Instituto Superior de Economia e Gestão
UNIVERSIDADE TÉCNICA DE LISBOA
DESDE 1911

MESTRADO

GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO

DISSERTAÇÃO

AVALIAR E MELHORAR A QUALIDADE DE DADOS COM IMPACTO NO NEGÓCIO NUM PROCESSO DE MIGRAÇÃO DE DADOS ENTRE ERPS

NELSON EDGAR MOÇO SOARES

OUTUBRO – 2015

Instituto Superior de Economia e Gestão
UNIVERSIDADE TÉCNICA DE LISBOA
DESDE 1911

MESTRADO EM
GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO
DISSERTAÇÃO

AVALIAR E MELHORAR A QUALIDADE DE DADOS COM IMPACTO NO NEGÓCIO NUM PROCESSO DE MIGRAÇÃO DE DADOS ENTRE ERPS

NELSON EDGAR MOÇO SOARES

ORIENTAÇÃO: PROFESSORA ENGENHEIRA ANA MARIA MARQUES RIBEIROS DOS SANTOS LUCAS

OUTUBRO – 2015

AGRADECIMENTOS

Não sendo este um trabalho individual, gostaria de reconhecer o apoio e empenhamento de algumas pessoas durante o período em que decorreu esta investigação, sem o qual não teria sido possível a sua a conclusão.

Em primeiro lugar, agradeço à Professora Engenheira Ana Lucas pelo apoio dado e pela sua disponibilidade em orientar a presente investigação.

Agradeço também à *RetailPC,* nas pessoas do Dima e do Pedro, por terem permitido a realização desta investigação, sendo seu o mérito de reconhecer a qualidade de dados como um fator crítico para o seu sucesso empresarial.

À Catarina Ortigão pela sua disponibilidade e pelos seus valiosos contritos numa fase crucial deste trabalho.

Aos meus amigos, por terem suportado as minhas ausências neste último ano.

Ao Pedro, à Anabela, à Maria e à Joana por servirem de fonte de inspiração, pela partilha de momentos e pelo carinho que recebo.

Aos meus «manos», por tudo o que representam para mim.

Aos meus pais, pelos sacrifícios que fizeram durante toda a sua vida em prol dos filhos e que hoje me permitem ser quem sou, por todos os princípios que me incutiram, por me fazerem acreditar que não importa de onde vens, mas para onde caminhas…

À *Puma* que é a maior!

A ti Catarina, porque este trabalho é dedicado só a ti.

"Somos o que repetidamente fazemos. A excelência, portanto, não é um feito, mas um hábito."

Aristóteles

LISTA DE SIGLAS E ACRÓNIMOS

API	*Application Programming Interface*
DB	*Database*
DP	*Data Profiling*
DS	*Data Staging*
ERP	*Enterprise Resource Planning*
ETL	*Extract, Transform Load*
IP	*Information Product*
IP-MAP	*Information Product Maps*
DC	*Data cleaning*
DM	*Data migration*
DQI	*Data Quality Improvement*
NIF	Número de Identificação Fiscal
OEDQ	*Oracle Enterprise Data Quality*
PBS	*Primavera Business Suite*
DQP	*Data Quality Problems*
PSP/IQ	*Product and Service Performance Model for Information Quality*
DQ	Data Quality
SBO	*SAP Business One*
IS	*Information System*
TDQM	*Total Data Quality Management*
XML	*eXtensible Markup Language*

RESUMO

Apesar de toda a literatura publicada sobre a melhoria da qualidade de dados, problemas de qualidade de dados continuam a afetar a operacionalidade das empresas e dos seus sistemas de decisão. Reconhecendo este facto, a *RetailPC,* uma empresa de comércio a retalho de equipamento informático, aceitou a realização da presente investigação, a qual teve como objetivo a avaliação e melhoria da qualidade de dados da sua entidade *Cliente*, durante um processo de migração de dados entre dois ERPs (Primavera Professional para SAP Business One).

Para o efeito, foi utilizada a metodologia *Action Research,* uma vez que permite ao investigador assumir um papel intervencionista na resolução do problema da qualidade de dados. No caso concreto deste trabalho, foi avaliada e melhorada a qualidade de dados durante a migração entre ERPs e alterados processos de recolha dos mesmos, tendo sido disponibilizados meios de diagnóstico para futuros ciclos de *Action Research*.

No final, foi possível constatar que a qualidade de dados foi melhorada significativamente. Foi possível corrigir todos os erros detetados nos atributos *ShipType* (Modo de expedição)*, PymCode* (Formas de pagamento)*, Currency* (Moeda) *e LangCode* (Língua da documentação enviada para o cliente); 98,53% dos erros detetados em sujeitos passivos coletivos com respeito ao atributo *LicTradNum* (NIF); 56,67% das moradas com erros do atributo *ZipCode* (Código Postal) e 99,65% dos tuplos que continham valores no atributo *IntrntSite*, uma vez estava a ser utilizado para um fim diferente do previsto pelo ERP, tendo esses valores sido migrados para o atributo *E_Mail* para posterior tratamento. Foram ainda detetados e eliminados 323 tuplos de entidades que se encontravam duplicados.

Palavras-chave: Qualidade de dados, Migração de dados, SAP Business One, TDQM, Dimensões de qualidade de dados, Problemas de qualidade de dados

ABSTRACT

Despite all the published literature on data quality enhancement, data quality problems continue to affect the company's operation and their decision systems. In recognition of that, *RetailPC*, an IT equipment retail trading company, accepted to be part of the present research, which aimed the assess and improve data quality of its *Customer* entity during a data migration process between ERPs (Primavera Professional to SAP Business One).

For this purpose, it was used the Action Research methodology, as it allows the researcher to assume an interventional role in the resolution of the data quality problem. In the specific case of this research, has been assessed and enhanced data quality during data migration and changed data collection processes, were made available diagnostic methods for future cycles of Action Research.

At the end it was perceived that the quality of data is improved significantly. It was possible to correct all the errors detected in attributes *ShipType* (Delivery mode), *PymCode* (Payment Methods), *Currency* (Currency) and *LangCode* (Language of documents sent to customer); correct 98.53% of detected errors in collective taxpayers with respect to *LicTradNum* attribute (Tax ID); 56.67% of addresses with errors in *ZipCode* attribute (Postal Code) and 99.65% of tuples that contain values in *IntrntSite* attribute, as was being used for a different purpose from that defined by the ERP, and these values were migrated to *E_Mail* attribute for further processing. They were also detected and eliminated 323 tuples of entities that were duplicated.

Keywords: Data quality, Data migration, SAP Business One, TDQM, Data quality dimensions, Data quality problems

ÍNDICE

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

1 INTRODUÇÃO

A modernização de sistemas de gestão constitui um dos grandes desafios em sistemas de informação (IS) sendo, por vezes, necessário proceder à sua substituição em resultado do surgimento de novas tecnologias ou de imposições legais ou até, devido à evolução do domínio das aplicações (Haller et al., 2011). Com a substituição de ERPs (do inglês *Enterprise Resource Planning*) obtêm-se funcionalidades melhoradas ou IS mais adequados aos processos de negócio.

A migração de dados (DM) tem um papel fundamental na substituição de sistemas antigos por novos ERPs. É uma tarefa complexa e, de acordo com Thalheim & Wang (2013), pode tornar-se na principal causa de insucesso de um projeto.

Existem várias situações que podem aumentar a complexidade do processo de DM. Em primeiro lugar, um sistema antigo pode ter diversas fontes de dados que utilizam diferentes ferramentas de modelação, ou com semânticas distintas, o que requer um conhecimento profundo sobre os dados existentes no que respeita às restrições e as interligações entre as várias fontes de dados. Em segundo lugar, as fontes de dados do sistema antigo podem estar incorretas, incompletas, duplicadas ou inconsistentes e ainda, pode ser exigido que os novos sistemas necessitem que os dados migrados cumpram determinados requisitos. Em consequência, colocar os dados com a qualidade necessária para serem migrados pode significar elevados custos, em tempo e em dinheiro, para uma empresa. Por último, as principais tarefas de DM precisam de ser executadas interactivamente e, frequentemente, as especificações dessa migração podem ter de ser alteradas para solucionar os problemas detetados. Em suma o fundamental é que, durante o processo de DM, seja garantida a qualidade de dados do sistema de destino (Manjunath & Hegadi, 2013).

Problemas com a qualidade de dados (DQP) serão sempre questões por responder. Investigações da década de 1980 (Ballou & Pazer, 1985), bem como da década de 1990, já após o início do desenvolvimento de uma *framework* de análise de qualidade de dados (DQ) (Strong et al., 1997; Wang et al., 1995b), não conseguiram despertar as empresas para a sua real importância, apesar dos profissionais em IS já reconhecerem o impacto económico e organizacional de uma baixa DQ (Redman, 1995).

Para que as empresas possam obter melhores resultados dos seus IS é necessária uma elevada DQ (Wang et al., 1995b), sendo esta um fator crítico nas organizações (Lee et al., 2002).

Cao & Zhu (2013) demonstraram que mesmo empresas com ERPs implementados podem ter DQP. Por sua vez, Baškarada & Koronios (2014) argumentam que as organizações dependem da DQ para as suas operações do dia-a-dia. Já Ballou & Tayi (1989), defendem que os IS só podem ser eficazes se os dados necessários possuírem um nível de integridade compatível com os requisitos de processamento e de utilizador.

Tendo como suporte a literatura existente nas áreas de DQ e DM, definiu-se o objetivo desta investigação: melhorar a qualidade de dados da entidade *Cliente* da empresa *RetailPC* (nome fictício), no âmbito de um processo de DM entre ERPs. Para o efeito, foi utilizada a metodologia *Action Research*, a qual foi utilizada numa área aplicacional definida como de maior importância para o objeto em estudo. Esta metodologia mostrou-se adequada porque envolve a resolução de problemas organizacionais através da intervenção dos investigadores e, ao mesmo tempo, contribui para o conhecimento existente (Davison et al., 2012).

Quanto à sua estrutura, o presente relatório é composto por cinco capítulos, sendo o primeiro deles destinado à introdução. No segundo capítulo é justificada a relevância do estudo, atendendo ao referido na literatura sobre DM e DQ. No capítulo terceiro, é explicada a metodologia e a sua adequação ao trabalho desenvolvido. Por sua vez, no

capítulo quarto são expostos os resultados da intervenção efetuada na *RetailPC*. Por último, no capítulo quinto são discutidas as principais conclusões, as limitações da dissertação e sugerem-se temas para futuras investigações.

2 REVISÃO DA LITERATURA

No que se refere a DQ, grande parte dos estudos anteriores abordaram o tema sem se preocupar com o contexto em que os dados são criados ou utilizados, dando ênfase às características intrínsecas o que, só por si, não permite que se resolvam problemas organizacionais que dependam da DQ. No entanto, no final da década de 90, a comunidade científica despertou para a relevância das características contextuais dos dados (Strong et al., 1997; Wang et al., 1995a; Wang & Strong, 1996).

Wang et al. (1995b) efetuaram uma pesquisa minuciosa à literatura existente na altura e propuseram uma *framework* para organizar e servir de guia a futuras investigações. Abordaram as dimensões da DQ propondo uma extensão ao modelo relacional de Codd (1970, 1990) que permitiria etiquetar os dados com indicadores de qualidade, realçando duas dimensões: a interpretabilidade e a credibilidade. Apesar de esta *framework* ser compreensiva, Madnick et al. (2009) evidenciaram a falta de alguns termos para simplificar a caracterização da DQ, não sendo, por isso, fácil de utilizar. A incorporação destes metadados seria uma tarefa que consumiria muito tempo e seria muito dispendiosa, sendo importante conhecer as caraterísticas dos potenciais utilizadores que poderiam beneficiar com a sua utilização (Fisher et al., 2003). No que se refere a definições de DQ, a literatura é bastante extensa. Sucintamente, algumas dessas definições são apresentadas na Tabela I, neste trabalho será utilizada a definição dada por Strong (1997, p.104).

TABELA I: DEFINIÇÕES DE QUALIDADE DE DADOS

Autor	Definição
(Strong et al., 1997, p.104)	*"We define high-quality data as data that is fit for use by data consumers (…)This means that usefulness and usability are important aspects of quality"*
(Wand & Wang, 1996, p.91)	*"A real-world system is said to be properly represented if: (1) there exists an exhaustive mapping, Rep: RWL→ ISL, and (2) no two states in RWL are mapped into the same state in ISL (the inverse mapping is a function)."*
(Lee et al., 2004, p.88)	*"(…) we define data quality as that are fit for use (…)"*
(Ballou & Pazer, 1985, p.151)	*"(…) data quality is a relative rather than an absolute term and can most usefully be defined in the context of end use (…)"*
(Wang & Strong, 1996, p.22)	*"(…) high-quality data should be intrinsically good, contextually appropriate for the task, clearly represented, and accessible to the data consumer"*
(Kahn et al., 2002, p.184)	*"(…) quality as conformance to specification and as exceeding consumer expectations"*
(Manjunath & Hegadi, 2013, p.101)	*"Data Quality is the measure of accuracy of data which meets the business requirements and supports to the decision makings."*

2.1 DIMENSÕES DE QUALIDADE DE DADOS

No que concerne à definição das dimensões de DQ existentes, bem como à sua classificação não existe um consenso alargado. Wang e Strong (1996, p.6) definiram como dimensões de DQ «um conjunto de atributos de DQ que representam um único espeto ou um constructo da DQ», e classificaram estas dimensões em quatro categorias: intrínseca, acessibilidade, contextual e representacional. As 16 dimensões de DQ que Strong et al. (1997) apresentaram são classificadas em quatro categorias, as quais podem ser consultadas no Anexo I.

Kahn et al.(2002) melhoraram a classificação de Strong et al. (1997) ao incluírem características de entrega de serviço. Advogam que, tanto o produto, como a entrega de serviço, são aspetos importantes da DQ, enquanto a visão anterior (Wang & Strong, 1996; Strong et al., 1997) era preferencialmente orientada para o produto. Elaboraram uma matriz em que reorganizaram as dimensões de DQ apresentadas anteriormente no

modelo *Product and Service Performance Model for Information Quality* (PSP/IQ), conforme é mostrado na Tabela II.

TABELA II: MODELO PSP/IQ

	Conforme especificação	Satisfaz expectativas dos utilizadores
Qualidade do produto	Informação robusta • Exatidão • Representação concisa • Completude • Representação consistente	Informação útil • Quantidade apropriada • Relevância • Compreensibilidade • Interpretabilidade • Objetividade
Qualidade do serviço	Informação segura • Oportunidade • Segurança	Informação utilizável • Credibilidade • Acessibilidade • Operacionalidade • Reputação • Valor acrescentado

(Kahn et al., 2002, p.188)

2.2 GESTÃO TOTAL DA QUALIDADE DE DADOS

Total Data Quality Management (TDQM) é uma extensão da gestão de qualidade total (TQM) que identifica quatro fases: Definir, Avaliar, Analisar e Melhorar, as quais apresentadas na Tabela III. No final da última fase, o ciclo de TDQM repete-se, pois as atividades de melhoria de qualidade de dados (DQI) podem voltar a ser executadas (Lee et al., 2004).

TABELA III: FASES DO CICLO TDQM

Fase	Descrição
Definir	Identificar as dimensões de DQ e dos requisitos que devem satisfazer para os consumidores de dados. Tipicamente um ciclo de DQI restringe-se a um subconjunto de dimensões.
Avaliar	Criar métricas para que seja possível avaliar a DQ nas dimensões escolhidas na fase anterior e de modo a respeitar os requisitos definidos.
Analisar	Analisar os valores obtidos na fase anterior e decidir como melhorar a DQ, investigando as causas da baixa DQ e os relacionamentos entre as mesmas.
Melhorar	Decidir como devem ser corrigidos os dados e modificar os processos que originam dados de baixa qualidade.

(Wang, 1998)

Lee (2004) propôs ligar a integridade de dados aos processos de negócio, estendendo o modelo relacional (Codd, 1970, 1990) através da adição de metadados, para registo de informação respeitante à DQ, em consequência, suportar a natureza dinâmica e global dos dados organizacionais.

2.3 GESTÃO DE QUALIDADE DE DADOS COMO UM PRODUTO DE INFORMAÇÃO (IP)

As organizações frequentemente veem os dados como se de um subproduto se tratasse. Nesse âmbito, e para tratar os dados com a importância devida, será necessário seguir quatro princípios: (i) entender as necessidades dos utilizadores; (ii) gerir os dados como produto de um processo bem definido; (iii) gerir os dados como um produto com o seu ciclo de vida e (iv) definir um gestor de DQ responsável pela gestão do processo e do seu resultado final (Wang et al., 1998).

O processo de transformação de dados brutos em informação pode ser visto como um processo industrial (Tabela IV), apesar de não ser possível atribuir algumas características diretamente aos processos de fabrico (Wang, 1998), em virtude das especificidades de cada um dos sistemas. Por exemplo, as dimensões credibilidade e oportunidade respeitam somente a dados, enquanto o consumo de matérias-primas apenas está relacionado com os produtos.

TABELA IV: ANALOGIA ENTRE A FABRICAÇÃO DE PRODUTOS E FABRICAÇÃO DE DADOS

	Fabricação de produtos	Fabricação de dados
Entrada	Matéria-prima	Dados em bruto
Processo	Processamento de materiais	Processamento de dados
Saída	Produto físico	Produtos de dados

(Wang, 1998, p.59)

Shankaranarayanan et al. (2000) apresentam um constructo para modelar o processo de produção de informação. O *Information Product Maps* (IP-MAP) é uma extensão do sistema de produção de informação (Ballou et al., 1998). Este constructo oferece um conjunto de vantagens, entre as quais: possibilidade de visualizar o processo de fabrico

dos dados; implementação de melhoria contínua dos dados na sua fonte e a medição da DQ utilizando as dimensões de DQ adequadas. De forma que seja possível capturar e representar todos os passos do processo de criação do IP, e registar toda a informação associada a cada um desses passos, o modelo suporta blocos de DQ e um repositório para capturar os metadados associados a cada bloco. Os constructos e os símbolos estão representados na Figura 1, enquanto na Figura 2 se apresenta um exemplo da sua aplicação.

SYMBOL	REPRESENTS
<Name>	**Data Source / Data Vendor / Point-of-Origin**
< Process Identifier>	**Process**
<Storage Name>	**Data / Information Storage**
<Criteria>	**Decision**
Quality Criteria	**Quality / Evaluation / Check**
< Current / New System Names >	**Information System Boundary** - used when a data unit (raw data, component data) changes from one system (paper or computerized) to another (paper or computerized)
<Current / New Org/dept names>	**Organizational Boundary** - used when a data unit (raw, component) moves across departments or across organizations
<Name>	**Data Sink / Consumer Block / Point-of-Destination.**

Figura 1: Constructos e símbolos utilizados na sua representação
(Fonte: Shankaranarayanan et al. (2000), p 7 e 8)

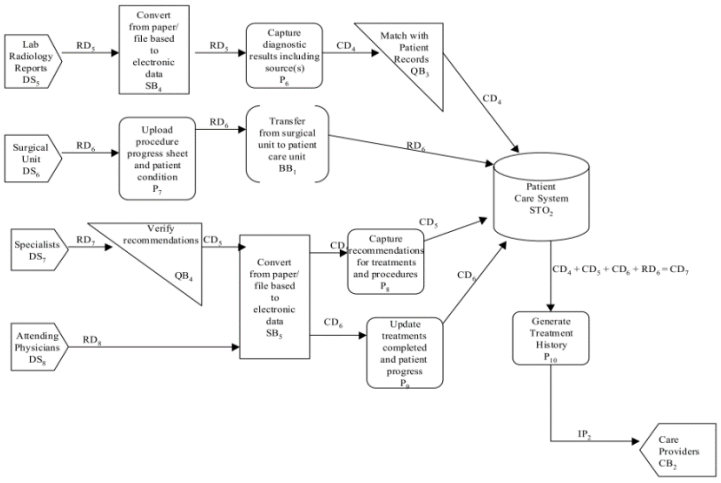


Figura 2: Exemplo de aplicação de IP-MAP
(Fonte: Shankaranarayanan (2000) , p.14)

2.4 TAXONOMIA DOS PROBLEMAS DE QUALIDADE DE DADOS

Muitas empresas começam a reconhecer os dados que têm ao seu dispor como um ativo valioso na estratégia da empresa. Apesar disso, continuam a não prestar a necessária

atenção à existência de DQP, nem a aplicar metodologias que assegurem uma elevada DQ, «como tal, diz-se que existem DQP se o utilizador ou a aplicação fica com um resultado incorreto ou não consegue obter um resultado devido a problemas inerentes aos dados» (Kim et al., 2003, p.82).

A taxionomia definida por Oliveira & Rodrigues (2005) foi formada agregando os DQP de trabalhos anteriores (Kim et al., 2003; Müller & Freytag, 2003; Rahm & Do, 2000). Estes problemas estão estruturados em 6 níveis de granularidade, do problema mais simples (um único atributo e num único tuplo), ao problema mais complexo (relacionados com múltiplas fontes de dados), conforme se enumera no Anexo II.

2.5 MELHORIA DA QUALIDADE DE DADOS ORIENTADA A DADOS

Não existe uma definição unânime e abrangente na literatura sobre DQI, a qual pode variar consoante a área onde se pretende aplicar. As principais áreas que incluem definições de DQI nos seus processos são, por um lado, *Data Warehouse* e *Data Mining* e, por outro, TDQM e IP-MAP. No âmbito desta dissertação recorrer-se-á a estes últimos para melhorar a DQ na empresa *RetailPC*, razão pela qual já foram apresentados no início deste capítulo, em 2.2.e 2.3.

2.5.1 Melhoria da qualidade de dados (DQI)

Uma intervenção reativa de DQI incide sobre os dados existentes numa determinada base de dados (DB) e envolve o desenvolvimento de um conjunto de atividades para identificar e corrigir os DQP. A identificação dos DQP é a primeira atividade a ser executada, recorrendo a ferramentas de *profiling* de dados (DP). As atividades seguintes têm como objetivo a correção dos DQP identificados através do DP (Oliveira, 2008).

2.5.1.1 Migração de dados

Na literatura existente é relativamente consensual a definição de DM. Na Tabela V reúnem-se algumas dessas definições e respetivos autores. Para o presente trabalho irá ser utilizada a definição de Agrawal (2008, p.1277).

Haller et al. (2011) apresentam uma arquitetura genérica de DM composta por 3 DB intermédias: (i) DB de origem temporária, para que o processo de DM não danifique a DB original e não torne a aplicação demasiado lenta; (ii) DB de transformação, para guardar os resultados intermédios de migração de dados e (iii) DB de destino temporária, para guardar os dados prontos para serem inseridos na DB de destino.

TABELA V: DEFINIÇÕES DE MIGRAÇÃO DE DADOS

Autor	Definição
(Thalheim & Wang, 2013, p.1)	*"In general data migration is the process of moving data from legacy data sources of a legacy system into new data sources of a target system, in which legacy and new systems have different data structures."*
(Agrawal et al., 2008, p.1277)	*"Data Migration refers to the process of transforming data from one or more old (possibly legacy) systems to a newer system providing new service capabilities and efficiency improvement."*
(Haller et al., 2011, p.1)	*"(…) data migration is understood as a tool- supported one-time process which aims at migrating formatted data from a source structure to a target data structure whereas both structures differ on a conceptual and/or technical level."*

No estado final (fase (iii) da arquitetura genérica) deve ser definido o nível mínimo de DQ exigido pelo sistema de destino, bem como o conjunto de testes de controlo que serão utilizados para efetuar a verificação da DQ. Caso estas condições não se verifiquem, segundo Haller et al. (2011), é possível que se incorra nos seguintes custos e/ou dificuldades associados a: (i) deteção de erros; (ii) duplicação de tarefas; (iii) prevenção de erros; (iv) atraso nas operações; (v) custos respeitantes a atrasos em processamentos; (vi) tomada de decisões incorretas; (vii) inconsistência global dos dados na empresa.

A transformação de dados é uma fase importante da DM. É um processo complexo devido a vários fatores: (i) diferenças existentes entre os esquemas de origem e de destino; (ii) dados corrompidos ou inconsistentes; (iv) dados que têm de ser conciliados entre várias fontes. Por exemplo, um sistema pode ter a morada num único atributo, enquanto o novo sistema pode requerer vários atributos para guardar a mesma informação (rua, nº de porta, código postal, localidade, país). Agrawal et al.(2008) propõem uma *framework* que estenda o processo de mapeamento de atributos entre o antigo e o novo esquema através de conhecimento de domínio. O esquema de destino define o objetivo a atingir, já que identifica os dados exigidos para que fique funcional.

2.5.1.2 Data Profiling

O *Data Profiling* (DP) corresponde a um conjunto de técnicas que servem para analisar dados, permitindo obter informação sobre os metadados, o tipo de dados, a completude e a unicidade dos atributos, as chaves primárias, as chaves estrangeiras e os valores padrão. As análises mais difíceis de efetuar são as que envolvem mais do que um atributo, como é o caso das dependências funcionais. A utilização do DP justifica-se pela necessidade de preparar os dados para tarefas subsequentes como a otimização de consultas, a limpeza de dados (DC), a integração de dados e a análise de dados.

Pelas suas características, Naumann (2014) considera que o DP é um processo complexo por 3 motivos: (i) os resultados são computacionalmente complexos; (ii) a característica de descoberta do *profiling* exige uma verificação complexa de todos os atributos e combinações de atributos e (iii) são executados em grandes quantidades de dados. Segundo Olsen (2003), as principais fases do DP são: (i) análise das propriedades do atributo; (ii) análise da estrutura e (ii) análise das restrições.

2.5.1.3 Limpeza de dados

Os autores são relativamente consensuais acerca da definição de limpeza de dados (DC), sendo vista como um processo que tem como objetivo a deteção e correção de DQP (Milano, 2005). Müller & Freytag (2003) referem-se à DC como sendo um processo recursivo que é executado num conjunto de dados, defendendo que é composto por 4 atividades: (i) DP – o qual já foi desenvolvido em 2.5.1.2 do presente trabalho; (ii) especificação e sequência de tarefas de DC a executar; (iii) execução das tarefas e (iv) pós processamento e controlo. Depois de cada processo de DC os dados são alterados, assim como os metadados, pelo que outro DP deve ser executado. Por exemplo, somente depois de uma correção da sintaxe dos códigos postais pode ser detetada uma dependência funcional (Naumann, 2014).

Em *Data Warehouse,* a DC é aplicada na resolução de entidades durante o processo de juntar várias DB (Hernández & Stolfo, 1998), bem como na ligação de registos e integração semântica (Elmagarmid et al., 2007). É geralmente o primeiro passo no pré-processamento de dados. Inspeciona, deteta a falta ou a incorreção de dados (Simoudis et al., 1995) e, por exemplo, analisa os dados que poderão causar problemas estatísticos como valores atípicos (Marcus et al., 2001).

2.5.2 Gestão de qualidade de dados orientada aos processos

As fases de DQI apresentadas em 2.5.1 (DM, DP e DC) são orientadas para os dados e por essa razão, não se focam na origem dos DQP. Se a DQ for melhorada sem ajustar os processos que estão na origem desses DQP, os sistemas irão continuar a produzir dados de qualidade inferior. Para que a DQI se torne consistente é necessário também aplicar métodos que se centrem nos processos que geram dados permitindo, assim, que os processos se tornem mais eficientes a longo prazo.

A gestão de DQ orientada para os processos foca-se na otimização destes, identificando e eliminado as causas dos erros e tornando a DQI num processo sustentável. Tal como

defendem Glowalla & Sunyaev (2013), o controlo de DQ pode ser acrescentado aos processos de negócio como uma tarefa adicional.

Shankaranarayanan (2000) apresenta o IP-MAP para modelar os processos organizacionais, permitindo incluir blocos de DQ (vide 2.3). Esta ferramenta de modelagem apresenta várias vantagens: (i) todo o processo de fabricação de dados pode ser facilmente visualizado, bem como as fases que podem afetar a DQ; (ii) a representação conceptual permitirá detetar estrangulamentos no processo de fabricação de dados e estimar o tempo entrega dos mesmos; (iii) o IP-MAP, baseado em princípios de melhoria contínua, possibilitará identificar os responsáveis dos processos assim como implementar processos de qualidade na origem dos dados; (iv) o IP-MAP irá permitir perceber a estrutura organizacional, assim como os limites do sistema de informação utilizado pelos diferentes processos de produção de dados e (v) possibilitará a medição da DQ, nas dimensões adequadas, em várias fases da sua produção.

3 METODOLOGIA

Para atingir o objetivo proposto por esta dissertação, e após ter sido efetuada a revisão de literatura em DM e DQ, iniciou-se um processo colaborativo com a empresa *RetailPC* (nome fictício), recorrendo aos conceitos presentes na metodologia *Action Research*.

A escolha desta metodologia justifica-se por duas ordens de razões. A primeira, por permitir que o investigador assuma um papel intervencionista na resolução de DQP na *RetailPC,* com o propósito de melhorar a DQ da entidade *Cliente* durante o processo de DM entre ERPs. A segunda, por possibilitar preparar a organização para futuros ciclos de *Action Research,* dando continuidade ao processo de DQI.

A área de IS é a ideal para a utilização da A*ction Research* pois atende, simultaneamente, à resolução de problemas reais e à expansão do conhecimento (Baskerville & Wood-Harper, 1996; Checkland & Holwell, 1998; Hult & Lennung, 1980). É descrita como sendo

a preferida dentro das metodologias pós-positivistas. É empírica, experimental e observacional e, no entanto, interpretativa, multivariada e intervencionista (Baskerville & Wood-Harper, 1996). Como se pretende intervir na organização durante o processo de investigação, considerou-se ser a metodologia ideal para desenvolver a pesquisa.

Checkland (1985) apresenta uma contextualização intelectual dos elementos base de qualquer investigação quando é utilizada a *Action Research* (Figura 3). Os componentes essenciais desse modelo são: uma *framework* (F), que serve de suporte e ligação entre ideias; uma metodologia (M), que especifica a forma de aplicação dessas ideias; e uma área de aplicação (A), onde é realizada a pesquisa. O objetivo é definir um enquadramento que capacite os intervenientes da investigação e que forneça conhecimento útil para agir analogamente em situações futuras.

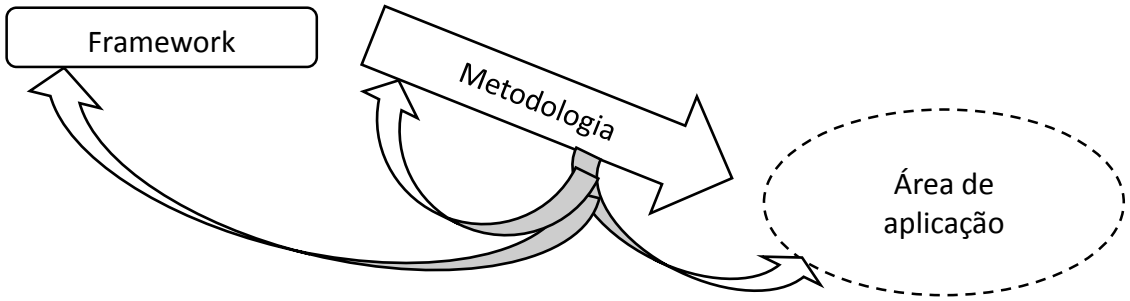


Figura 3: O uso organizado de pensamento racional.
(Fonte: Checkland (1985), p.758)

No caso concreto da *RetailPC*, a *framework* foi fundamentada em teorias e conceitos sobre a aplicabilidade de tarefas de DQ em processos de DM, evidenciada na revisão da literatura. A *Action Research* foi a metodologia utilizada e a área de aplicação consistiu no processo de avaliação DQI da entidade *Cliente* num processo de DM entre ERPs.

Sempre que é utilizada a *Action Research* são esperadas modificações na *framework*, na metodologia, mas principalmente, na área de aplicação. Segundo Checkland & Holwell (1998), a participação do investigador, enquanto interveniente ativo, envolvendo-se em situações e problemas reais, é o princípio mais importante desta metodologia. Defendem que, sem uma clara definição de cada um destes elementos, torna-se difícil

que o resultado da pesquisa seja proveitoso e que seja possível atingir os objetivos propostos.Tendo em conta que a *Action Research* procura que os resultados obtidos sejam relevantes, visando a resolução de problemas específicos, será necessário ter especial atenção às seguintes situações: (i) falta de imparcialidade do investigador; (ii) falta de disciplina; (iii) confusão com consultoria e (iv) dependência do contexto da pesquisa, que torna difícil generalizar as soluções encontradas.

Para assegurar rigor e relevância na pesquisa e evitar os perigos identificadas por Baskerville & Wood-Harper (1996), definiu-se um conjunto de critérios, conforme se apresenta de seguida.

Qual o papel do investigador e da *RetailPC* e como evoluem ao longo da pesquisa?

Na pesquisa em causa o investigador não foi um mero observador e o papel interventivo foi sempre sujeito a análise e reflexão do próprio. As responsabilidades de cada um dos intervenientes no projeto foi definida no início do mesmo.

Que informação foi recolhida para suportar a resolução do problema e atingir o objetivo da investigação?

Observou-se o funcionamento do Primavera Business Suite (PBS) e as regras de qualidade de dados existentes na *RetailPC*, especialmente no que se refere à entidade *Cliente*. Realizaram-se *workshops* internos sobre conceitos de DQ e criou-se um registo das ideias e dos assuntos debatidos.

Como foi estabelecida a relação entre o investigador e a *RetailPC*?

A iniciativa surgiu de um trabalho colaborativo entre o investigador (enquadramento teórico) e a *RetailPC* (componente prática). Sempre que existiam decisões a tomar, essa autoridade estava atribuída à *RetailPC*.

Como foi reconhecida a utilidade da solução para o problema em causa?

A definição da utilidade dos resultados obtidos suporta a imparcialidade da investigação e ajuda a criar uma base para desenvolvimentos futuros (Baskerville &

Wood-Harper, 1996). O sucesso da investigação foi reconhecido pela *RetailPC* na medida em que decidiu utilizar a mesma metodologia para as restantes entidades.

Como é definida a *framework* teórica que suporta a investigação?

Para garantir a imparcialidade da pesquisa é necessário ligar os resultados obtidos com a *framework* que suporta a investigação, distinguindo-a, assim, de consultoria (Baskerville & Wood-Harper, 1996). Reconhecer que ocorreram alterações nos processos e que o conhecimento adquirido pode ser utilizado em futuros ciclos, ajudou a defender a metodologia escolhida.

Seguiram-se as orientações propostas na *Action Research*, envolvendo uma componente teórica e uma componente prática, que foram desenvolvidas em conjunto para permitir avanços na pesquisa. A primeira consiste em estudar a real mudança pretendida na organização, enquanto a segunda consiste em conduzir um processo de investigação interativo que envolverá as 5 fases da *Action Research.*

Quando surgiu o problema de DQI, durante o processo de DM foi criado um acordo formal com a *RetailPC*, no qual ficou definido o objetivo da investigação e a metodologia a ser utilizada. O acordo deu autoridade à equipa de investigação para tomar decisões em benefício da organização, definiu o âmbito da pesquisa e a responsabilidade de cada uma das partes (Baskerville & Wood-Harper, 1996). Este projeto teve o patrocínio e a concordância da gestão de topo, sem a qual seria difícil implementar as medidas necessárias para obter dados de qualidade (Redman, 1995). A pesquisa foi conduzida de uma forma colaborativa e iterativa, com o diagnóstico, mitigação do problema e reflexão como atividades principais. As tarefas definidas na Tabela VI foram compostos por um conjunto de atividades propostas por Susman (1978), que permitiram suportar os processos de DM e DQI. O envolvimento do investigador foi de um agente facilitador na aplicação dos processos de DQI.

TABELA VI: TAREFAS A REALIZAR DURANTE O PROJETO DE INVESTIGAÇÃO

Tarefas	Fases de cada tarefa
• DM para o DS • DQI • DM para SBO • Alteração da forma de recolha de dados	Diagnóstico
	Planeamento da solução
	Implementação da solução
	Avaliação
	Reflexão e aprendizagem

Foram planeadas dois tipos de ações: (i) tomar medidas corretivas sobre DQP detetados e (ii) alterar os processos de negócio para corrigir as fontes da baixa DQ.

Para a intervenção reativa recorreu-se ao Oracle Data Quality Enterprise (ODQE), um *software* de *DP e DQI*. Para os processos de negócio que tiveram de ser redesenhados, foi utilizado o constructo sugerido por Lee et al.(2006). A reflexão e aprendizagem, apesar de formalmente constituir a última fase de cada ciclo, é um processo contínuo que decorre durante toda a investigação.

Para concluir, pode considerar-se que o conhecimento adquirido durante a pesquisa utilizando *Action Research* pode ser direcionado para três fins: a restruturação de processos que reflitam o conhecimento adquirido pela organização durante a pesquisa, disponibilizar meios de diagnóstico em futuros ciclos de *Action Research*, e disponibilizar conhecimento adicional para investigações futuras.

4 APRESENTAÇÃO DO CASO

A *RetailPC* é uma empresa de comércio a retalho de equipamento informático. Conta com três lojas físicas e uma *online*. Utiliza atualmente o PBS, mas este demonstra-se claramente insuficiente para apoiar a evolução dos processos de negócio. A *RetailPC* optou por mudar de ERP e iniciou o processo de implementação do SAP Business One (SBO), integrado com a solução da Magento para lojas *online*. Apostada em aumentar o seu crescimento e a sua visibilidade *online*, a *RetailPC* reconhece a extrema importância da DQ, em especial na entidade *Cliente*. Pois é considerada pela *RetailPC* a entidade com

mais DQP e com maior impacto no negócio, ao nível das vendas, processos logísticos de expedição, devoluções e falhas de entrega.

A DM é uma componente essencial na mudança de um ERP e uma das tarefas mais importantes. Tem um elevado grau de dificuldade associado e, na opinião de Thalheim & Wang (2013), pode colocar em risco o sucesso de um projeto. É neste cenário que se proporciona a investigação sobre a DQI durante uma DM. Neste contexto foi sugerida uma solução que permitisse realizar o proposto sem, contudo, afetar a operacionalidade da *RetailPC*. Para isso estruturou-se o processo em quatro ciclos de tarefas (Figura 4).

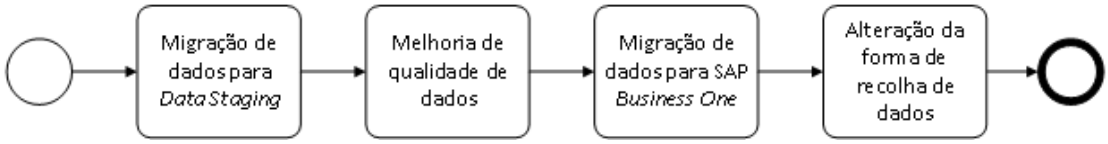


Figura 4: Tarefas a desenvolver durante a investigação.
(Fonte: Elaboração própria)

Para suportar o processo de DM foram criadas três DB intermédias, com diferentes objetivos, conforme se especifica na Tabela VII (Haller et al., 2011).

TABELA VII: DBS QUE COMPÕEM O *DATA STAGING*

DB	Objetivo
PBS_Staging	Previne que o processo de DM e DQI possa danificar a DB original ou que as operações com o PBS sejam mais lentas, por ser uma cópia da DB original.
SBO_Staging	Guarda os resultados intermédios do processo de DM. Todas as operações de DQI foram efetuadas sobre esta DB.
SBO_Final	Guarda os dados prontos a serem utilizados no ERP SBO. Os dados são inseridos nesta DB recorrendo ao Data Transfer Workbench, um *software* da SAP que utiliza a API do SBO.

4.1 CICLO DE MIGRAÇÃO DE DADOS PARA O *DATA STAGING*

O primeiro ciclo centrou-se na DM da DB PBS_Staging para a DB SBO_Staging. Este processo de DM foi dividido em duas tarefas principais: (i) mapeamento dos atributos das diferentes estruturas de dados; (ii) DM da DB PBS_Staging para a DB SBO_Staging.

4.1.1 Diagnóstico

Na análise da DB PBS e SBO foram utilizados processos de engenharia inversa de DB, a partir do qual se obteve um esquema das tabelas da DB. Estes processos englobaram duas fases: (i) a estrutura da DB foi recuperada como um esquema lógico baseado na implementação física da DB; (ii) o esquema lógico foi convertido num esquema de tabelas, conforme imagens que podem ser consultadas nos Anexo III e IV (Thalheim & Wang, 2013).

4.1.2 Planeamento da ação

Optou-se por executar o mapeamento entre os atributos do esquema de dados PBS e do esquema de dados do SBO manualmente (vide Anexos V e VI). Apesar de ser uma tarefa que consumiu bastante tempo, o profundo conhecimento do esquema de dados do PBS por parte da *RetailPC*, e do esquema de dados do SBO por parte do investigador, tornou o processo mais eficiente, tal como é defendido por Agrawal et al. (2008).

4.1.3 Implementação da solução

Os dados da DB PBS_Staging foram migrados para a DB SBO_Staging. Esta tarefa satisfez duas condições fundamentais: (i) foram extraídos todos os dados da DB PBS_Staging para a DB SBO_Staging; (ii) o modelo desenhado para o Data Staging (DS) manteve todas as restrições definidas sobre os objetos do sistema PBS (Thalheim & Wang, 2013). Para a DM foram definidas tarefas no SQL Server Integration Services, uma ferramenta de *Extraction, Transformation Loading* (ETL) da Microsoft.

4.1.4 Avaliação

Para verificar se a DM decorreu com sucesso foram realizados testes de completude e de correspondência, que verificaram se todos os registos foram migrados e se existiam registos na DB *SBO_Staging* que não tivessem correspondência na DB *PBS_Staging*. O esquema disponível no Anexo IV especifica que são necessários dois tuplos na tabela *CRD1* para cada tuplo na tabela *OCRD*. Os resultados são evidenciados na Tabela VIII.

TABELA VIII: RESULTADOS DOS TESTES DE DM DA DB PBS_STAGING PARA A DB SBO_STAGING

DB PBS_Staging	Nº de tuplos	DB SBO_Staging	Nº de tuplos
CondPag	10	OCTG	10
Contactos	55	COPR	55
Vendedores	34	OSLP	34
Moedas	3	OCRN	3
ModosExp	4	OSHP	4
Distritos	30	OCST	30
Clientes	78.753	OCRD	78.753
Clientes	78.753	CRD1	157.506

4.1.5 Reflexão e aprendizagem

Os resultados validaram que o conhecimento do domínio é um fator crítico para o correto mapeamento de atributos (Agrawal et al., 2008). Também se concluiu pela necessidade de efetuar testes semânticos depois da DM para a DB final assim que for possível consultar os dados no ERP SBO (Haller et al., 2011).

4.2 CICLO DE MELHORIA DA QUALIDADE DE DADOS

O segundo ciclo da investigação foi dedicado à DQI na DB SBO_Staging, de acordo com a taxionomia apresentada em 2.4(Oliveira & Rodrigues, 2005). Foi necessário identificar previamente os atributos mais importantes para o negócio e o tempo necessário para resolver cada um dos DQP identificados.

4.2.1 Diagnóstico

Após uma análise com a *RetailPC* foram identificados os atributos com impacto no negócio e foi determinado o tempo necessário para resolver cada um dos DQP.

O atributo *E_Mail* foi considerado pela *RetailPC* como tendo impacto no negócio pois, segundo a RetailPC, «serve como principal meio de comunicação com o cliente e será a identificação do utilizador na loja *online*».

Uma análise do DP permitiu identificar que o atributo *IntrntSite* da tabela *OCRD* está a ser utilizado para registar o endereço de correio eletrónico em lugar da página de internet do cliente. Esta situação verificava-se em 99,97% dos tuplos.

O atributo *ShipType* foi considerado pela *RetailPC* como tendo impacto no negócio, pois «é necessário que esteja definida a forma de envio dos produtos para os clientes». Depois da análise do DP apurou-se que, em apenas 7,64% dos tuplos, o atributo estava preenchido. Não foram encontrados dados que violassem os valores do domínio.

O atributo *PymCode* representa a forma de pagamento predefinida para um cliente e foi considerada com impacto no negócio porque «por exemplo, sempre que um cliente paga através de referência multibanco, só depois do pagamento efetuado é que o processo logístico é libertado». A análise ao DP deste atributo permitiu constatar que, em 2,33% dos tuplos, o atributo não estava preenchido e os restantes valores não pertenciam a nenhum domínio estabelecido.

O atributo *Currency* representa a moeda de transação com o cliente e o seu preenchimento é obrigatório. O DP do atributo evidenciou a existência de tuplos com diversos valores, segundo a *RetailPC* «todas as transações serão efetuadas em Euros».

O atributo *LangCode* determina a língua da documentação enviada para o cliente e o seu preenchimento é obrigatório. O DP deste atributo evidenciou a existência de tuplos com atributos não preenchidos.

Devido a alterações de requisitos legais o impacto do atributo *LicTradNum*, que regista o número de identificação fiscal (NIF), foi considerado elevado. O âmbito desta investigação limitou-se a analisar DQP de NIFs portugueses. A análise ao DP do atributo permitiu identificar 3 tipos de DQP a nível do atributo: (i) 92 tuplos com o atributo *LicTradNum* com uma quantidade de caracteres diferente de 9, sendo 42 de sujeitos passivos singulares e 50 de sujeitos passivos coletivos; (ii) 196 tuplos com o atributo *LicTradNum* com domínio inválido, verificação efetuada através do dígito de controlo utilizando o algoritmo Módulo 11 [1], correspondendo estes a 74 sujeitos passivos

[1] Este algoritmo certifica a validade do dígito de controlo, sendo através do Decreto-lei n.º 14/2013, de 28 de janeiro, que se encontram definidos os procedimentos a adotar na atribuição e gestão do NIF.

singulares e a 122 sujeitos passivos coletivos; (iii) 1610 tuplos com o atributo *LicTradNum* vazio, sendo 1557 de sujeitos passivos singulares e 53 sujeitos passivos coletivos. As regras de integridade do SBO obrigam que o atributo *LicTradNum* tenha como prefixo o código ISO[2] do país do cliente. Uma análise ao DP do atributo permitiu verificar que essa situação só se verifica em 0.61% dos casos. Foi também analisada a existência de valores duplicados no atributo *LicTradNum*, tendo sido identificados 53 tuplos.

O atributo *ZipCode*, cuja finalidade é registar os códigos postais, foi considerado um atributo com impacto no negócio. Tal como é referido pela *RetailPC* «sem código postal correto as encomendas não vão chegar ao destino». O âmbito desta investigação é restrito aos tuplos que correspondam a endereços portugueses. Sobre estes tuplos foi efetuado o DP ao atributo *ZipCode*. A análise ao padrão dos valores do atributo mostrou que somente 88,01% dos atributos respeitavam a sintaxe definida, 7,41% dos atributos tinham uma sintaxe incorreta e 4,58% dos atributos não estavam preenchidos.

Para a análise dos DQP de violação de domínio, recorreu-se a uma DB dos CTT que contém a lista de todos os códigos postais portugueses. O atributo *ZipCode,* de tuplos correspondentes a endereços portugueses[3], deverá pertencer ao conjunto de valores do atributo *CodigoPostal* da tabela *XX_CodigoPostal*. A análise ao DP permitiu verificar que existiam 10642 tuplos que violam esta regra.

TABELA IX:LISTA DE ATRIBUTOS COM IMPACTO NO NEGÓCIO

Tabela	Atributo	Impacto	Custo
OCRD	IntrntSite	Baixo	Baixo
OCRD	E_Mail	Elevado	Médio
OCRD	ShipType	Elevado	Baixo
OCRD	PymCode	Elevado	Baixo
OCRD	Currency	Baixo	Baixo
OCRD	LangCode	Médio	Baixo

[2] Disponivel em https://www.iso.org/obp/ui/
[3] http://www.ctt.pt/feapl_2/app/restricted/postalCodeSearch/postalCodeDownloadFiles.jspx.

OCRD	LicTradNum	Elevado	Elevado
CRD1	ZipCode	Elevado	Elevado

Depois de analisados os atributos da tabela OCRD e CRD1 individualmente, procedeu-se ao DP para verificar a existência de tuplos duplicados. Para considerarmos que um tuplo está duplicado elegemos um atributo da tabela OCRD – *CardName* – e três atributos da tabela CRD1 – *ZipCode*, *City*, *Street*. Utilizando este conjunto de atributos foi efetuada uma comparação com todas as entidades. Sempre que o valor da métrica Jaro- Winkler (1999) era superior a 0.80, o tuplo foi considerado duplicado. Através desta análise foi possível identificar 303 tuplos que provavelmente representam os mesmos clientes.

4.2.2 Planeamento da ação

Os primeiros DQP que foram abordados pelos processos de DQI são os que se enquadravam em valores de um atributo de um único tuplo (vide Anexo II). Com base nos diagnósticos efetuados anteriormente e, em conjunto com a *RetailPC,* foram delineados planos para mitigar os DQP.

Em relação ao atributo *IntrntSite* foi decidido: (i) mover os dados que têm uma elevada probabilidade de se serem endereços de correio eletrónico válidos para o atributo *E_Mail*; (ii) executar tarefas de DQI no atributo *E_Mail*.

Para os atributos *ShipType, Currency e LangCode* foram fixados valores predefinidos e decidiu preencher-se os tuplos que não tinham valores.

O atributo *PymCode* deve estar preenchido em todos os tuplos, sendo necessário converter os dados existentes para valores do domínio. Para efetuar essa conversão, foi discutida com a *RetailPC* a lista de equivalências entre os valores existentes e os valores do domínio.

Utilizando o diagnóstico efetuado para o atributo *LicTradNum* procurou-se uma solução para diminuir a quantidade de NIFs inválidos na DB. Decidiu separar-se os processos de DQI dos sujeitos passivos singulares dos processos de DQI dos sujeitos passivos

coletivos, porque os requisitos legais e a frequência com que cada tipo de sujeito passivo efetua compras são diferentes. No que diz respeito aos sujeitos passivos singulares portugueses, optou-se por eliminar os valores dos atributos que não tenham 9 caracteres de comprimento ou que não pertençam ao domínio válido para o atributo. Em relação aos sujeitos passivos coletivos foram definidos processos de DQI para corrigir os NIFs incorretos através de uma tripla validação dos dados dos clientes: (i) pesquisa do nome do sujeito passivo em serviços comerciais[4] obtendo uma lista de possíveis NIFs; (ii) esta lista é, posteriormente, sujeita a uma ordenação, onde é selecionado o melhor resultado obtido através da aplicação da métrica Jaro-Winkler (1999). Se nenhum resultado obtido para esta métrica for superior a 0.9 o valor será eliminado por se considerar que não existe nenhum NIF suficientemente semelhante para proceder à substituição; (iii) validação do NIF selecionado na fase (ii) através da página de internet da Comunidade Europeia – *webservice* do projeto VIES[5] ou no SICAE[6].

O último passo para corrigir os DQP de sintaxe do atributo foi decidido colocar o código ISO do país como prefixo do atributo *LicTradNum*.

Outro dos DQP a resolver enquadra-se no contexto de valores de um único atributo, segundo a taxonomia de Oliveira (2005). Os tuplos da tabela *OCRD* considerados duplicados na fase de diagnóstico foram sujeitos a um processo de escolha do melhor tuplo. Com a eliminação de um dos tuplos duplicados foi necessário assegurar a regras de integridade referencial com as tabelas que estão relacionadas.

Utilizando os resultados do DP do atributo *ZipCode* foi delineado um plano para minimizar a existência de valores errados na DB. A primeira fase desse plano consiste em corrigir os atributos que apresentam erros de sintaxe que possam ser corrigidos para

[4] Através do endereço eletrónico http:// www.nif.pt.
[5] Na página de internet da Comunidade Europeia – projeto VIES – apenas é possível validar NIFs que constem da BD por terem efetuado transações intracomunitárias.
[6] Cujos endereços são: http://ec.europa.eu/taxation_customs/vies/?locale=pt e www.sicae.pt.

a sintaxe correta, por exemplo "2440210" para "2440-210". Na segunda fase será efetuada uma consulta à página de internet dos CTT[7] através de um serviço *HTTP Request* para se obter informação completa do código postal em formato XML (CTT, 2014). Esta consulta será efetuada utilizando os atributos *Street, Block e City*. Os dados serão corrigidos sempre que a quantidade de tuplos devolvidos seja 1. Na última fase efetuar-se-á uma validação de violação de domínio, verificando a existência de valores do atributo *ZipCode* na tabela *XX_CodigoPostal*. Esta tabela foi obtida através da página de internet dos CTT.

4.2.3 Implementação da ação

Nesta presente secção descrevem-se as ações para resolver os DQP. Relativamente aos DQP em contexto de atributo de um único tuplo, apresentam-se na Tabela X os atributos e as tipologias de erros que foram corrigidos.

TABELA X: LISTA DE ATRIBUTOS E TIPOLOGIAS DE DQP EM CONTEXTO DE ATRIBUTO DE UM ÚNICO TUPLO

Tabela	Atributo	Valor em falta	Violação de sintaxe	Erro ortográfico	Violação de domínio
OCRD	IntrntSite		X		
OCRD	E_Mail		X		X
OCRD	ShipType	X			X
OCRD	PymCode	X			X
OCRD	LangCode	X			
OCRD	LicTradNum		X		X
CRD1	ZipCode	X	X		X

O atributo *E_Mail* foi atualizado com os dados do atributo *IntrntSite* que tinham a uma elevada probabilidade de se serem endereços de correio eletrónico válidos. Foram executadas as tarefas de DQI sobre o atributo *E_Mail* e foram aplicadas regras de validação de correio eletrónico.

[7] Disponível em http://www.ctt.pt/pdcp/xml_pdcp?

Foram executadas as tarefas previstas para os atributos *ShipType*, *PymCode*, *Currency* e *LangCode*, que decorreram conforme previsto.

De acordo com o planeado, foram executados os processos de DQI sobre o atributo *LicTradNum*. Relativamente aos sujeitos passivos individuais foram eliminados os dados dos atributos que não respeitavam a sintaxe ou o domínio. Para os atributos respeitantes a sujeitos passivos coletivos foi desenvolvido um *script* para tentar corrigir os valores que não respeitassem a sintaxe, o domínio ou que não estivessem preenchidos. Depois deste processo, os valores do atributo *LicTradNum* foram atualizados com o prefixo do país.

O atributo *ZipCode* foi sujeito às tarefas de DQI que foram planeadas. Inicialmente foram corrigidos os valores dos atributos cujos padrões eram facilmente identificáveis. Seguidamente, corrigiram-se os valores dos atributos que estavam *NULL* ou cuja sintaxe estivesse incorreta. Para se efetuar esta correção o atributo *Street* foi sujeito a tarefas de limpeza para retirar excertos inválidos dos nomes de rua (p.e. "Nº" ou "Lote"). Após esta operação foi desenvolvido um *script* que procurou obter um valor único de código postal na página de internet dos CTT. A aplicação deste *script* foi feita em vários ciclos, tendo sido otimizados os critérios de correção entre cada um deles.

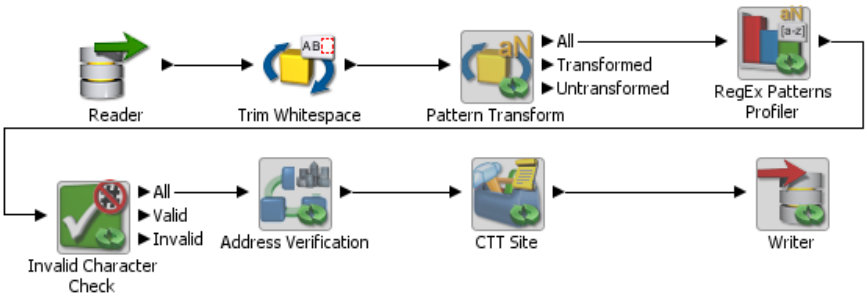


Figura 5: Exemplo do processo de DQI do atributo *ZipCode.*
(Fonte: Elaboração própria)

Relativamente aos DQP em contexto de valores de um único atributo, só foi identificado o atributo *LicTradNum* por violação de valor único. De acordo com o planeado para a resolução deste DQP foram elaborados processos de resolução de entidades e de

escolha do melhor tuplo (Figura 6). Foram, ainda, criados mecanismos para garantir a integridade referencial das tabelas relacionadas com os tuplos eliminados.

Figura 6: Exemplo do processo de deduplicação do OEDQ.
(Fonte: Elaboração própria)

4.2.4 Avaliação

Os atributos *IntrntSite* e *E_Mail* foram corrigidos em conjunto. Os dados do atributo *IntrntSite* foram melhorados conseguindo-se corrigir 96,43% dos endereços de correio eletrónico que encontravam em formatos incorretos. Foi também melhorada a qualidade de dados do atributo *IntrntSite* mas com um impacto reduzido para o negócio.

TABELA XI: RESULTADO DE DQI DO ATRIBUTO *E_MAIL*

Valor	Antes de DQI		Depois de DQI	
	Nº de tuplos	%	Nº de tuplos	%
Correio eletrónico em formato correto	45.469	99,31	45.586	99,65
Correio eletrónico com dados suspeitos	164	0,36	155	0,34
Correio eletrónico em formato desconhecido	108	0,24	5	0,01
Correio eletrónico em formato que pode ser sujeito a operações simples	32	0,07	0	0,00
Possivelmente endereço de página de internet	14	0,03	0	0,00

TABELA XII: EXEMPLOS DE ENDEREÇOS DE CORREIO CORRIGIDOS PELO OEDQ

Valor do atributo antes de DQI	Valor do atributo depois de DQI
João Mendes <joao.mendes@socem-its.pt>	joao.mendes@socem-its.pt
r.oliveirabraga1969€gmail.com	r.oliveirabraga1969@gmail.com
suporte@falx.pt/candidaterra@falx.pt	suporte@falx.pt
	candidaterra@falx.pt
rfdvalente91@gmail	rfdvalente91@gmail.com

Para os atributos *ShipType*, *PymCode, Currency* e *LangCode* foi melhorada a DQ a 100%. O DQI do atributo *LicTradNum* foi executado em duas fases distintas. Na primeira fase foram realizados os processos de DQI de NIF de sujeitos passivos individuais e, numa segunda fase, foram realizadas os processos de DQI dos sujeitos passivos coletivos. No caso dos sujeitos passivos coletivos foi possível corrigir 98.53% dos DQP. A Tabela XIII

apresenta alguns exemplos de NIFs que foram tratados por processo de DQI para violações de sintaxe. Foi utilizada a métrica de Winkler (1999) para obter o grau de semelhança. A resolução do DQP de violação de valor único no tuplo *LicTradNum* permitiu eliminar 52 tuplos.

TABELA XIII: EXEMPLOS DE AÇÕES DE DQI SOBRE O ATRIBUTO *LICTRADNUM*

NIF incorreto	Melhor NIF sugerido	Grau de semelhança	Ação sobre o atributo
503979051	503979058	0,956	Valor Corrigido
502287102	502287110	0,974	Valor Corrigido
501075212	500075212	0,978	Valor Corrigido
506600309	506600319	0,978	Valor Corrigido
501406783	501456783	0,956	Valor Corrigido
503985399	503985775	0,889	Valor Eliminado

O atributo *ZipCode* foi corrigido com a execução das tarefas de DQI que estavam planeadas. A primeira tarefa de DQI permitiu corrigir 56.67% dos atributos que tinham erros de sintaxe. A pesquisa de informação na página de internet dos CTT permitiu corrigir 1728 atributos. A Tabela XIV apresenta alguns exemplos de códigos postais que foram corrigidos e a Tabela XV apresenta os principais resultados obtidos das tarefas de DQI executadas.

TABELA XIV: EXEMPLOS DE CÓDIGOS POSTAIS CORRIGIDOS

Rua	Cidade	Cód. Postal na DB PBS	Cód. Postal corrigido
rua escritor julião quintinha, nº27A	beja	7800-61	7800-061
Quinta do Galo, lote 24 4ºEsq.	Viseu	3500	3500-849
Rua Principal, nº9	Lameiro	2425	2425-362
Rua Das Achadas Nº8	Meirinhas	3100	3105-462
Urb. do Quintalao, Lot 8 2 esq	Cartaxo	20701-153	2070-153

O significado dos valores da coluna “Padrão” são os seguintes: “N” representa um caracter numérico; “p” representa um “-“ e “_” representa um espaço.

TABELA XV: RESULTADO DA CORREÇÃO DOS PRINCIPAIS PADRÕES DO ATRIBUTO *ZIPCODE*

Padrão	Quantidade de tuplos antes de DQI	%	Quantidade de tuplos depois de DQI	%
NNNNpNNN	126.778	88,01	135.962	94,40

[vazio]	6.594	4,57	6.154	4,30
NNNN_p_NNN	6.444	4,47	0	0,00
NNNN	2.480	1,72	1.424	1,00
NNNNpNNN_	472	0,33	0	0,00
NNNN_pNNN	224	0,16	0	0,00
NNNN_p_NNN_	176	0,12	0	0,00

4.2.5 Reflexão e aprendizagem

A falta de um atributo específico para registar o correio eletrónico dos clientes e a falta de procedimentos levou a que se utilizasse de uma forma incorreta o atributo *IntrntSite*. A *RetailPC* aprendeu que deve definir melhor os processos internos e que deve criar novos atributos sempre que os existentes não sejam adequados.

Os atributos *ShipType, PymCode*, *Currency* e *LangCode* têm impacto no negócio, pelo que foi sugerido que o seu preenchimento fosse obrigatório e tivessem valor predefinido.

Atualmente o atributo *LicTradNum* não está sujeito a nenhuma validação de negócio, nem existe nenhum procedimento interno sobre as regras a implementar. A *RetailPC* aprendeu que este atributo tem impacto no negócio, nomeadamente a nível do cumprimento das obrigações fiscais. Foi sugerido incluir mecanismos de validação dos valores do atributo.

No que se refere ao *ZipCode* não foi possível corrigir todos os valores do atributo que estavam incorretos. Atualmente não existe nenhuma regra de negócio que defina os critérios de preenchimento para este atributo, mas foram sugeridas procedimentos que podem assegurar o seu correto preenchimento na tabela CRD1. Sendo difícil manter na DB os códigos postais existentes a nível mundial é, pelo menos, aconselhável definir processos que validem os códigos postais nacionais.

Pode concluir-se, quanto ao conhecimento científico, que foi possível percecionar a importância da definição taxinómica dos DQP e da sua relação com as dimensões de DQ

definidas. Cada projeto de DQI deverá atender as dimensões de DQ necessárias para a resolução dos DQP identificados.

4.3 CICLO DE MIGRAÇÃO DE DADOS PARA SAP BUSINESS ONE

O terceiro ciclo da investigação foi dedicado à DM da DB SBO_Staging para a DB SBO_Final. Este processo está condicionado pela ferramenta de DM que o SBO exige que seja utilizada: o *Data Transfer Workbench* (DTW), um *software* da SAP criado especificamente para o processo de DM.

4.3.1 Diagnóstico

O mapeamento dos atributos e a DM entre a DB PBS_Staging e a SBO_Staging foi apresentado em 4.1. Esta tarefa deixou a DB num esquema igual ao da DB SBO_Final, pelo que não é necessário efetuar nenhuma tarefa adicional. A DB SBO_Staging foi sujeita a um processo de DQI que deixou os dados preparados para serem migrados para a DB SBO_Final.

4.3.2 Planeamento da ação

Para que os dados sejam migrados com sucesso para a DB SBO_Final é importante respeitar a sequência que é definida pelas regras de integridade referencial (vide Anexo IV). Por exemplo é necessário migrar a tabela *OSHP* - que representa os modos de entrega - antes da tabela *OCRD*, que representa os clientes. Assim a migração de dados será efetuada pela seguinte sequência: *OSHP*; *OSLP*; *OCTG*; *OCST*; *OCRD*; *CRD1*; *OCPR*. As tabelas *OCRG, OLNG*, *OCRN*, *OPLN* e *OCRY* não serão alvo de nenhuma operação porque fazem parte dos dados que vêm preenchidos por defeito com o SBO. A tabela *CRD2* não será importada, já que o seu preenchimento é efetuado pela API do SBO.

4.3.3 Implementação da ação

De acordo com o planeado foi executada a DM para DB SBO_Staging. Para garantir que a DM não ficava incompleta e com erros, caso ocorresse um erro inesperado, cada tabela foi incluída em uma transação.

4.3.4 Avaliação

À semelhança da DM para a DB SBO_Staging, foram realizados testes de completude e de correspondência. Através dos mesmos foi possívesl verificar que todos os tuplos foram migrados com sucesso, conforme é apresentado na

Tabela XVI.

Para efetuar os testes semânticos (Haller et al., 2011) recorreu-se a colaboradores da *RetailPC*, que são quem melhor conhece o domínio e mais facilmente podem identificar erros. Foi definida uma amostra aleatória, composta por 100 tuplos da tabela de OCRD, que foram validados com sucesso. A título de exemplo exibe-se na

Figura 7 o ecrã de um cliente validado com sucesso.

TABELA XVI: RESULTADOS DOS TESTES DE DM DA DB SBO_STAGING PARA A DB SBO_FINAL

Tabela	Nº de tuplos na DB SBO_Staging	Nº de tuplos na DB SBO_Final
OCTG	10	10
OCPR	55	55
OSLP	34	34
OSHP	4	4
OCST	30	30
OCRD	78.700	78.700
CRD1	157.400	157.400

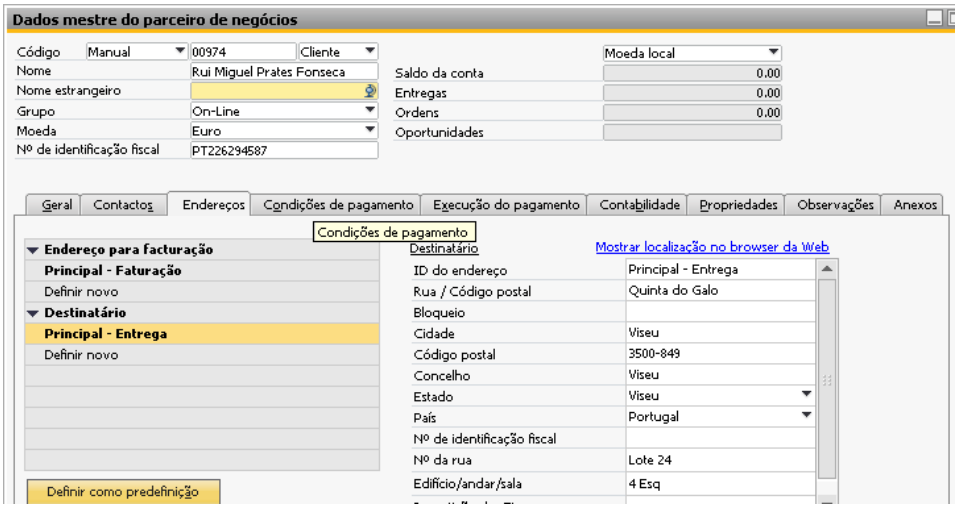


Figura 7: Exemplo de teste visual a um cliente no ERP SBO.
(Fonte: Elaboração própria)

4.3.5 Reflexão e aprendizagem

No final deste ciclo foi possível apreender que recorrendo à arquitetura sugerida por Haller et al. (2011), o processo de DM da DB SBO_Staging para a DB SBO_Final não se tratou de uma tarefa complexa. O trabalho desenvolvido no ciclo de DM da DB PBS_Staging para a DB SBO_Staging (4.1) e no ciclo DQI (4.2) deixaram os dados com um nível de qualidade exigido para que este ciclo decorresse sem anomalias.

4.4 CICLO ALTERAÇÃO DA FORMA DE RECOLHA DE DADOS

O quarto e último ciclo da investigação é dedicado à alteração da forma de recolha dos dados com impacto no negócio. Em 4.2 procedeu-se à DQI dos dados existentes e apesar da melhoria significativa na DQ, as correções efetuadas não impedem a entrada de dados de baixa qualidade na DB. Optou-se por isso, incorporar pontos de validação da DQ nos processos de negócio da *RetailPC*.

4.4.1 Diagnóstico

Iniciou-se a identificação dos processos que originam o registo de dados da entidade *Cliente* analisando a documentação existente em estreita colaboração com a *RetailPC.* Esta análise permitiu clarificar o processo na sua globalidade e elaborar uma lista dos seus componentes mais críticos. Os atributos considerados para essa lista foram: *ZipCode* e *LicTradNum*. A duplicação de tuplos também foi considerado um erro com impacto no negócio, por isso também deverá ser sujeito a validação no momento de criação do cliente. A recolha de dados é efetuada em dois sistemas distintos. Os clientes introduzem os seus dados através da página de internet e o departamento comercial introduz os dados diretamente no ERP. É necessário garantir a harmonização dos processos de validação de dados entre as duas plataformas.

4.4.2 Planeamento da ação

Em conjunto com a *RetailPC*, foi decidido implementar um conjunto de regras disponíveis para consulta no Anexo VII. Foram definidos dois tipos de ações, que podem

ocorrer quando, um erro é detetado: (i) exibição de um alerta, que permite continuar o registo do cliente; (ii) uma mensagem de erro, que obriga a correção da regra violada, antes de prosseguir com o registo do cliente. De forma a representar o fluxo de informação associado ao processo de negócio e aos pontos de controlo de erros construi-se um diagrama IP-MAP (Shankaranarayanan et al., 2000), conforme se apresenta na Figura 8.

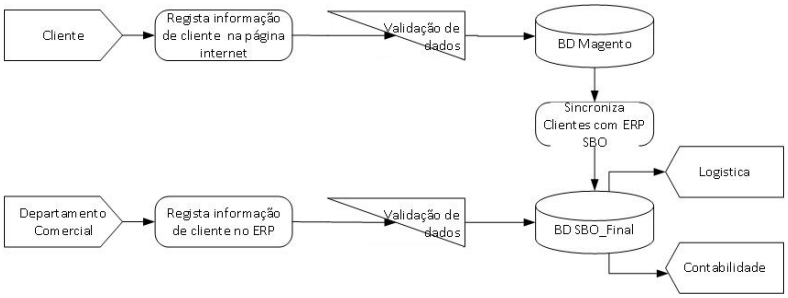


Figura 8: Fluxograma IP-MAP.
(Fonte: Elaboração própria)

4.4.3 Implementação da ação

Conforme planeado e em conjunto com a *RetailPC* foi desenvolvida uma plataforma única, de validação de informação acessível por todos os sistemas das empresas. Para tal foi criado um *webservice* que disponibiliza funções de validação e efetua sugestões de preenchimento de dados.

O atributo *LicTradNum* está sujeito a validação para prevenir a entrada na DB de tuplos com dados errados, tal como se exemplifica na Figura 9.

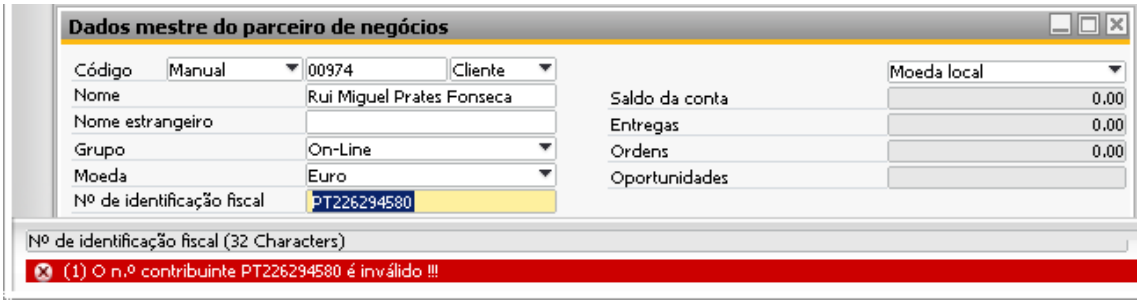


Figura 9: Exemplo de erro no atributo *LicTradNum.*
(Fonte: Elaboração própria)

Em relação ao atributo *ZipCode* as alterações foram mais abrangentes. Optou-se por alterar a forma de recolha de informação e colocou-se o preenchimento do atributo *ZipCode* em primeiro lugar. Todos os valores de atributos que se podem obter a partir

do código postal inserido no atributo *ZipCode* são preenchidos automaticamente. Por exemplo, se se colocar o valor "1700-131" no atributo *ZipCode*, o atributo *State* será preenchido com o valor "10" que corresponde a "Lisboa", o atributo *County* será preenchido com o valor "Lisboa", o atributo *Street* será preenchido com o valor " Rua Dom Alberto Bramão" e para o atributo *StreetNo* será sugerida uma lista de valores numéricos impares de 1 a 15. Foi ainda adicionada uma validação de inconsistência entre valores dos atributos, que verifica se o valor dos atributos *State* (que armazena os dados do distrito), *County* (que armazena os dados do Concelho) e *City* são consistentes com os valores do atributo *ZipCode*.Por último foi implementada uma validação de duplicação de entidade avaliando as semelhanças entre os atributos *CardName* da tabela OCRD e *ZipCode*, *City* e *Street* da tabela CRD1.

4.4.4 Avaliação

Para verificar os procedimentos de validação foi criado e validado, com a *RetailPC,* um conjunto alargado de testes. Os resultados dos testes foram os esperados. Foram introduzidos métricas para avaliar a execução dos testes e foi possível verificar que 10,03% dos tuplos foram bloqueados pelas regras de validação à primeira tentativa, sendo ainda possível apurar que 17 utilizadores desistiram do processo de registo, o que equivale a 0,56% dos registos criados durante o primeiro mês de utilização das regras de validação.

4.4.5 Reflexão e aprendizagem

A integração de blocos de verificação de DQ nos processos de negócio permitiu, à *RetailPC*, obter uma visão holística dos dados e do fluxo que os origina. Garante ainda que os dados refletem as iniciativas das mudanças organizacionais de DQI. A alteração da forma de recolha de dados e a constante monitorização da qualidade de dados é um processo de melhoria contínua da DQ.

Relativamente ao conhecimento científico podemos concluir que é necessário vincular os processos de negócio às regras de DQ. As características intrínsecas de DQ, só por si, não garantem a resolução dos problemas organizacionais, só se torna possível a resolução eficaz de DQP incorporando processos de melhoria contínua de DQ.

5 CONCLUSÃO

Constituiu objetivo desta investigação avaliar e melhorar a DQ, durante a DM entre ERPs, nos atributos da entidade *Cliente*, que a *RetailPC* elegeu como sendo a entidade com mais DQP e com maior impacto no negócio, ao nível das vendas, processos logísticos de expedição, devoluções e falhas de entrega.

Para o efeito, foi utilizada a *Action Research,* considerada a mais adequada para esta pesquisa, já que permitia que se tivesse um papel interventivo na organização e nos seus processos de negócio. O conhecimento adquirido durante a investigação foi direcionado para restruturação de processos da *RetailPC* e, simultaneamente, para a preparação de futuros ciclos de *Action Research*. Seguindo esta metodologia foram efetuados as seguintes tarefas: (i) DM para a DB SBO_Staging, (ii) DQI na DB SBO_Staging, (iii) DM para a DB SBO_Final e, finalmente, (iv) alteração dos processos de recolha de dados. As decisões tomadas durante a investigação foram fundamentadas na *framework* obtida com a revisão da literatura. Com respeito à DM foi utilizada a arquitetura sugerida por Haller et al. (2011), para o processo de DQI recorreu-se aos princípios da TDQM sugerida por Lee et al. (2004) e utilizou-se a classificação apresentada por Oliveira & Rodrigues (2005) para determinar a taxonomia dos erros a corrigir.

A DM foi realizada com sucesso tendo os resultados sido validados pela *RetailPC* através de uma amostra alargada. No que respeita à DQI, foi possível melhorar significativamente a DQ dos atributos da entidade cliente com impacto para o negócio, com especial relevância para os atributos *LicTradNum* e *ZipCode*, tendo melhorado,

respetivamente, 98,53% dos NIF das pessoas coletivas e 56,67% dos códigos postais que se encontravam errados. Relativamente aos atributos *ShipType, PymCode, Currency* e *LangCode* foi possível corrigir 100% dos erros detetados. Quanto ao atributo *IntrntSite*, e uma vez que estava a ser utilizado para um fim diferente do previsto pelo ERP, foi possível corrigir 99,65% dos tuplos que continham valores. Foi ainda efetuada uma análise e limpeza de tuplos duplicados, tendo sido eliminados 323 tuplos de duplicados na entidade *Cliente.* No que diz respeito às alterações dos processos de recolha de dados, foram implementados e monitorizados controlos de forma a impedir que os erros detetados se voltassem a verificar, tornando assim o processo de DQI mais consistente. Como resultado da investigação, a *RetailPC* instituiu processos formais de avaliação e garantia da DQ e ficou, ainda, demonstrada a sua intenção em aplicar a mesma metodologia nas restantes entidades do ERP.

Foi possível percecionar a importância da definição taxinómica dos DQP e da sua relação com as dimensões de DQ definidas. Cada projeto de DQI deverá atender as dimensões de DQ necessárias para a resolução dos DQP identificados. As tarefas de DQI que se concentram nos dados não são, só por si, suficientes para garantir a completa resolução dos DQP. É necessário vincular os processos de negócio às regras de DQ.

No que respeita às limitações do presente estudo destaca-se o facto de, não ter sido possível melhorar a DQ para todos os tuplos, visto existirem alguns que não tinham nenhum atributo da morada preenchido. Para outros tuplos não foi possível aprofundar, no tempo disponível, as tarefas de DQ, pelo facto de, p. e., existirem observações adicionais no atributo *Street*. Adicionalmente, e uma vez que não se tinha experiência na utilização do OEDQ, não foram aproveitadas integralmente as capacidades deste *software* na DQI.

Como sugestão de trabalho futuro julga-se que seria de alargar o trabalho efetuado (DM e DQI) a outras entidades da DB da *RetailPC,* nomeadamente, à entidade *Fornecedores*

e à entidade *Artigos*. No que respeita à entidade *Clientes,* julga ainda possível desenvolver novos ciclos de DQI e de alteração de processos para outros atributos, visto que "*Data quality is not a project, it is a lifestyle*" (David Wells)

6 BIBLIOGRAFIA

Agrawal, H., Chafle, G., Goyal, S., Mittal, S. & Mukherjea, S. (2008). An enhanced extract-transform-load system for migrating data in telecom billing. In: *Proceedings - International Conference on Data Engineering*. 2008, pp. 1277–1286.

Ballou, D.P. & Pazer, H.L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*. 31 (2). p.pp. 150–162.

Ballou, D.P. & Tayi, G.K. (1989). Methodology for allocating resources for data quality enhancement. *Communications of the ACM*. 32 (3). p.pp. 320–329.

Ballou, D.P., Wang, R.Y., Pazer, H.L. & Tayi, G.K. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science*. 44 (4). p.pp. 462–484.

Baškarada, S. & Koronios, A. (2014). A Critical Success Factor Framework for Information Quality Management. *Information Systems Management*. 31 (4). p.pp. 276–295.

Baskerville, R.L. & Wood-Harper, A.T. (1996). A critical perspective on action research as a method for information systems research. *Journal of Information Technology*. 11. p.pp. 235–246.

Cao, L. & Zhu, H. (2013). Normal Accidents: Data Quality Problems in ERP-Enabled Manufacturing. *Journal of Data and Information Quality*. 4 (3). p.pp. 1–26.

Checkland, P. (1985). From optimizing to learning: A development of systems thinking for the 1990s. *Journal of the Operational Research Society*. 36 (9). p.pp. 757–767.

Checkland, P. & Holwell, S. (1998). Action research: its nature and validity. *Systemic Practice and Action Research*. 11 (1). p.pp. 9–21.

Codd, E.F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*. 13 (6). p.pp. 377–387.

Codd, E.F. (1990). *The Relational Model for Database Management : Version 2*.

CTT (2014). Manual Técnico de Utilização - Validação de Códigos Postais. *CTT*. [Online]. p.pp. 1–3. Available from: https://www.ctt.pt/contentAsset/raw-data/09ae68cd-de81-40f1-9dd6-28728b0bbefd/ficheiro/f9b2b545-2002-44a6-9282-d04d6dcab63a/export/man_util_xml_v16.pdf.

Davison, R., Martinsons, M.G. & Ou, C.X.J. (2012). The roles of theory in canonical action research. *MIS Quarterly: Management Information Systems*. 36 (3). p.pp. 763–786.

Elmagarmid, A.K., Ipeirotis, P.G. & Verykios, V.S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*. 19 (1). p.pp. 1–16.

Fisher, C.W., Chengalur-Smith, I. & Ballou, D.P. (2003). The impact of experience and time on the use of data quality information in decision making. *Information Systems Research*. 14 (2).

Glowalla, P. & Sunyaev, A. (2013). Process-Driven Data Quality Management Through Integration of Data Quality into Existing Process Models. *Business & Information Systems Engineering*. 5 (6). p.pp. 433–448.

Haller, K., Matthes, F. & Schulz, C. (2011). Testing & Quality Assurance in Data Migration Projects. In: *27th IEEE International Conference on Software Maintenance ICSM*. 2011.

Hernández, M.A. & Stolfo, S.J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*. 2 (1). p.pp. 9–37.

Hult, M. & Lennung, S. (1980). Towards a definition of action research: a note and bibliography. *Journal of Management Studies*. 25 (6).

Kahn, B.K., Strong, D.M. & Wang, R.Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*. 45 (4ve). p.pp. 184–192.

Kim, W., Choi, B.J., Hong, E.K., Kim, S.K. & Lee, D. (2003). A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery*. 7 (1). p.pp. 81–99.

Lee, Y.W., Pipino, L.L., Funk, J.D. & Wang, R.Y. (2006). *Journey to data quality*.

Lee, Y.W., Pipino, L.L., Strong, D.M. & Wang, R.Y. (2004). Process-embedded data integrity. *Journal of Database Management*. 15 (1). p.pp. 87–103.

Lee, Y.W., Strong, D.M., Kahn, B.K. & Wang, R.Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management*. 40 (2). p.pp. 133–146.

Madnick, S., Wang, R.Y., Lee, Y.W. & Zhu, H. (2009). Overview and framework for data and information quality research. *Journal of Data and Information Quality*. 1 (1). p.pp. 1–22.

Manjunath, T.N. & Hegadi, R.S. (2013). Data Quality Assessment Model for Data Migration Business Enterprise. *International Journal of Soft Computing*. 5 (1). p.pp. 101–109.

Marcus, A., Maletic, J.I. & Lin, K.-I. (2001). Ordinal association rules for error identification in data sets. *Proceedings of the tenth international conference on Information and knowledge management - CIKM'01*. p.p. 589.

Milano, D. (2005). Using Ontologies for XML Data Cleaning. In: *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops*. pp. 562–571.

Müller, H. & Freytag, J. (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing. *Challenges*. (HUB-IB-164). p.pp. 1–23.

Naumann, F. (2014). Data profiling revisited. *ACM SIGMOD Record*. 42 (4). p.pp. 40–49.

Oliveira, P. (2008). *Detecção e Correcção de Problemas de Qualidade dos Dados: Modelo, Sintaxe e Semântica*. p.p. 383.

Oliveira, P. & Rodrigues, F. (2005). A taxonomy of data quality problems. *2nd Int. Workshop on Data and Information Quality*.

Olsen, J. (2003). *Data Quality: The Accuracy Dimension*.

Rahm, E. & Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* 23 (4). p.pp. 3–13.

Redman, T.C. (1995). Improve Data Quality for Competitive Advantage. *Management*. 36. p.pp. 99–107.

Shankaranarayanan, G., Wang, R.Y. & Ziad, M. (2000). IP-MAP: Representing the Manufacture of an Information Product. In: *Proceedings of the 2000 Conference on Information Quality*. 2000, pp. 1–16.

Simoudis, E., Livezey, B. & Kerber, R. (1995). Using Recon for Data Cleaning. In: *Proc. 1995 Int. Conf. Knowledge Discovery and Data Mining (KDD'95)*. 1995, pp. 282–287.

Strong, D.M., Lee, Y.W. & Wang, R.Y. (1997). Data quality in context. *Communications of the ACM*. 40 (5). p.pp. 103–110.

Susman, G.I. & Evered, R.D. (1978). An Assessment of the Scientific Merits of Action Research. *Administrative Science Quarterly*. 23 (4). p.pp. 582–603.

Thalheim, B. & Wang, Q. (2013). Data migration: A theoretical perspective. *Data and Knowledge Engineering*. 87. p.pp. 260–278.

Wand, Y. & Wang, R.Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*. 39 (11). p.pp. 86–95.

Wang, R.Y. (1998). A product perspective on total data quality management. *Communications of the ACM*. 41 (2).

Wang, R.Y., Lee, Y.W., Pipino, L.L. & Strong, D.M. (1998). Manage your information as a product. *Sloan Management Review*. 39 (4). p.pp. 95–105.

Wang, R.Y., Reddy, M. & Kon, H.B. (1995a). Toward quality data: An attribute-based approach. *Decision Support Systems*. 13. p.pp. 349–372.

Wang, R.Y., Storey, V.C.V.C. & Firth, C.P. (1995b). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*. 7 (4). p.pp. 623–640.

Wang, R.Y. & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*. 12 (4). p.pp. 5–34.

Winkler, W.E. (1999). The State of Record Linkage and Current Research Problems. *Statistical Research Division US Census Bureau*. p.pp. 1–15.

ANEXOS

ANEXO I - CATEGORIAS E DIMENSÕES DE QUALIDADE DE DADOS

Categoria de DQ	Dimensões de DQ
Intrínseca	Exatidão
	Objetividade
	Credibilidade
	Reputação
Contextual	Relevância
	Valor acrescentado
	Oportunidade
	Completude
	Quantidade apropriada
Representacional	Interpretabilidade
	Compreensibilidade
	Representação concisa
	Representação consistente
Acessibilidade	Acessibilidade
	Segurança
	Operacionalidade

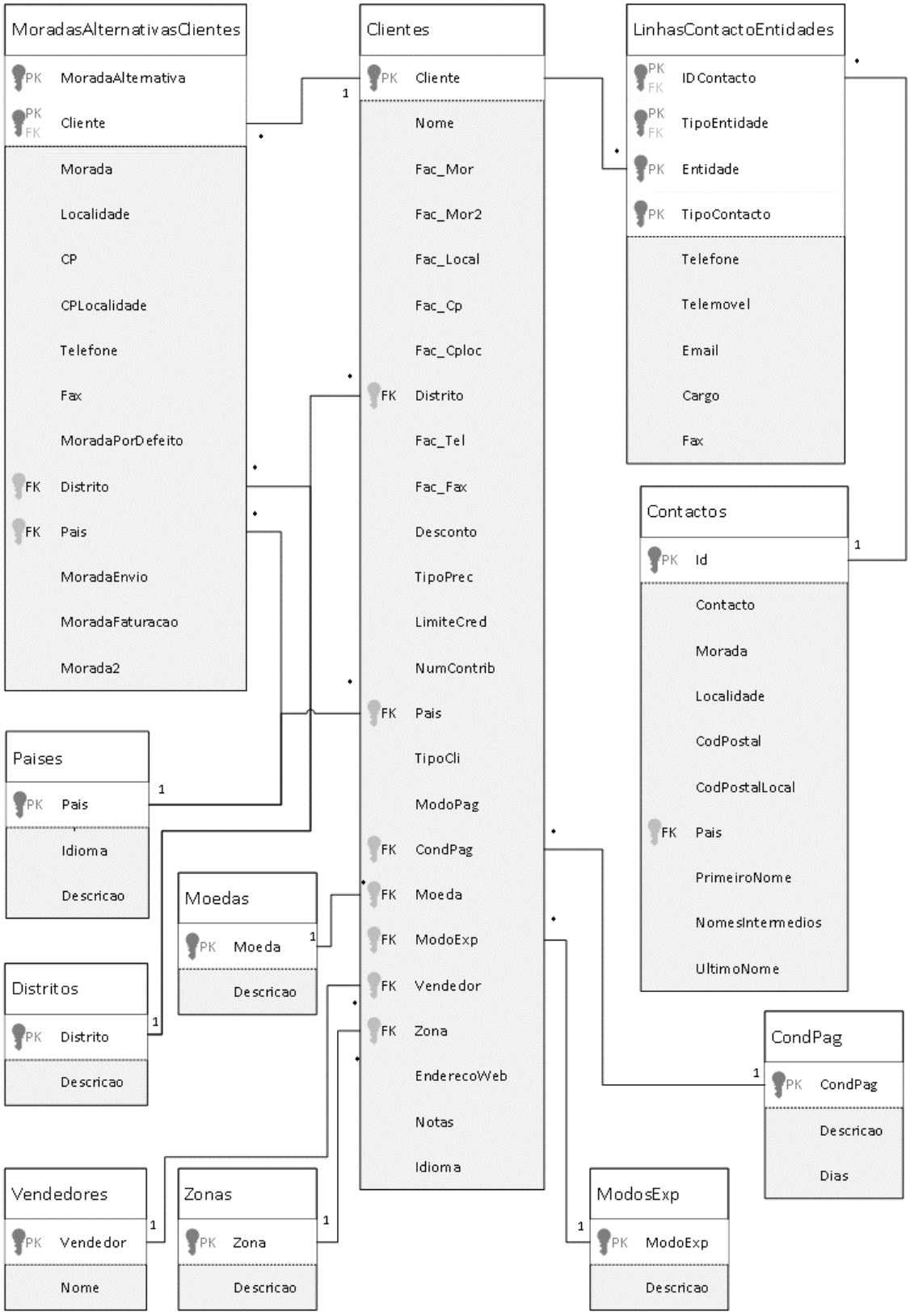
(Strong et al., 1997, p.104).

ANEXO II – TAXONOMIA DE PROBLEMAS DE QUALIDADE DE DADOS

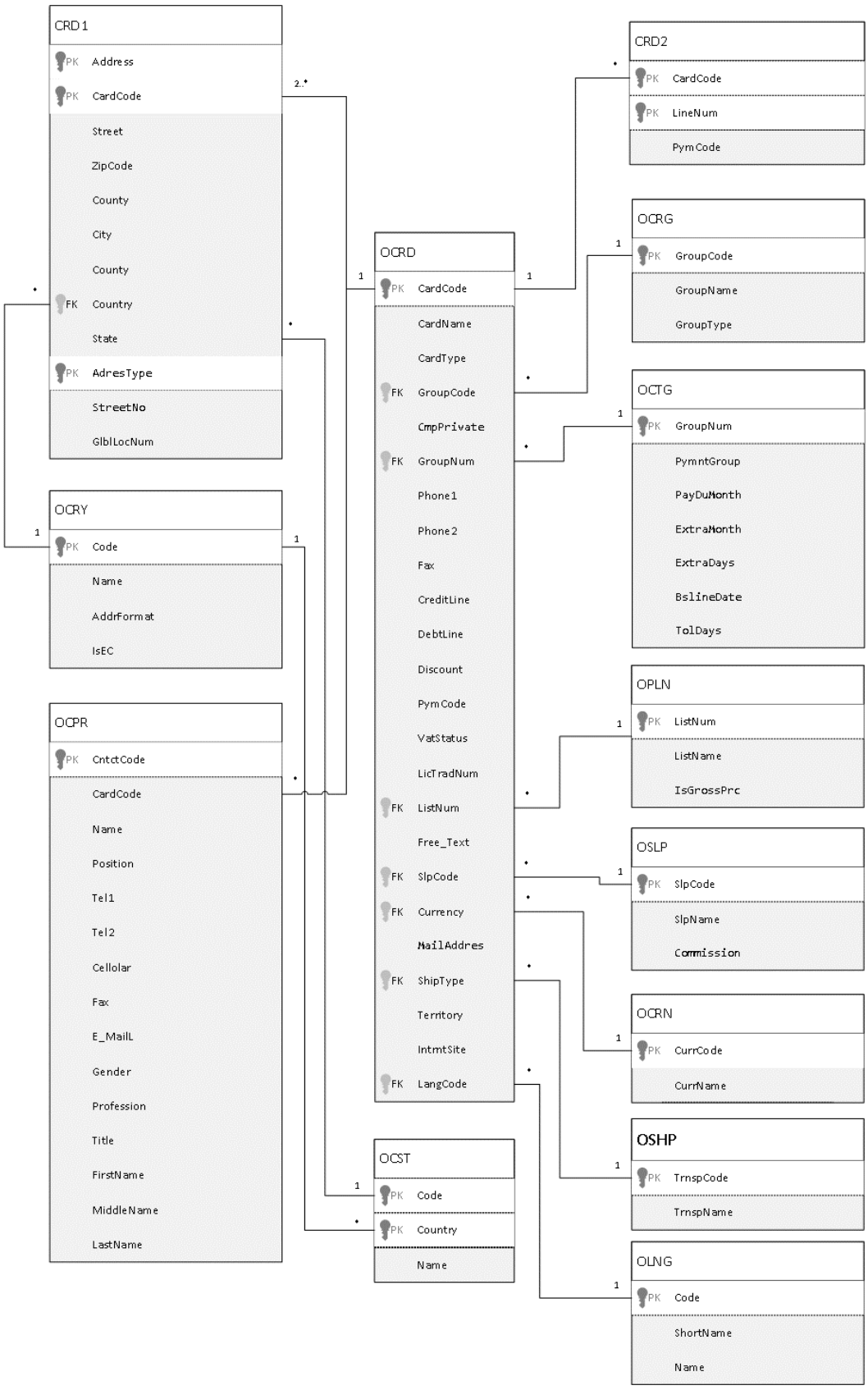
Nível de DQP	DQP
O valor de um atributo de um único tuplo	Valor em falta
	Violação da sintaxe
	Valor desatualizado
	Violação do intervalo
	Violação de domínio
	Erro ortográfico
	Valor inadequado para o contexto
	Valor para além do pretendido
	Valor sem significado
	Valor impreciso ou com vários significados
	Violação de restrição de domínio
Os valores de um único atributo	Violação de valor único
	Existência de sinónimos
	Violação de restrição de domínio
Os valores dos atributos de um único tuplo	Tuplo parcialmente vazio
	Inconsistência entre valores dos atributos
	Violação de restrição
Os valores dos atributos de vários tuplos	Redundância entre entidades
	Inconsistência entre entidades
	Violação de restrição de domínio
Relacionamentos entre múltiplas relações	Violação de dependência funcional
	Referência desatualizada
	Inconsistência da sintaxe
	Inconsistência entre atributos relacionados
	Referências circulares entre tuplos
Múltiplas fontes de dados	Inconsistência da sintaxe
	Diferentes unidades de medida
	Representação de dados inconsistente
	Diferentes níveis de agregação de dados
	Existência de sinónimos
	Existência de homónimos
	Redundância entre entidades
	Inconsistência entre entidades
	Violação de restrição

Adaptado de Oliveira & Rodrigues (2005)

ANEXO III – SEGMENTO DO ESQUEMA DE TABELAS DA DB PBS



ANEXO IV – SEGMENTO DO ESQUEMA DE TABELAS RELACIONAMENTO DA DB SBO



ANEXO V – MAPEAMENTO DE TABELAS ENTRE AS DBS PBS_STAGING E SBO_ STAGING

Tabela PBS	Tabela SBO	Observações
CondPag	OCTG	Objeto da tabela de destino: *PaymentTerms*
Contactos	OCPR	Objeto da tabela de destino: *ContactPersons.* O mapeamento das duas tabelas da DB PBS_Staging é realizado para uma única tabela na DB SBO_Staging. O tipo de dados do atributo *SexoMasculino* da DB PBS_Staging é diferente do tipo de dados do atributo *Gender* da DB SBO_Staging. Pela análise do *profiling* dos dados origem foi possível desenhar a regra de transformação de dados necessária para a DM.
LinhasContacto_ Entidades		
Vendedores	OSLP	Objeto da tabela de destino: *SalesPerson.* O tipo de dados da chave primária da tabela da DB PBS_Staging é diferente do tipo de dados da tabela da DB SBO_Staging. Na DB SBO_Staging a chave primária é uma *surrogate key*, por isso foi adicionado um atributo a esta tabela para guardar o código do vendedor da DB original.
Moedas	OCRN	Objeto da tabela de destino: *CurrencyCodes*
ModosExp	OSHP	Objeto da tabela de destino: *DeliveyTypes* Esta tabela foi objeto de um tratamento idêntico à tabela *Vendedores*
Distritos	OCST	Objeto da tabela de destino: *States.* A chave primária da tabela da DB SBO_Staging é composta por dois atributos (*Code* e *Country*), enquanto na tabela da DB PBS_Staging a chave primária tem um único atributo. Por esse motivo foi necessário adicionar a constante "PT" para preencher o atributo *Country* na DB SBO_Staging.
Clientes	OCRD	Objeto da tabela de destino: *BusinessPartners* e *Addresses.* A tabela *OCRD* (*Business Partners*) guarda os dados do cliente enquanto a tabela *CRD1* (*Business Partners - Addresses*) guarda os dados da morada de faturação e das moradas de entrega, pelo que foi necessário duplicar os registos das moradas. O tipo de dados do atributo *Idioma* da DB PBS_Staging é diferente do tipo de dados do atributo *LangCode* da DB SBO_Staging, por isso, foi necessário obter os valores correspondentes da tabela OLNG da DB SBO_Staging.
	CRD1	

ANEXO VI – MAPEAMENTO DE ATRIBUTOS DA TABELA CLIENTES

Atributo	Tabela de destino	Atributo de destino
Cliente	OCRD	CardCode
Nome	OCRD	CardName
Fac-Mor	CRD1	Street
Fac_Mor2	CRD1	Street
Fac_Local	CRD1	County
Fac_Cp	CRD1	ZipCode
Fac_CpLco	CRD1	City
Distrito	CRD1	State
Fac_Tel	OCRD	Phone1
Fac_Fax	OCRD	Fax
Desconto	OCRD	Discount
TipoPrec	OCRD	ListNum
LimiteCred	OCRD	CreditLine
NumContrib	OCRD	LicTradNum
Pais	CRD1	Country
TipoCli	OCRD	VatStatus
ModoPag	OCRD	PymCode
CondPag	OCRD	GroupNum
Moeda	OCRD	Currency
ModoExp	OCRD	ShipType
Vendedor	OCRD	SlpCode
Zona	OCRD	Territory
EnderecoWeb	OCRD	IntrntSite
Notas	OCRD	Remarks
Idioma	OCRD	LangCode

ANEXO VII – LISTA DE VALIDAÇÕES DE DADOS A IMPLEMENTAR

Tabela	Atributo	Taxonomia	Regras	Ação
OCRD	*LicTradNum*	Violação de sintaxe	O NIF de sujeitos passivos coletivos deve ter como prefixo o código ISO do país. Por exemplo todos os números portugueses devem começar por "PT".	Erro
		Violação de domínio	O dígito de controlo para os NIF de sujeitos passivos portugueses deve respeitar o algoritmo "Módulo 11".	Erro
		Violação de valor único	O NIF deve ser único para cada cliente.	Alerta
		Valor em falta	O NIF de sujeitos passivos coletivos é de preenchimento obrigatório.	Erro
CRD1	*ZipCode*	Violação de sintaxe	O código postal de endereços portugueses deve ter o formato AAAA-AAA. Sendo "A" um inteiro de 0 a 9.	Erro
		Violação de domínio	O código postal de endereços portugueses deve pertencer ao domínio definido por uma tabela de códigos postais ainda a definir.	Erro
		Inconsistência entre valores dos atributos	O código postal de endereços portugueses deve validar os atributos disponibilizados pela granularidade da sua definição. Por exemplo, o código postal 1700-131 permite identificar a rua, enquanto o código postal 2440-210 só permite identificar a localidade.	Alerta
OCRD e CRD1		Violação de valor único	A entidade *Cliente* deve ser verificada quanto à possibilidade da existência de entidades duplicadas utilizando os atributos *CardName* da tabela OCRD e *ZipCode*, *City* e *Street* da tabela CRD1.	Alerta