10/11/2014
CO_Nome	*Master*	Nome do falecido. Ex: João Silveira Santos
CO_Nome_Mae	*Master*	Nome da mãe do falecido. Ex: Joana Silveira Santos

Campo	Categoria	Descrição
CO_UF_nasc	Referência	Estado de naturalidade do falecido. Ex: PB,AC,MA
IDENTIDADE	*Master*	Número de identidade do falecido. Ex: 892341
CO_Estado	Referência	Estado do falecido. Ex: PB,AC,MA
CO_Municipio	Referência	Municipio do falecido. Ex: Rio Preto, São João.

A classificação da categoria a qual os dados pertencem possibilita a estratégia de técnicas de implementação de qualidade a partir do tipo de dado (Tabela 2).

Ainda, percebe-se de acordo com a Figura 22, que o campo identidade encontra-se fora do padrão utilizado para o sistema. Isto é ocasionado, muitas vezes, por diferentes programadores ao manipular uma tabela sem uma Regra de Negócio. Sua utilização fora de uma padronização pode comprometer os dados numa atualização futura, bem como, dificultar a implementação e manutenção de sistemas que utilizam este campo. Por isso, esta falha torna-se uma necessidade e deverá ser atualizada na Figura 19 e documentada sua revisão, como pode ser conferido no modelo final do escopo de estratégia do Apêndice A.

Ao seguir o fluxo de atividades da Etapa 2, o próximo passo objetiva a especificação do dado a partir do repositório. Para isto, será utilizado, adaptado e preenchido o modelo de especificação do dado, conforme Figura 23.

Figura 23. Modelo de especificação de dados do SISOBI

Repositório:	BD_Obitos			
Data:	07/10/2014			
Responsável:	<responsável pela especificação>			
Padrão de dado:				
Existência de padronização?	Não.			
Campo	**Tipo de Chave**	**Nula**	**Operações**	**Relacionamento**

Campo	Tipo de Chave	Nula	Operações	Relacionamento
CO_CPF	Primária	Não	Inserção;Atualização;Seleção	
CO_Data_nasc		Não	Inserção;Atualização;Remoção,Seleção	
CO_Data_obito		Não	Inserção;Atualização;Seleção	
CO_Nome		Não	Inserção;Atualização;Remoção,Seleção	
CO_Nome_Mae		Não	Inserção;Atualização;Remoção,Seleção	
CO_UF_nasc		Sim	Inserção;Atualização;Remoção,Seleção	
IDENTIDADE		Não	Inserção;Atualização;Remoção,Seleção	
CO_Estado		Não	Inserção;Atualização;Remoção,Seleção	
CO_Municipio		Sim	Inserção;Atualização;Remoção,Seleção	
Verificação de qualidade: Ausência de características:		Intríseca: precisão, credibilidade Contextual: valor agregado, completude Acessibilidade: segurança		

A partir da finalização deste documento, têm-se reunidas as informações necessárias para a próxima etapa. A verificação de qualidade, realizada conforme Figuras 10 e 11, proporciona, a partir de sua ausência de características, a diretriz na escolha de soluções às necessidades.

4.3 Aplicação da Etapa 3: desenvolvimento e ações

Nesta etapa são utilizadas as informações colhidas na aplicação da Etapa 1 e 2. A partir do modelo da Figura 23, tem-se a ausência das seguintes características que compõem um dado de qualidade:

- Intríseca: precisão, credibilidade
- Contextual: valor agregado, completude
- Acessibilidade: segurança

Será realizada então, a análise das técnicas apresentadas no item 3.4 e sua adaptação ao problema de maneira que representem a melhor solução.

A ausência de Regras de Negócio pôde ser notada desde a aplicação da Etapa 1. É necessária a criação e aplicação destas a todos os campos utilizados para o cadastro. Apesar do CPF ser um bom parâmetro único para o cruzamento de existência de cadastros, este campo pode ser desconhecido no Sistema único de Benefícios, então outros parâmentros deverão ser utilizados no SISOBI. Por isso, a criação das Regras de Negócio a todos os campos são demonstradas na Tabela 5:

Tabela 5: Regras de Negócio para o SISOBI

Regra	Campo	Descrição
RN01	CO_Nome	sequência de caracteres. No campo Nome do falecido, se o nome for desconhecido, deve ser utilizada unicamente a expressão “Desconhecido”. Expressões como: “Não identificado”, “Homem”, “mulher”, entre outros, reproduzem inconsistências no Sistema.
RN02	CO_Nome	o SISOBI antes de validar o cadastro, deverá verificar se o campo está vazio. Se estiver, o sistema retorna ao cadastro e avisa ao usuário que este deve digitar “Desconhecido”
RN03	CO_Nome_Mae	sequência de caracteres . No campo Nome da mãe do falecido, se o nome for desconhecido, deve ser utilizada unicamente a expressão

		Desconhecido. Expressões como: “Falecida”, “Ambos Falecidos”, “Não identificada”, entre outros, reproduzem inconsistências no Sistema.
RN04	CO_Nome_Mae	o SISOBI antes de validar o cadastro, deverá verificar se o campo está vazio. Se estiver, o sistema retorna ao cadastro e avisa ao usuário que este deve digitar digitar “Desconhecido”.
RN05	CO_Data_nasc	o campo data de nascimento do falecido deverá ser unicamente do formato numérico : dia mês ano. Ex: 24 05 1980.
RN06	CO_Data_nasc	o SISOBI deverá validar os campos de data seguindo a formatação da RN1_CO_Data_nasc.
RN07	CO_Data_nasc	anos menores que 1900 deverão retornar erro ao usuário.
RN08	CO_Data_obito	o campo data de óbito do falecido deverá ser unicamente do formato numérico : dia mês ano. Ex: 24 05 1980.
RN09	CO_Data_obito	o SISOBI deverá validar os campos de data seguindo a formatação da RN1_CO_Data_obito.
RN10	CO_Data_obito	datas anteriores a 24 abril 1986 não poderão ser aceitas pelo SISOBI.
RN11	CO_CPF	para o campo CPF apenas um campo numérico de 11 posições será aceito.
RN12	CO_CPF	O campo CPF deverá ser verificado no site da Receita Federal ou a partir do cálculo dos dois último dígitos verificadores.
RN13	CO_CPF	o SISOBI deverá validar o campo de acordo com a RN11 e RN12.
RN14	CO_Identidade	Campo numérico de 8 dígitos.
RN15	CO_Identidade	O SISOBI deverá validar o campo de acordo com a RN15.
RN16	CO_Estado	sigla da Unidade Federativa. Deverá conter um dos seguintes valores, em letras maiúsculas: AC, AL, AM, AP, BA, CE, DF, ES, GO, MA, MG, MS, MT, PA, PB, PE, PI, PR, RJ, RN, RO, RR, RS, SC, SE, SP e TO.
RN17	CO_Estado	o SISOBI deverá oferecer ao usuário uma lista com os valores da RN16.

RN18	CO_Municipio	deverá conter um conjunto de caracteres compondo municípios. a partir da escolha de estado.
RN19	CO_Municipio	o SISOBI deverá oferecer ao usuário uma lista com os valores válidos da RN18.
RN20	Todos	atualização e remoção da tabela deverá ser permitida apenas ao administrador(es) do sistema.
RN21	Todos	o nome da tabela deverá seguir a formatação CO_Nome1_nome2_nome3.
RN22	Todos	novas funcionalidades deverão respeitar as Regras de Negócio e verificadas em relação ao Sistema Único de Benefícios.

O critério na criação das Regras de Negócio incluiu técnicas de higienização de dados (item 3.4.2) que melhor definem uma estratégia para diminuir consideravelmente as falhas percebidas na Tabela 4.

Para a correção dos dados que se encontram com erros, deve-se realizar o tratamento a partir da comparação de entradas do SISOBI em relação ao Sistema de Registro Civil (SRC). O SRC apresenta uma saída consistente e foi implementado seguindo técnicas de qualidade de dados. Logo, para os cadastros inconsistentes, avalia-se qual é o campo do erro e utiliza-se outro na verificação. Por exemplo, se o campo CPF estava com apenas 5 números (no que o correto seriam 11), será utilizado o campo Identidade e Nome para o cruzamento de informações com o SRC. Se existir a igualdade, o cadastro será replicado, obedecendo os campos do SISOBI. Esta técnica, a comparação, deve ser utilizada também para referência na excluisão de registros duplicados.

4.4 Aplicação da Etapa 4: suporte pós produção

O cadastro do SISOBI é realizado por funcionários de cartórios de registro civis. A implementação das regras de negócio visou diminuir e excluir a dependência que o sistema tinha do usuário para armazenar dados de qualidade. Porém, para garantir maior qualidade ao processo é importante que os usuários sejam treinados no que se

refere às modificações no sistema, principalmente o modelo de entrada de dados. Para sistemas que necessitam de precisão nas entradas de dados é necessário que o funcionário realize o cadastro mantendo a exatidão dos dados informados. Uma das maneiras que possibilita a exatidão é a obtenção de dados a partir de um documento de referência, como por exemplo o Registro Geral ou Identidade. Nesse caso, elimina-se problemas de comunicação e entendimento entre o requerente e o funcionário.

Após a implementação de qualidade aos dados, inclusive o treinamento, deve ser realizada a verificação da qualidade do sistema. Uma aferição estatística deve ser realizada no período de 1 mês seguindo o modelo apresentado na Tabela 4. Se o percentual de falhas ainda for grande é necessário retornar à Etapa 1 e seguir novamente o fluxograma da Etapa 2 apresentado na Figura 6. Caso contrário deverá ser entregue a documentação final do projeto. A documentação é essencial devido à alta rotatividade de funcionários na organização – à medida que fornece também diretrizes que devem ser seguidas na manutenção ou escalabilidade do Sistema de Controle de Óbitos.

Capítulo 5

Considerações finais

Este capítulo descreve as considerações finais do trabalho, assim como apresenta a lista de trabalhos futuros a serem realizados. As conclusões deste trabalho foram obtidas a partir do confronto das falhas encontradas no SISOBI com a aplicação do processo proposto nesta monografia.

5.1 Conclusões

Todo projeto que envolva dados deve estabelecer alguns compromissos de qualidade para que sua implantação se torne um caso de sucesso. Para isso, deve-se utilizar metodologias que incrementem um nível de qualidade no gerenciamento e na entrega dos dados, mas também, a definição e acompanhamento da aplicação de técnicas de qualidade durante a continuidade do sistema alvo, ou seja, após a aplicação do processo desenvolvido neste trabalho.

O processo desenvolvido busca abranger qualquer organização na orientação da implementação de qualidade. Mas algumas limitações podem ser encontradas na adaptação das técnicas, ferramentas e instruções do processo à medida que precisam ser adequadas ao problema. Ou seja, requerem um conhecimento prévio e experiência em seu ajuste. Ainda, é importante destacar que embora as dificuldades inseridas durante o processo de implementação – informações privadas, tempo requerido e a necessidade de um esforço comum de várias pessoas envolvidas no projeto – a qualidade de dados é crucial para o sucesso de uma organização, e esta característica é mais do que suficiente para justificar o uso de técnicas que implementem esta qualidade à medida que ela provê eficiência. Dessa forma, a implementação do processo foi dada de maneira a seguir a literatura conhecida procurando ressaltar os pontos cruciais no sucesso desta prática.

Após a aplicação do processo para a qualificação dos dados do SISOBI, efetuando a padronização dos dados e a criação de regras de negócio com a finalidade de torná-lo um sistema integrável e de confiança, foi possível atingir os

objetivos propostos. O uso da metodologia TDQM foi essencial ao ajuste dos dados a partir do embasamento teórico sobre as exigências às quais atende um dado de qualidade.

Portanto, o processo poderá ser utilizado por profissionais de gerência de dados – ou ainda, qualquer profissional administrador de dados, que pretende desenvolver a vantagem competitiva em sua organização de modo a executar os processos de negócios com competência a partir de dados com qualidade.

5.2 Trabalhos futuros

O uso do processo em um exemplo real apresentado nesta monografia teve seu foco na aplicação de qualidade de dados em um sistema já existente. Com base nos resultados obtidos e no desempenho preliminar do processo, pode-se listar algumas melhorias em alguns trabalhos futuros:

- Desenvolvimento e aplicação do processo para sistemas em fase de concepção.
- Construção de uma ferramenta que realize a varredura dos dados informando os dados estatísticos do problema.
- Construção de uma ferramenta que implemente mineração de dados na correção de problemas.
- Aplicação do processo em outras organizações de ramos de negócio, tamanho e características diferentes do que foi apresentado nesta monografia.

Referências

[1] BALLOU, D. P.; TAYI, G. K. Enhancing Data Quality in Data Warehouse Enviroments. Communications of the ACM, New York, v. 42, n. 1, 1999.

[2] BALLOU, D. P.; TAYI, K. G. Methodology for allocating resources for. Communications of the ACM, New York, v. 32, n. 3, p. 320-329, 1989.

[3] BATINI, C., S. M. Data Quality: Concepts, Methodologies and Techniques. Berlim: Springer, v. 1, 2006.

[4] BELL, J. Re-Engineering Case Study – Analysis of Business Rules and Recomendations for Treatment of Rules in a Relational Database Environment. US West Information Technologies Group. [S.l.]. 1990.

[5] BEUREN, I. M. Gerenciamento da informação: um recurso estratégico no processo de gestão empresarial. 2. ed. São Paulo: Atlas, 200.

[6] CAMPELLO, A. V. C. MeGIQ - Metodologia de Geração de Informações de Qualidade para a Tomada de Decisão Executiva. Recife: Universidade Federal de Pernambuco, 2007.

[7] DataCare. Disponivel em: <http://www.assesso.com.br/qualidade-de-dados/datacare>. Acesso em: 11 novembro 2014.

[8] DATAPREV - Qualidade de dados. Disponivel em: <http://portal.dataprev.gov.br/wp-content/uploads/2009/12/QUALIDADEdeDADOS.pdf>. Acesso em: 11 novembro 2014.

[9] DAVENPORT, T. H. . P. L. Conhecimento empresarial. Rio de Janeiro: Campus, 1998.

[10] DEMING, E. W. Out of the Crisis. 1. ed. Massachussets: MIT, 1986.

[11] DOBYNS, L. . C.-M. C. Quality or Else: The Revolution. Boston: Houehton Mifflin, 1991.

[12] DRESCHER, S. Inteligência O que você sabe sobre seus dados? Venda Mais, 2004. Disponivel em: <http://www.vendamais.com.br/php/materia.php?id=36070>. Acesso em: 20 outubro 2014.

[13] ECKERSON, W. W. Data Quality and The Botttom Line: Achieving Business Success through a Commitment to High Quality Data. The Data Warehousing Institute Report Series. Disponivel em: <http://www.stuart.iit.edu/courses/im510/spring2002/dqreport.pdf>. Acesso em: 12 setembro 2014.

[14] ENGLISH, L. P. Information quality for business intelligence and data mining: assuring quality for strategic information uses, 2005. Disponivel em: <http://support.sas.com/news/users/LarryEnglish_0206.pdf>. Acesso em: 12 setembro 2014.

[15] GOLDSCHMIDT, R. . P. E. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. 1. ed. São Paulo: Elsevier, v. 1, 2005.

[16] HAUG, A. . F.; LIEMPD, D. The costs of poor data quality. Journal of Industrial Enginnering and Management, Denmark, v. 4(2), p. 168-193, 2011.

[17] INTERNATIONAL Association for Information and Data Quality. Disponivel em: <http://www.iaidq.org/main/>. Acesso em: 20 outubro 2014.

[18] IVANOV, K. Quality-control of information: On the concept of accuracy of information in data banks and in management information systems. The University of Stockholm and The Royal Institute of Technology. Stockholm. 1972.

[19] KOTONYA, G. . S. I. Requirements engineering: processes and techniques. England: Chichester, 1998.

[20] LAUDON, K. C. . L. J. P. Sistemas de informação gerenciais. 7. ed. São Paulo: Pearson Prentince Hall, v. 1, 2007.

[21] LOSHIN, D. Considering a Services Approach for Data Quality. Disponivel em: <http://www.pbinsight.com/files/resource-library/resource-files/serv-approach-for-dq-us-0910.pdf>. Acesso em: 12 setembro 2014.

[22] MADNICK, S. . W. R. Y. Introduction to Total Data Quality Management (TDQM) Research Program. MIT Sloan School of Management, Massachussets, 1992.

[23] MATOS, G. . C. R. . C. O. Metodología para la extracción del conocimiento empresarial a partir de los datos. [S.l.]: Inf Tecnol, v. 17(2), 2006.

[24] MATTIODA, R. A.; FAVARETTO, F. Qualidade da informação em duas empresas que utilizam Data Warehouse na perspectiva do consumidor de informação – um estudo de caso. Gestão & Produção, São Carlos, v. 16, n. 4, Dezembro 2009.

[25] MCGEE, J. V. . P. L. Gerenciamento Estratégico da Informação. 10. ed. Rio de Janeiro: Ernest & Young, 1994.

[26] MCGILVRAY, D. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. 1. ed. [S.l.]: Morgan Kaufmann, v. 1, 2008.

[27] MIT'S Total Data Quality Management (TDQM). Disponivel em: <http://web.mit.edu/tdqm/www/>. Acesso em: 12 set. 2014.

[28] NGUYEN, T. The Value of ETL and Data Quality., 2008. Disponivel em: <www2.sas.com/proceedings/sugi28/161-28.pdf>. Acesso em: 20 outubro 2014.

[29] Oracle Enterprise Data Quality. Disponivel em: <http://www.rittmanmead.com/2012/08/introducing-edq/>. Acesso em: 11 novembro 2014.

[30] Quality Center Enterprise. Disponivel em: <http://www8.hp.com/us/en/software-solutions/quality-center-quality-management/try-now.html?jumpid=reg_r1002_usen_c-001_title_r0001>. Acesso em: 11 novembro 2014.

[31] ROSS, R. G. Business Rule Concepts: Getting to the Point of Knowledge, 2009. Disponivel em: <http://www.brsolutions.com/b_concepts.php>. Acesso em: 20 outubro 2014.

[32] SANTOS, I. M. F. Uma proposta de governança de dados baseada em método de desenvolvimento de arquitetura empresarial. UFRJ, Rio de Janeiro, 2010.

[33] SETZER, V. W. Dado, Informação, Conhecimento e Competência. Disponivel em: <http://www.ime.usp.br/~vwsetzer/dado-info.html>. Acesso em: 20 outubro 2014.

[34] STOREY, V. C.; WANG, R. Y. Modeling Quality Requirements in Conceptual Database Design. TDQM Publications, Massachussets, October 1998.

[35] SUITE, B. Bizagi. Disponivel em: <http://www.bizagi.com/>. Acesso em: 07 novembro 2014.

[36] TAVARES, R. S. Bancos de Dados Qualificados Podem Reduzir Perdas e Aumentar os Ganhos em CRM. São Paulo: (Monografia(MBA)) , 2003.

[37] TURBAN, E. . . Tecnologia da informação para gestão. Transformado os negócios da economia digital. 3. ed. Porto Alegre: Bookman, 2004.

[38] TRIBUNAL DE CONTAS DA UNIÃO, TCU. Auditoria no Sistema Informatizado de Controle de Óbitos (Sisobi). Disponivel em: <http://portal2.tcu.gov.br/portal/pls/portal/docs/2057634.PDF>. Acesso em: 11 novembro 2014.

[39] VADUVA, A. Rule Development for Active Database System, Zurich, 1999.

[40] WANG, R. Y., S. D. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, v. 12, n. 4, p. 5-33, 1996.

[41] WATTS, S. G.; SHANKARANARAYANAN, A. E. Data quality assessment in context: a cognitive perspective. Decision Support Systems, v. 48, p. 202-211, 2009.

[42] WOO, R. 7 Facts About Data Quality [Infographic]. INSIGHTSQUARED, 2012. Disponivel em: <http://www.insightsquared.com/2012/01/7-facts-about-data-quality-infographic/>. Acesso em: 20 outubro 2014.

[43] ZANUSSO, M. Data Mining. Disponivel em: <http://www.dct.ufms.br/~mzanusso/Data_Mining.htm>. Acesso em: 20 outubro 2014.

[44] ZEISS, G. Estimating the economic and financial impact of poor data quality, 2013. Disponivel em: <http://geospatial.blogs.com/geospatial/2013/05/estimating-the-economic-impact-of-poor-data-quality.html>. Acesso em: 12 setembro 2014.

[45] ZHU, H., M. S. E., L. Y. W., W. R. Y. Data and Information Quality Research: Its Evolution and Future. 1. ed. [S.l.]: Taylor & Francis Group, v. 1, 2014.

[51] PALMER, J. D. Traceability. In: THAYER, R.; DORFMAN, M. (Ed.). Software Requirements Engineering. Los Alamitos, Ca, USA: IEEE Computer Society Press, 2000. p. 412–422.

Apêndice A

Modelo do escopo da estratégia após atualização

Este apêndice apresenta o modelo de documento do escopo da estratégia após a aplicação das etapas do processo na implementação de qualidade de dados no sistema SUB.

Nome do projeto:	ImQob – Implementação de qualidade no cadastro de óbitos
Data:	09/10/2014
Responsável:	<responsável pelo projeto>
Histórico de revisão	
Data:	25/11/2014
Responsável:	<responsável pela revisão>
Mudanças:	Estabelecimento de regras de negócio (Tabela 5)
Técnicas utilizadas:	Regras de negócio, higienização dos dados, replicação dos dados do de cadastro do SUB para os dados incompletos existentes no SISOBI.
Definição do projeto:	Implementação de qualidade de dados no Sistema de Controle de Óbitos.
Prioridade:	Alta
Consequências dos problemas atuais:	Custos desnecessários estimados em R$2,3 bilhões. Perda da credibilidade do sistema Perda da credibilidade da organização

<table>
<tr><td colspan="2">Benefícios esperados:</td><td colspan="2">Reverter as falhas encontradas; garantir integridade da informação.</td></tr>
<tr><td colspan="2">Escopo do projeto:</td><td colspan="2"></td></tr>
<tr><td>Objetivos</td><td>Área de atuação</td><td>Situação atual</td><td>Necessidade</td></tr>
<tr><td>Estabelecer padronização na entrada de dados</td><td>Cadastro, atualização.</td><td>Cadastro permite a entrada de dados a partir da inexistência de regras de negócio.</td><td>Implementar Regras de Negócio.</td></tr>
<tr><td colspan="2">Condições do projeto

Observada dependência com os sistemas:</td><td colspan="2">Cadastro Nacional de Informações Sociais

Sistema Único de Benefícios</td></tr>
<tr><td colspan="2">Treinamento:</td><td colspan="2">Essencial a apresentação das mudanças no sistema, a padronização de entradas e a validação dos dados de entrada por requerimento de documento(s) oficial(is).</td></tr>
</table>

Apêndice B

Relação de casos estudados – diagnóstico preliminar

Cheias em Pernambuco em 2010	As cheias que assolaram a mata sul de PE e o desabamento de diversas barreiras, devido às fortes chuvas que caíram no grande Recife no inverno, causaram a morte de dezenas de pessoas. A causa principal foram as fortes chuvas registradas, entretanto, muitas mortes poderiam ter sido evitadas com a divulgação das informações sobre as precipitações previstas. A descontinuidade dos dados (má qualidade de dados) fornecidos pela APAC - Agência Pernambucana de Águas e Chuvas, agência responsável pela previsão das chuvas no estado, contribui de forma significativa para a ocorrência das tragédias. Seria possível retirar as pessoas dos locais afetados, com antecedência, caso as informações fossem fornecidas em tempo hábil.
Dados equivocados em grandes empresas	A Companhia Pernambucana de Saneamento fez com que um cidadão não usuário de seus serviços tivesse seu nome negativado indevidamente junto aos órgãos de restrições cadastrais. O fato ocorreu devido aos dados incorretos da Compesa em relação ao mapeamento da região onde o não consumidor, de maneira autônoma, utilizava outros meios de saneamento e abastecimento de água independentes da Companhia. Ao generalizar a métrica de cobrança por região as informações cadastrais são induzidas ao erro e geram prejuízos jurídicos e financeiros à empresa e morais e materiais ao cidadão.
Consequências de dados desatualizados em órgãos públicos	Após auditoria feita pelo Tribunal de Contas da União no Instituto Nacional de Seguro Social foi estimado um prejuízo aos cofres públicos de aproximadamente R$ 2 bilhões de reais apenas devido aos pagamentos irregulares de benefícios. Cerca de 1 milhão de pessoas já falecidas ainda constavam como vivas no Sistema Informatizado de Controle de Óbitos (SISOBI). A causa apurada foi que os dados de óbito provenientes de cartórios não eram, ou demoravam a serem cadastrados no sistema do INSS.