



Universidade do Minho
Escola de Engenharia

Agostinho Filipe Fernandes Salgado

REPOSITÓRIO GENEALÓGICO NACIONAL:
INTEGRAÇÃO E CONSOLIDAÇÃO DE DADOS

REPOSITÓRIO GENEALÓGICO NACIONAL:
INTEGRAÇÃO E CONSOLIDAÇÃO DE DADOS

Agostinho Filipe Fernandes Salgado

UMinho | 2016

outubro de 2016



Universidade do Minho
Escola de Engenharia

Agostinho Filipe Fernandes Salgado

REPOSITÓRIO GENEALÓGICO NACIONAL:
INTEGRAÇÃO E CONSOLIDAÇÃO DE DADOS

Dissertação de Mestrado
Ciclo de Estudos Integrados Conducentes ao Grau de
Mestre em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação da
Professora Doutora Maribel Yasmina Santos

AGRADECIMENTOS

Após a conclusão da presente dissertação, gostaria de agradecer a todas as pessoas que sempre me apoiaram e acompanharam e que tornaram possível a sua concretização.

Agradeço à professora Maribel Santos, o apoio, a paciência e a disponibilidade demonstradas, bem como o conhecimento partilhado e a simpatia com que sempre me recebeu.

Quero deixar também um agradecimento à professora Norberta Amorim e ao professor Antero Ferreira pela constante motivação, acompanhamento e dedicação demonstrados em todo este projeto.

Aos colegas de curso, em especial à Ana Pereira, companheira incansável em todo este percurso académico.

À minha família pelo constante incentivo e pelas horas de atenção que tive de lhes subtrair.

Ao Filipe, pelo apoio constante, desde a preparação para o meu ingresso no percurso académico até à conclusão da presente etapa.

RESUMO

Desde há cerca de vinte anos que o Grupo de História das Populações (GHP) tem vindo a desenvolver, na Universidade do Minho, uma linha de investigação centrada no estudo de comunidades históricas numa perspetiva micro analítica. Estes trabalhos assentam em bases de dados paroquiais (BDP) constituídas a partir dos registos paroquiais (RP) de batismo, casamento e óbito, organizadas segundo uma metodologia desenvolvida por Maria Norberta Amorim (Amorim, 1991), que permite acompanhar o percurso de vida de cada residente da comunidade em encadeamento genealógico. Atualmente, estas BD, que se encontram isoladas - existe uma BD por paróquia - reúnem mais de 1 milhão de registos de indivíduos, com uma representação geográfica concentrada principalmente no Norte de Portugal e em duas ilhas do arquipélago dos Açores, para além de núcleos de menor dimensão nos distritos do Porto, Aveiro, Lisboa e Évora. Este volume de informação exige a concretização de um sistema centralizado que reúna os dados das diferentes comunidades e que possibilite ao investigador acompanhar o percurso dos indivíduos em áreas geográficas mais alargadas. Esta necessidade acentua-se nos estudos sobre espaços urbanos ao longo de vários séculos, considerando a elevada mobilidade dos indivíduos e das famílias.

Para a concretização deste sistema centralizado torna-se necessário proceder à integração dos dados das diversas bases de dados locais numa base de dados central (BDC) que, com um modelo de dados unificado, permita a integração, consolidação e análise dos dados disponíveis e a reconstituição, por exemplo, de genealogias familiares.

Na presente dissertação analisou-se, em primeira instância, o modelo de dados da BDP, tendo-se averiguado junto dos investigadores do GHP, as limitações que o mesmo apresenta. Com base na informação recolhida, estudou-se, propôs-se e implementou-se a BDC, cujo modelo de dados detém a capacidade de, por um lado, suprimir as limitações identificadas e, por outro, corresponder aos requisitos que a fusão das BDP exige. Idealizou-se e implementou-se, ainda, um conjunto de processos de extração, transformação e carregamento de dados, capaz de, em primeiro lugar, avaliar e tratar das inconsistências dos dados presentes em cada uma das BDP, procedendo depois às transformações de entidades e dados necessárias, para que correspondam aos formatos definidos na BDC. Estes processos realizam, de seguida, o carregamento dos dados para a BDC, garantindo a preservação de todos os registos e os atributos consistentes, presentes em cada uma das BDP.

Criou-se ainda uma funcionalidade para a deteção de possíveis registos de indivíduos duplicados, ajustada ao presente contexto de dados e às necessidades do GHP que se revelou de elevada eficácia. A combinação destes elementos resulta na concretização da BDC e de um conjunto de procedimentos capazes de integrar e fundir cada uma das BDP para este repositório único, conforme o desejado pelos investigadores do GHP, para o desenvolvimento de pesquisas e análises mais abrangentes, possíveis apenas com esta realidade.

Palavras-Chave: Fusão de dados, Integração de dados, Qualidade de dados, *Record Linkage*, Demografia Histórica

ABSTRACT

For about twenty years, the Grupo de História das Populações (GHP) has been developing at the University of Minho a line of research focused on the study of historical communities in a micro-analytic perspective. The works developed from these investigations are based on parochial databases (PDB) built from parish registers (PR) of baptism, marriage and death. The organization of these data bases follows a methodology that was developed by Maria Norberta Amorim (Amorim, 1991). This approach allows to track the life path of each resident of a certain community with genealogical linkage. Currently, there are more than 1 million individuals in isolated databases (there is one data base for each parish), with a geographical representation mainly from the North of Portugal and two islands of the Azores archipelago. Other nucleus of a smaller dimension from the districts of Oporto, Aveiro, Lisbon and Évora are as well represented. This volume of information requires the creation of a central system able to gather data from different communities and to enable the researcher to follow the life path of the individuals, in wider geographical areas. This need is more noticeable in studies about urban areas over the centuries that comprises the high mobility from families and individuals.

For the implementation of this centralized system it is necessary to integrate data from the multiple local databases in a central database (CDB) that, with a unified data model, allows the integration, consolidation and analysis of available data and the reconstruction, for example, of family genealogies.

In this dissertation is has been studied, on the first place, the data model of the PDB. Also, the GHP researchers have been inquired about the limitations of this model. Based on the collected information, the CDB has been studied, proposed and implemented, with a data model that has the capacity to, on one hand, eliminate the identified limitations and, on the other hand, satisfy to the requirements that the merge of the PDB demands. A set of processes of extraction, transformation and loading of data, capable of, firstly, assess and deal with the inconsistencies of the existing data in each one of the PDB, proceeding then to the necessary transformations of the entities and the data, in order to match the formats defined in the CDB, have been conceived and implemented. In the subsequent phase, these processes load the resulting data to the BDC, guaranteeing the preservation of all the consistent records and attributes in each one of the PDB.

Also, it has been developed a functionality for the detection of possible duplicate records, adjusted to the present data context and to the needs of the GHP which has proved to be of high efficiency.

The combination of these elements results in the implementation of CDB and of a set of procedures able to integrate and merge each one of the PDB to this central repository, as sought by investigators of the GHP, for the development of more comprehensive research and analyses, possible only on this new reality.

KEYWORDS: *Data Fusion, Data Integration, Data Quality, Record Linkage, Historical demography*

ÍNDICE

Agradecimentos.....	v
Resumo.....	vii
Abstract.....	ix
Lista de Figuras.....	xv
Lista de Tabelas	xvii
Lista de Abreviaturas, Siglas e Acrónimos	xix
1 Introdução	1
1.1 Enquadramento e Motivação	1
1.2 Finalidade e Objetivos.....	4
1.3 Abordagem Metodológica	5
1.4 Organização do Documento.....	6
2 Enquadramento Conceptual e Tecnológico	9
2.1 Integração de Dados	9
2.1.1 Objetivos da Integração de Dados	10
2.1.2 Desafios e Passos da Integração de Dados.....	13
2.1.2.1 Mapeamento de Esquemas de Dados.....	14
2.1.2.2 Detecção de Duplicados.....	16
2.1.2.3 Fusão de Dados.....	18
2.2 Qualidade dos Dados	23
2.2.1 Dimensões de Qualidade dos Dados	24
2.2.1.1 Precisão	24
2.2.1.2 Completude	26
2.2.1.3 Consistência	27
2.2.1.4 Tempo.....	27
2.3 <i>Record Linkage</i>	28
2.3.1 A aplicação de <i>Record Linkage</i>	30
2.3.2 Desafios de <i>Record Linkage</i> em Genealogia.....	31
2.3.3 <i>Record Linkage</i> e Ficheiros Genealógicos	33

2.4	<i>Extract, Transform and Load</i>	34
2.5	Tecnologias Consideradas.....	43
3	O Sistema de Reconstituição de Paróquias e a Sua Base de Dados	49
3.1	O Sistema de Reconstituição de Paróquias	49
3.1.1	O Modelo de Dados do Sistema de Reconstituição de Paróquias	51
3.1.2	Descrição das Principais Tabelas do Sistema de Reconstituição de Paróquias	53
3.1.3	Limitações do Sistema de Reconstituição de Paróquias.....	59
3.2	Base de Dados Central Para o Repositório Genealógico Nacional	61
3.2.1	O Modelo de Dados Para a Base de Dados do Repositório Genealógico Nacional.....	62
3.2.2	Principais Tabelas da Base de Dados do Repositório Genealógico Nacional	64
4	A Integração e Consolidação de Dados no Repositório Genealógico Nacional	71
4.1	Avaliação da Qualidade dos Dados.....	73
4.2	Estratégias de Limpeza e Tratamento dos Dados.....	75
4.2.1	Valores Admissíveis no Contexto da Base de Dados do Sistema de Reconstituição de Paróquias.....	75
4.2.2	Tratamento dos Nomes dos Indivíduos	78
4.3	A Integração dos Dados no Repositório Genealógico Nacional	79
4.3.1	A Extração dos Dados.....	82
4.3.2	A Limpeza de Dados	84
4.3.3	Mapeamento (Transformação) e Carregamento para o Repositório Genealógico Nacional	90
4.3.3.1	Refreshamento das Tabelas Auxiliares.....	90
4.3.3.2	Carregamentos das Entidades Principais e dos Registos Associados.....	93
4.3.4	A Detecção de Duplicados.....	100
5	Conclusões	109
5.1	Síntese	109
5.2	Resultados.....	110
5.3	Contribuições.....	110
5.4	Trabalho Futuro	111

Referências bibliográficas	113
Anexos	117
Diagrama de Entidades e Relacionamentos da Base de Dados Paroquial	118
Diagrama de Entidades e Relacionamentos da Base de Dados Central	119
Modelação <i>Entity Mapping Diagram</i>	122
Primeiros 100 resultados da comparação da função SQL	126
Primeiros 100 resultados da comparação R- RECORDLINKAGE	139

LISTA DE FIGURAS

Figura 1 - Design Science Research Methodology (DSRM) Process Model	5
Figura 2 - Integração de fontes de dados.....	11
Figura 3 - Resultados possíveis de combinação de dados	12
Figura 4 - Combinação de dados com completude <i>intencional</i> e <i>extensional</i>	12
Figura 5 - Os três passos da integração de dados.....	14
Figura 6 - Dois esquemas para serem mapeados	15
Figura 7 - <i>Schema matching</i>	15
Figura 8 - Tabelas com tabelas relacionadas e com atributos correspondentes	17
Figura 9 - Tabelas com correspondência de atributos e duplicados detetados	19
Figura 10 - Classificação de estratégias para a resolução de inconsistências nos dados	21
Figura 11 - O ambiente dos processo de ETL	35
Figura 12 - A <i>framework</i> genérica de EMD	37
Figura 13 - Metamodelo da EMD.....	38
Figura 14 - Tipos de transformações na EMD.....	39
Figura 15 - Construtores gráficos da EMD	40
Figura 16 - Esquema relacional da DS1	40
Figura 17 - Esquema relacional da DS2	41
Figura 18 - Esquema em estrela para o DW1	41
Figura 19 - Cenário EMD para a dimensão Produto	42
Figura 20 - Classificação de Empresas fornecedores de soluções de DI	44
Figura 21 - Quadro de ponderação dos parâmetros das empresas.....	45
Figura 22 - Quadro comparativo de características de produto	47
Figura 23 - Interface INDIVÍDUO no SRP	49
Figura 24 - Interface FAMÍLIA no SRP.....	50
Figura 25 - Diagrama de Entidades e Relacionamentos da BD do SRP	52
Figura 26 - Diagrama de Entidades e Relacionamentos do novo modelo de dados	63
Figura 27 - Perspetiva global da integração e consolidação de dados no RGN	72
Figura 28 - Exemplo de <i>script</i> para a identificação de erros nos dados.....	74
Figura 29 - Exemplo de <i>script</i> para a remoção de registos inconsistentes	75

Figura 30 - O domínio "Sexo" no SSDQS.....	76
Figura 31 - Modelação EMD de alto nível.....	80
Figura 32 - Cenário EMD para a entidade Indivíduo	81
Figura 33 - O fluxo da integração de dados nos SSIS	82
Figura 34 - Fluxo de dados da BDP para a BD SRP_STAGE	83
Figura 35 - Fluxo de dados da BD SRP_STAGE para IMPORTED_SRP.....	84
Figura 36 – Fluxo de dados para a limpeza da tabela Assinaturas	85
Figura 37 - Tarefa de limpeza de dados da tabela Assinatura.....	86
Figura 38 - Fluxo de dados para a limpeza da tabela INDIVIDUO	87
Figura 39 - <i>Script</i> para a separação dos elementos do nome	87
Figura 40 – Tratamento das partes do nome.....	88
Figura 41 - Script para a agregação dos elementos do nome	89
Figura 42 - Package 3.....	90
Figura 43 - Fluxo para o refrescamento da tabela LOCAL.....	91
Figura 44 - Extração dos sítios com associação do lugar.....	92
Figura 45 - O fluxo de refrescamento da tabela PROFISSAO	92
Figura 46 - Fluxo para o carregamento da tabela INDIVIDUO	93
Figura 47 - Comando SQL para leitura da tabela INDIVIDUO.....	94
Figura 48 - <i>Stored Procedure</i> para obtenção do local.....	95
Figura 49 - Fluxo para o carregamento da tabela ASSINATURAS.....	96
Figura 50 - Fluxo para o carregamento da tabela FAMILIA	97
Figura 51 - Fluxo para o carregamento da tabela CASAMENTO	98
Figura 52 - Fluxo para o carregamento das tabelas RESIDENCIA	99
Figura 53 - Resultados do ETL	100
Figura 54 - Resultados do primeiro ensaio R - RecordLinkage	101
Figura 55 - Resultados do segundo ensaio R - RecordLinkage	102
Figura 56 - Consulta para seleção de atributos para deteção de duplicados.	104
Figura 57 - Função para a validação das datas dos eventos do indivíduo.....	105
Figura 58 - Consulta para a avaliação de duplicados	106

LISTA DE TABELAS

Tabela 1 - Comparação de resultados de algoritmos de cálculo similaridade de <i>Strings</i>	18
Tabela 2 - Estratégias de resolução de conflitos.....	20
Tabela 3 - Funções de resolução de conflitos	22
Tabela 4 - Relação de filmes	25
Tabela 5 - Definições de completude.....	26
Tabela 6 - Tabela INDIVIDUO na BDP.....	54
Tabela 7 - Tabela FAMILIA na BDP.....	56
Tabela 8 - Tabela FAMILIAS na BDP.....	57
Tabela 9 - Tabela INFOCOMPLEMENTAR na BDP.....	57
Tabela 10 - Tabela CONCELHO na BDP	58
Tabela 11 - Tabela PAROQUIA na BDP	58
Tabela 12 - Tabela RESIDENCIA no SRP	59
Tabela 13 - Tabela INDIVIDUO na BDC	64
Tabela 14 - Tabela FAMILIA na BDC.....	66
Tabela 15 - Tabela CASAMENTO na BDC	67
Tabela 16 - Tabela ALCUNHA na BDC.....	69
Tabela 17 - Tabela RESIDENCIAINDIVIDUO na BDC	69
Tabela 18 - Resultado da validação do Domínio "NomeProprio"	79
Tabela 19 - Resultados comparação em SQL	107
Tabela 20 - Resultados comparação na ferramenta "R-RECORDLINKAGE"	108

LISTA DE ABREVIATURAS, SIGLAS E ACRÓNIMOS

Neste documento são utilizadas um conjunto de siglas, apresentadas de seguida.

BD	Base de Dados
BDC	Base de Dados Central
BDP	Base de Dados Paroquial
BI	<i>Business Intelligence</i>
CLR	<i>Common Language Runtime</i>
DQ	<i>Data Quality</i>
DW	<i>Data Warehouse</i>
ETL	<i>Extract, Transform, Load</i>
FD	Fusão de Dados
GHP	Grupo de História das Populações
GPL	<i>General Public License</i>
ID	Integração de Dados
MRP	Metodologia de Reconstituição de Paróquias
MS	<i>Microsoft</i>
RGN	Repositório Genealógico Nacional
RL	<i>Record Linkage</i>
RP	Registos Paroquiais
SEED	Sistema para o Estudo da Evolução Demográfica
SQL	<i>Structured Query Language</i>
SRP	Sistema de Reconstituição de Paróquias
SSDQS	<i>SQL Server Data Quality Services</i>
SSIS	<i>SQL Server Integration Services</i>
SSMS	<i>SQL Server Management Studio</i>
TI	Tecnologias da Informação
UML	<i>Unified Modeling Language</i>
XML	<i>eXtensible Markup Language</i>

1 INTRODUÇÃO

Este capítulo apresenta o enquadramento e a motivação para a realização da presente dissertação, bem como a finalidade e os objetivos da mesma, introduzindo-se ainda o tema da Demografia Histórica. Referem-se aqui, também, a metodologia de investigação a seguir, bem como a metodologia de implementação adotada. No último ponto deste capítulo apresenta-se a estrutura do documento.

1.1 Enquadramento e Motivação

O Grupo de Investigação «História das Populações»¹ (GHP), com uma história com mais de 20 anos, constitui-se em 2007 como componente do Centro de Investigação Transdisciplinar "Cultura, Espaço e Memória" (CITCEM). Este grupo tem como objetivo o “desenvolvimento de projetos de investigação interdisciplinar a partir das informações sistematizadas em bases demográficas construídas por aplicação de métodos e técnicas da Demografia Histórica(...)”¹, e integra atualmente 37 investigadores de diversas áreas de conhecimento que vão desde a História, Demografia Histórica, Sociologia, Ciências e Tecnologias da Computação, Antropologia Biológica, às Ciências Sociais.

O GHP tem vindo a desenvolver, na Universidade do Minho, uma linha de investigação centrada no estudo de comunidades históricas numa perspetiva micro analítica. Estes trabalhos assentam em bases de dados (BD) constituídas a partir dos registos paroquiais (RP) de batismo, casamento e óbito, organizadas segundo a Metodologia de Reconstituição de Paróquias (MRP), desenvolvida por Norberta Amorim (Amorim, 1991), que permite acompanhar o percurso de vida de cada residente da comunidade em encadeamento genealógico.

Em 1997, o GHP, na altura com a designação de Núcleo de Estudos da População e Sociedade (NEPS), em colaboração com Departamento de Informática da Universidade do Minho, criam o projeto SEED - Sistema para o Estudo da Evolução Demográfica², para apoio aos estudos em Demografia Histórica, suportados pelo Método de Reconstituição de Paróquias (MRP). No âmbito deste projeto, é desenvolvida a aplicação Sistema de Reconstituição de Paróquias – SRP, desenvolvida no contexto do Módulo de

¹ <http://www.ghp.ics.uminho.pt/>

² <http://www4.di.uminho.pt/~gepl/SEED/index.html>

Aquisição de Dados do sistema SEED, para suportar o processo de alimentação de Bases de Dados Paroquiais (BDP) a partir de RP.

Da utilização do SRP, a aplicação ainda em uso para a recolha de dados, resulta uma base de dados em formato *Microsoft Access* por cada paróquia cujos RP foram recolhidos por esta aplicação. Atualmente, estas BD, reúnem mais de 1 milhão de indivíduos, com uma representação geográfica concentrada principalmente no Norte de Portugal e em duas ilhas do arquipélago dos Açores, para além de núcleos de menor dimensão nos distritos do Porto, Aveiro, Lisboa e Évora.

Com este volume de informação, o grupo pretende a criação de um sistema centralizado que reúna os dados das diferentes comunidades (atualmente em BD diferentes) e que possibilite ao investigador acompanhar o percurso dos indivíduos em áreas geográficas mais alargadas. Esta necessidade acentua-se nos estudos sobre espaços urbanos ao longo de vários séculos, considerando a elevada mobilidade dos indivíduos e das famílias.

Pretende-se, agora, criar o Repositório Genealógico Nacional (RGN) com uma Base de Dados Central (BDC) alargada ao espaço nacional, que possibilite o estudo de diferentes indicadores demográficos (fecundidade, nupcialidade, mortalidade e mobilidade). Contudo, para a realização deste projeto, torna-se necessária a resolução de várias dificuldades que vão desde a fusão das várias BDP, à criação de um novo interface de introdução de dados, com possibilidades acesso multiutilizador e multiposto, passando ainda pela extração de informação e construção de genealogias.

A presente dissertação assenta na resolução do primeiro problema apresentado, ou seja, na integração e fusão das várias BDP numa BDC, com capacidade para acessos múltiplos em concorrência, e na resolução de todos os problemas e dificuldades que uma tarefa desta envergadura apresenta.

Demografia Histórica

A demografia histórica nasce na década de 50 quando *Louis Henry* desenvolve uma metodologia que permitia o estudo de um tipo particular de fontes, os registos paroquiais, existentes na generalidade dos países europeus a partir do século XIV (Ferreira, 2002). A demografia, que a *Infopédia* define como o “estudo das populações humanas, particularmente a sua densidade, volume, distribuição e estatísticas básicas (nascimentos, casamentos, doenças, mortes, etc.) ao longo de um dado período”¹, vê assim o

¹ <http://www.infopedia.pt/dicionarios/lingua-portuguesa/demografia>

seu âmbito de análise alargado para períodos anteriores aos recenseamentos modernos abrindo caminho a um melhor conhecimento passado, especialmente do homem comum (Ferreira, 2004).

O facto de os RP existirem em quase todos os países cristãos, com características idênticas, possibilitou a aplicação das mesmas metodologias, com resultados comparáveis, resultando numa multiplicação de estudos por toda a Europa. Contudo, conforme refere Ferreira em (Ferreira, 2002), os RP não apresentam a mesma qualidade em todos os países, o que originou metodologias alternativas, adaptadas às especificidades dos mesmos.

Demografia Histórica em Portugal

Em Portugal, Norberta Amorim desenvolveu nos anos 70 a MRP, ajustada às fontes portuguesas, que permite acompanhar em encadeamento genealógico o percurso de vida de cada indivíduo (Ferreira, 2004). Este método, na altura manual, teve consequências académicas em 1971, permitindo uma primeira informação para o país sobre variáveis demográficas do Antigo Regime (Amorim & Ferreira, 2006). O recurso à informática para suporte à metodologia surge em 1986 com o desenvolvimento de uma aplicação em *dBaseIII Plus*, em colaboração com investigadores do Departamento de Informática da Universidade do Minho: Luís Lima, Cecília Moreira e Pedro Henriques. A aplicação teve uma produtividade tão significativa que passou a ser a base de um vasto conjunto de trabalhos de investigação (Ferreira, 2004). O recurso a estas novas tecnologias, conforme refere o mesmo autor, veio alargar os horizontes temporais e espaciais da investigação. Norberta Amorim afirma “*Não só era possível atingir circunscrições mais vastas do que a freguesia, como era possível acompanhar em muito longa duração a sucessão das gerações, sem detença no século XIX. Era possível chegar ao presente e acompanhar os mais significativos ritmos de mudança. Era possível construir em muito longa duração bases de dados com percursos individuais em encadeamento genealógico.*”¹

Com a introdução de uma ferramenta informática facilitadora dos trabalhos destes investigadores, surgiu naturalmente a tentação de cruzar outras fontes nominais, aliando às fontes religiosas (os RP) outros tipos de fontes, tais como fontes fiscais, militares, judiciais, entre outras. No entanto, aqui começam a evidenciar-se as limitações do trabalho em *dBaseIII Plus*. Perante bases de dados cada vez mais volumosas, as pesquisas apresentavam muita lentidão e o facto de a implementação feita nesta

¹ Amorim, Norberta, “Da Genealogia à História da Família. O contributo da Demografia Histórica.”, comunicação no Encontro de Genealogia realizado em Abril de 2002, em Lisboa.

tecnologia não possuir um relacionamento automático entre as várias tabelas conduzia, invariavelmente, a situações de redundância e de inconsistência de dados. Daqui, surgiu a necessidade de avançar para uma tecnologia mais atualizada que possibilitasse a resolução destes e de outros constrangimentos, fazendo-se então a ligação ao projeto SEED, ao abrigo do qual foi construída uma nova aplicação: o SRP, ainda em uso pelo GHP.

O SRP apresenta-se como um forte aliado ao investigador na recolha de dados. Resolve questões como as dificuldades dos cruzamentos dos indivíduos e da criação de ligações entre os indivíduos familiares, pais e filhos, no entanto, dada a continuada investigação do GHP, surgem novas necessidades. Face ao crescente número de RP levantados, a aplicação começa a revelar alguns problemas de desempenho, apresentando uma navegação cada vez mais lenta entre os formulários indivíduos e de famílias. É também intenção do GHP poder cruzar a informação dos indivíduos e famílias presente nas várias paróquias onde originaram RP e que estará distribuída por BDP distintas. Pretende ainda que vários colaboradores possam trabalhar simultaneamente a mesma comunidade.

1.2 Finalidade e Objetivos

A finalidade deste trabalho consiste na concretização de uma base de dados central (BDC) que integre, consolide e funda todas as BDP existentes e que ultrapasse as limitações identificadas no modelo das BDP, tais como, a impossibilidade de registar algumas informações presentes nos RP ou o acesso multiutilizador em simultâneo.

O grande desafio deste projeto está na identificação dos indivíduos com registo em BDP distintas, detendo, em cada uma delas, informação tendencialmente complementar. Este processo torna-se complexo pelo facto de a tradição portuguesa não definir regras de transmissão de apelidos (o indivíduo pode receber apenas o sobrenome do pai ou da mãe...). Por outro lado, um indivíduo aparece muito frequentemente com referência incompleta ao seu nome no registo do seu batismo normalmente com indicação apenas do primeiro nome (“João”), surgindo referência ao nome (mais) completo nos registos de casamento, óbito, ou nos de batismo dos filhos que vier a ter (“João Silva Pereira”). No caso das mulheres esta questão agrava-se ainda mais com a adoção, ou não, do(s) apelido(s) do marido no ato do casamento.

São então objetivos deste projeto:

1. Modelar uma base de dados central (BDC) capaz de reunir os dados de todas as BDP, criadas pelo SRP, existentes e capaz de dar resposta às limitações identificadas
2. Desenhar os processos de ETL para a integração das BDP na BDC
3. Criar rotinas de integração e validação de dados com deteção de registos duplicados
4. Validar as rotinas de integração de dados implementadas

1.3 Abordagem Metodológica

Para o desenvolvimento da presente dissertação será adotada a metodologia “*A Design Science Research Methodology for Information Systems Research*” (Peppers et al., 2008), ilustrada na Figura 1, dadas a sua adequação a projetos da área de Sistemas de Informação, a sua flexibilidade, possibilitando diversos pontos de entrada na investigação, e as definições claras das etapas e dos resultados do processo que apresenta.

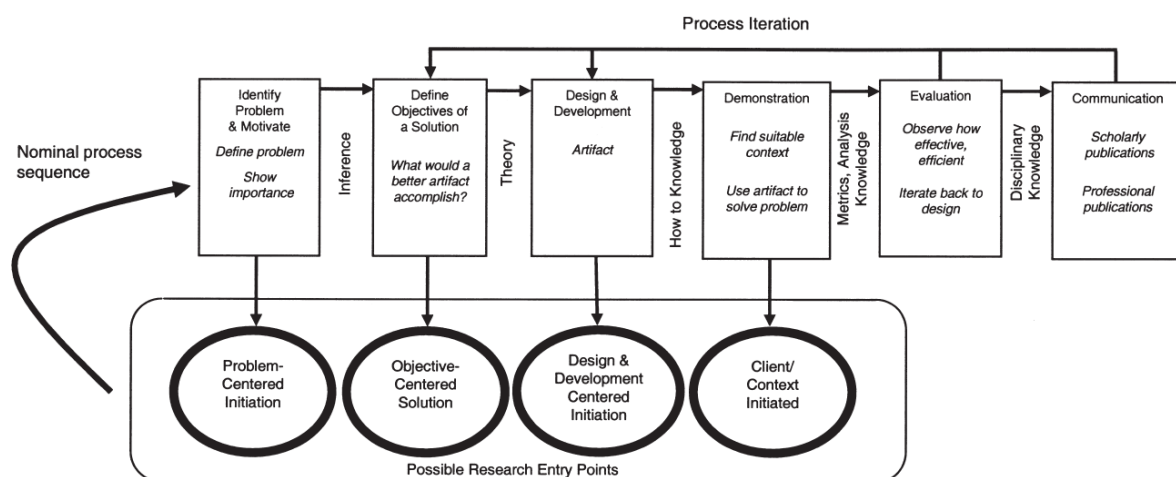


Figura 1 - Design Science Research Methodology (DSRM) Process Model – Retirado de (Peppers et al., 2008)

O desenvolvimento do presente projeto de dissertação iniciou-se com a elaboração do plano de trabalhos, num primeiro momento, tendo sido depois realizados os enquadramentos conceptual e tecnológico para uma ambientação da temática a ser tratada e das ferramentas disponíveis para tratarem o problema. Em seguida, deu-se início às fases da metodologia propriamente ditas, conforme breve descrição apresentada a seguir:

- **Identificação do problema** - Definição conceptual do problema em questão de modo a apreender a complexidade do mesmo e conseqüentemente, facilitar o desenvolvimento de uma solução adequada.

- **Definição de objetivos da solução** – Enumeração dos objetivos da solução para o problema, com base na definição do mesmo e no conhecimento relativamente ao que é exequível e praticável.
- **Conceção e desenvolvimento** - Criação de artefacto decorrente da investigação (desenho e implementação de uma solução que deverá, recorrendo a técnicas de integração de dados, fusão de dados, e de qualidade de dados, resolver o problema em tratamento).
- **Demonstração** - Experimentação e aplicação da solução desenvolvida.
- **Avaliação** - Observar e medir o nível de adequação da solução desenvolvida como resposta ao problema.
- **Comunicação** - Escrita da Dissertação e de artigos científicos.

Atendendo a que a presente dissertação contempla uma implementação de um sistema de integração de dados e que a mesma pode ser concretizada recorrendo a processos de ETL, serão ainda seguidas as orientações da metodologia *Entity Mapping Diagram* (EMD) desenvolvida por El-Sappagh et al. (2011). Esta metodologia foi selecionada uma vez que se apresenta como simples, sendo de fácil entendimento para o modelador do *Data Warehouse* (DW), completa, dado que permite a representação de todas as atividades dos processos de ETL, e flexível, uma vez que pode ser aplicada em ambientes de DW distintos. Por outro lado, conforme referem El-Sappagh et al. (2011), foi elaborada tendo por base metodologias anteriores, tendo esta sido enriquecida para suportar algumas lacunas encontradas. Esta metodologia está brevemente descrita no ponto 2.4.

1.4 Organização do Documento

Detalha-se, no presente documento, todo o trabalho realizado na presente dissertação, tendo-se estruturado o mesmo de acordo com as orientações do guia de dissertação disponibilizado pela coordenação do Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação.

O documento segue a seguinte estrutura:

- No capítulo 1 encontra-se a introdução onde se procede a um enquadramento e define a motivação para o problema a investigar, a finalidade e os objetivos da presente dissertação. Apresenta-se ainda uma contextualização da Demografia Histórica, área de negócio do cliente do presente projeto. De seguida expõe-se a abordagem metodológica a seguir, identificando-se ainda a metodologia de implementação adotada.

- No capítulo 2 procede-se ao enquadramento conceptual das temáticas abordadas na presente dissertação, bem como ao enquadramento tecnológico onde são avaliadas tecnologias existentes de relevo para o desenvolvimento deste projeto.
- No capítulo 3 procede-se a uma descrição da aplicação SRP e do modelo de dados da BDP que a suporta, identificando-se ainda aqui as limitações que o referido modelo apresenta. Apresenta-se também o modelo de dados da BDC para onde serão migradas todas as BDP.
- O capítulo 4 refere-se ao processo de integração e consolidação de dados no RGN, estando aqui detalhados todos os passos deste processo, desde a avaliação da qualidade dos dados, às transformações de dados efetuadas, bem como o processo de carregamento dos mesmos para a BDC, apresentando-se aqui, também, a avaliação à eficácia destes processos. No último ponto, detalha-se a funcionalidade de deteção de duplicados.
- Por último, no capítulo 5 apresentam-se os resultados obtidos assim como umas breves conclusões sobre o trabalho realizado.

2 ENQUADRAMENTO CONCEPTUAL E TECNOLÓGICO

Neste capítulo procede-se a um enquadramento conceptual dos temas associados à presente dissertação. Pretende-se aqui proporcionar um melhor conhecimento dos conceitos das temáticas de integração de dados, qualidade de dados, *Record Linkage* e dos processos de *Extract, Transform e Load* (ETL), que permitirão a concretização da integração dos dados.

Para a realização deste capítulo foi analisada literatura disponibilizada pela comunidade científica e técnica que trabalha estes mesmos temas, tendo a mesmo sido selecionada através da pesquisa em repositórios *online* de referência, acessíveis dados os protocolos da Universidade do Minho com estas instituições.

Foram então realizadas pesquisas pelas palavras-chave: *Data Integration, Data Fusion, Data Quality Record Lynkage, ETL*, tendo sido selecionada a literatura cujo título e respetivo resumo, se aproximam mais da temática em questão, priorizando-se sempre artigos com maior número de citações e a contemporaneidade dos mesmos.

Analisou-se ainda literatura recomendada e disponibilizada pelo GHP.

2.1 Integração de Dados

Conforme referem Brazhnik e Jones (2007), a era da produção de dados em que o volume de informação gerado impressionava qualquer audiência converteu-se na era da integração de dados (ID). Neste novo paradigma o foco não está nos desafios apresentados pelas questões técnicas da gestão de tais volumes de dados, mas no valor da informação que estes dados contêm. Para se conseguir a extração de informação e a geração de conhecimento esta profusão de dados necessita ser reunida e integrada. Qualquer tarefa de processamento de dados aspira produzir informação de confiança, confiança essa que poderá ser medida pelo quão bem a informação representa a realidade (Brazhnik & Jones, 2007). No *Website*¹ da IBM encontramos que “ID é a combinação de processos técnicos e de negócio utilizados para combinar dados de fontes díspares em informação significativa e de valor”.

Este processo é um desafio de grande complexidade, dado que terá de recolher dados a partir de um conjunto diverso de fontes autónomas e heterogéneas, mas é fundamental em organizações com vastas fontes de dados, em trabalhos científicos em que diferentes investigadores produzem conjuntos de dados

¹ <http://www.ibm.com/analytics/us/en/technology/data-integration>

diferentes e independentes, em agências governamentais (especialmente aquelas que tenham algum tipo de cooperação), entre outros contextos (Halevy & Ordille, 2006).

2.1.1 Objetivos da Integração de Dados

Um dos objetivos da ID pode ser definido como a agregação dos dados numa representação completa e concisa. A informação está **completa** se nenhum objeto é esquecido no resultado e é **concisa** se nenhum objeto está representado mais do que uma vez e a informação não contém contradições (Bleiholder & Naumann, 2008).

A **completude** (*completeness*) dos dados avalia a quantidade de informação disponível e pode ser dividida em duas dimensões:

- *extensional* - ao nível dos dados;
- *intencional* - ao nível do esquema de dados.

A completude *extensional* refere-se ao número de representações únicas de objetos em relação ao número de objetos do mundo real. Já a completude *intencional* refere-se ao número de atributos únicos de cada objeto (Bleiholder & Naumann, 2008).

A **concisão** refere-se à inexistência de dados redundantes pela fusão das entradas duplicadas e dos atributos comuns, num só. Divide-se também nas dimensões:

- *extensional* - ao nível dos dados (não contém objetos redundantes);
- *intencional* - ao nível do esquema (não contém atributos redundantes).

Para melhor compreensão destes conceitos atente-se no exemplo apresentado em (Bleiholder & Naumann, 2008) a seguir explicado. Na Figura 2 a fonte *Source S* apresenta três representações de objetos (1, 2, 3) identificados pelo atributo ID. Cada um destes objetos tem, nesta fonte, dois atributos (A e B). Na fonte *Source T* constam outras 3 representações de objetos (2, 3, 4) identificadas por um atributo também de designação ID. Estas representações têm também mais dois atributos (B e C). Nestas representações, o símbolo \perp refere-se a valores nulos para os atributos.

Source T			Source S		
B	C	ID	A	B	ID
n	m	2	x	y	1
k	⊥	3	z	⊥	2
⊥	m	4	⊥	x	3

Figura 2 – Integração de fontes de dados – Adaptado de (Bleiholder & Naumann, 2008)

Se se integrarem estes dados sem se realizar qualquer identificação dos objetos ou mapeamento de atributos obtemos a combinação (a) da Figura 3. Esta combinação apresenta uma completude elevada dado que não se perderam quaisquer objetos ou atributos. No entanto, é inconsistente, dado que os objetos com ID 2 e 3 estão duplicados, assim como o atributo B.

Se se conhecerem as correspondências de atributos e se realizar um mapeamento do esquema de dados, poderemos obter a combinação (b) da Figura 3. Neste caso, obtemos uma combinação com concisão *intencional*, dado que nenhuma propriedade (atributo) do mundo real está duplicada, no entanto, temos aqui objetos duplicados.

Se se conhecer e mapear algum identificador para os objetos, neste caso o ID, poderemos alcançar a combinação (c) da Figura 3. Aqui está mapeado o atributo ID (e só este), pelo que nenhum objeto se encontra duplicado. No entanto, temos uma duplicação do atributo B. Esta é uma situação de concisão *extensional*.

A combinação (d) da Figura 3 acontece depois de conhecidos e mapeados os objetos e os atributos comuns. Temos uma representação única por objeto do mundo real e todos os objetos representados, todos os atributos representados, representados uma única vez e com um valor único. Esta é uma combinação com concisão *intencional* e *extensional*, que é o objetivo último da integração e fusão de dados. No entanto, se atentarmos no atributo B do objeto com ID 3, verificamos que apresenta o valor “x”. Se consultarmos a fonte de dados T verificamos “k” é também um valor admissível para este atributo. Encontramo-nos perante uma situação de conflito, comum em tarefas de ID e que será analisada no ponto 2.1.2.3.1.

S.A	S.B	S.ID	T.ID	T.B	T.C
x	y	1	⊥	⊥	⊥
z	⊥	2	⊥	⊥	⊥
⊥	x	3	⊥	⊥	⊥
⊥	⊥	⊥	2	n	m
⊥	⊥	⊥	3	k	⊥
⊥	⊥	⊥	4	⊥	m

(a) Sem mapeamento e sem identificador de objeto

S.A	B	ID	T.C
x	y	1	⊥
z	⊥	2	⊥
⊥	x	3	⊥
⊥	n	2	m
⊥	k	3	⊥
⊥	⊥	4	m

(b) Mapeamento de (S.B ↔ T.B; S.ID ↔ T.ID) sem identificador de objeto

S.A	S.B	ID	T.B	T.C
x	y	1	⊥	⊥
z	⊥	2	n	m
⊥	x	3	k	⊥
⊥	⊥	4	⊥	m

(c) Mapeamento parcial (S.ID ↔ T.ID), com identificador de objeto (S.ID ↔ T.ID)

S.A	B	ID	T.C
x	y	1	⊥
z	n	2	m
⊥	x	3	⊥
⊥	⊥	4	m

(d) Mapeamento total (S.B ↔ T.B; S.ID ↔ T.ID) com identificador de objeto (S.ID ↔ T.ID)

Figura 3 - Resultados possíveis de combinação de dados – Adaptado de (Bleiholder & Naumann, 2008)

Uma outra combinação possível seria a apresentada na Figura 4 em que temos uma situação de completude *intencional* e *extensional*. Temos uma representação única e de todos os objetos e atributos. No entanto, temos múltiplos valores para alguns atributos.

Atributos comuns,
dado um mapeamento de esquema

S.A	B	ID	T.C
x	y	1	⊥
z	⊥/n	2/2	m
⊥	x/k	3/3	⊥
⊥	⊥	4	m

Completude *intencional*

Objetos comuns,
dado um identificador de objeto

Figura 4 - Combinação de dados com completude *intencional* e *extensional* – Adaptado de (Bleiholder & Naumann, 2008)

Dependendo do objetivo de ID, o foco, os modelos e as ferramentas utilizados poderão variar, no entanto, conforme referido em Brazhnik e Jones (2007), de uma forma generalista, existem duas abordagens para a ID. Uma primeira variante é normalmente adotada em realidades em que se conhecem os dados disponíveis e quais as questões a colocar aos mesmos. Esta é uma situação em que se importam, das

diversas fontes de dados, os campos necessários para uma BD desenhada para o efeito. Um exemplo de aplicação poderá ser um estudo em que uma empresa pretende analisar as vendas das várias filiais. Numa outra abordagem de ID procura-se avaliar o valor de dados existentes em fontes de dados com graus de confiabilidade distintos. Esta abordagem, em que se recorre a técnicas de *data mining*, centra-se na procura de correlações entre os dados e na descoberta de novo conhecimento potencialmente presente nos mesmos. São projetos em que se pretende responder a questões interdisciplinares complexas com base na informação que existe, na qualidade dos dados disponíveis e, ao mesmo tempo, perceber que informação adicional é necessária para este fim. Esta abordagem é muitas vezes utilizada em projetos relacionados com a área da saúde, em que, por exemplo, se podem cruzar dados de escolas, hospitais e farmácia, entre outros, para estudar um surto epidémico (Brazhnik & Jones, 2007). Qualquer que seja a abordagem, a ID terá de ser capaz de analisar os dados com uma variedade de métodos e técnicas e apresentar resultados de acordo com os requisitos definidos pelos futuros utilizadores da informação (Brazhnik & Jones, 2007).

2.1.2 Desafios e Passos da Integração de Dados

Com níveis de complexidade diferentes, os projetos de ID podem variar entre a (simples) união de conjuntos de dados com estrutura similar, que poderão ter sido produzidos em tempos ou locais diferentes, e a integração multidisciplinar. No entanto, em todos eles reúnem-se dados diversificados através do mapeamento e fusão de conceitos, modelos, vocabulários controlados e conjuntos, elementos e valores de dados (Brazhnik & Jones, 2007).

Naumann e Bleiholder (2006) apresentam a ID como uma tarefa bastante complexa que compreende a resolução de vários problemas que são, segundo estes autores, cada um por si só, formidáveis. O acesso aos dados, possivelmente localizados remotamente, poderá apresentar desde logo um desafio técnico. Por outro lado, esquemas de dados heterogéneos, de fontes de dados diferentes, terão de ser alinhados. Mais ainda, representações múltiplas de objetos de mundo real (duplicados) terão de ser detetadas, e por último, estes duplicados têm de ser fundidos de maneira a apresentar um resultado preciso e consistente ao utilizador. Estes autores propõem uma abordagem para a ID, num processo de três fases, conforme a Figura 5.

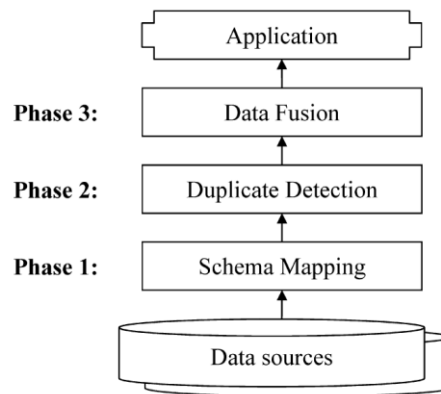


Figura 5 - Os três passos da integração de dados – Retirado de: (Bleiholder & Naumann, 2008)

Numa primeira fase, é necessário identificar correspondências de atributos que representam a mesma informação nas fontes, resultando daqui um mapeamento de esquema de dados que permitirá transformar os dados presentes nas fontes numa representação comum. Executam-se aqui operações de transformação e/ou renomeação dos dados. Na fase seguinte, recorrendo a técnicas de deteção de duplicados, procede-se à identificação e ao alinhamento dos objetos descritos nas fontes de dados. Nesta fase são encontradas as representações múltiplas, possivelmente inconsistentes, dos mesmos objetos do mundo real (duplicados). Na última fase, fusão de dados, resolvem-se as questões das inconsistências dos dados e as representações duplicadas são fundidas numa só (Bleiholder & Naumann, 2008).

Cada uma destas fases é de seguida descrita com mais detalhe.

2.1.2.1 Mapeamento de Esquemas de Dados

Pode-se definir mapeamento de esquemas de dados (*Schema Mapping*) como o resultado de uma operação de correspondência de esquemas (*Schema matching*). Um mapeamento de esquemas consiste num conjunto de elementos de mapeamento que indicam que determinados atributos de um esquema A estão relacionados com determinados atributos de um esquema B. Na Figura 6 podem-se ver dois esquemas de dados *PO* e *POrder*. Nestes esquemas, o atributo *PO.Lines.Item.Line* poderá ser mapeado com o atributo *POrder.Items.Item.ItemNumber* e representado com um elemento de mapeamento que poderá ser, por exemplo, uma expressão que pode até especificar alguma semântica, ou seja: *PO.Lines.Item.Line = POrder.Items.Item.ItemNumber*.

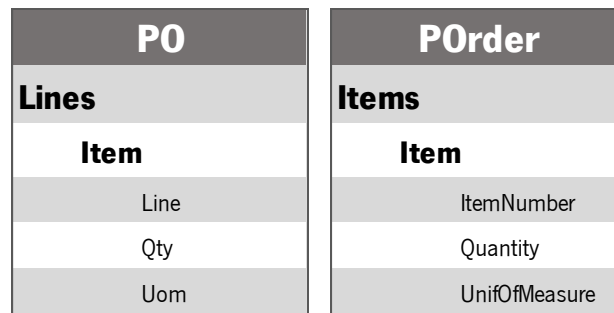


Figura 6 - Dois esquemas para serem mapeados – Adaptado de (Madhavan, Bernstein, & Rahm, 2001)

Numa tarefa de integração de dados de fontes autónomas é expectável que as mesmas apresentem esquemas de dados diferentes, pelo que, a resolução da heterogeneidade destes esquemas será o primeiro passo a realizar. Nesta fase, procede-se à identificação de elementos semanticamente equivalentes dos esquemas e, se necessário, à transformação de dados com a finalidade de os converter para um esquema único e comum. Correspondência de esquemas (*Schema matching*) pode ser definido como “o processo (semi) automático de deteção de correspondências de atributos entre dois esquemas de dados heterogéneos” (Naumann & Bleiholder, 2006). *Schema*, ou em português “esquema”, é uma estrutura formal que representa um artefacto de engenharia, tal como um *Structured Query Language* (SQL) *schema*, ou um diagrama de entidades e relacionamentos. Uma correspondência é uma relação entre dois ou mais elementos de um esquema e um ou mais elementos de outro esquema. Na Figura 7 pode-se observar um exemplo de correspondências que representam o mesmo conceito em dois esquemas diferentes. Estas correspondências são tipicamente do tipo 1 para 1 (1:1), ou seja, a um elemento de um esquema equivale outro elemento do outro esquema. No entanto, como podemos ver na mesma figura, o atributo *author* do esquema “Books” equivale a dois atributos do esquema “AuthorInfo”, nomeadamente: *LastName* e *FirstName*.

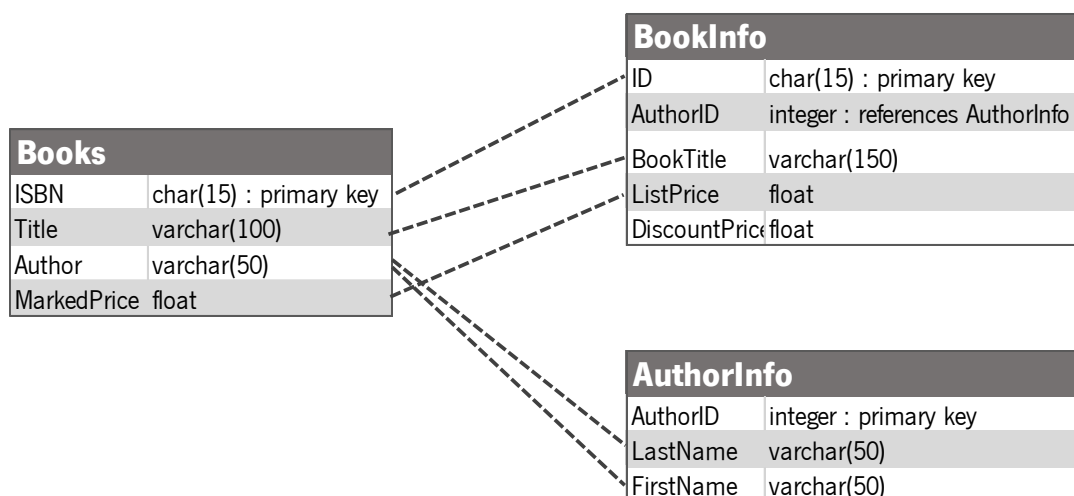


Figura 7 - *Schema matching* - Adaptado de (Bernstein, Madhavan, & Rahm, 2011)

2.1.2.2 Detecção de Duplicados

Nesta fase, recorrendo a técnicas adequadas, procede-se à identificação de possíveis múltiplas representações (potencialmente inconsistentes) do mesmo objeto do mundo real (Bleiholder & Naumann, 2008).

A problemática da deteção de duplicados é uma área de pesquisa com uma longa tradição, remontando aos trabalhos de *Record Linkage* de Newcomb et al. (1959). Na literatura aparece muitas vezes referenciada com outras designações, tais como: *record linkage*, *object identification*, *reference reconciliation*, entre outras.

Teoricamente, este processo aparenta ser simples. Bastará comparar pares de objetos recorrendo a medidas de similaridade e definir um limite (*threshold*). Se a medida de similaridade é maior que o limite definido, então o par é considerado um duplicado. No entanto, existem duas grandes questões neste processo que têm de ser resolvidas: a eficácia, que resulta grandemente da medida de similaridade e da escolha do limite, e a eficiência, afetada em grande escala pela dimensão do conjunto de dados a ser analisado (Bleiholder & Naumann, 2008).

Um processo de deteção de duplicados passará, naturalmente, pela escolha dos atributos a serem comparados. Estes atributos, conforme referido em (Naumann & Bleiholder, 2006), deverão ter 3 propriedades: 1 – estarem relacionados com o objeto a ser comparado; 2 – serem apropriados para a aplicação medida (função) de similaridade; 3 – serem passíveis de distinguir duplicados de não-duplicados. Num modelo de dados relacional, além dos atributos da tabela referente ao objeto, poderão ainda ser considerados atributos de tabelas relacionadas (tabelas que contêm chaves estrangeiras que referenciam a tabela em questão). Na Figura 8 pode-se ver um exemplo de duas tabelas, *MOVIE* e *FILM* com tabelas relacionadas, cujos atributos poderão ser utilizados para o processo de deteção de duplicados. Neste exemplo, os atributos *NAME* das tabelas *ACTOR* ou *ACTRESS*, relacionadas com a tabela *MOVIE*, poderão ser comparados com o atributo *NAME* da tabela *ACTORS*, relacionada com a tabela *FILM*, já que representam a mesma propriedade destes objetos.

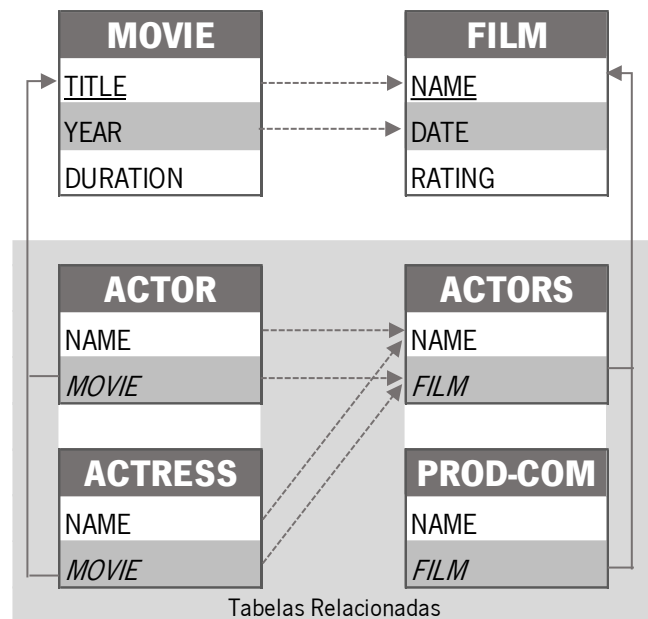


Figura 8 – Tabelas com tabelas relacionadas e com atributos correspondentes – Adaptado de (Naumann & Bleiholder, 2006)

Finda a seleção de atributos dá-se início ao processo de comparação, em que os tuplos são comparados recorrendo a uma medida de similaridade com dois limites definidos, um para ter a certeza de duplicado e outro para ter a indicação de possível duplicado. Desta avaliação resulta um de três resultados possíveis:

- Se o valor da similaridade é maior ou igual ao limite de certeza de duplicado, então considera-se um duplicado;
- Se está entre os limites, então tem-se um possível duplicado;
- Se está abaixo ou é igual ao limite de possível duplicado, regista-se um não duplicado.

No caso dos possíveis duplicados, os tuplos serão apresentados ao utilizador, de forma ordenada para que este os possa classificar manualmente (Naumann & Bleiholder, 2006).

Na Tabela 1 pode-se observar o resultado da comparação de nomes, com recurso a diferentes algoritmos de cálculo de similaridade de *strings*. Nestes cálculos a similaridade é maior quanto mais próximo o valor estiver de 1.000. Se se definir um limite para certeza de similaridade de valor 0.900 e um limite de possível duplicado de valor 0.800, de acordo com o algoritmo *Jaro* pode-se classificar os pares “SHACKLEFORD – SHACKELFORD”, “NICHLESON - NICHULSON”, “JERALDINE – GERALDINE”, “MARHTA – MARTHA” e “JON - JOHN” como **duplicados**. Os pares “DUNNINGHAM – CUNNINGHAM”, “MASSEY – MASSIE”, “ABROMS - ABRAMS”, “MICHELLE – MICHAEL”, “JULIES – JULIUS”, “TANYA –

TONYA” e “DWAYNE - DUANE” classificar-se-iam como **possíveis duplicados** e os restantes pares como **não duplicados**.

Tabela 1 - Comparação de resultados de algoritmos de cálculo de similaridade de *Strings* Adaptado de - (Winkler, 2006)

	Indivíduo A	Indivíduo B	Algoritmo			
			Jaro	Jaro-Winkler	Bigram	Edit
Sobrenome	SHACKLEFORD	SHACKELFORD	0.970	0.982	0.925	0.818
	DUNNINGHAM	CUNNIGHAM	0.896	0.896	0.917	0.889
	NICHLESON	NICHULSON	0.926	0.956	0.906	0.889
	JONES	JOHNSON	0.790	0.832	0.000	0.667
	MASSEY	MASSIE	0.889	0.933	0.845	0.667
	ABROMS	ABRAMS	0.889	0.922	0.906	0.833
	HARDIN	MARTINEZ	0.000	0.000	0.000	0.143
	ITMAN	SMITH	0.000	0.000	0.000	0.000
Nome	JERALDINE	GERALDINE	0.926	0.926	0.972	0.889
	MARHTA	MARTHA	0.944	0.961	0.845	0.667
	MICHELLE	MICHAEL	0.869	0.921	0.845	0.625
	JULIES	JULIUS	0.889	0.933	0.906	0.833
	TANYA	TONYA	0.867	0.880	0.883	0.800
	DWAYNE	DUANE	0.822	0.840	0.000	0.500
	SEAN	SUSAN	0.783	0.805	0.800	0.400
	JON	JOHN	0.917	0.933	0.847	0.750

Num cenário de ID considerar-se-iam não só o valor de cada par, mas a combinação das cotações para os dois campos; o Nome e o Sobrenome. Neste caso, considerando-se, por exemplo, a média das cotações, apenas o indivíduo “JERALDINE SHACKLEFORD / GERALDINE SHACKELFORD” seria considerado um duplicado, dado que o valor médio da medida de similaridade é de 0.948 valores.

Este tema da deteção de duplicados, por ser fulcral no desenvolvimento da presente dissertação vai ser analisado com maior detalhe na seção 2.3, associada ao *Record Linkage*.

2.1.2.3 Fusão de Dados

Após a deteção de duplicados, encontrados na fase anterior, procede-se à operação de fusão de dados (FD). Bleiholder e Naumann (2008) definem FD como o processo de fusão de múltiplos registos que retratam o mesmo objeto do mundo real, numa representação única, simples, completa e consistente. Neste processo as múltiplas representações de um objeto são fundidas numa só, resolvendo as inconsistências nos dados. Estas representações múltiplas de objetos levarão, invariavelmente, a situações de conflito de dados, que poderão ser resolvidas recorrendo a estratégias de resolução de conflitos nos dados, conforme se pode verificar no ponto 2.1.2.3.1.

Referem ainda Bleiholder e Naumann (2008) que a literatura apresenta outras denominações para este procedimento, designadamente: *record data merging*, *data consolidation*, *entity resolution* ou *finding representation/survivors*.

Na literatura, alguns autores consideram os conceitos de integração de dados e de fusão de dados como atividades equivalentes. No entanto, na presente dissertação, considera-se integração de dados como um processo global de integração e, fusão de dados, como uma das etapas desse processo, conforme se pode verificar no ponto 2.1.2.

2.1.2.3.1 Conflitos

Na fase de FD surgirão, naturalmente, uma série de conflitos de dados, resultantes das representações duplicadas dos objetos encontradas. Encontra-se um conflito de dados, por exemplo, quando um atributo, semanticamente equivalente, oriundo de fontes distintas apresenta discordância no valor. Imagine-se uma entidade do tipo pessoa em que, numa fonte de dados, o ano de nascimento é 1975 e noutra fonte para a mesma pessoa, o ano de nascimento armazenado é de 1978.

Ao nível dos dados podemos encontrar dois tipos de conflitos: (i) incerteza; (ii) contradição.

A incerteza prende-se com o valor do atributo, conseqüente da inexistência dessa informação em algumas fontes, enquanto noutras está presente (conflito entre nulo e não nulo). Já a contradição resulta da existência de valores não nulos diferentes para o mesmo atributo (Bleiholder & Naumann, 2008).

Na Figura 9 podemos verificar uma contradição relativamente ao ano de realização do filme *Snatch*, já que a tabela da esquerda apresenta o valor 2000 e a da direita apresenta o valor 1999. Já para o filme *Troy* encontramos uma incerteza quanto ao género desta produção (Bleiholder & Naumann, 2006). Nesta representação, o símbolo “⊥” refere-se a valores nulos para os atributos.

Title	Year	Director	Genre	ID
Snatch	2000	Ritchie	Crime	1
Troy	2004	Peterson	⊥	2
Vanilla Sky	2001	Crowe	Sci-Fi	3
Shrek	2001	Adamson	Anim.	4
The Matrix	1999	Wachowski	Fantasy	5

ID	Titel	Jahr	Rating	Genre
1	Snatch	1999	R	Crime
2	Troja	2004	R	History
3	Vanilla Ski	2001	R	Sci-Fi
3	Vanilla Sky	2000	16	Comedy
5	Matrix	1999	16	Fantasy

Figura 9 - Tabelas com correspondência de atributos e duplicados detetados – Adaptado de (Bleiholder & Naumann, 2006)

A resolução de inconsistências nos dados pode ser alcançada através de estratégias de resolução de conflitos. Estas estratégias definem o que fazer com estes dados inconsistentes e podem até definir as ações a tomar, como por exemplo: qual o valor a considerar, como combinar o valor ou como calcular um novo valor para criar uma representação única e consistente para o mesmo. Bleiholder e Naumann (2006) apresentam uma definição e classificação deste tipo de estratégias que foram resumidas na Tabela 2. Como se pode verificar, existem várias alternativas para a resolução de conflitos. Em algumas estratégias, como a *Pass it on* ou a *Consider all possibilities* cabe ao utilizador a decisão da ação tomar. Noutras, mediante determinada condição é aplicada uma regra, como por exemplo a estratégia *Take the information* que, numa situação de incerteza, rejeita sempre o valor nulo, considerando o valor não nulo. Algumas estratégias são mais sofisticadas e tentam resolver o conflito com recurso a alguma análise da informação disponível, como por exemplo a estratégia *Cry with the wolves* que procura o valor mais comum para aquele atributo, de entre os valores em conflito, aplicando-o, ou a *Keep up the date*, que, caso exista, recorre à informação temporal, por exemplo, a data de alteração dos valores, para seleccionar o valor mais atual.

Tabela 2 - Estratégias de resolução de conflitos – Adaptado de (Bleiholder & Naumann, 2006)

Estratégia	Classificação	Descrição
<i>Pass it on</i>	<i>Ignoring</i>	Passa todos os valores em conflito para o utilizador/aplicação decidirem a ação a tomar.
<i>Consider all possibilities</i>	<i>Ignoring</i>	Enumera ou combina todas as possibilidades para determinado valor e apresenta-os ao utilizador para decidir qual o valor a seleccionar.
<i>Take the information</i>	<i>Avoiding, instance based</i>	Utilizada em situações de incerteza (valor vs <i>NULL</i>), selecciona o valor e rejeito o nulo.
<i>No gossiping</i>	<i>Avoiding, instance based</i>	Ignora toda a informação inconsistente e apresenta apenas os factos consistentes.
<i>Trust your friends</i>	<i>Avoiding, metadata based</i>	Em caso de conflito, selecciona sempre a informação presente em determinada fonte, previamente definida.
<i>Cry with the wolves</i>	<i>Resolution, deciding</i>	Selecciona o valor mais comum entre aqueles que estão em conflito.
<i>Roll the dice</i>	<i>Resolution, deciding</i>	Escolhe aleatoriamente um valor entre todos os valores em conflito.

Estratégia	Classificação	Descrição
Meet int the middle	<i>Resolution, mediating</i>	Calcula um novo valor, o mais próximo possível dos valores presentes (por exemplo média).
Keep up to date	<i>Resolution, deciding</i>	Utiliza o valor mais recente (necessita de informação temporal ou então, em casos de <i>stream</i> de dados, utiliza a ordem de chegada da informação).

Bleiholder e Naumann classificam as estratégias de resolução de conflitos em três tipos, a seguir explicados, conforme a Figura 10, a seguir apresentada:

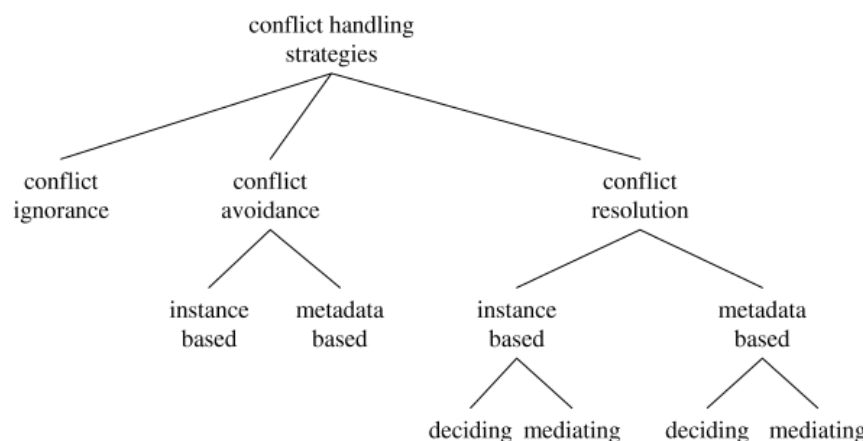


Figura 10 - Classificação de estratégias para a resolução de inconsistências nos dados – Retirado de (Bleiholder & Naumann, 2006)

- **Conflict ignorance** (Ignorar o conflito) – As estratégias deste tipo não tomam qualquer decisão/ação e muitas vezes nem sabem da existência destes conflitos. São estratégias que deixam a decisão a cargo do utilizador.
- **Conflict avoidance** (Evitar o conflito) – Estas estratégias não resolvem conflitos apesar de conhecerem a sua possível existência. Tratam as inconsistências de um modo genérico, ou seja, aplicam uma decisão única para todos os casos. Por exemplo: dão prioridade à preferência de determinada fonte de dados. Estas decisões podem ser tomadas tendo em consideração os dados (*instance based*) ou os metadados (*metadata based*).
- **Conflict resolution** (Resolver o conflito) – Com este tipo de estratégias procura-se uma resolução para os conflitos através da análise dos dados e dos metadados. Esta categoria de estratégias pode ainda ser dividida em dois tipos: (i) decisão, em que a técnica decide qual o

valor, dentro dos presentes, a utilizar; e (ii) mediação, em que se calcula um novo valor a aplicar com base nos valores apresentados (por exemplo a média desses valores).

As estratégias de resolução de conflitos são aplicadas com recurso a um conjunto de funções de resolução de conflitos presentes nas ferramentas de integração. Em Bleiholder e Naumann (2005), os autores apresentam uma breve descrição de algumas destas funções, que podem ser visualizadas na Tabela 3.

Tabela 3 - Funções de resolução de conflitos – Adaptado de (Bleiholder & Naumann, 2005)

Função	Descrição
<i>Count</i>	Conta o nº de valores não nulos distintos
<i>Min / Max</i>	Devolve o valor mínimo ou máximo de um atributo numérico
<i>Sum / Avg / Median</i>	Calcula a soma, a média ou a mediana dos valores não nulos presentes
<i>Variance / Stddev</i>	Calcula a variância ou o desvio padrão
<i>Random</i>	Escolhe um valor aleatório entre os valores não nulos
<i>Choose</i>	Escolhe um valor de uma fonte específica
<i>Coalesce</i>	Devolve o primeiro valor não nulo
<i>First / Last</i>	Devolve o primeiro ou último valor, mesmo se nulo
<i>Vote</i>	Devolve o valor mais repetido
<i>Group</i>	Devolve todos os valores (a decisão fica a cargo do utilizador)
<i>Shortest / Longest</i>	Devolve o valor mais curto ou mais longo, de acordo com uma medida de comprimento
<i>(Annotated) Concat</i>	Devolve os vários valores concatenados (podem incluir anotações)
<i>Highest Quality</i>	Devolve o valor com maior qualidade (necessita de um modelo de avaliação de qualidade)
<i>Most Recent</i>	Devolve o valor mais atual
<i>Most Active</i>	Devolve o valor mais acedido e utilizado
<i>Choose Corresponding</i>	Seleciona um valor que pertence ao valor escolhido para outra coluna

Função	Descrição
<i>Most Complete</i>	Seleciona o valor da fonte que tiver menos valores nulos
<i>Most Distinguishing</i>	Seleciona o valor mais distinto de entre os presentes
<i>Most General Concept / Most Specific Concept</i>	Devolve o valor mais geral com base numa taxonomia ou ontologia

Estas funções podem ter uma equivalência direta às estratégias apresentadas e por isso, uma aplicação facilitada. Se atentarmos na estratégia *Pass it on*, as funções *Group* ou *Concat* lidam facilmente com esta questão. Já a função *Coalesce* seria indicada para a estratégia *Take the information*. Podem existir diferentes funções adequadas à mesma estratégia e, dependendo do contexto, algumas podem ser mais apropriadas do que outras (Bleiholder & Naumann, 2006).

2.2 Qualidade dos Dados

Poder-se-á dizer que os dados têm qualidade se estão em conformidade com os *standards* que foram definidos para os mesmos e, portanto, poderão ser utilizados para a finalidade a que se destinam (Herzog et al., 2007). O conceito de qualidade dos dados (QD) é complexo e pode diferir consoante o contexto em que é aplicado. Alguns problemas de QD podem ser facilmente identificados como, por exemplo, erros de digitação de texto ou a presença de valores fora dos intervalos possíveis. Mas noutros casos, em que temos valores admissíveis mas incorretos para determinadas propriedades, esta tarefa torna-se bastante mais complexa (Batini & Scannapieco, 2006).

Segundo Batini e Scannapieco (2006), o nível de qualidade dos dados tem consequências significativas a nível da eficácia e eficiência das organizações e empresas. Um exemplo dado é o custo que os problemas relacionados com QD tem para as empresas norte americanas. De acordo com um relatório do *Data Warehousing Institute* (Eckerson, 2002), esse custo pode ser avaliado na ordem dos 600 mil milhões de dólares por ano.

Em (Herzog et al., 2007) é referido que, quando os níveis de qualidade dos dados são muito baixos, estes podem provocar baixos níveis de satisfação tanto dos clientes como dos trabalhadores. Estes baixos níveis de satisfação dos trabalhadores, podem até, de acordo com o mesmo autor, provocar alta rotatividade do fator do trabalho nas organizações e empresas e, conseqüente, aumento de custos. Quando se fala em dados financeiros, o baixo nível de qualidade dos mesmos pode tornar difícil apreender a verdadeira situação/viabilidade financeira de uma empresa. Por outro lado, os mesmos

autores evidenciam que níveis elevados de qualidade de dados traduzem-se em vantagem competitiva sobre os concorrentes ou numa oportunidade de negócio, visto que outras empresas/organizações terão interesse em informação que é útil e fidedigna.

Na literatura, como as anteriormente mencionadas, encontram-se também referências a dimensões de QD que permitem avaliar a mesma em diferentes grandezas, que serão desenvolvidas a seguir.

2.2.1 Dimensões de Qualidade dos Dados

Será natural que, dada a existência de literatura diversa sobre este tema, a classificação das dimensões de QD possa diferir, em certas abordagens, de acordo com a relevância das mesmas para esses trabalhos. De sublinhar que até a definição das próprias dimensões não tem um entendimento comum. Contudo Batini et al. (2009) agregaram um conjunto de dimensões comumente usadas pela maioria dos autores, deixando a ressalva de que estas dimensões não configuram um entendimento na literatura, sendo elas: Precisão, Completude, Consistência e Tempo (nas vertentes atualidade, volatilidade e periodicidade).

2.2.1.1 Precisão

Existem, como já foi anteriormente referido, várias definições dependendo do autor. Batini et al. (2009) citam alguns exemplos destas definições como em Wang e Strong (1996) que definem precisão como “até que ponto os dados são corretos, de confiança e certificados” ou em Ballou e Panzer (1985) que considera que os dados são precisos desde que os dados contidos na base de dados correspondam aos dados do mundo real.

Em Batini et al. (2009) e Batini & Scannapieco (2006) apresenta-se a definição para precisão como sendo “a medida de proximidade entre um valor v e um valor v' sendo v' o valor correto, no mundo real, e v o que pretende representar v' na base de dados”.

Desta definição resultam dois tipos de precisão: a sintática, “que é o grau de proximidade de um valor v com os elementos do domínio de definição correspondente D ”, e semântica, “o grau de proximidade entre o valor v e o valor no mundo real v' ”. Para se entender melhor estes tipos de precisão considere-se a Tabela 4, a seguir apresentada, com uma relação de filmes.

Tabela 4 – Relação de filmes – Adaptado de (Batini & Scannapieco, 2006)

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	<i>null</i>
3	Rman Holiday	Wylder	1953	0	<i>null</i>
4	Sabrina	<i>null</i>	1964	0	1985

No exemplo dado por Batini e Scannapieco (2006) é indicado que *Rman Holiday* (v) no filme 3 é sintaticamente incorreto por não ter qualquer correspondência com o nome de um filme, sendo *Roman Holiday* o título de filme mais aproximado a v. Com a diferença na inserção do carácter “o” considera-se que a distância na edição é de 1, refletindo-se num grau de precisão sintática de 1. Segundo os mesmos autores, poder-se-á definir uma função C em que o grau de precisão será o valor mínimo da função C de um valor v quando comparado com todos os valores do domínio D, sendo que a função C está compreendida entre [0, ... , n], sendo n o valor máximo do número de comparações que a função permite.

Em relação à Precisão Semântica, os mesmos autores recomendam que se troquem entre si os nomes dos diretores para os filmes 1 e 2. Apesar de *Weir* poder ser considerado um valor admissível no conjunto de nomes e, como tal, sintaticamente correto para o filme 1, este não é, na realidade, o diretor deste filme e como tal, apresenta-se aqui um erro de precisão semântica.

Em Batini et al. (2009) é indicado que entre os dois tipos de precisão, a mais utilizada, quando se fala em QD, é a precisão sintática talvez devido ao elevado grau de complexidade de cálculo que a precisão semântica exige, tal como é distinguido por Batini e Scannapieco (2006).

2.2.1.2 Completude

Em (Batini et al., 2009) são resumidas numa tabela algumas definições de completude que se podem observar na Tabela 5, sendo que, em todas elas faz-se referência à quantidade de informação disponível ou à capacidade de representação de um objeto do mundo real.

Tabela 5 - Definições de completude – Adaptado de (Batini et al., 2009)

Referência	Definição
Wand e Wang (Wand & Wang, 1996)	Capacidade de um sistema de informação para representar todos os estados significativos de um sistema do mundo real.
Wang e Strong (Wang & Strong, 1996)	Medida em que os dados têm largura, profundidade e alcance para a tarefa em execução.
Bovee et al. (Bovee et al., 2003)	Informação que possui todas as partes requeridas para a descrição de uma entidade.
Naumann (Naumann, 2002)	Rácio entre o número de valores não nulos numa fonte e o tamanho da relação universal.

Contudo em (Herzog et al., 2007) encontra-se uma explicação simples e direta do conceito de completude, ou seja, entende-se como sendo o facto de que não existe nenhum registo omissos na base de dados e que nenhum dos elementos que compõem esses registos está também em falta. Esta definição é o estado ótimo de completude, pelo que, quanto mais próximos os dados registados estiverem desta definição, maior é o grau de completude e, conseqüentemente, melhor será a qualidade dos dados. Em (Batini & Scannapieco, 2006) são identificados 3 tipos de completude: **completude de esquema** que considera o grau de completude em que os conceitos e propriedades dos mesmos não se encontram omissos do esquema de dados; **completude de coluna**, que é uma unidade de medida que verifica valores omissos de determinada propriedade ou coluna; e, **completude de população**, que controla se existem valores omissos da população em análise.

A completude refere-se, portanto, à quantidade de informação disponível relativamente à informação possível. De realçar que na completude dos dados pode-se verificar facilmente a ausência de valores em algumas propriedades, mas se todo um tuplo estiver em falta, esta ausência já não será tão facilmente detetável (Batini & Scannapieco, 2006).

2.2.1.3 Consistência

Em relação à consistência, tanto em Batini et al. (2009) como em Batini e Scannapieco (2006), esta é definida como “o nível de violação de regras semânticas detetadas num conjunto de dados”. Estes autores apresentam como exemplo dessas regras semânticas, que validam a consistência da qualidade de dados, no caso da teoria relacional, as restrições de integridade e no caso da estatística, a edição de dados.

Considerando as restrições de integridade, para a teoria relacional, pode-se ainda distinguir em restrições intrarelacionais e interrelacionais. As intrarelacionais definem o conjunto de valores admissíveis de um determinado atributo ou de vários atributos de uma só relação. Por exemplo “a idade deve estar compreendida entre 0 e 120 anos ou, se “*anosDeServico* < 3 , então o salário não pode ser maior que 25000 *u.m.* por ano”. As interrelacionais são restrições de atributos com mais do que uma relação. Considere-se, por exemplo, a Tabela 4 e considere-se uma nova relação, *Óscares* que indica quais os Óscares que determinado filme ganhou e em que ano. Desde logo, um exemplo de restrição interrelacional será, como indicado pelos autores, “Ano do filme = Ano do Óscar” uma vez que ambas as propriedades têm de referir a um mesmo ano civil.

Contudo, segundo Batini e Scannapieco (2006), mesmo quando os dados não são relacionais, as regras de consistência também são aplicáveis, como por exemplo, no caso da estatística, as regras de semântica da edição de dados (*data edits*), definidas pelos autores como: “um processo que garante que certas inconsistências são detetadas através de um conjunto de regras impostas, para todas as respostas validadas como corretas”. Um exemplo dado para esta regra semântica, edição de dados, será no caso de definir uma idade mínima para o estado civil casado: “se o estado civil for casado a idade não pode ser inferior a 14”. Os erros, conforme referem os autores, podem ser difíceis de detetar, mesmo com a aplicação de rotinas de verificação de consistência dos dados.

2.2.1.4 Tempo

Esta dimensão da qualidade dos dados assenta na análise do nível de qualidade de variáveis relacionadas com o tempo, isto porque, os dados estão em constante atualização e mudança. Apesar de em (Batini et al. (2009) existir um conjunto de definições para estas diferentes variáveis relacionadas com o tempo (atualidade, volatilidade e periodicidade) recolhidas de diferentes autores, é em Batini e Scannapieco (2006) que encontramos uma definição resumida destas mesmas variáveis. Assim sendo, pode-se considerar como **atualidade**, o nível de rapidez com que os dados são atualizados de modo a que a informação mais recente seja espelhada. Um exemplo de atualidade dos dados pode ser a atualização de uma morada de um cliente. Se corresponde à morada atual em que a pessoa vive, então a atualidade

dos dados é elevada. **Volatilidade** remete para a variação dos dados no tempo, ou seja, quanto mais os dados variarem no tempo, mais voláteis eles serão, o que poderá significar menor qualidade dos mesmos, visto que a informação poderá não ser a mais real. Os autores dão como exemplo dados relacionados com a data de nascimento, que terão níveis de volatilidade de 0, dado que esta informação, se correta, nunca mais sofre alterações. Os preços das ações, por outro lado, terão níveis de volatilidade elevados, dado que podem variar a cada segundo que passa. Finalmente, **periocidade**. Esta variável surge do pressuposto de que existe a possibilidade de, segundo os autores, a informação atual ser redundante se ela for atualizada demasiado tardiamente para a tarefa em questão. Os autores dão como exemplo os horários de aulas dos cursos das universidades, se estes por alguma razão, só forem disponibilizados após o início das aulas então esta informação é considerada redundante para a tarefa visto que não estava disponível antes do início das aulas.

2.3 Record Linkage

Record Linkage (RL), também designado por *computerized matching*, é o nome atribuído às técnicas utilizadas para identificar uma mesma entidade (pessoa, negócio, etc.) recorrendo a quasi-identificadores, tais como o nome, a morada ou uma data. Isolados, estes quasi-identificadores não serão capazes de identificar uma correspondência unívoca mas, combinados, poderão alcançar este objetivo (Winkler, 2014). O nome de um indivíduo, por si só, dificilmente identificará univocamente uma pessoa, dado que outro indivíduo poderá ter a mesma denominação. No entanto, conforme refere Winkler, combinando o nome do indivíduo com a sua data e/ou local de nascimento esta identificação já poderá ser possível (Winkler, 2014).

O termo *Record Linkage* foi utilizado pela primeira vez por Dunn (1946) num artigo em que descreve a ideia de que cada indivíduo cria um “livro de vida”, que tem início com o seu registo de nascimento e término com o respetivo registo de óbito, e cujo conteúdo é preenchido com os registos da interação desse indivíduo com, entre outros, os serviços de saúde, da segurança social e registos militares. Dunn apresenta RL como “o processo de reunir as páginas deste livro num só volume”.

Christen (2012) refere que a utilização de *Record Linkage* teve as primeiras aplicações na área de medicina em estudos e investigações em que houve necessidade de ligação dos pacientes aos respetivos registos médicos. Considerando que a tarefa se baseava em encontrar correspondências entre registos, ainda em papel, através de quasi-identificadores, como nomes e datas de nascimento, que podiam conter

erros ortográficos, apresentar informação em formatos diferentes ou estarem presentes em registos manuais de difícil de leitura, a mesma, tornava-se, claramente “herculeana”.

Newcombe, tal como explica Christen (2012) , através dos seus dois *papers* para a *Science* (Newcombe et al., 1959) e para a *Communications of the Association of Computing Machinery* (Newcombe & Kennedy, 1962) apresenta uma proposta para a utilização de computadores para automatizar a ligação dos registos, tendo desenvolvido as ideias base para a abordagem probabilística de RL. Nesta abordagem, comparando os quasi-identificadores dos registos a atribuindo-lhes uma cotação baseada numa medida de similaridade, os registos que apresentassem valor acima de um determinado limite superior seriam tidos como correspondentes. De outro modo, se se encontrassem abaixo de um limite inferior, seriam considerados não correspondentes. Os resultados encontrados entre esses limites seriam alvo de revisão e, caso fossem detetados erros, de correção.

Estes investigadores da área da medicina genética e da biomédica tinham (entre outros) interesse em acompanhar grupos amplos de indivíduos e avaliar a fertilidade diferencial das famílias considerando a presença ou ausência de doenças hereditárias e, para tal, consideraram que o historial de um indivíduo poderia ser obtido através do cruzamento dos vários registos vitais de várias agências governamentais (Herzog et al., 2007). Assim, numa primeira fase, desenvolveram um método para cruzar registos de nascimentos e de casamentos. Recorreram aos nomes/apelidos dos progenitores nos ficheiros dos casamentos e aos nomes completos, naturalidade, idade, profissões, etc. do pai ou da mãe nos ficheiros dos nascimentos, pontuando positivamente cada concordância e negativamente cada discordância. O registo de casamento com a pontuação mais elevada era então ligado ao registo de nascimento em análise (Ferreira, 2004).

O modelo desenvolvido por Newcombe tem o seu posterior reconhecimento em 1969 através da teorização do modelo matemático de Fellegi e Sunter (1969) para RL probabilístico o que vem comprovar a sua aplicabilidade e robustez e que tem sido amplamente utilizado até aos dias de hoje.

Gu e Baxter (2003), reforçam RL, como um tema de grande importância principalmente para projetos de investigação em que se necessitam realizar constantemente comparações de informação acerca de uma entidade. Segundo os mesmos autores, a visão ideal seria poder aplicar modelos de RL determinísticos. Contudo, para que isso fosse possível, seria necessário que todas as diferentes bases de dados, independente da fonte, tivessem um identificador único, visto que a aplicação determinística de RL assume a inexistência de erros ou incoerências. Na prática é raro observar essa condição devido a questões que vão desde erros de digitação, níveis de completude baixos ou porque as identificações dos campos mudam ao longo do tempo, por cada intervenção efetuada.

NeSmith afirma na sua obra que: “desde que exista uma grande base de dados de ficheiros genealógicos, o problema de duplicação de dados para o mesmo indivíduo ou família irá sempre existir” (NeSmith, 1992). Esta autora classifica a junção destes dados, que comumente têm erros e que usam identificadores diferentes para cada uma da base de dados, como um “pesadelo” para a área da genealogia. Wilson (2008) acrescenta que, num contexto de globalização, o desafio aumenta por se ter de considerar o sistema de escrita, de linguística e de construção de nomes, situação essa provocada pelas diferentes culturas. Ivie et al (2007) sublinham a diferença entre o caso particular de projetos que têm de lidar com genealogias e os restantes. Para estes autores os projetos não ligados às genealogias têm, normalmente, que lidar com um número pequeno e finito de atributos em que as respetivas fontes de dados não são tão dispersas como nesta área, nem se comparam com o largo volume de atributos que as caracterizam.

2.3.1 A aplicação de *Record Linkage*

Neiling em (1998) apresenta um método de aplicação de RL. Este autor divide o processo de RL em duas partes, uma primeira, responsável pelo pré-processamento dos dados e uma segunda parte, pelo tratamento dos dados, ou seja, o RL propriamente dito.

Para a primeira parte, segundo o autor, é necessário dispor de um conjunto de dados de treino, já classificados, realizando-se de seguida:

- A divisão em dois conjuntos de registos – um conjunto de pares com correspondência (*Link*) e um conjunto de pares sem correspondência (*Non-link*);
- Definir uma função de comparação para os dois grupos anteriores - esta função deverá indicar o grau de concordância entre cada um dos pares criados, com base no cálculo do valor de similaridade dos atributos definidos para a comparação, por exemplo, a média do valor de similaridade de todos os atributos em análise.
- Analisar os resultados do grau de concordância que resultam da aplicação função anterior aos conjuntos do primeiro ponto. Verifica-se aqui o valor de similaridade obtido, quer para os casos de correspondência e de não correspondência e definem-se os limites de concordância superior e inferior para posteriores classificações.

Na segunda parte, depois de definidas a função de comparação e de encontrados os limites de concordância para efetuar a classificação, realiza-se o RL propriamente dito, executando as seguintes tarefas:

- Aplicar a função de comparação nos conjuntos de dados a serem analisados - compara-se cada um dos registos de uma fonte de dados com cada um dos registos de outra fonte e aplica-se a função de comparação.
- Classificar todos os pares de acordo com a categoria que se inserem: *Link*, *Non-link* e *Possible Link*; com base na classificação obtida no ponto anterior e nos limites de similaridade definidos.
- Analisar todos os pares que recaem na categoria *Link* e, se necessário, *Possible Link* para determinar quais as melhores correspondências;
- Construir a base de dados com o resultado das correspondências do passo anterior.

2.3.2 Desafios de *Record Linkage* em Genealogia

Na área da genealogia, conforme refere Wilson (2008), um desafio de destaque prende-se com a questão da variação dos dados. Quer sejam nomes, datas ou lugares, a forma como os mesmos dados podem ser escritos pode provocar “ruído”, o que por sua vez leva à ocorrência de erros durante o processo de classificação e, conseqüentemente, a resultados pouco precisos. No limite, leva a níveis de qualidade de dados muito baixos, dado que, certas correspondências encontradas poderão ser tidas como verdadeiras, quando na realidade, serão falsas e vice-versa. Veja-se, por exemplo, o caso de um indivíduo que pode ter o registo de nascimento com determinado nome e ter, entretanto, alterado o mesmo ou, ter optado pela adoção do apelido do cônjuge no seu casamento. Nos registos podem ainda aparecer as alcunhas, ou então estes podem conter erros de escrita.

Outros problemas poderão ser, por exemplo: diferentes formas de escrever datas, cidades com o mesmo nome em diferentes países ou a variação do nome de ruas ao longo do tempo, que levam a que a mesma localização pode assumir vários nomes.

Contudo, para as variações de dados mais comuns, existe a possibilidade de fazer normalização e estandardização de dados, permitindo assim uma classificação mais robusta dos mesmos. Do ponto de vista do mesmo autor, a normalização refere-se a “tratamento da pontuação, ou passar todos os nomes para letra minúscula”, enquanto a estandardização trata de formatações, como conversão de datas para um formato comum, como “dia/mês/ano”, ou na atribuição de um ID numa base já existente de forma a poder definir um ID comum (Wilson, 2008).

Outros desafios estão ligados à situação de globalização, já anteriormente referida, e apresentada por este mesmo autor, mais concretamente:

- Escrita - embora o *Unicode* ajude a resolver esta situação existem, ainda assim, idiomas em que é necessário ter que lidar com questões de caracteres, por exemplo, os caracteres chineses *vs* japoneses *vs* coreanos;
- Ordem dos nomes - na cultura ocidental o nome próprio vem primeiro e depois o apelido, isto sem esquecer casos específicos, como segundo nome e apelidos maternos e paternos, enquanto que, na cultura oriental, usualmente, acontece o inverso;
- O uso de espaços – mais uma vez a cultura ocidental e oriental diferem no uso de espaços em nomes. No oriente o uso de espaço no nome não é importante, enquanto no ocidente é necessário usar como divisão, como em Filipe [espaço] Salgado;
- Abreviaturas e nomes adquiridos por matrimónio - muitas vezes os registos, principalmente em idiomas como o inglês ou similares, são utilizadas abreviaturas como *Mrs.*, *Mr.* ou *Dr.* que não correspondem a nomes. Também, muitas vezes, por questões culturais, o registo utiliza o nome e apelido do marido: *Mrs. Jonh Smith* apesar de este não ser o nome do indivíduo. Em última análise, isto significaria que o nome e apelido são desconhecidos. A mesma situação acontece na cultura oriental em que, após o casamento, o apelido de família da esposa “cai” dando lugar ao do marido;
- Honorífico – muito usados nos países orientais. Além de cada país ter o seu próprio honorífico para o mesmo significado, existe ainda a possibilidade de a mesma palavra poder ter diferentes significados dentro da mesma cultura, o que por sua vez é replicado no registo nas bases de dados. Em Wilson (Wilson, 2008) podem-se observar vários exemplos, como o caso japonês em que, por norma, o apelido de solteiro é conhecido mas o nome próprio é desconhecido. Neste caso, é usualmente registado um honorífico junto do apelido que pode ser lido como *Miss/Mr* ou *Filha/Filho*. Por exemplo, segundo o autor, um indivíduo do sexo feminino de nome Suzuki irá ser listado como *Miss Suzuki* ou *Suzuki-filha*;
- Patronímicos – esta característica é mais comum nos países escandinavos e pode dificultar bastante o processo de integração de dados, principalmente na área da genealogia. Culturalmente, é comum usar o nome do pai na construção do nome dos filhos. No exemplo dado pelo autor imagine-se que *Olaf* tem um pai com o nome próprio *Sven* então *Olaf* poderia chamar-se *Olaf Svensen* que se traduz como *Olaf- filho de Sven*. Se *Olaf* tiver, por exemplo, uma

filha com o nome próprio *Inga*, o nome dela seria *Inga Olafsdotter* em vez de *Inga Svensen*, ou seja, utilizam o nome próprio em vez do apelido, o que pode dificultar o processo de correspondências verdadeiras (Wilson, 2008);

- Patronímicos Cirílicos – na Rússia e nos países de leste europeus a utilização desta particularidade de nomes patronímicos afeta essencialmente os apelidos porque na construção do nome dos filhos, o género é considerado. Por existir uma terminação diferente para o género feminino e masculino pode-se estar na situação em que se assume um erro. Considere-se o indivíduo masculino *Ivan Popov* e o facto de o nome patronímico ser usado no meio do nome dos filhos com a diferença de *-vich* para o género masculino e *-yevna*, *-ovna* ou *-ichna* para o género feminino. O nome do filho poderia ser um *Sergey Ivanovich Popov* e o da filha poderia ser *Tatjana Ivanova Papova*.
- Nomes de locais nos países asiáticos - anteriormente já se referiu que nos países asiáticos como a China, o Japão ou a Coreia, o uso de espaços ou outros delimitadores no uso de nomes não tem a mesma importância que nos países ocidentais. Esta situação, no caso de nomes de locais, pode dificultar o processo de RL na medida em que são escritos do geral (País ou Distrito) para o específico (cidade ou aldeia) sem qualquer espaço.

Estes desafios num mundo cada vez mais globalizado e com bases de dados cada vez maiores e temporalmente mais distantes faz com que as problemáticas aqui expostas sejam por demais desafiantes para o RL, pelo que um pequeno erro durante o processo pode originar situações que diminuem a fiabilidade dos dados e, tal como já foi anteriormente referido, em última análise a redução do nível de qualidade dos mesmos.

2.3.3 Record Linkage e Ficheiros Genealógicos

Nesmith (1992) indica que RL, na prática e de forma simples, é “um programa de computador que utiliza um algoritmo com elevado nível de detalhe, baseado na probabilidade, para determinar se os dois registos que estão a ser comparados representam o mesmo indivíduo”. Este algoritmo é bastante adaptável pelo que pode ser afinado de forma a responder a certas especificidades culturais como as apresentadas no capítulo anterior. É claro que isto implica que este algoritmo possa ser ajustado para cada uma das áreas geográficas. Para a realização de RL, a autora indica que as técnicas mais utilizadas são:

- **Blocking** – refere-se ao sistema usado ao efetuar uma pesquisa acerca de determinado registo e encontrar a sua correspondência. Este utiliza um sistema de índices em que só se devolve aquele que tem maior probabilidade de correspondência com o registo pedido, “bloqueando” os restantes ou, se considerarmos outra perspectiva, “recuperando” a informação que irá corresponder ao pedido. O autor indica também que a eficácia deste parâmetro está relacionada com as métricas “recall” e “precisão”. *Recall* refere-se a uma medida que determina quantos registos foram considerados relevantes pela técnica de *blocking*. Precisão é uma outra medida que determina quantos dos registos considerados como correspondência pela técnica de *blocking* são relevantes;
- **Weight Calculations** – após as correspondências serem obtidas através do passo anterior, este processo irá comparar as mesmas com o registo em análise, dando um determinado peso de acordo com o resultado desta comparação. Irá ser atribuído um valor positivo quando existe concordância no resultado da comparação. Quando a concordância é parcial esse valor positivo é menor e quando não há concordância o peso atribuído é negativo. Este parâmetro assume automaticamente o valor zero quando existe falta de informação num dos campos dos registos que estão a ser comparados.
- **Threshold Determination** – Após ter sido obtido o conjunto de correspondências mais prováveis de satisfazer a pesquisa e depois de ter sido dado um peso a cada uma dessas correspondências, será neste passo decidido se efetivamente existe ou não correspondência. Será definido um limite superior e inferior para determinar se os registos são correspondentes, correspondentes possíveis ou não correspondentes. A determinação destes limites deverá ser adequada ao projeto em questão de modo a evitar falsos positivos ou falsos negativos

2.4 Extract, Transform and Load

Qualquer operação de integração de dados passará por um conjunto de tarefas de extração, manipulação e carregamento dos mesmos para uma BD de destino, a que se dá normalmente o nome de ETL.

Os processos de ETL referem-se a um conjunto de atividades que, com recurso a artefactos de *software* realizam tarefas de extração de dados de múltiplas fontes, limpeza e transformação dos mesmos e, posteriormente, a sua inserção num destino de dados, tendencialmente, um DW (Vassiliadis et al., 2005). Neste processo a informação é extraída das fontes de dados originais, limpa, transformada e tratada de modo a corresponder ao esquema de dados de destino e, por fim, carregada para o mesmo. Dada a

variedade e o número elevado de atividades diferentes que os integram, estes processos são, por norma, bastante complexos, dado que aqui se realizam operações que vão desde a limpeza de dados, deteção de duplicados, verificação de restrições de integridade, filtros, ordenação, agregações, cálculos ou mesmo a aplicação de funções integradas na ferramenta de ETL ou com recurso a *scripts* numa linguagem declarativa (Vassiliadis et al. 2009).

Na Figura 11 podemos visualizar uma representação de uma *framework* genérica para processos ETL descrita em (Vassiliadis, Simitsis, & Skiadopoulos, 2002). Na camada inferior estão representados todos os repositórios de dados inerentes ao processo, nomeadamente, as fontes dos dados originais (*Sources*), do lado esquerdo; a área de estágio dos dados (*Data Staging Area*), onde se procede à limpeza e transformação dos dados; e o DW onde são armazenados os dados tratados. Na camada superior estão representadas as operações relativas ao processo.

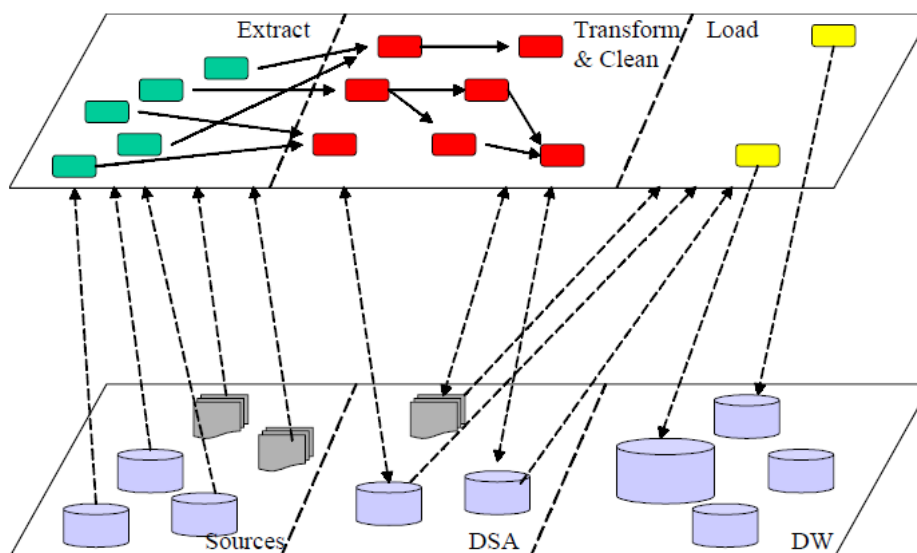


Figura 11 - O ambiente dos processo de ETL – Retirado de (Vassiliadis et al., 2002)

O processo de ETL apresenta várias fases, conforme descrito em (El-Sappagh et al., 2011) de seguida brevemente explicadas.

Extract – Nesta fase procede-se à obtenção de dados das fontes originais tornando-os disponíveis para serem processados. Estes dados poderão ter origem em bases de dados e/ou aplicações diversas e terão possivelmente formatos distintos e que terão de ser tratados de forma a conseguir-se obter uma extração de dados eficiente.

Transform – Nesta fase procede-se à aplicação de um conjunto de regras e de funções sobre os dados, com o intuito de os transformar e preparar para serem carregados para o DW na fase seguinte. Alguns

dados poderão não necessitar de quaisquer tratamentos, sendo carregados tal e qual como estão nas fontes de dados. Executam-se aqui tarefas de duas naturezas:

- *Clean* (Limpeza) - Em que se realizam operações de normalização de dados, conversão de valores (por exemplo, nulos para um valor *standard*), geração de identificadores unívocos (por exemplo: “Masculino/Feminino/Desconhecido”, “M/F/*null*”, “Homem/Mulher” são convertidos para um formato *standard* “Masculino/Feminino/Desconhecido”), validação de dados segundo um formato *standard* (por exemplo: Cód. Postal, N° de telefone)
- *Transform* (Transformação) – Em que se executam operações de transformação propriamente dita, por exemplo: conversões de medidas para uma dimensão comum (milímetros para centímetros), fusão de dados (de várias fontes), agregações, derivação de valores, ordenação, geração de *surrogate keys* (chaves substitutas), aplicação de regras de validação avançadas, entre outras.

Load – Procede-se aqui ao carregamento dos dados, depois de tratados, para o DW. É necessário garantir que o carregamento é feito corretamente e com o menor consumo de recursos do sistema possível, para garantir um bom desempenho do processo global. Uma sugestão pode ser a desativação das restrições e dos índices na base de dados de destino, tratando-se estas restrições no próprio processo de ETL.

Conforme referem El-Sappagh et al. (2011), apesar da elevada importância dos processos de ETL, e da existência de algumas abordagens para a resolução deste problema, tal como a de Vassiliadis et al. (2002), não existe ainda um modelo conceptual padrão para a representação de uma forma simplificada dos mesmos. Estes autores estudaram as abordagens existentes e, com base em alguma integração e extensão das mesmas, propõem uma nova abordagem para a modelação destes processos, o *Entity Mapping Diagram* (EMD). Este novo modelo, representado na Figura 12, responde aos seguintes requisitos:

1. Suporta a integração de múltiplas fontes de dados
2. É robusta na perspectiva de alteração das fontes de dados
3. Suporta transformações flexíveis
4. Pode ser facilmente implementado num ambiente adequado
5. Suporta as várias operações de extração, transformação e carregamento
6. É de simples criação e manutenção

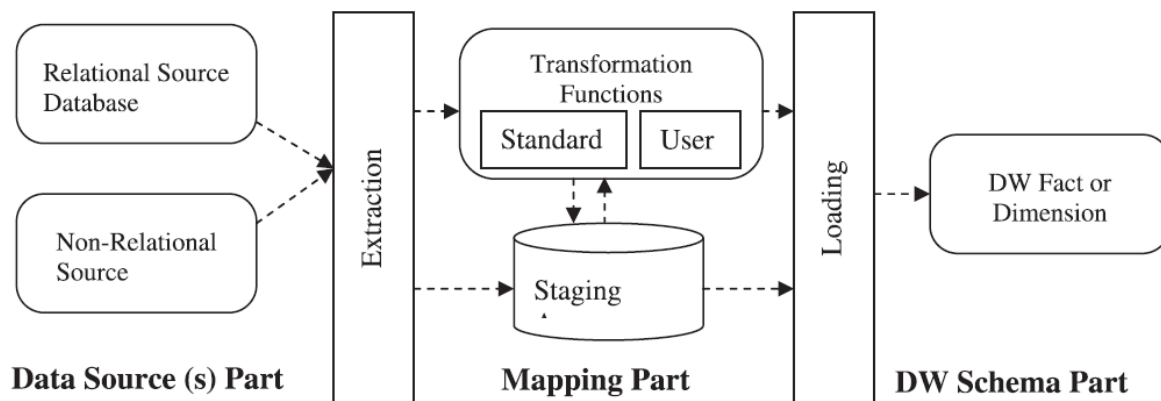


Figura 12 - A *framework* genérica de EMD - Retirado de: (El-Sappagh et al., 2011)

- Na parte de fonte(s) de dados (*Data Source(s) Part*) desenham-se as fontes de dados envolvidas que podem referir-se a bases de dados estruturadas ou a fontes não estruturadas, existindo, neste caso, a necessidade conversão dos dados para uma forma estruturada.
- No momento da extração (*Extraction*) poderá existir a necessidade de criação de tabelas temporárias para armazenar o resultado da estruturação de dados das fontes não estruturadas. Esta fase pode contemplar dois cenários, um para a extração inicial e outro para os refrescamentos, o que pode implicar a construção de dois modelos, um para cada cenário.
- Na parte do esquema do DW (*DW Schema Part*) é desenhado o esquema de dados das tabelas do DW, numa estrutura de modelo relacional.
- A fase de mapeamento (*Mapping*) contempla o desenho das funções de transformação que se realizam na área de estágio dos dados.
- A área de estágio (*Staging area*) dos dados representa um repositório que contempla as tabelas temporárias necessárias para a transformação dos dados, sendo o resultado destas transformações guardados nas mesmas.
- O carregamento (*Loading*) representa o envio dos dados, depois de transformados para formato adequado, para a tabela de destino.

Os autores deixam ainda a ressalva de que, antes do desenho do EMD quer as fontes de dados, quer o DW de destino devem estar já claramente definidos.

Para este modelo conceptual, os autores apresentam também a arquitetura do metamodelo, que pode ser visualizado na Figura 13. Este metamodelo é composto por duas camadas, a camada de abstração

(*abstraction layer*) e a camada do modelo (*Template layer*). A camada de abstração contempla cinco tipos de objetos - repositório de dados (*data container*), função (*function*), entidade(*entity*), relação (*relationship*) e atributo (*attribute*) – que representam uma visão de alto nível dos componentes ou objetos que podem ser utilizados para o desenho do cenário do EMD. A camada modelo apresenta-se como uma expansão da camada de abstração, podendo as ligações entre as camadas representadas ser consideradas ligações de agregação.

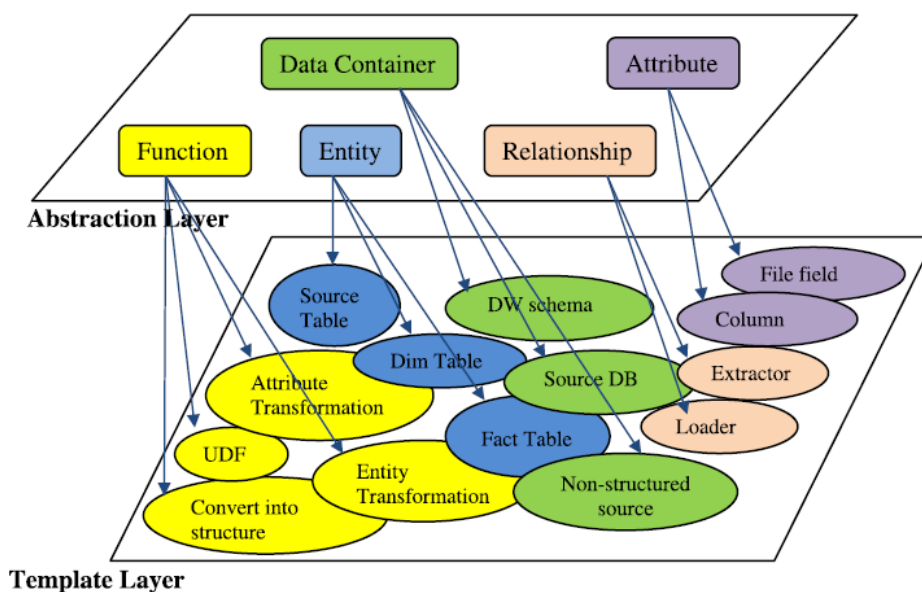


Figura 13 - Metamodelo da EMD - Retirado de: (El-Sappagh et al., 2011)

Para os objetos do tipo função, El-Sappagh et al. (2011), referem ainda um conjunto de transformações, com a respetiva notação, para aplicação no contexto da EMD, que podem ser visualizadas na Figura 14. Relativamente às entidades, podem ser realizadas, entre outras, operações de união, intersecção, junção. Os atributos podem sofrer operações de conversão de tipo, de formato ou de valor. As transformações definidas pelo utilizador contemplam qualquer transformação não presente no modelo, geradas para a resolução de um problema específico do fluxo de dados. Já a conversão em estrutura refere-se às operações de transformação de dados provenientes de fontes não estruturadas, em dados estruturados.

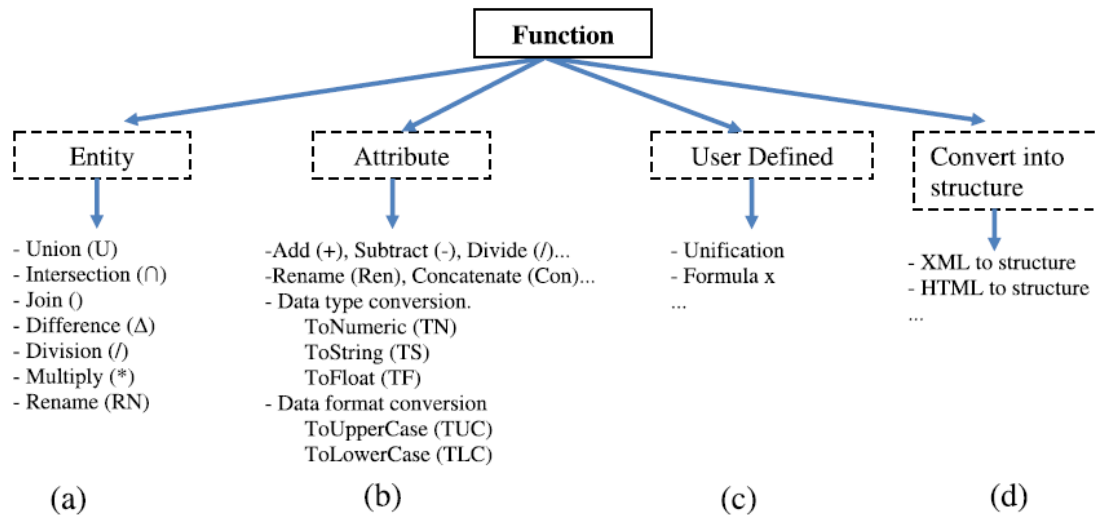


Figura 14 - Tipos de transformações na EMD - Retirado de: (El-Sappagh et al., 2011)

Para o desenho da EMD os autores apresentam um conjunto de construtores gráficos, conforme a Figura 15, que são a seguir detalhados:

- Relação de carregamento (*loader relationship*) – utilizada para estabelecer ligação entre o último elemento fonte de dados (a fonte de dados atual ou uma temporária) e o elemento de destino.
- Relação de carregamento opcional (*optional loader relationship*) – utilizada quando os dados a carregar podem ter proveniência em duas fontes distintas.
- Conversão em estrutura (*convert into structure*) – usada para representar, quando necessário, a transformação de dados não estruturados em dados estruturados.
- Transformação de entidade (*entity transformation operation*) – utilizada para a representação das transformações das entidades através dos operadores indicados na Figura 14 (a).
- Transformação de atributo (*attribute transformation operation*) – usada para a representação de transformações comuns dos atributos, como as apresentadas em Figura 14 (b).
- Função definida pelo utilizador (*user defined function UDF*) – para os casos em que as transformações necessárias não se encontram definidas, utiliza-se este construtor para representar as transformações criadas pelo utilizador.
- Fonte não estruturada (Non-structured source) – utilizada para representar uma fonte de dados que não possua uma estrutura relacional, tal como ficheiros *eXtensible Markup Language* (XML).






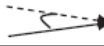





Mapping Construct		To Represent
Name	Shape	
Cylinder		Schema
Rectangle		Entity
Oval		Attribute
Diamond with rounded arrow		Convert into structure
Solid arrow		Loader Relationship
Connected arrows		Optional Loader Relationship
Square with rounded edge		Attribute Transformation
Square with triangle edge		User Defined Function (UDF)
Hexagon		Entity Transformation operation
Document		Non-structured source
Rectangle with folded corner		User Note

Figura 15 - Construtores gráficos da EMD - Retirado de: (El-Sappagh et al., 2011)

Os autores apresentam ainda um exemplo de aplicação da EMD para uma organização que pretenda construir um DW com informação dos processos de venda de duas filiais.

A primeira filial vende livros, estando o respetivo modelo de dados (DS1) desta fonte representado na Figura 16.

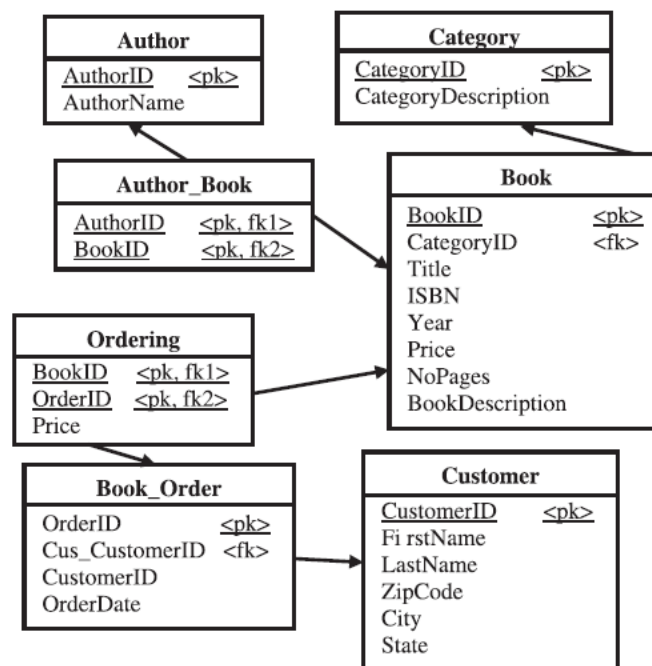


Figura 16 - Esquema relacional da DS1 - Retirado de: (El-Sappagh et al., 2011)

A segunda filial vende produtos generalizados, entre eles livros, e representa uma fonte de dados cujo modelo pode ser visualizado na Figura 17.

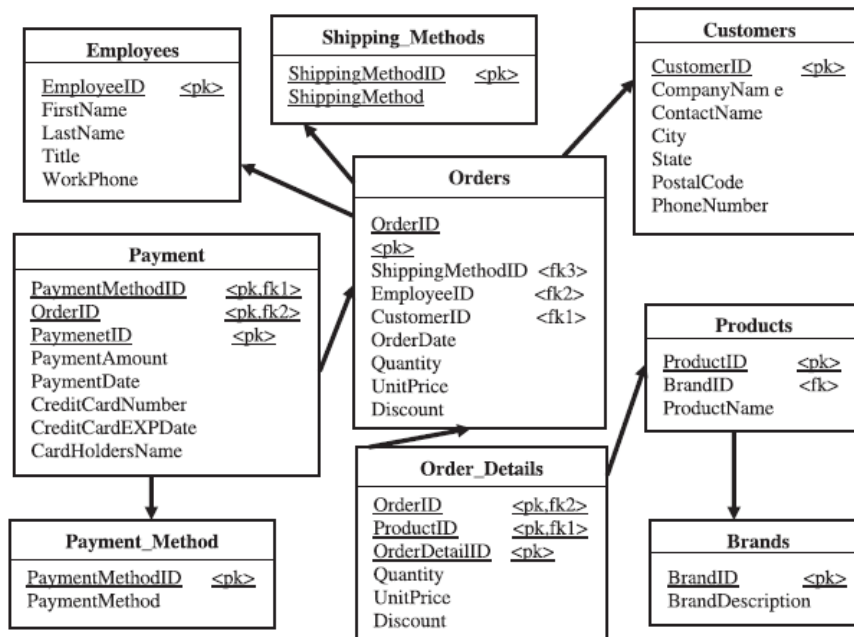


Figura 17 - Esquema relacional da DS2 - Retirado de: (El-Sappagh et al., 2011)

Para a construção do DW os autores propõem o esquema em estrela representado na Figura 18.

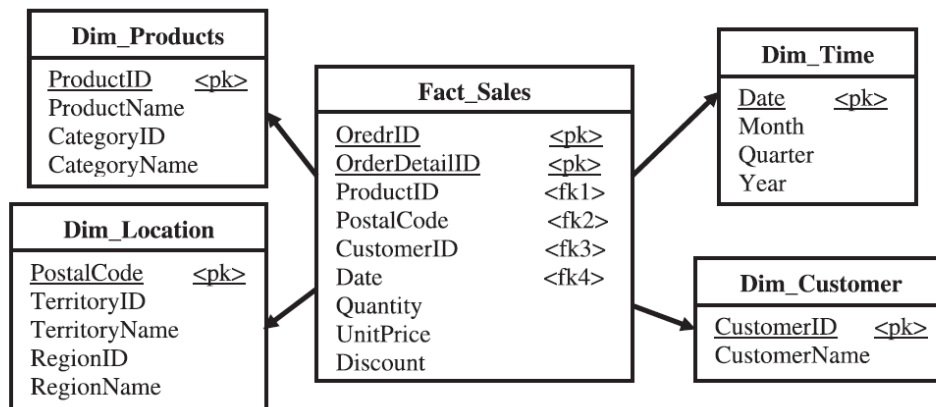


Figura 18 - Esquema em estrela para o DW1 - Retirado de: (El-Sappagh et al., 2011)

Com base neste cenário, em que dispõem de duas fontes de dados distintas, depois de identificadas todas as operações necessárias para o processo de ETL da dimensão Produto (*Dim_Product*) os autores desenharam o cenário da Figura 19.

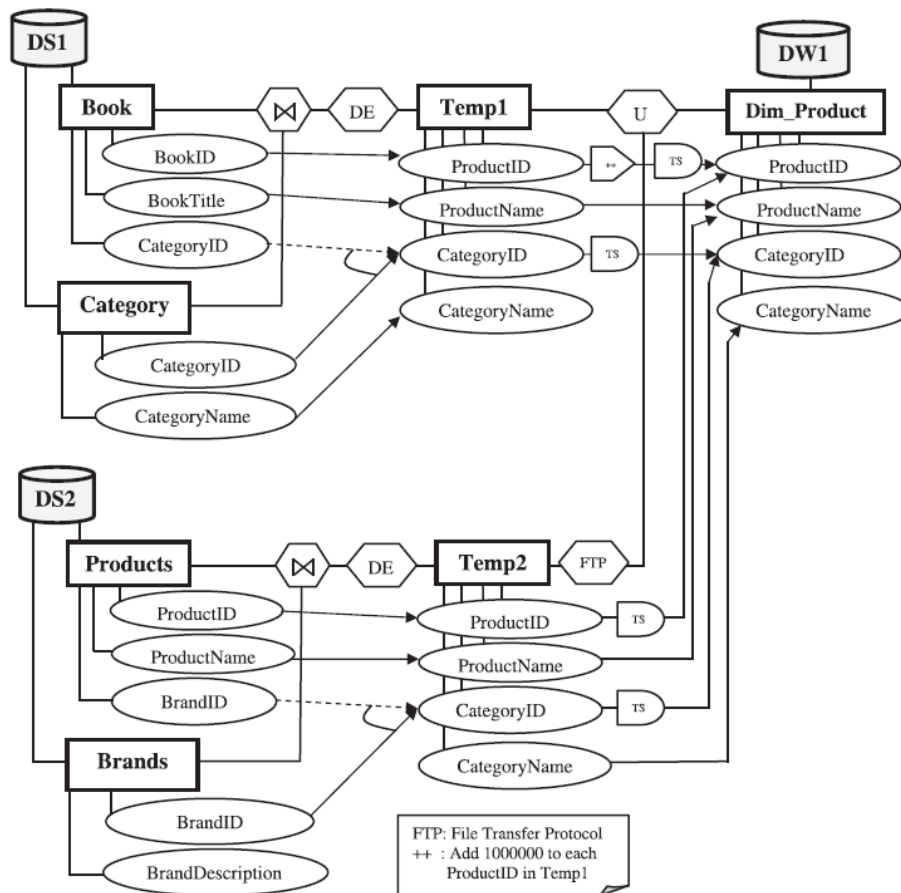


Figura 19 - Cenário EMD para a dimensão Produto - Retirado de: (El-Sappagh et al., 2011)

Neste cenário as fontes DS1 e DS2 referem-se às fontes de dados livros e produtos, respetivamente. Em cada uma das fontes estão definidas duas entidades, para a DS1, LIVRO (*Book*), com os atributos *BookID*, *BookTitle*, *CategoryId* e CATEGORIA (*Category*), com os atributos *CategoryId* e *CategoryName*. Na DS2 existem as entidades PRODUTOS (*Products*) com os atributos *ProductId*, *ProductName* e *BrandId* e MARCAS (*Brands*) com os atributos *BrandId* e *BrandDescription*. O DW de destino dos dados encontra-se definido no construtor DW1. Na parte interior do diagrama estão definidos um conjunto de processos de mapeamento que se iniciam com uma operação de junção (*Join*) das tabelas LIVRO e CATEGORIA seguidos de uma operação de eliminação de duplicados (DE), sendo os dados de seguida enviados para uma tabela temporária TEMP1. Os dados provenientes de DS2 passam por um processo similar sendo de seguida enviados para a tabela TEMP2. Após estas operações realizam-se algumas transformações de dados, como caso representado em TS, em que se convertem os valores para o tipo de dados *String*. Finalmente define-se uma operação de união (U) das tabelas TEMP1 e TEMP2, desenhando-se ainda os conectores de carregamento ligam cada um dos atributos das tabelas TEMP ao atributo correspondente na tabela DIM_PRODUCT do DW1.

2.5 Tecnologias Consideradas

Segundo Eckerson e White (2003), nos últimos anos as necessidades de *Business Intelligence* (BI) têm mudado substancialmente. As fontes de dados são agora mais diversas o que provoca um volume de dados cada vez maior e complexo. É neste contexto de dinamismo que as plataformas de integração de dados surgem como uma evolução natural das ferramentas de ETL contribuindo com novas características como, segundo o mesmo documento, captura avançada de dados, atualizações incrementais e qualidade de dados.

No relatório *Vendor Landscape: Data Integration Tools* elaborado pela *Info~Tech Research Group* publicado em 2013 é apresentado um conjunto de empresas que disponibilizam soluções de plataformas de integração de dados. Desse relatório surge a Figura 20 onde são classificadas a posição das empresas no mercado, resultado de uma análise ponderada da própria empresa segundo parâmetros de:

- Viabilidade, que inclui a performance financeira, tamanho da base de dados de clientes e presença no mercado
- Estratégia, pondera o tipo de estratégia da empresa para com o mercado e para os diferentes tipos de clientes (pequenos, médios e grandes)
- Alcance, capacidade de suporte pós venda da empresa a nível global
- Canais, nível de comunicação com cliente parceiro e respetivas estratégias para o reforço dessa relação;

E do produto:

- Características, se o produto é diferenciador e competitivo
- Funcionalidade, referente nível de componentes de software incorporado, capacidade de integração de dados e nível de utilização intuitiva;
- Preço, custo total de aquisição do produto a 3 anos
- Arquitetura, nível de integração com as restantes ferramentas já existentes no cliente, facilidade de lançamento e nível de aplicabilidade da plataforma

¹ <http://www.infotech.com/research/ss/vendor-landscape-data-integration-tools>

Desta ponderação resultam 4 classificações:

- Campeões, tanto a avaliação do produto como da empresa são acima da média;
- Inovadores, a avaliação do produto é acima da média mas a empresa teve um resultado abaixo da média;
- Empresas Consolidadas, o produto é avaliado abaixo da média mas a empresa está avaliada acima da média;
- Empresas Emergentes, tanto a avaliação do produto como da empresa são abaixo da média.

Importa também referir que a média considerada é relativa aos resultados do grupo que a Info~Tech considerou para o relatório.

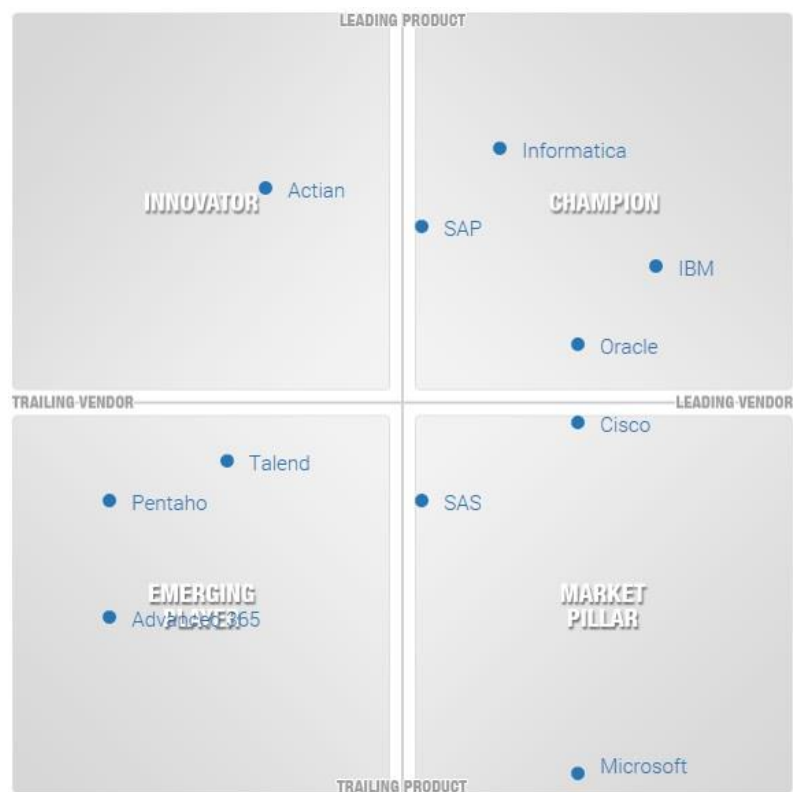


Figura 20 - Classificação de Empresas fornecedoras de soluções de DI – Retirado de <http://www.infotech.com/research/ss/vendor-landscape-data-integration-tools> (2013)

Da observação da figura as soluções comercializadas pela SAP, IBM, Oracle e Informatica seriam as classificadas como *Campeãs* ou tidas como exemplares. Contudo, uma vez que estas soluções apresentam um custo elevado para o projeto em questão, a decisão iria recair entre a *Microsoft*, *Talend* e *Pentaho*. A primeira opção por estar disponível gratuitamente devido à existência de protocolos com a

Universidade do Minho em relação aos seus produtos e as últimas duas porque oferecem uma solução grátis. Pelo que faz todo o sentido expor o resultado do relatório da Info~Tech acerca destas 3 empresas em vez de fazer da totalidade do grupo que foi analisado.

A Info~Tech caracteriza estas empresas da seguinte forma:

- “Microsoft – uma empresa na liderança na programação e desenvolvimento de *software*, com solução de integração de dados através da ferramenta *Microsoft SQL Server*, tem um conjunto de ferramentas ETL que suportam a integração de aplicações especializadas no processamento de dados e a extração e transformação de dados de uma grande variedade de fontes;
- Pentaho - empresa especializada e líder em integração de *Big Data* para o uso de análises;
- Talend – empresa que oferece soluções *open-source* e de subscrição, que permitem a integração de dados em tempo real através da utilização de ferramentas ETL para integração de dados, depuração, migração e sincronização e BI.

Da observação da Figura 21, atentando especificamente a análise ao produto, o que sobressai é que apesar de a *Microsoft* ser uma empresa consolidada no mercado não é aquela que oferece a melhor solução segundo os parâmetros avaliados. De facto, a *Info~Tech* destaca a *Pentaho* como uma empresa vanguardista por se especializar na integração de dados para BI e de utilizar uma arquitetura que facilita a integração do *Big Data*. Já pela análise ao produto conclui-se que a *Talend* é a que se destaca.

Se considerarmos o aspeto das vantagens de parceria o cenário explicado anteriormente é totalmente oposto a esta análise. O *know-how* e experiência de mercado da *Microsoft* como *big player* claramente fica destacado aqui e se a prioridade na escolha da solução for a parceria e apoio a *Microsoft* seria a escolha óbvia.

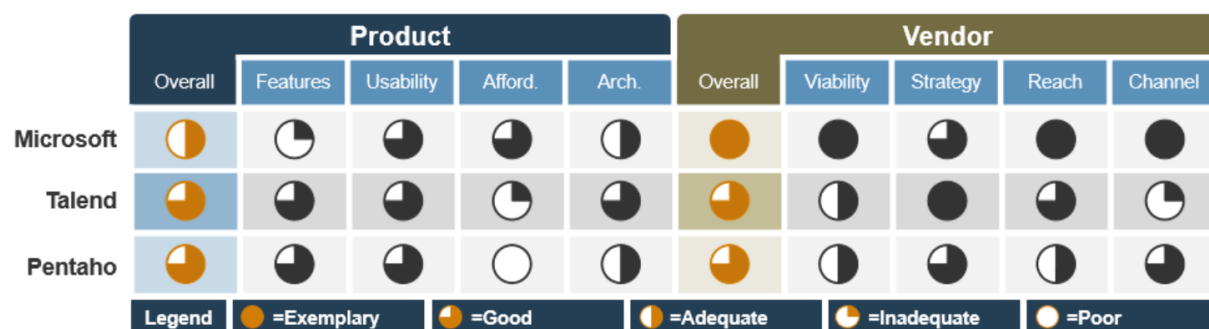


Figura 21 - Quadro de ponderação dos parâmetros das empresas – Adaptado de <http://www.infotech.com/research/ss/vendor-landscape-data-integration-tools> (2013)

O passo seguinte será, naturalmente, observar o resultado da análise ao produto segundo o conjunto de características definidas pela Info~Tech que por motivos de contextualização se passa agora a expor o que eles significam para essa empresa:

- *Big Data Connectivity* – capacidade em integrar com tecnologia relacionada com *Big Data* e bases de dados *NoSQL*;
- *Cloud Deployment*, - capacidade de periodicamente disponibilizar dados integrados na *cloud*;
- *Data Virtualization* – capacidade de criar visualizações e fundir dados das fontes mais diversas e dispares;
- *Mobile* – capacidade de disponibilizar soluções de aplicações móveis ou de HTML5 para funções administrativas;
- *Data Quality and Profiling*- capacidade de gerir, fundir, caraterizar e reconciliar dados de várias fontes e em lote;
- *Change Data Capture* (CDC) – capacidade de monitorizar alterações nas fontes dos dados sem ser invasiva;
- *Data Masking*- ser capaz de ocultar dados de informação sensível de forma anónima ou com o uso de caracteres especiais;
- *Exception Handling* – ser capaz recuperar dados após erros;
- *Middleware Compatible* – compatibilidade com as principais aplicações de dados e de comunicação;
- *Data Semantics/ Context* – capacidade de resolver conflitos de semântica e contexto entre as várias fontes de informação

A Figura 22 contribui para uma melhor compreensão acerca do resultado que a *Microsoft* obteve ao nível do produto. Apesar de ser uma empresa consolidada no mercado, não conseguiu destacar-se, a este nível, das empresas emergentes *Talend* e *Pentaho*. Na realidade, da observação da figura a *Microsoft* só está ao mesmo nível que as restantes empresas na vertente de *Data Quality and Profiling*, quando nas restantes características, comparativamente à *Talend* e *Pentaho*, estão ausentes, parcialmente presentes ou em desenvolvimento. A *Talend* e a *Pentaho* são bastante idênticas em termos de oferta de características dos seus produtos. As áreas de diferença serão *Cloud deployment* e *Mobile* onde a *Pentaho* se destaca e na *Change Data Capture* onde neste caso a *Talend* tem a característica presente no seu produto.

		Evaluated Features									
		Big Data Connectivity	Cloud Deployment	Data Virt.	Mobile	Data Quality & Profiling	CDC	Data Masking	Exception Handling	Middleware Compatible	Data Semantics
Microsoft		●	●	●	●	●	●	●	●	●	●
Talend		●	●	●	●	●	●	●	●	●	●
Pentaho		●	●	●	●	●	●	●	●	●	●
Legend		● =Feature fully present			● =Feature partially present/pending			● =Feature absent			

Figura 22 - Quadro comparativo de caraterísticas de produto - Adaptado de <http://www.infotech.com/research/ss/vendor-landscape-data-integration-tools> (2013)

A *Info Tech* descreve que a *Microsoft* (*SQL Server*) e a *Pentaho* são mais adequadas para pequenas e médias empresas enquanto que a *Talend* já consegue fazer escalonamento a nível das grandes empresas.

Atendendo a que a estrutura tecnológica para o presente projeto é disponibilizada pelo Departamento de Tecnologias e Sistemas de Informação (DTSI) da Universidade do Minho, a seleção da plataforma de integração para a realização do projeto, ficou a cargo desta entidade, tendo sido a solução *SQL SERVER 2012 Business Intelligence Edition* da *Microsoft* o recurso selecionado, o que vai também de encontro às recomendações da *Info Tech*, conforme descrito no parágrafo anterior.

3 O SISTEMA DE RECONSTITUIÇÃO DE PARÓQUIAS E A SUA BASE DE DADOS

Neste capítulo apresenta-se a aplicação SRP, ferramenta que alimenta as BDP a serem integradas na BDC. Descrimina-se ainda o modelo de dados da BDP, com uma breve exposição das tabelas que a constituem. Realiza-se ainda o levantamento das restrições que este modelo apresenta no registo de informação dos RP. Por último expõe-se ainda o modelo de dados proposto para a BDC, realizando-se uma breve descrição das principais tabelas deste modelo.

3.1 O Sistema de Reconstituição de Paróquias

O SRP (Sistema de Reconstituição de Paróquias) é uma ferramenta informática desenvolvida por Fernanda Faria (Faria, 2004) para servir de instrumento de suporte à MRP, permitindo o registo e o cruzamento de informação recolhida dos RP.

A aplicação apresenta dois formulários principais (Figura 23 e Figura 24) para tratar as duas entidades principais do sistema: Indivíduo e Família.

The screenshot displays the 'Indivíduo' form within the SRP application. The window title is 'SRP - Sistema de Reconstituição de Paróquias / BDP: Sé e São Pedro'. The menu bar includes 'BDParoquial', 'Outras Fontes', 'Manutenção', 'Ajuda', and 'Sair'. The form fields are as follows:

- Nº Ind.:** 26
- Nome:** António
- Indivíduo:** 18
- Tipo Ind.:** 1 (Indivíduo nascido na paróquia)
- Sexo:** M (Male)
- Família de Origem:** [Empty field]
- Data Nasc.:** 19-10-1737
- Local Nasc.:** Sé e São Pedro - F- 070520
- Filiação:** Exposto
- Intervenientes:** [Empty field]
- Pai (Information at birth):**
 - Idade: [Empty field]
 - Estado civil: [Empty dropdown]
- Mãe (Information at birth):**
 - Idade: [Empty field]
 - Estado civil: [Empty dropdown]
- Observações (Obs.):** [Empty text area]
- Buttons on the right:** Profissões, Residências, Assinaturas, Apadrinham.
- Buttons at the bottom:** [Navigation icons], Atribuir, Localizar, Inserir, Alterar, Eliminar, Fechar.
- Footer:** Desenvolvido por Fernanda Faria - NEPS/DI - Universidade do Minho (31/08/2004 - V. 1.0.6)

Figura 23 - Interface INDIVÍDUO no SRP

O formulário “Individuo” permite a recolha da informação relativa aos registos de nascimento e óbito de um indivíduo. Para os RP de nascimento, permite registar, entre outros dados, o nome, sexo, data e local do ato, bem como alguma informação relativa aos pais no momento no nascimento. Permite-nos ainda associá-lo a uma família (família de origem).

No caso dos RP de óbito recolhe-se a informação da data e local de óbito correspondente, o estado civil do indivíduo e ainda alguma informação relativa à cerimónia fúnebre.

Em ambos os casos existirá informação adicional relativa ao ato e/ou ao Indivíduo, como por exemplo, as testemunhas do ato, profissão do indivíduo, ou a residência do mesmo, que podem ser também registadas a partir deste formulário.

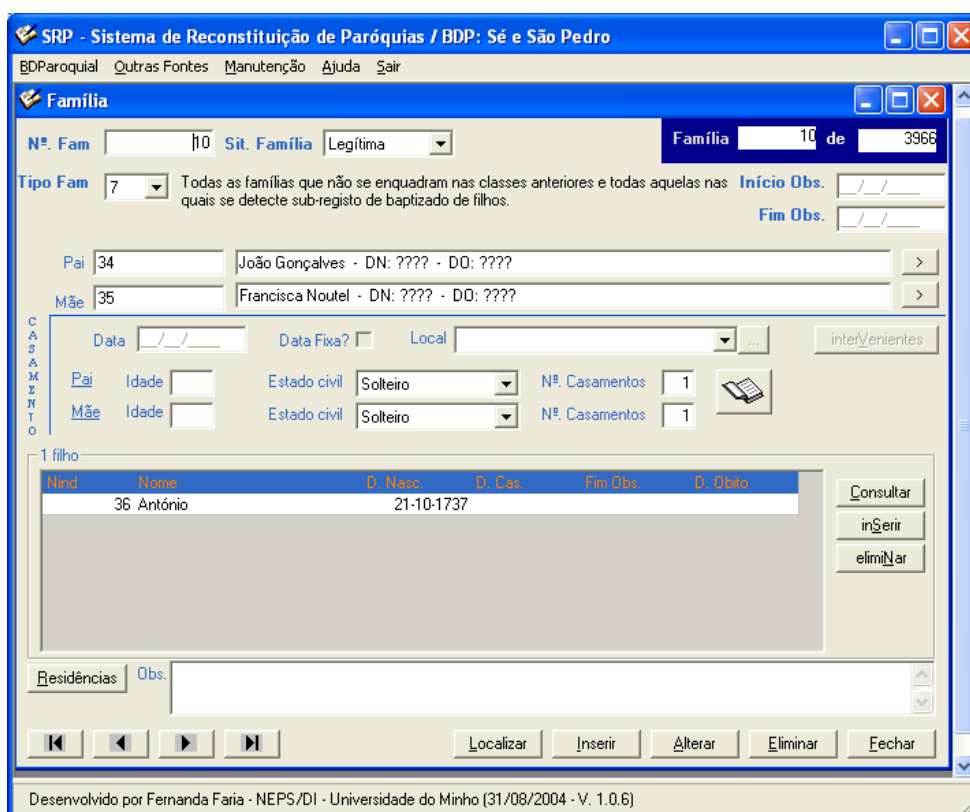


Figura 24 - Interface FAMÍLIA no SRP

O formulário “Família” permite estabelecer a criação de uma família através da criação de uma relação entre dois indivíduos, com informação oriunda, tendencialmente, dos RP de Casamento. No caso de ambos serem naturais da mesma paróquia procura-se identificar as respetivas fichas de indivíduo, criando assim uma relação entre dois indivíduos já representados na BDP. Caso um ou os dois indivíduos não estejam ainda registados, deverá ser introduzida toda a informação referente aos mesmos (no formulário “Individuo”) criando-se depois esta relação familiar neste formulário.

Em qualquer dos atos poderá existir a necessidade de constituir novas fichas de indivíduo e/ou de família, consoante os indivíduos ou famílias envolvidas nos atos estejam, ou não, já registadas no sistema. Por exemplo, num RP de Casamento além da identificação dos noivos, estará também presente a identificação dos pais. O primeiro passo será o da verificação, através dos mecanismos de pesquisa do SRP, da existência de cada um desses indivíduos na BDP. Se algum estiver presente, completa-se a nova informação na ficha adequada, se não existir ainda, cria-se novo registo, com toda a informação disponível no RP. Assumindo que o registo de casamento se refere a uma noiva já presente na BDP e um noivo, oriundo de outra paróquia, não registado ainda, criar-se-ia nova ficha de indivíduo para o noivo, sendo este depois associado à ficha de Família relativa a este ato de casamento, em conjunto com a ficha da noiva. Estando presente a identificação dos pais do noivo, e assumindo novamente a inexistência na BDP dos registos dos mesmos, criar-se-ia uma ficha de indivíduo para cada um dos progenitores que seriam depois associados numa família, que estaria depois referenciada como família de origem para o noivo. Deste modo estabelecem-se sempre que possível, as relações do indivíduo com encadeamento genealógico.

3.1.1 O Modelo de Dados do Sistema de Reconstituição de Paróquias

Como referido anteriormente, da aplicação SRP resulta uma base de dados em formato *Microsoft Access* por cada paróquia tratada. Na Figura 25 podemos visualizar o Diagrama de Entidades e Relacionamentos (DER) da referida BD.

O modelo de dados da BDP assenta, naturalmente, nas duas entidades principais do contexto da MRP, Indivíduo e Família. A informação da entidade indivíduo armazena-se na tabela INDIVIDUO, existindo um conjunto de tabelas ligadas, como é o caso de PROFISSOES e RESIDENCIAS, para a inserção da respetiva informação nos diferentes momentos em que surgem nos RP.

Neste modelo de dados, uma família é uma associação de um par de indivíduos (marido e mulher). Para o registo desta entidade existem duas tabelas, FAMILIA e FAMILIAS. A tabela FAMILIA permite o registo de informação relativa à constituição da mesma, tal como a data, o local e a situação. Na tabela FAMILIAS efetua-se a associação dos indivíduos que constituem o casal à respetiva família. Esta tabela contempla, à partida, duas linhas para cada família, cada uma delas com a identificação da família a que pertence e com a identificação do indivíduo associado.

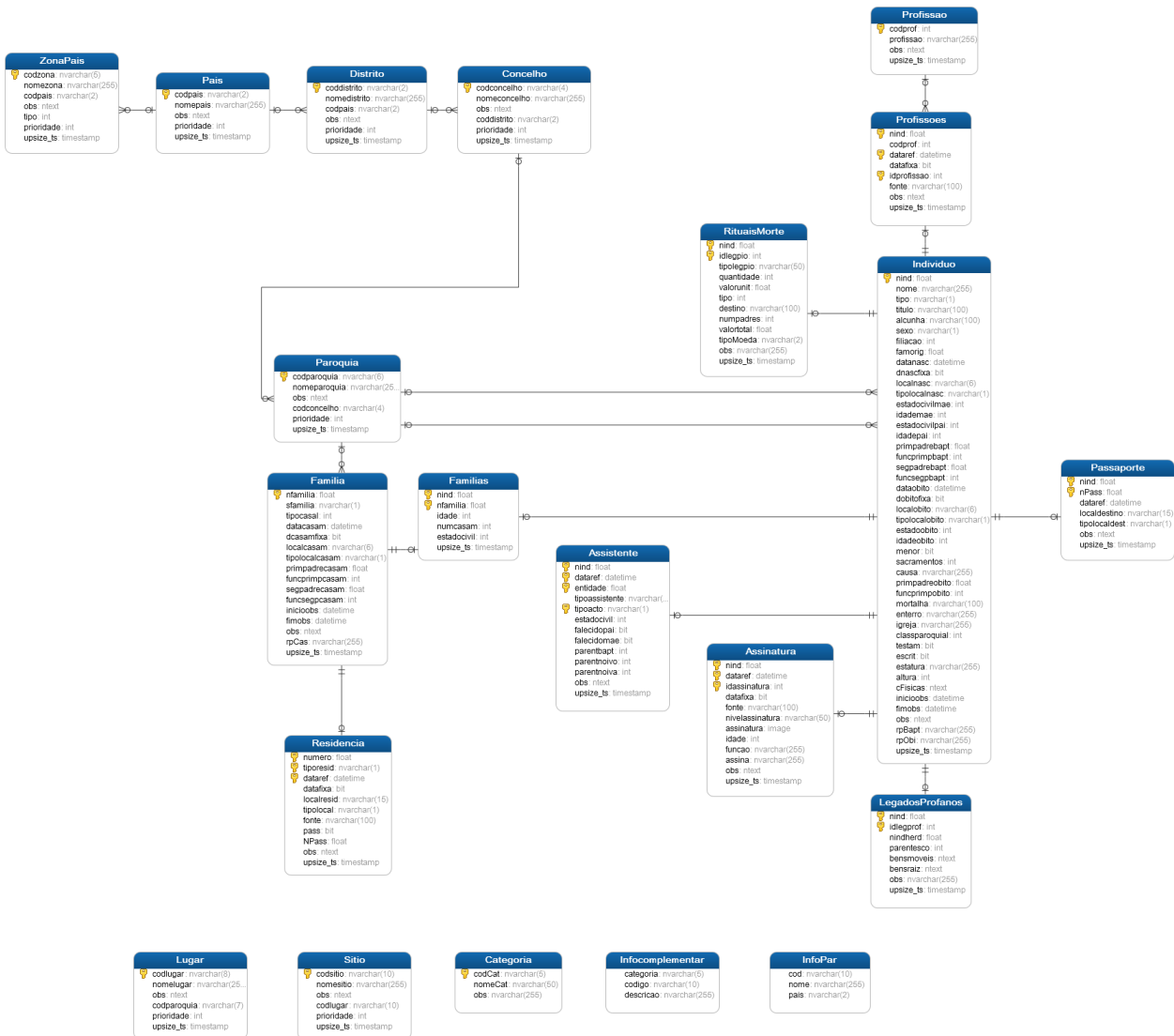


Figura 25 - Diagrama de Entidades e Relacionamentos da BD do SRP

A relação entre o indivíduo (filho) e a respetiva família de origem (pais), estabelece-se através do campo *famorig* na tabela INDIVIDUO, onde se indica o número da família (*nfamilia*) que os pais constituíram. Cada um dos pais é, por sua vez, um indivíduo no sistema que terá também a indicação da família de origem, no respetivo campo, estabelecendo-se assim o encadeamento genealógico.

Todas as datas no sistema são armazenadas recorrendo a campos do tipo Data, no entanto, no presente contexto, há situações em que surgem referências incompletas a datas nos RP, por exemplo, a alguém nascido em 1880. Dado que para o armazenamento de um atributo do tipo data numa BD é necessário

fornecer sempre uma data completa, por cada data presente nas tabelas criou-se um campo “*datafixa*” do tipo booleano, definido por defeito para Verdadeiro. Caso a data a registar se apresente incompleta, completa-se a informação em falta, devendo o campo “*datafixa*” respetivo ser assinalado como Falso para sinalizar esta situação. Esta solução resolve apenas parte do problema dado que, não existe possibilidade de se distinguir quais os componentes que estão corretos, ou seja, não é possível distinguir entre uma referência com mês e ano e outra apenas com o ano, corretos.

Para o armazenamento das referências espaciais, foi utilizada uma estrutura em hierarquia que contempla em tabelas ligadas PAÍS, DISTRITO, CONCELHO, PARÓQUIA, LOCAL e SÍTIO, cada uma delas ligada ao nível hierárquico imediatamente superior. As tabelas desde o nível PAÍS até PARÓQUIA são estáticas, podendo o utilizador registar para cada paróquia 100 novos locais e para cada local, 100 novos sítios. Existe ainda uma tabela ZONAPAIS que permite o armazenamento de locais relativos a outros países que não Portugal.

Nas diversas tabelas da BDP, as referências aos locais fazem-se pela associação de dois campos, um denominado “*tipoLocal*” que identifica na hierarquia o nível geográfico a que se refere, por exemplo, “L” para local ou “C” para concelho, e outro com o código do local com o valor da chave estrangeira da tabela a que se refere. Permite-se assim o registo dos locais em diferentes níveis, consoante a informação patente nos RP.

3.1.2 Descrição das Principais Tabelas do Sistema de Reconstituição de Paróquias

Apresentam-se neste ponto as principais tabelas da BDP usada pela aplicação SRP com uma breve descrição do tipo de informação que armazenam e dos atributos que contêm.

A tabela INDIVIDUO, representada na Tabela 6, permite o armazenamento da informação de um conjunto de atributos do indivíduo como o nome, título, alcunha, sexo, datas e locais nascimento e óbito, bem como a identificação dos padres intervenientes nos atos. Registam-se aqui também a classificação do tipo de indivíduo e algumas informações relativas aos pais, no momento do seu nascimento.

Nesta representação, as chaves primárias e estrangeiras estão identificadas com as siglas PK e FK, respetivamente, colocadas junto no nome do campo correspondente.

Tabela 6 - Tabela INDIVIDUO na BDP

Tabela	Individuo	
Descrição	Propriedades do individuo	
Campo	Tipo	Descrição
nind : PK	double	Nº do individuo
nome	nvarchar	Nome do individuo
tipo : FK	nvarchar	Classificação da MRP
titulo	nvarchar	Título do individuo
alculha	nvarchar	Alculha do individuo
sexo	nvarchar	Sexo do individuo
filiacao : FK	int	Filiação do individuo (Legítimo, Ilegítimo...)
famorig : FK	Float	Família de origem
datanasc	datetime	Data de nascimento
dnascfixa	Bit	Indica se a data de nascimento é conhecida é completa (01/01/1800) ou incompleta (Jan de 1800)
localnasc	nvarchar	Local de nascimento
tipolocalnasc	nvarchar	Tipo de local de nascimento (F => Paróquia, C => Concelho,..., P => País)
estadocivilmae : FK	int	Estado civil da mãe ao nascimento deste individuo
idademae	int	Idade da mãe ao nascimento deste individuo (indicada no registo de batismo)
estadocivilpai : FK	int	Estado civil do pai ao nascimento deste individuo
idadepai	int	Idade do pai ao nascimento deste individuo (indicada no registo de batismo)
primpadrebapt : FK	Double	Primeiro padre do batismo
funcprimpbapt : FK	int	Função do primeiro padre do batismo
segpadrebapt : FK	Float	Segundo padre do batismo
funcsegpbapt : FK	int	Função do Segundo padre do batismo
dataobito	Datetime	Data de óbito
dobitofixa	Bit	Indica se a data de óbito é conhecida é completa (01/01/1800) ou incompleta (Jan de 1800)

Tabela	Individuo	
Descrição	Propriedades do individuo	
Campo	Tipo	Descrição
localobito : FK	nvarchar	Local do óbito
tipolocalobito	nvarchar	Tipo de local do óbito (F => Paróquia, C => Concelho,..., P => País)
estadoobito : FK	Int	Estado civil do individuo ao óbito
idadeobito	Int	Idade ao óbito (indicada no registo de óbito)
menor	bit	Indica se o individuo era menor no momento do óbito
sacramentos : FK	Int	Sacramentos aplicados
causa	nvarchar	Causa de morte
primpadreobito : FK	Float	Primeiro padre do óbito
funcprimobito : FK	Int	Função do primeiro padre do óbito
mortalha	nvarchar	Descrição da mortalha
enterro	nvarchar	Descrição do enterro
igreja	nvarchar	Identificação da igreja
classparoquial : FK	Int	Classificação do individuo por parte do padre
testam	Bit	Deixou testamento?
escrit	Bit	Deixou escritura?
inicioobs	Datetime	Início de observações
fimobs	Datetime	Fim de observações
obs	nvarchar	Observações
estatura	nvarchar	Estatura
altura	Int	Altura
rpBapt	nvarchar	Localização do RP de batismo
rpObi	nvarchar	Localização do RP de óbito
cFisicas	nvarchar	Descrição física

A entidade Família é armazenada na tabela FAMILIA, conforme Tabela 7. Para cada família introduzida no sistema é registada uma linha nesta tabela onde são registadas as informações da classificação da mesma, o local e data de casamento, bem como a identificação dos padres intervenientes no ato.

Tabela 7 - Tabela FAMILIA na BDP

Tabela	Família	
Descrição	Informação da família	
Campo	Tipo	Descrição
nfamilia : PK	float	Nº da família
sfamilia : FK	nvarchar	Legítima, Ilegítima
tipocasal : FK	int	Classificação da Prof Norberta
datacasam	Datetime	Data de casamento
dcasamfixa	Bit	Indica se a data de casamento é conhecida é completa (01/01/1800) ou incompleta (Jan de 1800)
localcasam : FK	nvarchar	Local de casamento
tipocalcasam	nvarchar	Identifica o tipo de local de casamento (F => Paróquia, C => Concelho,..., P => País)
primpadrecasam : FK	float	Primeiro padre do casamento
funcprimpcasam : FK	int	Função do primeiro padre do casamento
segpadrecasam : FK	float	Segundo padre do casamento
funcsegpcasam : FK	int	Função do Segundo padre do casamento
inicioobs	datetime	Início de observações
fimobs	datetime	Fim de observações
obs	nvarchar	Observações
rpCas	nvarchar	Localização do RP de casamento

A Tabela 8 contempla a associação dos indivíduos a uma família. Por cada família registada em FAMILIA existirão duas linhas na tabela FAMILIAS, cada uma com a identificação do indivíduo associado à família, bem como da família respetiva. Para cada um dos indivíduos regista-se ainda a o estado civil e a idade ao casamento, se declarados no RP.

Tabela 8 - Tabela FAMILIAS na BDP

Tabela	Famílias	
Descrição	Associação dos pares de indivíduos numa família (cada família tem dois registos nesta tabela, um para cada indivíduo)	
Campo	Tipo	Descrição
nind : PK	Float	Nº do indivíduo
nfamilia: PK	Float	Nº da família
idade	int	Idade do individuo ao casamento
numcasam	int	Nº de casamentos
estadocivil : FK	int	Estado civil do individuo ao casamento

No presente modelo de dados existe um conjunto de categorias com valores normalizados para o preenchimento de determinados atributos nas tabelas principais, tais como as classificações dos indivíduos e famílias, ou o estado civil. A tabela INFOCOMPLEMENTAR, Tabela 9, contempla o conjunto de valores admissíveis para cada uma dessas categorias.

Tabela 9 – Tabela INFOCOMPLEMENTAR na BDP

Tabela	Infocomplementar	
Descrição	Descrição propriedades diversas dos indivíduos e família	
Campo	Tipo	Descrição
categoria	nvarchar	Identificação da categoria
codigo	nvarchar	Código da categoria
descricao	nvarchar	Descrição da categoria

Nas Tabela 10 e Tabela 11 expõem-se as tabelas CONCELHO e PAROQUIA, respetivamente, presentes na BDP. Cada uma delas apresenta um valor *cod...* para a chave primária, a descrição do respetivo local e um campo onde se podem registar algumas observações. Cada uma delas tem, como já indicado, uma chave estrangeira ligada ao nível hierárquico geográfico imediatamente anterior, de forma a identificar o nível superior a que pertencem. As tabelas PAIS, DISTRITO, LUGAR, SITIO e ZONAPAIS apresentam estrutura semelhante.

Tabela 10 - Tabela CONCELHO na BDP

Tabela	Concelho	
Descrição	Lista dos Concelhos	
Campo	Tipo	Descrição
codconcelho : PK	nvarchar	Código do concelho
nomeconcelho	nvarchar	Descrição do concelho
obs	nvarchar	Observações
coddistrito : FK	nvarchar	Código do distrito a que pertence

Tabela 11 - Tabela PAROQUIA na BDP

Tabela	Paróquia	
Descrição	Lista das Paróquias	
Campo	Tipo	Tabela:Valor (Chaves estrangeiras)
codparoquia : PK	nvarchar	Código do Paróquia
nomeparoquia	nvarchar	Descrição do Paróquia
obs	nvarchar	Observações
codconcelho : FK	nvarchar	Código do concelho a que pertence

O registo das residências quer dos indivíduos, quer das famílias é realizado na tabela RESIDENCIA, que ostenta o modelo apresentado na Tabela 12. Identifica-se aqui o código da entidade, bem como o tipo de entidade (indivíduo ou família) a que se refere. Regista-se ainda, se disponível a data da residência, com possibilidade de se indicar uma data completa e ainda o local a que se refere, com o par de atributos *localresid* para o código do local e *tipolocal* para a identificação da tabela correspondente.

Tabela 12 - Tabela RESIDENCIA no SRP

Tabela		
Residência		
Descrição		
Registo das residências dos indivíduos e das famílias		
Campo	Tipo	Descrição
numero : PK	double	Código do indivíduo ou família a que se refere
tiporesid	nvarchar	Tipo de entidade a que se refere: Indivíduo ou Família
dateref : PK	Datetime	Data de referência
datafixa	Bit	Identificador de data completa
localresid : FK	nvarchar	Código do local
tipolocal	nvarchar	Tipo de local
obs	nvarchar	Observação

3.1.3 Limitações do Sistema de Reconstituição de Paróquias

Esta aplicação foi concebida para acompanhar o investigador na recolha de informação relativa a uma paróquia e embora se tenha apresentado como uma ferramenta de muito valor no suporte à metodologia MRP, apresenta, no entanto, algumas limitações, umas relativas à aplicação em si, outras relacionadas com o modelo de dados, que se pretendem suprimir, nomeadamente:

- É uma aplicação monoposto e monoutilizador
- Acesso apenas local (na máquina onde está instalada)
- Cria uma base de dados isolada por paróquia (não permite ligações a outras paróquias, o que apresenta um obstáculo ao acompanhamento dos indivíduos registados no sistema, dada a mobilidade crescente das populações)
- A pesquisa está limitada a um conjunto de campos que não cobrem todas as necessidades dos investigadores
- As pesquisas não têm filtros pelo que podem apresentar uma gama de resultados demasiado alargada
- A pesquisa é sensível a caracteres acentuadas (“João” não devolve “Joao”)
- Não permite pesquisar em intervalos de datas
- Não permite o registo de datas incompletas
- Permite apenas o registo de 100 locais por cada paróquia
- Dificuldades na alteração da associação de um indivíduo a uma família
- Para os países estrangeiros não permite adicionar hierarquia de locais
- Não permite o registo de evolução de alguns estados da Família ou do Indivíduo (Se indivíduo era ilegítimo e foi legitimado, não é possível registar esta evolução)
- Não regista qual o utilizador que inseriu ou editou a informação

- Não regista as alterações efetuadas na base de dados (*logs*)
- Não permite o registo de mais do que uma residência para um indivíduo/família para uma mesma data
- Permite o registo de apenas um título e uma alcunha por cada indivíduo
- Não contempla atributos para o registo de algumas informações presentes nos RP
 - Nascimento
 - Permite apenas o registo da data de nascimento ou de batismo
 - Permite apenas a inserção de 2 padres como intervenientes no batismo
 - Passaporte
 - Não permite a indicação de "Passaporte coletivo"
 - Não permite o registo da "Origem do indivíduo"
 - Não permite o registo do navio
 - Não permite o registo de "Acompanhantes"
 - Não permite o registo normalizado de características físicas
 - Olhos
 - Cabelo
 - Sobrancelhas
 - Rosto
 - Nariz
 - Boca
 - Cor
 - Óbito
 - Causa da morte não normalizada (Texto livre)
 - Idade ao óbito admite apenas valores inteiros em anos (não permite registar idades inferiores a um ano)
 - Não permite o registo do local da sepultura (Igreja ou outro)
 - Mortalhas não normalizadas (Texto livre)
 - Impossibilidade do registo do "Acompanhamento do funeral"
 - N° de padres
 - N° de côcos
 - Tumba
 - Valor
 - Caixão
 - Descrição
 - Irmandades
 - Acompanhantes
 - Observações
 - Família
 - Não permite o registo da "Dissolução do casamento" caso aconteça

3.2 Base de Dados Central Para o Repositório Genealógico Nacional

A convergência das BDP para uma BDC terá, por partida, o estudo de um modelo de dados capaz de garantir a preservação de toda a informação consistente existente. Pretende-se assegurar, conforme apresentam Batini e Scannapieco (2006), níveis máximos de completude de esquema - presença de todas as propriedades - e de população - presença de todos os registos. Após o estudo detalhado do modelo de dados da BDP, que teve como finalidade o conhecimento em detalhe do tipo de informação armazenada e o modo como os dados e entidades no modelo se relacionam, e atentando-se nas limitações apresentadas pelo cliente, desenhou-se e propôs-se o modelo de dados para a BDC.

Este modelo apresenta-se como uma evolução do modelo da BDP, tirando-se assim proveito dos pontos fortes que o mesmo ostenta, permitindo-se ainda uma mais fácil adaptação das técnicas de análise já implementadas pelo GHP, a esta nova realidade dos dados ao mesmo tempo que se suprimem os constrangimentos anteriormente identificados.

Este novo modelo foi depois implementado no MS SQL SERVER 2012 que, além do elevado desempenho que apresenta, suporta o acesso multiutilizador, bem como o acesso remoto à mesma.

Optou-se pela codificação de caracteres para a BDC como *Latin1_General_CI_AI* o que permite que as pesquisas em campos de texto não sejam sensíveis nem a maiúsculas/minúsculas nem a caracteres acentuados, facilitando as tarefas de pesquisa dos investigadores.

Procedeu-se a uma implementação do registo de todas as datas referentes aos RP em campos separados, Anos, Mês, Dia e quando aplicável, Hora e Minuto, permitindo armazenar fielmente a informação presente nos RP.

As classificações de tipo Individuo (*Legítimo, Ilegítimo,...*) e de situação da Família passam a ser registadas em tabelas à parte, de modo a poderem ser criados vários registos para esta propriedade e, consequentemente registrar-se a evolução destes estados, caso aconteçam.

Os títulos e alcunhas do indivíduo passam a ser registado em tabelas próprias, permitindo o armazenamento de vários valores para cada indivíduo.

Mantem-se a mesma estrutura de locais desde o nível País até Local, tendo-se eliminado o registo do sítio, que passa agora a ser registado como um local, de modo a facilitar o registo destas referências.

Todas as referências a locais passam a ser feitas com granularidade ao local.

Foram criadas novas tabelas para permitir o registo de algum tipo de informação, não prevista no modelo das BDP, nomeadamente para os registos de óbito, cerimónias fúnebres e passaportes.

Remodelaram-se as tabelas FAMILIA e FAMILIAS, registando-se agora informação correspondente nas tabelas FAMILIA, que contém a identificação dos elementos do casal, bem como algumas propriedades

dos mesmos no momento da constituição da família, e na tabela CASAMENTO, para o caso das famílias Legítimas, onde se regista a informação relativa ao ato (Data, local). Permite-se agora, também na tabela CASAMENTO, o registo de informação da dissolução do mesmo, caso aconteça.

Todas as tabelas passam a dispor de campos para o registo do utilizador que registou e alterou a informação da mesma, com a indicação do momento em que tal aconteceu.

Por cada uma das propriedades normalizadas do modelo (estado civil, sexo, classificação do indivíduo,...) foi criada uma tabela, estabelecendo-se depois, nas tabelas principais, restrições de integridade através da definição de chaves estrangeiras referentes a estas tabelas.

3.2.1 O Modelo de Dados Para a Base de Dados do Repositório Genealógico Nacional

O modelo de dados proposto para a BDC do RGN pode ser visualizado no diagrama de entidades e relacionamentos da Figura 26, cuja representação não contempla os atributos das tabelas.

Dada a dimensão do esquema, este será incluído em anexo, em formato de maior dimensão. No entanto, pode-se desde já averiguar a expansão realizada ao modelo das BDP através do número de novas tabelas implementadas.

No centro do modelo encontram-se representadas as entidades principais, INDIVIDUO e FAMILIA, estando todas as outras tabelas ligadas a estas entidades ou a tabelas de registos de informações auxiliares das mesmas, como é o caso das PROFISSOES, RESIDENCIAINDIVIDUO, RESIDENCIAFAMILIA, FILIACAO ou TIPOFAMILIA.

Apresentam-se, no ponto seguinte, em detalhe, as principais tabelas deste novo modelo de dados, com a identificação dos seus atributos, o tipo de dados dos mesmos, bem como uma breve descrição dos dados que representam.

Nesta representação, as chaves primárias e estrangeiras estão identificadas com as siglas PK e FK, respetivamente, colocadas junto no nome do campo correspondente.

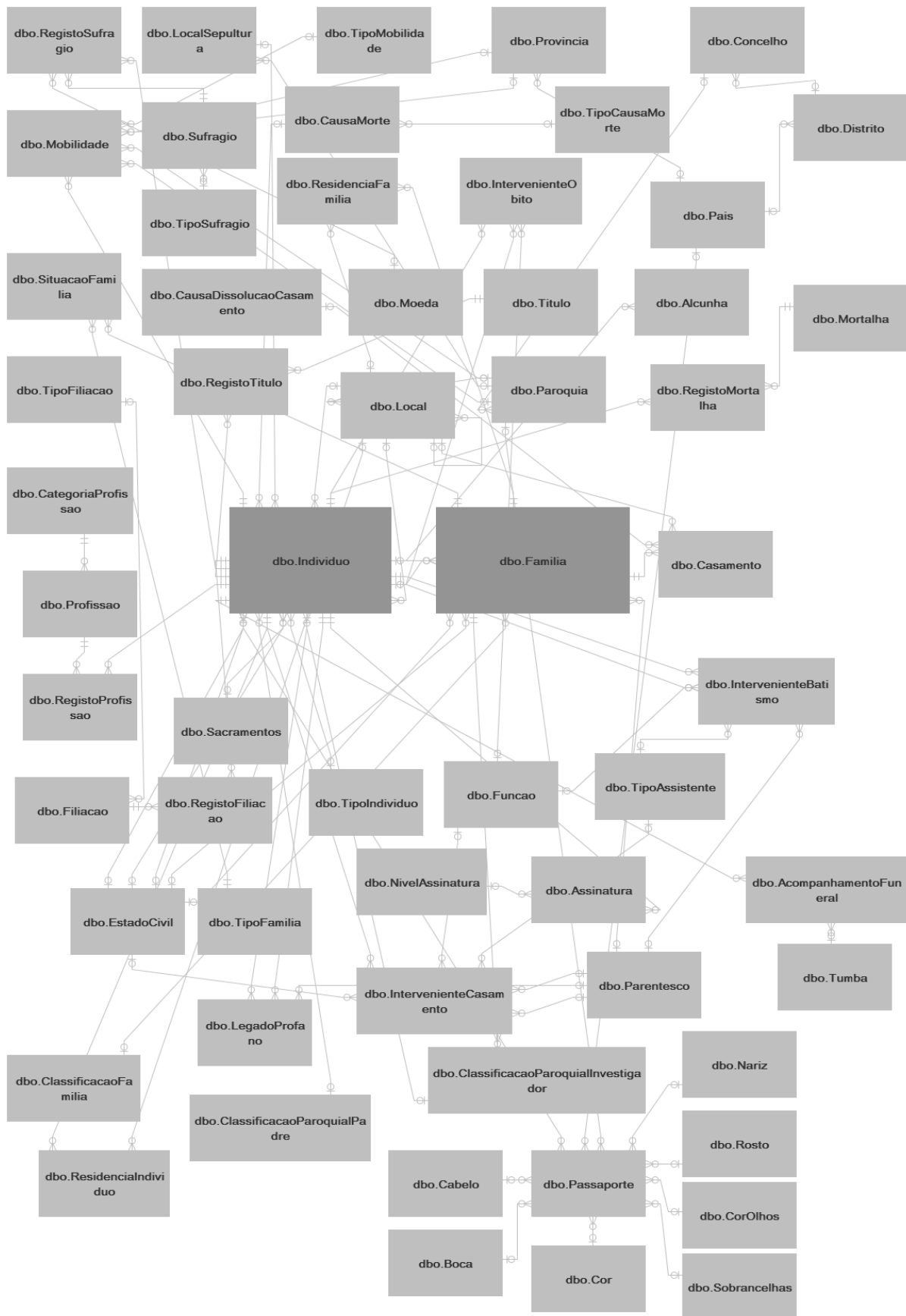


Figura 26 - Diagrama de Entidades e Relacionamentos do novo modelo de dados

3.2.2 Principais Tabelas da Base de Dados do Repositório Genealógico Nacional

A tabela INDIVIDUO foi expandida, relativamente ao modelo da BDP, com a inclusão de novos atributos, tal como apresentado na Tabela 13. Todas as datas são agora armazenadas por componentes (ano, mês, dia) em campos separados, permitindo registar fielmente a informação disponibilizada nos RP. Os locais de nascimento e óbito foram definidos para uma granularidade ao Local, devendo-se, nos casos em que não se conhece a informação com este nível de detalhe, registar o evento como Local desconhecido para a respetiva paróquia, Concelho ou Distrito.

Foram criados campos de observações distintos para se poderem separar os comentários relativos ao nascimento e ao nascimento.

Preservam-se ainda os campos *nind* e *famorig*, preenchidos caso o registo tenha sido importado de uma BDP, para possibilitar a consulta, caso necessária, destes registos na BDP de origem.

Tabela 13 - Tabela INDIVIDUO na BDC

Tabela	Individuo	
Descrição	Propriedades do indivíduo	
Campo	Tipo	Descrição
IdIndividuo : PK	int	Número do indivíduo na BDC
IdIndividuoSRP	int	Número do indivíduo na BDP
NomeIndividuo	Nvarchar	Nome do indivíduo
Sexo : FK	Char	Sexo do indivíduo
LocalNascimento : FK	Int	Local de nascimento
FamiliaOrigem : FK	Int	Família de origem na BDC
FamiliaOrigemSRP	Int	Família de origem na BDP
TipoIndividuo : FK	Int	Tipo de indivíduo
LocalObito : FK	Int	Local óbito
NumCasamentos	Int	Número de casamentos
IdadeAoObito	Decimal	Idade ao óbito
MenorAoObito	Bit	Menor ao óbito
CausaMorte : FK	Int	Causa da morte
ObservacaoNascimento	Texto	Observações ao nascimento
ObservacaoMorte	Texto	Observações da morte

Tabela	Individuo	
Descrição	Propriedades do indivíduo	
Campo	Tipo	Descrição
Sacramentos : FK	Int	Sacramentos aplicados
CausaSacramentos	Texto	Causa da aplicação dos sacramentos
Testamento	Bit	Deixou testamento
Escritura	Bit	Deixou escritura
EstadoCivilObito : FK	Int	Estado civil ao óbito
ClassificacaoParoquialPadre : FK	Int	Classificação paroquial do padre
ClassificacaoParoquialInvestigador : PK	Int	Classificação paroquial do investigador
EstadoCivilMaeAoNascimento : FK	Int	Estado civil da mãe ao nascimento
IdadeMaeAoNascimento	Int	Idade da mãe ao nascimento
EstadoCivilPaiAoNascimento : FK	Int	Estado civil do pai ao nascimento
IdadePaiAoNascimento	Int	Idade do pai ao nascimento
AnoNascimento	Int	Ano do nascimento
MesNascimento	Int	Mês do nascimento
DiaNascimento	Int	Dia do nascimento
HoraNascimento	Int	Hora do nascimento
MinutoNascimento	Int	Minuto do nascimento
AnoObito	Int	Ano de óbito
MesObito	Int	Mês do óbito
DiaObito	Int	Dia do óbito
HoraObito	Int	Hora do óbito
MinutoObito	Int	Minuto do óbito
LocalSepultura : FK	Int	Local da sepultura
AnoIniObservacao	Int	Ano de início de observação
MesIniObservacao	Int	Mês de início de observação
DiaIniObservacao	Int	Dia de início de observação
AnoFimObservacao	Int	Ano de fim de observação
MesFimObservacao	Int	Mês de fim de observação

Tabela	Individuo	
Descrição	Propriedades do indivíduo	
Campo	Tipo	Descrição
DiaFimObservacao	Int	Dia de fim de observação
AnoBatismo	Int	Ano do batismo
MesBatismo	Int	Mês do batismo
DiaBatismo	Int	Dia do batismo
HoraBatismo	Int	Hora do batismo
MinutoBatismo	Int	Minuto do batismo
CUser	Nvarchar	Criador o registo
CDate	Datetime	Data de criação
UUser	Nvarchar	Atualizador do registo
UDate	Datetime	Data de atualização
ParoquiaSRP : FK	Int	BDP de origem
LocalNascimentoDescr	Nvarchar	Local de nascimento descritivo
LocalObitoDescr	Nvarchar	Local de óbito descritivo

A tabela FAMILIA, Tabela 14, contempla agora a identificação dos indivíduos constituintes da mesma, bem como de alguma informação relativa aos mesmos, no momento da sua constituição, tal como o seu estado civil, ou a idade declarada no RP.

Tabela 14 - Tabela FAMILIA na BDC

Tabela	Família	
Descrição	Propriedades da família	
Campo	Tipo	Descrição
IdFamilia : PK	int	Número da família na BDC
IdFamiliaSRP	int	Número da família na BDP
IdIndMarido : FK	int	Número do marido na BDC
IdIndMaridoSRP	int	Número do marido na BDP
EstadoCivilMarido : FK	int	Estado civil do marido ao casamento

Tabela	Família	
Descrição	Propriedades da família	
Campo	Tipo	Descrição
IdadeMaridoCasamento	int	Idade do marido ao casamento
NumCasamentoMarido	int	Número de casamentos
IdIndMulher : FK	int	Número da mulher na BDC
IdIndMulherSRP	int	Número da mulher na BDP
IdadeMulherCasamento	int	Idade da mulher ao casamento
EstadoCivilMulher : FK	int	Estado civil da mulher ao casamento
NumCasamentosMulher	int	Número de casamentos
CUser	nvarchar	Criador o registo
CDate	datetime	Data de criação
UUser	nvarchar	Atualizador do registo
UDate	datetime	Data de atualização
ParoquiaSRP : FK	Int	BDP de origem

Caso a família seja legítima, existirá nos RP informação relativa ao ato de casamento, que fica preservada na Tabela 15, tal como a data e local do ato. Permite-se agora o registo, caso aconteça, da data e da causa de dissolução do casamento. Existem aqui dois campos de observações, um para o casamento e outro para a dissolução.

Tabela 15 - Tabela CASAMENTO na BDC

Tabela	Casamento	
Descrição	Propriedades do Casamento	
Campo	Tipo	Descrição
IdCasamento : PK	int	Número do casamento
IdFamilia : FK	int	Número da família a que corresponde
AnoIniObservacao	int	Ano de início de observação
MesIniObservacao	int	Mês de início de observação
DiaIniObservacao	int	Dia de início de observação

Tabela	Casamento	
Descrição	Propriedades do Casamento	
Campo	Tipo	Descrição
AnoFimObservacao	int	Ano de fim de observação
MesFimObservacao	int	Mês de fim de observação
DiaFimObservacao	int	Dia de fim de observação
ClassificacaoFamilia : FK	int	Classificação da família
AnoCasamento	int	Ano do casamento
MesCasamento	int	Mês do casamento
DiaCasamento	int	Dia do casamento
LocalCasamento	int	Local do casamento
ObsCasamento	text	Observações do casamento
ParentescoCasal	int	Parente entre o casal
AnoDissolucao	int	Ano da dissolução do casamento
MesDissolucao	int	Mês da dissolução do casamento
DiaDissolucao	int	Dia da dissolução do casamento
CausaDissolucao : FK	int	Causa da dissolução do casamento
ObsDissolucao	text	Observações da dissolução do casamento
CUser	nvarchar	Criador o registo
CDate	datetime	Data de criação
UUser	nvarchar	Atualizador do registo
UDate	datetime	Data de atualização
LocalCasamentoDresc	nvarchar	Local de casamento descritivo

Conforme referido, algumas propriedades do indivíduo e da família passam a ser registadas em tabelas separadas de forma a possibilitar o armazenamento de mais do que uma ocorrência dessa propriedade, como é o caso da Tabela 16, onde se registam as várias alcunhas com que cada indivíduo pode aparecer referenciado, bem como a data da sua ocorrência.

Tabela 16 - Tabela ALCUNHA na BDC

Tabela	Alcunha	
Descrição	Registo das alcunhas do indivíduo	
Campo	Tipo	Descrição
IdAlcunha : PK	int	Número da alcunha
IdIndividuo : FK	int	Número do indivíduo a que corresponde
Alcunha	nvarchar	Alcunha
Observacao	Texto	Observações
AnoAlcunha	Int	Ano da alcunha
MesAlcunha	Int	Mês da alcunha
DiaAlcunha	Int	Dia da alcunha
CUser	nvarchar	Criador o registo
CDate	datetime	Data de criação
UUser	nvarchar	Atualizador do registo
UDate	datetime	Data de atualização

As residências das famílias e dos indivíduos passam a ser armazenadas em tabelas separadas, tal como a Tabela 17, onde se registam estas ocorrências para os indivíduos, sempre com uma associação a um local e a uma data.

Tabela 17 - Tabela RESIDENCIAINDIVIDUO na BDC

Tabela	ResidenciaIndividuo	
Descrição	Registo das residências do indivíduo	
Campo	Tipo	Descrição
IdResidenciaIndividuo : PK	int	Número da residência
IdIndividuo : FK	int	Número do indivíduo a que pertence
IdLocal : FK	int	Local da residência
LocalTxt	nvarchar	Descrição do local
Observacao	Texto	Observações
AnoResidenciaIndividuo	Int	Ano da residência
MesResidenciaIndividuo	Int	Mês da residência

Tabela	ResidencialIndividuo	
Descrição	Registo das residências do indivíduo	
Campo	Tipo	Descrição
DiaResidencialIndividuo	Int	Dia da residência
CUser	nvarchar	Criador o registo
CDate	datetime	Data de criação
UUser	nvarchar	Atualizador do registo
UDate	datetime	Data de atualização

4 A INTEGRAÇÃO E CONSOLIDAÇÃO DE DADOS NO REPOSITÓRIO GENEALÓGICO NACIONAL

Após a modelação e implementação da BDC para o RGN, reuniram-se condições para a preparação do processo de integração das BDP nesta nova base. Este processo, conforme referem Naumann & Bleiholder (2006), pode apresentar-se bastante complexo, dado que será necessário revolver um conjunto de problemas para se garantir uma representação completa e concisa dos dados, conforme demonstrado no ponto 2.1.1.

Deparamos então com uma situação em que é necessário fundir um conjunto vasto de BDP, todas com a mesma estrutura, sendo necessário, em primeira instância, garantir a preservação de toda a informação presente em cada uma delas. Sabe-se, à partida, da possibilidade da existência de problemas com os dados, pelo que é necessário proceder a uma avaliação e correção, se necessária, dos mesmos. Numa realidade em que se trabalham isoladamente as paróquias e em que é permitido que o utilizador introduza novos valores para algumas tabelas auxiliares, como as de locais e de profissões, é também expectável a existência de registos em BDP diferentes que se referem a uma mesma entidade e que detêm um valor de chave primária distinto, podendo até apresentar uma ligeira variação na descrição da entidade. Torna-se, por isso, necessária a identificação destas possíveis representações, procedendo-se à inserção na BDC apenas dos valores novos, associando-se os outros ao valor correspondente já registado, de modo a evitar representações redundantes dos dados.

É também conhecida a existência, nas populações das paróquias, de uma dinâmica de mobilidade dos indivíduos e famílias. Por exemplo, um indivíduo que nasceu em determinada paróquia poderá vir a realizar os atos de casamento e de óbito numa outra paróquia qualquer. Este indivíduo deixará a informação do seu percurso de vida distribuída por todas as paróquias onde realizou qualquer ato registado nos RP, o que, de acordo com a metodologia apresentada de recolha de informação, resultará na criação de um registo para esse mesmo indivíduo, em cada uma das BDP associadas às paróquias recolhidas. Torna-se, portanto, necessária a implementação de uma estratégia para a identificação destes indivíduos, de forma a permitir o posterior tratamento da respetiva informação.

Após uma reflexão relativa ao conjunto de tarefas a ser executadas, foi delineado um conjunto de operações, divididas em seis momentos, conforme se pode facilmente verificar na Figura 27.

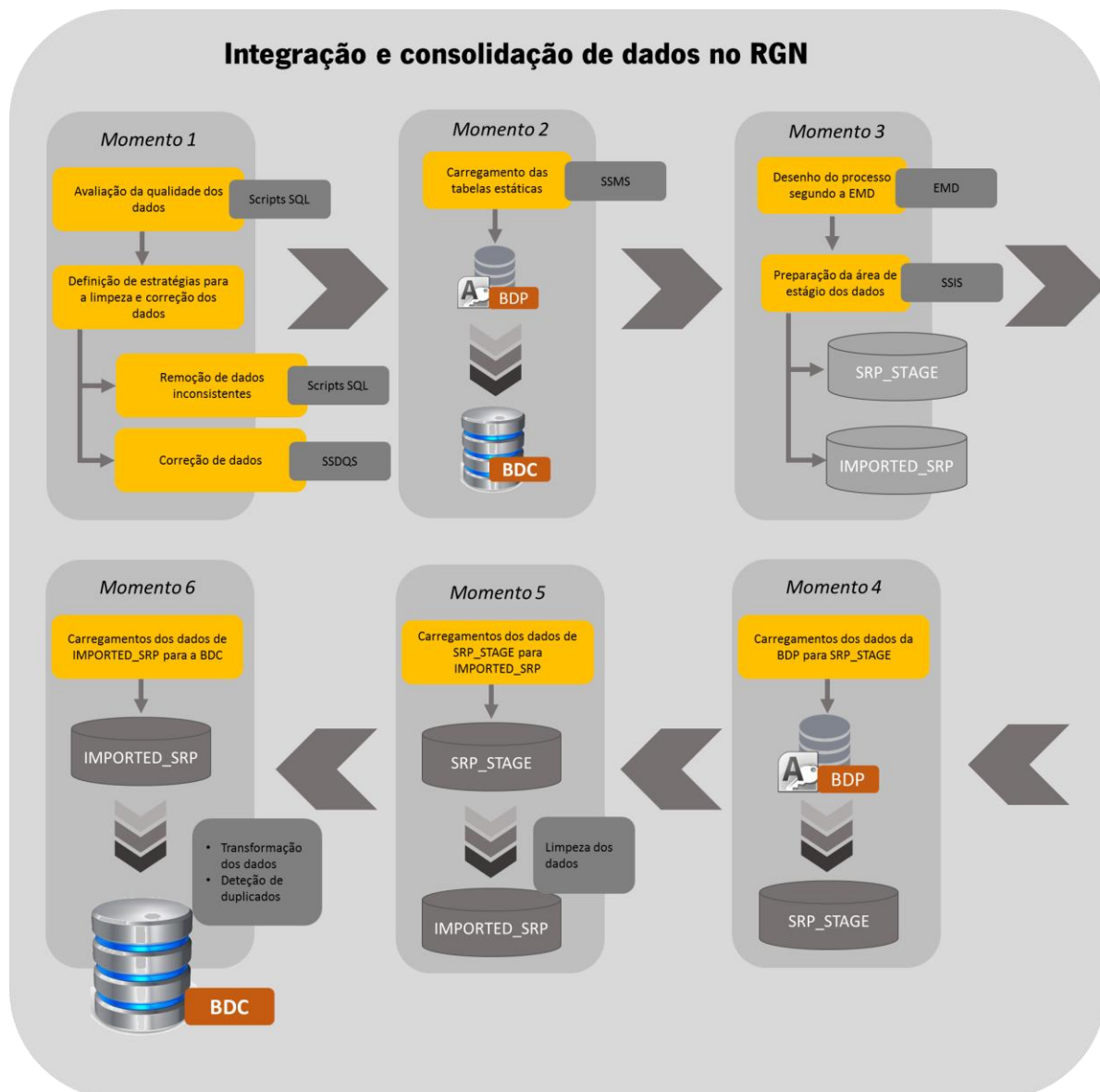


Figura 27 - Perspetiva global da integração e consolidação de dados no RGN

O momento 1, relativo à primeira fase do trabalho, teve início com uma averiguação da qualidade dos dados nas BDP, no sentido de se identificarem potenciais problemas nos mesmos e de se definirem estratégias para a sua correção. De seguida, no momento 2, procedeu-se ao carregamento das tabelas estáticas da BDC. No terceiro momento, desenhou-se o modelo conceptual para os processos de ETL de acordo com a metodologia EMD, tendo-se ainda preparado a área de estágio dos dados. Os momentos 4 e 5 referem-se ao carregamento dos dados para a área de estágio e às operações de limpeza dos mesmos. Por último, no momento 6, procede-se às transformações dos dados necessárias, bem como à deteção de possíveis duplicados, realizando-se posteriormente o carregamento para a BDC. Estas operações estão detalhadas nos pontos a seguir apresentados.

4.1 Avaliação da Qualidade dos Dados

A primeira fase da integração dos dados passará, naturalmente, por uma fase de pré-processamento com vista à otimização da qualidade dos mesmos e, conseqüentemente, à melhoria do desempenho e da eficácia dos processos de integração.

O primeiro momento desta fase passou pela avaliação da qualidade dos dados presentes nas BDP no sentido de se averiguar a sua consistência e a presença de omissões e de erros.

Apesar dos esforços por parte da programação da aplicação e das regras definidas para o Sistema de Gestão e Base de Dados (SGBD), com algumas consultas simples às tabelas das BDP geradas pelo SRP pode-se verificar que alguns campos apresentam valores fora do domínio do contexto em questão, por exemplo, no modelo de dados da BDP, o campo sexo do indivíduo deveria contemplar apenas os valores “m” para indivíduos do sexo masculino, “f” para indivíduos do sexo feminino e “d”, desconhecido, para os casos em que não se consegue identificar o sexo do indivíduo. No entanto, em algumas destas BD consultadas, surgem, para este campo, valores como “0”, “1”, “2”, “i”, “u”, (...). Estes valores podem ser explicados pelo facto de algumas destas BD terem sido criadas por aplicações anteriores ao SRP que contemplavam um modelo de dados próprio, tendo as mesmas, sido já alvo de um processo de migração em que poderão ter ocorrido algumas falhas. Existem ainda alguns problemas, já identificados pelos investigadores do GHP, relacionados com a eliminação dos registos, em que, por exemplo, quando se elimina um indivíduo do sistema que tenha alguma(s) profissão(ões) registada(s) na tabela *PROFISSOES*, este(s) registos de profissão(ões) não são eliminados, ficando a BD com um registo de profissão referente a um indivíduo que não existe.

Para a identificação das situações em que se verificam violações da integridade dos dados, foi criado um conjunto de *scripts* em SQL, como o que pode ser verificada na Figura 28, em que se identificam referências a lugares que não existem na tabela *LUGAR* referenciada.

```

58 --Individuos/Familia com Lugar de nascimento/obito/casamento que não existe na tabela Lugar
59 SELECT INDIVIDUO.nind, INDIVIDUO.localnasc
60 FROM INDIVIDUO
61 LEFT JOIN LUGAR ON INDIVIDUO.localnasc = LUGAR.codlugar
62 WHERE INDIVIDUO.tipolocalnasc = '1'
63        AND (LUGAR.codlugar IS NULL);
64
65 SELECT INDIVIDUO.nind, INDIVIDUO.localobito
66 FROM INDIVIDUO
67 LEFT JOIN LUGAR ON INDIVIDUO.localobito = LUGAR.codlugar
68 WHERE (
69        (INDIVIDUO.tipolocalobito = '1')
70        AND ((LUGAR.codlugar) IS NULL)
71        );
72
73 SELECT FAMILIA.nfamilia, FAMILIA.localcasam
74 FROM FAMILIA
75 LEFT JOIN LUGAR ON FAMILIA.localcasam = LUGAR.codlugar
76 WHERE (
77        (FAMILIA.tipolocalcasam = '1')
78        AND ((LUGAR.codlugar) IS NULL)
79        );

```

Figura 28 - Exemplo de *script* para a identificação de erros nos dados

Com base na análise modelo de dados da BDP e nas respectivas as relações, os *scripts* concebidos permitiram identificar os seguintes problemas:

- Valores para os locais de nascimento, casamento, óbito e residência, presentes nas tabelas *INDIVIDUO*, *FAMILIA* e *RESIDENCIA* como chaves estrangeiras, não têm valores nas tabelas *SITIO*, *LOCAL*, *FREGUESIA*, *CONCELHO*, *DISTRITO*, *PAIS* ou *ZONAPAIS* a que se referem.
- Indivíduos com referência a família de origem não existe na tabela *FAMILIA*
- Assinatura de individuo que não existe em *INDIVIDUO*
- Assistente ou entidade (assistido) que não existe em *INDIVIDUO* ou em *FAMILIA*
- Legados profanos de individuo que não existe em *INDIVIDUO*
- Profissão que não existe em *PROFISSOES* ou de individuo que não existe em *INDIVIDUO*
- Residências de indivíduos que não existem em *INDIVIDUO* ou de famílias que não existem em *FAMILIA*
- Rituais de morte de individuo que não existe em *INDIVIDUO*
- Famílias sem indivíduos
- Famílias (na tabela *FAMILIAS*) que não existem na tabela *FAMILIA*
- Família que não tem famílias associada
- Indivíduos casados sem sexo definido
- Indivíduos com pais do mesmo sexo

Apresentam-se no ponto seguinte as estratégias implementadas para a limpeza e correção dos dados.

4.2 Estratégias de Limpeza e Tratamento dos Dados.

Após um primeiro momento de identificação dos problemas que os dados podem apresentar nas BDP tornou-se necessário definir um conjunto de operações que permitam a limpeza dos mesmos.

Para o tratamento dos problemas da inconsistência dos dados nos casos em que se verificam violações das restrições de integridade, atendendo a que estes registos deverão ser eliminados, criou-se um conjunto de *scripts* que os eliminam e que podem ser aplicados a todos os problemas de inconsistência detetados, à exceção dos casos em que se identificam indivíduos casados sem sexo definido ou indivíduos com pais do mesmo sexo, situação em que se deve corrigir a informação na ficha de indivíduo correspondente. Na Figura 29 pode observar-se um exemplo de um *script* em que se removem registos inconsistentes da tabela *FAMILIA*. No presente modelo de dados, por cada família registada na tabela *FAMILIA* terá sempre de existir pelo menos uma linha na tabela *FAMILIAS*, em que se efetua a associação de um indivíduo a esta mesma família.

```

106 -- Apaga Familia sem Familias
107 DELETE *
108 FROM FAMILIA
109 WHERE FAMILIA.nfamilia IN
110 ( SELECT FAMILIA.nfamilia
111 FROM FAMILIA
112 LEFT JOIN FAMILIAS
113 ON FAMILIA.nfamilia = FAMILIAS.nfamilia
114 WHERE (( ( FAMILIAS.nfamilia ) IS NULL )) );
115

```

Figura 29 - Exemplo de *script* para a remoção de registos inconsistentes

Para os casos em que se verifica a existência de campos com valores fora do domínio e atendendo à necessidade de correção destes valores, optou-se pelo recurso aos “SQL Server Data Quality Services”, conforme explicado no ponto seguinte.

4.2.1 Valores Admissíveis no Contexto da Base de Dados do Sistema de Reconstituição de Paróquias

O modelo de dados da BDP contempla, na tabela *INFOCOMPLEMENTAR*, um conjunto de valores admissíveis para o preenchimento de determinados campos nas demais tabelas da BD, como por exemplo, para o campo *filiação* da tabela *INDIVIDUO*, que só pode assumir os valores {1, 2 ou 3}, consoante o indivíduo seja considerado “Legítimo, Ilegítimo ou Exposto”, respetivamente. Atendendo à situação verificada no ponto 4.1, em que se verifica a existência de valores não admissíveis para este

contexto, foi necessário implementar uma estratégia para correção destes valores, tendo-se recorrido, para este efeito, aos serviços para a qualidade dados integrados no SQL Server, os “SQL Server Data Quality Services”¹ (SSDQS), uma vez que já se encontram presentes na estrutura que suporta o RGN. Nestes serviços é possível a criação de bases de dados de conhecimento em que se constroem domínios, domínios estes que são um conjunto de valores admissíveis para determinado contexto. Assim, é possível criar, por exemplo, um domínio de nome “Sexo” que contém uma lista de valores admissíveis para esta dimensão, neste caso, {m, f, d}. É ainda possível definir regras para a correção destes valores na BD em tratamento, por exemplo, todos os registos com valor {‘i’} deverão ser convertidos no valor “d” como pode ser verificado na Figura 30.

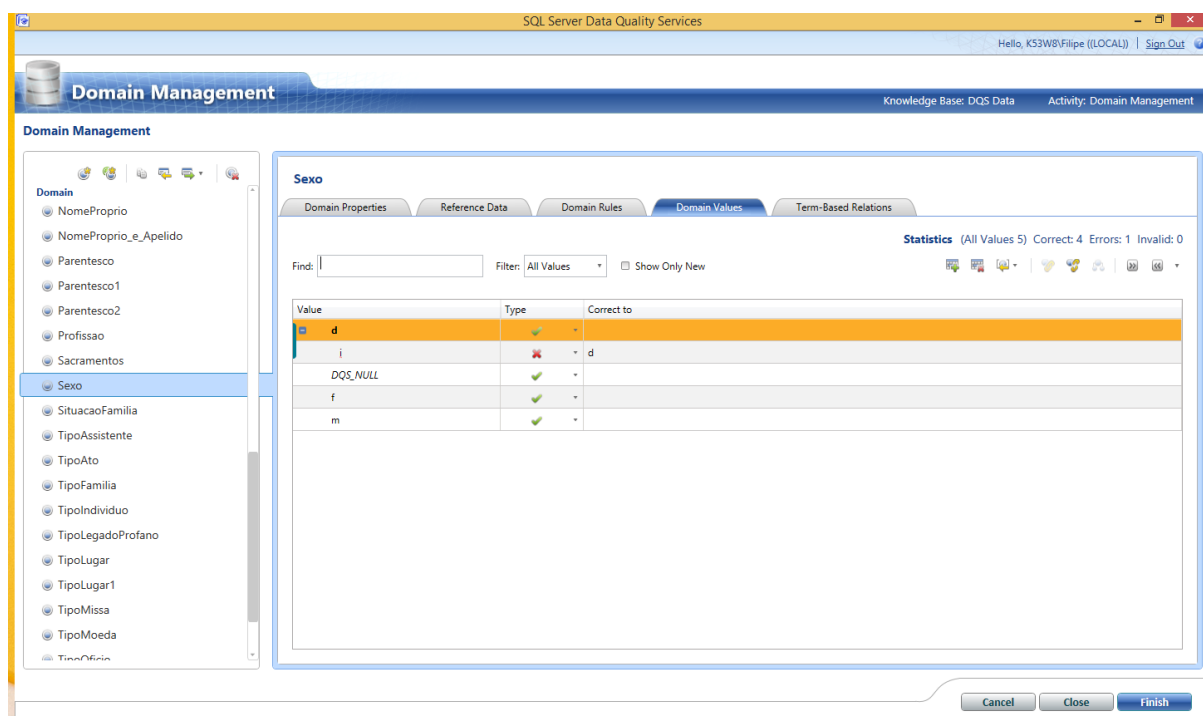


Figura 30 - O domínio "Sexo" no SSDQS

Depois de analisada a BDP e de verificados todos os campos que necessitam de ser validados criaram-se os seguintes domínios com os respetivos valores admissíveis, bem como com as respetivas regras de validação, nos SSDQS:

- ClassificacaoParoquial
- EstadoCivil

¹ [https://msdn.microsoft.com/en-us/library/ff877925\(v=sql.110\).aspx](https://msdn.microsoft.com/en-us/library/ff877925(v=sql.110).aspx)

- Filiacao
- NivelAssinatura
- Parentesco
- Profissao
- Sacramentos
- Sexo
- SituacaoFamilia
- TipoAssistente
- TipoIndividuo
- TipoLegadoProfano
- TipoLugar
- TipoMissa
- TipoMoeda
- TipoOficio
- TipoResidencia

Os SSDQS analisam os dados e, para cada campo comparado, atribuem-lhe um estado de *Corret* (correto), caso este valor esteja em conformidade, *Corrected* (Corrigido), caso o valor tenha sido retificado por ação de alguma das regras definidas, ou *Invalid* (inválido) caso o valor seja ilegal no contexto e não exista nenhuma regra para a correção do mesmo. Estes serviços podem ser utilizados dentro da ferramenta cliente dos SSDQS, onde se cria um projeto em que se comparam os valores de uma fonte de dados que contenha campos que se pretendam validar com estes domínios, ou como uma operação dentro de uma tarefa de fluxo de dados, dos *Sql Server Integration Services* (SSIS).

Em qualquer das modalidades que os serviços sejam invocados, por cada análise efetuada só é possível comparar um campo de uma tabela por cada domínio. No entanto se se quiser avaliar, por exemplo, a tabela *INDIVIDUO*, da BDP, verifica-se uma situação em que temos três campos relativos ao estado civil, um para o estado civil do individuo ao óbito e outros dois para o registo do estado civil de cada um dos pais, no momento do seu nascimento. Para resolver esta questão, os SSDQS possibilitam a criação de domínios ligados (*Linked Domains*) que assumem todos os valores e regras do domínio de origem, sendo todas as alterações posteriormente efetuadas, quer no domínio de origem quer no domínio ligado, replicadas entre si. Assim, para os casos em que houve necessidade, foram criados os domínios ligados necessários. No caso do estado civil, foram criados dois domínios ligados.

4.2.2 Tratamento dos Nomes dos Indivíduos

Considerando que no passado não existiam regras claras de transmissão dos apelidos, a MRP assenta no cruzamento nominativo a partir do nome próprio dos indivíduos, aquele que se mantém mais constante ao longo da trajetória de vida. No entanto, uma breve consulta a qualquer BDP permite-nos detetar variações na grafia nome, que tanto podem resultar de dificuldades na transcrição dos registos como de inevitáveis erros de digitação. Torna-se, por isso, essencial tratar estes primeiros nomes para se conseguir maiores índices de qualidade e conseqüentemente melhores resultados na deteção de duplicados.

Utilizando os SSDQS criou-se um domínio denominado “NomeProprio” preenchido com uma lista de 2.662 nomes admissíveis em Portugal, recolhida do Instituto dos Registos e Notariado¹. Esta lista contém apenas os nomes admissíveis atualmente, pelo que teria de ser enriquecida com nomes que, entretanto, caíram em desuso, mas que, para este contexto, estão corretos.

De seguida, recorrendo à ferramenta de “Descoberta de Conhecimento” dos SSDQS que consulta um campo de uma tabela de dados e o compara com um domínio, analisou-se uma tabela com os nomes próprios de uma BDP. Com base na frequência dos valores e na medida da sua similaridade que os serviços calculam, esta ferramenta propôs a adição ao domínio de um conjunto de valores que considerou corretos e, para os que assumiu como erros de digitação, a sua inserção no conjunto de regras de correção de valores.

No primeiro ensaio efetuado, os nomes “Anatónio” e “Ántónio” foram considerados erros de digitação do nome “António” que já constava no domínio, conforme se pode verificar na Tabela 18. Já relativamente ao nome “Antónia”, que não existia no domínio, foi proposta a sua adição com base na frequência com que aparece na BD, sugerindo-se ainda a adição de “Àntonia” à lista de regras de correção. Cabe ao utilizador aceitar as propostas feitas pela ferramenta, tendo liberdade para efetuar os ajustes que considere necessários.

¹ www.irn.mj.pt

Tabela 18 - Resultado da validação do Domínio "NomeProprio"

NomeProprio_Source	NomeProprio_Output	NomeProprio_Status	NomeProprio_Confidence
Alexandrinha	Alexandrina	Corrected	0.8148148
Cartarina	Catarina	Corrected	0.8421053
Cristovão	Cristóvão	Auto suggest	0.7272727
Cristovão	Cristóvão	Auto suggest	0.7272727
Cristovão	Cristóvão	Auto suggest	0.7272727
Domingos	Domingos	Correct	
Fracisca	Francisca	Auto suggest	0.7619048
Manuel	Manuel	Correct	
Maria	Maria	Correct	

Com este processo, recorrendo apenas a uma BDP, adicionaram-se mais 241 valores corretos ao domínio "NomeProprio". A sua posterior aplicação sobre um pequeno conjunto de BDP possibilitou, o enriquecimento deste domínio que conta, atualmente, 3060 valores corretos.

Criou-se ainda um domínio de nome "Apelido" para lidar com o tratamento do restante nome do indivíduo. Este domínio contempla todos os nomes do domínio "NomeProprio", dado que estes podem surgir noutras posições do nome que não a primeira, tendo sido enriquecido, à semelhança do "NomeProprio" com a ferramenta de descoberta de conhecimento dos SSDQS, contemplando, neste momento 4679 valores corretos. Considerando a necessidade do tratamento de cada um dos componentes do nome, criaram-se 19 domínios ligados ao domínio "Apelido", o que permitirá tratar nomes com 21 componentes, número considerado suficiente pelos responsáveis do GHP, para o respetivo contexto.

4.3 A Integração dos Dados no Repositório Genealógico Nacional

Avaliados os dados e definidas as estratégias para a limpeza dos mesmos, reúnem-se condições para se iniciar a modelação do processo de integração de dados no RGN.

Existem em ambos os modelos tabelas muito similares com valores que servem para normalizar determinados campos e que, além de se manterem iguais em ambos os modelos, são estáticos, não sofrendo, normalmente, qualquer atualização.

Assim sendo, e dada a natureza simples desta operação, optou-se pela importação direta dos valores destas tabelas, recorrendo para isso à ferramenta de importação de dados do *SQL Server Management Studio* (SSMS). Outras tabelas foram preenchidas manualmente, dados o reduzido número de registos que contemplam. Desta forma, foram estão povoadas as seguintes tabelas da BDC:

- *CLASSIFICACAOFAMILIA*
- *CLASSIFICACAOPAROQUIALINVESTIGADOR*
- *CLASSIFICACAOPAROQUIALPADRE*
- *CONCELHO*
- *DISTRITO*
- *ESTADOCIVIL*
- *FILIACAO*
- *MOEDA*
- *PAIS*
- *PARENTESCO*
- *PAROQUIA*
- *SACRAMENTOS*
- *SEXO*
- *SITUACAOFAMILIA*
- *TIPOASSISTENTE*
- *TIPOFAMILIA*
- *TIPOFILIACAO*
- *TIPOINDIVIDUO*
- *TIPOMOBILIDADE*
- *TIPOREGISTOPAROQUIAL*
- *TIPOSUFRAGIO*
- *TUMBA*

Após a operação anterior avançou-se para o desenho do processo de integração de dados de acordo com a metodologia EMD. Na Figura 31 pode-se observar a modelação dos processos em alto nível, estando aqui representadas as três etapas da metodologia, Extração, Mapeamento e Carregamento e a área de estágio dos dados.

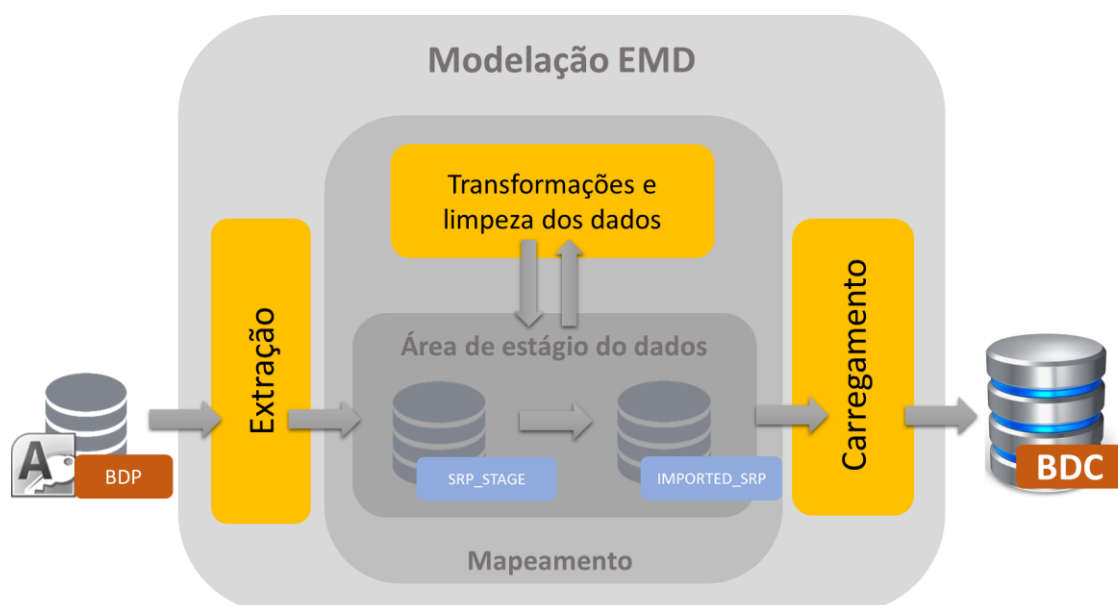


Figura 31 - Modelação EMD de alto nível

A Figura 32 apresenta o cenário EMD para a integração da entidade Indivíduo. Neste cenário, após a leitura dos dados da tabela INDIVIDUO da BDP, efetuam-se algumas transformações do tipo de dados de alguns atributos, como o caso do *nind* que é convertido do tipo *duplo* para *inteiro* (operador “Tint”), sendo o fluxo de dados remetido para a tabela INDIVIDUO na BD temporária SRP_STAGE. Na transformação seguinte, aplicam-se mecanismos de limpeza de dados (operador DC) aos atributos apropriados, como é o caso do nome, enviando-se de seguida os mesmos para a tabela INDIVIDUOS da, também temporária, BD IMPORTED_SRP. No último passo, realizam-se as transformações necessárias para converter os dados em formatos adequados à BDC, como é o caso da desagregação das datas em campos separados, “Ano, Mês, Dia” (operador “TdtP”), procedendo-se de seguida ao carregamento dos dados para a BDC.

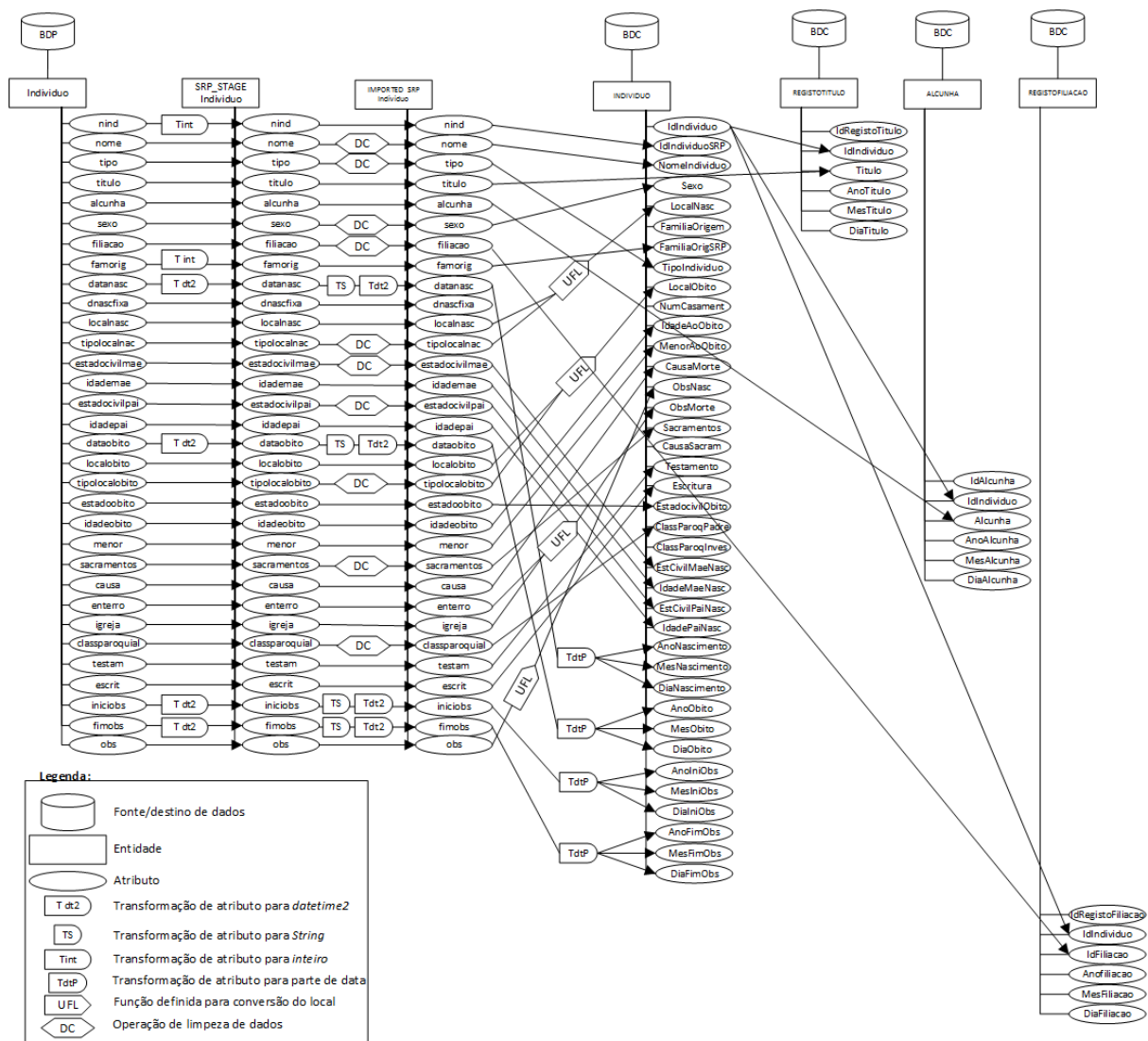


Figura 32 - Cenário EMD para a entidade Indivíduo

Depois da modelação conceptual dos processos, avançou-se para a implementação dos mesmos, que foi realizada com recurso aos SSIS¹ da Microsoft, integrados também na estrutura que suporta o RGN. A integração conseguiu-se com recurso à modelação de três *packages* diferentes num projeto de integração de dados dos SSIS, conforme o fluxo apresentado na Figura 33. Tal como se pode observar na mesma figura, foi criada a área de estágio dos dados no SQLSERVER, com as duas BD (SRP_STAGE e IMPORTED_SRP) com modelo de dados similar ao das BDP para onde se realizam as operações de extração, limpeza e transformação dos dados.

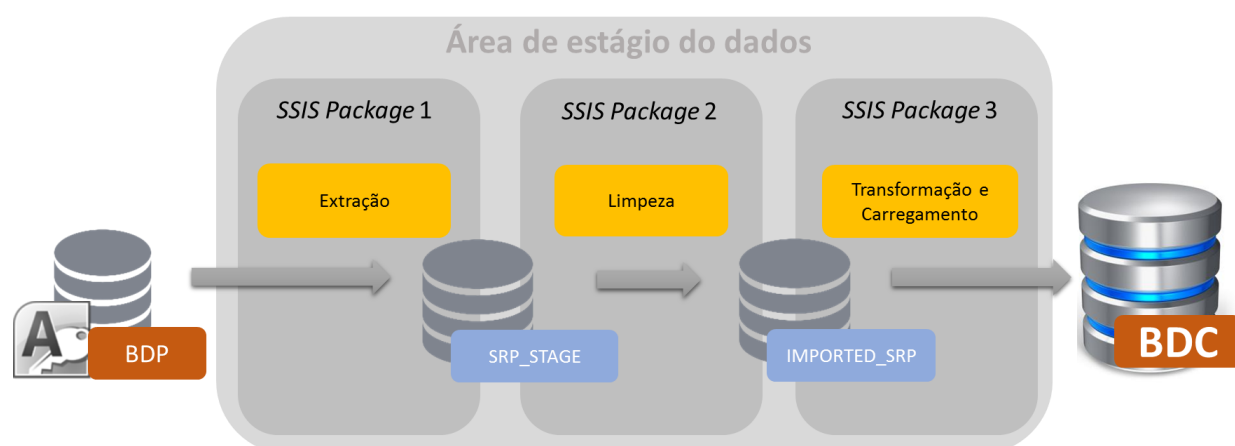


Figura 33 - O fluxo da integração de dados nos SSIS

Apresentam-se de seguida as operações de cada um dos momentos do fluxo.

4.3.1 A Extração dos Dados

Conforme já referido, a primeira tarefa de um processo de integração de dados passa pela extração dos mesmos a partir da fonte de dados onde residem. No presente contexto os dados encontram-se em BDP distintas (normalmente, uma por cada paróquia já recolhida) mas que apresentam a mesma estrutura. Criou-se, por isso, um primeiro *Package* responsável por fazer a extração de dados das tabelas da BDP, para a BD SRP_STAGE. Esta operação realiza-se com um mapeamento dos campos quase direto, realizando-se aqui apenas a conversão de alguns tipos de dados em alguns campos de algumas tabelas, nomeadamente, os campos referentes ao número de indivíduo (*nind*) e número de família (*nfamilia*) que, apesar de na aplicação SRP serem exibidos como valores inteiro, na BDP encontram-se definidos como

¹ [https://msdn.microsoft.com/pt-pt/library/ms141026\(v=sql.110\).aspx](https://msdn.microsoft.com/pt-pt/library/ms141026(v=sql.110).aspx)

duplo (*double*) e os campos de data que estão declarados com o tipo Data/Hora (*Datetime*) no MS ACCESS. Este tipo de dados no MS ACCESS tem uma amplitude de 01/01/100 até 31/12/9999¹. No SQL SERVER o tipo *Datetime* compreende apenas valores no intervalo 01/01/1753 até 31/12/9999² o que impossibilita a utilização do mesmo, dado que as BDP apresentam datas que se estendem até ao século XV. No SQL SERVER, torna-se, portanto, necessária a utilização do tipo *Datetime2*, para armazenar estas datas, uma vez que este tipo compreende valores entre 01/01/0001 até 31/12/9999³.

Este processo inicia-se com uma tarefa de execução de um comando em SQL que elimina, caso existam, os dados das tabelas da BD SRP_STAGE, assegurando que esta esteja vazia no momento da extração. Seguidamente é realizado um conjunto de operações de fluxo de dados (*Data Flow Task*), contendo cada uma delas um conjunto de tarefas de leitura de dados da tabela de origem, transformação (se necessária), e carregamento para a BD SRP_STAGE, conforme se pode verificar na Figura 34.

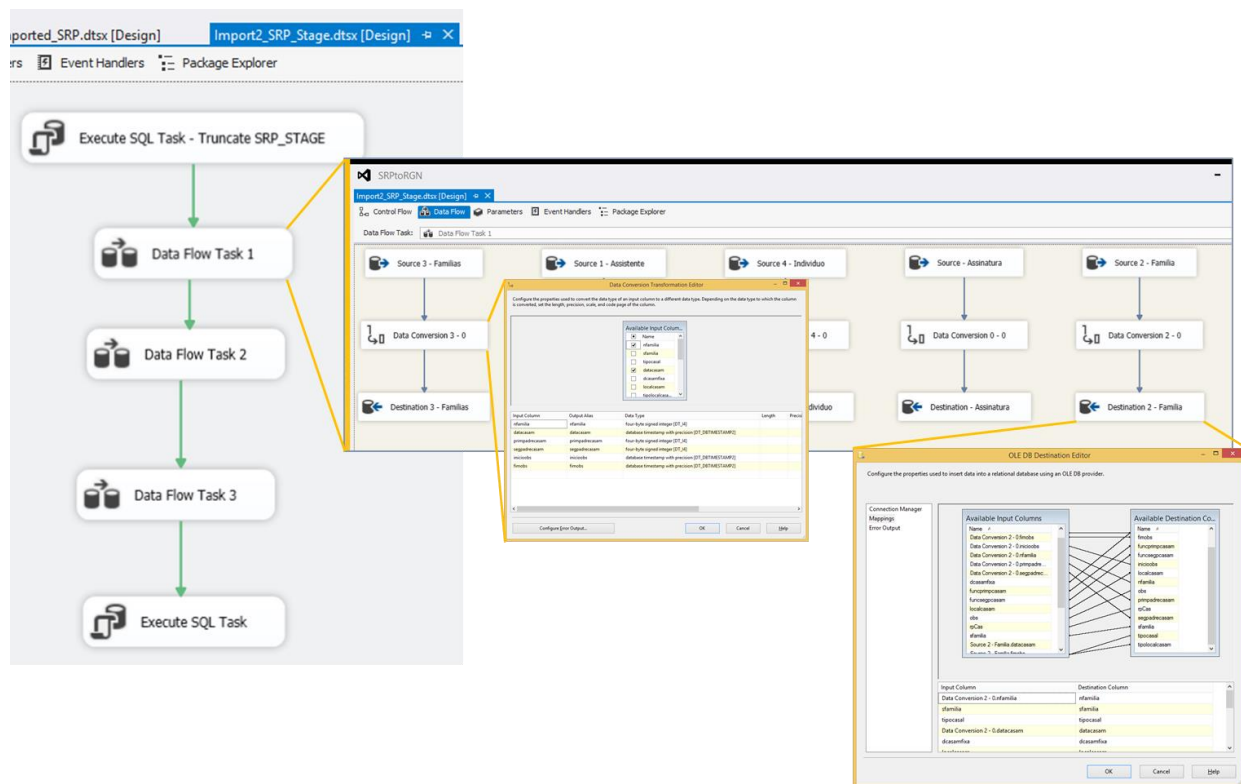


Figura 34 - Fluxo de dados da BDP para a BD SRP_STAGE

¹ <https://support.office.com/en-us/article/Format-the-date-and-time-field-in-Access-47fbbdc1-52fa-416a-b8d5-ba24d881b698>

² <https://msdn.microsoft.com/en-us/library/ms187819.aspx>

³ <https://msdn.microsoft.com/en-us/library/bb677335.aspx>

4.3.2 A Limpeza de Dados

Num segundo momento do processo de integração de dados realizam-se as operações de limpeza dos mesmos. Estas operações, modeladas no segundo *Package* do projeto criado para esta finalidade, avaliam a conformidade dos valores para os campos das tabelas e, se necessário, corrigem-nos para valores admissíveis para esse mesmo contexto. Foram então modeladas tarefas de fluxo de dados para cada uma das tabelas a ser integrada, conforme se pode verificar na Figura 35.

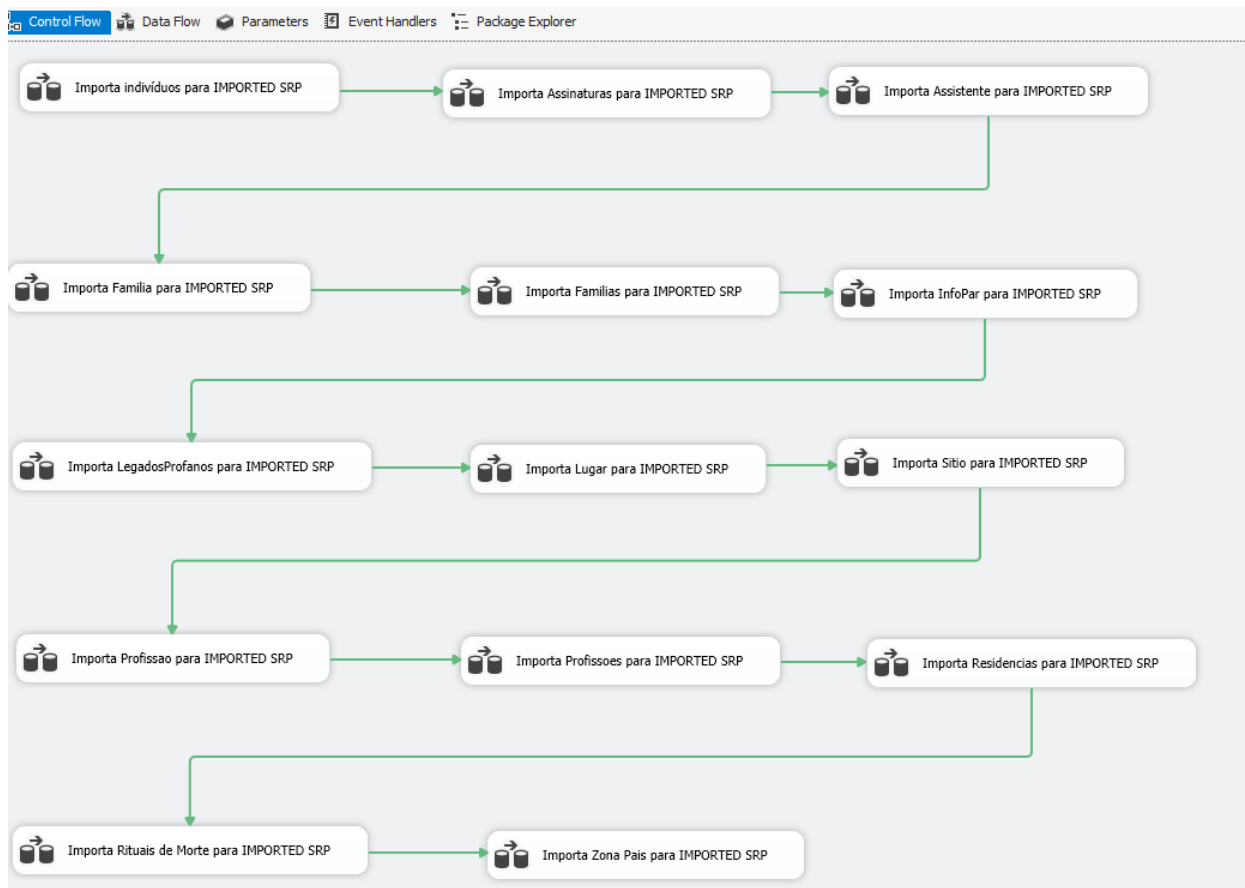


Figura 35 - Fluxo de dados da BD SRP_STAGE para IMPORTED_SRP

Cada uma destas tarefas extrai os dados de uma tabela da BD SRP_STAGE, efetua uma operação de limpeza de dados, nos campos em que se aplique, enviando-os de seguida para a tabela similar na BD IMPORTED_SRP. Na Figura 36 apresenta-se, em detalhe e a título de exemplo, este processo para a tabela ASSINATURAS. Temos um primeiro elemento (*ADO NET Source*) que é um conector à tabela na BD origem com dados a serem tratados. O segundo elemento (*Data Conversion*) realiza a conversão do campo *dataref*, que contém um campo do tipo *data*, que na tabela de origem está definido com o tipo *Datetime2*, para o tipo *string*, uma vez que os SSIS, não suportam o tratamento de dados do tipo

Datetime2, colocando estes dados com valores *NULL*, caso o fluxo de dados passe por qualquer um dos componentes de tratamento de dados da mesma.

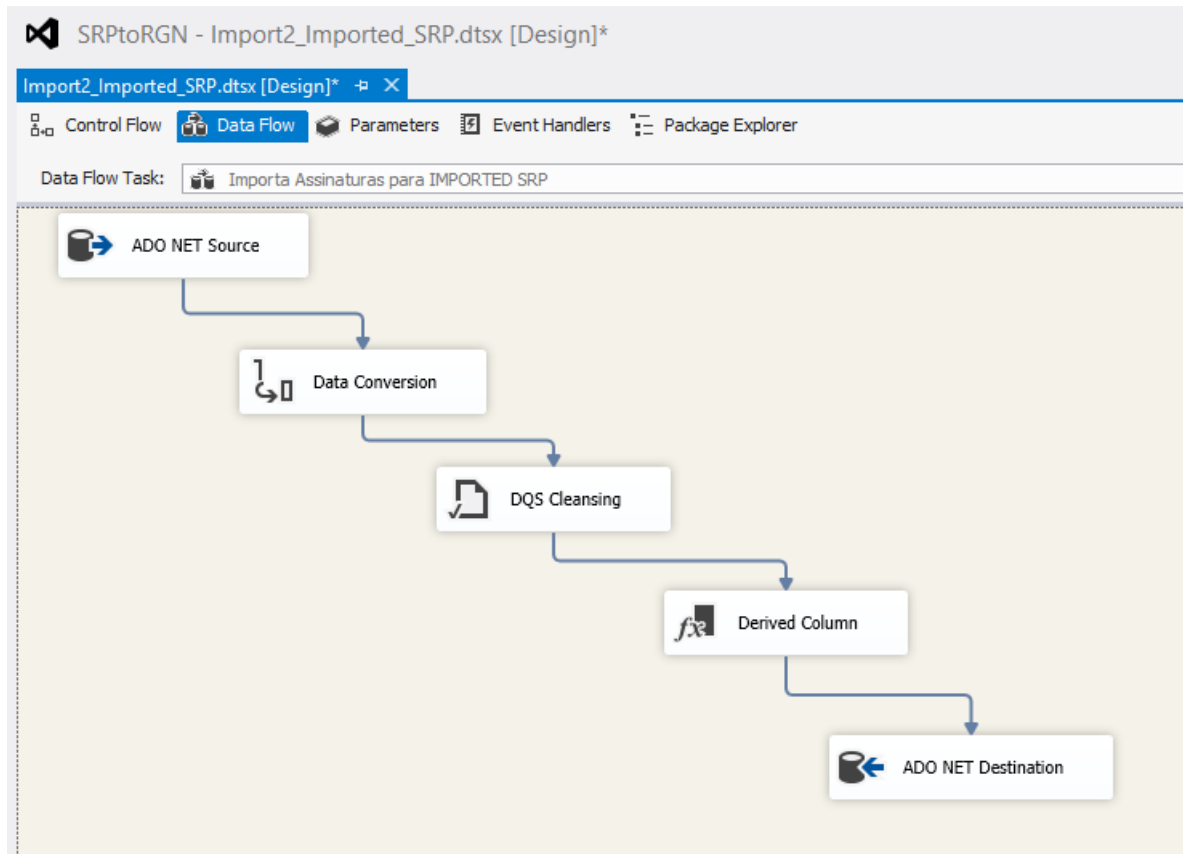


Figura 36 – Fluxo de dados para a limpeza da tabela Assinaturas

O terceiro elemento (*DQS Cleansing*) recorre aos SSQDS e com recurso aos domínios criados, conforme o explicado no ponto 4.2.1, efetuam a limpeza dos campos. Na Figura 37 podemos verificar a configuração deste componente, apresentando-se, na parte esquerda da imagem, a definição da conexão aos SSDQS e, na parte direita, o mapeamento do campo a ser tratado. Neste caso, pretende-se avaliar o campo *nivelassinatura*, que, seguindo as regras definidas no domínio *NivelAssinatura*, deverá contemplar um valor inteiro, compreendido no intervalo {1 a 5}. Caso o valor não esteja em conformidade, e de acordo com regra também definido no domínio, este campo deverá ser corrigido para o valor *NULL*.

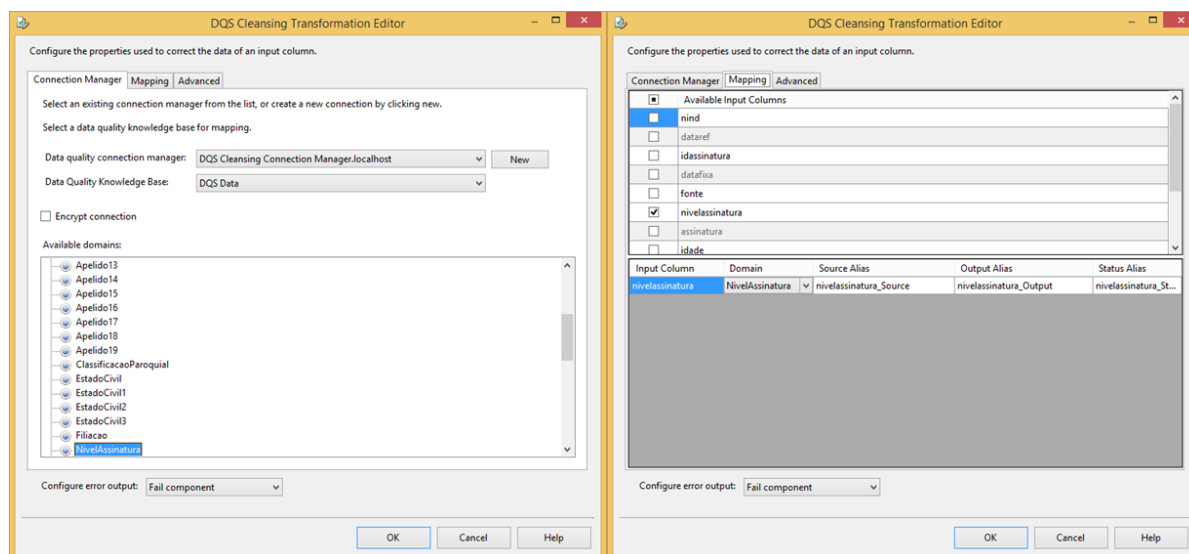


Figura 37 - Tarefa de limpeza de dados da tabela Assinatura

Prossegue-se o fluxo para o quarto elemento (*Derived Column*) onde se encontra uma tarefa de conversão de dados que volta a reverter o campo *dataref* para o tipo *Datetime2*, sendo este seguido do último elemento (*ADO NET Destination*), onde se efetua a ligação à tabela na BD de destino e o consequente envio dos dados.

Na Figura 38 apresenta-se o fluxo relativo à tabela INDIVIDUO. Este fluxo difere, em complexidade dos demais, uma vez que aqui se efetua o tratamento do nome do indivíduo, operação de grande importância para este contexto, conforme o verificado no ponto 4.2.2 e que se detalha a seguir.

No terceiro componente deste fluxo (*Multicast*) gera-se um fluxo duplicado dos dados, seguindo o que se dirige para a esquerda, para um conjunto de tarefas para o tratamento do nome. Assim, no elemento *Derived Column*, efetua-se uma remoção dos possíveis espaços existentes no início e no fim deste campo ou de espaços múltiplos entre as diversas partes do nome.

De seguida recorre-se a um *Script Component* que, fazendo uso da linguagem de programação C#, cria uma lista de componentes do nome, recorrendo aos espaços que o separam para os dividir, atribuindo cada uma destas partes a um novo campo, adicionado a este fluxo, conforme a Figura 39.

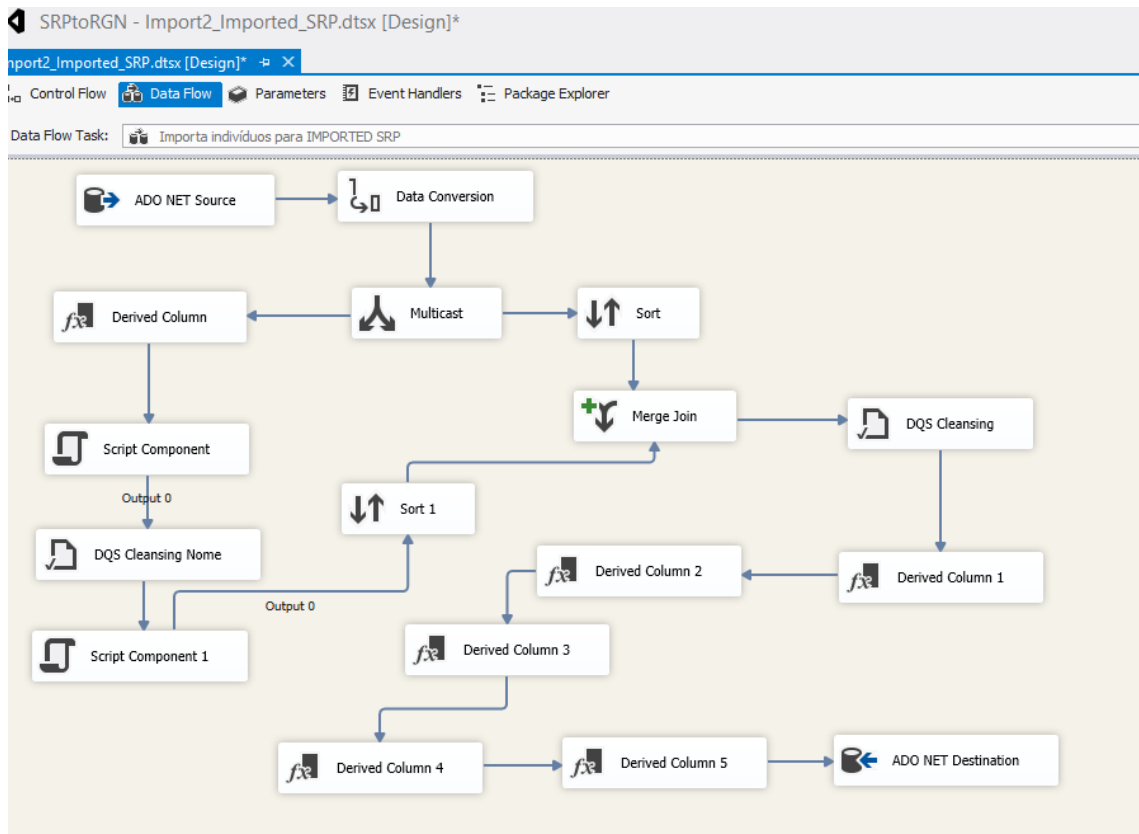


Figura 38 - Fluxo de dados para a limpeza da tabela INDIVIDUO

```

/// <param name="Row">The row that is currently passing through the component</param>
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    var nomeclean = Row.nome.Replace("\r", string.Empty).Replace("\n", string.Empty);

    var nomes = nomeclean.Split(' ');
    System.Collections.Generic.List<string> list = new List<string>(nomes);

    var count = nomes.Length;

    if (count > 0){
        Output0Buffer.AddRow();
        Output0Buffer.nome = Row.nome;
        Output0Buffer.nind = Row.nind;
        if (list.ElementAtOrDefault(0) == null) { Output0Buffer.N1 = null; return; } else { Output0Buffer.N1 = list.ElementAtOrDefault(0); }
        if (list.ElementAtOrDefault(1) == null) { Output0Buffer.N2 = null; return; } else { Output0Buffer.N2 = list.ElementAtOrDefault(1); }
        if (list.ElementAtOrDefault(2) == null) { Output0Buffer.N3 = null; return; } else { Output0Buffer.N3 = list.ElementAtOrDefault(2); }
        if (list.ElementAtOrDefault(3) == null) { Output0Buffer.N4 = null; return; } else { Output0Buffer.N4 = list.ElementAtOrDefault(3); }
        if (list.ElementAtOrDefault(4) == null) { Output0Buffer.N5 = null; return; } else { Output0Buffer.N5 = list.ElementAtOrDefault(4); }
        if (list.ElementAtOrDefault(5) == null) { Output0Buffer.N6 = null; return; } else { Output0Buffer.N6 = list.ElementAtOrDefault(5); }
        if (list.ElementAtOrDefault(6) == null) { Output0Buffer.N7 = null; return; } else { Output0Buffer.N7 = list.ElementAtOrDefault(6); }
        if (list.ElementAtOrDefault(7) == null) { Output0Buffer.N8 = null; return; } else { Output0Buffer.N8 = list.ElementAtOrDefault(7); }
        if (list.ElementAtOrDefault(8) == null) { Output0Buffer.N9 = null; return; } else { Output0Buffer.N9 = list.ElementAtOrDefault(8); }
        if (list.ElementAtOrDefault(9) == null) { Output0Buffer.N10 = null; return; } else { Output0Buffer.N10 = list.ElementAtOrDefault(9); }
        if (list.ElementAtOrDefault(10) == null) { Output0Buffer.N11 = null; return; } else { Output0Buffer.N11 = list.ElementAtOrDefault(10); }
        if (list.ElementAtOrDefault(11) == null) { Output0Buffer.N12 = null; return; } else { Output0Buffer.N12 = list.ElementAtOrDefault(11); }
        if (list.ElementAtOrDefault(12) == null) { Output0Buffer.N13 = null; return; } else { Output0Buffer.N13 = list.ElementAtOrDefault(12); }
        if (list.ElementAtOrDefault(13) == null) { Output0Buffer.N14 = null; return; } else { Output0Buffer.N14 = list.ElementAtOrDefault(13); }
        if (list.ElementAtOrDefault(14) == null) { Output0Buffer.N15 = null; return; } else { Output0Buffer.N15 = list.ElementAtOrDefault(14); }
        if (list.ElementAtOrDefault(15) == null) { Output0Buffer.N16 = null; return; } else { Output0Buffer.N16 = list.ElementAtOrDefault(15); }
        if (list.ElementAtOrDefault(16) == null) { Output0Buffer.N17 = null; return; } else { Output0Buffer.N17 = list.ElementAtOrDefault(16); }
        if (list.ElementAtOrDefault(17) == null) { Output0Buffer.N18 = null; return; } else { Output0Buffer.N18 = list.ElementAtOrDefault(17); }
        if (list.ElementAtOrDefault(18) == null) { Output0Buffer.N19 = null; return; } else { Output0Buffer.N19 = list.ElementAtOrDefault(18); }
        if (list.ElementAtOrDefault(19) == null) { Output0Buffer.N20 = null; return; } else { Output0Buffer.N20 = list.ElementAtOrDefault(19); }
    }
}
    
```

Figura 39 - Script para a separação dos elementos do nome

No passo seguinte (*DQS Cleansing Nome*), cada uma das partes do nome é avaliada pelos SSDQS, tratando o primeiro nome com o domínio *NomeProprio* e cada uma das outras partes com os domínios *Apelido1*, *Apelido2*, *Apelido3*, *Apelido4*, (...) conforme a Figura 40.

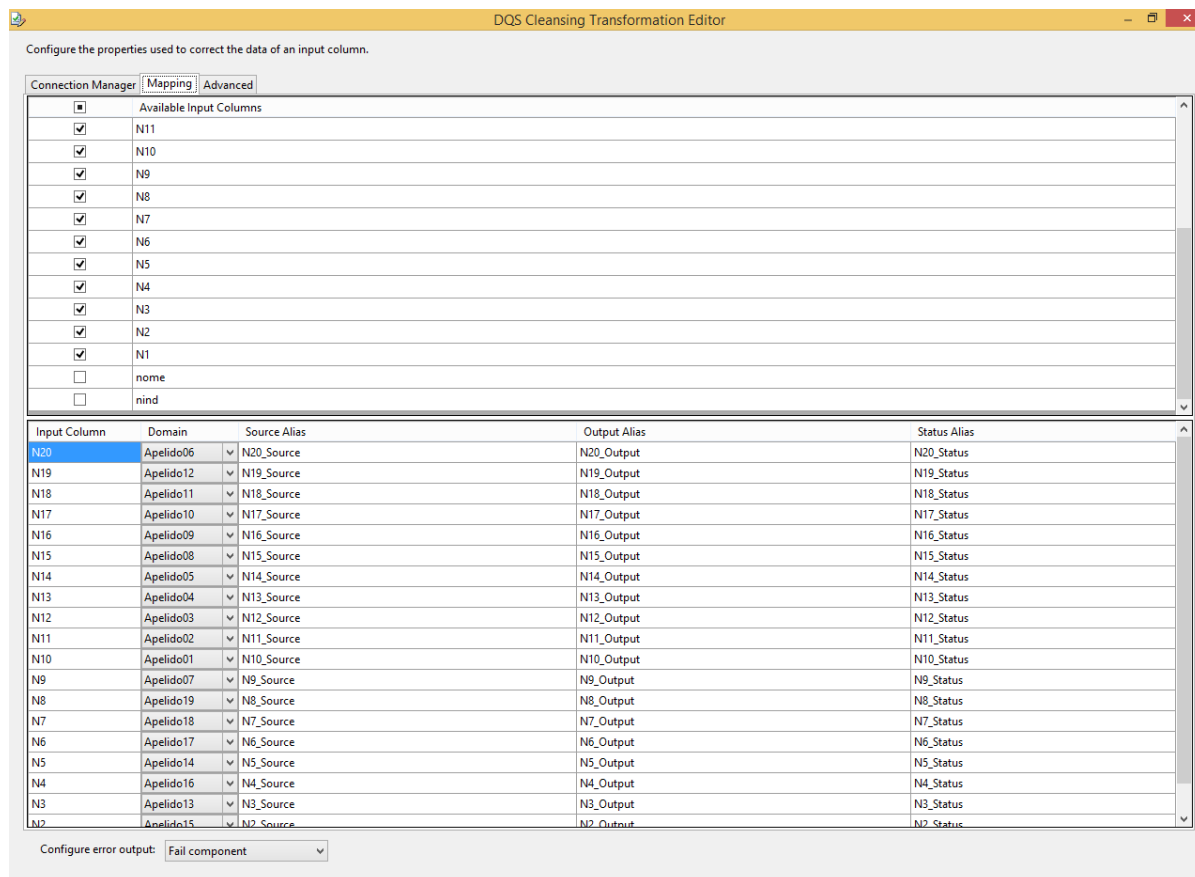


Figura 40 – Tratamento das partes do nome

No *Script Component 1* (Figura 41) volta-se a agregar cada uma das partes no nome, agora tratado, para uma nova coluna *nomeCorrigidoCompleto*. Além da agregação das partes do nome, este *script* efetua ainda a conversão das partículas do nome (de, da, e, ...) para minúsculas, uma vez que, segundo as regras dos domínios *NomeProprio* e *Apelido*, todos os nomes analisados são devolvidos numa formatação com inicial maiúscula. Avalia-se aqui também a existência do caractere '?' no nome, dado que é prática comum, aquando da recolha dos dados dos RP, caso parte do nome se apresente ilegível, ou suscite dúvidas quanto ao seu valor, a colocação deste sinal para assinalar esta situação. A limpeza ao nome realizada pelos SSDQS tendencialmente removerá este carácter, pelo que, neste *script*, verifica-se também se ocorreu tal situação e, em caso, afirmativo, assinala-se o *nomeCorrigidoCompleto* da mesma forma.

```

    /// <param name="Row">The row that is currently passing through the component</param>
    public override void Input0_ProcessInputRow(Input0Buffer Row)
    {
        //Concatena cada uma das partes do nome
        var nomeCompleto = Row.N10Output;

        if (!String.IsNullOrEmpty(Row.N20Output)) { nomeCompleto += " " + Row.N20Output; }
        if (!String.IsNullOrEmpty(Row.N30Output)) { nomeCompleto += " " + Row.N30Output; }
        if (!String.IsNullOrEmpty(Row.N40Output)) { nomeCompleto += " " + Row.N40Output; }
        if (!String.IsNullOrEmpty(Row.N50Output)) { nomeCompleto += " " + Row.N50Output; }
        if (!String.IsNullOrEmpty(Row.N60Output)) { nomeCompleto += " " + Row.N60Output; }
        if (!String.IsNullOrEmpty(Row.N70Output)) { nomeCompleto += " " + Row.N70Output; }
        if (!String.IsNullOrEmpty(Row.N80Output)) { nomeCompleto += " " + Row.N80Output; }
        if (!String.IsNullOrEmpty(Row.N90Output)) { nomeCompleto += " " + Row.N90Output; }
        if (!String.IsNullOrEmpty(Row.N100Output)) { nomeCompleto += " " + Row.N100Output; }
        if (!String.IsNullOrEmpty(Row.N110Output)) { nomeCompleto += " " + Row.N110Output; }
        if (!String.IsNullOrEmpty(Row.N120Output)) { nomeCompleto += " " + Row.N120Output; }
        if (!String.IsNullOrEmpty(Row.N130Output)) { nomeCompleto += " " + Row.N130Output; }
        if (!String.IsNullOrEmpty(Row.N140Output)) { nomeCompleto += " " + Row.N140Output; }
        if (!String.IsNullOrEmpty(Row.N150Output)) { nomeCompleto += " " + Row.N150Output; }
        if (!String.IsNullOrEmpty(Row.N160Output)) { nomeCompleto += " " + Row.N160Output; }
        if (!String.IsNullOrEmpty(Row.N170Output)) { nomeCompleto += " " + Row.N170Output; }
        if (!String.IsNullOrEmpty(Row.N180Output)) { nomeCompleto += " " + Row.N180Output; }
        if (!String.IsNullOrEmpty(Row.N190Output)) { nomeCompleto += " " + Row.N190Output; }
        if (!String.IsNullOrEmpty(Row.N200Output)) { nomeCompleto += " " + Row.N200Output; }

        //Passa as partículas dos nomes para iniciais minúsculas
        nomeCompleto = nomeCompleto.Replace(" De ", " de ");
        nomeCompleto = nomeCompleto.Replace(" Da ", " da ");
        nomeCompleto = nomeCompleto.Replace(" Do ", " do ");
        nomeCompleto = nomeCompleto.Replace(" Das ", " das ");
        nomeCompleto = nomeCompleto.Replace(" Dos ", " dos ");
        nomeCompleto = nomeCompleto.Replace(" E ", " e ");

        if ((Row.nome.Contains("?") && (!nomeCompleto.Contains("?"))) {
            nomeCompleto += " ?";
        }

        //Adiciona os valores ao output
        Output0Buffer.AddRow();
        Output0Buffer.NomeCorrigidoCompleto = nomeCompleto;
        Output0Buffer.nome = Row.nome;
        Output0Buffer.nind = Row.nind;
    }
}

```

Figura 41 - Script para a agregação dos elementos do nome

Findo o tratamento do nome, os dois fluxos de dados são, depois de ordenados, novamente associados num fluxo único (*Merge Join*) sendo de seguida tratados os demais campos do fluxo, passíveis de validação, recorrendo novamente aos SSDQS (*DQS Cleansing*). Preserva-se o nome original do indivíduo no campo de observações do nascimento do mesmo, para futuras validações.

Estando completa a fase de limpeza dos dados, passa-se para a fase de Mapeamento (transformação) e carregamento, detalhadas no ponto seguinte.

4.3.3 Mapeamento (Transformação) e Carregamento para o Repositório Genealógico Nacional

Após a limpeza dos dados, reúnem-se as condições para se realizarem as transformações dos mesmos, necessárias para corresponderem ao modelo de dados de destino a BDC do RGN.

Para este efeito foi criado o terceiro *Package* no projeto de integração de dados. Este *package* está dividido em dois *Containers*, conforme se pode verificar na Figura 42 e que se detalham nos pontos seguintes.

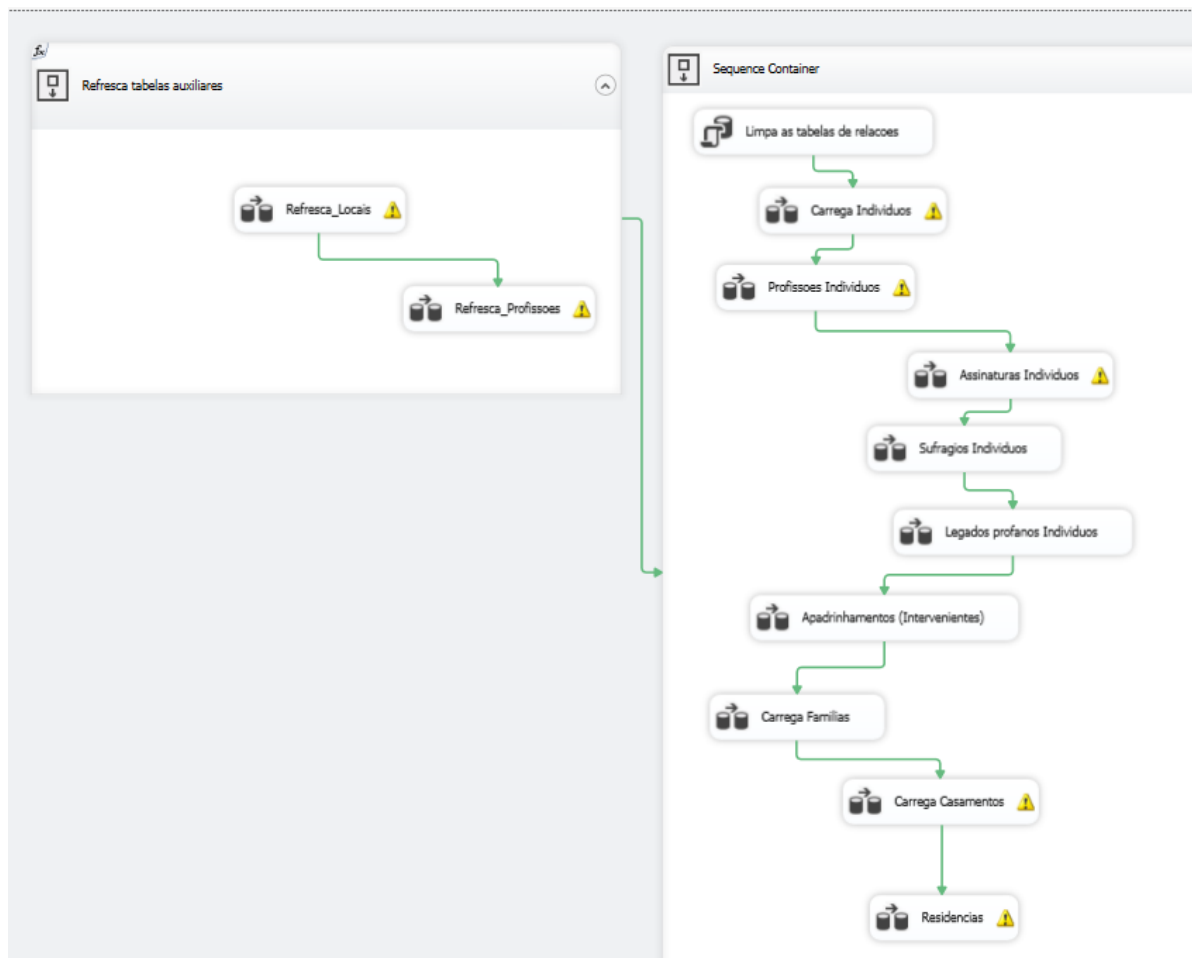


Figura 42 - Package 3

4.3.3.1 Refreshamento das Tabelas Auxiliares

Num primeiro momento realiza-se o refreshamento das tabelas *LOCAL* e *PROFISSAO*. Uma vez que são tabelas que podem ser alimentadas pelos investigadores no momento da recolha de dados. Atendendo a que as bases de dados se encontram isoladas, em bases de dados distintas ocorrerá naturalmente o registo dos mesmos lugares e/ou profissões que terão em cada uma delas, uma chave primária diferente, tornando-se, por isso, necessário convergir os valores destas tabelas para listas únicas.

Na BDP, como referido no ponto 3.1.1 a estrutura das referências difere em granularidade relativamente ao modelo do BDC, possibilitando o registo de vários sítios para cada local existente. Na BDC o nível mais baixo para as referências geográficas é o local, tornando-se assim necessário agregar todos os sítios provenientes nas BDP em locais. Optou-se por uma estratégia em que, existindo na BDP o sítio “Caminho de Baixo” referente ao lugar “Silveira” procede-se ao registo de um novo local na BDC com o nome “Silveira - Caminho de Baixo”. Assim sendo, procedeu-se à implementação de um fluxo apresentado na Figura 43 que permite a concretização desta operação e posterior refrescamento de dados na LOCAL.

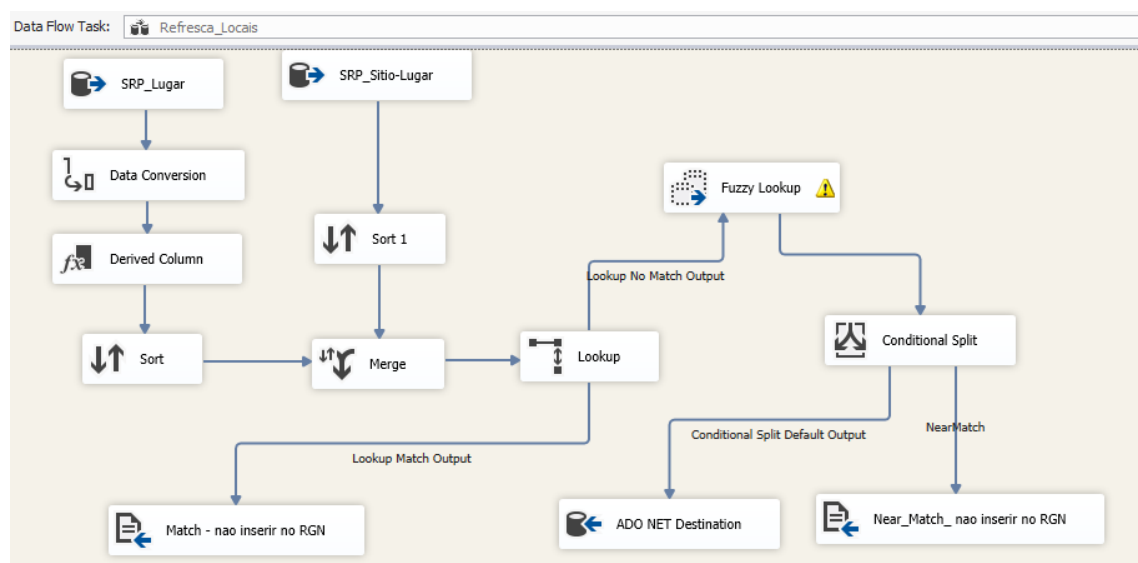


Figura 43 - Fluxo para o refrescamento da tabela LOCAL

Os dados são então extraídos das tabelas *LUGAR* e *SITIO*, sendo que, na tabela *SITIO* em vez da leitura integral da tabela, implementou-se uma *QuerySQL* (Figura 44) que devolve o nome do sítio já associado ao nome do lugar a que se refere, como se pretende que seja armazenado. Seguidamente avalia-se a existência destes lugares na tabela *LOCAL* do RGN, primeiro procurando uma correspondência exata (*Lookup*) e depois uma correspondência aproximada (*Fuzzy Lookup*) na lista de locais já registados para a paróquia a que se referem. Os valores com correspondência exata e muito aproximada são então descartados, uma vez que já existem na tabela de destino, sendo apenas inseridos os valores considerados novos.

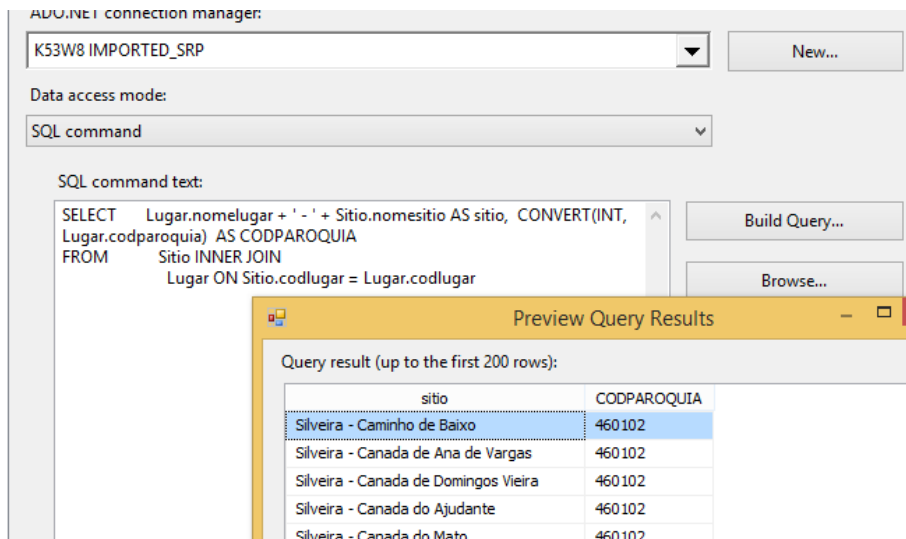


Figura 44 - Extração dos sítios com associação do lugar

Para o refrescamento da tabela profissão procede-se de um modo semelhante, tal como apresentado na Figura 45, em que se avalia a existência de potenciais novos valores a ser registados. No entanto, neste fluxo, faz-se também uso de um dos domínios dos SSDQS, utilizados aqui, à semelhança dos domínios “NomeProprio” e “Apelido” para efetuar uma correção na designação da profissão, dado que, poderão existir aqui também erros e/ou variações na grafia.

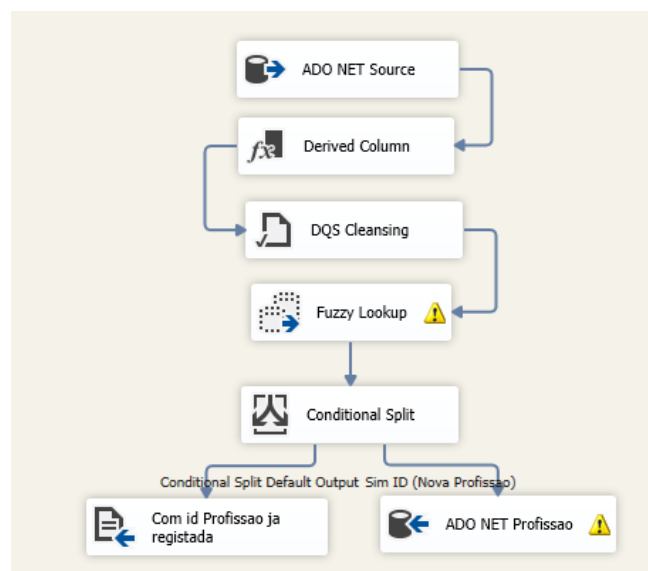


Figura 45 - O fluxo de refrescamento da tabela PROFISSAO

Se se identificarem profissões novas, são adicionadas à tabela PROFISSAO, sendo aquelas que foram identificadas como já presentes, descartadas.

4.3.3.2 Carregamentos das Entidades Principais e dos Registos Associados

Após o refrescamento das tabelas realizado no ponto anterior, reúnem-se condições para realizar o carregamento das tabelas relativas às entidades principais no sistema: *INDIVIDUO*, *FAMILIA* e das restantes tabelas com informações relativas a estas entidades, tais como, profissões e residências.

A primeira tabela a ser carregada tem de ser a *INDIVIDUO*, uma vez que todas as outras tabelas possuem referência a esta entidade. Apresenta-se este fluxo na Figura 46.

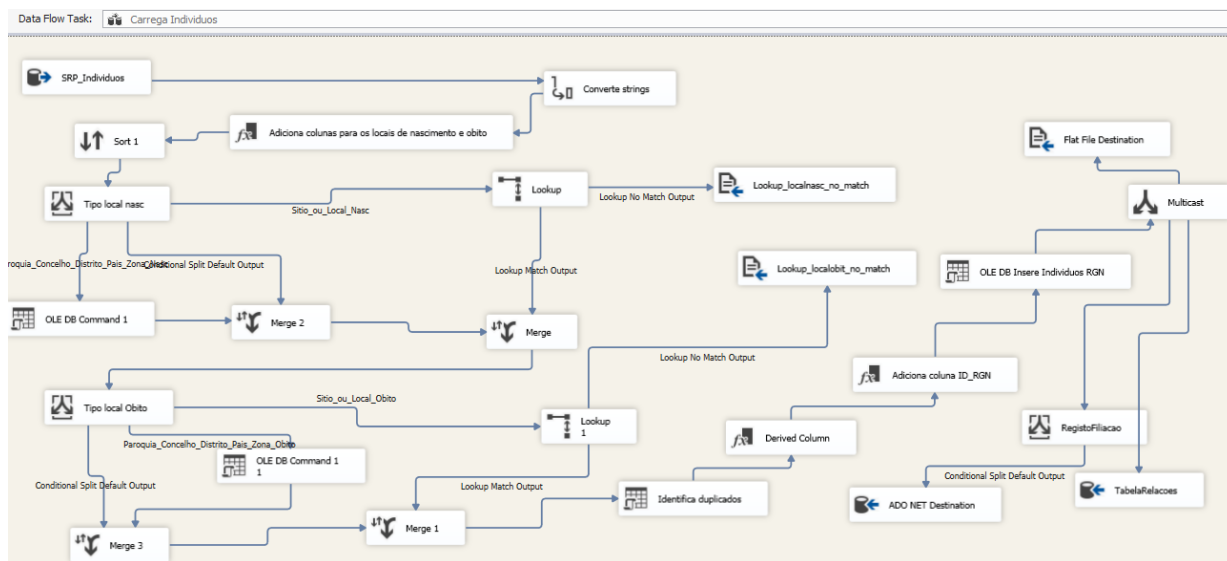


Figura 46 - Fluxo para o carregamento da tabela INDIVIDUO

A leitura da tabela de origem é realizada com recurso a um comando *SQL* (Figura 47) que possibilita a junção de informação relativa ao mesmo, que será necessária no decorrer do fluxo, nomeadamente, para os casos em que os locais de nascimento e /ou óbito são do tipo ‘lugar’, ‘sítio’, ou ‘zonaPais’ acrescenta-se a descrição destes lugares para posterior identificação. No caso das datas, dado o problema referido no ponto 4.3.1 e o facto de que o armazenamento deste tipo de dados na BDC é realizado em campos separados, efetua-se também neste momento a separação das mesmas nas componentes ‘Ano’, ‘Mês’ e ‘Dia’, possibilitando-se assim, por um lado, a preservação desta informação e, por outro lado, a extração da mesma num formato adequado para as tabelas de destino.

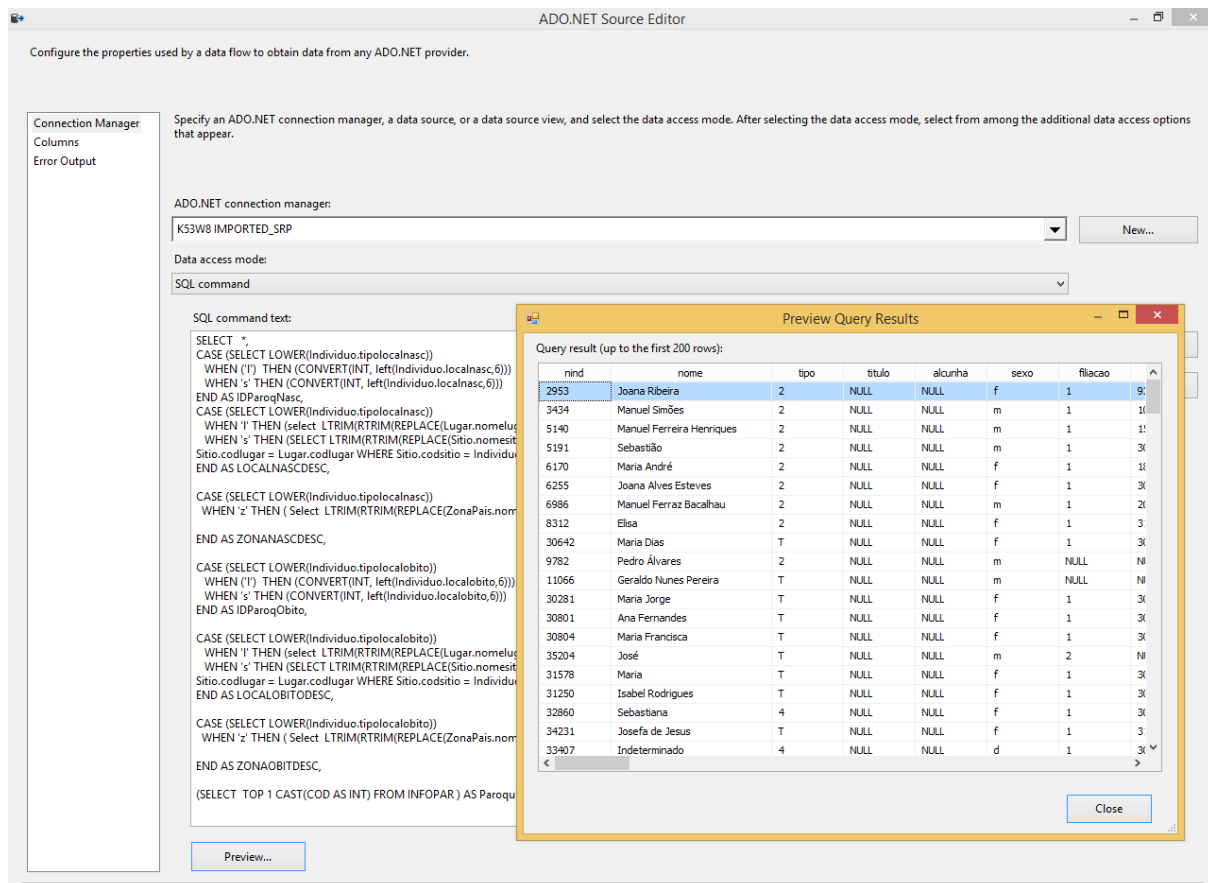


Figura 47 - Comando SQL para leitura da tabela INDIVIDUO

Seguidamente, trata-se da conversão da referenciação dos locais de nascimento e óbito. Na BDP, os locais são referenciados, como anteriormente indicado, por um par de campos, um com o código dos locais e outro com a indicação do tipo de local, por exemplo, o local de nascimento é '30601001' e o tipo de local de nascimento é 'L', o que indica que o código do local de nascimento se refere a um registo que se encontra na tabela LUGAR, permitindo registar, conforme a informação disponível que o indivíduo nasceu num lugar, paróquia, concelho... No RGN as referências geográficas foram implementadas com nível de granularidade ao local, sendo necessário armazenar estas referências sempre a este nível. Para permitir o registo de informação para outros níveis de granularidade foram gerados na BDC um distrito 'desconhecido' para cada um dos países registados. Para cada distrito foi gerado um concelho 'desconhecido', para cada concelho foi registada uma paróquia 'desconhecida' e cada uma das paróquias teve atribuído um local também 'desconhecido'. Deste modo, se os RP indicam o local de nascimento, faz-se o registo neste local, se se refere apenas a paróquia de nascimento, regista-se o ato no local 'desconhecido' da referida paróquia, ou se se conhece apenas o distrito, regista-se no local 'desconhecido', da paróquia 'desconhecida', do concelho 'desconhecido' do distrito indicado.

Seguidamente divide-se o fluxo consoante o tipo de local de nascimento. Se este campo for tipo 'sítio' ou 'local' procura-se o valor adequado com a ferramenta *lookup* que devolverá a chave primária do local correspondente. Para os casos 'Paroquia', 'Concelho', 'Distrito', 'Pais' ou 'ZonaPais' foi implementado, na BD do RGN, um *stored procedure*, que pode ser parcialmente visualizado na Figura 48, que, recebendo um código e a indicação do tipo de local, devolve o *id* do local "desconhecido", no nível de granularidade adequado, referente a essa combinação.

```

14  -- =====
15  -- Author:      <Author,,Name>
16  -- Create date: <Create Date,,>
17  -- Description: <Description,,>
18  -- =====
19
20
21  CREATE PROCEDURE [dbo].[sp_Get_LocalDesconhecido]
22  (
23      @IdLocalSRP AS nvarchar(255),
24      @TipoLocalSRP AS nvarchar(255),
25      @IDLocalRGN AS INT OUTPUT
26  )
27  AS
28  BEGIN
29
30
31  IF (LOWER(@TipoLocalSRP) = 'f')
32  BEGIN
33      SET @IDLocalRGN = (
34          select IdLocal
35          --, Local, Paroquia, Concelho, Distrito, Pais
36          from [dbo].Local
37          Join [dbo].Paroquia on [dbo].Local.IdParoquia = [dbo].Paroquia.IdParoquia
38          -- Join [dbo].Concelho on [dbo].Paroquia.IdConcelho = [dbo].Concelho.IdConcelho
39          -- Join [dbo].Distrito on [dbo].Concelho.IdDistrito = [dbo].Distrito.IdDistrito
40          -- Join [dbo].Pais on [dbo].Distrito.CodPais = [dbo].Pais.CodPais
41          where
42          [dbo].Local.Local = 'Desconhecido'
43          and [dbo].Paroquia.IdParoquia = CONVERT(int, @IdLocalSRP)
44          --and [dbo].Concelho.Concelho = 'Desconhecido'
45          --and [dbo].Distrito.Distrito = 'Desconhecido'
46          --and [dbo].Pais.CodPais = 'PT'
47          --and [dbo].Concelho.IdConcelho = CONVERT(int, @IdLocalSRP)
48          --and [dbo].Concelho.IdConcelho = CONVERT(int, @IdLocalSRP)
49          )
50
51  END
52

```

Figura 48 - *Stored Procedure* para obtenção do local

Tratado o local de nascimento, repete-se o procedimento para o local de óbito.

No passo seguinte avalia-se a possível existência de indivíduos duplicados na BDC. Todos os indivíduos que se apresentam nesta situação são sinalizados, armazenando-se o(s) *id*(s) dos possíveis duplicados no respetivo campo de observações para tratamento posterior. A técnica de deteção de duplicados, por ser de elevada proeminência na presente dissertação, encontra-se detalhada no ponto 4.3.4.

Seguidamente, e atendendo a que os indivíduos inseridos na BDC recebem um novo *id* (*IdIndividuo*) e que este valor é necessário para a posterior inserção de informação complementar aos mesmos, como é o caso das residências ou profissões, criou-se um outro *stored procedure* na BDC que permite o

carregamento dos indivíduos e a devolução do respetivo *id* atribuído, funcionalidade não presente nos conetores às BD de destino presentes nos SSIS. Depois da inserção, e com base no novo *id* procede-se ao registo da filiação, que neste novo modelo se faz numa tabela à parte, registando-se ainda na tabela “RELACOES” a associação do *nind* (BDP) com o *IdIndividuo* (BDC).

Seguidamente carregam-se as tabelas com as informações complementares dos indivíduos, com recurso a consulta à tabela RELACOES, nomeadamente das tabelas:

- Profissões
- Assinaturas
- Sufrágios
- Legados profanos
- Apadrinhamentos

Na Figura 49 pode-se observar, a título de exemplo, o fluxo da tabela ASSINATURAS.

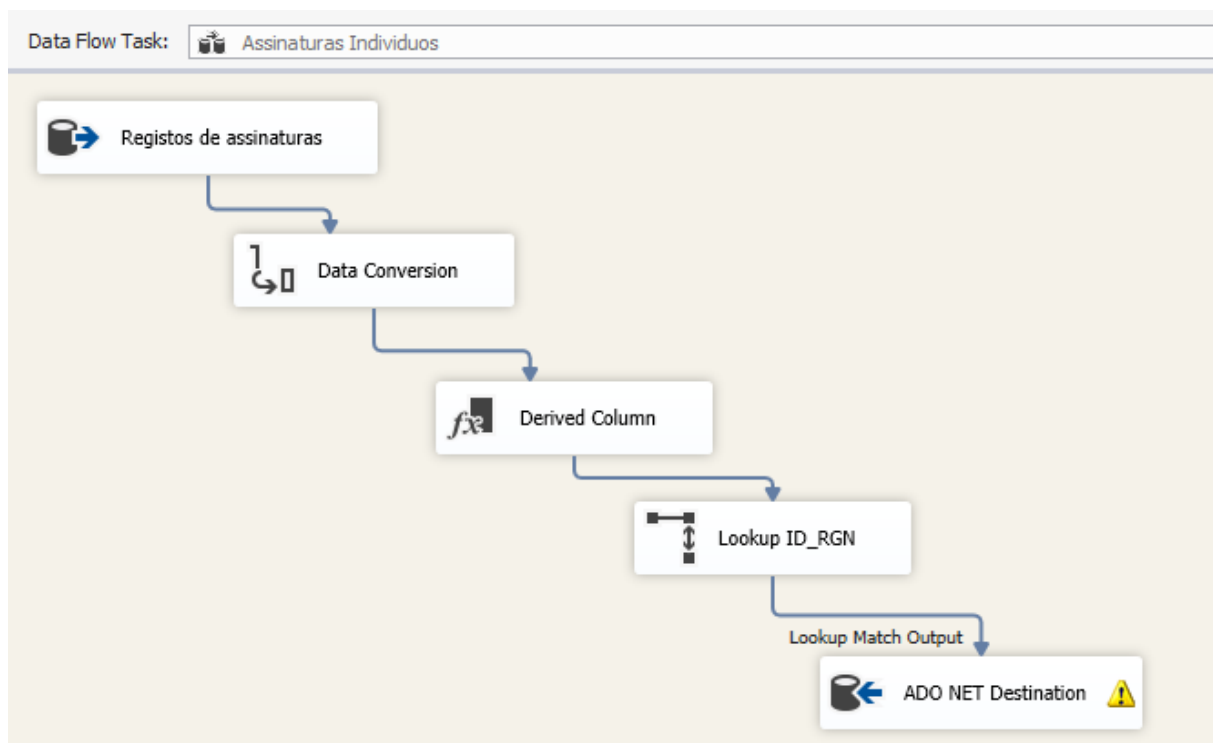


Figura 49 - Fluxo para o carregamento da tabela ASSINATURAS

O passo seguinte do fluxo (Figura 50) passa pelo carregamento da tabela *FAMILIA*. No modelo de dados da BDC cada registo relativo a uma família contempla a identificação (*IdIndividuo*) de cada um dos

indivíduos do casal, pelo que, o processo passa pela recolha da informação relativa a cada um deles na tabela *FAMILIAS*. Na BDP, por cada família registada na tabela *FAMILIA*, existirão duas linhas na tabela *FAMILIAS*, uma para cada elemento do casal, associando esse indivíduo à *FAMILIA*. Depois de colhida a informação, procede-se à recolha do *IdIndividuo* atribuído no RGN para cada um deles, procedendo-se de seguida à inserção da mesma na BDC, recorrendo também a um *stored procedure*, para permitir a recolha do novo id de família (*IdFamilia*) atribuído. Findo este passo, regista-se de seguida a situação da família, e, para utilização nas operações seguintes, armazena-se a associação *nfamilia* (BDP) *IdFamilia* (BDC) na tabela, RELACOESFAMILIARES.

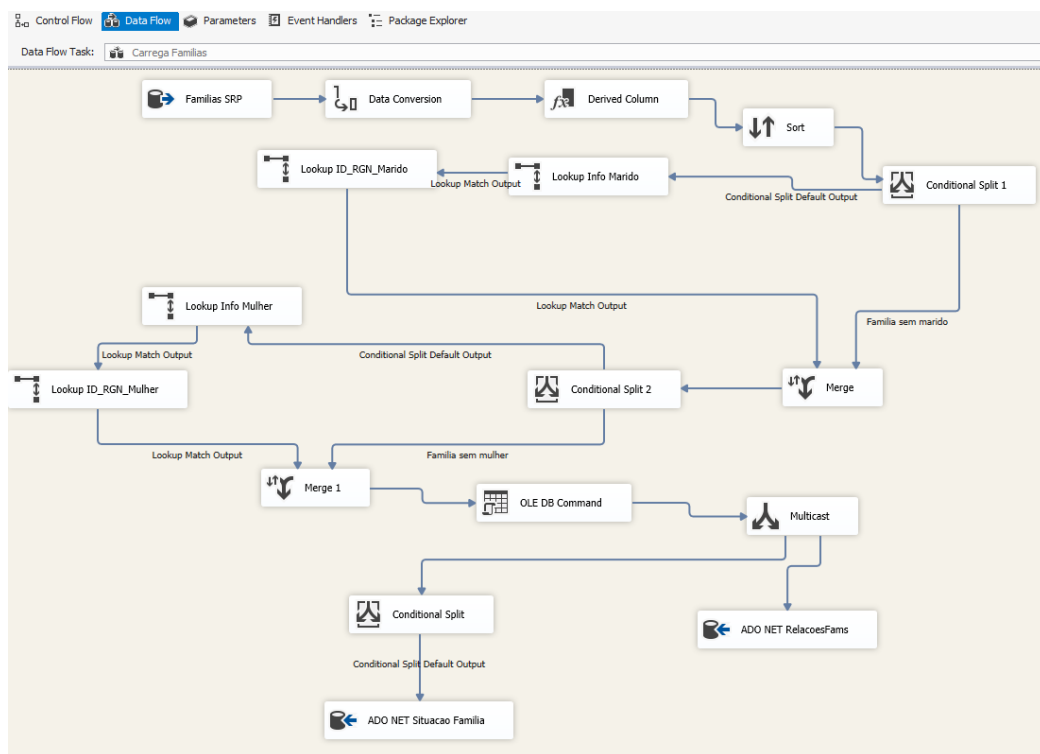


Figura 50 - Fluxo para o carregamento da tabela FAMILIA

No momento seguinte, procede-se ao carregamento da informação para a tabela CASAMENTO (Figura 51) da BDC onde, caso exista informação, se inserem os dados relativos ao casamento da família, nomeadamente, a data e o local do casamento. A associação do local de casamento segue um fluxo similar ao descrito para a associação aos locais de nascimento e óbitos, atrás descritos.

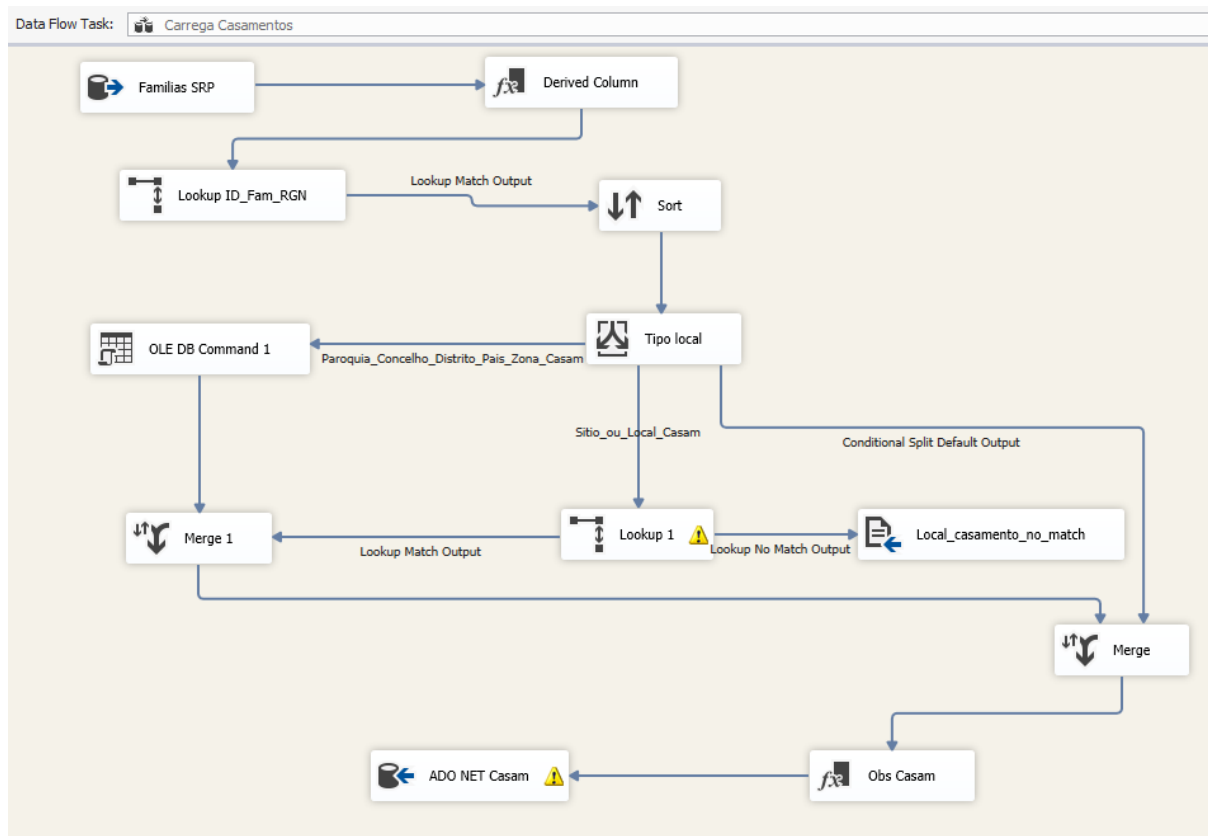


Figura 51 - Fluxo para o carregamento da tabela CASAMENTO

No último momento deste fluxo procede-se ao registo das residências, quer dos indivíduos, quer das famílias. Na BDP esta informação é armazenada numa tabela apenas, existindo um campo *tipoResidencia* com valor “i” ou “f”, para individuo ou família, respetivamente, identificar o tipo de entidade a que se refere, enquanto que na BDC, foram criadas duas tabelas, uma para as residências do individuo e outra para a das famílias. Assim, de acordo com o fluxo da Figura 52, após a identificação do código do lugar – procedimento similar a todas as referências de lugares – efetua-se uma divisão condicional do fluxo, consoante a entidade a que se refere o registo, sendo cada um deles encaminhado para a tabela de destino correspondente.

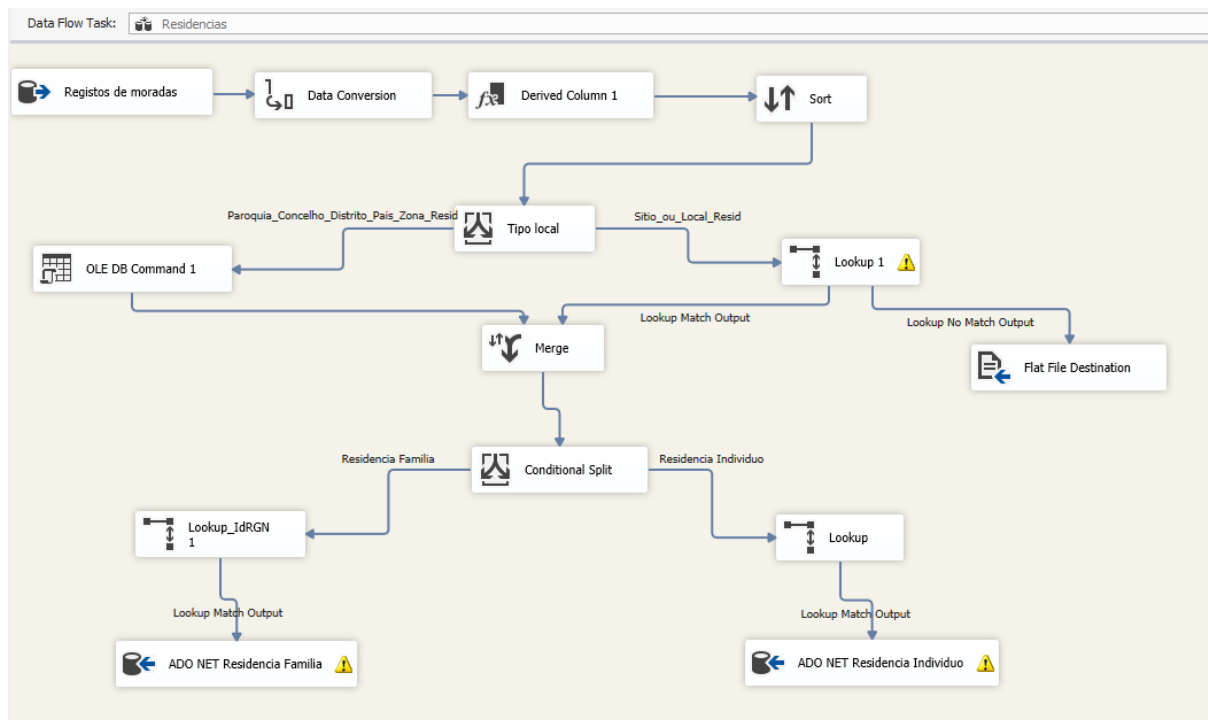


Figura 52 - Fluxo para o carregamento das tabelas RESIDENCIA

Finda a implementação dos processos de ETL, realizaram-se alguns testes em que se verificou a eficácia da integração. Verificou-se, com consultas às BDP de origem e à BDC a presença de todas as entidades principais (Indivíduos e Famílias) e secundárias (Profissões, Residências,...) na BDC, conforme se pode verificar na Figura 53, onde se apresentam os resultados do processo para a paróquia do Corvo.

Comparando-se os valores assinalados na parte esquerda da imagem com os valores assinalados com marcador da mesma cor, do lado direito da imagem, pode-se confirmar a presença do mesmo número de registos na BDP e na BDC, para a mesma paróquia, com a exceção do número de famílias na BDC. Esta diferença deve-se à presença de dois registos de família inconsistentes na BDP que, foram descartados pela aplicação das rotinas de limpeza atrás referidas.

Averiguou-se e confirmou-se, ainda, a presença de todos os valores de todos os atributos relativos a cada uma das entidades, bem como do seu estado de correção, validando-se assim a eficácia dos processos de ETL implementados.

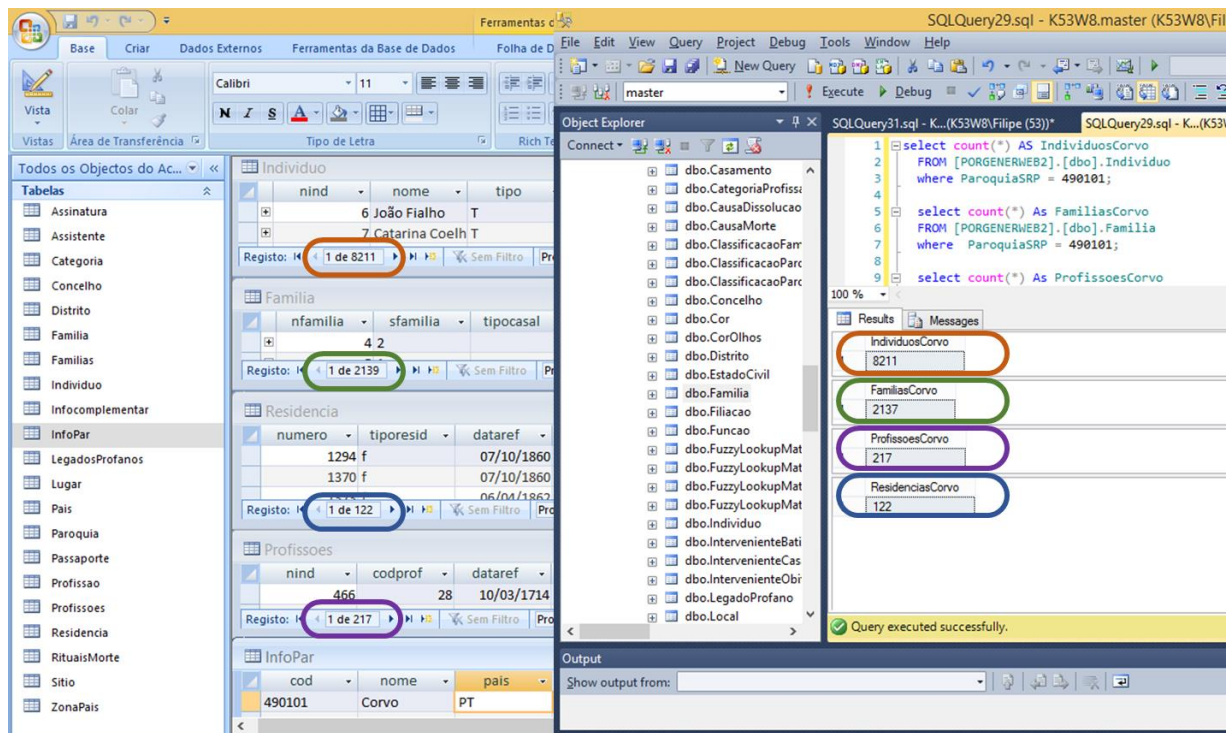


Figura 53 - Resultados do ETL

4.3.4 A Detecção de Duplicados

Existindo uma base de dados por cada paróquia recolhida é expectável que, dada a mobilidade dos indivíduos/famílias entre as paróquias, existam registos duplicados, sendo esta uma grande dificuldade que o GHP enfrenta no desenvolvimento das suas investigações.

Num primeiro momento averiguou-se, junto dos investigadores do GHP, a abordagem “manual” seguida para a realização desta tarefa, ou seja, procurando um indivíduo duplicado, que conjunto de condições o permitirão identificar. Verificou-se que a identificação de um indivíduo dificilmente poderá ser conseguida recorrendo apenas ao conjunto de atributos presentes na ficha de indivíduo, sendo necessário, quase sempre, recorrer à identificação dos pais, realizando-se depois, no enquadramento familiar a pesquisa, por exemplo, “João Pereira da Silva, filho de António da Silva e de Maria Pereira”. Se disponível, adiciona-se a esta pesquisa uma qualquer referência geográfica e/ou temporal, por exemplo, data e local de nascimento, no sentido de reduzir o número de resultados que, na ausência destas referências, pode ascender a centenas ou milhares, o que torna o processo muito moroso.

Tendo por base esta indicação apreendeu-se que os melhores atributos para a execução desta tarefa seriam então, o nome do indivíduo, o nome do pai, o nome da mãe, e, se disponíveis, as datas e locais de nascimento e óbito.

Desenhou-se então uma consulta em SQL que devolve para cada indivíduo este conjunto de atributos, no sentido de se poder testar a deteção de duplicados nas BDP. De seguida testou-se a ferramenta "RecordLinkage" (Borg & Sariyar, 2016), *software* livre com licença *General Public License* (GPL), desenvolvido para o ambiente de programação R, indicado para a realização destas operações.

No primeiro ensaio procurou-se identificar, seguindo as instruções em (Sariyar & Borg, 2010), indivíduos semelhantes utilizando os campos nome do indivíduo e os nomes dos pais, tendo-se aplicado como *blocking* os campos sexo e data de nascimento. Numa primeira operação, o sistema pesquisa os pares de acordo com as condições estabelecidas. De seguida os resultados são classificados automaticamente, com base no valor mínimo de similaridade previamente definido pelo utilizador, que apresente resultados satisfatórios. Definido o intervalo, pode-se extrair o conjunto de pares que apresenta maior verosimilhança.

O sistema devolveu os pares classificados que conseguiu cruzar ("Class=L"- link), indicando o seu nível de similaridade (Figura 54). Verificou-se a eficácia do cruzamento mesmo quando existem ligeiras variações nos nomes dos indivíduos, como se pode verificar no segundo par (linhas 5 e 6). No entanto, esta é uma situação invulgar, em que se dispõe da data de nascimento do indivíduo nos dois ficheiros, o que facilita uma identificação inequívoca.

	A	B	C	D	E	F	G	H
1	nind	nome	sexo	datanasc	painome	maenome	Class	Weight
2	11828	Damásia Peixoto Azevedo	f	16/10/1626 00:00:00	Damáσιο Peixoto Azevedo	Isabel Gomes		
3	19778	damásia peixoto azevedo	f	16/10/1626 00:00:00	damásio peixoto azevedo	isabel gomes	L	0.9332633
4								
5	13987	Jerónima Silva Carvalho	f	15/1/1693 00:00:00	Martinho da Silva	Jerónima da Silva		
6	1441	Jerónima da Silva	f	15/1/1693 00:00:00	martinho coelho	jerónima silva	L	0.9246201
7								
8	17580	Domingos	m	16/9/1703 00:00:00	Domingos Coelho Figueiredo	Maria Azevedo Rosado		
9	21923	domingos	m	16/9/1703 00:00:00	domingos coelho figueiredo	maria azevedo rogada	L	0.9229439
10								
11	253	Margarida Soares	f	13/2/1648 00:00:00	Domingos João	Andreia Lopes		
12	6398	margarida soares	f	13/2/1648 00:00:00	domingos joão	andrea lopes	L	0.9165198
13								

Figura 54 - Resultados do primeiro ensaio R - RecordLinkage

Tentou-se, num segundo ensaio com as mesmas BD, uma abordagem mais próxima da realidade em que habitualmente o grupo trabalha, sem dispor da data ao nascimento nas duas BD. De facto, como já explicado, a possibilidade de duplicação de indivíduos está muitas vezes associada à constituição de famílias em que um dos indivíduos é exterior à paróquia. Nestes casos, com frequência, dispõe-se de uma identificação completa do indivíduo (nome, local de nascimento, identificação dos pais), mas raramente se conhece a sua data de nascimento. Para este ensaio, utilizaram-se os nomes para

comparação utilizando apenas o sexo como filtro (*block*) da formação de pares, isto é, só serão analisados os pares do mesmo sexo avaliando-se a similaridade dos nomes do indivíduo, do pai e da mãe. Os restantes procedimentos são semelhantes aos referidos no primeiro ensaio.

Verificamos que, conforme a Figura 55, mesmo sem dispor da data de nascimento, foi possível estabelecer uma correspondência significativa entre vários indivíduos dos dois ficheiros, no entanto o número de ocorrências é de tal modo elevado que obrigaria o investigador a gastar muito tempo a analisar todos os casos.

	A	B	C	D	E	F	G	H
1	nind	nome	sexo	datanasc	painome	maenome	Class	Weight
2	48619	Maria de Lurdes	f		Francisco Pinto Pereira Cardoso	Carolina Elvira Amaral Fernandes		
3	26025	Maria de Lurdes	f	28/4/1885 00:00:00	Francisco Pinto Pereira Cardoso	Carolina Elvira do Amaral Ferreira	L	0.7044325
4								
5	29142	António Corvas Azevedo	m		Manuel Corvas Azevedo	Casimira Rosa		
6	25431	António	m	14/2/1874 00:00:00	Manuel Corvas Azevedo	Casimira Rosa	L	0.6940789
7								
8	49373	Bernardino Pereira Melo	m		Boaventura Pereira Melo	Ana Joaquina Martins		
9	26233	Bernardino	m	30/11/1893 00:00:00	Boaventura Pereira de Melo	Ana Joaquina Martins	L	0.6890335
10								
11	48379	Ana Corvas de Azevedo	f		Manuel Corvas Azevedo	Casimira Rosa		
12	25514	Ana	f	16/3/1875 00:00:00	Manuel Corvas Azevedo	Casimira Rosa	L	0.6794967
13								
14	49367	Alberto José Maria da Silva Cm	m		António Augusto Silva Carneiro	Cristina Amélia Castro		
15	25311	Alberto	m	31/7/1872 00:00:00	António Augusto da Silva Carneiro	Cristina Amélia de Castro Sampaio	L	0.6624956
16								
17	49367	Alberto José Maria da Silva Cm	m		António Augusto Silva Carneiro	Cristina Amélia Castro		
18	25837	Alberto	m	18/1/1880 00:00:00	António Augusto da Silva Carneiro	Cristina Amélia de Castro Sampaio	L	0.6624956
19								
20	48374	Armanda Alice de Castro San f	f		António Augusto Silva Carneiro	Cristina Amélia Castro		
21	26004	Armanda	f	19/4/1884 00:00:00	António Augusto da Silva Carneiro	Cristina Amélia de Castro Sampaio	L	0.6616045
22								
23	47804	Maria Georgina da Silva Carn f	f		António Augusto Silva Carneiro	Cristina Amélia Castro		
24	25955	Maria	f	14/5/1882 00:00:00	António Augusto da Silva Carneiro	Cristina Amélia de Castro Sampaio	L	0.6607452
25								
26	49369	José Maria da Silva Carneiro	m		António Augusto Silva Carneiro	Cristina Amélia Castro		
27	25689	José	m	25/10/1877 00:00:00	António Augusto da Silva Carneiro	Cristina Amélia de Castro Sampaio	L	0.6601314
--								

Figura 55 - Resultados do segundo ensaio R - RecordLinkage

Encontra-se aqui uma situação de compromisso entre a aplicação dos filtros (*blocking*), que leva a um menor número de ocorrências, mas que levará à omissão de comparações entre pares que potencialmente se referem ao mesmo indivíduo.

É sabido que as técnicas de *Record Linkage*, se baseiam, por norma, no cálculo da similaridade de *Strings* em campos como o nome do indivíduo, combinados com uma dimensão temporal e/ou espacial. Foi já referido também que, neste contexto, o nome de um indivíduo pode sofrer várias alterações, sendo de salientar o costume frequente de, no caso dos registos de batismo, indicar-se apenas o primeiro nome do batizando. Dado que, tipicamente, o primeiro nome do indivíduo não sofre alterações partiu-se para uma abordagem que faz uso apenas desta componente do nome.

Os ensaios efetuados na ferramenta anterior apresentaram bons resultados, no entanto, dada natureza do presente contexto de dados, sabemos que nos encontramos perante uma situação em que a informação não está tendencialmente repetida.

Atentando-se no contexto dos registos paroquiais e na recolha dos mesmos para BDP independentes, facilmente podem concluir-se duas realidades:

- Uma em que não houve mobilidade do indivíduo, estando, portanto, tendencialmente, toda a informação disponível, agregada na base de dados da paróquia em que viveu. Ou seja, será um indivíduo de que se conhece a família de origem, a data e o local de nascimento (proveniente do registo de batismo), se se casou, a data e o local de casamento (oriunda do registo de casamento), e ainda a data e o local de óbito (procedente do seu registo de óbito). Este indivíduo não deverá ter registo duplicado numa outra BDP.
- Outra situação acontece quando pelo menos um destes eventos acontece em paróquias distintas. Assumindo, por exemplo um indivíduo que nasceu na paróquia A e casou na paróquia B, este terá, por norma na BDP da paróquia A, a informação do seu nome (em alguns registos, eventualmente, apenas o nome próprio), da sua data e local de nascimento, bem como a identificação dos pais. Na paróquia B estará registado novamente, sem data de nascimento, associado a um ato de casamento, para o qual se conhece a data, estando, por norma, também identificados os seus pais.

Se se tentar encontrar este registo duplicado nas duas BDP recorrendo ao conjunto de atributos nome do indivíduo, nome do pai e nome da mãe, para a maioria dos nomes mais comuns, o sistema poderá devolver centenas ou milhares de possibilidades. Para se reduzir o número de comparações, aplicam-se, normalmente, técnicas de *blocking*, ou seja, definem-se atributos para os quais tem de existir uma igualdade para que se efetue uma comparação, por exemplo, o indivíduo terá de ter nascido na mesma paróquia e/ou terá de ter data de nascimento comum. Com a aplicação destas técnicas, vemos o número de possibilidades reduzir drasticamente, no entanto, conforme descrito no parágrafo anterior, a informação relativa a um mesmo indivíduo presente em diferentes BDP, tende a ser complementar, o que leva a que a aplicação das técnicas de *blocking* só poderá ser efetuada em atributos que à partida estarão presentes em todas as BDP, independentemente do ato a que se referem. No presente contexto apenas o nome e sexo do indivíduo apresentam esta realidade.

Sabendo-se ainda que, por norma, existirá uma data associada ao RP levantado, decidiu-se experimentar uma abordagem que, faz uso dos atributos sempre presentes, o nome e o sexo, e em que se avaliam as

datas associadas aos atos dos registos, não numa perspetiva de igualdade, mas numa perspetiva de complementaridade.

Enriqueceu-se a consulta às BD de forma a que devolva para cada individuo o seu nome próprio, o nome completo dos pais (se estiver registado um filho para a família, o nome dos pais deverá apresentar-se completo) e as datas de nascimento, de casamento (se houver mais do que uma a mais antiga), de óbito e de nascimento de um filho (se houver mais do que um, também o mais antigo), conforme a Figura 56.

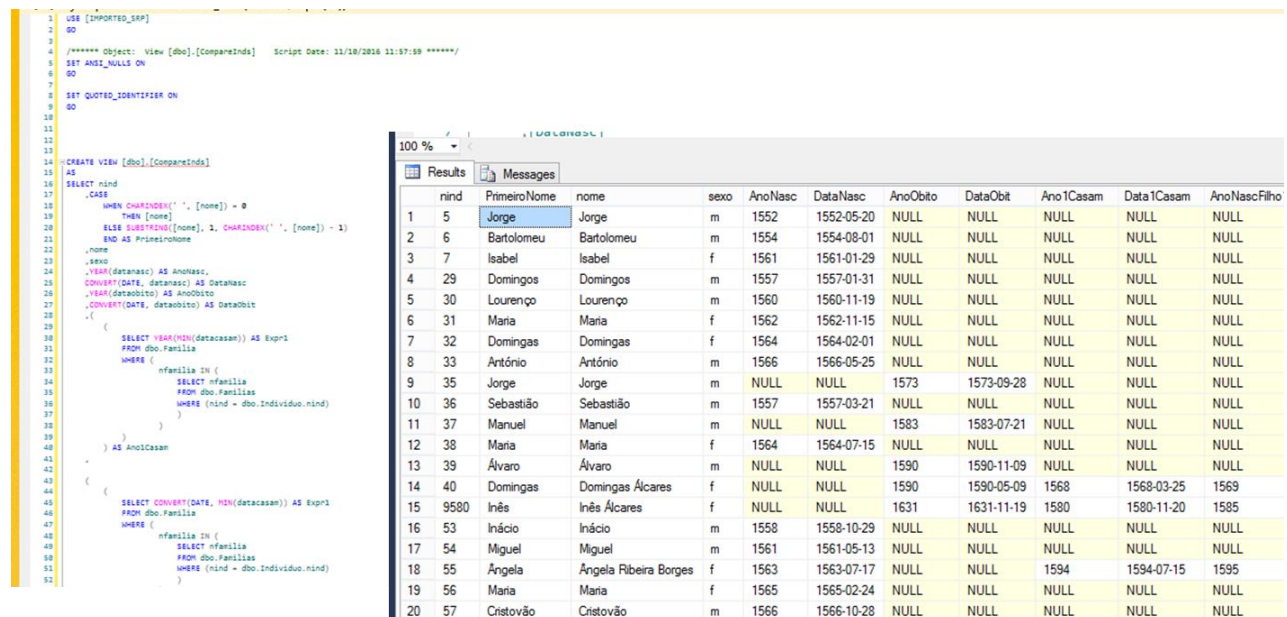


Figura 56 - Consulta para seleção de atributos para deteção de duplicados.

Definiu-se então uma função para SQL SERVER recorrendo às possibilidades da *Common Language Runtime*¹ (CLR) presentes na *framework* .NET, que calcula a similaridade de *strings* entre os nomes, recorrendo ao algoritmo de Jaro-Winkler (Winkler, 1990) que, conforme se pode verificar na Tabela 1, apresenta bons resultados na tarefa de comparação.

De seguida construiu-se uma nova função, Figura 57, que avalia as datas associadas aos mesmos, criando regras de exclusão para os casos em que as datas se apresentam incompatíveis, ou seja:

- A data de óbito de um indivíduo não pode ser anterior à sua data de nascimento, nem posterior em mais de 120 anos e vice-versa.
- As datas de casamento e de nascimento dos filhos têm de estar compreendidas entre as datas de nascimento e óbito (aplica-se uma folga de um ano à data de óbito, para incluir os casos em

¹ <https://msdn.microsoft.com/en-us/library/8bs2ecf4%28v=vs.110%29.aspx?f=255&MSPPErr=-2147217396>

que, para os indivíduos do sexo masculino, o nascimento da criança ocorra após a sua morte), e não podem acontecer nos primeiros 12 anos de idade do indivíduo. O nascimento do filho também não pode acontecer para os indivíduos com mais de 70 anos.

```
[Microsoft.SqlServer.Server.SqlFunction]
public static System.Data.SqlTypes.SqlDouble RGNsimiPONT2(
    int? anoNasca, int? anoCasamA, int? anoObitoA, int? anoNascFilhoA,
    int? anoNascB, int? anoCasamB, int? anoObitoB, int? anoNascFilhoB)
{
    SqlDouble RGN_MatchScore = 0.0;

    //Considerando que só podem viver no máximo até aos 120 anos e casar com 12
    //A idade máxima para ter filhos é aos 70(homens)

    //Se têm anos de nascimento/óbito diferentes não são o mesmo indivíduo
    if (((anoNasca != null && anoNascB != null) && (anoNascB != anoNasca))
        || ((anoObitoA != null && anoObitoB != null) && (anoObitoB != anoObitoA)))
    { return (SqlDouble)0.0; }

    //Se Um só tem data de nascimento e o Outro só tem data de obito e as datas não são compatíveis, não são o mesmo indivíduo
    if (null != anoNasca && anoObitoB != null && (anoObitoB < anoNasca || anoObitoB > anoNasca + 120) ||
        (anoNascB != null && anoObitoA != null) && (anoObitoA < anoNascB || anoObitoA > anoNascB + 120))
    { return (SqlDouble)0.0; }

    //Casamento
    //Se Um só tem data de nascimento e o Outro tem data de casamento e as datas não são compatíveis, não são o mesmo indivíduo
    if ( ((anoNasca != null && anoCasamB != null) && (anoCasamB < anoNasca + 12 || anoCasamB > anoNasca + 108)) ||
        ((anoNascB != null && anoCasamA != null) && (anoCasamA < anoNascB + 12 || anoCasamA > anoNascB + 108)))
    { return (SqlDouble)0.0; }

    //Se Um só tem data de óbito e o Outro tem data de casamento e as datas não são compatíveis, não são o mesmo indivíduo
    if ( (((anoObitoA != null && anoCasamB != null) && (anoCasamB > anoObitoA || anoCasamB < anoObitoA + 108)) ||
        ((anoObitoB != null && anoCasamA != null) && (anoCasamA > anoObitoB || anoCasamA < anoObitoB + 108)))
    { return (SqlDouble)0.0; }

    //Nascimento filho
    //Se Um só tem data de nascimento e o Outro tem data de nascimento de filho e as datas não são compatíveis, não são o mesmo ind
    if ( ((anoNasca != null && anoNascFilhoB != null) && (anoNascFilhoB < anoNasca + 12 || anoNascFilhoB > anoNasca + 70)) ||
        ((anoNascB != null && anoNascFilhoA != null) && (anoNascFilhoA < anoNascB + 12 || anoNascFilhoA > anoNascB + 70)))
    { return (SqlDouble)0.0; }

    //Se Um só tem data de óbito e o Outro tem data de casamento e as datas não são compatíveis, não são o mesmo indivíduo
    if ( (((anoObitoA != null && anoNascFilhoB != null) && (anoNascFilhoB > anoObitoA+1 || anoNascFilhoB < anoObitoA + 108)) ||
        ((anoObitoB != null && anoNascFilhoA != null) && (anoNascFilhoA > anoObitoB+1 || anoNascFilhoA < anoObitoB + 108)))
    { return (SqlDouble)0.0; }

    return (SqlDouble)1;
}
}
```

Figura 57 - Função para a validação das datas dos eventos do indivíduo

Com a aplicação desta função definiu-se uma janela temporal para o período de vida ocorrido ou expectável do indivíduo, comparando-se apenas aqueles que possuem datas para os atos presentes nos RP compreendidas dentro do limite de vida do mesmo.

De seguida realiza-se uma consulta, Figura 58, às BD a serem comparadas, utilizando as funções criadas, procurando pares em que se definem os valores mínimos de similaridade para cada um dos atributos com nome - nesta abordagem existe a possibilidade de se definir para cada um dos nomes o valor mínimo pretendido – e que obtenham o valor 1 como resultado da aplicação da função de validação das datas.

```

SELECT
t1.nind as NindA, t2.nind as NindB,
t1.Nome as NomeA, t2.Nome as NomeB,
t1.NomePai AS NomePaiA, t2.NomePai AS NomePaiB,
t1.NomeMae AS NomeMaeA, t2.NomeMae AS NomeMaeB,
t1.datanasc as DataNascA, t2.datanasc as DatanascB ,
t1.Ano1Casam as AnoCasameto1A, t2.Ano1Casam as AnoCasameto1B,
t1.DataNascFilho1 as DataNascFilho1A, t2.AnoNascFilho1 as DataNascFilho1B,
t1.dataobit as DataObitA, t2.dataobit as DataObitB,

[dbo].[StringDistance](t1.PrimeiroNome, t2.PrimeiroNome) as StringDistanceNome,
[dbo].[StringDistance](t1.NomePai, t2.NomePai) as StringDistancePai,
[dbo].[StringDistance](t1.NomeMae, t2.NomeMae) as StringDistanceMae

FROM [IMPORTED_SRP_GOLAES].[dbo].[CompareInds] as t1
INNER JOIN [IMPORTED_SRP_FAFE].[dbo].[CompareInds] as t2
ON

[dbo].[StringDistance](t1.PrimeiroNome, t2.PrimeiroNome) > 0.87 and
[dbo].[StringDistance](t1.NomePai, t2.NomePai) > 0.87 and
[dbo].[StringDistance](t1.NomeMae, t2.NomeMae) > 0.87 and
[dbo].[RGNSimiPONT2](t1.anonasc,t1.Ano1Casam,t1.anoobito,t1.AnoNascFilho1,
t2.anonasc,t2.Ano1Casam,t2.anoobito,t2.AnoNascFilho1) = 1

and t1.sexo = t2.sexo

```

Figura 58 - Consulta para a avaliação de duplicados

Para avaliar a eficácia desta abordagem experimentou-se a sua aplicação na comparação entre duas BDP de paróquias contíguas, Fafe e Golães, respetivamente, dado existir uma grande possibilidade de existência de mobilidade dos indivíduos entre elas. Estas BDP foram recolhidas pelo mesmo investigador que, sempre que conseguia identificar numa BDP indivíduos presentes na outra, procurava registar a informação correspondente nas duas BDP, o que permite uma maior validação do possível par.

Avaliaram-se então as duas BD, tendo-se definidos os seguintes parâmetros:

- A similaridade do nome próprio do indivíduo é superior a 0.87
- A similaridade do nome completo do pai do indivíduo é superior a 0.87
- A similaridade do nome completo da mãe é superior a 0.87
- A função de validação das datas tem de devolver o valor 1, o que significa que as datas que lhe estão associadas são compatíveis.

O procedimento devolveu apenas 182 resultados, cujos 10 primeiros pares podem ser observados na Tabela 19. Observando-se em detalhe cada par, pode-se verificar que todos eles cumprem todos os requisitos para serem considerados possíveis duplicados.

Tabela 19 – Resultados da comparação em SQL

NIND	Atributos comparados			Datas (anos) avaliados				Similarida.		
	NOME	PAI	MAE	NASC	CAS AM	NASCFILHO	OBITO	N	P	M
2160	Cristóvão de Castro	Jerónimo de Castro	Ana Gonçalves	1693-08-08	1728		1747-09-02	1	1	1
1669	Cristóvão de Castro	Jerónimo de Castro	Ana Gonçalves	1693-08-08	1728	1729	1747-09-02	1	1	1
2168	João de Castro Soares	João de Castro	Maria de Barros	1687-10-26	1730	1732	1755-05-25	1	1	1
1765	<i>João de Castro Soares</i>	<i>João de Castro</i>	<i>Maria de Barros</i>	<i>1687-10-26</i>		<i>1739</i>	<i>1755-05-25</i>	<i>1</i>	<i>1</i>	<i>1</i>
2152	Cristóvão da Costa	João Durães	Catarina da Costa	1690-11-11	1724	1720	1743-05-30	1	1	1
1786	Cristóvão da Costa	João Durães	Catarina da Costa	1690-11-11	1724	1720	1743-05-30	1	1	1
7553	Francisco de Freitas	Manuel de Freitas	Antónia Maria Pereira da Silva	1858-12-18	1887			1	1	0,9
1581977	Francisco Freitas	Manuel de Freitas	Antónia Maria Pereira			1889	1900-12-23	1	1	0,9
4196	Custódio José de Castro	João Baptista de Castro	Custódia de Castro Fernandes	1777-10-12	1798	1799		1	1	1
4734	Custódio José de Castro	João Baptista de Castro	Custódia de Castro Fernandes	1777-10-12		1799		1	1	1
4596	António José de Castro	João Baptista de Castro	Custódia de Castro Fernandes	1779-12-29	1809	1810		1	1	1
4735	António José de Castro	João Baptista de Castro	Custódia de Castro Fernandes	1779-12-29				1	1	1
3504	Maria Joana de Oliveira	Luís da Silva	Josefa de Oliveira da Silva	1770-11-10	1796	1796	1819-09-10	1	1	1
15856	Maria Joana de Oliveira	Luís da Silva	Josefa de Oliveira da Silva	1770-11-10	1796	1801	1819-09-10	1	1	1
2105	Luísa Fernandes	António Francisco	Senhorinha Fernandes	1713-07-31	1745	1745		1	1	1
15950	Luísa Fernandes	António Francisco	Senhorinha Fernandes			1753	1763-10-26	1	1	1
697	Maria de Araújo	Pedro de Araújo	Catarina Veloso	1661-01-14			1704-07-26	1	1	1
14426	Maria de Araújo	Pedro de Araújo	Catarina Veloso	1661-01-14	1685	1686	1704-07-26	1	1	1
2763	Mariana (Solteira)	António Soares	Catarina da Silva	1700-11-06		1741	1752-12-17	1	1	1
1149	Mariana	António Soares	Catarina da Silva	1700-11-06			1752-12-17	1	1	1

De seguida, recorreu-se novamente ao ambiente R e, utilizando-se uma vez mais o *package* “RecordLinkage”, avaliaram-se as mesmas BDP, aplicando-se uma definição de parâmetros semelhante: os nomes com similaridade maior que 0.87; aplicação de *blocking* no atributo sexo.

Após a execução da operação obtiveram-se, desta ferramenta, 26631₁ pares possíveis. Avaliou-se de seguida a presença dos pares encontrados com a função à medida, e confirmou-se a presença de todas as possibilidades. Seguidamente, assinalaram-se nos resultados do R aqueles que não estão presentes nos resultados da função à medida, para avaliar, mediante as datas presentes nos registos, se continham valores possíveis de ditar a exclusão dessa comparação.

A Tabela 20 apresenta os primeiros 11 pares resultantes dessa classificação, todos com medida de similaridade no valor máximo (1.0). Pode-se verificar imediatamente que o primeiro par não pode representar a mesma pessoa, dado que os dois indivíduos considerados apresentam datas de

nascimento diferentes. Os pares com *nind* 6334 e 1582083 também não são compatíveis dado que o indivíduo com *nind* 1582083 apresenta uma data de casamento posterior à data de óbito do indivíduo com *nind* 6334.

Avaliaram-se de seguida as primeiras 100 ocorrências, em detalhe e verificou-se que, de facto, os pares aqui apresentados que não constam dos resultados da função desenhada para o processo, se referem a falsas associações de indivíduos, o que comprova a eficiência da ferramenta desenvolvida.

Tabela 20 – Resultados da comparação na ferramenta "R-RECORDLINKAGE"

nind	PrimeiroNome	Válido	nome	NomePai	NomeMãe	Weight	DataNasc	DataObit	Data1Casam	DataNascFilho1	Classes
320	Maria	FALSO	Maria	Gonçalo Gonçalves	Maria Gonçalves		1646-04-15				
623	Maria	FALSO	Maria Frutuosa Gonçalves	Gonçalo Gonçalves	Maria Gonçalves	1.000000	1636-05-18	1709-05-01	1667-08-19	1670-03-08	L
348	Gonçalo	VERDADO	Gonçalo de Freitas	Jacinto Salgado	Isabel de Freitas		1648-03-15	1734-07-09			
1589828	Gonçalo	EIRO	Gonçalo de Freitas	Jacinto Salgado	Isabel de Freitas	1.000000		1734-07-09	1670-04-19	1670-06-05	L
526	João	VERDADO	João de Freitas	António Gonçalves	Maria de Freitas		1655-05-16	1720-08-08	1688-02-01	1689-01-09	
14573	João	EIRO	João de Freitas	António Gonçalves	Maria de Freitas	1.000000	1655-05-16	1720-08-08	1688-02-01		L
697	Maria	VERDADO	Maria de Araújo	Pedro de Araújo	Catarina Veloso		1661-01-14	1704-07-26			
14426	Maria	EIRO	Maria de Araújo	Pedro de Araújo	Catarina Veloso	1.000000	1661-01-14	1704-07-26	1685-02-25	1686-02-13	L
834	João	FALSO	João	Pedro Francisco	Maria de Freitas		1670-10-22				
14575	João	FALSO	João de Freitas	Pedro Francisco	Maria de Freitas	1.000000	1672-08-25	1735-07-22	1690-05-17		L
899	Domingos	VERDADO	Domingos Ribeiro	Sebastião Gonçalves	Maria Ribeiro				1681-11-05		
14461	Domingos	EIRO	Domingos Ribeiro	Sebastião Gonçalves	Maria Ribeiro	1.000000			1681-11-05	1682-11-26	L
932	Mateus	VERDADO	Mateus Gonçalves	Francisco Gonçalves	Catarina Ribeiro		1672-06-05	1706-04-20			
14605	Mateus	EIRO	Mateus Gonçalves	Francisco Gonçalves	Catarina Ribeiro	1.000000	1672-06-05	1706-04-20	1689-07-27	1697-04-09	L
938	João	VERDADO	João de Freitas	Pedro Francisco	Maria de Freitas		1672-08-25	1735-07-22	1690-05-17	1691-09-05	
14575	João	EIRO	João de Freitas	Pedro Francisco	Maria de Freitas	1.000000	1672-08-25	1735-07-22	1690-05-17		L
952	Ana	FALSO	Ana	Francisco Pires	Catarina Francisca		1673-09-15				
401	Ana	FALSO	Ana Francisca	Francisco Pires	Catarina Francisca	1.000000	1631-03-30	1694-11-04	1643-10-15	1650-09-21	L
6861	Clementina	VERDADO	Clementina Soares (Solteira)	Custódio Soares	Maria de Castro		1836-05-28			1864-03-19	
6318	Clementina	EIRO	Clementina Soares	Custódio Soares	Maria de Castro	1.000000	1836-05-28				L
6334	João	FALSO	João	José da Costa	Antónia Maria de Freitas		1851-03-27	1852-08-12			
1582083	João	FALSO	João Costa	José da Costa	Antónia Maria de Freitas	1.000000			1888-04-14	1890-01-30	L

5 CONCLUSÕES

Apresenta-se neste capítulo uma breve descrição do trabalho realizado, bem como uma síntese dos resultados obtidos tendo em consideração os objetivos e finalidade propostos. Apresentam-se, ainda, alguns aspetos, considerados passíveis de sofrerem melhorias em trabalhos futuros.

5.1 Síntese

A presente dissertação teve como finalidade a concretização de uma BDC com capacidade para integrar, consolidar e fundir todas as BDP detidas pelo GHP, e, ao mesmo tempo, resolver as limitações que o modelo de dados da BDP apresenta na recolha da informação disponível nos RP.

Tendo-se presente esta finalidade e os objetivos definidos para a execução da mesma, avançou-se para a realização do enquadramento conceptual. Estudou-se a temática da integração e fusão de dados, tendo-se averiguado os seus objetivos, desafios e dificuldades, bem como as abordagens a seguir na realização de tais tarefas. Analisaram-se as questões da qualidade dos dados, tendo-se refletido nas diversas dimensões associadas a este tópico. Investigou-se ainda o tema *Record Linkage*, tendo-se estudado a sua aplicação, os desafios que apresenta e a sua utilização em contextos com ficheiros genealógicos. Perspetivando-se a integração dos dados com recurso a processos de ETL, estudou-se também este conceito, tendo-se também selecionado e aplicado uma metodologia de implementação adequada a estes procedimentos.

Analisou-se, de seguida o modelo de dados da BDP, tendo-se averiguado junto dos investigadores do GHP, as limitações e constrangimentos que o mesmo apresenta. Com base na informação recolhida, estudou-se, propôs-se e implementou-se a BDC, cujo modelo de dados detém a capacidade de, por um lado, suprimir as limitações identificadas e, por outro, corresponder aos requisitos que a fusão das BDP compreende.

Tendo sempre presente os conceitos estudados no momento do enquadramento conceptual, idealizou-se, modelou-se e implementou-se, um conjunto de processos de extração, transformação e carregamento de dados, capaz de, em primeiro lugar, avaliar e tratar das inconsistências dos dados presentes em cada uma das BDP, procedendo depois às transformações de entidades e dados necessárias, para que correspondam aos formatos definidos na BDC. Estes processos realizam, de seguida, o carregamento dos dados para a BDC, garantindo a preservação de todos os registos e os atributos consistentes, presentes em cada uma das BDP.

Criou-se ainda uma funcionalidade para a deteção de possíveis registos de indivíduos duplicados, ajustada ao presente contexto de dados e às necessidades do GHP que se revelou de elevada eficácia.

5.2 Resultados

Os trabalhos realizados durante a presente dissertação resultaram num conjunto de elementos capazes de responder a todos os objetivos definidos, podendo-se considerar que a mesma teve resultados muito positivos. Foi, então, desenhada e construída a BDC, modelados e implementados os processos de ETL, bem como as rotinas de integração e deteção de registos duplicados, tendo as mesmas sido validadas. A combinação destes elementos resultou num artefacto que contempla a BDC e um conjunto de procedimentos capazes de integrar e fundir cada uma das BDP para este repositório único, conforme o desejado pelos investigadores do GHP, para o desenvolvimento de pesquisas e análises mais abrangentes, possíveis apenas com esta realidade.

Importa ainda referir que uma componente do presente trabalho foi já apresentada no *XI Congreso de la Asociación de Demografía Histórica, ADEH*, que se realizou em Cádiz, em junho de 2016 com a comunicação “Da reconstituição de famílias ao Repositório Genealógico - Uma via aberta para as ciências sociais” (Amorim et al., 2016).

5.3 Contribuições

Dada a especificidade e a natureza singular que qualquer projeto de fusão e integração de dados apresenta, poder-se-á considerar, por si só o artefacto gerado, um contributo.

De referir ainda que, o mecanismo de identificação de registos duplicados pode qualificar-se como inovador, uma vez que, partindo das abordagens mais comuns para realização de RL, se criou um instrumento que avalia não só a proximidade do valor dos atributos presentes nos dois conjuntos de dados, mas que estende esta avaliação para uma janela temporal do período de vida expectável de um indivíduo. Eliminam-se assim as ocorrências em que, apesar da similaridade dos atributos presentes, as datas dos atos dos RP não se enquadram. Dado o contexto de dados das BDP, em que se sabe que, para os casos em que houve mobilidade, estas informações dos atos, principalmente as datas dos mesmos, se encontram dispersas por BD distintas, esta funcionalidade apresenta-se como um mecanismo de elevado valor para a identificação dos registos com informação complementar. Isto representa indubitavelmente uma mais valia para a área de negócio em questão.

5.4 Trabalho Futuro

O trabalho desenvolvido na presente dissertação respondeu a todos os desafios colocados, no entanto, considera-se que existem alguns aspetos que poderão ser alvo de melhoria e, conseqüentemente, desenvolvidos em trabalhos futuros, nomeadamente:

- Poder-se-á criar a possibilidade de executar a ferramenta de deteção de duplicados em subconjunto de dados (por exemplo só nos nomes iniciados pelas letras de A a F) e iterar sobre estes subconjuntos ate correr toda a BD, dado que a BDC irá aumentar consideravelmente, o que levará a que o número de comparações aumente exponencialmente
- Poder-se-á ainda disponibilizar a funcionalidade de deteção de duplicados para a interface de recolha de dados de modo a avaliar a presença de duplicados no momento da transcrição dos RP
- Outra possibilidade, será a criação de uma interface que permita ao investigador visualizar a informação dos indivíduos duplicados e seleccionar, de cada par apresentado, a informação que pretende manter.

Estas sugestões visam aprimorar o trabalho desenvolvido com este tema de dissertação, embora o núcleo central tenha sido já implementado e devidamente testado. Com estas sugestões conseguir-se-á melhorar o desempenho do processo de deteção de duplicados, retratado neste documento, bem como, estender a sua aplicação ao contexto da recolha dos dados a partir dos RP.

REFERÊNCIAS BIBLIOGRÁFICAS

- Amorim, N. (1991). *Uma metodologia de reconstituição de paróquias*. Universidade do Minho. Retrieved from <https://books.google.pt/books?id=IE9bMwEACAAJ>
- Amorim, N., & Ferreira, A. (2006). Os registos paroquiais como fontes para a demografia histórica e história social.
- Amorim, N., Santos, M. Y., Ferreira, A., & Salgado, F. (2016). Da reconstituição de famílias ao Repositório Genealógico - Uma via aberta para as ciências sociais.
- Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, *31*(2), 150–162.
- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, *41*(3), 16:1–16:52. <http://doi.org/10.1145/1541880.1541883>
- Batini, C., & Scannapieco, M. (2006). *Data Quality*. Springer. Springer Berlin Heidelberg. <http://doi.org/10.1007/3-540-33173-5>
- Bernstein, P. a, Madhavan, J., & Rahm, E. (2011). Generic Schema Matching , Ten Years Later. *Proceedings of the VLDB Endowment*, *4*(11), 695–701.
- Bleiholder, J., & Naumann, F. (2005). Declarative data fusion - Syntax, semantics, and implementation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3631 LNCS, pp. 58–73).
- Bleiholder, J., & Naumann, F. (2006). Conflict Handling Strategies in an Integrated Information System. *Proceedings of the IJCAI Workshop on Information on the Web*, (197), 1–13.
- Bleiholder, J., & Naumann, F. (2008). Data fusion. *ACM Comput. Surv.*, *41*(1), 1–41. <http://doi.org/10.1145/1456650.1456651>
- Borg, A., & Murat Sariyar. (2016). *Package "RecordLinkage."* Retrieved from : <https://CRAN.R-project.org/package=RecordLinkage>
- Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. In *International Journal of Intelligent Systems* (Vol. 18, pp. 51–74).
- Brazhnik, O., & Jones, J. F. (2007). Anatomy of data integration. *Journal of Biomedical Informatics*, *40*(3), 252–269.
- Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and*

- duplicate detection. Change*. Springer. <http://doi.org/10.1007/978-3-642-31164-2>
- Dunn, H. L. (1946). Record linkage. *American Journal of Public Health*, 36(12), 1412–1416. <http://doi.org/10.1111/j.1365-2265.2010.03958.x>
- Eckerson, W. W. (2002). Data quality and the bottom line. *TDWI Report, The Data Warehouse Institute*.
- Eckerson, W., & White, C. (2003). Evaluating ETL and Data Integration Platforms. *TDWI Report Series*.
- El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences*, 23(2), 91–104. <http://doi.org/10.1016/j.jksuci.2011.05.005>
- Faria, F. (2004). *SRP - Sistema de Reconstituição de Paróquias*. SEED – Módulo de Aquisição de dados. Guimarães. NEPS/ Departamento de Informática da Universidade do Minho.
- Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210.
- Ferreira, A. (2002). Contributos da reconstituição de paróquias para a investigação genealógica.
- Ferreira, A. (2004). *Sistemas Informáticos para análise de dados demográficos: uma abordagem histórica*. SIAD'04: *Sistemas informáticos para a análise de dados demográficos*. Guimarães: Universidade do Minho. Núcleo de Estudos de População e Sociedade.
- Gu, L., & Baxter, R. (2003). Record linkage: Current practice and future directions. *Cmis*, 03/83. Retrieved from http://festivalofdoubt.uq.edu.au/papers/record_linkage.pdf
- Halevy, A., & Ordille, J. (2006). Data Integration : The Teenage Years. *Artificial Intelligence*, 41(1), 9–16. <http://doi.org/http://portal.acm.org/citation.cfm?id=1182635.1164130&coll=Portal&dl=GUIDE&CFID=38899936&CFTOKEN=23237860>
- Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. *Data Quality and Record Linkage Techniques*. Springer New York.
- Ivie, S., Pixton, B., & Giraud-Carrier, C. (2007). Metric-based data mining model for Genealogical Record Linkage. In *2007 IEEE International Conference on Information Reuse and Integration, IEEE IRI-2007* (pp. 538–543).
- Madhavan, J., Bernstein, P. a, & Rahm, E. (2001). Generic Schema Matching with Cupid. *VLDB Journal*, 10(1), 15.
- Naumann, F. (2002). *Quality-Driven Query Answering for Integrated Information Systems*. *Quality-Driven Query Answering for Integrated Information Systems* (Vol. 2261). Springer Berlin Heidelberg. <http://doi.org/10.1007/3-540-45921-9>
- Naumann, F., & Bleiholder, J. (2006). Data Fusion in Three Steps : Resolving Inconsistencies at Schema.

- Bulletin of the Technical Committee on Data Engineering*, 1–11.
- Neiling, M. (1998). Data fusion with record linkage. In *Proceedings of the Workshop 'Foederierte Datenbanken', Aachen*.
- NeSmith, N. P. (1992). Record Linkage and Genealogical Files. *Utah Genealogical Journal*, 20(3–4), 113–119.
- Newcombe, H. B., & Kennedy, J. M. (1962). Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information. *Communications of the ACM*, 5(11), 563–566. Retrieved from <http://portal.acm.org/citation.cfm?doid=368996.369026>
- Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, a P. (1959). Automatic linkage of vital records. *Science (New York, N.Y.)*, 130, 954–959.
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research methodology for Information Systems research. *Journal of Management Information Systems*, 24(3), 45–77. <http://doi.org/10.2753/MIS0742-1222240302>
- Sariyar, M., & Borg, A. (2010). The RecordLinkage Package: Detecting Errors in Data. *R Journal*, 2(2), 61–67. Retrieved from http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf
- Vassiliadis, P., Simitsis, A., & Baikousi, E. (2009). A taxonomy of ETL activities. *Proceedings of the ACM Twelfth International Workshop on Data Warehousing and OLAP*, 25–32. Retrieved from <http://dl.acm.org/citation.cfm?id=1651297>
- Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., & Skiadopoulou, S. (2005). A generic and customizable framework for the design of ETL scenarios. In *Information Systems* (Vol. 30, pp. 492–525).
- Vassiliadis, P., Simitsis, A., & Skiadopoulou, S. (2002). Conceptual modeling for ETL processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP - DOLAP '02* (pp. 14–21). Retrieved from <http://dl.acm.org/citation.cfm?id=583890.583893>
- Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86–95.
- Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Source Journal of Management Information Systems*, 12(4), 5–33. Retrieved from <http://www.jstor.org/stable/40398176><http://www.jstor.org/page/info/about/policies/terms.jsp>
- Wilson, D. R. (2008). Genealogical Record Linkage on International Data. *Eighth Annual Workshop on*

Technology for Family History and Genealogical Research.

- Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research, American Statistical Association*, 354–359. Retrieved from <http://files.eric.ed.gov/fulltext/ED325505.pdf>
- Winkler, W. E. (2006). Overview of record linkage and current research directions. *Current*, (2006–2), 1–28. <http://doi.org/10.1206/3728.2>
- Winkler, W. E. (2014). Matching and record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(5), 313–325.

ANEXOS

Diagrama de Entidades e Relacionamentos da Base de Dados Paroquial

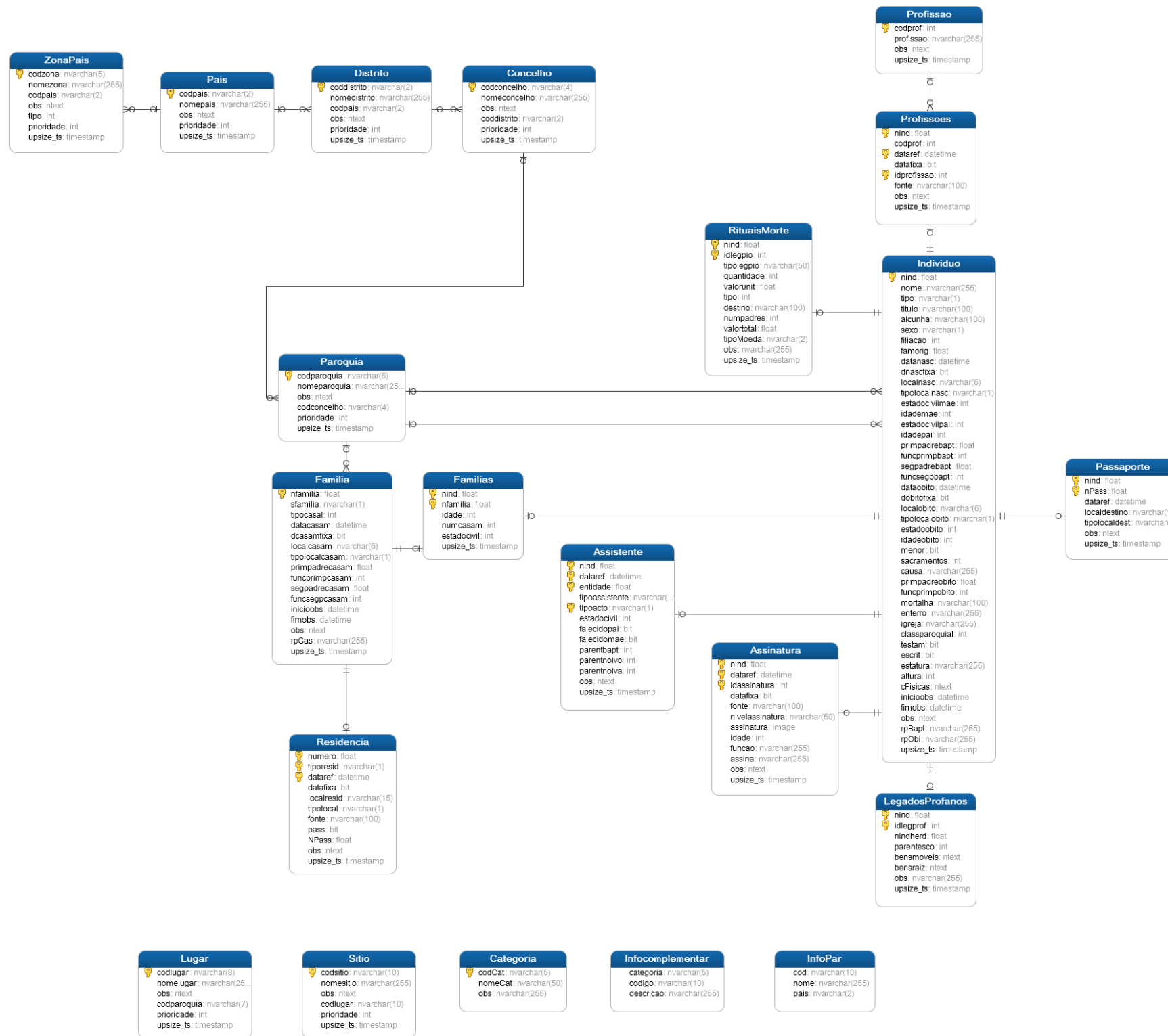
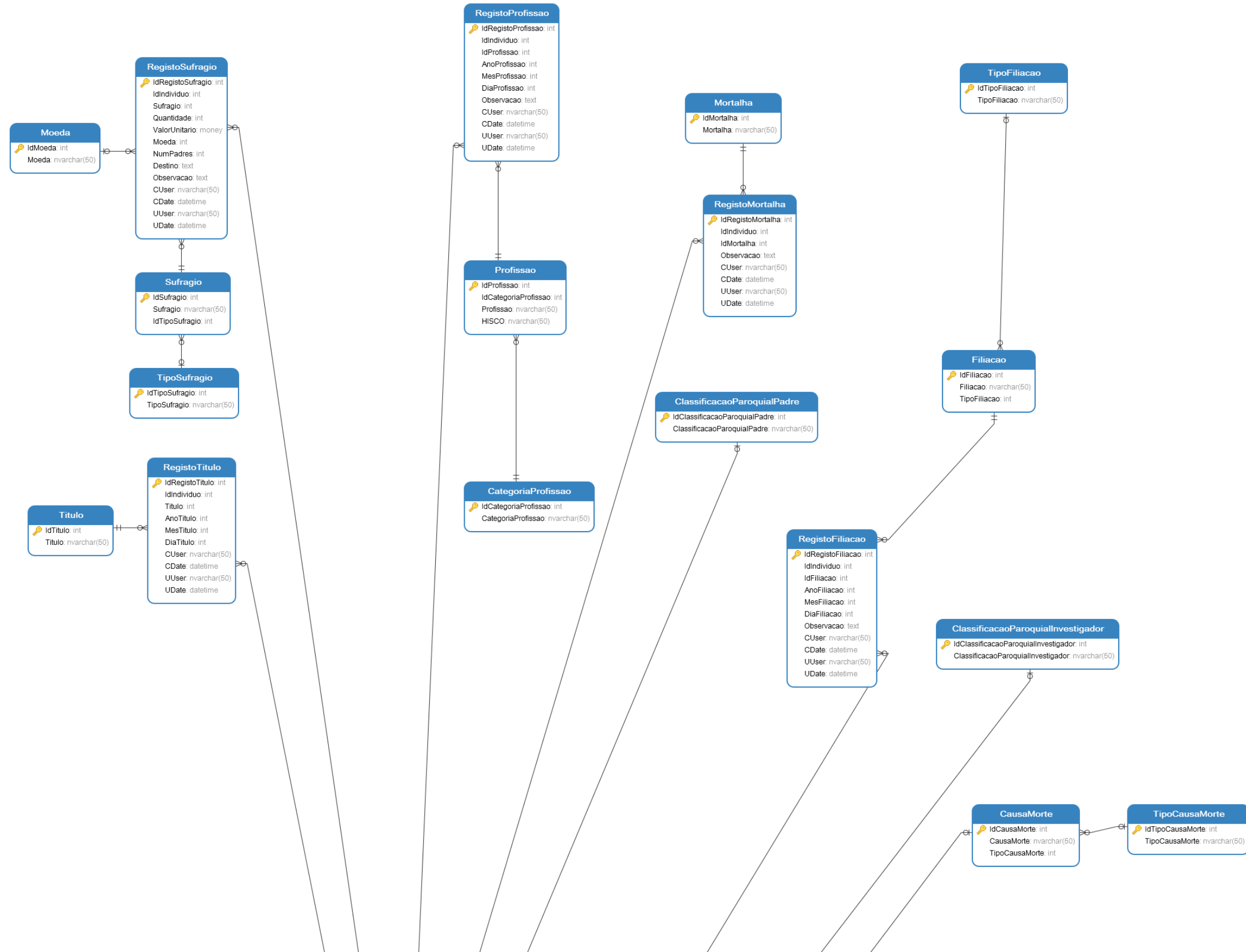
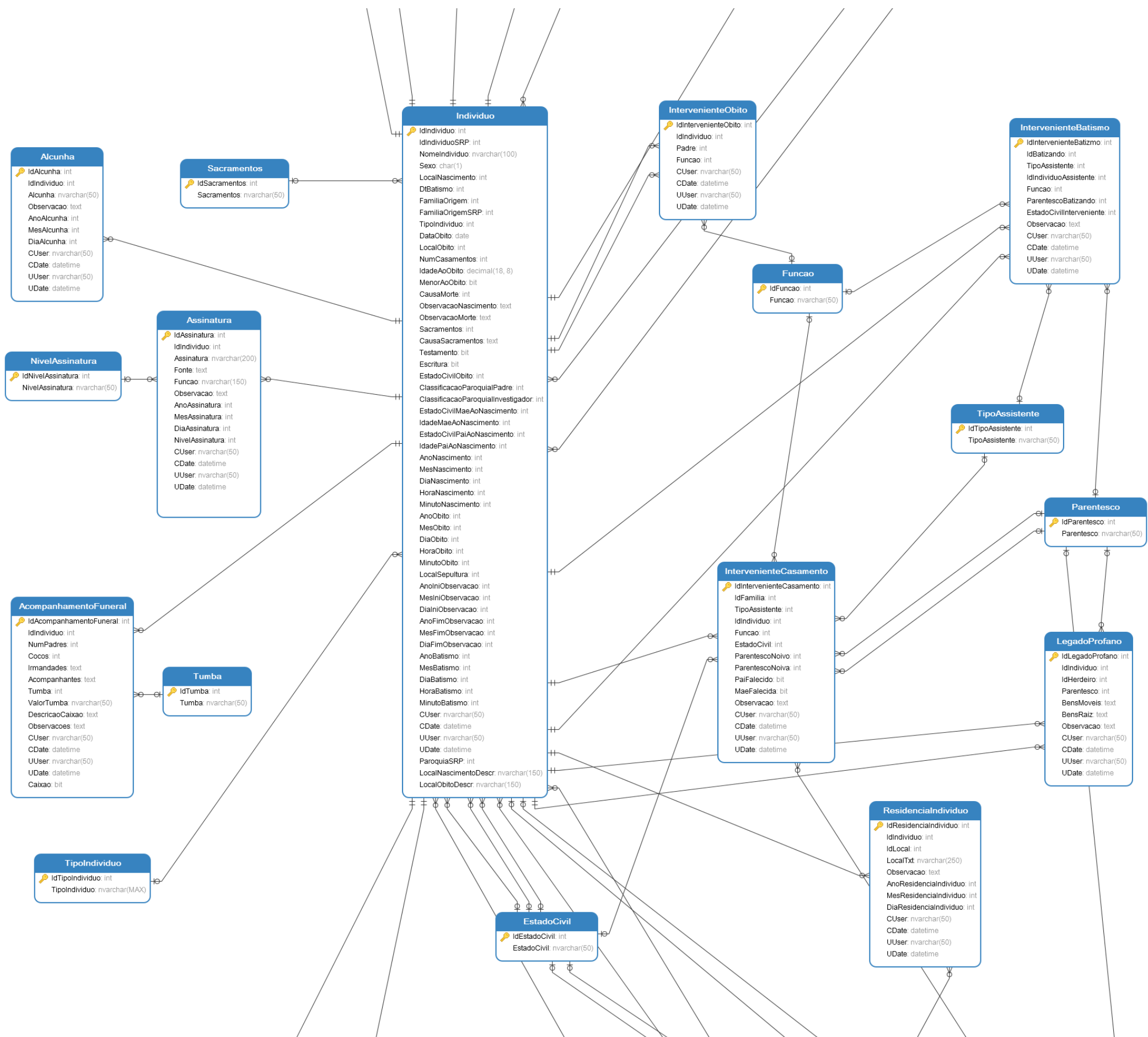
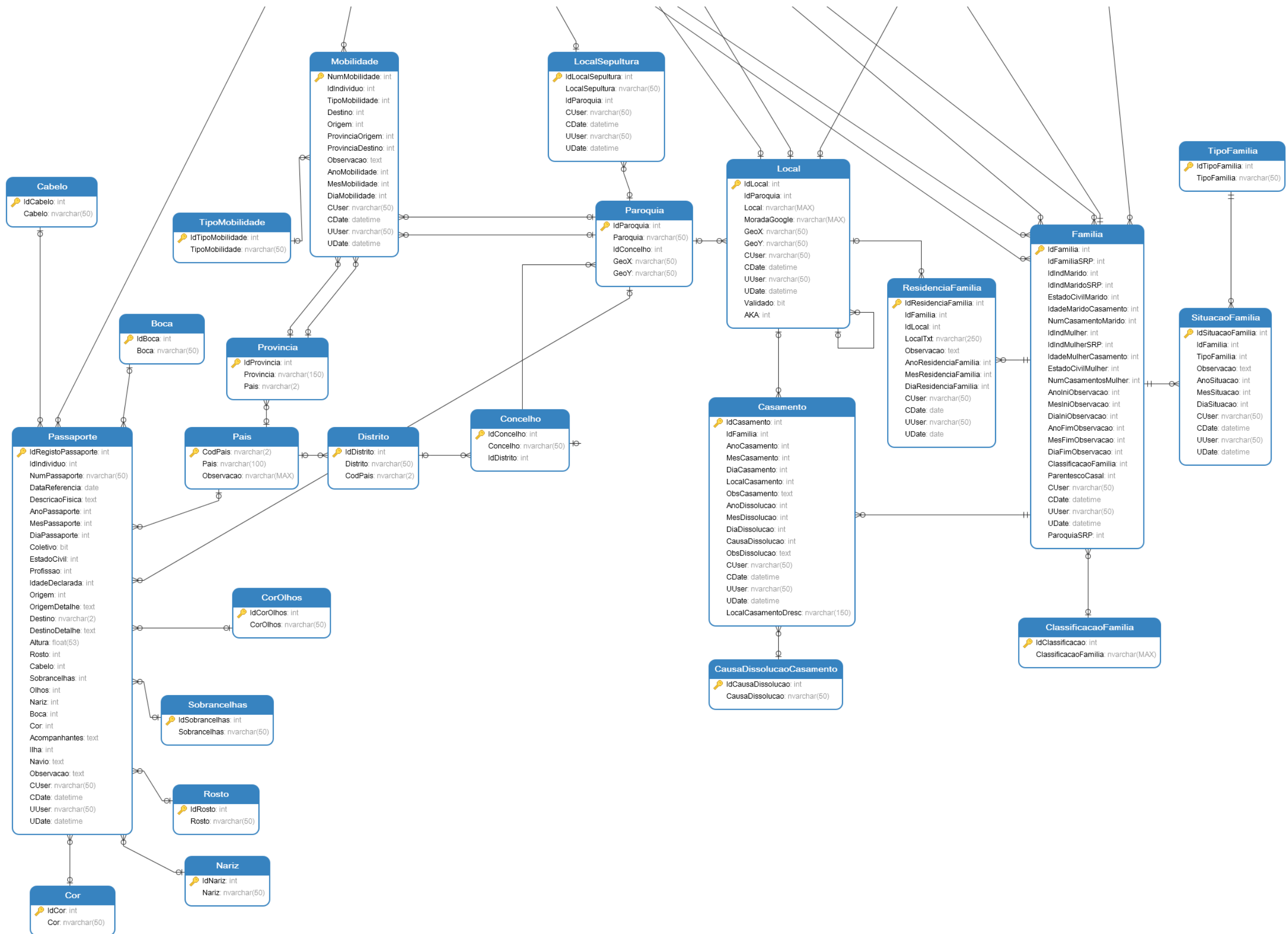


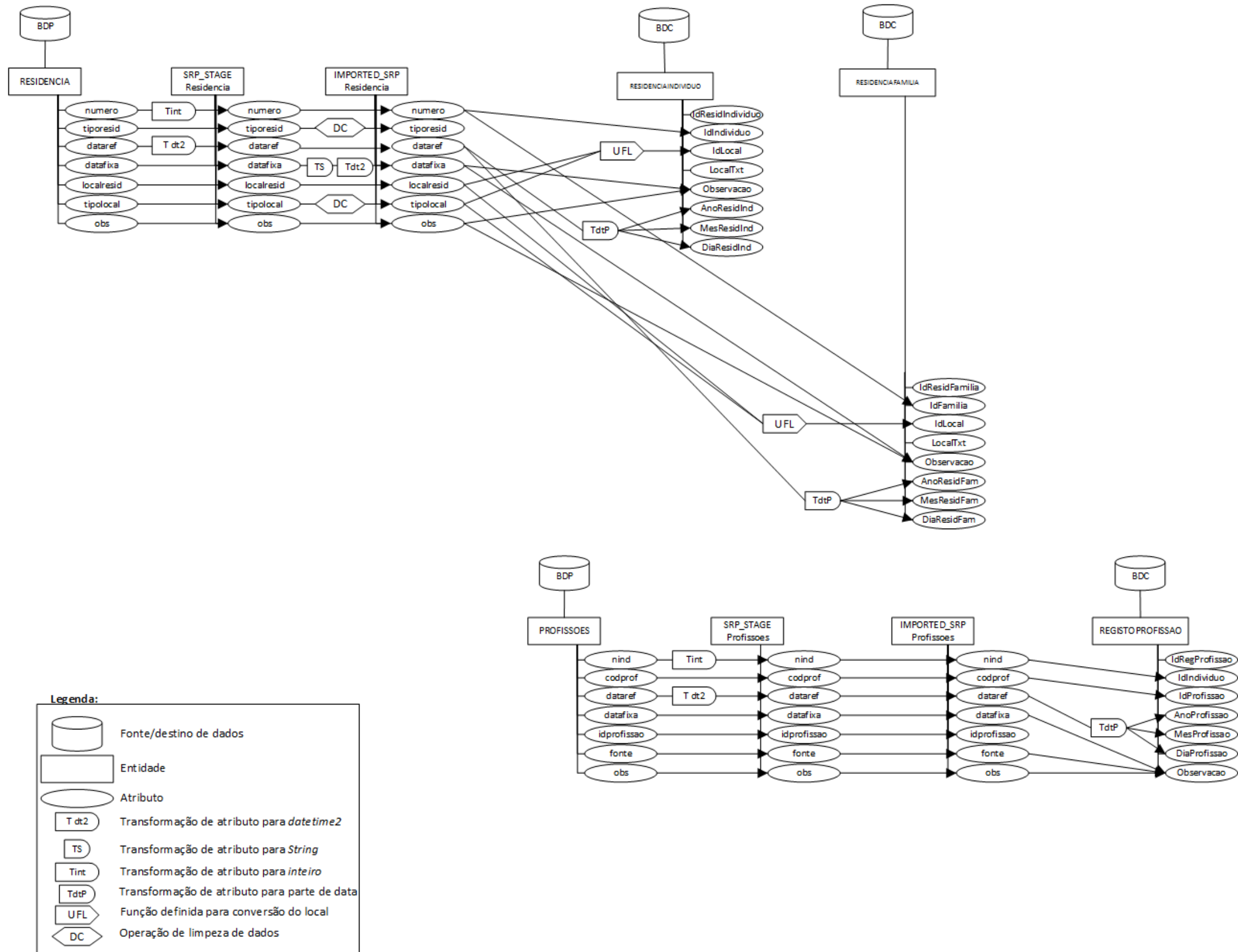
Diagrama de Entidades e Relacionamentos da Base de Dados Central

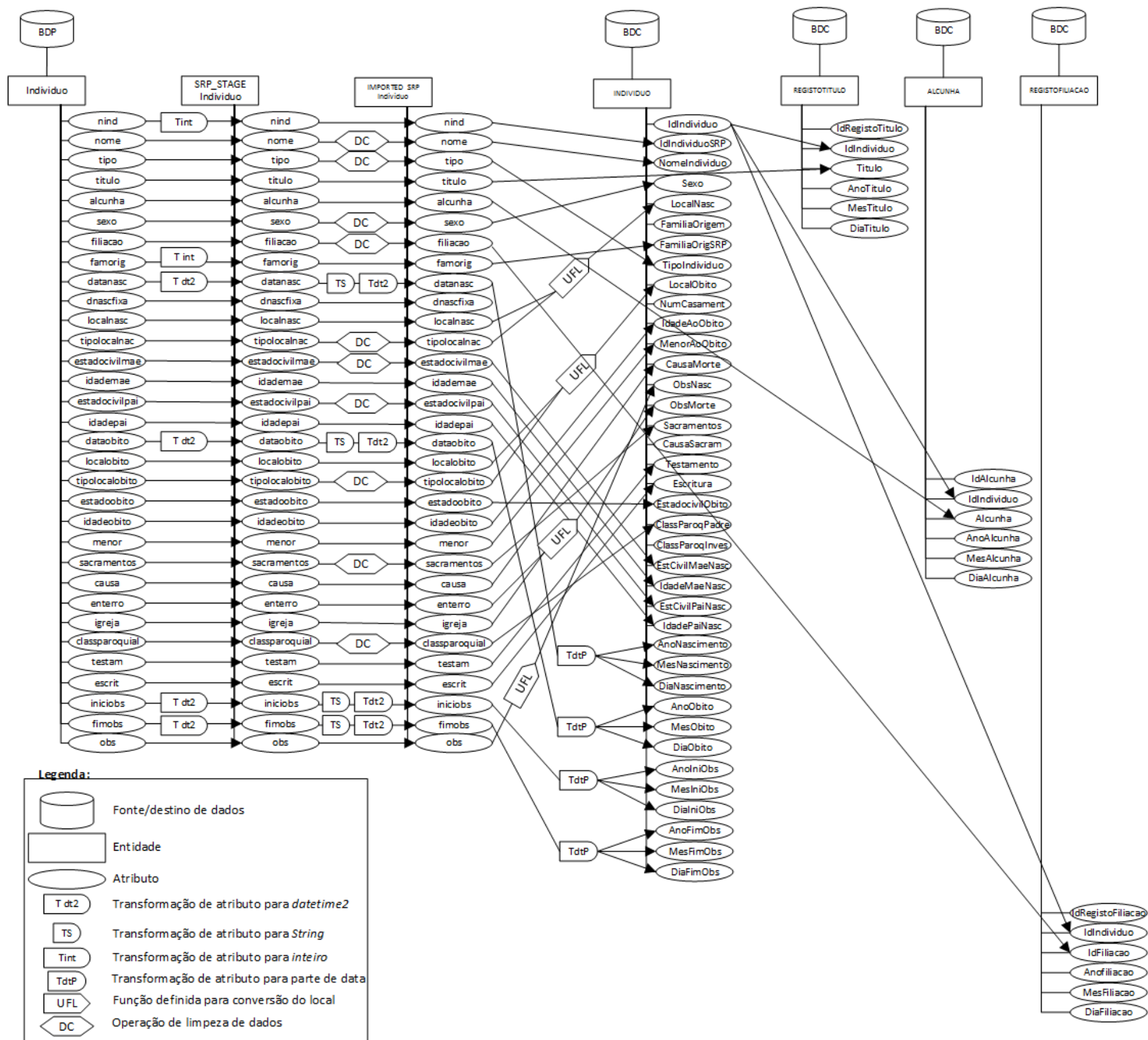






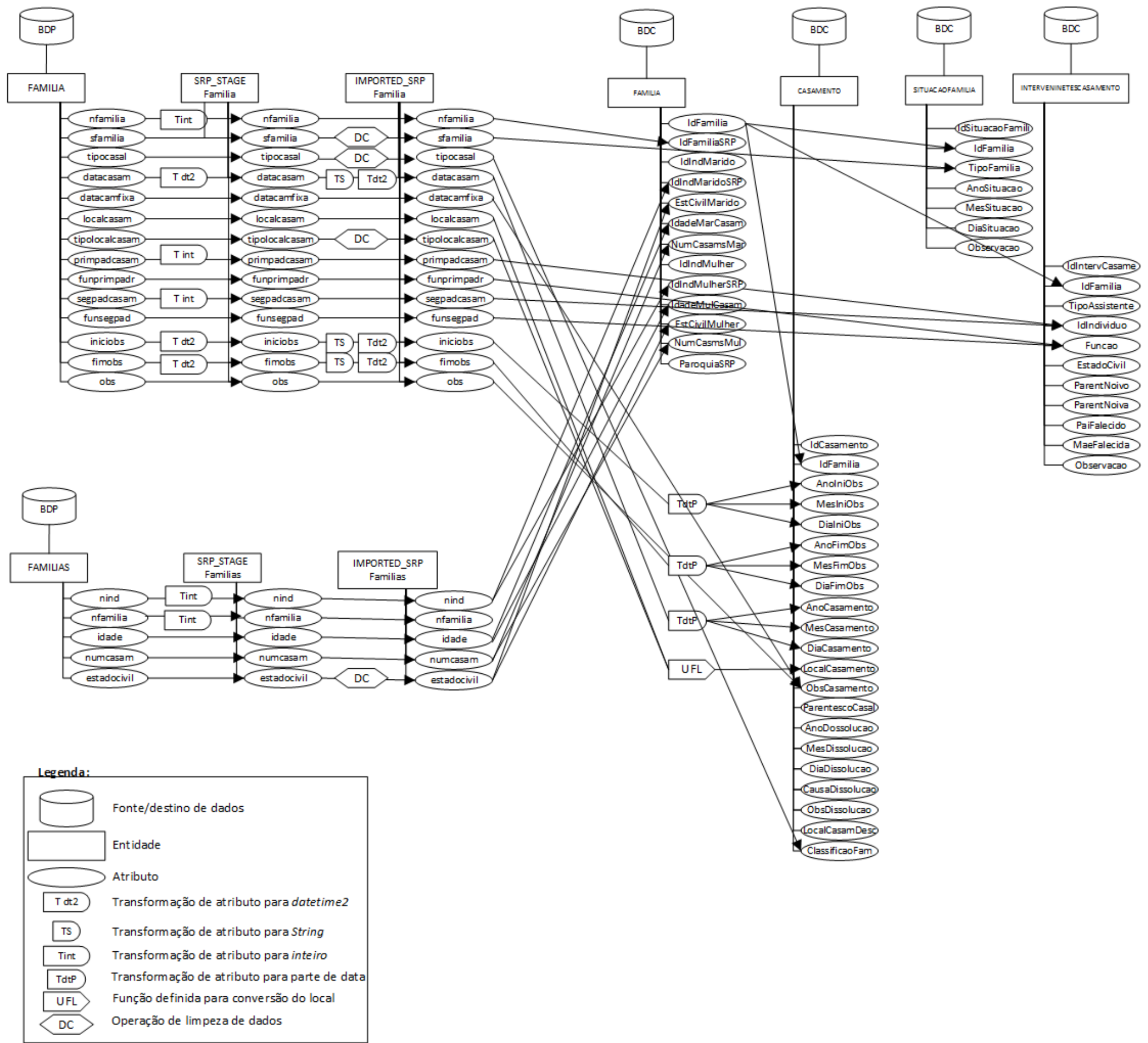
Modelação *Entity Mapping Diagram*

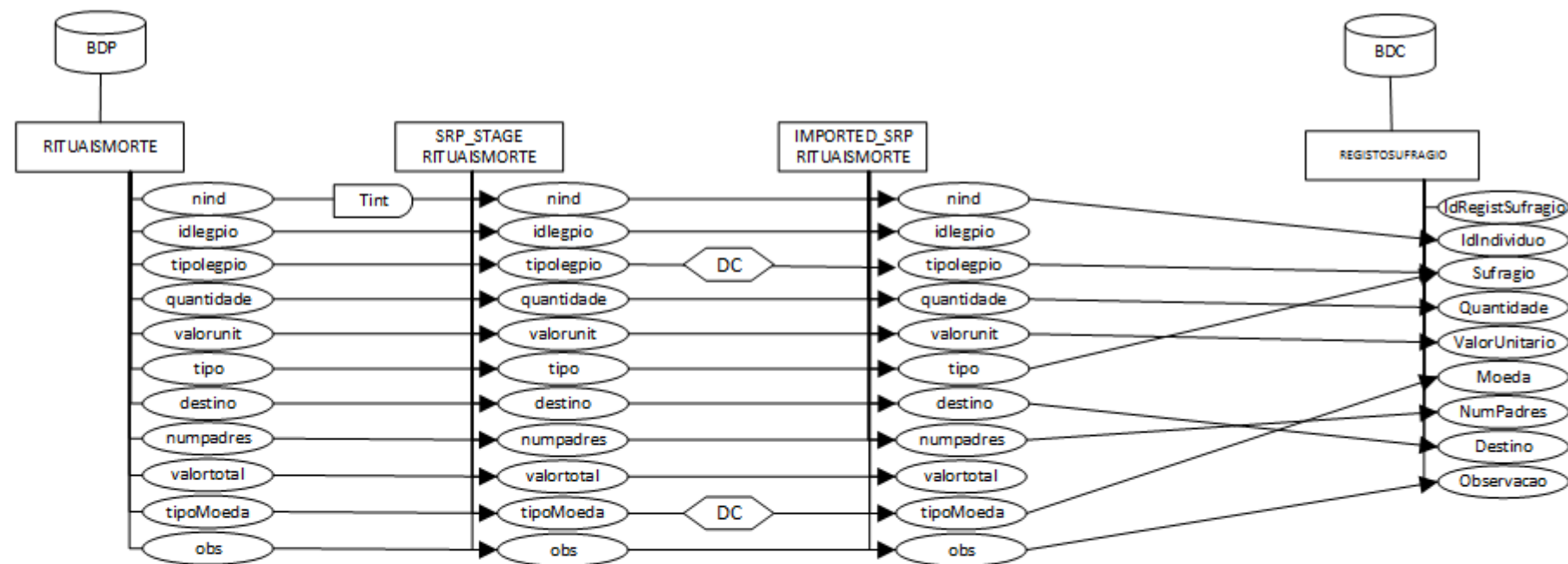
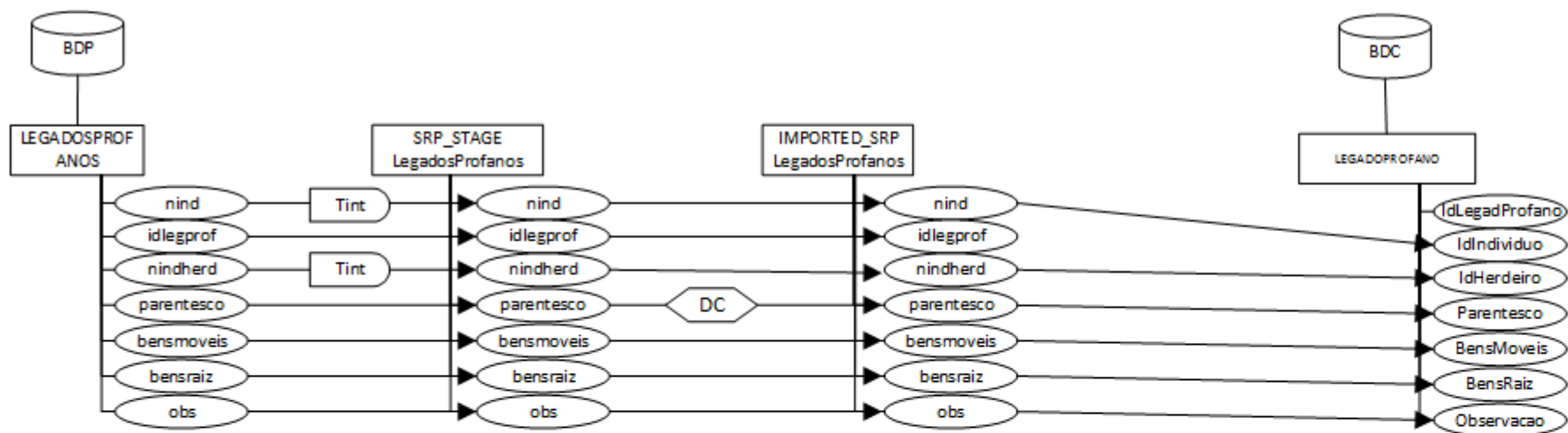




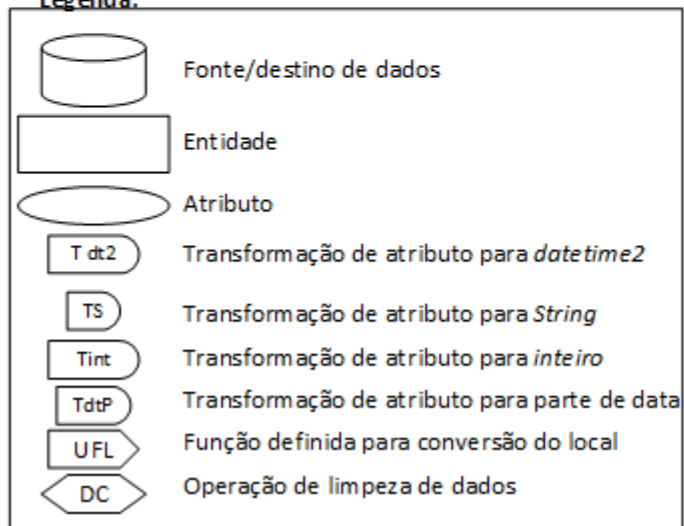
Legenda:

- Fonte/destino de dados
- Entidade
- Atributo
- Transformação de atributo para *datetime2*
- Transformação de atributo para *String*
- Transformação de atributo para *inteiro*
- Transformação de atributo para parte de data
- Função definida para conversão do local
- Operação de limpeza de dados





Legenda:



Primeiros 100 resultados da comparação da função SQL

	NI N D	VA LID O R	NOME	NOMEPAI	NOMEMAE	DA TA NA SC	AN OC ASA M	DATA NASC FILHO	DAT AO BIT O	DN	DP	DM
1	21 60	21 60	Cristóvão de Castro	Jerónimo de Castro	Ana Gonçalves	169 3- 08- 08	172 8	NULL	174 7- 09- 02	1	1	1
1	16 69	16 69	Cristóvão de Castro	Jerónimo de Castro	Ana Gonçalves	169 3- 08- 08	172 8	1729	174 7- 09- 02	1	1	1
2	21 68	21 68	João de Castro Soares	João de Castro	Maria de Barros	168 7- 10- 26	173 0	1732- 08-21	175 5- 05- 25	1	1	1
2	17 65	17 65	João de Castro Soares	João de Castro	Maria de Barros	168 7- 10- 26	NUL L	1739	175 5- 05- 25	1	1	1
3	21 52	21 52	Cristóvão da Costa	João Durães	Catarina da Costa	169 0- 11- 11	172 4	1720- 12-09	174 3- 05- 30	1	1	1
3	17 86	17 86	Cristóvão da Costa	João Durães	Catarina da Costa	169 0- 11- 11	172 4	1720	174 3- 05- 30	1	1	1
4	75 53	75 53	Francisco de Freitas	Manuel de Freitas	Antónia Maria Pereira da Silva	185 8- 12- 18	188 7	NULL	NUL L	1	1	0,9
4	15 81 97 7	15 81 97 7	Francisco Freitas	Manuel de Freitas	Antónia Maria Pereira	NU LL	NUL L	1889	190 0- 12- 23	1	1	0,9
5	41 96	41 96	Custódio José de Castro	João Baptista de Castro	Custódia de Castro Fernandes	177 7- 10- 12	179 8	1799- 07-11	NUL L	1	1	1
5	47 34	47 34	Custódio José de Castro	João Baptista de Castro	Custódia de Castro Fernandes	177 7- 10- 12	NUL L	1799	NUL L	1	1	1
6	45 96	45 96	António José de Castro	João Baptista de Castro	Custódia de Castro Fernandes	177 9- 12- 29	180 9	1810- 08-25	NUL L	1	1	1
6	47 35	47 35	António José de Castro	João Baptista de Castro	Custódia de Castro Fernandes	177 9- 12- 29	NUL L	NULL	NUL L	1	1	1
7	35 04	35 04	Maria Joana de Oliveira	Luís da Silva	Josefa de Oliveira da Silva	177 0- 11- 10	179 6	1796- 12-04	181 9- 09- 10	1	1	1
7	15 85 6	15 85 6	Maria Joana de Oliveira	Luís da Silva	Josefa de Oliveira da Silva	177 0- 11- 10	179 6	1801	181 9- 09- 10	1	1	1

8	21 05	21 05	Luísa Fernandes	António Francisco	Senhorinha Fernandes	171 3- 07- 31	174 5	1745- 03-26	NUL L	1	1	1
8	15 95 0	15 95 0	Luísa Fernandes	António Francisco	Senhorinha Fernandes	NU LL	NUL L	1753	176 3- 10- 26	1	1	1
9	69 7	69 7	Maria de Araújo	Pedro de Araújo	Catarina Velo	166 1- 01- 14	NUL L	NULL	170 4- 07- 26	1	1	1
9	14 42 6	14 42 6	Maria de Araújo	Pedro de Araújo	Catarina Velo	166 1- 01- 14	168 5	1686	170 4- 07- 26	1	1	1
1 0	27 63	27 63	Mariana (Solteira)	António Soares	Catarina da Silva	170 0- 11- 06	NUL L	1741- 10-13	175 2- 12- 17	1	1	1
1 0	11 49	11 49	Mariana	António Soares	Catarina da Silva	170 0- 11- 06	NUL L	NULL	175 2- 12- 17	1	1	1
1 1	25 90	25 90	Maria Soares da Silva	António Soares	Catarina da Silva	171 2- 05- 23	174 0	1741- 03-28	179 2- 08- 25	1	1	1
1 1	11 53	11 53	Maria Soares da Silva	António Soares	Catarina da Silva	171 2- 05- 23	NUL L	NULL	179 2- 08- 25	1	1	1
1 2	28 89	28 89	Joana Soares	António Soares	Catarina da Silva	171 8- 02- 07	174 7	1747- 09-24	180 3- 12- 06	1	1	1
1 2	11 55	11 55	Joana Soares	António Soares	Catarina da Silva	171 8- 02- 07	174 7	1747	180 3- 12- 06	1	1	1
1 3	57 13	57 13	Ana Joaquina da Fonseca	António José da Fonseca	Custódia Maria Soares	181 2- 10- 01	NUL L	1834- 04-17	NUL L	1	1	1
1 3	15 85 83 1	15 85 83 1	Ana Joaquina da Fonseca	António José da Fonseca	Custódia Maria Soares	181 2- 10- 01	NUL L	NULL	NUL L	1	1	1
1 4	42 08	42 08	Manuel José de Freitas	João de Freitas	Mariana de Castro dos Santos	177 3- 08- 07	180 0	NULL	181 3- 06- 15	1	1	1
1 4	47 46	47 46	Manuel José de Freitas	João de Freitas	Mariana de Castro dos Santos	177 3- 08- 07	180 0	1804	181 3- 06- 15	1	1	1
1 5	11 85	11 85	Bento João	António João	Catarina Francisca Fernandes	165 1- 10- 25	168 3	1684- 03-23	172 7- 10- 22	1	1	1
1 5	20 1	20 1	Bento João	António João	Catarina Francisca Fernandes	165 1- 10- 25	NUL L	NULL	172 7-	1	1	1

1 6	15 04	15 04	Francisco João	António João	Catarina Francisca Fernandes	10- 25 165 5- 06- 13	170 0	NULL	10- 22 NUL L	1	1	1
1 6	20 2	20 2	Francisco João	António João	Catarina Francisca Fernandes	165 5- 06- 13	NUL L	NULL	NUL L	1	1	1
1 7	32 10	32 10	Manuel José de Freitas	José Francisco de Freitas	Luísa Maria de Freitas	175 9- 04- 20	178 9	1793- 11-09	NUL L	1	1	1
1 7	15 80 80 4	15 80 80 4	Manuel José de Freitas	José Francisco de Freitas	Luísa Maria de Freitas	175 9- 04- 20	178 9	NULL	NUL L	1	1	1
1 8	18 58	18 58	João	António Francisco	Catarina de Araújo	170 7- 06- 04	NUL L	NULL	NUL L	1	1	1
1 8	24 71	24 71	João	António Francisco	Catarina de Araújo	NU LL	NUL L	NULL	172 6- 12- 21	1	1	1
1 9	38 48	38 48	Francisco José de Freitas	Manuel de Freitas	Maria Teresa	178 1- 05- 19	NUL L	NULL	NUL L	1	1	1
1 9	78 68	78 68	Francisco José de Freitas	Manuel de Freitas	Maria Teresa	178 1- 05- 19	182 6	1827	NUL L	1	1	1
2 0	18 94	18 94	Apolónia Antunes (Solteira)	Domingos Antunes	Catarina Nogueira	168 7- 04- 04	NUL L	1708- 09-05	NUL L	1	1	1
2 0	14 25	14 25	Apolónia Antunes	Domingos Antunes	Catarina Nogueira	168 7- 04- 04	NUL L	1716	NUL L	1	1	1
2 1	26 40	26 40	Maria de Castro (Solteira)	Domingos Francisco	Catarina de Castro	168 9- 06- 09	NUL L	1736- 01-09	NUL L	1	1	1
2 1	14 46	14 46	Maria	Domingos Francisco	Catarina de Castro	168 9- 06- 09	NUL L	NULL	NUL L	1	1	1
2 2	27 51	27 51	Maria Francisca de Castro	Domingos de Castro	Isabel Francisca	NU LL	NUL L	1721- 09-08	NUL L	1	1	1
2 2	14 67	14 67	Maria Francisca de Castro	Domingos de Castro	Isabel Francisca	169 5- 02- 24	171 8	1719	173 3- 03- 05	1	1	1
2 3	27 51	27 51	Maria Francisca de Castro	Domingos de Castro	Isabel Francisca	NU LL	NUL L	1721- 09-08	NUL L	0,9047 61904 76190 5	1	1
2 3	14 69	14 69	Mariana	Domingos de Castro	Isabel Francisca	170 1-	NUL L	1749	NUL L	0,9047 61904	1	1

						03-31				761905		
24	27	27	Maria Francisca de Castro	Domingos de Castro	Isabel Francisca	NU LL	NUL L	1721-09-08	NUL L	0,904761904761905	1	1
24	14	14	Mariana	Domingos de Castro	Isabel Francisca	1708-08-09	NUL L	NULL	NUL L	0,904761904761905	1	1
25	56	56	Maria de Castro	José de Castro Nogueira	Quitéria Maria de Castro de Freitas	1802-03-11	1841	NULL	1845-04-25	1	1	1
25	50	50	Maria de Castro	José de Castro Nogueira	Quitéria Maria de Castro de Freitas	1802-03-11	NUL L	NULL	1845-04-25	1	1	1
26	35	35	Inácio da Rocha	António Barbosa de Castro	Teresa de Castro da Rocha	1740-04-17	1773	NULL	NUL L	1	1	1
26	26	26	Inácio da Rocha	António Barbosa de Castro	Teresa de Castro da Rocha	1740-04-17	1773	1774	NUL L	1	1	1
27	50	50	António Luís	Custódio José de Oliveira	Rosa Maria de Freitas	1817-05-03	NUL L	NULL	NUL L	1	1	1
27	15	15	António José de Oliveira Guimaráes	Custódio José de Oliveira	Rosa Maria de Freitas	NU LL	1860	NULL	1861-07-31	1	1	1
28	57	57	Custódio José Gonçalves	Simão Gonçalves	Maria Ferreira	NU LL	NUL L	1834-11-11	NUL L	1	1	1
28	15	15	Custódio José Gonçalves Ferreira	Simão Gonçalves	Maria Ferreira	NU LL	1826	1830	NUL L	1	1	1
29	32	32	João da Maia	Bento de Freitas	Joana Benta da Maia Baptista	1743-09-17	1763	1765-03-24	1807-11-06	1	1	1
29	26	26	João da Maia	Bento de Freitas	Joana Benta da Maia Baptista	1743-09-17	NUL L	NULL	1807-11-06	1	1	1
30	47	47	Luís António Baptista Guimaráes	João Baptista Pereira	Maria Rosa Palhares	1808-05-17	1836	1837-12-26	NUL L	1	1	1
30	15	15	Luís António Baptista Guimaráes	João Baptista Pereira	Maria Rosa Palhares	1808-05-17	1836	NULL	NUL L	1	1	1
31	38	38	Manuel António de Freitas	José António de Freitas de Oliveira	Luísa Maria Fernandes de Oliveira	1782-07-14	1804	1805-04-01	1833-04-14	1	1	1

3	15	15	Manuel	José	Luísa Maria	178	180	NULL	183	1	1	1
1	81	81	António de	António de	Fernandes de	2-	4		3-			
	01	01	Freitas	Freitas de	Oliveira	07-			04-			
	1	1		Oliveira		14			14			
3	40	40	Maria	José de	Joana de	178	NUL	NULL	NUL	1	1	1
2	85	85		Oliveira	Freitas	8-	L		L			
						09-						
						26						
3	50	50	Maria de	José de	Joana de	NU	NUL	1811	NUL	1	1	1
2	60	60	Oliveira	Oliveira	Freitas	LL	L		L			
3	44	44	Luísa Teresa	Manuel	Maria de	NU	180	1803-	183	1	1	1
3	76	76	de Castro	Monteiro	Castro	LL	1	05-30	1-			
			Monteiro						09-			
									30			
3	53	53	Luísa Teresa	Manuel	Maria de	177	NUL	NULL	NUL	1	1	1
3	15	15		Monteiro	Castro	6-	L		L			
						01-						
						09						
3	39	39	Custódia	Tomé	Luísa da Silva	176	178	1787-	182	1	1	1
4	93	93	Francisca da	António de	Borges	8-	7	04-21	7-			
			Silva	Passos		08-			04-			
						12			12			
3	54	54	Custódia	Tomé	Luísa da Silva	176	NUL	NULL	182	1	1	1
4	95	95	Francisca da	António de	Borges	8-	L		7-			
			Silva	Passos		08-			04-			
						12			12			
3	61	61	Antónia	António	Maria Rosa de	184	NUL	NULL	NUL	1	1	1
5	44	44		José Pinto	Freitas	6-	L		L			
						11-						
						26						
3	58	58	Antónia de	António	Maria Rosa de	NU	186	1868	NUL	1	1	1
5	57	57	Freitas	José Pinto	Freitas	LL	7		L			
3	60	60	Teresa	António	Maria Rosa de	184	NUL	NULL	NUL	1	1	1
6	64	64		José Pinto	Freitas	4-	L		L			
						11-						
						06						
3	58	58	Teresa Rosa	António	Maria Rosa de	NU	187	1875	NUL	1	1	1
6	58	58	de Freitas	José Pinto	Freitas	LL	3		L			
3	68	68	Clementina	Custódio	Maria de	183	NUL	1864-	NUL	1	1	1
7	61	61	Soares	Soares	Castro	6-	L	03-19	L			
			(Solteira)			05-						
						28						
3	63	63	Clementina	Custódio	Maria de	183	NUL	NULL	NUL	1	1	1
7	18	18	Soares	Soares	Castro	6-	L		L			
						05-						
						28						
3	37	37	Ana Josefa	Cristóvão	Luísa Maria de	176	177	1779-	NUL	1	0,8845	1
8	88	88	da Costa	da Costa	Freitas	0-	6	04-07	L		31590	
						12-					41394	
						23					3	
3	42	42	Ana Josefa	Cristóvão	Luísa Maria de	176	177	NULL	NUL	1	0,8845	1
8	94	94	da Costa	da Costa	Freitas	0-	6		L		31590	
						12-					41394	
						23					3	
3	40	40	Manuel José	José de	Quitéria Maria	178	NUL	NULL	NUL	1	1	1
9	87	87	de Castro	Castro	de Castro de	8-	L		L			
				Nogueira	Freitas	10-						
						30						
3	15	15	Manuel José	José de	Quitéria Maria	178	181	NULL	NUL	1	1	1
9	81	81	de Castro	Castro	de Castro de	8-	1		L			
	01	01		Nogueira	Freitas	10-						
	6	6				30						

4 0	65 69	65 69	João	José da Costa	Antónia Maria de Freitas	185 9- 07- 02	NUL L	NULL	NUL L	1	1	1
4 0	15 82 08 3	15 82 08 3	João Costa	José da Costa	Antónia Maria de Freitas	NU LL	188 8	1890	NUL L	1	1	1
4 1	63 43	63 43	Marcelino da Costa Nobre	Miguel Joaquim da Costa Nobre	Clara Maria da Silva	185 1- 09- 28	188 1	NULL	NUL L	1	1	1
4 1	15 82 08 6	15 82 08 6	Marcelino da Costa Nobre	Miguel Joaquim da Costa Nobre	Clara Maria da Silva	185 1- 09- 28	188 1	1883	NUL L	1	1	1
4 2	36 06	36 06	Francisco José de Castro	Domingos Peixoto de Castro	Rosa Maria Mendes de Freitas	175 0- 09- 26	177 9	1780- 04-22	NUL L	1	1	1
4 2	28 59	28 59	Francisco José de Castro	Domingos Peixoto de Castro	Rosa Maria Mendes de Freitas	175 0- 09- 26	NUL L	NULL	NUL L	1	1	1
4 3	70 07	70 07	António Joaquim Ferreira	Manuel Ferreira	Teresa Fernandes	186 8- 01- 03	188 7	1888- 07-30	194 4- 10- 12	1	1	1
4 3	15 82 23 9	15 82 23 9	António Ferreira	Manuel Ferreira	Teresa Fernandes	NU LL	NUL L	1903	NUL L	1	1	1
4 4	34 8	34 8	Gonçalo de Freitas	Jacinto Salgado	Isabel de Freitas	164 8- 03- 15	NUL L	NULL	173 4- 07- 09	1	1	1
4 4	15 89 82 8	15 89 82 8	Gonçalo de Freitas	Jacinto Salgado	Isabel de Freitas	NU LL	167 0	1670	173 4- 07- 09	1	1	1
4 5	37 44	37 44	Maria Teresa Leite	Francisco Soares de Castro	Maria Leite de Sampaio	176 2- 04- 05	178 3	1785- 06-12	183 8- 06- 12	1	1	1
4 5	44 45	44 45	Maria Teresa Leite	Francisco Soares de Castro	Maria Leite de Sampaio	176 2- 04- 05	NUL L	NULL	183 8- 06- 12	1	1	1
4 6	38 15	38 15	Maria Teresa	Francisco Gomes	Custódia Maria	178 0- 02- 15	NUL L	NULL	NUL L	1	1	1
4 6	45 39	45 39	Maria Teresa Gomes	Francisco Gomes	Custódia Maria	NU LL	180 4	1806	NUL L	1	1	1
4 7	23 82	23 82	Maria Lopes	António de Sousa	Mariana Lopes	172 4- 12- 11	NUL L	NULL	NUL L	1	1	1
4 7	14 77 9	14 77 9	Maria Lopes	António de Sousa	Mariana Lopes	172 4- 12- 11	NUL L	1746	NUL L	1	1	1

48	2905	2905	Custódia Francisca de Castro	Silvestre Lopes Pinheiro	Ana de Castro	1725-01-26	1749	1751-04-02	NUL L	1	1	1
48	3410	3410	Custódia Francisca de Castro	Silvestre Lopes Pinheiro	Ana de Castro	1725-01-26	NUL L	NULL	NUL L	1	1	1
49	7269	7269	Albina	Joaquim de Freitas	Josefa Leite	1873-12-13	NUL L	NULL	1944-12-24	1	1	1
49	1581485	1581485	Albina de Freitas	Joaquim de Freitas	Josefa Leite	NU LL	1893	1894	NUL L	1	1	1
50	7988	7988	Florinda da Costa	José da Costa	Maria de Castro Borges	NU LL	NUL L	1890-12-07	NUL L	1	1	0,8939394
50	1581497	1581497	Florinda da Costa	José da Costa	Maria de Castro	NU LL	NUL L	1897	NUL L	1	1	0,8939394
51	4677	4677	Luísa da Silva	João Baptista de Castro da Silva	Maria Joaquina da Silva	1805-09-05	NUL L	NULL	NUL L	1	1	1
51	1583025	1583025	Luísa Baptista da Silva	João Baptista de Castro da Silva	Maria Joaquina da Silva	NU LL	NUL L	1833	NUL L	1	1	1
52	6956	6956	Emília Peixoto	Nicolau Peixoto	Luísa Lopes	1866-09-13	1892	NULL	1901-06-04	1	1	0,8787879
52	1581582	1581582	Emília Peixoto	Nicolau Peixoto	Luzia Lopes	1866-09-13	1892	1893	1901-06-04	1	1	0,8787879
53	7220	7220	Emília	Joaquim de Freitas	Josefa Leite	1872-04-11	NUL L	NULL	1960-02-14	1	1	1
53	1581624	1581624	Emília Freitas	Joaquim de Freitas	Josefa Leite	NU LL	1894	1895	NUL L	1	1	1
54	6977	6977	Maria Pereira	Luís Pereira	Maria Baptista	1867-02-17	1893	1894-06-27	NUL L	1	1	1
54	1581648	1581648	Maria Pereira	Luís Pereira	Maria Baptista	1867-02-17	1893	1900	NUL L	1	1	1
55	6192	6192	Maria	João Cardoso de Moura	Angélica Rosa Cardoso	1848-03-15	NUL L	NULL	NUL L	1	1	1
55	7200	7200	Maria José Cardoso de Moura	João Cardoso de Moura	Angélica Rosa Cardoso	NU LL	NUL L	1875	NUL L	1	1	1

5 6	62 86	62 86	Maria	João Cardoso de Moura	Angélica Rosa Cardoso	184 9- 10- 25	NUL L	NULL	NUL L	1	1	1
5 6	72 00	72 00	Maria José Cardoso de Moura	João Cardoso de Moura	Angélica Rosa Cardoso	NU LL	NUL L	1875	NUL L	1	1	1
5 7	63 51	63 51	Josefina	João Cardoso de Moura	Angélica Rosa Cardoso	185 2- 02- 23	NUL L	NULL	NUL L	1	1	1
5 7	72 03	72 03	Josefina Cardoso Moura	João Cardoso de Moura	Angélica Rosa Cardoso	NU LL	187 4	1875	NUL L	1	1	1
5 8	64 05	64 05	Emília	João Cardoso de Moura	Angélica Rosa Cardoso	185 3- 08- 21	NUL L	NULL	NUL L	1	1	1
5 8	72 04	72 04	Emília Rosa Cardoso Moura	João Cardoso de Moura	Angélica Rosa Cardoso	NU LL	187 6	1876	NUL L	1	1	1
5 9	54 38	54 38	Custódia	Manuel José Baptista	Maria Rosa Ribeiro de Freitas	182 8- 12- 14	NUL L	NULL	NUL L	1	1	0,8 735 632
5 9	17 34 6	17 34 6	Custódia Baptista Freitas	Manuel José Baptista	Maria Rosa Ribeiro	NU LL	186 1	1866	189 7- 09- 04	1	1	0,8 735 632
6 0	56 50	56 50	Joaquina Rosa da Silva Freitas	José de Freitas	Maria Rosa da Silva	183 1- 08- 06	NUL L	NULL	190 3- 01- 21	1	1	1
6 0	17 44 4	17 44 4	Joaquina Rosa da Silva Freitas	José de Freitas	Maria Rosa da Silva	183 1- 08- 06	185 8	1859	NUL L	1	1	1
6 1	49 15	49 15	Joaquina de Oliveira	Miguel José de Oliveira	Teresa Maria da Costa	180 2- 10- 08	182 0	NULL	NUL L	1	1	1
6 1	79 80	79 80	Joaquina de Oliveira	Miguel José de Oliveira	Teresa Maria da Costa	180 2- 10- 08	NUL L	NULL	NUL L	1	1	1
6 2	56 81	56 81	Rosa Mendes de Freitas	Miguel José de Oliveira	Teresa Maria da Costa	180 9- 02- 21	NUL L	1833- 01-23	NUL L	1	1	1
6 2	79 83	79 83	Rosa Mendes de Freitas	Miguel José de Oliveira	Teresa Maria da Costa	180 9- 02- 21	NUL L	NULL	NUL L	1	1	1
6 3	67 92	67 92	Antónia da Rocha	José Neto da Rocha	Luísa Ribeiro	NU LL	186 0	1862- 06-06	NUL L	1	1	1
6 3	17 55 1	17 55 1	Antónia da Rocha	José Neto da Rocha	Luísa Ribeiro	NU LL	186 0	1864	NUL L	1	1	1
6 4	26 22	26 22	Custódia Francisca	Bento Francisco	Mariana Francisca	173 4- 12- 09	NUL L	NULL	NUL L	1	1	1

6	26	26	Custódia	Bento	Mariana	173	NUL	NULL	NUL	1	1	1
4	22	22	Francisca	Francisco	Francisca	4-12-09	L		L			
6	15	15	Custódio	Francisco	Bernarda	NU	NUL	NULL	NUL	1	1	1
5	89	89	José de	José de	Maria de	LL	L		L			
	94	94	Oliveira	Oliveira	Oliveira							
	8	8										
6	38	38	Custódio	Francisco	Bernarda	178	180	1804-	184	1	1	1
5	72	72	José de	José de	Maria de	2-	3	08-29	1-			
			Oliveira	Oliveira	Oliveira	04-21			11-18			
6	15	15	Joaquim de	João de	Bernardina da	183	186	1871	NUL	1	1	1
6	81	81	Castro	Castro	Silva Nogueira	9-	6		L			
	34	34				03-						
	9	9				01						
6	58	58	Joaquim de	João de	Bernardina da	183	186	1871-	NUL	1	1	1
6	67	67	Castro	Castro	Silva Nogueira	9-	6	02-02	L			
						03-						
						01						
6	15	15	Miguel	Luís	Maria de	NU	187	1875	NUL	1	1	1
7	81	81	Baptista	António	Freitas	LL	4		L			
	43	43		Baptista								
	3	3		Guimarães								
6	64	64	Miguel	Luís	Maria de	185	NUL	NULL	NUL	1	1	1
7	56	56		António	Freitas	5-	L		L			
				Baptista		10-						
				Guimarães		23						
6	15	15	Manuel de	José de	Maria Rosa da	182	NUL	NULL	NUL	1	1	1
8	89	89	Freitas	Freitas	Silva	7-	L		L			
	11	11				09-						
	6	6				05						
6	53	53	Manuel de	José de	Maria Rosa da	182	185	1857-	NUL	1	1	1
8	90	90	Freitas	Freitas	Silva	7-	5	04-18	L			
						09-						
						05						
6	10	10	José de	António de	Luísa de Freitas	188	NUL	NULL	193	1	1	1
9	36	36	Oliveira	Oliveira		5-	L		4-			
	7	7				11-			01-			
						23			16			
6	86	86	José de	António de	Luísa de Freitas	188	190	NULL	193	1	1	1
9	19	19	Oliveira	Oliveira		5-	8		4-			
						11-			01-			
						23			16			
7	17	17	Francisco de	José de	Maria Rosa da	183	186	1866	NUL	1	1	1
0	41	41	Freitas	Freitas	Silva	9-	1		L			
	5	5				03-						
						26						
7	58	58	Francisco de	José de	Maria Rosa da	183	189	NULL	NUL	1	1	1
0	70	70	Freitas	Freitas	Silva	9-	7		L			
						03-						
						26						
7	15	15	Luísa	Francisco	Custódia	173	NUL	1760	NUL	1	1	0,9
1	83	83		Gomes	Francisca	5-	L		L			009
	71	71			Peixoto	02-						009
	8	8				09						
7	25	25	Luísa	Francisco	Custódia	173	NUL	NULL	NUL	1	1	0,9
1	63	63		Gomes	Francisca	5-	L		L			009
					Peixoto de	02-						009
					Freitas	09						
7	70	70	Maria	António da	Rosa Maria da	186	188	1888-	194	1	0,8985	1
2	20	20	Joaquina da	Silva	Silva	8-	7	08-16	2-		50724	
			Silva									

						03-20			04-14		63768	
7	11	11	Maria	António da	Rosa Maria da	NU	NUL	NULL	NUL	1	0,8985	1
2	51	51		Silva Andrés	Silva	LL	L		L		50724	
	7	7									63768	
											1	
7	45	45	Maria	João	Maria Joaquina	180	NUL	NULL	NUL	1	1	1
3	38	38		Baptista de	da Silva	1-	L		L			
				Castro da		11-						
				Silva		28						
7	73	73	Maria de	João	Maria Joaquina	NU	NUL	1825	NUL	1	1	1
3	93	93	Castro	Baptista de	da Silva	LL	L		L			
				Castro da								
				Silva								
7	71	71	Emília	Luís Pereira	Maria Baptista	186	NUL	NULL	194	1	1	1
4	38	38	Pereira			9-	L		0-			
						09-			02-			
						18			18			
7	15	15	Emília	Luís Pereira	Maria Baptista	186	189	1898	194	1	1	1
4	81	81	Pereira			9-	7		0-			
	77	77				09-			02-			
	5	5				18			18			
7	27	27	Catarina	Jacinto	Mariana	174	177	1768-	180	1	1	1
5	86	86	Francisca de	Ribeiro	Francisca de	2-	0	07-29	6-			
			Freitas		Castro	07-			08-			
						06			19			
7	16	16	Catarina	Jacinto	Mariana	174	177	1770	180	1	1	1
5	00	00	Francisca de	Ribeiro	Francisca de	2-	0		6-			
	1	1	Freitas		Castro	07-			08-			
						06			19			
7	61	61	Rosa de	Manuel	Custódia Maria	182	184	1847-	NUL	1	1	1
6	56	56	Freitas	José de	de Freitas	5-	9	03-29	L			
			Coelho	Castro (O	Coelho	07-						
				Coelho		15						
7	76	76	Rosa de	Manuel	Custódia Maria	182	NUL	NULL	NUL	1	1	1
6	18	18	Freitas	José de	de Freitas	5-	L		L			
				Castro (O	Coelho	07-						
				Coelho		15						
7	41	41	Maria Joana	António	Joana Luísa de	NU	179	1798-	NUL	1	1	1
7	88	88	Peixoto	José	Castro	LL	7	08-10	L			
				Peixoto	Fernandes							
7	16	16	Maria Joana	António	Joana Luísa de	NU	179	1805	NUL	1	1	1
7	47	47		José	Castro	LL	7		L			
	1	1		Peixoto	Fernandes							
7	29	29	Custódia	António	Ana de Freitas	174	176	1769-	182	1	1	1
8	25	25	Maria de	Francisco		7-	6	04-20	2-			
			Freitas			03-			08-			
						16			05			
7	16	16	Custódia	António	Ana de Freitas	174	176	1772	182	1	1	1
8	05	05	Maria de	Francisco		7-	6		2-			
	2	2	Freitas			03-			08-			
						16			05			
7	36	36	Custódia	Manuel	Catarina	177	NUL	NULL	180	1	1	1
9	80	80	Maria de	António de	Francisca da	6-	L		6-			
			Castro	Castro	Costa	02-			11-			
						24			21			
7	16	16	Custódia	Manuel	Catarina	177	180	1802	180	1	1	1
9	07	07	Maria de	António de	Francisca da	6-	0		6-			
	9	9	Castro	Castro	Costa	02-			11-			
						24			21			
8	53	53	Maria	Nicolau	Antónia Maria	179	182	1827-	186	1	1	0,8
0	98	98	Joaquina	António	Soares	6-	3	10-03	8-			833
			Soares	Correia								333

						04-17			01-18				
8	16	16	Maria	Nicolau	Antónia Maria	179	182	1826	186	1	1	0,8	
0	49	49	Joaquina	António		6-	3		8-			833	
	3	3	Soares	Correia		04-17			01-18			333	
8	52	52	Joaquina	António	Maria de	182	NUL	NULL	185	1	1	1	
1	03	03	Rosa de Castro	José de Castro	Freitas Fernandes	2-01-30	L		6-01-16				
8	16	16	Joaquina	António	Maria de	182	185	1851	185	1	1	1	
1	54	54	Rosa Castro	José de Castro	Freitas Fernandes	2-01-30	0		6-01-16				
	6	6											
8	62	62	Rosa da Costa	João da Costa	Clara Maria Fernandes	184	188	1886-	NUL	1	1	1	
2	16	16				8-12-02	5	08-01	L				
8	15	15	Rosa da Costa	João da Costa	Clara Maria Fernandes	184	188	NULL	NUL	1	1	1	
2	84	84				8-12-02	5		L				
	32	32											
	7	7											
8	55	55	Rosa Maria Lopes	João Lopes	Rosa Gonçalves Martins Carvalho	179	182	1831-	NUL	1	1	1	
3	05	05				4-02-11	1	04-03	L				
8	16	16	Rosa Maria Lopes	João Lopes	Rosa Gonçalves Martins Carvalho	179	182	1829	NUL	1	1	1	
3	62	62				4-02-11	1		L				
	6	6											
8	69	69	Rosa Maria	António Fernandes	Albina Rosa Fernandes	186	NUL	NULL	NUL	1	1	1	
4	10	10				5-10-29	L		L				
8	15	15	Rosa Fernandes	António Fernandes	Albina Rosa Fernandes	NU	188	NULL	NUL	1	1	1	
4	84	84				LL	5		L				
	44	44											
	0	0											
8	70	70	Balbina Rosa	António Fernandes	Joaquina Rosa Ribeiro	186	NUL	NULL	NUL	1	1	1	
5	18	18				8-03-16	L		L				
8	15	15	Balbina Rosa da Silva	António Fernandes	Joaquina Rosa Ribeiro	NU	188	NULL	NUL	1	1	1	
5	84	84				LL	7		L				
	45	45											
	5	5											
8	64	64	Francelina Rosa de Castro	José de Castro	Rosa de Freitas	185	187	NULL	NUL	1		0,8787	1
6	04	04				3-06-21	7		L			87878	
												78787	
												9	
8	15	15	Francelina Rosa de Castro	José de Castro	Rosa de Freitas	NU	188	NULL	NUL	1		0,8787	1
6	84	84				LL	7		L			87878	
	45	45										78787	
	8	8										9	
8	43	43	Antónia	António José Lopes	Quitéria Maria da Costa Fernandes de Sousa	179	NUL	NULL	NUL	1	1	0,9	
7	79	79				7-10-15	L		L			285	
												714	
8	16	16	Antónia Maria Fernandes Lopes	António José Lopes	Quitéria Maria da Costa Fernandes	179	183	1833	186	1	1	0,9	
7	76	76				7-10-15	0		4-04-11			285	
	3	3										714	

8	27	27	Joana de Castro	João de Castro Soares	Maria Francisca de Freitas	173	177	NULL	181	1	1	1
8	23	23				9-06-23	3		2-11-07			
8	16	16	Joana de Castro	João de Castro Soares	Maria Francisca de Freitas	173	177	1774	181	1	1	1
8	33	33				9-06-23	3		2-11-07			
8	41	41	Joaquina de Castro de Freitas	António José de Freitas de Oliveira	Maria Josefa de Castro	179	181	1812-08-23	183	1	1	1
9	34	34				0-04-15	0		7-12-26			
8	16	16	Joaquina de Castro de Freitas	António José de Freitas de Oliveira	Maria Josefa de Castro	179	NUL	1812	183	1	1	1
9	82	82				0-04-15	L		7-12-26			
9	38	38	Ana Maria de Freitas	António José de Freitas de Oliveira	Maria Josefa de Castro	178	180	NULL	NUL	1	1	1
0	33	33				0-07-28	5		L			
9	17	17	Ana Maria de Freitas	António José de Freitas de Oliveira	Maria Josefa de Castro	178	180	1809	NUL	1	1	1
0	00	00				0-07-28	5		L			
1	1	1										
9	60	60	Angélica Rosa Cardoso	João Rodrigues Cardoso	Antónia Maria	181	184	1845-11-17	188	1	1	1
1	99	99				8-11-19	0		4-07-12			
9	17	17	Angélica Rosa Cardoso	João Rodrigues Cardoso	Antónia Maria	181	184	1858	188	1	1	1
1	00	00				8-11-19	0		4-07-12			
9	37	37	Clara Maria	António de Oliveira	Joana Maria de Abreu de Oliveira	177	181	NULL	185	1	1	1
2	78	78				8-11-14	3		0-05-08			
9	17	17	Clara Maria	António de Oliveira	Joana Maria de Abreu de Oliveira	177	181	1815	185	1	1	1
2	01	01				8-11-14	3		0-05-08			
8	8	8										
9	51	51	Joaquina Maria Coelho	João Pereira	Maria Luísa Coelho	NU	182	1825-05-08	NUL	1	1	0,8
3	38	38				LL	2		L			703
9	17	17	Joaquina Maria Pereira	João Pereira	Maria Luísa	NU	NUL	1822	187	1	1	0,8
3	04	04				LL	L		9-05-18			703
1	1	1										704
9	39	39	Luísa Maria Lopes da Silva	João Lopes de Sousa	Custódia da Silva Pereira	178	NUL	NULL	183	1	1	1
4	28	28				4-10-02	L		4-04-07			
9	17	17	Luísa Maria Lopes da Silva	João Lopes de Sousa	Custódia da Silva Pereira	178	180	1806	183	1	1	1
4	05	05				4-10-02	5		4-04-07			
1	1	1										
9	31	31	Custódio de Freitas	Custódio de Freitas	Mariana Francisca	175	177	1779-04-07	180	1	1	1
5	04	04				4-07-18	6		9-12-03			
9	17	17	Custódio de Freitas	Custódio de Freitas	Mariana Francisca	175	177	NULL	180	1	1	1
5	62	62				4-07-18	6		9-12-03			
3	3	3										

96	65	65	José	Nicolau Peixoto	Luísa Lopes	185	NUL	NULL	193	1	1	0,8
	72	72				9-07-20	L		7-04-28			787
												879
96	15	15	José Peixoto	Nicolau Peixoto	Luzia Lopes	NU	189	1892	NUL	1	1	0,8
	81	81				LL	1		L			787
	71	71										879
	7	7										
97	71	71	José	António Fernandes	Maria Baptista	186	NUL	NULL	193	1	1	1
	31	31				9-08-04	L		5-06-12			
97	15	15	José Fernandes	António Fernandes	Maria Baptista	NU	189	1893	NUL	1	1	1
	81	81				LL	2		L			
	72	72										
	6	6										
98	56	56	Francisco José de Castro	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas	181	184	1843-10-21	NUL	1	1	1
	12	12				8-02-20	0		L			
98	56	56	Francisco José de Castro	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas	181	184	1843-10-21	NUL	1	1	1
	12	12				8-02-20	0		L			
99	76	76	José de Castro	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas	182	NUL	NULL	NUL	1	1	1
	17	17				2-03-18	L		L			
99	56	56	José de Castro	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas	182	184	1843-10-29	NUL	1	1	1
	24	24				2-03-18	3		L			
100	38	38	António	João Nogueira	Teresa Maria Fernandes	178	NUL	NULL	NUL	1	1	1
	07	07				0-02-06	L		L			
100	16	16	António José Nogueira	João Nogueira	Teresa Maria Fernandes	NU	NUL	1812	182	1	1	1
	40	40				LL	L		4-09-07			
	6	6										

Primeiros 100 resultados da comparação R- RECORDLINKAGE

P a r	X.U. FEFF .n id	Pr me iro No me	Vál ido	nome	NomePai	NomeMae	We ight	Dat aNa sc	Dat aOb it	Dat a1C asa m	Dat aNa scFil ho1	Class
1	320	Mar ia	FA LS O	Maria	Gonçalo Gonçalves	Maria Gonçalves		164 6- 04- 15				
1	623	Mar ia		Maria Frutuosa Gonçalves	Gonçalo Gonçalves	Maria Gonçalves	1.0 000 000	163 6- 05- 18	170 9- 05- 01	166 7- 08- 19	167 0- 03- 08	L
2	348	Gon çalo	VE R D A	Gonçalo de Freitas	Jacinto Salgado	Isabel de Freitas		164 8- 03- 15	173 4- 07- 09			
2	1589 828	Gon çalo	DE IR O	Gonçalo de Freitas	Jacinto Salgado	Isabel de Freitas	1.0 000 000		173 4- 07- 09	167 0- 04- 19	167 0- 06- 05	L
3	526	Joã o	VE R D A	João de Freitas	António Gonçalves	Maria de Freitas		165 5- 05- 16	172 0- 08- 08	168 8- 02- 01	168 9- 01- 09	
3	1457 3	Joã o	DE IR O	João de Freitas	António Gonçalves	Maria de Freitas	1.0 000 000	165 5- 05- 16	172 0- 08- 08	168 8- 02- 01		L
4	697	Mar ia	VE R D A	Maria de Araújo	Pedro de Araújo	Catarina Velo		166 1- 01- 14	170 4- 07- 26			
4	1442 6	Mar ia	DE IR O	Maria de Araújo	Pedro de Araújo	Catarina Velo	1.0 000 000	166 1- 01- 14	170 4- 07- 26	168 5- 02- 25	168 6- 02- 13	L
5	834	Joã o	FA LS O	João	Pedro Francisco	Maria de Freitas		167 0- 10- 22				
5	1457 5	Joã o		João de Freitas	Pedro Francisco	Maria de Freitas	1.0 000 000	167 2- 08- 25	173 5- 07- 22	169 0- 05- 17		L
6	899	Do min gos	VE R D A	Domingos Ribeiro	Sebastião Gonçalves	Maria Ribeiro				168 1- 11- 05		
6	1446 1	Do min gos	DE IR O	Domingos Ribeiro	Sebastião Gonçalves	Maria Ribeiro	1.0 000 000			168 1- 11- 05	168 2- 11- 26	L
7	932	Mat eus	VE R D A	Mateus Gonçalves	Francisco Gonçalves	Catarina Ribeiro		167 2- 06- 05	170 6- 04- 20			
7	1460 5	Mat eus	DE IR O	Mateus Gonçalves	Francisco Gonçalves	Catarina Ribeiro	1.0 000 000	167 2-	170 6-	168 9-	169 7-	L

								06-05	04-20	07-27	04-09		
8	938	João	VE R D A	João de Freitas	Pedro Francisco	Maria de Freitas		167 2-08-25	173 5-07-22	169 0-05-17	169 1-09-05		
8	14575	João	DE IR O	João de Freitas	Pedro Francisco	Maria de Freitas	1.0 000 000	167 2-08-25	173 5-07-22	169 0-05-17		L	
9	952	Ana	FA LS O	Ana	Francisco Pires	Catarina Francisca		167 3-09-15					
9	401	Ana		Ana Francisca	Francisco Pires	Catarina Francisca	1.0 000 000	163 1-03-30	169 4-11-04	164 3-10-15	165 0-09-21	L	
10	6861	Clementina	VE R D A	Clementina Soares (Solteira)	Custódio Soares	Maria de Castro		183 6-05-28			186 4-03-19		
10	6318	Clementina	DE IR O	Clementina Soares	Custódio Soares	Maria de Castro	1.0 000 000	183 6-05-28				L	
11	6910	Rosa	VE R D A	Rosa Maria	António Fernandes	Albina Rosa Fernandes		186 5-10-29					
11	1584440	Rosa	DE IR O	Rosa Fernandes	António Fernandes	Albina Rosa Fernandes	1.0 000 000			188 5-11-05		L	
12	6941	João	VE R D A	João de Oliveira	Bento José Gonçalves	Teresa Rosa de Oliveira		182 0-04-04	188 8-12-09	185 4-10-15	186 6-04-01		
12	17093	João	DE IR O	João de Oliveira	Bento José Gonçalves	Teresa Rosa de Oliveira	1.0 000 000	182 0-04-04	188 8-12-09	185 4-10-15	185 6-04-28	L	
13	6977	Maria	VE R D A	Maria Pereira	Luís Pereira	Maria Baptista		186 7-02-17		189 3-07-17	189 4-06-27		
13	1581648	Maria	DE IR O	Maria Pereira	Luís Pereira	Maria Baptista	1.0 000 000	186 7-02-17		189 3-07-17	190 0-08-30	L	
14	7000	Maria	VE R D A	Maria de Freitas	Joaquim de Freitas	Josefa Leite		186 7-12-23		188 9-04-10	189 6-02-12		
14	1582069	Maria	DE IR O	Maria de Freitas	Joaquim de Freitas	Josefa Leite	1.0 000 000				189 1-01-18	L	
15	7007	António	VE R D A	António Joaquim Ferreira	Manuel Ferreira	Teresa Fernandes		186 8-01-03	194 4-10-12	188 7-11-19	188 8-07-30		

15	1582239	António	DEIRO	António Ferreira	Manuel Ferreira	Teresa Fernandes	1.000000				1903-09-21	L	
16	7018	Balbina	VERDA	Balbina Rosa	António Fernandes	Joaquina Rosa Ribeiro		1868-03-16					
16	1584455	Balbina	DEIRO	Balbina Rosa da Silva	António Fernandes	Joaquina Rosa Ribeiro	1.000000			1887-03-31		L	
17	7045	Francisco	FALSO	Francisco António	António Dias	Joana de Oliveira		1868-11-15	1959-01-28				
17	1584752	Francisco	FALSO	Francisco António Dias	António Dias	Joana de Oliveira	1.000000		1939-01-28	1898-11-05		L	
18	7051	Rosa	VERDA	Rosa de Freitas (Solteira)	Joaquim de Freitas	Josefa Leite		1869-02-15	1961-05-13		1889-03-10		
18	1582017	Rosa	DEIRO	Rosa de Freitas	Joaquim de Freitas	Josefa Leite	1.000000			1890-07-10	1891-04-11	L	
19	7051	Rosa	VERDA	Rosa de Freitas (Solteira)	Joaquim de Freitas	Josefa Leite		1869-02-15	1961-05-13		1889-03-10		
19	1582536	Rosa	DEIRO	Rosa de Freitas	Joaquim de Freitas	Josefa Leite	1.000000				1908-12-27	L	
20	7131	José	VERDA	José	António Fernandes	Maria Baptista		1869-08-04	1935-06-12				
20	1581726	José	DEIRO	José Fernandes	António Fernandes	Maria Baptista	1.000000			1892-02-14	1893-01-27	L	
21	7138	Emília	VERDA	Emília Pereira	Luís Pereira	Maria Baptista		1869-09-18	1940-02-18				
21	1581775	Emília	DEIRO	Emília Pereira	Luís Pereira	Maria Baptista	1.000000	1869-09-18	1940-02-18	1897-10-03	1898-09-25	L	
22	7220	Emília	VERDA	Emília	Joaquim de Freitas	Josefa Leite		1872-04-11	1960-02-14				
22	1581624	Emília	DEIRO	Emília Freitas	Joaquim de Freitas	Josefa Leite	1.000000			1894-05-13	1895-03-10	L	

23	7238	Florinda	VE R D A	Florinda Gonçalves (Solteira)	José Gonçalves	Rosa da Cunha					187 3- 01- 13		
23	1584 715	Florinda	DE I R O	Florinda Gonçalves	José Gonçalves	Rosa da Cunha	1.0 000 000			187 9- 07- 12		L	
24	7269	Albina	VE R D A	Albina	Joaquim de Freitas	Josefa Leite		187 3- 12- 13	194 4- 12- 24				
24	1581 485	Albina	DE I R O	Albina de Freitas	Joaquim de Freitas	Josefa Leite	1.0 000 000			189 3- 08- 26	189 4- 01- 28	L	
25	7284	Joaquina	VE R D A	Joaquina Pereira	Luís Pereira	Maria Baptista		187 4- 04- 29	195 4- 09- 24				
25	1584 813	Joaquina	DE I R O	Joaquina Pereira	Luís Pereira	Maria Baptista	1.0 000 000	187 4- 04- 29	195 4- 09- 24	190 3- 08- 24		L	
26	7348	Custódio	VE R D A	Custódio Pereira	Luís Pereira	Maria Baptista		187 6- 10- 12					
26	1584 780	Custódio	DE I R O	Custódio Pereira	Luís Pereira	Maria Baptista	1.0 000 000	187 6- 10- 12		190 1- 04- 14		L	
27	7416	Teresa	VE R D A	Teresa de Freitas	Gaspar de Freitas Oliveira	Clara Maria Lopes Peixoto		187 8- 12- 30	197 0- 04- 08	190 5- 12- 16			
27	1582 475	Teresa	DE I R O	Teresa de Freitas	Gaspar de Freitas Oliveira	Clara Maria Lopes Peixoto	1.0 000 000	187 8- 12- 30	197 0- 04- 08	190 5- 12- 16	190 6- 04- 02	L	
28	7488	Amélia	VE R D A	Amélia	Manuel da Silva Maia	Rosalina Ferreira		188 1- 05- 17					
28	1582 315	Amélia	DE I R O	Amélia da Silva Maia	Manuel da Silva Maia	Rosalina Ferreira	1.0 000 000				190 9- 12- 05	L	
29	7676	Helena	VE R D A	Helena de Freitas	Joaquim de Freitas	Josefa Leite		188 3- 10- 18	194 1- 01- 08				
29	1584 807	Helena	DE I R O	Helena de Freitas	Joaquim de Freitas	Josefa Leite	1.0 000 000	188 3- 10- 18	194 1- 01- 08	190 2- 11- 30		L	
30	6064	Teresa	VE R D A	Teresa	António José Pinto	Maria Rosa de Freitas		184 4- 11- 06					

30	5858	Teresa	DEIRO	Teresa Rosa de Freitas	António José Pinto	Maria Rosa de Freitas	1.000000			1873-04-21	1875-12-13	L	
31	6099	Angélica	VERDA	Angélica Rosa Cardoso	João Rodrigues Cardoso	Antónia Maria		1818-11-19	1884-07-17	1840-03-17	1845-11-17		
319	1700	Angélica	DEIRO	Angélica Rosa Cardoso	João Rodrigues Cardoso	Antónia Maria	1.000000	1818-11-19	1884-07-12	1840-03-17	1858-04-25	L	
32	6102	João	VERDA	João Cardoso de Moura	José Cardoso da Silva	Maria Vitória			1893-02-21	1840-03-17	1845-11-17		
328	1700	João	DEIRO	João Cardoso de Moura	José Cardoso da Silva	Maria Vitória	1.000000		1893-02-21	1840-03-17	1858-04-25	L	
33	6144	Antónia	VERDA	Antónia	António José Pinto	Maria Rosa de Freitas		1846-11-26					
33	5857	Antónia	DEIRO	Antónia de Freitas	António José Pinto	Maria Rosa de Freitas	1.000000			1867-02-02	1868-04-05	L	
34	6156	Rosa	VERDA	Rosa de Freitas Coelho	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas Coelho		1825-07-15		1849-04-11	1847-03-29		
34	7618	Rosa	DEIRO	Rosa de Freitas	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas Coelho	1.000000	1825-07-15				L	
35	6192	Maria	VERDA	Maria	João Cardoso de Moura	Angélica Rosa Cardoso		1848-03-15					
35658	1582	Maria	DEIRO	Maria Rosa Cardoso Moura	João Cardoso de Moura	Angélica Rosa Cardoso	1.000000			1873-02-01	1874-01-09	L	
36	6192	Maria	VERDA	Maria	João Cardoso de Moura	Angélica Rosa Cardoso		1848-03-15					
36	7200	Maria	DEIRO	Maria José Cardoso de Moura	João Cardoso de Moura	Angélica Rosa Cardoso	1.000000				1875-11-04	L	
37	6202	Maria	VERDA	Maria de Freitas	Custódio José Joaquim de Oliveira	Custódia de Freitas			1885-08-18		1848-07-22		
37053	1587	Maria	DEIRO	Maria de Freitas	Custódio José Joaquim de Oliveira	Custódia de Freitas	1.000000		1885-08-18			L	
38	6216	Rosa	VERDA	Rosa da Costa	João da Costa	Clara Maria Fernandes		1848-		1885-	1886-		

			DA					12-02		01-07	08-01		
38	1584327	Rosa	DEIRO	Rosa da Costa	João da Costa	Clara Maria Fernandes	1.000000	1848-12-02		1885-01-07		L	
39	6284	Teresa	VERDA	Teresa	Francisco de Oliveira	Joaquina Ferreira		1849-10-13					
39	1582614	Teresa	DEIRO	Teresa de Oliveira	Francisco de Oliveira	Joaquina Ferreira	1.000000		1898-06-11	1881-04-21	1878-07-07	L	
40	6286	Maria	VERDA	Maria	João Cardoso de Moura	Angélica Rosa Cardoso		1849-10-25					
40	1582658	Maria	DEIRO	Maria Rosa Cardoso Moura	João Cardoso de Moura	Angélica Rosa Cardoso	1.000000			1873-02-01	1874-01-09	L	
41	6286	Maria	VERDA	Maria	João Cardoso de Moura	Angélica Rosa Cardoso		1849-10-25					
41	7200	Maria	DEIRO	Maria José Cardoso de Moura	João Cardoso de Moura	Angélica Rosa Cardoso	1.000000				1875-11-04	L	
42	6334	João	FALSO	João	José da Costa	Antónia Maria de Freitas		1851-03-27	1852-08-12				
42	1582083	João	FALSO	João Costa	José da Costa	Antónia Maria de Freitas	1.000000			1888-04-14	1890-01-30	L	
43	6343	Marcelino	VERDA	Marcelino da Costa Nobre	Miguel Joaquim da Costa Nobre	Clara Maria da Silva		1851-09-28		1881-08-14			
43	1582086	Marcelino	DEIRO	Marcelino da Costa Nobre	Miguel Joaquim da Costa Nobre	Clara Maria da Silva	1.000000	1851-09-28		1881-08-14	1883-10-01	L	
44	6351	Josefina	VERDA	Josefina	João Cardoso de Moura	Angélica Rosa Cardoso		1852-02-23					
44	7203	Josefina	DEIRO	Josefina Cardoso Moura	João Cardoso de Moura	Angélica Rosa Cardoso	1.000000			1874-12-18	1875-12-19	L	
45	6405	Emília	VERDA	Emília	João Cardoso de Moura	Angélica Rosa Cardoso		1853-08-21					
45	7204	Emília	DEIRO	Emília Rosa Cardoso Moura	João Cardoso de Moura	Angélica Rosa Cardoso	1.000000			1876-03-16	1876-01-06	L	

46	6425	José	FA LS O	José Fernandes	Joaquim Fernandes	Antónia Maria Pires		185 4- 05- 29		188 9- 04- 10	189 6- 02- 12		
46	1582 068	José		José Fernandes	Joaquim Fernandes	Antónia Maria Pires	1.0 000 000	186 9- 07- 24			189 1- 01- 18	L	
47	6456	Miguel	VE R D A	Miguel	Luís António Baptista Guimarães	Maria de Freitas		185 5- 10- 23					
47	1581 433	Miguel	DE IR O	Miguel Baptista	Luís António Baptista Guimarães	Maria de Freitas	1.0 000 000			187 4- 10- 03	187 5- 07- 28	L	
48	6541	Joaquina	VE R D A	Joaquina Maria	António José Francisco da Costa	Clara Maria		183 3- 03- 30		185 6- 01- 24	185 8- 09- 13		
48	1720 7	Joaquina	DE IR O	Joaquina Maria	António José Francisco da Costa	Clara Maria	1.0 000 000	183 3- 03- 30		185 6- 01- 24	185 7- 08- 18	L	
49	6569	João	VE R D A	João	José da Costa	Antónia Maria de Freitas		185 9- 07- 02					
49	1582 083	João	DE IR O	João Costa	José da Costa	Antónia Maria de Freitas	1.0 000 000			188 8- 04- 14	189 0- 01- 30	L	
50	6792	Antónia	VE R D A	Antónia da Rocha	José Neto da Rocha	Luísa Ribeiro				186 0- 02- 12	186 2- 06- 06		
50	1755 1	Antónia	DE IR O	Antónia da Rocha	José Neto da Rocha	Luísa Ribeiro	1.0 000 000			186 0- 02- 12	186 4- 09- 14	L	
51	4915	Joaquina	VE R D A	Joaquina de Oliveira	Miguel José de Oliveira	Teresa Maria da Costa		180 2- 10- 08		182 0- 05- 11			
51	7980	Joaquina	DE IR O	Joaquina de Oliveira	Miguel José de Oliveira	Teresa Maria da Costa	1.0 000 000	180 2- 10- 08				L	
52	5024	José	VE R D A	José António Ferreira	Manuel Ferreira	Antónia Maria de Figueiredo		178 0- 12- 18		180 0- 12- 04	181 6- 08- 19		
52	1690 9	José	DE IR O	José António Ferreira	Manuel Ferreira	Antónia Maria de Figueiredo	1.0 000 000	178 0- 12- 18		182 0- 07- 25	182 3- 01- 31	L	
53	5044	António	VE R D A	António Luís	Custódio José de Oliveira	Rosa Maria de Freitas		181 7- 05- 03					

53	1580887	António	DEIRO	António José de Oliveira Guimarães	Custódio José de Oliveira	Rosa Maria de Freitas	1.000000		1861-07-31	1860-12-31			L	
54	5203	Joaquina	VERDA	Joaquina Rosa de Castro	António José de Castro	Maria de Freitas Fernandes		1822-01-30	1850-01-16					
54	16546	Joaquina	DEIRO	Joaquina Rosa Castro	António José de Castro	Maria de Freitas Fernandes	1.000000	1822-01-30	1850-01-16	1850-12-26	1851-09-24		L	
55	5390	Manuel	VERDA	Manuel de Freitas	José de Freitas	Maria Rosa da Silva		1827-09-05		1855-10-05	1857-04-18			
55	1589116	Manuel	DEIRO	Manuel de Freitas	José de Freitas	Maria Rosa da Silva	1.000000	1827-09-05					L	
56	5481	João	VERDA	João da Silva	Luís da Silva	Catarina Fernandes				1829-09-06	1830-08-31			
56	17030	João	DEIRO	João Silva	Luís da Silva	Catarina Fernandes	1.000000		1883-03-26		1838-05-23		L	
57	5505	Rosa	VERDA	Rosa Maria Lopes	João Lopes	Rosa Gonçalves Martins Carvalho		1794-02-11		1821-07-29	1831-04-03			
57	16626	Rosa	DEIRO	Rosa Maria Lopes	João Lopes	Rosa Gonçalves Martins Carvalho	1.000000	1794-02-11		1821-07-29	1829-01-30		L	
58	5573	António	VERDA	António José Mendes de Freitas	Francisco Mendes de Freitas	Custódia Maria de Freitas		1782-08-17	1850-08-25	1837-02-02	1838-01-22			
58	4500	António	DEIRO	António José Mendes de Freitas	Francisco Mendes de Freitas	Custódia Maria de Freitas	1.000000	1782-08-17	1850-08-25				L	
59	5612	Francisco	VERDA	Francisco José de Castro	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas Coelho		1818-02-20		1840-01-10	1843-10-21			
59	7616	Francisco	DEIRO	Francisco José de Castro	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas Coelho	1.000000	1818-02-20					L	
60	5620	Maria	FALSO	Maria de Castro	José de Castro Nogueira	Quitéria Maria de Castro de Freitas		1802-03-11	1845-04-25	1841-02-18				
60	5040	Maria		Maria Joana	José de Castro Nogueira	Quitéria Maria de Castro de Freitas	1.000000	1794-07-23					L	

61	5620	Maria	VERDA	Maria de Castro	José de Castro Nogueira	Quitéria Maria de Castro de Freitas		180 2-03-11	184 5-04-25	184 1-02-18			
61	5042	Maria	DEIRO	Maria de Castro	José de Castro Nogueira	Quitéria Maria de Castro de Freitas	1.0000	180 2-03-11	184 5-04-25				L
62	5624	José	VERDA	José de Castro	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas Coelho		182 2-03-18		184 3-01-25	184 3-10-29		
62	7617	José	DEIRO	José de Castro	Manuel José de Castro (O Coelho)	Custódia Maria de Freitas Coelho	1.0000	182 2-03-18					L
63	5650	Joaquina	VERDA	Joaquina Rosa da Silva Freitas	José de Freitas	Maria Rosa da Silva		183 1-08-06	190 3-01-21				
634	1744	Joaquina	DEIRO	Joaquina Rosa da Silva Freitas	José de Freitas	Maria Rosa da Silva	1.0000	183 1-08-06		185 8-12-11	185 9-09-30		L
64	5679	Maria	VERDA	Maria Clara Teixeira Pereira Lopes (Dona)	Pedro Pereira Lopes	Maria Clara Teixeira		180 0-05-26		183 2-12-07	183 3-01-03		
647	1725	Maria	DEIRO	Maria Clara Teixeira Pereira Lopes	Pedro Pereira Lopes	Maria Clara Teixeira	1.0000	180 0-05-26		183 2-12-07	183 6-08-10		L
65	5681	Rosa	VERDA	Rosa Mendes de Freitas	Miguel José de Oliveira	Teresa Maria da Costa		180 9-02-21			183 3-01-23		
65	7983	Rosa	DEIRO	Rosa Mendes de Freitas	Miguel José de Oliveira	Teresa Maria da Costa	1.0000	180 9-02-21					L
66	5698	José	VERDA	José	Luís Soares	Joaquina Dias		183 3-08-06					
668	1684	José	DEIRO	José Soares	Luís Soares	Joaquina Dias	1.0000		187 9-06-18	185 4-10-09	185 5-09-15		L
67	5713	Ana	VERDA	Ana Joaquina da Fonseca	António José da Fonseca	Custódia Maria Soares		181 2-10-01			183 4-04-17		
67831	1585	Ana	DEIRO	Ana Joaquina da Fonseca	António José da Fonseca	Custódia Maria Soares	1.0000	181 2-10-01					L
68	5733	Custódio	VERDA	Custódio José Gonçalves	Simão Gonçalves	Maria Ferreira					183 4-11-11		

68	1580891	Custódio	DEIRO	Custódio José Gonçalves Ferreira	Simão Gonçalves	Maria Ferreira	1.000000			1826-08-13	1830-06-06	L	
69	5867	Joaquim	VERDA	Joaquim de Castro	João de Castro	Bernardina da Silva Nogueira		1839-03-01		1866-07-04	1871-02-02		
69	1581349	Joaquim	DEIRO	Joaquim de Castro	João de Castro	Bernardina da Silva Nogueira	1.000000	1839-03-01		1866-07-04	1871-02-02	L	
70	5870	Francisco	VERDA	Francisco de Freitas	José de Freitas	Maria Rosa da Silva		1839-03-26		1897-08-19			
70	17415	Francisco	DEIRO	Francisco de Freitas	José de Freitas	Maria Rosa da Silva	1.000000	1839-03-26		1861-08-29	1866-07-08	L	
71	2958	Luísa	FALSO	Luísa	Bento Francisco	Luísa Fernandes		1748-11-21					
71	4194	Luísa	FALSO	Luísa Maria	Bento Francisco	Luísa Fernandes	1.000000	1753-10-26				L	
72	3040	Antónia	VERDA	Antónia (Solteira)	Francisco Fernandes	Mariana do Vale					1753-06-27		
72	2896	Antónia	DEIRO	Antónia	Francisco Fernandes	Mariana do Vale	1.000000	1719-04-19				L	
73	3104	Custódio	VERDA	Custódio de Freitas	Custódio de Freitas	Mariana Francisca		1754-07-18	1809-12-03	1776-07-05	1779-04-07		
73	17623	Custódio	DEIRO	Custódio de Freitas	Custódio de Freitas	Mariana Francisca	1.000000	1754-07-18	1809-12-03	1776-07-05		L	
74	3210	Manuel	VERDA	Manuel José de Freitas	José Francisco de Freitas	Luísa Maria de Freitas		1759-04-20		1789-02-14	1793-11-09		
74	1580804	Manuel	DEIRO	Manuel José de Freitas	José Francisco de Freitas	Luísa Maria de Freitas	1.000000	1759-04-20		1789-02-14		L	
75	3244	José	VERDA	José António de Castro Fernandes	Anselmo de Castro	Maria Fernandes		1761-04-06		1780-09-04	1787-06-03		
75	16206	José	DEIRO	José António de Castro Fernandes	Anselmo de Castro	Maria Fernandes	1.000000	1761-04-06		1780-09-04	1790-10-15	L	

76	3280	João	VERDA	João da Maia	Bento de Freitas	Joana Benta da Maia Baptista		174 3-09-17	180 7-11-06	176 3-12-20	176 5-03-24		
76	2686	João	DEIRO	João da Maia	Bento de Freitas	Joana Benta da Maia Baptista	1.0000	174 3-09-17	180 7-11-06			L	
77	3340	João	VERDA	João Pereira	João Pereira Fernandes	Catarina Francisca		176 2-12-06		178 6-02-20	178 8-01-16		
77	16135	João	DEIRO	João Pereira	João Pereira Fernandes	Catarina Francisca	1.0000	176 2-12-06		178 6-02-20	178 9-12-07	L	
78	3471	Francisco	VERDA	Francisco Mendes de Freitas	Francisco Mendes	Maria de Freitas		173 5-01-07	180 0-12-19	176 6-02-02	176 9-04-20		
78	16051	Francisco	DEIRO	Francisco Mendes de Freitas	Francisco Mendes	Maria de Freitas	1.0000	173 5-01-07	180 0-12-19	176 6-02-02	177 2-02-25	L	
79	3504	Maria	VERDA	Maria Joana de Oliveira	Luís da Silva	Josefa de Oliveira da Silva		177 0-11-10	181 9-09-10	179 6-02-03	179 6-12-04		
79	15856	Maria	DEIRO	Maria Joana de Oliveira	Luís da Silva	Josefa de Oliveira da Silva	1.0000	177 0-11-10	181 9-09-10	179 6-02-03	180 1-06-17	L	
80	3557	Manuel	VERDA	Manuel Monteiro	Paulo Monteiro	Catarina Monteiro			182 1-02-26	177 3-03-19	177 3-06-03		
80	16278	Manuel	DEIRO	Manuel Monteiro	Paulo Monteiro	Catarina Monteiro	1.0000				177 6-01-09	L	
81	3561	Inácio	VERDA	Inácio da Rocha	António Barbosa de Castro	Teresa de Castro da Rocha		174 0-04-17		177 3-05-29			
81	2611	Inácio	DEIRO	Inácio da Rocha	António Barbosa de Castro	Teresa de Castro da Rocha	1.0000	174 0-04-17		177 3-05-29	177 4-03-17	L	
82	2021	João	VERDA	João	Pedro Teixeira	Maria Moreira		171 0-05-09					
82	15361	João	DEIRO	João Teixeira	Pedro Teixeira	Maria Moreira	1.0000		177 1-12-17	173 5-07-31	173 6-10-23	L	
83	2063	José	VERDA	José Teixeira	João Francisco	Ana Teixeira Martins		171 2-03-10	175 3-11-04				

83	15576	José	DEIRO	José Teixeira	João Francisco	Ana Teixeira Martins	1.000000	1712-03-10	1753-11-04	1742-04-29		L	
84	2105	Luísa	VERDA	Luísa Fernandes	António Francisco	Senhorinha Fernandes		1713-07-31		1745-02-17	1745-03-26		
84	15950	Luísa	DEIRO	Luísa Fernandes	António Francisco	Senhorinha Fernandes	1.000000		1763-10-26		1753-10-26	L	
85	2113	Custódio	VERDA	Custódio Pereira	João Pereira	Catarina Francisca		1713-11-29	1793-05-12	1737-03-18	1738-03-09		
85	15543	Custódio	DEIRO	Custódio Pereira	João Pereira	Catarina Francisca	1.000000	1713-11-29	1793-05-12	1745-07-19		L	
86	2152	Cristóvão	FAISO	Cristóvão da Costa	João Durães	Catarina da Costa		1690-11-11	1743-05-30	1724-07-13	1720-12-09		
86	1784	Cristóvão		Cristóvão	João Durães	Catarina da Costa	1.000000	1686-10-06				L	
87	2152	Cristóvão	VERDA	Cristóvão da Costa	João Durães	Catarina da Costa		1690-11-11	1743-05-30	1724-07-13	1720-12-09		
87	1786	Cristóvão	DEIRO	Cristóvão da Costa	João Durães	Catarina da Costa	1.000000	1690-11-11	1743-05-30	1724-07-13	1720-12-09	L	
88	2155	Domingos	VERDA	Domingos de Castro	Domingos de Castro	Isabel Francisca		1691-09-12	1773-02-06	1725-01-21			
88	1466	Domingos	DEIRO	Domingos de Castro	Domingos de Castro	Isabel Francisca	1.000000	1690-09-12	1773-02-06	1725-01-21	1725-11-29	L	
89	2160	Cristóvão	VERDA	Cristóvão de Castro	Jerónimo de Castro	Ana Gonçalves		1693-08-08	1747-09-02	1728-08-04			
89	1669	Cristóvão	DEIRO	Cristóvão de Castro	Jerónimo de Castro	Ana Gonçalves	1.000000	1693-08-08	1747-09-02	1728-08-04	1729-05-30	L	
90	2168	João	VERDA	João de Castro Soares	João de Castro	Maria de Barros		1687-10-26	1755-05-25	1730-07-10	1732-08-21		
90	1765	João	DEIRO	João de Castro Soares	João de Castro	Maria de Barros	1.000000	1687-10-26	1755-05-25		1739-06-23	L	
91	2231	João		João	Cosme da Costa	Catarina de Freitas		1717-					

			FA LS O					05- 25					
9 1	1537 1	João		João de Freitas	Cosme da Costa	Catarina de Freitas	1.0 000 000	172 0- 05- 24		174 6- 01- 22	174 6- 11- 17	L	
9 2	2287	João	VE R D A	João de Freitas	Cosme da Costa	Catarina de Freitas		172 0- 05- 24					
9 2	1537 1	João	DE IR O	João de Freitas	Cosme da Costa	Catarina de Freitas	1.0 000 000	172 0- 05- 24		174 6- 01- 22	174 6- 11- 17	L	
9 3	2299	Ter esa	VE R D A	Teresa da Costa	Cristóvão da Costa	Maria Ribeiro		172 0- 12- 09					
9 3	1591 511	Ter esa	DE IR O	Teresa da Costa	Cristóvão da Costa	Maria Ribeiro	1.0 000 000	172 0- 12- 09	179 9- 10- 05	174 8- 05- 26	174 9- 05- 03	L	
9 4	2370	Ma nue l	VE R D A	Manuel José Peixoto	Manuel Peixoto	Rosa Peixoto		172 4- 04- 07	176 9- 11- 05				
9 4	1547 5	Ma nue l	DE IR O	Manuel José Peixoto	Manuel Peixoto	Rosa Peixoto	1.0 000 000	172 4- 04- 07	176 9- 11- 05	174 9- 03- 14	175 0- 03- 14	L	
9 5	2382	Mar ia	VE R D A	Maria Lopes	António de Sousa	Mariana Lopes		172 4- 12- 11					
9 5	1477 9	Mar ia	DE IR O	Maria Lopes	António de Sousa	Mariana Lopes	1.0 000 000	172 4- 12- 11			174 6- 10- 28	L	
9 6	2423	Ant ónio	FA LS O	António	António de Sousa	Mariana Lopes		172 7- 05- 01					
9 6	1587 1	Ant ónio		António José de Sousa	António de Sousa	Mariana Lopes	1.0 000 000	173 2- 02- 04		175 5- 12- 28	175 6- 10- 03	L	
9 7	2473	Cus tódio	VE R D A	Custódio Fernandes	Domingos de Oliveira	Maria Fernandes		173 0- 04- 26	181 8- 10- 16				
9 7	1599 2	Cus tódio	DE IR O	Custódio Fernandes	Domingos de Oliveira	Maria Fernandes	1.0 000 000	173 0- 04- 26	181 8- 10- 16	175 1- 10- 11	175 3- 08- 24	L	
9 8	2504	Ant ónio	VE R D A	António José de Sousa	António de Sousa	Mariana Lopes		173 2- 02- 04					
9 8	1587 1	Ant ónio	DE IR O	António José de Sousa	António de Sousa	Mariana Lopes	1.0 000 000	173 2- 02- 04		175 5- 12- 28	175 6- 10- 03	L	

99	2559	Custódia	VERDA	Custódia Francisca Peixoto de Freitas	Manuel Peixoto	Rosa Peixoto				1734-03-08	1735-02-09		
99	15283	Custódia	DEIRO	Custódia Francisca Peixoto	Manuel Peixoto	Rosa Peixoto	1.000000			1734-03-08	1735-02-09	L	
100	2562	Francisco	VERDA	Francisco Gomes	António Francisco	Maria Fernandes		1710-06-28	1784-08-11	1734-03-08	1735-02-09		
100	1209	Francisco	DEIRO	Francisco Gomes	António Francisco	Maria Fernandes	1.000000	1710-06-28	1784-08-11	1734-03-08	1735-02-09	L	