

DEFINING DATA QUALITY METRICS FOR A MASTER DATA MANAGEMENT IMPLEMENTATION

Data Quality Assessment for a Financial Services Company

Mario Fernando Cervi,
Director, Data Quality and CRM Projects, Assesso Engenharia de Sistemas Ltda

In collaboration with Marcelo de Oliveira and Márcio Torelli, Project Managers, Assesso

Abstract. In order to manage customer data in an integrated and unique view, to be shared by all sides of operation, companies have been promoting the establishment of a Master Data Management, bringing together customer master information into a single database structure capable to provide a complete view of the customer and their relationship with the company. This has increased the relevance of Data Quality issues to a high priority level, considering the need for accuracy, timeliness and other aspects, in dealing with customer data. This article presents a method of Data Quality assessment and metrics definition, based on the concepts of the MIT Information Quality Program, and successfully applied in a project carried by Assesso Engenharia de Sistemas Ltda for one of their customers.

BACKGROUND

This paper is based on the experience and knowledge acquired by Assesso, a Brazilian technology enterprise which has been working with Data Quality and Customer Data Integration issues for over 20 years, conducting more than 100 projects.

In the last 7 years, Assesso has been practicing their methodology in the region, with special care of the principles proposed by the Information Quality Program carried by the Massachusetts Institute of Technology – MIT.

The experience described herein was successfully applied in a project for one of Assesso's client, to build a Master Data Management environment, designed to concentrate customer master data to be shared by all company's application systems.

Table of Contents

INTRODUCTION	3
SURVEY WITH THE INVOLVED AREAS	4
DATA QUALITY ASSESSMENT	12
DEFINITION OF DATA QUALITY RULES AND METRICS FOR MDM	22
RECOMMENDATIONS	37
CONCLUSION	38
APPENDIX I – DATA QUALITY ASSESSMENT – EXAMPLES OF DATA VALIDATION AND PROFILING REPORTS	39

INTRODUCTION

In order to allow the appropriated management of the customer master data, a Brazilian financial services company decided to implement a Master Data Management program, integrating all their application systems. The objective is to provide a comprehensive view of the customer and their relationship with the company, as well as to establish a platform to guarantee a good level of quality for the customer information, thus contributing to enhance the customer relationship programs.

An important issue in this project is to define and implement the rules and metrics for the master data, to facilitate monitoring and improvement of data quality levels. This task itself became a project in the MDM project, which will be referred here as the DQ Project.

The DQ Project was carried in four steps:

- A survey with the involved areas to determine the relevant data for each one and the their perception of the quality level;
- A data quality assessment of the related legacy systems;
- The definition of the rules and metrics for data quality to be implemented in the MDM structure; and
- The indication of recommended actions to support the implementation of the program.

The next topics detail the project steps.

SURVEY WITH THE INVOLVED AREAS

The objective of the survey was to understand, from the point-of-view of both management and operations, the specific needs of each department, as well as their perception of the information quality.

The survey was based on 17 interviews and a final consolidation review with IT and Marketing departments. The areas involved were: Corporate Credit, Consumer Credit, Billing, Customer Services, Operations Management and Strategic Planning. The interviews followed the script below:

- Interviews with Management:
 - Understand the role of the area within the company
 - Identification of the information flow
 - Understand the expectations for short and long term
- Interviews with Operations:
 - Detail of the information flow
 - Identification of the relevant master data
 - Identification of the information quality perception
- Consolidation with IT and Marketing
 - Selection of the relevant master data for individual and corporate customers

In this process, each department indicated the relevant master data for their operation. The charts below show the selected master data. Due to specific business characteristics, separated analyses were produced for individual and corporate customers.

Chart 1 – Person Master Data (individual customer)

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	X	X	X		X	X	X
Name	X	X	X	X	X	X	X
Address	X	X	X	X	X	X	X
Zip code	X	X	X	X	X	X	X
Telephone		X	X	X	X	X	X
E-mail			X	X	X	X	X
Gender	X					X	X
Birth date	X	X			X	X	X
Mother's name		X				X	X
Income	X	X				X	X
Profession	X	X				X	X
Job		X				X	X
Marital status	X	X				X	X
Bank Account		X			X	X	X
Bank Account Age		X				X	X

Chart 2 – Company Master Data (corporate customer)

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	X	X	X		X	X	X
Name	X	X	X	X	X	X	X
Address	X		X	X	X	X	X
Zip Code	X	X	X	X	X	X	X
Telephone			X	X	X	X	X
E-mail			X	X	X	X	X
Foundation date	X	X				X	X
Gross income	X	X				X	X
Fleet without onus	X	X				X	X
Fleet with onus	X	X				X	X
Activity code 1	X	X				X	X
Size	X					X	X
Activity code 2			X			X	X
Bank Account	X	X			X	X	X
Bank Account Age	X	X				X	X

In order to assess DQ perception, a selection of the subjective dimensions proposed by the Total Data Quality Management – TDQM, developed by MIT Information Quality Program, was discussed. The chart below shows the data quality dimensions proposed by TDQM:

Chart 3 – Data Quality Dimensions – TDQM:

Category	Dimension	Type
Intrinsic	Accuracy	Objective
	Objectivity	Subjective
	Believability	Subjective
	Reputation	Subjective
Accessibility	Access	Subjective
	Security	Subjective
Contextual	Relevancy	Subjective
	Value-Added	Subjective
	Timeliness	Subjective
	Completeness	Objective
	Amount of data	Objective
	Ease of manipulation	Subjective
Representation	Interpretability	Subjective
	Ease of understanding	Subjective
	Concise representation	Subjective
	Consistent representation	Subjective

To identify the data quality perception in the areas surveyed, four subjective dimensions were selected:

Chart 4 – Selected Subjective Data Quality Dimensions

Category	Dimension
Intrinsic	Reputation
Accessibility	Security
Contextual	Timeliness
	Ease of manipulation

Each person interviewed could classify every master data in the four dimensions with the following scale: very good, good, regular, bad, very bad or no perception (in this case, when the information is not available to them).

It is important to say that the areas use different computer systems, with a considerable variation of technology, whether or not these are state-of-the-art. Each application system has its own customer database with a low level of integration to one another.

The charts below show the information quality perception of the different areas for the dimensions chosen.

Chart 5 – DQ Perception – Dimension: Reputation – Individual Customer

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	-	Good	Good	Good	Very good	Bad	Bad
Name	-	Good	Good	Good	Good	Bad	Bad
Address	-	Regular	Good	Good	Regular	Bad	Bad
Zip code	Good	Regular	Good	Bad	Regular	Bad	Bad
Telephone	-	Regular	Good	Bad	Regular	Bad	Bad
E-mail	-	-	Bad	-	Bad	Bad	Bad
Gender	Good	-	-	-	-	Bad	Bad
Birth date	Good	Good	-	-	Good	Bad	Bad
Mother's name	-	Regular	-	-	-	Bad	Bad
Income	Good	Regular	-	-	-	Bad	Bad
Profession	-	Regular	-	-	-	Bad	Bad
Job	-	Regular	-	-	-	Bad	Bad
Marital status	Good	Regular	-	-	-	Bad	Bad
Bank Account	-	Regular	-	-	Bad	-	-
Bank Account Age	-	Regular	-	Good	Bad	-	-

Chart 6 – DQ Perception – Dimension: Reputation – Corporate Customer

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	-	Good	Good	-	Very good	Bad	Bad
Name	-	Good	Good	Good	Good	Bad	Bad
Address	-	-	Good	Good	Regular	Bad	Bad
Zip Code	Good	-	Good	Good	Regular	Bad	Bad
Telephone	-	-	Good	Bad	Regular	Bad	Bad
E-mail	-	-	Bad	Bad	Bad	Bad	Bad
Foundation date	Good	Regular	-	-	-	Bad	Bad
Gross income	Good	Regular	-	-	-	Bad	Bad
Fleet without onus	-	Regular	-	-	-	Bad	Bad
Fleet with onus	-	Regular	-	-	-	Bad	Bad
Activity code 1	-	Bad	-	-	-	Bad	Bad
Size	-	-	-	-	-	Bad	Bad
Activity code 2	-	-	-	-	-	Bad	Bad
Bank Account	-	Regular	-	-	Bad	-	-
Bank Account Age	-	Regular	-	-	Bad	-	-

Chart 7 – DQ Perception – Dimension: Security – Individual Customer

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	-	Bad	Very good	-	Regular	Bad	Bad
Name	-	Bad	Very good	Good	Regular	Bad	Bad
Address	-	Bad	Very good	Good	Regular	Bad	Bad
Zip code	Good	Bad	Very good	Good	Regular	Bad	Bad
Telephone	-	Bad	Very good	Good	Regular	Bad	Bad
E-mail	-	-	Very good	Bad	Regular	Bad	Bad
Gender	Good	-	-	-	-	Bad	Bad
Birth date	Good	Bad	-	-	Regular	Bad	Bad
Mother's name	-	Bad	-	-	-	Bad	Bad
Income	Good	Bad	-	-	-	Bad	Bad
Profession	-	Bad	-	-	-	Bad	Bad
Job	-	Bad	-	-	-	Bad	Bad
Marital status	Good	Bad	-	-	-	Bad	Bad
Bank Account	-	Bad	-	-	Regular	-	-
Bank Account Age	-	Bad	-	-	Regular	-	-

Chart 8 – DQ Perception – Dimension: Security – Corporate Customer

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	-	Bad	Very good	-	Regular	Bad	Bad
Name	-	Bad	Very good	Good	Regular	Bad	Bad
Address	-	-	Very good	Good	Regular	Bad	Bad
Zip Code	Good	-	Very good	Good	Regular	Bad	Bad
Telephone	-	-	Very good	Good	Regular	Bad	Bad
E-mail	-	-	Very good	Bad	Regular	Bad	Bad
Foundation date	Good	Bad	-	-	-	Bad	Bad
Gross income	Good	Bad	-	-	-	Bad	Bad
Fleet without onus	-	Bad	-	-	-	Bad	Bad
Fleet with onus	-	Bad	-	-	-	Bad	Bad
Activity code 1	-	Bad	-	-	-	Bad	Bad
Size	-	-	-	-	-	Bad	Bad
Activity code 2	-	-	-	-	-	Bad	Bad
Bank Account	-	Bad	-	-	Regular	-	-
Bank Account Age	-	Bad	-	-	Regular	-	-

Chart 9 – DQ Perception – Dimension: Timeliness – Individual Customer

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	-	Good	Good	-	Very good	Bad	Bad
Name	-	Good	Good	Good	Good	Bad	Bad
Address	-	Regular	Good	Good	Regular	Bad	Bad
Zip code	Good	Regular	Good	Good	Regular	Bad	Bad
Telephone	-	Regular	Good	Bad	Regular	Bad	Bad
E-mail	-	-	Bad	Bad	Bad	Bad	Bad
Gender	Good	-	-	-	-	Bad	Bad
Birth date	Good	Good	-	-	Good	Bad	Bad
Mother's name	-	Regular	-	-	-	Bad	Bad
Income	Good	Regular	-	-	-	Bad	Bad
Profession	-	Regular	-	-	-	Bad	Bad
Job	-	Regular	-	-	-	Bad	Bad
Marital status	Good	Regular	-	-	-	Bad	Bad
Bank Account	-	Regular	-	-	Bad	-	-
Bank Account Age	-	Regular	-	-	Bad	-	-

Chart 10 – DQ Perception – Dimension: Timeliness – Corporate Customer

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	-	Good	Good	-	Very good	Bad	Bad
Name	-	Good	Good	Good	Good	Bad	Bad
Address	-	-	Good	Good	Regular	Bad	Bad
Zip Code	Good	-	Good	Good	Regular	Bad	Bad
Telephone	-	-	Good	Bad	Regular	Bad	Bad
E-mail	-	-	Bad	Bad	Bad	Bad	Bad
Foundation date	Good	Regular	-	-	-	Bad	Bad
Gross income	Good	Regular	-	-	-	Bad	Bad
Fleet without onus	-	Regular	-	-	-	Bad	Bad
Fleet with onus	-	Regular	-	-	-	Bad	Bad
Activity code 1	-	Bad	-	-	-	Bad	Bad
Size	-	-	-	-	-	Bad	Bad
Activity code 2	-	-	-	-	-	Bad	Bad
Bank Account	-	Regular	-	-	Bad	-	-
Bank Account Age	-	Regular	-	-	Bad	-	-

Chart 11 – DQ Perception – Dimension: Ease of Manipulation – Individual Customer

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	-	Good	Good	-	Regular	Good	Good
Name	-	Good	Good	Good	Regular	Good	Good
Address	-	Good	Good	Good	Regular	Good	Good
Zip code	Good	Good	Good	Good	Regular	Good	Good
Telephone	-	Good	Good	Good	Regular	Good	Good
E-mail	-	-	Good	Good	Regular	Good	Good
Gender	Good	-	-	-	-	Good	Good
Birth date	Good	Good	-	-	Regular	Good	Good
Mother's name	-	Good	-	-	-	Good	Good
Income	Good	Good	-	-	-	Good	Good
Profession	-	Good	-	-	-	Good	Good
Job	-	Good	-	-	-	Good	Good
Marital status	Good	Good	-	-	-	Good	Good
Bank Account	-	Good	-	-	Regular	-	-
Bank Account Age	-	Good	-	-	Regular	-	-

Chart 12 – DQ Perception – Dimension: Ease of Manipulation – Corporate Customer

Information	Corporate Credit	Consumer Credit	Billing	Customer Services	Operations Management	Strategic Planning	Marketing
Income tax id code	-	Good	Good	-	Regular	Good	Good
Name	-	Good	Good	Good	Regular	Good	Good
Address	-	-	Good	Good	Regular	Good	Good
Zip Code	Good	-	Good	Good	Regular	Good	Good
Telephone	-	-	Good	Good	Regular	Good	Good
E-mail	-	-	Good	Good	Regular	Good	Good
Foundation date	Good	Good	-	-	-	Good	Good
Gross income	Good	Good	-	-	-	Good	Good
Fleet without onus	-	Good	-	-	-	Good	Good
Fleet with onus	-	Good	-	-	-	Good	Good
Activity code 1	-	Good	-	-	-	Good	Good
Size	-	-	-	-	-	Good	Good
Activity code 2	-	-	-	-	-	Good	Good
Bank Account	-	Good	-	-	Regular	-	-
Bank Account Age	-	Good	-	-	Regular	-	-

The following observations can be highlighted:

- Most areas consider Reputation and Timeliness good or regular, except for Marketing and Strategic Planning.
- There is a variation in the perception of Security, allegedly due to the variation of technology and the difficulty in changing old application systems.
- Ease of manipulation is a positive characteristic for most users.
- Except for Ease of manipulation, Marketing and Strategic Planning classified as bad the other dimensions. It must be noted that these areas are indirect users of the data coming from the application systems (all other areas have made their analysis based on their day-by-day operational systems).

DATA QUALITY ASSESSMENT

The objective of this step was to assess the quality of information identified as relevant in the legacy systems databases.

The result of this data profiling process was the basis for specifying the rules and objective metrics for data quality in the MDM structure being designed. The metrics used in the diagnosis were not exactly the same as suggested for the MDM, since they were introduced in the context of the different application systems.

The assessment included data from six application systems which will feed the MDM. For each data source, the following types of data validation were applied:

- Content investigation (ABC curve), to check for completeness and identify suspect repetition of values;
- Type of data, domain, interval, check digit validation;
- Investigation of suspect content in name;
- Address validation, using the Brazilian Post Office master file;
- Telephone validation, checking prefix vs. area code, area code vs. ZIP code;
- E-mail address validation, checking syntax and frequent misspelling.

Besides the validation rules applicable to each attribute, as listed above, a combination of each of them was generated to produce a quality category, using to the “RYG Method” (Red, Yellow, Green), as follows:

- ★ Red – Bad quality level. Data should not be used. Corrective actions must be taken.
- ★ Yellow – Suspect quality level. Possible risk in using the data. Plan corrective actions (e. g. at receptive contact).
- ★ Green – Good quality level. No action is required.

The chart below shows the RYG rules defined for each attribute in the DQ assessment. The same RYG rules were applied for all legacy systems.

Chart 13 – RYG Rules for Individual Customer







Person Data				
Income Tax Id Code	Present AND Valid check digit AND No excessive repetition	–		Not present OR Invalid check digit OR Excessive repetition
Name	Present AND Valid (according to Brazilian standards)		Present AND Suspect content	Not present
Address	Present AND Address confirmed		Present AND Confirmed with correction	Not present OR Unrecognized
Telephone	Present AND Area code and prefix confirmed AND No excessive repetition		Present AND (Area code or prefix corrected) AND No excessive repetition	Not present OR (Area code or prefix unrecognized) OR Excessive repetition
Email	Present AND No syntax errors AND No domain correction		Present AND No syntax errors AND Domain corrected	Not present OR Syntax errors
Gender	Present AND Valid value AND Compatible with name		Present AND Valid value AND Incompatible with name	Not present OR Invalid value
Birth date	Present AND Valid date AND No excessive repetition AND Before today AND Age between 18 and 100		Present AND Valid date AND No excessive repetition AND Before today AND Age over 100	Not present OR Invalid OR Excessive repetition OR After today OR Age under 18
Mother's Name	Present AND Valid (according to Brazilian standards)		Present AND Suspect content	Not present
Income	Present AND (Value between 100 and 1,000,000 OR = 0)	–		Not present OR (Value < 100 or > 1,000,000) AND Value <> 0
Profession	Present AND Valid value AND Value <> "Other"	–		Not present OR Invalid Value OR Value = "Other"
Job	Present AND Valid value AND Value <> "Other"	–		Not present OR Invalid Value OR Value = "Other"
Marital status	Present AND Valid value AND Value <> "Other"	–		Not present OR Invalid Value OR Value = "Other"
Bank Account	Present AND Valid value	–		Not present OR Invalid value
Bank Account Age	Present AND Valid month AND Year >= 1900	–		Not present OR Invalid month OR Invalid Year OR Year < 1900

Chart 14 – RYG Rules for Corporate Customer

Company Data			
Income Tax Id Code	Present AND Valid check digit AND No excessive repetition	–	Not present OR Invalid check digit OR Excessive repetition
Name	Present AND Valid (according to Brazilian standards)	Present AND Suspect content	Not present
Address	Present AND Address confirmed	Present AND Confirmed with correction	Not present OR Unrecognized
Telephone	Present AND Area code and prefix confirmed AND No excessive repetition	Present AND (Area code or prefix corrected) AND No excessive repetition	Not present OR (Area code or prefix unrecognized) OR Excessive repetition
Email	Present AND No syntax errors AND No domain correction	Present AND No syntax errors AND Domain corrected	Not present OR Syntax errors
Foundation date	Present AND Valid date AND No excessive repetition AND Before today AND Age between 18 and 100	Present AND Valid date AND No excessive repetition AND Before today AND Age over 100	Not present OR Invalid OR Excessive repetition OR After today OR Age under 18
Income	Present AND Value >= 1,000	–	Not present OR Value < 1,000
Fleet without onus	Present AND Value > 0		Not present OR Value <= 0
Fleet with onus	Present AND Value > 0		Not present OR Value <= 0
Activity code 1	Present AND Valid value	–	Not present OR Invalid Value
Size	Present AND Valid value	–	Not present OR Invalid Value
Bank Account	Present AND Valid value	–	Not present OR Invalid value
Bank Account Age	Present AND Valid month AND Year >= 1900	–	Not present OR Invalid month OR Invalid Year OR Year < 1900

The following charts show the distribution of the RYG rules applied to all records of each legacy system.

Chart 15 – Legacy System A – RYG Rules Distribution for Individual Customer




Person Data			
Income Tax Id Code	99,9 %	-	0,1 %
Name	99,8 %	0,2 %	-
Address	65,5 %	20,4 %	14,1 %
Telephone	90,9 %	7,2 %	1,9 %
Email	3,0 %	0,1 %	96,9 %
Gender	87,9 %	12,0 %	0,1 %
Birth date	99,9 %	0,01 %	0,99 %
Mother's Name	88,0 %	0,6 %	11,4 %
Income	97,5 %	-	2,5 %
Profession	64,1 %	-	35,9 %
Job	54,9 %	-	45,1 %
Marital status	93,7 %	-	6,3 %
Bank Account	5,9 %	-	94,1 %
Bank Account Age	63,4 %	-	36,6 %

Chart 16 – Legacy System A – RYG Rules Distribution for Corporate Customer




Company Data			
Income Tax Id Code	99,9 %	-	0,1 %
Name	99,7 %	0,3 %	-
Address	72,0 %	15,1 %	12,9 %
Telephone	89,9 %	10,1 %	2,0 %
Email	9,2 %	0,1 %	90,7 %
Foundation date	96,1 %	0,3 %	3,6 %
Income	62,7 %	-	37,3 %
Fleet without onus	100 %	-	-
Fleet without onus	100 %	-	-
Activity code 1	87,6 %	-	12,4 %
Size	100 %	-	-
Bank Account	5,8 %	-	94,2 %
Bank Account Age	54,7 %	-	45,3 %

Chart 17 – Legacy System B – RYG Rules Distribution for Individual Customer




Person Data			
Income Tax Id Code	99,7 %	-	0,3 %
Name	99,7 %	0,3 %	-
Address	55,1 %	24,3 %	20,6 %
Telephone	29,4 %	26,0 %	44,6 %
Email	1,6 %	0,01 %	98,39 %
Gender	75,8 %	10,4 %	13,8 %
Birth date	58,1 %	0,01 %	41,89 %
Mother's Name	-	-	-
Income	95,1 %	-	4,9 %
Profession	4,4 %	-	95,6 %
Job	-	-	-
Marital status	83,3 %	-	16,7 %
Bank Account	-	-	-
Bank Account Age	-	-	-

Chart 18 – Legacy System B – RYG Rules Distribution for Corporate Customer




Company Data			
Income Tax Id Code	99,3 %	-	0,7 %
Name	99,4 %	0,6 %	-
Address	59,1 %	24,3 %	16,6 %
Telephone	37,6 %	44,4 %	18,0 %
Email	-	-	100 %
Foundation date	-	-	-
Income	-	-	-
Fleet without onus	-	-	-
Fleet without onus	-	-	-
Activity code 1	-	-	-
Size	-	-	-
Bank Account	-	-	-
Bank Account Age	-	-	-

Chart 19 – Legacy System C – RYG Rules Distribution for Individual Customer




Person Data			
Income Tax Id Code	94,9 %	-	5,1 %
Name	99,7 %	0,3 %	-
Address	26,9 %	56,4 %	16,7
Telephone	48,3 %	44,8 %	6,9 %
Email	1,8 %	0,01 %	98,19 %
Gender	84,6 %	11,6 %	3,8 %
Birth date	95,5 %	0,01 %	4,49 %
Mother's Name	36,2 %	0,2 %	63,6 %
Income	99,7 %	-	0,3 %
Profession	13,1 %	-	86,9 %
Job	-	-	-
Marital status	92,9 %	-	7,1 %
Bank Account	2,8 %	-	97,2 %
Bank Account Age	-	-	-

Chart 20 – Legacy System C – RYG Rules Distribution for Corporate Customer




Company Data			
Income Tax Id Code	99,1 %	-	0,9 %
Name	99,5 %	0,5 %	-
Address	61,2 %	22,6 %	16,2 %
Telephone	40,5 %	38,7 %	20,8 %
Email	2,89 %	0,01 %	97,1 %
Foundation date	22,8 %	-	77,2 %
Income	-	-	-
Fleet without onus	-	-	-
Fleet without onus	-	-	-
Activity code 1	27,5 %	-	72,5 %
Size	50,8 %	-	49,2 %
Bank Account	2,4 %	-	97,6 %
Bank Account Age	-	-	-

Chart 21 – Legacy System D – RYG Rules Distribution for Individual Customer




Person Data			
Income Tax Id Code	99,5 %	-	0,5 %
Name	99,98 %	0,1 %	0,01 %
Address	42,1 %	30,9 %	27 %
Telephone	22,3 %	61,5 %	16,3 %
Email	2,1 %	0,01 %	97,89 %
Gender	90,6 %	9,4 %	-
Birth date	92,3 %	0,1 %	7,6 %
Mother's Name	-	-	-
Income	-	-	-
Profession	-	-	-
Job	-	-	-
Marital status	92,3 %	-	7,7 %
Bank Account	-	-	-
Bank Account Age	-	-	-

Chart 22 – Legacy System D – RYG Rules Distribution for Corporate Customer




Company Data			
Income Tax Id Code	97,1 %	-	2,9 %
Name	99,6 %	0,4 %	-
Address	36,9 %	37,2 %	25,9 %
Telephone	32,2 %	51,7 %	16,1
Email	-	-	-
Foundation date	-	-	-
Income	-	-	-
Fleet without onus	-	-	-
Fleet without onus	-	-	-
Activity code 1	-	-	-
Size	-	-	-
Bank Account	-	-	-
Bank Account Age	-	-	-

Chart 23 – Legacy System E – RYG Rules Distribution for Individual Customer




Person Data			
Income Tax Id Code	100 %	-	-
Name	99,9 %	0,1 %	-
Address	55,2 %	23,0 %	21,8 %
Telephone	90,1 %	4,3 %	5,6 %
Email	70,8 %	1,6 %	27,6 %
Gender	71,6 %	28,4 %	-
Birth date	35,3 %	-	64,7 %
Mother's Name	30,4 %	0,1 %	69,5 %
Income	92,4 %	-	7,6 %
Profession	-	-	-
Job	24,6 %	-	75,4 %
Marital status	32,4 %	-	67,6 %
Bank Account	-	-	-
Bank Account Age	-	-	-

Chart 24 – Legacy System E – RYG Rules Distribution for Corporate Customer




Company Data			
Income Tax Id Code	99,4 %	-	0,6%
Name	99,5 %	0,5 %	-
Address	44,6 %	18,3 %	37,1 %
Telephone	86,1 %	11,8 %	2,1 %
Email	70,8 %	1,6 %	27,6 %
Foundation date	94,9 %	-	5,1 %
Income	-	-	-
Fleet without onus	-	-	-
Fleet without onus	-	-	-
Activity code 1	40,7 %	-	59,3 %
Size	100 %	-	-
Bank Account	-	-	-
Bank Account Age	-	-	-

Chart 25 – Legacy System F – RYG Rules Distribution for Individual Customer







Person Data			
Income Tax Id Code	99,99 %	-	0,01 %
Name	99,9 %	0,1%	-
Address	63,2 %	19,4 %	17,4 %
Telephone	74,0 %	20,0 %	6,0 %
Email	0,7 %	0,01 %	99,29 %
Gender	-	-	-
Birth date	76,7 %	0,1 %	23,2 %
Mother's Name	0,2 %	-	99,8 %
Income	99,99 %	-	0,01 %
Profession	-	-	-
Job	-	-	-
Marital status	-	-	-
Bank Account	-	-	-
Bank Account Age	-	-	-

Chart 26 – Legacy System F – RYG Rules Distribution for Corporate Customer

Company Data			
Income Tax Id Code	99,99 %	-	0,01 %
Name	99,8 %	0,2 %	-
Address	64,7 %	17,0 %	18,3 %
Telephone	75,5 %	18,9 %	5,6 %
Email	1,8 %	0,01 %	98,19 %
Foundation date	99,2 %	-	0,8 %
Income	33,1 %	0,1 %	66,8 %
Fleet without onus	-	-	-
Fleet without onus	-	-	-
Activity code 1	99,8 %	-	0,2 %
Size	-	-	-
Bank Account	-	-	-
Bank Account Age	-	-	-

Besides the data validation tools, the assessment also included entity resolution processes to determine the level of duplicates among household addresses, people and companies.

The deduplication processes have tested a variety of matching keys and scoring configuration, combining master data as name, income tax id code, birth date, address, telephone number and e-mail address. The configuration was refined in a series of test cycles, in accordance with the TDQM method (Define, Measure, Plan, Analyze). The charts below show the level of duplicates found in each legacy system and in the consolidated inter-system view.

Chart 27 – Duplicate Level in the Legacy Systems:

Entity	System A	System B	System C	System D	System E	System F
Person	25%	4%	1%	1%	0%	0,1%
Company	76%	5%	1%	6%	7%	0%
Household	33%	46%	35%	26%	6%	13%

Chart 28 – Duplicate Level Overall:

Entity	Duplicate Level
Person	57%
Company	74%
Household	62%

Appendix I shows examples of data validation and the various reports produced in this step of the project.

DEFINITION OF DATA QUALITY RULES AND METRICS FOR MDM

Data Quality rules must consider the multiple aspects related to each piece of information. Some rules consider only the piece of information itself. Other rules require a cross-validation between two or more attributes of a customer record. There are still rules that involve two or more records of the same customer.

For this reason, the definition of objective Data Quality rules and metrics can be divided into three levels:

- Level 1 – Per information (a single attribute, e.g. birth date, or a set of attributes, e.g. address):
 - Presence
 - Age (when the information was collected)
 - Domain integrity
 - Column integrity
 - Entity integrity
 - Business rules (or user-defined integrity) applicable to the single information

- Level 2 – Inter-attributes validation, relating two or more attributes in a customer record:
 - Referential integrity
 - Cardinality (e. g., minimum or maximum addresses per customer)
 - Business rules involving two or more attributes (e. g., birth date vs. contract start date)

- Level 3 – Inter-record rules, relating different records of the same customer:
 - Entity resolution (or de-duplication) rules
 - Merge & purge rules

The chart below shows an example of rules and metrics in level 1:

Chart 29 – Example of Rules and Metrics Defined in Level 1 for Document Number:

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Check Digit	M3 – Valid Check Digit	Yes / No
Check Repetition Level	M4 – Too Many Repetitions	Yes / No

For level 1, each rule corresponds to an individual metric. Additionally, a new metric may be created combining the individual metrics, to produce a data quality score for each attribute, varying from 0 to 10. In this way, the average of the combined metric defines the level of quality for each attribute in the database. The chart below shows the combined metric defined for the example above.

Chart 30 – Example of Combined Metric for Document Number:

Combined Metric	Score
If M3 = Yes and M4 = No	10
If M3 = Yes and M4 = Yes	5
If M1 = No or M3 = No	0

Some type of information changes as time goes by, for instance, the customer address. In this case, it is recommended to make an adjustment of the combined Data Quality metric to reflect the risk of obsolescence of the information. Such adjustment can be made using a reduction factor as shown in the example below:

Chart 31 – Example of Age-based Reduction Factor:

Age of Data	Reduction Factor
Less than 1 year	* 1
From 1 to 3 years	* 0,8
From 4 to 6 years	* 0,7
From 7 to 10 years	* 0,6
More than 10 years	* 0,5

In level 2, rules and metrics are defined for a combination of two or more attributes of the same customer record. The chart below shows an example of rules and metrics defined in level 2:

Chart 32 – Example of Rules and Metrics Defined in Level 2 for a Person:

Rule	Metric	Result
Validate minimum age on first relationship date	M1 – Minimum age OK	Yes / No
Check compatibility of name and gender	M2 – Name and gender compatible	Yes / No
Compare address location and telephone area code location	M5 – Distance from address and area code locations	<= 100 Km > 100 Km

In the course of the project, this method was adopted to define the metrics and rules for the MDM. The next charts show all level 1 and 2 metrics defined in this project.

It is important to emphasize that entity and referential integrity rules are not in the context of this project. They will be defined in the DBMS (Data Base Management System) context.

LEVEL 1

Chart 33 – Level 1 Rules and Metrics for Income Tax Id Code

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Check Digit	M3 – Valid Check Digit	Yes / No
Check Repetition Level	M4 – Too many repetitions	Yes / No
Check Info in Data Supplier 1	M5 – Checked Data Supplier 1	Yes / No
Check Info in Data Supplier 2	M6 – Checked Data Supplier 2	Yes / No

Note: although fifth and sixth rules above use information from other source (Data Supplier files), they are considered level 1 in this context, because there is a regular process to aggregate the check indicator to the customer record.

Combined Metric	Score
If M5 = Yes or M6 = Yes	10
If M3 = Yes and M4 = No	9
If M3 = Yes and M4 = Yes	2
If M1 = No or M3 = No	0

Chart 34 – Level 1 Rules and Metrics for Name

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Name Content	M3 – Name Validation Ret Code	Valid; Suspect; Invalid
Check Info in Data Supplier 1	M4 – Checked Data Supplier 1	Yes / No
Check Info in Data Supplier 2	M5 – Checked Data Supplier 2	Yes / No

Combined Metric	Score
If M4 = Yes or M5 = Yes	10
If M3 = Valid	9
If M3 = Suspect	5
If M3 = Invalid	2
If M1 = No	0

Chart 35 – Level 1 Rules and Metrics for Address

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate address with Post Office Master File	M3 – Address Validation Ret Code	Confirmed; Confirmed with small correction; Confirmed with big correction; Unrecognized
Check Info in Data Supplier 1	M4 – Checked Data Supplier 1	Yes / No
Check Info in Data Supplier 2	M5 – Checked Data Supplier 2	Yes / No

Combined Metric	Score
If M3 = Confirmed and (M4 = Yes or M5 = Yes)	10
If M3 = Confirmed or M4 = Yes or M5 = Yes	9
If M3 = Confirmed with small correction	8
If M3 = Confirmed with big correction	5
If M3 = Unrecognized	2
If M1 = No	0

A reduction factor shall be applied to the score, according to the Age (M2):

Age (M2)	Reduction Factor for Person	Reduction Factor for Company
Less than 1 year	* 1	* 1
From 1 to 3 years	* 0,8	* 1
From 4 to 6 years	* 0,7	* 1
From 7 to 10 years	* 0,6	* 0,9
More than 10 years	* 0,5	* 0,8

Chart 36 – Level 1 Rules and Metrics for Telephone Number

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Area Code and Prefix with Official Master Files	M3 – Telephone Validation Ret Code	Confirmed; Confirmed with correction; Unrecognized; Cell phone Toll free number
Check Contact Result	M4 – Contact Ret Code	Confirmed (customer contacted); Suspect (contact not made); Incorrect (telephone belongs to other person); No contact attempted

Combined Metric	Score
If M4 = Confirmed	10
If M4 = (Suspect or No contact) and M3 = Confirmed	9
If M4 = (Suspect or No contact) and M3 = Confirmed with Correction	7
If M4 = (Suspect or No contact) and M3 = Cell phone	6
If M4 = (Suspect or No contact) and M3 = Toll free number	2
If M1 = No or M4 = Incorrect	0

A reduction factor shall be applied to the score, according to the Age (M2):

Age (M2)	Reduction Factor for Person	Reduction Factor for Company
Less than 1 year	* 1	* 1
From 1 to 3 years	* 0,8	* 1
From 4 to 6 years	* 0,6	* 0,8
From 7 to 10 years	* 0,4	* 0,5
More than 10 years	* 0,2	* 0,2

Chart 37 – Level 1 Rules and Metrics for E-mail Address

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Syntax	M3 – E-mail Validation Ret Code	Valid; Domain corrected; Invalid
Check Contact Result	M4 – Contact Ret Code	Confirmed (customer contacted); Suspect (no response); Incorrect (answer from server or other person); No contact attempted

Combined Metric	Score
If M4 = Confirmed	10
If M4 = (Suspect or No contact) and M3 = (Valid or Domain Corrected)	8
If M1 = No or M3 = Invalid or M4 = Incorrect	0

A reduction factor shall be applied to the score, according to the Age (M2):

Age (M2)	Reduction Factor for Person	Reduction Factor for Company
Less than 1 year	* 1	* 1
From 1 to 3 years	* 0,8	* 1
From 4 to 6 years	* 0,6	* 0,8
From 7 to 10 years	* 0,4	* 0,5
More than 10 years	* 0,2	* 0,2

Chart 38 – Level 1 Rules and Metrics for Gender

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Content (Domain)	M3 – Valid Value	Yes / No

Combined Metric	Score
If M3 = Yes	10
If M1 = No or M3 = No	0

Chart 39 – Level 1 Rules and Metrics for Birth and Foundation Date

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Syntax / Existence	M3 – Valid Date	Yes / No
Validate Interval	M4 – Valid Age	Yes / No
Validate Repetition Level	M5 – Suspect Date	Yes / No

Combined Metric	Score
If M4 = Yes and M5 = No	10
If M4 = Yes and M5 = Yes	5
If M3 = Yes and M4 = No	2
If M1 = No or M3 = No	0

Chart 40 – Level 1 Rules and Metrics for Mother’s Name

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Name Content	M3 – Name Validation Ret Code	Valid; Suspect; Invalid
Check Name’s Gender	M4 – Name’s Gender	Male Female Undefined (name used for both)

Combined Metric	Score
If M3 = Valid and M4 = Female	10
If M3 = Valid and M4 = Undefined	9
If M3 = Valid and M4 = Male	7
If M3 = Suspect	4
If M3 = Invalid	2
If M1 = No	0

Chart 41 – Level 1 Rules and Metrics for Income / Gross Income / Fleet With or Without Onus

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Value >= 0	M3 – Value >= 0	Yes / No
Validate Interval	M4 – Value within Interval	Yes / No

Combined Metric	Score
If M4 = Yes	10
If M4 = No and M3 = Yes	5
If M1 = No or M3 = No	0

A reduction factor shall be applied to the score, according to the Age (M2):

Age (M2)	Reduction Factor for Person	Reduction Factor for Company
Less than 1 year	* 1	* 1
From 1 to 3 years	* 0,8	* 1
From 4 to 6 years	* 0,6	* 0,8
From 7 to 10 years	* 0,4	* 0,5
More than 10 years	* 0,2	* 0,2

Chart 42 – Level 1 Rules and Metrics for Profession / Job / Marital Status

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Content (Domain)	M3 – Valid Value	Yes / No
Check for “Other”	M4 – Value = “Other”	Yes / No

Combined Metric	Score
If M3 = Yes and M4 = No	10
If M4 = No and M3 = Yes	5
If M1 = No or M3 = No or M4 = Yes	0

A reduction factor shall be applied to the score, according to the Age (M2) for Job and Marital Status (does not apply to Profession):

Age (M2)	Reduction Factor
Less than 1 year	* 1
From 1 to 3 years	* 0,8
From 4 to 6 years	* 0,6
From 7 to 10 years	* 0,4
More than 10 years	* 0,2

Chart 43 – Level 1 Rules and Metrics for Activity Code and Size

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Content (Domain)	M3 – Valid Value	Yes / No
Check for “Other”	M4 – Value = “Other”	Yes / No

Combined Metric	Score
If M3 = Yes and M4 = No	10
If M4 = No and M3 = Yes	5
If M1 = No or M3 = No or M4 = Yes	0

A reduction factor shall be applied to the score, according to the Age (M2):

Age (M2)	Reduction Factor
Less than 1 year	* 1
From 1 to 3 years	* 1
From 4 to 6 years	* 0,8
From 7 to 10 years	* 0,7
More than 10 years	* 0,6

Chart 44 – Level 1 Rules and Metrics for Bank Account (Bank and Branch):

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Content (Domain)	M3 – Valid Value	Yes / No

Combined Metric	Score
If M3 = Yes	10
If M1 = No or M3 = No	0

A reduction factor shall be applied to the score, according to the Age (M2):

Age (M2)	Reduction Factor
Less than 1 year	* 1
From 1 to 3 years	* 1
From 4 to 6 years	* 0,8
From 7 to 10 years	* 0,7
More than 10 years	* 0,6

Chart 45 – Level 1 Rules and Metrics for Bank Account Age (Month / Year):

Rule	Metric	Result
Validate Presence (Not Null)	M1 – Present	Yes / No
Check Last Update Date	M2 – Age of Last Update	Age
Validate Syntax / Existence	M3 – Valid Date	Yes / No
Validate Interval	M4 – Valid Age	Yes / No

Combined Metric	Score
If M4 = Yes	10
If M3 = Yes	5
If M1 = No or M3 = No	0

LEVEL 2

Chart 46 – Level 2 Rules and Metrics for Individual Customers

Rule	Metric	Result
Validate minimum customer age vs. first relationship date	M1 – Valid age	Yes / No
Validate customer name vs. gender	M2 – Name and gender compatible	Yes / No
Check customer's number of addresses	M3 – Customer with no addresses	Yes / No
Check customer's number of telephones	M4 – Customer with no telephones	Yes / No
Check distance from telephone area code to address	M5 – Distance	<= 100 Km > 100 Km

Chart 47 – Level 2 Rules and Metrics for Corporate Customers

Rule	Metric	Result
Validate foundation date vs. first relationship date	M1 – Foundation prior to relationship	Yes / No
Validate size vs. gross income	M2 – Size and gross income coherent	Yes / No
Check customer's number of addresses	M3 – Customer with no addresses	Yes / No
Check customer's number of telephones	M4 – Customer with no telephones	Yes / No
Check distance from telephone area code to address	M5 – Distance	<= 100 Km > 100 Km

As a method to establish the level of quality, as well as the goals to be achieved, the average of the combined metrics will be calculated for each attribute in MDM database, to be classified according to the “RYG Method” (Red, Yellow, Green).

The charts below show the RYG ranges (combined metric categories) defined for each attribute in the MDM database.

Chart 48 – Combined Data Quality Metric Categories for Individual Customers:







Person Information			
Income tax id code	From 9 to 10	From 7 to 8,9	Less than 7
Name	From 9 to 10	From 7 to 8,9	Less than 7
Address	From 7 to 10	From 5 to 6,9	Less than 5
Telephone number	From 7 to 10	From 5 to 6,9	Less than 5
E-mail address	From 7 to 10	From 5 to 6,9	Less than 5
Gender	From 9 to 10	From 7 to 8,9	Less than 7
Birth date	From 9 to 10	From 7 to 8,9	Less than 7
Mother's name	From 9 to 10	From 7 to 8,9	Less than 7
Income	From 7 to 10	From 5 to 6,9	Less than 5
Profession	From 9 to 10	From 7 to 8,9	Less than 7
Job	From 7 to 10	From 5 to 6,9	Less than 5
Marital status	From 7 to 10	From 5 to 6,9	Less than 5
Bank, branch, account	From 7 to 10	From 5 to 6,9	Less than 5
Account age	From 7 to 10	From 5 to 6,9	Less than 5

Chart 49 – Combined Data Quality Metric Categories for Corporate Customers:

Company Information			
Income tax id code	From 9 to 10	From 7 to 8,9	Less than 7
Company name	From 9 to 10	From 7 to 8,9	Less than 7
Address	From 7 to 10	From 5 to 6,9	Less than 5
Telephone number	From 7 to 10	From 5 to 6,9	Less than 5
E-mail address	From 7 to 10	From 5 to 6,9	Less than 5
Date of foundation	From 9 to 10	From 7 to 8,9	Less than 7
Gross income	From 7 to 10	From 5 to 6,9	Less than 5
Fleet without Onus	From 7 to 10	From 5 to 6,9	Less than 5
Fleet with Onus	From 7 to 10	From 5 to 6,9	Less than 5
Activity code 1	From 9 to 10	From 7 to 8,9	Less than 7
Size	From 9 to 10	From 7 to 8,9	Less than 7
Activity code 2	From 9 to 10	From 7 to 8,9	Less than 7
Bank, branch, account	From 7 to 10	From 5 to 6,9	Less than 5
Account age	From 7 to 10	From 5 to 6,9	Less than 5
Partner join date	From 9 to 10	From 7 to 8,9	Less than 7
% Partner share	From 7 to 10	From 5 to 6,9	Less than 5

LEVEL 3

At level 3, the entity resolution (deduplication) rules were recommended after a series of test cycles using the legacy systems data. The recommendation is presented in the chart below.

Chart 50 – Level 3 – Entity Resolution Criteria:

Entity	Criteria
Household	<ul style="list-style-type: none"> • Street name phonetically similar • Address number identical • Address complement identical (apt, floor, etc.) • ZIP code identical (5 first digits)
	<ul style="list-style-type: none"> • Name phonetically similar • Income tax id identical
	OR
Person	<ul style="list-style-type: none"> • Name phonetically similar • Birth date identical • ZIP code identical (3 first digits)
	OR
	<ul style="list-style-type: none"> • Name phonetically similar • Street name phonetically similar • Address number identical • Address complement identical (apt, floor, etc.) • ZIP code identical (5 first digits)
Company	<ul style="list-style-type: none"> • Name phonetically similar • Income tax id identical

Still at level 3, the merge & purge rules were recommended considering the following aspects:

For data stored as a single occurrence (e. g. birth date), the priority in the merge & purge process shall consider:

- Reputation of the sources
- Integrity (objective metrics)
- Recency of each occurrence

For attributes with multiple occurrences (e. g. address), all shall be kept whenever storage availability allows. In the need of discarding occurrences, the same recommendation above applies.

The charts below show the priority for each attribute, based on reputation. Complementarily, integrity and recency must be checked for each piece of data.

Chart 51 – Level 3 – Merge & Purge Priority Based on Reputation – Individual Customers:

Information	System A	System B	System C	System D	System E	System F
Income tax id code	1	1	1	1	1	1
Name	1	1	1	1	1	1
Gender	1	2	1	1	2	-
Birth date	1	4	2	2	5	3
Mother's name	1	-	2	-	3	-
Profession	1	-	-	-	-	-
Job	1	-	-	-	2	-
Marital status	1	2	1	1	3	-

Chart 52 – Level 3 – Merge & Purge Priority Based on Reputation – Corporate Customers:

Information	System A	System B	System C	System D	System E	System F
Income tax id code	1	1	1	1	1	1
Name	1	1	1	1	1	1
Foundation date	1	-	2	-	1	-
Activity code 1	1	-	2	-	-	-
Size	1	-	-	-	1	-

Note: in the charts above, the lowest value corresponds to the highest priority.

RECOMMENDATIONS

In addition to implementing data quality rules and metrics, it is recommended to adopt complementary actions which will help to achieve a good level of information quality in the company.

Adopt a DQ Management Policy

The adoption of a data quality management policy is a critical factor for the success of the project in long term. Such policy demands:

- Centralized coordination of efforts
- Participation of all departments
- Clear definition of roles and responsibilities
- Education and training in all levels

A suggested step-by-step method to implement the data quality management is:

- Find a critical DQ problem for the company to motivate the whole team
- Identify the stakeholders and produce a DQ assessment with them
- Create a Data Quality Group
- Define the Information Product Manager
- Create corrective actions
- Create preventive actions
- Establish monitoring processes, with DQ metrics and goals for both subjective and objective dimensions

This must be an unending policy and have the commitment of top management.

Enhance Data Collection and Enrichment

Most data quality problems can be avoided by implementing basic data validation at collection time or using external sources to check and update internal databases. Examples of good practices here are:

- Give special attention to customer contact data: name, address, telephone number and e-mail address.
- Offer alternatives to address validation, e.g.: search street and city using ZIP code; or search ZIP code using street and city names.
- Use adequate level of validation severity: inhibit the transaction conclusion when strictly necessary. Consider a back-office operation to do the rest.
- Prevent data duplication: identify duplicates whenever possible (households or customers).
- Use data acquired by other processes (e. g. Credit Bureau information) to validate or update customer contact data.

CONCLUSION

Data assessment and definition of rules and metrics are fundamental to create a data quality program in every company. They constitute the mechanism to understand the problems and weaknesses, allowing the adoption of direct and correct actions to avoid them.

However, the basis for this program is the commitment of top management and the involvement of all areas. Data quality culture and care is mandatory. Only this can guarantee the resources required to make the program permanent and effective.

APPENDIX I – DATA QUALITY ASSESSMENT – EXAMPLES OF DATA VALIDATION AND PROFILING REPORTS

The data quality assessment on the six application databases that will be source to the MDM produced a series of reports for each source. The reports below are representative of the various types of report produced. The assessment used the data quality software DataCare®, a trade mark of Assesso Engenharia de Sistemas Ltda.

EXAMPLES OF DATA VALIDATION

Chart I-1 – Examples of Name Validation:

PERSON NAME	VALIDATION
TESTE CLIENTE	Suspect words (TESTE=test, CLIENTE=customer)
SONHO JOSE DE SOUZA	Suspect words (SONHO=dream)
VICTORIO VICTORIO	Just two identical words
FRANCISCO FERREIRA DOS SANTOS	Three consecutive equal letters (RRR)
GBDGHF FSPRT	No vowels
JOSE	Just one word
LOURIVAL D1 ALMEIDA	Numbers (1)
VANESSA ALVES S	Last name abbreviated
CARLOS SANTOS; SILVA	Special characters (;)

Chart I-2 – Examples of Address Validation:

REC	ADDRESS	DISTRICT	CITY	ST	Zip Cd	VALIDATION
IN	Rua Josefina Cince Ragaini 39	Itaim Paulista	S. Paulo	SP	08140060	District, zip code corrected; address standardized
OUT	R Josefina Cince Ragaini 39	Vila Morgadouro	São Paulo	SP	08140260	
IN	Rua Fernando Moreira 1293	Chácara	Santo Antonio	SP	04716003	District and city corrected and street type standardized
OUT	R Fernandes Moreira 1293	Chácara Santo Antonio	São Paulo	SP	04716003	
IN	Avenida Cláudio Dantas 300 apto 10	Vila Mariana	Salvador	BA	51300000	Unidentified street; zip code does not belong to the city
OUT	Av Cláudio Dantas 300 ap 10	Vila Mariana	Salvador	BA	51300000	

Chart I-3 – Examples of E-mail Address Validation:

INPUT	OUTPUT	VALIDATION
CARLOS MARCONDES@BOL.COM.BR	carlos marcondes@bol.com.br	Contains spaces
ANA.BARTIRA@GRANDCORP.COM;BR	ana.bartira@grandcorp.com;br	Invalid character (;)
FELIX_CRUZ@HOTMAIL.COM.BR	felix_cruz@hotmail.com	Domain corrected
GALMEIDA@GMAIL.COM.BR	galmeida@gmail.com	Domain corrected
MARIACARVALHO@ZAZ.COM.BR	mariacarvalho@terra.com.br	Domain corrected
MARCOS.PORTELLA@TERR	marco.portella@terr	Incomplete domain
JOSE_CURVELINO@COM.BR	jose_curvelino@com.br	Reserved domain
99753280	99753280	Missing @
D_ANTUNES.AOL.COM	d_antunes.aol.com	Missing @
ROSANGELA.BARROSO@	rosangela.barroso@	Missing domain
BRASRIIBEIRO@.COM	brasriibeiro@.com	@ followed by .

Chart I-4 – Examples of Birth Data Validation: (format day/month/year)

BIRTH DATE	VALIDATION
25/13/1995	Invalid date
01/01/1970	Too many repeated
02/05/2015	Future date
23/01/2007	Age under 18
10/08/1901	Age over 100
27/01/1956	Valid date
	Empty field

EXAMPLES OF DATA QUALITY PROFILING REPORTS



ASSESSO

Process: SOURCE_A_VALIDATION
Step: CONTENT_INVESTIGATION

Date: 30/10/2009
Time: 16:44:15

TABLE: TB_PCR_CLI_PF **ATTRIBUTE:** BIRTH_DATE

TOOL: ContentFrequency_5

Content Frequency

Sorted by Value (asc)		Sorted by Quantity (desc)	
Quantity	Value	Quantity	Value
37	00000000	1.065	01011970
7	01011900	321	22041966
1	01011901	315	12121970
1	01011911	266	28091962
2	01011914	263	01011960
2	01011918	257	06031978
1	01011919	249	09091977
2	01011920	247	20011968
1	01011921	245	13061972
2	01011922	245	01011959
7	01011923	243	10051964
3	01011924	240	28081972
3	01011925	240	12101963
8	01011926	239	21071971
6	01011927	237	04041964
13	01011928	236	19031964
8	01011929	236	18021972
18	01011930	236	06101978
19	01011931	236	03091965
17	01011932	235	12061976
23	01011933	235	01011963
46	01011934	234	10101965
36	01011935	233	01051970
30	01011936	233	01011965
31	01011937	232	19041974
40	01011938	232	04061970
33	01011939	231	10061967
71	01011940	231	05051968
58	01011941	230	26041976
49	01011942	230	03031964
82	01011943	229	27091968
69	01011944	229	25061974
91	01011945	229	10051974
86	01011946	229	03111965
105	01011947	228	19031976
121	01011948	227	25081981
136	01011949	227	23031963
182	01011950	227	05091963
155	01011951	225	10101972
147	01011952	225	01091972

Process: SOURCE_A_VALIDATION
Step: PERSONAL_DATA_VALIDATION

Date: 23/11/2009
Time: 09:47:25

TABLE: TB_CLI_PF **ATTRIBUTE:** PROFESSION

TOOL: DomainValidation_12

Group: Category

Description	#	%
Valid	2.240.771	99,8%
Invalid	4.188	0,2%
Empty	52	0,0%
TOTAL	2.245.011	

Valid

Ret Code	Description	#	%
0	Valid	2.240.771	100,0%
TOTAL		2.240.771	

Invalid

Ret Code	Description	#	%
1	Invalid	4.188	100,0%
TOTAL		4.188	

Empty

Ret Code	Description	#	%
999	Empty field	52	100,0%
TOTAL		52	

Process: SOURCE_D_VALIDATION
Step: PERSONAL_DATA_VALIDATION

Date: 01/12/2009
Time: 14:16:51

TABLE: TB_CLI_PF **ATTRIBUTE:** BIRTH_DATE

TOOL: Business_Rule_16

Group: Category

Description	#	%
Red	8.720	23,2%
Yellow	29	0,1%
Green	28.853	76,7%
TOTAL	37.602	

Red

Ret Code	Description	#	%
1	Empty field	0	0,0%
2	Value too many repeated	6.898	79,1%
3	Invalid date	0	0,0%
4	Future date	2	0,0%
5	Age under 18	1.820	20,9%
TOTAL		8.720	

Yellow

Ret Code	Description	#	%
6	Age over 100	29	100,0%
TOTAL		29	

Green

Ret Code	Description	#	%
900	Valid date	28.853	100,0%
TOTAL		28.853	

Process: SOURCE_F_VALIDATION
Step: PERSONAL_DATA_VALIDATION

Date: 01/12/2009
Time: 14:16:51

TABLE: TB_CLI_PF	ATTRIBUTE: TAX_ID_CODE
-------------------------	-------------------------------

TOOL: TaxIdCheckDigitValidation_3

Group: Category

Description	#	%
Valid	37.601	100,0%
Invalid	1	0,0%
Empty	0	0,0%
TOTAL	37.602	

Valid

Ret Code	Description	#	%
0	Valid Income Tax Id Code	37.601	100,0%
TOTAL		37.601	

Invalid

Ret Code	Description	#	%
1	Invalid Income Tax Id Code	1	100,0%
TOTAL		1	

Empty

Ret Code	Description	#	%
500	Empty field	0	0,0%
TOTAL		0	

Process: SOURCE_A_VALIDATION
Step: PERSONAL_DATA_VALIDATION

Date: 23/11/2009
Time: 09:47:25

TABLE: TB_CLI_PF **ATTRIBUTE:** CUSTOMER_NAME

TOOL: NameValidation_8

Group: Category

Description	#	%
Valid	2.241.053	99,8%
Suspect	1.382	0,1%
Invalid	2.576	0,1%
TOTAL	2.245.011	

Valid

Ret Code	Description	#	%
0	Valid	2.077.667	92,7%
1	Valid but first name only initial	159	0,0%
2	Valid but first name unrecognized	163.227	7,3%
TOTAL		2.241.053	

Suspect

Ret Code	Description	#	%
11	Suspect words	689	49,9%
12	Only two words and identical	2	0,1%
13	Three consecutive equal letters	188	13,6%
14	More than three consecutive equal letters	4	0,3%
15	Name with less than 3 letters	0	0,0%
18	Only one word	159	11,5%
19	Only two words and last one abbreviated	1	0,1%
20	Last word abbreviated	339	24,5%
TOTAL		1.382	

Invalid

Ret Code	Description	#	%
16	No vowels	55	2,1%
17	Contains numbers	1.141	44,3%
21	Contains special characters	1.380	53,6%
500	Empty field	0	0,0%
TOTAL		2.576	

Process: SOURCE_A_VALIDATION
Step: ADDRESS_VALIDATION

Date: 23/11/2009
Time: 10:18:18

TABLE: TB_CLI_END **ATTRIBUTE:** ADDRESS

TOOL: AddressValidation_14

Group: Category

Description	#	%
Confirmed or corrected	2,464,888	85,9%
Unrecognized	405,103	14,1%
TOTAL	2,869,991	

Confirmed or Corrected

Ret Code	Description	#	%
0	Confirmed and standardized address	1,880,386	76,3%
1	ZIP code suffix corrected	156,654	6,4%
2	ZIP code corrected	252,352	10,2%
4	City corrected	6,149	0,2%
5	ZIP code suffix and city corrected	1,561	0,1%
6	ZIP code and city corrected	124	0,0%
8	State corrected	966	0,0%
9	ZIP code suffix and state corrected	221	0,0%
10	ZIP code and state corrected	13	0,0%
12	City and state corrected	273	0,0%
13	ZIP code suffix, city and state corrected	58	0,0%
14	ZIP code, city and state corrected	0	0,0%
16	Confirmed but not standardized address	109,771	4,5%
17	ZIP code suffix corrected - not standardized	9,087	0,4%
18	ZIP code corrected - not standardized	15,550	0,6%
20	City corrected - not standardized	2,090	0,1%
21	ZIP code suffix and city corrected - not standardized	892	0,0%
22	ZIP code and city corrected - not standardized	0	0,0%
24	State corrected - not standardized	48	0,0%
25	ZIP code suffix and state corrected - not standardized	16	0,0%
26	ZIP code and state corrected - not standardized	0	0,0%
28	City and state corrected - not standardized	14	0,0%
29	ZIP code suffix, city and state corrected - not standardized	13	0,0%
30	ZIP code, city and state corrected - not standardized	0	0,0%
90	P.O. box or farm - city corrected	0	0,0%
91	P.O. box or farm - zip code suffix corrected	51	0,0%
92	P.O. box or farm - zip code corrected	1,259	0,1%
93	P.O. box or farm - state and city corrected	0	0,0%
94	P.O. box or farm - state corrected	2	0,0%
95	P.O. box or farm - zip code suffix and state corrected	0	0,0%
97	P.O. box or farm - zip code and state corrected	0	0,0%
99	P.O. box or farm - zip code and city compatible	27,338	1,1%
TOTAL		2,464,888	

Unrecognized

Ret Code	Description	#	%
113	Unidentified street - zip code and city compatible	191,940	47,4%
115	Unable to solve street tie - zip code and city compatible	48,307	11,9%
117	Address number out of range - zip code and city compatible	9,453	2,3%
213	Unidentified street - zip code and city incompatible	91,278	22,5%
215	Unable to solve street tie - zip code and city incompatible	42,227	10,4%
217	Address number out of range - zip code and city incompatible	3,668	0,9%
299	P.O. box or farm - zip code and city incompatible	4,465	1,1%
311	Unable to solve city tie	755	0,2%
312	Unidentified city	10,994	2,7%
399	P.O. box or farm - unidentified city	516	0,1%
420	Foreign address - not validated	0	0,0%
500	Empty field	4	0,0%
501	No valid character	1,087	0,3%
502	No valid component	409	0,1%
TOTAL		405,103	

Process: SOURCE_C_VALIDATION
Step: TELEPHONE_VALIDATION

Date: 30/11/2009
Time: 15:32:18

TABLE: TB_TEL_PF ATTRIBUTE: TELEPHONE

TOOL: TelephoneValidation_6

Group: Category

Description	#	%
Confirmed or corrected	2.115.086	66,1%
Suspect	60.847	1,9%
Unrecognized	1.022.054	32,0%
TOTAL	3.197.987	

Confirmed or Corrected

Ret Code	Description	#	%
Ret code 00x - Telephone and address in the same city			
000	Area code and prefix OK	623.882	29,5%
001	Area code corrected	4.490	0,2%
002	Prefix corrected	903.251	42,7%
003	Area code and prefix corrected	63.428	3,0%
Ret code 02x - Telephone and address within 100 km			
020	Area code and prefix OK	115.162	5,4%
021	Area code corrected	272	0,0%
022	Prefix corrected	69.150	3,3%
023	Area code and prefix corrected	2.952	0,1%
Ret code 05x - Zip code not informed			
050	Area code and prefix OK	39.782	1,9%
051	Area code corrected	797	0,0%
052	Prefix corrected	136.025	6,4%
053	Area code and prefix corrected	11.306	0,5%
Ret code 07x - Area code not informed			
070	Area code and prefix OK	0	0,0%
071	Area code corrected	38.464	1,8%
072	Prefix corrected	0	0,0%
073	Area code and prefix corrected	106.047	5,0%
Ret code 09x - Area code and zip code not informed			
090	Area code and prefix OK	0	0,0%
091	Area code corrected	78	0,0%
092	Prefix corrected	0	0,0%
093	Area code and prefix corrected	0	0,0%
TOTAL		2.115.086	

Suspect

Ret Code	Description	#	%
Ret code 10x - Telephone and address over 100 km distant			
100	Area code and prefix OK	28.379	46,6%
101	Area code corrected	180	0,3%
102	Prefix corrected	29.338	48,2%
103	Area code and prefix corrected	2.950	4,8%
TOTAL		60.847	

Unrecognized

Ret Code	Description	#	%
Ret code 2xx+ - Invalid or not validated			
201	Prefix does not exist in area code	173.426	17,0%
202	Prefix unknown	52.332	5,1%
203	Unable to solve prefix tie	2.571	0,3%
302	Invalid telephone number	49.537	4,8%
303	Unidentified zip code city	0	0,0%
304	Unidentified area code city	0	0,0%
305	Invalid area code number	10	0,0%
400	Toll free number	3	0,0%
401	Cell phone - not validated	217.562	21,3%
500	Empty field	0	0,0%
501	No valid character	526.613	51,5%
TOTAL		1.022.054	

Process: SOURCE_A_VALIDATION
Step: EMAIL_VALIDATION

Date: 23/11/2009
Time: 09:47:25

TABLE: TB_CLI_PF **ATTRIBUTE:** EMAIL

[Menu](#)

TOOL: EmailValidation_30

Group: Category

Description	#	%
Valid	70.188	3,1%
Invalid	2.174.823	96,9%
TOTAL	2.245.011	

Valid

Ret Code	Description	#	%
0	Valid e-mail address	68.946	98,2%
101	Domain corrected	1.242	1,8%
	TOTAL	70.188	

Invalid

Ret Code	Description	#	%
201	Missing "@"	6.545	0,3%
202	More than one "@"	44	0,0%
203	Starting with "@"	3.394	0,2%
204	Starting with "."	0	0,0%
205	Ending with "."	29	0,0%
206	Missing domain	34	0,0%
207	Contains spaces	569	0,0%
208	"@" after "."	46	0,0%
209	"." after @ or "."	129	0,0%
210	Unknown character	0	0,0%
211	Invalid character	124	0,0%
301	Domain ended by a single letter	22	0,0%
302	Domain incomplete	877	0,0%
303	Domain reserved to Brazilian internet authorities	19	0,0%
304	Brazilian domain with a letter different from B E N or S	12	0,0%
305	Brazilian domain numeric	6	0,0%
306	Brazilian domain with more than 26 letters	0	0,0%
307	Brazilian domain starting or ending with "-"	2	0,0%
308	American domain with a letter different from Q X ou Z	77	0,0%
309	Inexistent Top Level Domain (TLD)	109	0,0%
401	Invalid character for Brazil	9	0,0%
402	Invalid character for USA	0	0,0%
500	Empty field	2.162.776	99,4%
	TOTAL	2.174.823	