

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 => Roberto Ssônia => Sônia Jjosé => José Ddenilson => 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 => Barros Annna => 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 => Roberto Ssônia => Sônia Jjosé => José Ddenilson => 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 => Barros Annna => 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	Padronização informações de complemento de logradouro escritas de formas distintas ou inválidas Quando iniciar com SL e logo após a letra possuir espaço, substituir por "SALA" Quando iniciar com S e após o espaço a direita houver um caracter diferente da letra N , substituir por "SALA" Quando campos possuir SN, S/N, S N, remover da string

Padronização Complemento Zero a Esquerda CPF	Padronização Complemento Zero a Esquerda CPF	Complementar com zero a esquerda do CPF quando conteúdo for inferior a 11 dígitos. Quando o CPF possuir quantidade inferior a 11 dígitos, incluir zeros a esquerda completando o número em 11 dígitos
Padronização Completude Sufixo e Prefixo Nome Pessoa Física	Padronização Completude Sufixo e Prefixo Nome Pessoa Física	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 Exemplo: "DR. JOAO DA SILVA JR" -> "JOAO DA SILVA JUNIOR" , o prefixo DR. será movido para o campo de nome de tratamento
Padronização Conteúdo SN	Padronização Conteúdo SN	Padronização informação "Sem Número" escrita de formas distintas. 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"
Padronização Correção Abreviações Bairro	Padronização Correção Abreviações Bairro	Correção de abreviações comuns para nome do bairro. Substituir: . 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 Aplicar rotina QualityStage para completude e padronização da informação do nome do bairro

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 -> gmail gmai -> gmail hgotmail -> hotmail hhotmail -> 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</p> <p>Exemplo: "com.ltda" -> ".com.br" "comm.br" -> ".com.br" ".cvom.br" -> ".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 Óbito Inconsistente	Padronização Data Óbito Inconsistente	<p>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</p>
Padronização Espaçamento de Strings	Padronização Espaçamento de Strings	<p>Define a forma padrão de espaçamento de strings. O excesso de espaçamento deve ser removido.</p> <p>Espaço no início ou no final da string também deve ser removido, exemplo: " JOAO DA SILVA SOARES " -> "JOAO DA SILVA SOARES"</p>
Padronização Formato de Data	Padronização Formato de Data	<p>As datas devem ser armazenadas seguindo um formato padrão. Armazenar no HUB MDM as datas no formato: YYYY-MM-DD HH:MM:SS</p>

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	Não é permitido a existência de determinados caracteres antes e após o @ e caracteres especiais em sequência. Conforme relação abaixo, devemos substituir : @@ por @ -@ por @ @- por @ .@ por @ @. por @ _@ por @ @_ por @ -- por - .. por .
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. Exemplo: Rio de Janeiro Capital - > 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:" Exemplo: "e-mail: joaodasilva@email.com.br" -> "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 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 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>

<p>Padronização Separação Conteúdo Número Logradouro</p>	<p>Padronização Separação Conteúdo Número Logradouro</p>	<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 apósesses caracteres, remover todo oconteúdo da string após o primeironúmero Ex: 123 AP 456 -> 123 permaneceria em número logradouro e AP 456 seria migrado para complemento logradouro</p>
<p>Padronização Separação de E-mail</p>	<p>Padronização Separação de E-mail</p>	<p>Separa as ocorrências de vários e-mails em uma mesma string a partir dos caracteres delimitadores "\>< , ; # - ".</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 -> 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 -> 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 -> BELO HORIZONTE</p> <p> MOJIMIRIM -> 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
-------------------------------	-------------------------------	---

Revision #8

Created 22 July 2022 12:28:20

Updated 13 December 2022 10:40:57