

# 4. Esteira de Qualidade de Dados

A esteira de qualidade de dados do MDM aplica regras de qualidade de dados que incluem rotinas de crítica, validação de dados e padronização de dados através de robustas regras de manipulação de Tokens, algoritmos de enriquecimento baseado em regras internas ou mesmo através de cruzamentos com bases externas como bureaus de crédito ou bases confiáveis como Correios e Anatel. Com os dados devidamente validados, padronizados e enriquecidos a esteira de qualidade do MDM segue para a etapa de resolução de identidade aplicando um mecanismo de agrupamento e correlação entre registros capaz de identificar dados que se referem a mesma pessoa. Nesta etapa de "Matching" são aplicados algoritmos de comparação probabilística com mecanismos transparentes de gestão e parametrização. Estes registros então serão logicamente interligados. Ou seja, os cadastros referentes ao mesmo indivíduo que estavam presentes na camada Integrada, poderão ser conectados a um único registro mestre em nível corporativo. Denominado "Golden Record", permitindo uma visão corporativa de todas as informações existentes sobre ela nos sistemas da empresa que estiverem conectados ao MDM.

- 4.1. Regras de Padronização
- 4.2. Enriquecimento de Dados
- 4.3. Crítica de Dados
- 4.4. Matching
- 4.5. Sobrevivência

# 4.1. Regras de Padronização

Processo que identifica, remove e ou corrige registros de dados imprecisos para garantir qualidade e consistência. É um processo fundamental para o gerenciamento de dados mestre (MDM).

O produto possui uma extensa biblioteca, contendo regras de padronização visando a adequação dos registros oriundos dos legados, reduzindo possíveis inconsistências que impactem nos processos de formação do golden record.

A solução também aplica a metodologia de padronização de dados da plataforma IBM, contemplando bibliotecas padrão para diversos países, inclusive o Brasil.

A MD2 enriqueceu estas rotinas com experiência de atuação de mais de 1 década implantando a solução de MDM em grandes empresas do mercado nacional e traz estes artefatos como aceleradores de projetos, além de rotinas regionalizadas e preparadas para tratar diversos tipos de dado.

Abaixo temos a tabela com algumas Regras de Padronização disponíveis:

Nome de Recurso	Nome	Descrição Detalhada
Padronização Acentuação	Padronização Acentuação	Define a forma padrão de armazenamento de strings dentro do MDM, onde os caracteres devem ser persistidos sem acentuação. Os caracteres acentuados devem ser substituídos pelo caractere correspondente sem acento
Padronização Adequação Gênero	Padronização Adequação Gênero	Adequação dos valores de gênero para M ou F. Se os valores de gênero estiverem descritivos ou numéricos, os mesmos deverão ser convertido para M ou F
Padronização Agência Bancária	Padronização Agência Bancária	São retirados dígitos não numéricos. Se necessário, complementa-se o registro com zeros à esquerda até completar 4 caracteres. Caso o registro contenha 5 dígitos, tratam-se os 4 primeiros dígitos como código da agência e o 5º dígito como verificador
Padronização Banco	Padronização Banco	O processo insere zeros a esquerda até se atingir 3 dígitos

Padronização CaracteresConsecutivos	Padronização CaracteresConsecutivos	<p>Não é permitido três ou mais caracteres iguais consecutivos conforme regra abaixo:</p> <p>. Não é permitido 2 caracteres iguais consecutivos no início de nomes ou sobrenomes, exceto para vogais, o excesso deve ser excluído, deixando apenas um caractere.</p> <p>Exemplo: Rroberto =&gt; Roberto  Ssônia =&gt; Sônia  Jjosé =&gt; José  Ddenilson =&gt; Denilson</p> <p>. Não é permitido 3 ou mais caracteres iguais consecutivos no meio dos nomes ou sobrenomes, o excesso deve ser excluído, deixando apenas dois caracteres</p> <p>Exemplo: Barrrros =&gt; Barros  Annna =&gt; Anna</p>
Padronização Caracteres Consecutivos Endereço	Padronização Caracteres Consecutivos Endereço	<p>Não é permitido três ou mais caracteres iguais consecutivos, exceto números romanos e sequência numérica</p> <p>. Não é permitido 2 caracteres iguais consecutivos no início de nomes ou sobrenomes, exceto para vogais, o excesso deve ser excluído, deixando apenas um caractere.</p> <p>Exemplo: Rroberto =&gt; Roberto  Ssônia =&gt; Sônia  Jjosé =&gt; José  Ddenilson =&gt; Denilson</p> <p>. Não é permitido 3 ou mais caracteres iguais consecutivos no meio dos nomes ou sobrenomes, o excesso deve ser excluído, deixando apenas dois caracteres</p> <p>Exemplo: Barrrros =&gt; Barros  Annna =&gt; Anna</p>
Padronização Caracteres Permitidos Bairro e Cidade	Padronização Caracteres Permitidos Bairro e Cidade	<p>Todos os caracteres devem respeitar a relação de caracteres permitidos para nome do bairro e cidade</p> <p>Caracteres diferentes da lista abaixo devem ser removidos:</p> <p>0123456789abcdefghijklmnopqrstuvwxyz</p>
Padronização Caracteres Permitidos CEP	Padronização Caracteres Permitidos CEP	<p>Todos os caracteres devem respeitar a relação de caracteres permitidos para CEP</p> <p>São permitidos os caracteres numéricos 0123456789.</p> <p>Caracteres diferentes desta lista devem ser removidos.</p>

Padronização Caracteres Permitidos CPF	Padronização Caracteres Permitidos CPF	Todos os caracteres devem respeitar a relação de caracteres permitidos para CPF São permitidos os caracteres numéricos 0123456789. Caracteres diferentes desta lista devem ser removidos.
Padronização Caracteres Permitidos Latitude e Longitude	Padronização Caracteres Permitidos Latitude e Longitude	Todos os caracteres devem respeitar a relação de caracteres permitidos para Latitude e Longitude Lista de caracteres permitidos 0123456789 .- Caracteres diferentes desta lista devem ser removidos.
Padronização Caracteres Permitidos Logradouro	Padronização Caracteres Permitidos Logradouro	Todos os caracteres devem respeitar a relação de caracteres permitidos para Logradouro (Nome, Número e Complemento) Caracteres diferentes da lista abaixo devem ser removidos: 0123456789abcdefghijklmnopqrstuvwxyz ° º , . º
Padronização Caracteres Permitidos Município e UF IBGE	Padronização Caracteres Permitidos Município e UF IBGE	Todos os caracteres devem respeitar a relação de caracteres permitidos para município e UF IBGE São permitidos os caracteres numéricos 0123456789. Caracteres diferentes desta lista devem ser removidos.
Padronização Caracteres Permitidos para E-mail	Padronização Caracteres Permitidos para E-mail	Todos os caracteres devem respeitar a relação de caracteres permitidos para E-mail São permitidos os caracteres alfabéticos abcdefghijklmnopqrstuvwxyz , numéricos 0123456789 e também os especiais . _ - @ Caracteres diferentes desta lista devem ser removidos.
Padronização Caracteres Permitidos para Nome do País	Padronização Caracteres Permitidos para Nome do País	Todos os caracteres devem respeitar a relação de caracteres permitidos para nome do país Caracteres diferentes da lista abaixo devem ser removidos: " 0123456789abcdefghijklmnopqrstuvwxyz
Padronização Caracteres Permitidos para Nome Pessoa Física	Padronização Caracteres Permitidos para Nome Pessoa Física	Todos os caracteres devem respeitar a relação de caracteres permitidos para nome de Pessoa Física Caracteres diferentes da lista abaixo devem ser removidos: " ABCDEFGHIJKLMNOPQRSTUVWXYZ"

Padronização Caracteres Permitidos para Telefone	Padronização Caracteres Permitidos para Telefone	Todos os caracteres devem respeitar a relação de caracteres permitidos para Telefone (DDI, DDD, Telefone e Ramal) São permitidos os caracteres numéricos 0123456789. Caracteres diferentes desta lista devem ser removidos.
Padronização Caracteres Permitidos RG e Passaporte	Padronização Caracteres Permitidos RG e Passaporte	Todos os caracteres devem respeitar a relação de caracteres permitidos para RG e Passaporte São permitidos os caracteres numéricos 0123456789 e alfabéticos ABCDEFGHIJKLMNOPQRSTUVWXYZ Caracteres diferentes desta lista devem ser removidos.
Padronização Caracteres Permitidos UF	Padronização Caracteres Permitidos UF	Todos os caracteres devem respeitar a relação de caracteres permitidos para UF Caracteres diferentes da lista abaixo devem ser removidos: ABCDEFGHIJKLMNOPQRSTUVWXYZ
Padronização Case Sensitive	Padronização Case Sensitive	Define a forma padrão de armazenamento de strings dentro do MDM, onde os caracteres devem ser persistidos em caixa alta (maiúsculas)
Padronização Case Sensitive E-mail	Padronização Case Sensitive E-mail	Define a forma padrão de armazenamento de strings de e-mail dentro do MDM, onde os caracteres devem ser persistidos em caixa baixa (minúsculas)
Padronização CEP Genéricos	Padronização CEP Genéricos	Não é permitido a existência de conteúdo genérico de CEP. Se conteúdo genérico, inferir nulo Ex: 00000000, 11111111 ... 99999999
Padronização Complemento CEP de São Paulo	Padronização Complemento CEP de São Paulo	Complemento com zero a esquerda para CEP de São Paulo. Concatenar um zero a esquerda quando a UF='SP' e o CEP contiver 7 dígitos

Padronização Complemento Logradouro	Padronização Complemento Logradouro	<p>Padronização informações de complemento de logradouro escritas de formas distintas ou inválidas</p> <p>Quando iniciar com SL e logo após a letra possuir espaço, substituir por "SALA"</p> <p>Quando iniciar com S e após o espaço a direita houver um caracter diferente da letra N , substituir por "SALA"</p> <p>Quando campos possuir SN, S/N, S N, remover da string</p>
Padronização Complemento Zero a Esquerda CPF	Padronização Complemento Zero a Esquerda CPF	<p>Complementar com zero a esquerda do CPF quando conteúdo for inferior a 11 dígitos.</p> <p>Quando o CPF possuir quantidade inferior a 11 dígitos, incluir zeros a esquerda completando o número em 11 dígitos</p>
Padronização Completude Sufixo e Prefixo Nome Pessoa Física	Padronização Completude Sufixo e Prefixo Nome Pessoa Física	<p>Aplicar rotina QualityStage de padronização de nomes para completude de sufixo e prefixo do nome, corrigindo as principais abreviaturas e movendo o prefixo do nome para nome de tratamento</p> <p>Exemplo: "DR. JOAO DA SILVA JR" -&gt; "JOAO DA SILVA JUNIOR" , o prefixo DR. será movido para o campo de nome de tratamento</p>
Padronização Conteúdo SN	Padronização Conteúdo SN	<p>Padronização informação "Sem Número" escrita de formas distintas.</p> <p>Quando conteúdo contiver "SNUMERO, SN, S N,S/NUMERO, SN, S/N, S N, S/NR, SNR" entre espaços, substituir a string por "S/N"</p>

Padronização Correção Abreviações Bairro	Padronização Correção Abreviações Bairro	<p>Correção de abreviações comuns para nome do bairro. Substituir:</p> <p>. Z. ou Z por Zona . VL V.L. VL. V.L maiúsculas ou minúsculas por VILA . STA STA. STª Sta maiúsculas ou minúsculas por SANTA . RS RES RES. Res. Res por RESIDENCIAL . PRQ PQUE Pque PQ PQ. Pq Pq. por PARQUE . JD. JD Jdim JDIM Jd Jd. por Jardim . Dist. Dist Distr. DISTR DIS DIS. Dis Dis. por DISTRITO . CPO por CAMPO . COND COND. por CONDOMINIO</p> <p>Aplicar rotina QualityStage para completude e padronização da informação do nome do bairro</p>
Padronização Correção Provedores de E-mail	Padronização Correção Provedores de E-mail	<p>Completude e padronização da informação de e-mail para tradução de erros comuns de provedores de e-mail</p> <p>Exemplo gmael -&gt; gmail gmai -&gt; gmail hgotmail -&gt; hotmail hhotmail -&gt; hotmail</p> <p>Acerto no final do e-mail onde após o provedor não existir ".com", ".com.br", "br"</p>
Padronização Correção Erros Comuns Final E-mail	Padronização Correção Erros Comuns Final E-mail	<p>Correção dos erros comuns no final da string de E-mail Exemplo: "com.ltda" -&gt; ".com.br" "comm.br" -&gt; ".com.br" ".cvom.br" -&gt; ".com.br"</p>
Padronização Data Nascimento Inconsistente	Padronização Data Nascimento Inconsistente	<p>Os valores contidos na data de nascimento devem ser consistentes, ou seja, devem possuir um intervalo de valores mínimo e máximo. Valores superiores a data atual e inferiores a 1900-01-01 devem ser anulados</p>

Padronização Data ÓbitoInconsistente	Padronização Data ÓbitoInconsistente	Os valores contidos na data de óbito devem ser consistentes, ou seja, devem possuir um intervalo de valores mínimo e máximo. Valores superiores a data atual , inferiores a 1900-01-01 ou inferiores a data de nascimento devem ser anulados
Padronização Espaçamento de Strings	Padronização Espaçamento de Strings	Define a forma padrão de espaçamento de strings. O excesso de espaçamento deve ser removido.  Espaço no início ou no final da string também deve ser removido, exemplo: " JOAO DA SILVA SOARES " -> "JOAO DA SILVA SOARES"
Padronização Formato de Data	Padronização Formato de Data	As datas devem ser armazenadas seguindo um formato padrão. Armazenar no HUB MDM as datas no formato: YYYY-MM-DD HH:MM:SS
Padronização Inclusão de Dígitos	Padronização Inclusão de Dígitos	Inclusão do nono dígito para telefone celular e dígito três para telefone fixo Se possuir oito dígitos e Iniciado por 6, 7, 8 ou 9: . Incluir o 9 a esquerda do número do telefone para telefones que não sejam NEXTEL conforme tabela ANATEL Se possuir sete dígitos e o campo tiver data de alteração/inclusão anterior ao ano de 2006, incluir o número 3 no início do número
Padronização Quantidade Máxima de Caracteres Número Logradouro	Padronização Quantidade Máxima de Caracteres Número Logradouro	O número de logradouro não deve ser maior que 14 caracteres. Caso ultrapasse o valor máximo, o conteúdo do número do logradouro deverá ser anulado
Padronização Quantidade Máxima de Números Ramal	Padronização Quantidade Máxima de Números Ramal	O número do ramal deve respeitar a quantidade máxima de caracteres. Os valores que não respeitarem essa restrição deverão ser anulados
Padronização Quantidade Mínima de Caracteres Bairro	Padronização Quantidade Mínima de Caracteres Bairro	O nome do bairro não deve ser menor que 2 caracteres. Caso seja inferior ao valor mínimo, o conteúdo do nome do bairro deverá ser anulado
Padronização Quantidade Mínima de Caracteres Complemento Logradouro	Padronização Quantidade Mínima de Caracteres Complemento Logradouro	O complemento de logradouro não deve ser menor que 2 caracteres. Caso seja inferior ao valor mínimo, o conteúdo do complemento do logradouro deverá ser anulado



Padronização Remoção Caracteres Indesejados E-mail	Padronização Remoção Caracteres Indesejados E-mail	<p>Não é permitido a existência de determinados caracteres antes e após o @ e caracteres especiais em sequência.</p> <p>Conforme relação abaixo, devemos substituir :</p> <p>@@ por @</p> <p>-@ por @</p> <p>@- por @</p> <p>.@ por @</p> <p>@. por @</p> <p>_@ por @</p> <p>@_ por @</p> <p>-- por -</p> <p>.. por .</p>
Padronização Remoção de Dígitos	Padronização Remoção de Dígitos	Remover zeros a esquerda do Telefone, DDD e DDI
Padronização Remoção Espaço E-mail	Padronização Remoção Espaço E-mail	Não é permitido espaços em branco na string de e-mail. Os espaços entre strings, no início e no final da string devem ser removidos
Padronização Remoção Palavra Indesejada para Nome Cidade	Padronização Remoção Palavra Indesejada para Nome Cidade	<p>Palavras indesejadas devem ser removidas do conteúdo Nome Cidade Possuindo a palavra Capital no final da string, a mesma deverá ser excluída.</p> <p>Exemplo: Rio de Janeiro Capital - &gt; Rio de Janeiro</p> <p>Possuindo a string 'N D' , substituir por nulo</p>
Padronização Remoção Pontuação no Início e Final da String	Padronização Remoção Pontuação no Início e Final da String	A string de e-mail não pode iniciar ou terminar com caractere .(ponto). Os pontos no início e no final da string devem ser removidos, caso existam
Padronização Remoção String Indesejada E-mail	Padronização Remoção String Indesejada E-mail	<p>Remover do conteúdo a string "e-mail:"</p> <p>Exemplo: "e-mail: joaodasilva@email.com.br" -&gt; "joaodasilva@email.com.br"</p>
Padronização Remoção String Indesejada Nome Logradouro	Padronização Remoção String Indesejada Nome Logradouro	<p>Strings indesejadas devem ser removidas do conteúdo Nome Logradouro</p> <p>Quando conteúdo contiver "S/NUMERO, SN, S/N, S N, S/NR, SNR" entre espaços, remover da string</p>
Padronização Remoção Zero a Esquerda CEP	Padronização Remoção Zero a Esquerda CEP	Remoção zero a esquerda do CEP caso contenha 9 dígitos e o primeiro dígito for zero

Padronização Separação Conteúdo Nome Logradouro	Padronização Separação Conteúdo Nome Logradouro	<p>Separação Tipo, Número e Complemento Logradouro do Nome Logradouro</p> <p>Aplicar rotina QualityStage para completude e padronização da informação do nome do logradouro , tipo de logradouro, número do logradouro e complemento, separando as informações caso estejam presentes na string de Nome Logradouro</p>
Padronização Separação Conteúdo Número Logradouro	Padronização Separação Conteúdo Número Logradouro	<p>Separação Complemento Logradouro do Número Logradouro.</p> <p>Quando número do logradouro iniciar com AP, APTO ou APT e houver números após esses caracteres, remover da string</p> <p>Quando número do logradouro iniciar com AP, APTO ou APT e houver números após esses caracteres, retirar do campo “número” e acrescentar ao campo complemento sem excluir o que já existe nesse campo. Se a informação retirada no campo “número” for igual a presente no campo “complemento”, descartar informação</p> <p>Quando número do logradouro iniciar com número da esquerda para a direita e após esses tiver AP, APTO, APT, CASA ou CS e houver números da esquerda para a direita após esses caracteres, remover todo o conteúdo da string após o primeiro número</p> <p>Ex: 123 AP 456 -&gt; 123 permaneceria em número logradouro e AP 456 seria migrado para complemento logradouro</p>
Padronização Separação de E-mail	Padronização Separação de E-mail	<p>Separa as ocorrências de vários e-mails em uma mesma string a partir dos caracteres delimitadores "\&gt;&lt; , ; # -".</p> <p>Os dados entre eles devem ser quebrados em linhas para análise e tratamento unitário</p>

Padronização Separação de Telefone	Padronização Separação de Telefone	<p>Separa as ocorrências de vários telefones em uma mesma string a partir das seguintes regras:</p> <p>Quando possuir os caracteres delimitadores " ; / OU ", eliminar o caractere delimitador, separando o conteúdo de telefone em linhas distintas. Exemplo:  32227856ou991913455  32227856;991913455 , ficará:  32227856 991913455 32227856 991913455, sendo cada número de telefone um novo registro</p> <p>Para campos com 15 caracteres, somente numéricos, dividi-los em duas partes (8 dígitos e 7 dígitos), separando em linhas distintas de telefone.</p> <p>Para campos com 16 caracteres, somente numéricos, dividi-los em duas partes iguais com 8 caracteres cada em linhas distintas de telefone.</p>
Padronização Separação Município da UF IBGE	Padronização Separação Município da UF IBGE	<p>Separação Município da UF IBGE quando o conteúdo estiver em uma mesma string</p> <p>Realizar a separação do código da UF e do município nos casos em que o campo "código UF IBGE" contem 7 dígitos. E o "código município IBGE" não contenha conteúdo</p> <p>Recuperar os dois primeiros dígitos para código UF IBGE</p> <p>Recuperar os 5 últimos dígitos para código município IBGE</p>
Padronização Separação RG, UF e Órgão Emissor	Padronização Separação RG, UF e Órgão Emissor	<p>Separa o número RG, UF e Órgão Emissor contidos em uma mesma string</p> <p>Exemplo:  MG 102030 SSP -&gt; MG (UF Emissor), 102030 (Número RG), SSP (Órgão Emissor)</p>

Padronização Separação UF do Nome Cidade	Padronização Separação UF do Nome Cidade	<p>Separação da UF do Nome da Cidade quando o conteúdo contiver as duas informações</p> <p>A separação da UF deve ocorrer a partir da aplicação da regra abaixo: Se o terceiro caractere for traço “-” ou barra “/”(desconsiderando os espaços) e a direita dele possuir dois caracteres alfabéticos, remover traço “-” ou barra “/” . Os dois caracteres posteriores serão removidos do Nome da Cidade e movidos para UF</p> <p>Ex: São Paulo - SP -&gt; Cidade ficaria com o conteúdo "São Paulo" e UF ficaria com o conteúdo "SP"</p>
Padronização Substituição Dígito Dois ou Caractere Asterico por Arroba	Padronização Substituição Dígito Dois ou Caractere Asterico por Arroba	<p>Substituir o dígito 2 pelo caractere @ quando:</p> <p>O conteúdo do campo e-mail conter somente uma ocorrência do número 2 e o campo e-mail não possuir @</p> <p>Substituir o caractere * pelo caractere @ quando:</p> <p>O conteúdo do campo e-mail conter somente uma ocorrência do * e o campo e-mail não possuir @</p>
Padronização Tamanho Padrão de Caracteres CEP	Padronização Tamanho Padrão de Caracteres CEP	<p>Os valores de CEP devem respeitar o tamanho padrão de 8 dígitos numéricos.</p> <p>Se quantidade de dígitos do CEP for diferente de 8, inferir Nulo</p>
Padronização Tamanho Padrão PIS/PASEP/NIT	Padronização Tamanho Padrão PIS/PASEP/NIT	<p>Define o tamanho padrão de 11 dígitos para armazenamento das informações de PIS/PASEP/NIT</p> <p>Campos inferiores a 11 caracteres, completar com zero a esquerda</p>
Padronização Tradução Nome Cidade	Padronização Tradução Nome Cidade	<p>Tradução das abreviações e correção de erros comuns de digitação do Nome Cidade</p> <p>Aplicar rotina QualityStage para tradução das abreviações e correção de erros comuns de digitação do Nome Cidade</p> <p>Exemplo: BH -&gt; BELO HORIZONTE</p> <p>                  MOJIMIRIM -&gt; MOGI MIRIM</p>
Padronização Validação Bairro DNE	Padronização Validação Bairro DNE	<p>Efetuar validação conjunta da UF, CIDADE, BAIRRO com a tabela DNE_BAIRRO dos Correios.</p> <p>Para as informações que não forem consistentes, aplicar rotina QualityStage de matching para comparação aproximada do BAIRRO com a tabela DNE_BAIRRO dos Correios</p>

Padronização Validação Cidade DNE	Padronização Validação Cidade DNE	Efetuar validação da UF e CIDADE com a tabela DNE_CIDADE dos Correios. Para as informações que não forem consistentes, aplicar rotina QualityStage de matching para comparação aproximada da CIDADE com a tabela DNE_CIDADE dos Correios
Padronização Validação DDD Anatel	Padronização Validação DDD Anatel	Verificar se o DDD é válido na Anatel. Caso os valores não sejam válidos, os mesmos deverão ser anulados
Padronização Validação DDI Anatel	Padronização Validação DDI Anatel	Verificar se o DDI é válido na Anatel. Caso os valores não sejam válidos, os mesmos deverão ser anulados
Padronização Validação Nome Logradouro DNE	Padronização Validação Nome Logradouro DNE	Efetuar validação conjunta da UF, CIDADE, BAIRRO e NOME LOGRADOURO com a tabela DNE_LOGRADOURO dos Correios. Para as informações que não forem consistentes, aplicar rotina QualityStage de matching para comparação aproximada do NOME LOGRADOURO com a tabela DNE_LOGRADOURO dos Correios
Padronização Validação Nome País DNE	Padronização Validação Nome País DNE	Realizar a validação do Nome do País com a tabela DNE_PAIS. Informações que não forem consistentes deve-se inferir Nulo no campo
Padronização Validação Prefixo Telefone e DDD Anatel	Padronização Validação Prefixo Telefone e DDD Anatel	Verificar se o prefixo do Telefone e DDD são válidos na Anatel a) Para telefones celulares (primeiro dígito=9), devemos pegar os 5 primeiros dígitos como prefixo b) Para telefone fixo (primeiro dígito 2, 3, 4 ou 5 ), pegar os 4 primeiros dígitos como prefixo Validar o DDD e PREFIXO com a tabela BCR_PREFIXO_ANATEL através dos campos NUMERO_DDD e NUMERO_PREFIXO.
Padronização Validação Tipo Logradouro DNE	Padronização Validação Tipo Logradouro DNE	Efetuar validação do Tipo de Logradouro com o DNE dos Correios. Validar campo descritivo tipo logradouro na tabela de dominio DNE_TIPO_LOGRADOURO, inferir nulo caso não seja válido

Padronização Validação UF DNE	Padronização Validação UF DNE	Efetuar validação do campo UF com a tabela DNE_UF dos Correios, as informações que não forem consistentes, inferir Nulo
-------------------------------	-------------------------------	---

## 4.2. Enriquecimento de Dados

Enriquecimento de dados refere-se a processos usados para aprimorar, refinar ou melhorar os dados. No mundo do MDM, o enriquecimento de seus dados mestres pode acontecer pela inclusão de dados de terceiros para obter uma visão mais completa, por exemplo.

O MD2 MDM aplica diversas técnicas de enriquecimento de dados evoluídas a partir da larga experiência na implementação de soluções robustas nesse contexto em diversas empresas do âmbito nacional. As técnicas podem ser classificadas como atômicas, regras internas ou mesmo consulta a bases externas de referência.

Exemplos de regras de enriquecimento de dados:

Atômica:

- Inferência de Gênero pelo Nome
- Eliminação de Duplo @@ em e-mails

Internas:

- Ajustes de Domínios de Dados Corporativos (De-Para de Domínios)
- Regras de Negócio ou Bases Internas de Referência

Externas

- Validação de Titularidade de Documentos (em Bureaus Externos)
- Enriquecimento de Endereços baseado no DNE (Diretório Nacional de Endereçamento Postal)

Exemplificando esse tópico, abaixo uma tabela com as regras de enriquecimento de dados referentes ao DNE implementadas em nossa solução:

Nome de Recurso	Nome	Descrição Detalhada
Enriquecimento Bairro DNE	Enriquecimento Bairro DNE	Enriquecimento do Bairro a partir do número do CEP considerando a seguinte regra: Se campo Bairro não estiver preenchido, recuperar o Bairro pelo intervalo de CEP usando as tabelas do DNE para bairros pertencentes a cidades que não possuem CEP único

Enriquecimento CEP DNE	Enriquecimento CEP DNE	<p>Enriquecimento do CEP a partir daUF, Cidade, Bairro e Logradouro considerando a seguinte regra:</p> <p>Se campo CEP não estiver preenchido, recuperar o CEP através da UF, Cidade, Bairro e Logradouro usando as tabelas do DNE para cidades que não possuem CEP Único</p> <p>Se campo CEP não estiver preenchido e Cidade com CEP único, recuperar CEP da tabela DNE_CIDADE</p>
Enriquecimento Cidade DNE	Enriquecimento Cidade DNE	<p>Enriquecimento da Cidade a partir do número do CEP considerando a seguinte regra:</p> <p>Se campo Cidade não estiver preenchida, recuperar a Cidade pelo intervalo de CEP usando as tabelas do DNE</p> <p>Para Cidades com CEP único, se Cidade não estiver preenchida, recuperar a Cidade pelo CEP usando a tabela DNE_CIDADE</p>
Enriquecimento DDD	Enriquecimento DDD	<p>Enriquecimento do DDD a partir do número do telefone considerando a seguinte regra:</p> <p>. Número do telefone contendo dez caracteres, verificar se os dois primeiros caracteres da esquerda para a direita se enquadram da relação da Anatel dos códigos de Discagem Direta a Distância, caso positivo, separa-los no campo DDD, e os 8 caracteres restantes permanecerão no campo número do telefone</p> <p>. Número do telefone contendo onze caracteres verificar se os dois primeiros caracteres da esquerda para a direita se enquadram da relação da Anatel dos códigos de Discagem Direta a Distância, caso positivo, separa-los no campo DDD, e os 9 caracteres restantes permanecerão no campo número do telefone</p>
Enriquecimento DDI	Enriquecimento DDI	<p>Enriquecimento do DDI a partir da validação dos dados na Anatel</p> <p>Quando os campos DDD e Telefone estiverem preenchidos e válidos na Anatel, preencher automaticamente o DDI do Brasil (55).</p>



Enriquecimento Gênero	Enriquecimento Gênero	<p>Enriquecimento do Gênero a partir do nome da pessoa utilizando a regra de padronização do QualityStage que possui a lista com o de-para nome x gênero.</p> <p>O enriquecimento só deverá ocorrer se o Gênero estiver sem conteúdo e a regra de padronização QualityStage efetuar a geração do Gênero a partir do nome</p>
Enriquecimento Logradouro DNE	Enriquecimento Logradouro DNE	<p>Enriquecimento do Nome Logradouro a partir do número do CEP considerando a seguinte regra: Se campo Logradouro não estiver preenchido, recuperar o Logradouro pelo CEP usando as tabelas do DNE</p>
Enriquecimento Nacionalidade	Enriquecimento Nacionalidade	<p>Enriquecimento da Nacionalidade com a string "BR" quando a Nacionalidade não estiver preenchida e os dados de UF ou Nome da Cidade forem válidos no DNE</p>
Enriquecimento UF Cidade DNE	Enriquecimento UF Cidade DNE	<p>Enriquecimento da UF a partir do nome da cidade do DNE Correios Quando a UF não estiver preenchida, recuperar a UF da tabela DNE_CIDADE a partir do nome da cidade.</p> <p>O enriquecimento deverá ser realizado somente quando o nome da cidade for exclusivo para uma única UF em todo o Brasil</p>
Enriquecimento UF DNE	Enriquecimento UF DNE	<p>Enriquecimento da UF a partir do número do CEP considerando a seguinte regra: Se campo UF não estiver preenchido, recuperar a UF a pelo intervalo de CEP usando as tabelas do DNE</p>

A solução parte do pressuposto que o enriquecimento garante dados mais confiáveis e fidedignos, auxiliando por consequência em todo processo de construção do Golden Record. Por meio de sua arquitetura, incluindo processos e modelo de dados, possibilita-se a adequação visando a utilização de diversas técnicas para enriquecer os dados.

## 4.3. Crítica de Dados

Os registros provenientes dos sistemas legados podem estar repletos de anomalias. Os motivos podem ser adversos, inclusive detectados em tempo de projeto. É preciso criar estruturas rígidas de críticas e filtros das informações da origem, para que sua entrada no MDM seja autorizada. Caso contrário estes problemas serão migrados para a base qualificada, afetando sua utilização estratégica.

Nesta etapa, é necessário aplicar as regras de crítica e validação da informação com objetivo de estabelecer critérios de qualidade para que a informação possa ser persistida no MDM. Alguns exemplos de regras de crítica:

- Nome de Pessoa Física deve conter mais que uma palavra;
- Nome de Pessoa Física não pode conter palavrões;
- CPF, CNPJ, CNH devem possuir o dígito verificador válido;
- O E-mail deve possuir sintaxe válida;
- O Telefone não pode conter mais que onze dígitos.

Abaixo temos a tabela com as Críticas de Dados:

Nome de Recurso	Nome	Descrição Detalhada
Crítica CNH	Crítica CNH	A crítica de CNH segue as seguintes regras: 1- Verificação de tamanho da CNH, sendo 9 dígitos validamos de acordo com os padrões da CNH Antiga, sendo 11 dígitos validamos conforme os padrões da CNH nova 2- Verifica se todos dígitos são numéricos 3- Verifica se o documento não é viciado. 4- Valida-se dígito verificador Como adendo, caso a categoria da CNH seja inválida cria-se um alerta, porém caso o CNH seja válido o dado é somado ao Golden Record
Crítica CTPS	Crítica CTPS	A crítica de CTPS invalida um registro nos casos em que: 1- O número de série tiver tamanho maior que 5 dígitos 2- O número do documento tiver tamanho maior que 8 dígitos 3- O número do documento estiver viciado

Crítica DDD e DDI Nulo	Crítica DDD e DDI Nulo	Gerar alerta se o número do DDD ou DDI possuírem conteúdo nulo, vazio ou somente espaços em branco
Crítica Documento com Vício Preenchimento	Crítica Documento com Vício Preenchimento	<p>O número do documento não pode conter vício de preenchimento.</p> <p>O número do documento não pode conter vício de preenchimento, ou seja, não pode conter a repetição de um mesmo caractere em todo o conteúdo da string Exemplo: XXXXX,ZZZZZ,11111,22222 Os registros que não respeitarem essa restrição deverão ser invalidados</p>
Crítica E-mail com Vício de Preenchimento	Crítica E-mail com Vício de Preenchimento	<p>O e-mail não pode conter vício de preenchimento, ou seja, não pode conter a repetição de um mesmo caractere em todo o conteúdo da string Exemplo: XXXXX,ZZZZZ</p>

Crítica E-mail Fora do Padrão	Crítica E-mail Fora do Padrão	<p>Os e-mails devem obedecer uma estrutura padrão de conteúdo, conforme regra abaixo:</p> <ul style="list-style-type: none"> <li>• À esquerda do @: <ul style="list-style-type: none"> <li>. O primeiro caractere deve ser alfanuméricos abcdefghijklmnopqrstuvwxyz ou 0123456789 . Ex: Exemplo: maria-rossi@hotmail.com ,maria-rossi2000@hotmail.com, 123maria@hotmail.com</li> <li>. É obrigatório conter pelo menos um caractere alfabético</li> </ul> </li> <li>• À direita do @: <ul style="list-style-type: none"> <li>. Deve haver no mínimo dois caracteres alfanuméricos abcdefghijklmnopqrstuvwxyz 0123456789</li> <li>. Não é permitido iniciar com caractere diferente dos alfanuméricos Ex: nmneto@.md2.com.br (não permitido)</li> <li>. Não é permitida a finalização com qualquer caractere diferente do alfabético abcdefghijklmnopqrstuvwxyz Ex: nmneto@md2.com.br.. (não permitido) nmneto@md2.com.br1 (não permitido)</li> <li>. Deve possuir pelo menos um ponto e no máximo 3 pontos.</li> <li>. Entre ou após os pontos deve conter no mínimo dois caracteres alfanuméricos</li> <li>. Não é permitido dois ou mais pontos sequenciais, Ex: nmneto@md2..com.br =&gt; (não permitido)</li> </ul> </li> </ul> <p>Ex. permitido:    Ex. não permitido xpto@r7.br    xpto@_r1.com xpto@r-7.aaa.com.br    xpto@r7.aa-.a xpto@r_7.us    xpto@aa-.a_aa</p> <p>Os registros que violarem o formato padrão deverão ser invalidados</p>
-------------------------------	-------------------------------	--

Crítica Email Genérico	Crítica Email Genérico	<p>O e-mail não pode conter valores genéricos</p> <p>Aplicar rotina QualityStage para completude e padronização da informação de e-mail para identificação de e-mails genéricos mais comuns.</p> <p>Exemplo:</p> <p>clientenaopossui  clientenaopossuiem  clientenaotem  clientenaotememail  naopossuiemail  naopossuiemailpessoal</p> <p>Os registros identificados como e-mails genéricos deverão ser invalidados</p>
Crítica Endereço Nulo	Crítica Endereço Nulo	<p>As informações de endereço devem estar preenchidas</p> <p>Pelo menos um dos campos de endereço, tais como nome do logradouro, bairro, cidade ou CEP devem estar preenchidos.</p> <p>Os registros que não respeitarem essa restrição deverão ser invalidados.</p>
Crítica Estrutura Padrão Passaporte	Crítica Estrutura Padrão Passaporte	<p>Os valores de Passaporte devem respeitar a estrutura padrão.</p> <p>Campo composto por no máximo 11 algarismos, composto por números e letras, porém é permitido letras antes e após os números, não podendo possuir letras entre os números</p> <p>Ex. não permitido: W123456AB789</p> <p>Ex. permitido: W123456Z</p> <p>Registros que não atenderem a regra estrutural, deverão ser invalidados</p>
Crítica Estrutura Padrão RG	Crítica Estrutura Padrão RG	<p>Os valores de RG de estados diferente de RJ e SP devem respeitar o tamanho padrão entre 2 e 11 dígitos. Para os estados de RJ e SP o tamanho permitido é de 9 caracteres e ainda é utilizado uma regra que valida o dígito verificador. Será gerando uma alerta para registros que não atenderem essa condição</p>

Crítica Nome Blocklist	Crítica Nome Blocklist	<p>O nome não pode conter palavras existentes na blocklist de nomes. A blocklist de nomes contém uma relação de palavrões e xingamentos que podem estar ocultos no nome da pessoa.</p> <p>Os registros que não respeitarem essa restrição deverão ser invalidados.</p>
Crítica Nome com Vício de Preenchimento	Crítica Nome com Vício de Preenchimento	<p>O nome da Pessoa Física ou Jurídica não pode conter vício de preenchimento, ou seja, não pode conter a repetição de um mesmo caractere em todo o conteúdo da string Exemplo: XXXXX, ZZZZZ</p> <p>Os registros que não respeitarem essa restrição deverão ser invalidados</p>
Crítica Nome Nulo	Crítica Nome Nulo	<p>O nome da Pessoa Física ou Jurídica deve estar preenchido O nome da Pessoa Física ou Jurídica não pode ser nulo, vazio ou somente espaços em branco. Os registros que não respeitarem essa restrição deverão ser invalidados.</p>
Crítica Nome Pai e Mãe com Vício de Preenchimento	Crítica Nome Pai e Mãe com Vício de Preenchimento	<p>O nome do Pai e Mãe não pode conter vício de preenchimento, ou seja, o não pode conter a repetição de um mesmo caractere em todo o conteúdo da string Exemplo: XXXXX, ZZZZZ</p> <p>Os nomes de Pai e Mãe que não respeitarem essa restrição deverão ser anulados</p>
Crítica Nome Pai e Mãe na Blocklist	Crítica Nome Pai e Mãe na Blocklist	<p>O nome do Pai e Mãe não pode conter palavras existentes na blocklist de nomes. A blocklist de nomes contém uma relação de palavrões e xingamentos que podem estar ocultos no nome da pessoa.</p> <p>Os nomes que não respeitarem essa restrição deverão ser anulados e um alerta será gerado informando essa situação</p>
Crítica Pessoas Diferentes Com o Mesmo CPF	Crítica Pessoas Diferentes Com o Mesmo CPF	<p>O CPF não pode estar associado a mais de uma pessoa. Caso contrario, o registro deverá ser invalidado</p>

Crítica Quantidade Mínima de Caracteres Nome Pai e Mãe	Crítica Quantidade Mínima de Caracteres Nome Pai e Mãe	O nome do Pai e Mãe deve respeitara quantidade mínima de 3caracteres. Os nomes que não respeitarem essa restrição deverão ser anulados e um alerta será gerado informando essa situação
Crítica Quantidade Mínima de Palavras Nome Pessoa Física	Crítica Quantidade Mínima de Palavras Nome Pessoa Física	O nome de Pessoa Física deve respeitar a quantidade mínima de palavras Os registros com Nome de Pessoa Física que não respeitarem as regras abaixo deverão ser invalidados: Nome com apenas uma palavra Nome com duas palavras, com um caractere em um das palavras
Crítica Telefone com Vício de Preenchimento	Crítica Telefone com Vício de Preenchimento	O número do telefone não pode conter vício de preenchimento, ou seja, não pode conter a repetição de um mesmo caractere em todo o conteúdo da string Exemplo: 11111,22222 Os registros que não respeitarem essa restrição deverão ser invalidados
Crítica Telefone Fora do Padrão	Crítica Telefone Fora do Padrão	Os telefones devem obedecer uma estrutura padrão de conteúdo, conforme regra abaixo: Telefones devem respeitar a seguinte regra estrutural: Fixo: deve conter 8 caracteres numéricos e iniciados (da esquerda para a direita) pelos números 2 (dois), 3 (três), 4 (quatro) ou 5 (cinco); Móvel:deve conter 9 caracteres numéricos iniciados pelo número 9 Especial: iniciar com 0800 e possuir até 11 números  Os registros que violarem o formato padrão deverão ser invalidados
Crítica Telefone Nulo	Crítica Telefone Nulo	O número do telefone não pode ser nulo, vazio ou somente espaços em branco. Os registros que não respeitarem essa restrição deverão ser descartados.
Crítica Validação CPF	Crítica Validação CPF	Validação CPF pelo dígito verificador. Caso documento não seja válido , o registro deverá ser invalidado
Crítica Validação PIS	Crítica Validação PIS	Validação PIS pelo dígito verificador. Caso documento não seja válido , o registro deverá ser invalidado

## 4.4. Matching

Uma vez os dados padronizados, criticados e muitas vezes enriquecidos chega-se ao momento em que deve-se comparar e agrupar os registros que se correspondem. Para fazer essa **comparação** ou **matching**, o processo calcula a probabilidade de um registro estar relacionado a outro. Ele envolve uma configuração de quão semelhantes são os registros, usando limiares (limites). Os limites definem qual o valor necessário para que os registros sejam considerados duplicados ou não.

Através dessas regras, é possível realizar uma comparação probabilística e estabelecer notas para validar se duas ou mais ocorrências se tratam de um mesmo registro, mesmo que possuam divergências entre si, como por exemplo **Igor, Ygor ou Higor**. A solução leva em consideração muito mais do que apenas o NOME, pois é possível unificar **homônimos**, nem apenas o CPF pois pode-se unificar **irmãos ou marido e mulher** que compartilharam daquele documento, situações essas bastante comuns encontradas nos sistemas.

Para serem mais assertivos, as regras de matching/comparações, aplicadas na solução MD2 MDM são mais inteligentes e robustas, possuindo inúmeros passos, com diferentes regras que foram sendo desenvolvidas e melhoradas durante anos de prática nos inúmeros projetos entregues. A cada nova release essas regras podem ser revisadas pela equipe de desenvolvimento que estão constantemente revisando e melhorando os motores de qualidade.

Uma vez feita essa comparação, cada registro é agrupado com seus similares e ficam à disposição para seguir a diante na esteira de qualidade.




## 4.5. Sobrevivência

Os registros de dados devidamente criticados e validados, após terem sido padronizados e possivelmente enriquecidos, são comparados contra a base de dados alvo e devidamente agrupados com a visão corporativa.

No entanto, a gravação na base única é feita usando o registro conhecido como Golden Record (Melhor registro). O Golden Record é composto de informações de quaisquer das instâncias desta entidade, podendo conter dados de diversos registros cuja informação tenha sido definida por meio de uma regra de sobrevivência estabelecida.

As regras de sobrevivência de registros visam estabelecer critérios para que o registro resultante do processo de unificação contemple as melhores informações possíveis (mais adequadas ao negócio), de forma automatizada;



Prof	Antonio J Fernandes	19450815	20030514
Dr	Antonio José Fernandes		20021220
Prof	Antonio José Fernandes	19450815	<b>20080630</b>
Prof	Antonio José Fernandes	19540815	20080425

<b>Prof</b>	<b>Antonio José Fernandes</b>	<b>19450815</b>	<b>20080630</b>
-------------	-------------------------------	-----------------	-----------------

Como cada empresa possui sua particularidade, a solução permite a customização das regras para a definição/escolha da melhor informação. Essa escolha pode levar em consideração diversos critérios como:

- **Menor / Maior (quantidade de caracteres);**
- **Mais frequente / Mais frequente não nulo;**
- **Igual a / Diferente de;**
- **Maior que / Menor que;**
- **Pelo menos um;**
- **Nomeação de atributo por meio de verificação dos sistemas de origem;**

A partir da aplicação das regras é possível verificar os resultados e, caso necessário, alterá-las para sanar qualquer inconsistência verificada nos golden records.

Além dessas regras mais simples, é possível customizar regras mais complexas que levarão em consideração mais de uma informação ao mesmo tempo como por exemplo a regra abaixo:

Valor **não nulo**, conforme priorização abaixo. Para desempate levar em consideração o registro **mais atual**

1º - **SISTEMA\_RH**

2º - **SISTEMA\_MARKETING**

3º - **SISTEMA\_FINANCEIRO**

Vamos entender a regra acima:

- Primeiro é priorizado a informação **não nula**, independente da origem;
- Com o grupo de registros que possuem a informação preenchida é priorizado por exemplo a origem da informação. O usuário de negócio com o conhecimento corporativo, definiu por exemplo que o nome da pessoa no **SISTEMA\_RH** tem uma qualidade superior aos outros sistemas pois são validados por exemplo no E-Social.
- Por fim, caso esse sistema possua registros duplicados para desempate, o analista por exemplo definiu que o registro **mais atual** deve prevalecer.

Outro ponto importante e destacado na solução é que essas regras são customizadas por grupo de informação. Para a escolha do **NOME** é possível criar algo semelhante ao exemplo acima, porém para a escolha do melhor **E-MAIL** é possível definir uma regra diferente. Nesse caso pode ser que o sistema de marketing tenha uma informação mais atual e mais confiável. E é essa composição que formará o **Golden Record**.