Data quality improvement in a production environment

Cremers, R.C.V.

Award date:
2016

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
Eindhoven, July 2016

Data quality
improvement
in a production environment

by
Raimond Cremers

Student identity number 0829160

In partial fulfilment of the requirements for the degree of

Master of Science

In Operations Management and Logistics

Supervisors:

Dr. Ir. H. Eshuis, TU/e, IS
Dr. P.P.F.M. van de Calseyde TU/e, HPM

Company Supervisors:

M.D.J. Janssen, VDL Nedcar
L.H.J.E. Heiligers, VDL Nedcar
TUe School of Industrial Engineering

Series Master Theses Operations Management and Logistics

Subject headings: Industrial engineering, data quality, ERP-system, automation tool.
Abstract
This research focused on the improvement of data quality in an ERP-system within a production environment. The study researches if data quality can be improved by developing an automation tool that uses the model of Haug et al. (2009). The research provides insight in how implementing an automation tool that uses the model of Haug et al. (2009) benefits the organization.
Management summary
This study focuses on data quality problems in Enterprise Resource Planning (ERP) system SAP/R2 within the Supply Chain Engineering (SCE) department of VDL Nedcar. The goal is to investigate the data quality problems in SAP and significantly decrease these problems. The research provides a solution design that considers all aspects defined in the analysis. This paper underlines why it is important to investigate which data quality problems occur and how these specified data quality problems can be solved.

Problem description
Supply chain data engineers work with SAP, which uses a huge amount of parameters with respect to the logistical process. The logistical process consists of the physical, non-physical, in-house, and out-house flow. Information systems are needed to lead these flows in the right direction. In this logistical process, several problems occur due to data quality issues in SAP. The data quality problems in the SAP data are mainly intrinsic with respect to meaningfulness and correctness and these data errors are above the level of 5% of the Bills Of Material (BOM) of 12255 lines. The impacts can be defined as costs that can be categorised into delay, extra paperwork, transportation operations, and dissatisfaction of the employees. These aspects can be summarized into the following problem statement: ‘The current percentage of data quality errors in SAP with respect to the BOM is above 5%, which results in increased costs that are categorised into delay, extra logistical operations, and dissatisfaction of the employees, and does not meet Nedcar’s expectations of data quality with a desired error percentage below 0.5%.’

Research approach
The research was conducted in three phases: the data collection phase, the data analysis phase, and the solution design phase.

In the data collection phase, all of the necessary information that was needed to conduct an analysis was gathered. This was done by observing the floor processes within VDL Nedcar, where the actions and communications of the Material Handling Floor employees and the Material Handling Engineer were observed. Furthermore, information was gathered by extracting data from SAP. An SCE expert explained the extraction. The extraction provided four lists that were used to control the process and as a reference to the correct data. To understand the data extracted for analysis, further knowledge about these data needed to be obtained by conducting interviews with stakeholders and experts. The interviews among stakeholders served to define the general problems in more detail. To obtain basic and advanced knowledge regarding the four data lists, interviews were conducted with the experts. For every list, basic knowledge was gathered from the expert of SCE to ensure a well-performed analysis. The basic knowledge consists of important headings for the data, and the definitions of these important headings. The advanced knowledge consists of the fixed combinations between the data columns. This is expressed with IF-THEN rules. The interviews with the stakeholders were standardised, open-ended interviews, while the interviews with the experts used a general interview guide approach. This approach was chosen to ensure that the same general areas of information were covered, but that there was still room for discussion. To gain process knowledge of the problem, seven stakeholders (including two experts) from the Material Handling Engineer (MHE) and the SCE departments completed questionnaires. The questionnaire treated topics that are important to decision-making in the solution design. Also interviews with the SCE expert, SCE stakeholders, and an information management expert were conducted to ensure the scope of the project would be remained.

Subsequently, both a process-oriented and an empirical analysis were conducted using the gathered information. Both analyses had the purpose of identifying a solution design. The empirical analysis focused on the data themselves, and the process-oriented analysis focused on the business process to
avoid unexpected problems caused by that business process. The process-oriented analysis began with
the results of the questionnaires, which were conducted to obtain process knowledge of the problem.
Next, the unload PSAs were analysed, which were unloading gates. Here the most used unload PSAs
were determined, and subsequently a sample of these unload PSAs was analysed. In this analysis, costs
made per hour were defined by dividing the labour tariff per year plus surcharge by the individual
working hours per year. These costs made per hour were multiplied by the duration of work, so the
outcome became the costs made per day. Finally, the process flow was analysed, exposing the
bottlenecks in the process due to data quality errors.

The empirical analysis started with the detection of the symptoms of the data quality problems. To
accomplish this, the interviews were analysed. The causes and consequences of the conducted
interviews that the stakeholders cited more than three times were considered the main cause and
consequence. Furthermore, the problem statement analysis could be conducted. This was done using
the master data lists and the process knowledge of two experts. Furthermore, a refined cause-and-
effect tree was developed of the preliminary cause-and-effect tree made in the problem definition. With
the results of the interviews, the preliminary cause-and-effect diagram was further analysed and a
comparison of the new cause-and-effect tree (Ishikawa diagram) and the preliminary cause-and-effect
diagram was made. The master data lists were first validated with the use of the pivot table method.
Next, these lists were analysed by using a Pareto chart.

The solution design was developed based on the conclusions drawn from the analysis. This addressed
the requirements, data quality classification, analysis conclusions, design, data improvement tool
solution design, automation tool actual designing, user manual, and implementation plan. The
requirements were gathered from the interviews with the SCE expert, the SCE stakeholders and the
information management expert. With the SCE expert, mainly the Functional and Boundary
requirements were set. The User requirements were set from the interview with the SCE stakeholders.
The Design restrictions were gathered from the Information management expert. Based on the data
quality classification model by Anders Haug, Jan Stentoft Arlbjorn, and Anne Pedersen (2009), a data
quality classification for improvement was developed. The analysis conclusions were categorised within
Haug et al.’s (2009) classification model. Next, for each conclusion made based on the analysis, a
solution was developed. These solutions were provided in comparison with Haug et al.’s (2009) model,
and the influences were elaborated. Based on these explorations, the data improvement tool solution
design was created using the design approach by COMET (Berre et al., 2006).

First the stakeholders of the data improvement tool were identified for development. Second, a goal
hierarchy model was developed. Third, a business resource model was made. Fourth, a process model
was developed. Finally, a refined process model was provided. With these models, the design of the
data improvement tool was specified. After specification of the design, the actual designing of the
automation tool was addressed. This part consisted of synthesis-evaluation iterations. The main goal of
the tool was to check the daily data on errors and provide a results list with these errors presented in a
clear and understandable manner. The synthesis-evaluation iterations can be explained by discussing
three main concept versions of the tool. The first version served to create a basis: obtain data, check
data, and present the results of the data checked on one list. The second concept version served to
expand all the rules of one list. The third concept version aimed to add the other lists and combine these
into one tool. All of the automation codes were written in VBA. Finally, a user manual and an
implementation plan were created.
Results
The SCE expert and the MHE expert are of the opinion that the tool is a solution to the problem, because every time it will be used and the provided data errors will partially be solved, the data quality will increase. At first, they believed that the data quality improvement would cost a great deal of time and that this would be achieved in steps. Subsequently, it will be a matter of keeping the data quality up-to-date; the tool will be used weekly/monthly to maintain the data quality. The experts’ opinion is that the tool represents the data in such a way that it will be possible to change the data more easily than before. Furthermore, the awareness of data errors will increase among Supply Chain Engineers.

The tool was also validated by the experts and the stakeholders, who will be using it. The two experts were asked to use the tool for a period of time and to provide feedback with regard to the requirements and goals set. Each expert was asked to use the tool for detecting errors in SAP.

Below, the feedback provided by the experts is stated:

Expert 1: ‘Today, once again, we sat together and discussed, viewed, and tested the automation tool made by Raimond Cremers. I have carried out various checks with the tool and the output of these tests was satisfactory. I expect a big improvement in data quality after the introduction of this tool.’

Expert 2: ‘I have used the automation tool for several errors and found the errors with ease. The automation tool found the same results as when I do the check manually. The automation tool will save time in checking the data for SCE.’

Based on these citations, it can be concluded that the tool works properly and the requirements and goals that were set have been met and accomplished. Furthermore, both experts predict that the tool will accomplish the reach target of <0.5% data quality error percentage.

As mentioned in Section 3.3.2.1.3, N. Brand and H. van der Kolk (1995) provide four dimensions with which to evaluate the data improvement tool. These dimensions are as follows: Time, Costs, Quality, and Flexibility. The time that employees spend correcting data quality errors is expected to decrease by 2.755 hours/day in the long run. The costs incurred by correcting data quality are expected to decrease by 111.25 hours/day in the long run. The quality of the data is expected to increase by 90%, and the flexibility will increase by providing an option regarding how to check the data for data quality errors.

Conclusions
The main research question, which is derived from the problem statement, is the following: ‘How can data quality in an ERP system (SAP R/3) be significantly improved in a production environment by implementing an automation tool to keep the percentage of inconsistencies in data under 0.5% of the BOM of 12255 lines in the long run?’

The data quality in SAP R/3 can be increased significantly in the long run, as is validated in Chapter 7, by means of Haug et al.’s (2009) model, integrated in the automation tool, and the implementation of the automation tool, mentioned in Chapter 6. The tool extracts and checks the SAP data with the help of process knowledge and provides an overview of data errors that have to be adapted. In combination with the authorisation protocol, the communication protocol, and the workflow of the process, this covers an important part of Haug et al.’s (2009) model, which can be seen in Chapter 2. As a result, the data quality will significantly increase and the target of fewer than 0.5% wrong entries will be reached, as stated by the experts in Section 7.3. The main recommendation is that the tool can be made more useful by expanding its reach to other databases within VDL Nedcar. Furthermore, the SCE expert should manage the tool.


Contents

Abstract .............................................................................................................................. iii

Management summary ................................................................................................... iv

   Problem description .................................................................................................... iv

   Research approach ..................................................................................................... iv

   Results ......................................................................................................................... vi

   Conclusions ................................................................................................................. vi

List of Figures .................................................................................................................. xii

List of Tables ....................................................................................................................... xiii

Abbreviations .................................................................................................................... xiv

1. Introduction ................................................................................................................. 1

   1.1 Introduction to the organisation ........................................................................... 1

   1.2 Problem definition ............................................................................................... 1

      1.2.1 Factory operations ....................................................................................... 2

      1.2.2 Supply chain engineering – data engineer ................................................... 3

      1.2.3 Cause-and-effect tree .................................................................................. 4

      1.2.4 Problems ....................................................................................................... 6

      1.2.5 Problem statement ....................................................................................... 7

   1.3 Research question ................................................................................................. 7

      1.3.1 Main research question ................................................................................. 7

   1.4 Structure of report ............................................................................................... 7

2. Literature .................................................................................................................... 9

   2.1 Data quality ......................................................................................................... 9

   2.2 Data quality and ERP systems ............................................................................ 11

   2.3 Implementation in VDL Nedcar ......................................................................... 13

3. The research approach .............................................................................................. 14

   3.1 Conceptual research design ................................................................................ 14

   3.2 Methodology ........................................................................................................ 14

      3.2.1 The problem solving cycle .......................................................................... 15

4. Data collection .......................................................................................................... 20

   4.1 Questionnaire ...................................................................................................... 20

   4.2 Observation of the floor processes and the employees ...................................... 21

vii
4.3 Extracting data ........................................................................................................... 21
4.4 Interviews ................................................................................................................... 21
  4.4.1 Stakeholder interviews ......................................................................................... 21
  4.4.2 Expert stakeholder interviews ............................................................................. 21
  4.4.3 Knowledge collection ......................................................................................... 22
5 Data analysis ................................................................................................................ 26
  5.1 Process-oriented analysis ....................................................................................... 26
      5.1.1 Questionnaire ................................................................................................. 26
      5.1.2 Unload PSAs ................................................................................................ 26
      5.1.3 Process flow .................................................................................................. 28
  5.2 Empirical analysis ................................................................................................... 29
      5.2.1 Stakeholders interviews ................................................................................... 29
      5.2.2 Problem statement analysis ............................................................................ 29
  5.3 Conclusions from the analysis .............................................................................. 35
6 Solution design ............................................................................................................. 36
  6.1 Requirements .......................................................................................................... 36
  6.2 Data quality classification model ............................................................................ 36
  6.3 Analysis conclusions and design ............................................................................ 37
      6.3.1 Decreasing time ............................................................................................... 38
      6.3.2 Decreasing costs ............................................................................................. 39
      6.3.3 Increasing process knowledge ......................................................................... 40
      6.3.4 Authorisation ................................................................................................ 40
      6.3.5 Communication ............................................................................................. 40
      6.3.6 Controllability ................................................................................................ 40
      6.3.7 Main focus on the wrong entries ..................................................................... 41
      6.3.8 Use of the 80-20 heuristic .............................................................................. 41
      6.3.9 Conclusion ...................................................................................................... 41
  6.4 Data improvement tool solution design .................................................................... 41
      6.4.1 Context statement ........................................................................................... 42
      6.4.2 Goal hierarchy model ..................................................................................... 42
      6.4.3 Business resource model ............................................................................... 44
      6.4.4 Process model ................................................................................................ 45
6.4.5 Work Analysis Refinement Model (WARM) diagrams.........................................................45
6.5 Designing the DIT..................................................................................................................46
  6.5.1 VBA..................................................................................................................................46
  6.5.2 Version 1 ..........................................................................................................................46
  6.5.3 Version 2 ..........................................................................................................................47
  6.5.4 Version 3 ..........................................................................................................................47
6.6 User manual ..........................................................................................................................47
6.7 Implementation plan .............................................................................................................48
7 Verification and Validation of the tool....................................................................................49
  7.1 Devil's quadrangle................................................................................................................49
    7.1.1 Time ...............................................................................................................................49
    7.1.2 Costs ..............................................................................................................................49
    7.1.3 Quality ............................................................................................................................50
    7.1.4 Flexibility .........................................................................................................................50
  7.2 Pivot table method ...............................................................................................................50
  7.3 Expert and stakeholder validation ......................................................................................51
    7.3.1 Experts ............................................................................................................................51
    7.3.2 Stakeholders....................................................................................................................51
  7.4 Conclusion ...........................................................................................................................51
8 Conclusion, discussion, and recommendations.......................................................................52
  8.1 Conclusions..........................................................................................................................52
    8.1.1 Operational level.............................................................................................................52
    8.1.2 Tactical level....................................................................................................................52
    8.1.3 Strategic level...................................................................................................................52
  8.2 Problem statement conclusion ............................................................................................53
  8.3 Theoretical contribution .......................................................................................................53
  8.4 Limitations ............................................................................................................................53
  8.5 Recommendations ...............................................................................................................53
  8.6 Future research directions ....................................................................................................54
Bibliography .................................................................................................................................55
Appendix .........................................................................................................................................57
I. Organograms..............................................................................................................................58
## VDL Group

VDL Nedcar

### II. History

- History VDL Nedcar
- Nedcar
- VDL Group

### III. Production stream

- Press shop
- Body shop
- Paint shop
- Final Assembly shop
- Test track & Yard

### IV. Additional information

- IV.I Design research guidelines
- IV.II Operational research plan
- IV.III The cost of the project
- IV.IV The organization of the project
  - Rik Eshuis
  - Phillipe Calseyde
  - Marcel Janssen
  - Leon Heiligers

### V. Data collection files

- V.I Questionnaire

### VI. Process models

### VII. Data Improvement Tool (DIT) user manual

- V.I Visual Basic for Applications
- V.II Activation
- V.III Main menu
- V.IV Get data Process
- V.V Check data process
- V.VI Checking logistic process
  - V.VI.I A check category
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ground plan shops in VDL Nedcar</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Cause-and-effect diagram</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>Conceptual project design</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>The problem solving cycle by Aken et al. (2007)</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>Stakeholders’ questionnaire results</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>Pareto chart of unload PSAs</td>
<td>27</td>
</tr>
<tr>
<td>7</td>
<td>Process flow Supply Chain Engineers</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>Pareto chart of wrong entries</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>Pareto categorised wrong entries</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>Ishikawa diagram</td>
<td>34</td>
</tr>
<tr>
<td>11</td>
<td>Merging three steps</td>
<td>39</td>
</tr>
<tr>
<td>12</td>
<td>Insertion of the checking data step in the process flow</td>
<td>39</td>
</tr>
<tr>
<td>13</td>
<td>Context statement model</td>
<td>42</td>
</tr>
<tr>
<td>14</td>
<td>Goal hierarchy model</td>
<td>43</td>
</tr>
<tr>
<td>15</td>
<td>Business resource model</td>
<td>44</td>
</tr>
<tr>
<td>16</td>
<td>Process model</td>
<td>45</td>
</tr>
<tr>
<td>17</td>
<td>WARM activity diagram</td>
<td>45</td>
</tr>
<tr>
<td>18</td>
<td>Organogram VDL Group</td>
<td>58</td>
</tr>
<tr>
<td>19</td>
<td>Organogram VDL Nedcar</td>
<td>59</td>
</tr>
<tr>
<td>20</td>
<td>Operational research planning</td>
<td>65</td>
</tr>
<tr>
<td>21</td>
<td>Process flow LSP</td>
<td>68</td>
</tr>
<tr>
<td>22</td>
<td>Product flow</td>
<td>68</td>
</tr>
<tr>
<td>23</td>
<td>Process flow Material Handling</td>
<td>69</td>
</tr>
<tr>
<td>24</td>
<td>DIT command button</td>
<td>70</td>
</tr>
<tr>
<td>25</td>
<td>DIT tool menu</td>
<td>70</td>
</tr>
<tr>
<td>26</td>
<td>DIT tool notification</td>
<td>71</td>
</tr>
<tr>
<td>27</td>
<td>DIT tool getting data menu</td>
<td>71</td>
</tr>
<tr>
<td>28</td>
<td>DIT tool checking data menu</td>
<td>71</td>
</tr>
<tr>
<td>29</td>
<td>DIT tool checking data logistic process menu</td>
<td>72</td>
</tr>
<tr>
<td>30</td>
<td>DIT tool checking data checking packing instructions menu</td>
<td>73</td>
</tr>
<tr>
<td>31</td>
<td>DIT tool checking data checking packing instructions &amp; logistics process menu</td>
<td>73</td>
</tr>
<tr>
<td>32</td>
<td>Sequence diagram menu</td>
<td>76</td>
</tr>
<tr>
<td>33</td>
<td>Sequence diagram check</td>
<td>77</td>
</tr>
<tr>
<td>34</td>
<td>Sequence diagram manual</td>
<td>78</td>
</tr>
</tbody>
</table>
List of Tables
Table 1 Intrinsic data quality categorisation ................................................................. 10
Table 2 Relevant heading definitions ............................................................................... 23
Table 3 Experts’ advanced knowledge regarding the data lists ........................................ 24
Table 4 Clustered questionnaire results ............................................................................ 26
Table 5 Data quality error correction times ...................................................................... 27
Table 6 Pivot table example ............................................................................................ 30
Table 7 Fixed combination of the experts check ............................................................... 30
Table 8 Calculation table of errors .................................................................................... 31
Table 9 Data quality issues ............................................................................................... 35
Table 10 Requirements ...................................................................................................... 36
Table 11 Framework of the data classification model by Haug et al. (2009) ..................... 37
Table 12 Classified framework of the data classification model by Haug et al. (2009) ....... 38
Table 13 Solutions versus data quality classification model .............................................. 41
Table 14 Extraction of Table 3 .......................................................................................... 50
Table 15 Pivot table of values from Table 14 .................................................................... 50
Table 16 Error percentage summary ................................................................................ 50
Table 17 Example of the results presentation list .............................................................. 50
Table 18 Design research guidelines ............................................................................... 63
Table 19 Relationship IF-Then example .......................................................................... 72
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Car type A MM</td>
</tr>
<tr>
<td>AccAssCatP</td>
<td>Account Assignment Category PSA</td>
</tr>
<tr>
<td>B</td>
<td>Car type B MM</td>
</tr>
<tr>
<td>BMW</td>
<td>Bayerische Motoren Werke</td>
</tr>
<tr>
<td>BOM</td>
<td>Bill Of Materials</td>
</tr>
<tr>
<td>BPR</td>
<td>Business process reengineering</td>
</tr>
<tr>
<td>C</td>
<td>Car type C MM</td>
</tr>
<tr>
<td>CostCntrPS</td>
<td>Cost Centre PSA</td>
</tr>
<tr>
<td>DIT</td>
<td>Data improvement Tool</td>
</tr>
<tr>
<td>ERP</td>
<td>Enterprise Resource Planning</td>
</tr>
<tr>
<td>ExtEmbCdeP</td>
<td>External Emballage Code Packing Instruction (PI)</td>
</tr>
<tr>
<td>FLC</td>
<td>Flowcode MM</td>
</tr>
<tr>
<td>FPS</td>
<td>Field Problem Solving</td>
</tr>
<tr>
<td>GenItCatGr</td>
<td>Generic Item Category Group MM</td>
</tr>
<tr>
<td>GsPsA</td>
<td>Goods Supplier PSA</td>
</tr>
<tr>
<td>HandlingUnit1CalculatedVolPI</td>
<td>Volume Dimension Calculated With Handling Unit</td>
</tr>
<tr>
<td>HandlingUnit1DimLength</td>
<td>Length Dimension in Handling Unit</td>
</tr>
<tr>
<td>HandlingUnit1DimWidth</td>
<td>Width Dimension in Handling Unit</td>
</tr>
<tr>
<td>HandlingUnit1Height</td>
<td>Height Dimension in Handling Unit</td>
</tr>
<tr>
<td>hghtDimPI</td>
<td>Height Dimension PI</td>
</tr>
<tr>
<td>IPC</td>
<td>Internal Process Code MM</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>LghtDimPI</td>
<td>Length Dimension PI</td>
</tr>
<tr>
<td>LO</td>
<td>Laboratory Office MM</td>
</tr>
<tr>
<td>LSP</td>
<td>Logistics Service Provider</td>
</tr>
<tr>
<td>MatDescMM</td>
<td>Material Description MM</td>
</tr>
<tr>
<td>MDM</td>
<td>Master Data Management</td>
</tr>
<tr>
<td>MHE</td>
<td>Material Handling Engineer</td>
</tr>
<tr>
<td>MHF</td>
<td>Material Handling Floor employee</td>
</tr>
<tr>
<td>MM</td>
<td>Material Master</td>
</tr>
<tr>
<td>Pack.obj.</td>
<td>Packaging Object Number MM</td>
</tr>
<tr>
<td>PI</td>
<td>Packing Instruction</td>
</tr>
<tr>
<td>ProjNrMM</td>
<td>Project Number MM</td>
</tr>
<tr>
<td>PSA</td>
<td>Product Scheduling Agreement</td>
</tr>
<tr>
<td>QtyPerPMat</td>
<td>Quantity (Ref) Material per 1 Packing material</td>
</tr>
<tr>
<td>SAP</td>
<td>Systemen, Anwendungen und Produkte</td>
</tr>
<tr>
<td>SCE</td>
<td>Supply Chain Engineering</td>
</tr>
<tr>
<td>SlocPSA</td>
<td>Storage Location PSA</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical Package for the Social Sciences</td>
</tr>
<tr>
<td>SsiMM</td>
<td>Storage Section Indicator MM</td>
</tr>
<tr>
<td>StkPlcmmntM</td>
<td>Stock Placement MM</td>
</tr>
<tr>
<td>StkRmvlMM</td>
<td>Stock Removal MM</td>
</tr>
<tr>
<td>Sut1MM</td>
<td>Storage Unit Type 1 MM</td>
</tr>
<tr>
<td>Sut1QtyMM</td>
<td>Loading Equipment Quantity 1 MM</td>
</tr>
<tr>
<td>Sut3MM</td>
<td>Storage Unit Type 3 MM</td>
</tr>
<tr>
<td>Sut3QtyMM</td>
<td>Loading Equipment Quantity 3 MM</td>
</tr>
<tr>
<td>SUTLEC</td>
<td>Storage Unit Type in List of Emballage Code</td>
</tr>
<tr>
<td>TotVolPI</td>
<td>Total Volume Dimension PI</td>
</tr>
<tr>
<td>TrEmbCdePI</td>
<td>Transport Emballage Code Packing Instruction</td>
</tr>
<tr>
<td>TrgQtyRMat</td>
<td>Target Quantity of Ref Material PI</td>
</tr>
<tr>
<td>TTF</td>
<td>Task Technology Fit</td>
</tr>
<tr>
<td>UnloadPSA</td>
<td>Unload Gate PSA</td>
</tr>
<tr>
<td>VBA</td>
<td>Visual Basics for Applications</td>
</tr>
<tr>
<td>VDL</td>
<td>Van Der Leegte</td>
</tr>
<tr>
<td>WARM</td>
<td>Work Analysis Refinement Model</td>
</tr>
<tr>
<td>WdthDimPI</td>
<td>Width Dimension PI</td>
</tr>
</tbody>
</table>
1. Introduction
This master thesis about improving data quality in an Enterprise Resource Planning (ERP) system is executed within VDL Nedcar in the department of Supply Chain Engineering, in the sub-department of Data Engineering. The organogram is attached in Appendix I.

1.1 Introduction to the organisation
VDL Nedcar was formerly Nedcar, but since 2012 it is owned by VDL Group. The history of both companies before the takeover by the VDL group is presented in Appendix II. VDL Nedcar is an independent vehicle contract manufacturer. BMW VDL Nedcar’s main producer; VDL Nedcar produces the MINI. The company’s number of employees has grown to around 2,400, with more than 20 nationalities. VDL Nedcar is also responsible for thousands of fulltime jobs in the region. The factory area in Sittard-Geleen (Born) is ca. 927,000 m², of which 330,000 m² consist of the factory building. This building comprises four shops where cars are built or pre-work is done: the Press shop, Body shop, Paint shop, and the Final Assembly shop. The factory has continuously improved, and there are now fully automatic flexible assembly lines, which have a production capacity of 200,000 vehicles per year in a two-shift system. In addition, a high degree of automation is accomplished with the assistance of around 1,600 robots. In the Press shop, there are some of the most advanced presses in all of Europe.

Figure 1 Ground plan shops in VDL Nedcar

1.2 Problem definition
Supply chain data engineers work with the ERP system SAP/R2, which they fill with parameters with respect to the logistical process. This logistical process starts at the unload gates, where material arrives at the factory, and ends at the storage location in the factory. From the storage location, the activities are taken over by the Material Handling department.

In this logistical process, several problems occur due to data quality issues in SAP. This thesis will underline why it is important to investigate which data quality problems occur and how these specified data quality problems can be solved.

The purpose of this research within VDL Nedcar is to relate theoretical knowledge to real world problems. To define these problems, this section first provides a proper view of the company. Then, it describes all characteristics of the specific problem investigated within VDL Nedcar.
1.2 Factory operations
To define a proper problem definition, the factory operations of the Supply Chain Engineering process should be described. The whole process is divided into the following streams:

- Data stream
- Part stream:
  o Out-house
  o In-house
- Production stream

The streams are explained within the scope of the research, in order to keep the focus on the significant and important problems.

1.2.1 Data stream
BMW develops a bill of material for a new car and sends this list to VDL Nedcar. Within VDL Nedcar, this list is engineered by several departments to develop a process that transforms the materials mentioned in the list into the car. This list is maintained and stored in SAP. One of these departments is Supply Chain Engineering, which fills the huge amount of parameters with respect to the logistical process and the packing instructions.

1.2.1.1 Part stream
Two part streams are defined: the out-house part stream and the in-house part stream.

1.2.1.1.1 Out-house
VDL Nedcar works towards an end product. Therefore, a Material Requirements Planning (MRP) is produced to set the production process’s sell and production plan. With this MRP, a planning is made to purchase raw materials and components to fabricate the vehicle. These raw materials and components are not specified for unique vehicle parts, but are applicable to many cars. The orders placed with the suppliers are made in batches and can be stored in VDL Nedcar’s central warehouse, the so-called Warehouse/direct delivery.

VDL Nedcar also works with the pearl chain process. The underlying principle of the pearl chain process is that production and logistics flows follow one plan in order to ensure efficient operations and the lowest possible stock level. Although the pearl chain is fixed six weeks prior to delivery, there is some flexibility to allow for changes in customer demand (i.e. regarding order characteristics and sequence).

Logistics can place batch deliveries based on these pearl chains. This is called the Just In Time (JIT) principle. The delivered batches are brought to the Warehouse On Wheels (WOW), which is a yard with delivery truck trailers, where the deliveries are positioned according the JIT and when necessary can be taken to the dock where the in-house stream starts. The deliveries according the JIT arrives at the trailer park five days before the deliveries are necessary, and these deliveries are then officially taken into the factory when they are needed. The necessity of the deliveries is determined by the in-house stream, which requests the deliveries because the stock is almost empty.

If the delivery of the raw materials and the components is placed in the actual sequence of the pearl chain, this is called Just In Sequence (JIS). Like the JIT deliveries, the JIS deliveries also arrive five days before the deliveries are needed.
Finally, deliveries are done by the suppliers, who can deliver within two hours after an order is made; this is called Supply In Line Sequence (SILS). A forecast is sent five days before the car is produced. The order to deliver is made when the car enters the Final Assembly Line (FAS).

Supply chain engineers have to process this information in the SAP system to ensure the correct process flow.

1.2.1.2 In-house

With JIT orders, each packing has only one article number and the two-bin concept is used. This means that two packings are always located at the production line. If they do not fit into the line because there are too many product variables (different types of the same product group) the parts are commissioned in the commissioning area. This means that the parts are set in the right order of production. This commissioned packing is then installed along the line. The truck trailers are emptied and placed on the footprints stationed at the correct gate. The empty packings are placed back into the truck trailer. These trailers are filled based on the pearl chain mentioned in the previous paragraph.

With JIS orders, each packing has more article numbers of one product variable. The products are pre-packed conform the pearl chain. Due to this work method, only two packings along the production line are ever needed. If the production order differs from the pearl chain, the JIS orders are re-sequenced in the re-sequence area. The process of unloading is done according to a strictly agreed upon order. The loading is the same as for the JIT order.

With SILS, each product is delivered in the realised production order. This means that no re-sequencing has to be done. The supplier has to be able to deliver the products in a couple of hours, as mentioned in the previous paragraph. The packings are unloaded from the trailers to a footprint and are then brought directly to the production line.

Small parts delivered in standardised cradles are handled by the small box warehouse. Special racks with a unique barcode are placed along the production line. Each box has its one article number, which is processed in a barcode. If a box is empty, this is scanned by an employee who collects the boxes, and an automatic order is placed to the employee who delivers these boxes from the small box warehouse.

As can be seen, it is important to fill in the right data to ensure that the products are delivered following the right in-house stream.

1.2.1.3 Production stream

Eventually, all of the streams discussed above have as purpose to serve the production stream such that it has an optimal flow. VDL Nedcar’s goal is to deliver MINIs. Because the MINI building process is interesting but is beyond the scope of this thesis, the production streams can be found in Appendix III. If one of the above streams does not perform as expected, this can have important consequences for the cost, quality, and efficiency of logistic processes.

1.2.2 Supply chain engineering – data engineer

The supply chain data engineers fill in the parameters. These parameters also relate to the packing instructions regarding the ordered parts. The filling in of these data ranges from adding parts to existing packing instructions, to developing completely new packing instructions. This is also done by the data engineers. Packing instructions give information about the parts, the supplier, and the packing itself.
In this logistical process, several problems occur due to data quality issues in SAP. The following section will underline why it is important to investigate which data quality problems occur and how these specified data quality problems can be solved.

1.2.3 Cause-and-effect tree
Based on the intake meeting and the interviews with employees of VDL Nedcar, a preliminary cause-and-effect tree was developed. This preliminary cause-and-effect tree shows all causes and effects of the data quality problem, and is presented on the next page. Furthermore, the completeness, the ambiguousness, and the communication between the involved departments are also investigated.

The part of the cause-and-effect tree within the dashed boundary is the scope of this research. The scope of the project is within the Master Data Management (MDM) in the SCE department. This means that the data quality in the MDM is invested only for the parameters filled by the supply chain engineers. The scope will be on inserting and adjusting the data directly in the MDM.

In the intake meetings and interviews with employees of VDL Nedcar, it was determined that wrong entries and wrong adjustments were made by employees who work with SAP.
Figure 2 Cause-and-effect diagram
1.2.4 Problems
As discussed above, there are many flows within VDL Nedcar; these are physical, non-physical, in-house, and out-house flows. Information systems are needed to lead these flows in the right direction. This study investigates the information system SAP, which is directly related to the data stream and the part stream, both in- and outbound, and is therefore also related to the last phase of the production line. As previously indicated, the running process of the logistic flows is managed with SAP. In this information system, there are data quality problems with important parts of these data; these have a negative impact on the in- and out-house part streams. The problems are inconsistencies between parameters within SAP. Examples of these inconsistencies are:

- Inconsistencies between parameters of shops and unloading gates;
- Inconsistencies between parameters of internal process codes and unloading gates;
- Inconsistencies between parameters related to transport; and
- Inconsistencies between parameters of transport packing codes.

According to Redman (1998), data quality is of high importance on the operational, tactical, and strategic aspects of the organisation. The examples provided above about the inconsistencies between parameters within SAP influence directly the operational and the tactical and indirectly the strategical aspects of VDL Nedcar.

As can be seen in the cause-and-effect diagram, data quality problems lead to operational difficulties that increase costs, which can be categorised into delay, extra paperwork, transportation operations, and dissatisfaction of the employees. For example, there are inconsistencies between parameters of shops and unloading gates. As can be seen in Figure 1, there are four shops at VDL Nedcar. In the near of each shop there are unload gates to ensure the highest efficiency with deliveries. This real world information is mapped into an information state in the data with the help of parameters. Because of the inconsistencies between parameters of shops and unload gates, the Press shops receives sometimes a delivery which belongs to the Final Assembly Shop because the unload gate is wrong. The press shop employees have to interrupt their activities to report the wrong delivery to their supervisors. The supervisors have to correct the parameters that were wrong or contact the Supply Chain Engineering department to correct the parameters that were wrong. Subsequently, the supervisor has to order an employee to internally transport the delivery to the right unload gate. For almost each of the provided examples of inconsistencies, these consequences are the same. In case of the inconsistencies between parameters of internal process codes and unloading gates, the wrong parameters can be detected further in the process and therefore can do more damage because the materials will be necessary at the production line in a shorter time. In the worst case scenario, the production line of the car will stand still.

English (1999) states that the cost of poor data quality is strongly context-dependent, as opposed to the cost of a data quality program. Evaluation of data quality is therefore difficult. English (1999) classifies costs into two categories: process costs and opportunity costs. Process costs are caused by data errors, whereby re-execution of the process is needed. Opportunity costs are incurred due to lost and missed revenues. In the case of VDL Nedcar, the costs previously discussed are process costs and the opportunity costs will be explained next.

The tactical problems are the poor decisions making and forecasting of costs to develop a car for future customers. Due to inconsistencies between parameters, the costs of producing a car can be calculated lower or higher than it really is. The forecasting is necessary for making a quote for a car customer. If the car customer is pleased with the quote, they could decide to sign a contract with VDL Nedcar to produce the car. If the cost to produce a car is forecasted lower, VDL Nedcar is gaining less profit than it has calculated. If the cost to produce a car is forecasted higher, the car manufacturer may not sign the contract and search for another car manufacturer.
The strategic problem is caused by the operational and tactical problem because the manager cannot set a strategy based on data which is not consistent. Also the manager has to solve too many issues regarding these problems, according to what the manager explained in the intake meeting, which leads to less focus towards new car customers.

The data consists of a BOM of 12255 lines. The supply chain engineering department manager’s expectations are that the inconsistencies in the data cannot be more than 0.5% of the 12255 lines of the BOM. However, these inconsistencies are currently above 5%, which is far too high and leads to too many costs in the process. Each error stands for one mistake in one line. If there are more errors in one line, these will be counted as one error line. The data errors are caused by incorrect entries or adjustments of the employees of the supply chain engineering and the material handling departments.

1.2.5 Problem statement
The discussion above is summarized as the following problem statement:

‘The current percentage of data quality errors in SAP with respect to the BOM is above 5%, which results in increased costs that are categorised into delay, extra logistical operations, and dissatisfaction of the employees, and does not meet Nedcar’s expectations of data quality with a desired error percentage below 0.5%.’

1.3 Research question
The data quality problems within the ERP system SAP are discussed in the section 1.2.3. These problems are currently mainly solved using employees’ knowledge of the desired product locations and flows. These faults in the data are intrinsic data quality problems, because the independency of the data quality goals and because the data say something about the data themselves. The purpose is to solve the problems that occur due to poor data quality. To reach a solution, a research question and several sub-questions are defined.

From the above formulation, the research question can be defined. This will have several sub-questions, which answer aspects of the research question and will provide handles to the methodology.

1.3.1 Main research question
‘How can data quality in an ERP system (SAP R/3) be significantly improved in a production environment by implementing an automation tool to accomplish in the long run a percentage of inconsistencies below 0.5% for the BOM?’

1.3.1.1 Sub-questions
- What is the current data quality situation (as-is situation)?
- What are the important causes and consequences of the data quality problem?
- How can the automation tool be designed?
- Is this automation tool applicable, and what is the detailed solution design?
- How can the automation tool be implemented?

1.4 Structure of report
The remainder of this report is organised as follows. In chapter 2, the literature will be explained, where the definition data quality is elaborated, data quality in ERP systems is explained and finally the literature is concluded. In chapter 3, the research approach is explained, in which the conceptual
research design and the methodology is explained. Before the data analysis can be done in chapter 5, chapter 4 will explain how all the information is gathered to do this research. After the data analysis in chapter 5 a solution design will be explained in chapter 6. This solution design will subsequently be validated in chapter 7. Finally, in chapter 8 a conclusion, a discussion, and recommendations are made.
2. Literature

Given the problem described in the previous chapter, a literature study was performed to gain insights into the theory and solution methods regarding data quality problems. The purpose of the literature review was to garner insights into solutions methods and to find gaps in the literature related to data quality.

This chapter highlights the relevant data quality theories and solution methods that were identified. Subsequently, it discusses the data quality within ERP systems.

2.1 Data quality

The literature presents many different ways to describe data quality. Several researchers have defined data quality differently in the past years. Ballou and Pazer (1985) define data quality as ‘a relative rather than an absolute term that can most usefully be defined in the context of end use’. Wang and Strong (1996) define data quality as ‘the level of data that are fit for use by data consumers’. Ballou and Kumar Tayi (1999) define data quality as ‘Fitness for use’, while Lederman, Shanks, and Gibbs (2003) define it as ‘fitness for purpose’. Whereas Ballou and Kumar Tayi (1999) refer to the users who need particular data, Reeva Lederman et al. (2003) refer to the purpose for which data are needed. Given the great number of implementations of ERP systems in organisations, as stated by Myreteg (2015), Park and Kusiak (2005) propose an ERP-specific definition of data quality. Park and Kusiak (2005) define data quality as ‘the measure of the agreement between the data views presented by ERP and that same data in the real world’.

Data quality is also defined as a multiple dimensional concept (Wand & Wang, 1996). Other literature explains data quality by dividing it into dimensions. These dimensions vary in different studies, but Ballou and Pazer (1985) propose the following four dimensions as a basis: accuracy, completeness, consistency, and timeliness.

By accuracy, Ballou and Pazer (1985) mean that the recorded value that is given in the data is in conformity with the actual value that is given in the real world. Timeliness is defined as the recorded value that is not out of date, meaning that it is reviewed/refreshed within the critical time that is standard for reviewing/refreshing the value. Completeness means that all of the values are recorded that fall within the scope of the necessary data. Finally, Consistency indicates that the representation of the data value is the same in all cases. These dimensions are all given characteristics of the data themselves; they are independent of data quality goals. This means that these dimensions are intrinsic data quality dimensions. Because the supervisors in the intake interview stated that the problems were with the ERP data themselves, the study focuses on these intrinsic data quality dimensions. In research published after Ballou and Pazer’s (1985), dimensions have been added, the dimensions have changed, and other perspectives of dimensions have been created. Table 1 presents an overview of how subsequent literature has changed or added perspectives on these dimensions.
Table 1 Intrinsic data quality categorisation

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>Completeness</th>
<th>Consistency</th>
<th>Timeliness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ballou &amp; Pazer (1985)</td>
<td>meaningfulness</td>
<td>completeness</td>
<td>objectivity</td>
<td>temporal correlation</td>
</tr>
<tr>
<td>Wang &amp; Strong (1996)</td>
<td>reliability</td>
<td>completeness</td>
<td>consistency</td>
<td>renewable</td>
</tr>
<tr>
<td>Weidema &amp; Wesnaes (1996)</td>
<td>accuracy</td>
<td>completeness</td>
<td>consistency</td>
<td></td>
</tr>
<tr>
<td>Redman (1998)</td>
<td>correctness</td>
<td>completeness</td>
<td>consistent representation</td>
<td></td>
</tr>
<tr>
<td>Levitin &amp; Redman (1998)</td>
<td>correctness</td>
<td>completeness</td>
<td>consistency</td>
<td></td>
</tr>
<tr>
<td>Hoxmeier 1998</td>
<td>correctness</td>
<td>completeness</td>
<td>consistency</td>
<td></td>
</tr>
<tr>
<td>Shanks &amp; Corbitt (1999)</td>
<td>accuracy</td>
<td>completeness</td>
<td>completeness</td>
<td></td>
</tr>
<tr>
<td>Kahn, Strong, &amp; Wang (2002)</td>
<td>free of error</td>
<td>completeness</td>
<td>completeness</td>
<td></td>
</tr>
</tbody>
</table>

Based on Table 1, the conclusion can be drawn that four papers discuss intrinsic data quality dimensions that fall outside of the fundamental dimensions defined by Ballou and Pazer (1985). These papers are by Wang and Strong (1996), Levitin and Redman (1998), Hoxmeier (1998), and Kahn et al. (2002). Two other papers have used the intrinsic data quality dimensions proposed by Wang and Strong (1996), who define believability and reputation as intrinsic. However, these are not intrinsic data quality dimensions because they are not all characteristics of the data themselves and the dimensions are dependent on data quality goals set by users. In addition, depreciability, share-ability (Levitin & Redman, 1998), and concise representation (Kahn et al., 2022) are also dimensions that are dependent on data quality goals set by users. These dimensions are dependent on the opinion of the users to be set as right or wrong.

Based on the arguments presented above, these four papers have not the same interpretation of intrinsic as Balou and Pazer (1985) have. Therefore, these papers will not be used with regard to the intrinsic data quality dimensions in this study. Second, because timeliness is difficult to categorise if it is intrinsic, the papers that categorise timeliness in the intrinsic dimension are also eliminated. Following these two eliminations, two papers have the basic intrinsic dimensions of Ballou and Pazer (1985), excluding the timelines dimension. These two papers have to be compared in order to choose which of the two has the best intrinsic data quality dimensions. These are the papers by Shanks and Corbitt (1999) and Wand and Wang (1996).

Shanks and Corbitt (1999) use the semiotic theory, in which the Syntactic and Semantic can be categorised as intrinsic. The two levels proposed Shanks and Corbitt (1999) have an accuracy-categorised dimension as intrinsic, together with a completeness and a consistency dimension. Thus, if a data quality error is defined in a certain level, it is not certain, for example, whether it is an accuracy or a completeness error. Furthermore, Shanks and Corbitt (1999) use the dimension of accuracy, which means that the recorded value that is given in the data is in conformity with the actual value that is given in the real world.

Wand and Wang (1996) use four independent dimensions. Two intrinsic dimensions define whether the recorded value is in conformity with the actual value that is given in the real world. Here the two dimensions specify mistakes from a vague definition to a specific measurable error. With Wand and Wang’s (1996) intrinsic dimensions, each data quality error can be categorised into one dimension.

Based on the comparison of these two articles, Wand and Wang's (1996) intrinsic data quality dimensions will be used for the benefits stated above. It is important to define the real world state and the information system state. The real world is defined as an application domain. The information system is defined as a representation of a real world system. Wang and Wand (1996) state that there is completeness if each real world state is mapped to an information system state. Thus, useful or not useful, the data have to be recorded in the concerned information state. This is in contrast to Ballou and Pazer’s (1985) dimension completeness, in which there is completeness if all necessary data are
recorded. Unambiguousness occurs when no two real world states are mapped to the same information system state. Meaningfulness is defined as: no meaningless information states. This means real world states are mapped to meaningful information systems. Correctness means that a real world state is mapped to a correct information state. All these dimensions can be defined as errors if the real world is wrongly mapped in an information state. In this case, Wang and Wand (1996) call the last two dimensions errors, meaningless and correctness errors, garbling. ‘Typically, garbling occurs due to incorrect human actions during system operation (e.g., erroneous data entry, or failure to record changes in the real world)’ (Wang and Wand, 1996).

2.2 Data quality and ERP systems

Due to the great number of implementations of ERP systems in organisations, as is noted by Myreteg (2015), data quality within or for these systems is becoming important. Therefore, several authors have elaborated implementation guidelines for ERP systems. This literature review focuses on the improvement of data quality within an existing ERP system. Therefore, the focus lies on post-implementation phases of the ERP systems.

Glowalla and Sunyaev (2014) facilitate an understanding of ERP systems and data quality interdependency by presenting the use of ERP systems for data quality management. Task Technology Fit (TTF) is a theory developed in order to assess linkages between information system (IS) use and individual performance depending on the IS’s fit for tasks (Goodhue and Thompson, 1995). TTF was applied in an explorative study, in which semi-structured expert interviews were conducted with participants in information technology strategic design making. Glowalla and Sunyaev (2014) present current practices of ERP system use in the insurance sector. The following main conclusions are made by Glowalla and Sunyaev (2014):

- Main use of ERP systems for administrative (standard) functions allows drawing on existing ERP system experiences and research from other (e.g. manufacturing) sectors.
- ERP system use, particularly for accounting, supports data quality management to comply with regulations in large insurance organisations.
- ERP systems provide a starting point for data analysis if data quality is reassessed for the new task and context.
- When focussing on interdependent, complex tasks (e.g. data analysis), sector specific approaches are more important and ERP systems and their data need to be considered within a broader organizational setting and system landscape.
- ERP system misfits arise continuously. Future research needs to be aware of ERP systems being embedded into increasingly complex information technology (IT) and organisational structures.

According to Glowalla and Sunyaev (2014), one possible solution process for data quality management is automation. Automated systems are essential in complex and challenging environments, such as Command and Control. Among the benefits related to automation, Breton and Bossé (2002) note: ‘The reduction of the operator’s workload’. Next, with automated systems, operators’ attentional resources can be allocated to other tasks executed concurrently. Furthermore, the reduction of the stress factor is induced by the stakes of the situation. In addition, there is a reduction of the fatigue factor, and automated systems provide a certain level of stability in the execution of a task. Finally, automated systems eliminate human errors.

Unfortunately, some cognitive costs are also related to the introduction of automated systems into the Command and Control environment. Manual skills may weaken in the presence of long periods of automation (Wickens, 1992). Automation removes the human from the loop, producing significant decreases in situation awareness (Sarter & Woods, 1992). Finally, over-reliance on automation may make the human less aware of what the system is doing, leaving the human ill-equipped to deal with
system failures (Scerbo, 1996). However, a potential solution to tackle the automated system introduction is to train the human to adequately supervise the system functioning.

The question is then how the data quality within an ERP system can be evaluated. Much literature discusses methodologies to improve data quality (Batini et al., 2009). Frameworks are needed that evaluate data quality within an ERP system with the intrinsic data quality used by Wand and Wang (1996). In the present study, two papers were identified that increase the understanding of the ERP system (( Xu & Nord, 2002) & (Haug et al., 2009)). For the following reasons, Haug et al. (2009) have the best framework with which to evaluate the ERP system based on data quality. First, Haug et al. (2009) use Wang and Wand’s (1996) intrinsic data quality, whereas Xu and Nord (2002) use the intrinsic data quality dimension timeliness. This dimension was defined in section 2.21 as a dimension that was hard to categorised as intrinsic. Second, Haug et al. (2009) evaluate an ERP system in the post-implementation phase, while Xu and Nord (2002) do so in the implementation phase. Finally, Haug et al. (2009) validate their results by conducting three case studies that confirm data quality improvement due to the evaluation framework.

Haug et al. (2009) state that the most relevant data quality categories when evaluating ERP system data are:

1. Intrinsic data quality dimensions: completeness, unambiguousness, meaningfulness, and correctness based on Wand & Wang (1996)
2. Data accessibility dimensions: access rights, storage in ERP system, representation barriers
3. Data usefulness dimensions: relevance, value-adding

To summarise, Haug et al. (2009) propose a classification model for evaluating data quality in ERP systems and defined the main causal relationships between categories of data quality dimensions. Three case studies conducted at three different companies, of which one had a SAP R3 ERP system, confirm that the classification model captures the most important aspects of describing ERP data quality and that the defined causalities between categories of data quality dimensions correspond to practice.

However, there are more ways to improve the data quality in ERP systems than only evaluate the actual data. Myreteg (2015) reviews the literature on organisational learning in the context of ERP systems in the post-implementation phase. There is a heavy dominance of studies concerning how to use the ERP system itself, rather than investigating how IT can support learning processes that could have operational, managerial, strategic, or organisational benefits. Myreteg (2015) identifies two patterns over time: first, a shift from the use of case or field studies to the use of surveys as the chosen research method; and second, a shift from organisational learning as a process to organisational learning as a critical success factor. He notes that the former influenced the latter. However, he also mentions that it is difficult to validate whether the observed patterns represent an actual trend. A learning process is to evaluate implementations of SAP R/3 and search for improvements by illuminating the main reasons of failures. Al-Mashari and Al-Medimigh (2003) do this by a case study of a failed implementation of SAP R/3 to re-engineer the business processes of a major manufacturer. The main reasons explained by Al-Mashari & Al-Medimigh (2003) are the following: scope creep, lack of ownership and transference of knowledge, lack of change management, lack of communication, lack of performance measurement, and propensity to isolate IT from business affairs. Al-Mashari and Al-Medimigh (2003) conclude that the following five core competencies are necessary: change strategy development and deployment; enterprise-wide project management; change management techniques and tools; BPR integration with IT; and strategic, architectural, and technical aspects of SAP installation.
2.3 Implementation in VDL Nedcar

Based on above findings, the evaluation model of Haug et al. is a great model where the SAP R3 ERP-system of VDL Nedcar can be evaluated. Based on the intake meeting and the interviews with employees of VDL Nedcar combined with the studied literature, it was noted that incorrectness and meaningless dimensions of intrinsic data quality errors occurred within the SCE department of VDL Nedcar. The main focus will lie therefore on the intrinsic data quality dimensions, which were defined above as the best choice. Also according to Haug et al. (2009) the accessibility dimensions are in the main focus that will have together with the intrinsic dimensions a causal relationship with the usefulness dimensions.

To eventually improve the data quality problems exposed with the model of Haug et al. (2009), findings of several literature on ERP-systems will be used ((Glowalla and Sunyaev, 2014), (Breton and Bossé, 2002), (Myreteg, 2015), and (Al-Mashari & Al-Medimigh, 2003)). Glowalla and Sunyaev (2014) shows that data quality misfits arise continuously and that the ERP-system at VDL Nedcar can thus be a starting point for data analysis; however, a reassessment is necessary to improve the data quality of new tasks and context. Within the article of Glowalla and Sunyaev (2014) also one possible solution process for data quality management is automation. Breton and Bossé (2002) show that in a complex and challenging environment, which the environment of VDL Nedcar is, automation have several benefits, if the user will be trained in such way that they keep their cognitive manual skills. Myreteg (2015) shows there are two patterns over time; a shift from the use of case or field studies to the use of survey as choice of research method; and a shift from organizational learning (OL) as a process to organizational learning as critical success factors (CSF). Finally, Al-Mashari & Al-Medimigh (2003) show that, however they mention prevention steps taken for preventing implementation of an ERP-system of failing, change of management techniques and tools can be redesigned afterwards. These can have positive effects on the working of the ERP system within VDL Nedcar.

This paper will use several conclusions made in this literature in the methodology to conduct the research. As main conclusion an automation tool will be used to reassess the ERP system at VDL Nedcar which is concluded from Glowalla and Sunyaev (2014). This tool will use the evaluation model of Haug et al. (2009) to do the empirical analysis. Furthermore, surveys will be chosen as research method as stated by Myreteg (2015). The results of the case study by Al-Mashari & Al-Medimigh (2003) will be keep on mind by the change of management techniques and tools, which has positive effects on the working of the ERP system. The paper combines all findings and creates a unique data improvement tool, which will contribute to the theory of data quality improvement in ERP systems in a production environment.

This chapter has highlighted the definition of data quality, data quality problems in ERP systems and how these can be assessed. In the next chapter the research approach will be elaborated to be able to do a good analysis.
3. The research approach
This section presents the research approach used in this study. It discusses the conceptual project design, the methodology, and the operational research plan. This provides an overview of the steps that were taken in this study.

3.1 Conceptual research design
The conceptual research design presents the outline of the research in an abstract way. In the conceptual research design, the subject of the analysis is defined. The subjects of the present analysis are the ERP system and the business process of the Supply Chain engineering department with regard to data quality. Next, the conceptual research design presents the theoretical perspectives applied in the analysis. It is an unrealistic to expect that all relevant theoretical perspectives can be combined in one integrated, homogeneous theory. Thus, the theoretical perspectives used in this study must be defined. By defining the theoretical perspectives, a clearer view is created of the scope of the problem, thus a conceptual research design is created. Furthermore, the deliverables of the research are provided: a diagnosis and an exploration of solution directions. Finally, a comparison is made between the theoretical perspectives, the subject of analysis, and the deliverables of the research. The following model presents the conceptual research design.

![Conceptual project design](image)

Figure 3 Conceptual project design

3.2 Methodology
To solve the data quality problems mentioned in Section 1.2.3, an automation tool was developed with the aim to ‘support operations, management, analysis, and decision-making functions in an organisation’, which according to Davis and Olson (1985) is a characteristic of artefacts in general. These data quality problems have practical as well as knowledge-related problems that need to be solved. Wieringa (2009) analyses the mutual nesting of practical problems and knowledge problems and derives methodological guidelines from this analysis. Practical problems call for a change of the world so that it better matches stakeholders’ goals. Knowledge problems call for change in knowledge about the logistical world.

In this research, the practical problem is to find a solution, such as an automation tool, whereas the knowledge problem is to define which knowledge rules should be implemented in this tool and which action should be taken to obtain a significant result.
Wieringa (2009) states that however the practical (research) and knowledge (design) problems may be different, they are closely related activities. For example, top-level questions are always practical, but in order to solve these practical problems, it might be necessary to first solve a knowledge problem. For practical problems, Wieringa (2009) uses the regulative cycle by Pieter J. Van Strien (1997), which is also used in the literature by Joan van Aken, Berends, and van der Bij (2012) to develop the problem solving cycle. The cycle steps have remained essentially the same as those originally proposed by Van Strien (1997); however, the later authors have added the analysis and the learning part to the diagnosis and the intervention phases, respectively, and have a problem mess as a starting point. For the knowledge problems, Wieringa (2009) uses the design guidelines described by Hevner, March, Park, and Ram (2010). This applies to the present project because the data quality problem is a problem mess and can be solved by following the steps proposed by the problem solving cycle, whereas the knowledge problems that occur can be solved with Hevner et al.'s (2010) design guidelines.

3.2.1 The problem solving cycle
Field Problem Solving research (FPS research) makes use of the problem solving cycle. This cycle is used to solve a business performance problem in the material world of action. The cycle consists of five steps, which are explained below.

![The problem solving cycle by Aken et al. (2007)](image)

3.2.1.1 Analysis and diagnosis
The first step after the problem definition of the problem solution cycle is the analysis and diagnosis. The purpose of the diagnosis is to validate the data quality problem, to explore and validate the causes and consequences of the data quality problem, and to develop preliminary ideas about alternative directions to solve the problem. This step can be divided into two approaches that are helpful in producing a diagnosis.

1. **Empirical analysis.** Here, the symptoms of the data quality problems, their potential causes, such as wrong entries, and their potential consequences had to be identified. In addition, evidence to support the analysis had to be gathered by interviewing several stakeholders. The problem statement had to be validated with factual information. This was done by analysing the master data management list with the pivot method. Furthermore, the problem statement was validated with, for example, stories of situations in which the data quality problem occurred. These were obtained through interviews with stakeholders. Once the validation of the problem was established, its causes could be investigated. Important input for this diagnosis was the cause-and-effect tree. However, it was unlikely that the orientation
phase had already provided an exhaustive overview of potential causes that only needed to be verified. The potential causes of the data quality problems needed to be validated, so each element and relationship between elements in the cause-and-effect diagram had to be examined. In order to validate these cause-and-effect relationships, reliable and valid data were needed. This was solved by using data or people within VDL Nedcar who provided evidence of a cause. In order to claim a certain level of reliability, some inter-subjective agreement had to be targeted when interviewing people within the organisation. This means that there had to be at least two people from different departments who shared the relevant beliefs. In addition, the problem was further analysed. All of the errors that occurred were presented in an overview and categorised with the help of the work of Anders Haug et al. (2009), which was found in the literature review to be the most appropriate method for improving quality in an ERP system.

2. Process-oriented analysis. If the focus stayed too much on the cause-and-effect tree, there was a high risk of staying away from the real business processes and producing a superficial analysis. The idea of the process-oriented approach was to develop a general description of the process of VDL Nedcar, including operational processes and the control system, as well as performance norms. The purpose was to gather a background understanding of the business process, such as knowledge sharing. Therefore, 7 respondents from different disciplines answered questionnaires regarding the data quality issue. With the resulting information, the process knowledge of the problem was enlarged, which was useful in creating the solution design. With the main stakeholders, it was possible to determine knowledge about certain combinations in which the data needs to meet the specifications of the process flow. Because most employees are not familiar with all of these combinations, these interviews were conducted with two main stakeholders. These stakeholders were situated in two different departments to ensure objectivity.

3.2.1.2 Solution design
The next step in the problem solving cycle is the solution design. Analysis and design are two activities that differ in character. In an analysis, the dominant logic is from question to answer, using a closed question. In design, it is from solution to design requirements, and the question is open-ended. A design is not an end in itself but an input for the next step in the creation process. A design is a model of a possible future reality, and professional designing involves the development of two designs:

- Object design, which is the model of the system or process, to be realised. This is the automation tool that needs to be realised.
- Process design, which is a design of the process of the analysis, and design itself that is used to produce the object and realisation design. This section will discuss which steps were taken to develop a solution.

First, a design was chosen to refer to a design solution that was successful and influential, based on the literature review.

In the process design, a process flow was created to research the completeness of the data in the ERP system by comparing all the real world states with the information states in the as-is situation. If a real world state is not defined in the information system, the data will not be complete.

With the help of the process flow, the ambiguousness could also be researched. If for example two real world states are listed in one information state, this is called ambiguous. The value of this information state cannot be determined correctly due to the lack of detailed information.
Meaningfulness and correctness are discussed together because they both describe a different information state than is stated in the real world. Often this happens due to mapping the real world state into a wrong information state; Wand and Wang (1996) call this garbling, as stated in Section 2.1 Data quality. By finding the garbling, the automation tool will be used such that the data quality will improve. Therefore, the automation tool will henceforth be called the data improvement tool in this thesis.

To develop this data improvement tool, all of the knowledge had to be determined. This knowledge was gathered by interviewing the two main stakeholders (experts) of the data, who represented two separate departments. Furthermore, this had to be validated. The pivot method was used to analyse the data and find the relationships between these data, which could then be used to validate the knowledge. With this information, a Pareto chart could be developed to determine the most important rules.

With the gathered knowledge, an object design could be created to illustrate how the data improvement tool would work the most efficiently. The process flow could also been used to determine the program steps that had to be taken in the development phase.

Furthermore, it was established in which system the tool would be made. If this choice was made together with stakeholders within VDL Nedcar, it could be determined how this tool could be implemented to check the SAP data.

Finally, the data improvement tool could be developed with the help of the aforementioned knowledge and process flows. This data improvement tool was constructed, taking account its accessibility and usability.

The actual designing started with design requirements: the demands that the realised tool had to meet. Two categories of such requirements were determined:

- Functional requirements, which described what the system had to do; these constituted the core of the requirements, in the form of performance demands on the object to be designed;
- Non-functional requirements, which described how the system would work; these constituted user requirements for representation.

These design requirements will be described in section 6.1 using the MOSCOW method, which contains the following requirements:

- Must have: requirements that are critical to the project. Without these requirements, the project is a failure.
- Should have: requirements that are important, but not critical to the project. While these can be as important as must-have requirements, they can be solved differently if necessary and solved in the future as required.
- Could have: requirements that are desirable but not necessary.
- Won’t have: requirements of such little importance that they are not planned.

These design requirements were identified using interviews with the main stakeholders. The interviews were developed based on the findings of the diagnosis. Together with the design requirements, the development of the change plan began. The change plan will be elaborated after the design process. After the design requirements, the synthesis will take place.
The basic actions in the actual designing process were synthesis-evaluation iterations. First, a synthesis of rules was made, and then this synthesis was evaluated with concepts and logical reasoning. Thus, the data improvement tool was created and all of the requirements and conditions were evaluated. If the evaluation was negative, meaning that the tool produced an error, the synthesis step started again. This was repeated until the evaluation step gave a positive answer. An error can be the result of a simple type-error or an error in the design requirements according to time boundaries. The synthesis step cannot be logically derived from the input of the diagnosis and analysis step. According to Charles Sanders Peirce (1923), the synthesis always involves an abduction. Charles Sanders Peirce (1923) states that there are three modes of inference: deduction, induction, and abduction. Whereas deduction and induction follow an unbroken chain of reasoning, abduction involves an element of guessing. This is the creative jump that the designer takes to come to an answer.

Thus, the following design process could be determined and followed to reach to a well-engineered data improvement tool.

- Problem analysis: the data quality was investigated with empirical, theoretical, and process-oriented analyses; this was done in greater depth than in the problem definition phase, with the use of SPSS.

- Specifying the design requirements: these were identified in the interviews with the stakeholders.

- Sketching: a solution concept was created by modelling a design.

- Creating the outline design:
  - Determining the design requirements of this solution concept;
  - Determining the values of the design requirements;
  - Making an overall assessment of the outline design; and

- Detailing: creating the various detail designs.

3.2.1.3 Validation of the tool
The outcome of the design, the data improvement tool, had to be validated at the end of the design process. Because a design existed in the as-is situation, the developed design is a redesign. Jansen-Vullers, Looschilder, Kleingeld, and Reijers (2007) investigate which factors should be used to evaluate a redesign. They suggest validating the results of the redesign on the four dimensions of the Devil’s quadrangle: time, costs, quality, and flexibility. This is strengthened by the fact the Devil’s quadrangle (Brand & van der Kolk, 1995) was used to evaluate a redesign in the lectures of Business Process Management (BPM) (Dr. Ir. Vanderfeesten, 2014). Validation uses the results of the evaluation steps of the actual design process, but is different from these evaluation steps, as they are taken from the perspective of the developer. Validation of the data improvement tool was done together with the organisation, in order to decide whether or not to implement the design solution. This validation can be divided into four basic steps that needed to be executed:

- An evaluation using the four dimensions of the Devil’s quadrangle by N. Brand and H. van der Kolk (1995), which consists of time, costs, quality, and flexibility.

- A validation check, with the use of the pivot method; and

- A report in which experts and stakeholders confirmed that the tool is easy to work with and works properly and why the experts are of the opinion that the tool is a solution to the problem;
3.2.1.4 Additional information

Appendix IV discusses the design research guidelines, the operational research plan, the cost of the project, and the organisation of the project.

Now it is clear what the assignment exactly is and a method is designed to solve this assignment, the next chapter will explain the meaning of data quality.
4. Data collection
Before the problem could be properly analysed, information was gathered that influenced the problem scope and provided insight that could lead to a better analysis. This was done by observation of floor processes, extracting data from SAP, and interviewing stakeholders about the data. This chapter will provide the different techniques used to collect the data for the research. First, the questionnaire is explained. Second, the floor processes and the employees are observed. Third, the downloading of the data is explained and finally all the interviews are clarified.

4.1 Questionnaire
To gain process knowledge regarding the problem, a survey was conducted. This was concluded based on the findings of Myreteg (2015). Seven stakeholders (including two experts) were asked to answer a questionnaire; the respondents divided over the MHE and the SCE departments. The questionnaire was used to obtain an insight into the impact of the data quality problems. The questionnaire addressed topics that are important to decision-making in the solution design. The topics were divided over 15 questions and were data awareness, process knowledge, communication, knowledge sharing, authorisation, change management, change process, controllability, and accessibility.

The questionnaire was administered digitally to provide all stakeholders with anonymity and to give them the freedom to plan their participation. Every question could be answered on a scale of 1 to 5. 1 meant that the level was bad, 2 meant that the level was weak, 3 meant that the level was enough, 4 meant that the level was good, and 5 meant that the level was very good. The questionnaire is provided in appendix V.

The results of the questionnaires are shown in Figure 5.

![Business knowledge questionnaire](image)

Figure 5 Stakeholders’ questionnaire results
4.2 Observation of the floor processes and the employees
The researcher visited the floor with a material handling engineer (MHE) and watched the proceedings executed by the Material Handling Floor employees (MHF). He also observed the communication between the MHE and the MHF employees to gather significant information regarding the data quality problem. By observing the communication of the MHE and the MHF employees, information was gathered that confirmed several of the causes of data quality problems mentioned previously. It became clear that many MHF employees adjust data if they have to be corrected; several did not know all codes regarding the process flow; and several MHF complained about the communication with Supply Chain Engineers (SCE).

4.3 Extracting data
The data that were analysed were extracted from SAP. Four lists were needed, and two of the four had to be taken from SAP. These were two lists that are constantly updated with information regarding the aspects of the materials to ensure a good throughput of the process. The other two lists were reference lists, in which all correct data for certain processes were stated. The entire lists are indicated below:

- 431 list (Master Data Management data, SAP)
- 921 list (Master Data Management data, SAP)
- Handlings units (reference list)
- List of emballage codes (reference list)

For the scope of the data quality for the Supply Chain Engineering department, not all columns of these lists were important. The important aspects were identified with the use of interviews.

4.4 Interviews
To gather the right information knowledge about the data quality problems, interviews were held with stakeholders who have to work with SAP. For more specific knowledge about the data quality problem, interviews were also conducted with experts.

4.4.1 Stakeholder interviews
Interviews were conducted separately with each of the five stakeholders. Because of activities, four of the stakeholders fill in data in the SAP system, while the last stakeholder manages these data in the Supply Chain Engineering department. The interviews were standardised, open-ended interviews. In these interviews, all five stakeholders of the SCE department who work with SAP were asked the same questions, and were free to choose how to answer them. One expert was among these five stakeholders. This made it easier to analyse the results, as they could more easily be compared. The questions that were asked concerned the following topics:

1. Demographic: what is the function of the stakeholder?
2. Behaviour: how does the person act with the system when he works?
3. Opinions/values: what does the stakeholder think of the data quality issue?
4. Sensory: what has the stakeholder experienced?

4.4.2 Expert stakeholder interviews
Interviews with three experts (the SCE, the information management and the MHE) were conducted. These interviews were of the general interview guide approach type; this ensured that the same general areas of information were collected from each interviewee. Thus, there was a focus on the
problem, but there was also freedom and adaptability in gaining information from the interviewees. The interviews were conducted in a setting with little distraction, and most of the time (if the rooms were available) in the interviewee’s department, in order to foster the comfort of the interviewee.

The purpose of the interview was then explained and the length of the interview was specified. Furthermore, the interviewees were told that if they had questions, they could always ask them in the interview due the general interview guide approach that was used. Written notes were taken of the interview with permission from the interviewee; these were then supplemented by extensions on those notes immediately following the end of the interviews. Van Aken et al. (2012) state that interviews can be considered reliable if the experts are in different departments and do not influence each other.

4.4.3 Knowledge collection
To understand the downloaded lists discussed in Section 4.3, and to be able to properly analyse them, knowledge needed to be gathered using interviews with the stakeholders and the experts.

4.4.3.1 Basic knowledge
For two of the downloaded lists – the 431 and 921 lists mentioned in Section 4.3, the basic knowledge was gathered from the interview with the SCE expert to ensure a well-performed analysis. The basic knowledge consists of knowing which headings are useful to analyse and what all of the headings mean. All of these lists describe several aspects of each material, so that the researcher knew what had to be done where and at which moment, as well as which references were important for the scope of this project. The basic knowledge that was needed is presented in Table 2. In the left column, all of the headings that are relevant in this scope are given, while in the right column the meaning of these headings is stated.

Most of the headings have definitions that are clear and easy to understand, such as the dimensions of the packing. However, other headings needed more explanation to understand their use in the research. The headings and their definitions are as follows.

**Flowcode Material Master (MM):** The flowcode indicates which of the in-house streams the material will follow. These streams consist of the SILS, JIS, JIT, and the WH, which were explained in section 1.2.1.2.2 In-house.

**Internal Process Code MM:** There are several codes that define the internal process. These codes indicate more detailed process steps that must be taken on the floor.

**Laboratory office MM:** This refers to the different shops where the cars are built, which were mentioned earlier, in the introduction to the organisation. The shops are defined as follows:

- 001: Final Assembly Shop
- 003: Body Shop
- 008: Paint Shop

**Goods Supplier Product Scheduling Agreement (PSA):** A unique number of the supplier.

**Storage Location PSA:** The location where the material is stored. For different materials there are different locations to deal with these materials.

**Unload Gate PSA:** The gate where the material has to be brought into the factory to ensure the shortest and most efficient route to its destination.
Cost Centre PSA: This number refers to the person paying. The only value that this column can have is 53400.

With this basic knowledge about the headings that were given, the knowledge about the specific right combination of several headings could be taught by the experts to develop a reference list containing the right data.

Table 2 Relevant heading definitions

<table>
<thead>
<tr>
<th>Relevant Heading</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Material Master</td>
<td>MM</td>
</tr>
<tr>
<td>MatDescMM</td>
<td>Material Description MM</td>
</tr>
<tr>
<td>A</td>
<td>Car type A MM</td>
</tr>
<tr>
<td>B</td>
<td>Car type B MM</td>
</tr>
<tr>
<td>C</td>
<td>Car type C MM</td>
</tr>
<tr>
<td>FLC</td>
<td>Flowcode MM</td>
</tr>
<tr>
<td>IPC</td>
<td>Internal Process Code MM</td>
</tr>
<tr>
<td>LO</td>
<td>Laboratory Office MM</td>
</tr>
<tr>
<td>GsPsA</td>
<td>Goods Supplier PSA</td>
</tr>
<tr>
<td>ProjNrMM</td>
<td>Project Number MM</td>
</tr>
<tr>
<td>AccAssCatP</td>
<td>Account Assignment Category PSA</td>
</tr>
<tr>
<td>SlocPSA</td>
<td>Storage Location PSA</td>
</tr>
<tr>
<td>UnloadPSA</td>
<td>Unload Gate PSA</td>
</tr>
<tr>
<td>CostCntrPS</td>
<td>Cost Centre PSA</td>
</tr>
<tr>
<td>Pack.obj.</td>
<td>Packaging Object Number MM</td>
</tr>
<tr>
<td>TrEmbCdePI</td>
<td>Transport Emballage Code Packing Instruction</td>
</tr>
<tr>
<td>GenItCatGr</td>
<td>Generic Item Category Group MM</td>
</tr>
<tr>
<td>ExtEmbCdeP</td>
<td>External Emballage Code Packing Instruction (PI)</td>
</tr>
<tr>
<td>StkRmvlMM</td>
<td>Stock Removal MM</td>
</tr>
<tr>
<td>StkPlcmntM</td>
<td>Stock Placement MM</td>
</tr>
<tr>
<td>SsiMM</td>
<td>Storage Section Indicator MM</td>
</tr>
<tr>
<td>Sut1MM</td>
<td>Storage Unit Type 1 MM</td>
</tr>
<tr>
<td>Sut3MM</td>
<td>Storage Unit Type 3 MM</td>
</tr>
<tr>
<td>TrgQtyRMat</td>
<td>Target Quantity of Ref Material PI</td>
</tr>
<tr>
<td>QtyPerPMat</td>
<td>Quantity (Ref) Material per 1 Packing material</td>
</tr>
<tr>
<td>Sut1QtyMM</td>
<td>Loading Equipment Quantity 1 MM</td>
</tr>
<tr>
<td>Sut3QtvyMM</td>
<td>Loading Equipment Quantity 3 MM</td>
</tr>
<tr>
<td>LghtDimPI</td>
<td>Length Dimension PI</td>
</tr>
<tr>
<td>WdthDimPI</td>
<td>Width Dimension PI</td>
</tr>
<tr>
<td>hghtDimPI</td>
<td>Height Dimension PI</td>
</tr>
<tr>
<td>TotVolPI</td>
<td>Total Volume Dimension PI</td>
</tr>
<tr>
<td>SUTLEC</td>
<td>Storage Unit Type in List of Emballage Code</td>
</tr>
<tr>
<td>HandlingUnit1DimLength</td>
<td>Length Dimension in Handling Unit</td>
</tr>
<tr>
<td>HandlingUnit1DimWidth</td>
<td>Width Dimension in Handling Unit</td>
</tr>
<tr>
<td>HandlingUnit1Height</td>
<td>Height Dimension in Handling Unit</td>
</tr>
<tr>
<td>HandlingUnit1CalculatedVolPI</td>
<td>Volume Dimension Calculated With Handling Unit</td>
</tr>
</tbody>
</table>

4.4.3.2 Advanced knowledge
From the results of the interviews held with the two experts, advanced knowledge was derived regarding the fixed combinations between the data columns. This means that if a column has a specific value, the value in another column can only be one value or a range that is narrower than
before. These are rules what the stakeholders of the SCE department need to follow. Interviews were held with the two experts separately to ensure that all of the knowledge about these rules was obtained. Forty-four possible relations were identified; they are shown in Table 3. For each row, a group could be defined with the help of the fixed relation. This list could be used as a reference list to analyse the data lists downloaded and find the number of mismatches to validate the stated problem.

Table 3 Experts’ advanced knowledge regarding the data lists

<table>
<thead>
<tr>
<th>Relation</th>
<th>IF</th>
<th>Then</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO_IPC_FLC_VS_STOCRMVL</td>
<td>LO:001 + IPC: E8,E2</td>
<td>StKRMVL: Blanks</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>E7</td>
<td>60LC + M75, M10, (M70?)</td>
</tr>
<tr>
<td>LO_IPC_FLC_VS_STOCRMVL</td>
<td>LO:001_FLC:WH_IPC:E7_UNLOAD PSA:M75</td>
<td>623,626,6TB</td>
</tr>
<tr>
<td>SStM VS LO</td>
<td>LO: 008 Ssim: F01-F04</td>
<td>Are not allowed to occur</td>
</tr>
<tr>
<td>LO_IPC_FLC_VS_STOCRMVL</td>
<td>LO:001_IPC:E7_FLC:WH_UNLOAD PSA:M70</td>
<td></td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>E5</td>
<td>60SH + all M except M-75, M10 + no C,V</td>
</tr>
<tr>
<td>LO_IPC_FLC_VS_STOCRMVL</td>
<td>LO:001 + IPC: E1 +FLC: SILS</td>
<td>Blank V INC</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>C1</td>
<td>20LC+all C named UNLOAD PSA</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>D3</td>
<td>40LC AND V-15</td>
</tr>
<tr>
<td>AccAssCatP_VS_IPC &amp; CostCntr &amp; StocRplcmnt</td>
<td>AccAssCatP: k</td>
<td>IPC: D1,D2 &amp; CostCntrPS: 53400 &amp; StocPLcmnt: 430</td>
</tr>
<tr>
<td>LO_UNLOAD PSA</td>
<td>1</td>
<td>all M + SILLS+TWMC+V15</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>E2,E3,E8</td>
<td>all M except M-75 + M10 no C,V</td>
</tr>
<tr>
<td>LO_IPC</td>
<td>1</td>
<td>3A,D3, all E’s</td>
</tr>
<tr>
<td>LO_UNLOAD PSA</td>
<td>8</td>
<td>all V’s</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>E1+SILS</td>
<td>6A00xNot M70vM75vM10vV-XXvC</td>
</tr>
<tr>
<td>LO_IPC_FLC_VS_STOCRMVL</td>
<td>LO:003_IPC:C1_UNLOAD PSA:C25xx</td>
<td></td>
</tr>
<tr>
<td>LO_IPC</td>
<td>8</td>
<td>D1,D2,D3</td>
</tr>
<tr>
<td>LO_IPC_FLC_VS_STOCRMVL</td>
<td>LO:001_FLC:WH_IPC:3A</td>
<td>645</td>
</tr>
<tr>
<td>LO_IPC_FLC_VS_STOCRMVL</td>
<td>LO:001 + IPC: E4 +FLC: WH</td>
<td>610</td>
</tr>
<tr>
<td>LO_UNLOAD PSA</td>
<td>3</td>
<td>never with UNLOAD PSA: M</td>
</tr>
<tr>
<td>LO_IPC_FLC_VS_STOCRMVL</td>
<td>LO:001_FLC:WH_IPC:E7_UNLOAD PSA:M10</td>
<td>626, 6TB</td>
</tr>
<tr>
<td>Sut3MM</td>
<td></td>
<td>Sut3MM: TBD</td>
</tr>
<tr>
<td>SutQTyMM</td>
<td></td>
<td>SutQTyMM: 999999</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>E1 + WH</td>
<td>TW01xTWMC</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>FSC</td>
<td>6A00 + SILLS</td>
</tr>
<tr>
<td>LO_IPC</td>
<td>3</td>
<td>C1,C5, D3</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>E4</td>
<td>M20 AND 60BL</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>D2</td>
<td>40LC AND V15</td>
</tr>
<tr>
<td>FSC_IPC</td>
<td>JIS</td>
<td>E2 v E8</td>
</tr>
<tr>
<td>FSC_IPC</td>
<td>SILS</td>
<td>E1</td>
</tr>
<tr>
<td>FSC_IPC</td>
<td>JIT</td>
<td>E3</td>
</tr>
<tr>
<td>FSC_IPC</td>
<td>WH_003</td>
<td>C1, C5, D3</td>
</tr>
<tr>
<td>IPC_VS_UNLOAD PSA_AND_SLOC</td>
<td>3A</td>
<td>TWMC AND TW01</td>
</tr>
</tbody>
</table>
All the information needed is gathered with the help of the above explained data collection methods and will be analysed in the next chapter to improve the knowledge about the problem to come to a well-defined solution design.
5 Data analysis

This chapter analyses the collected data and subsequently draws conclusions. The chapter is divided into an empirical analysis and a process-oriented analysis. The purpose of the empirical analysis is to validate the business problem and specify its characteristics, to explore the causes, and to validate those causes. The purpose of the process-oriented analysis is to obtain a background understanding of the business process, and with this background information come to a better solution design. The business process is analysed to gain insight into the impact of the data quality problem.

5.1 Process-oriented analysis

As mentioned in Section 3.3.2, the analysis phase started with the process-oriented analysis. The business process was analysed to gain insight into the impact the data quality problem. This was done in two ways: with the development of a process flow, and with a questionnaire about business process knowledge considering the data quality issues.

5.1.1 Questionnaire

The results given in Figure 5 are specific; therefore, conclusions can better be drawn after clustering the questions across the topics that were specified earlier. The clustering was done by taking the same topics of the questions together with the specific topics defined earlier, and taking the average of the original scores initially obtained for each question. With this clustering, the question topics were narrowed down from 15 to 9. Table 4 presents the results of this clustering.

Table 4 Clustered questionnaire results

<table>
<thead>
<tr>
<th>Business knowledge subjects (BKS)</th>
<th>Scores</th>
<th>Clustered BKS</th>
<th>Scores BKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data awareness</td>
<td>3.6</td>
<td>Data awareness</td>
<td>3.6</td>
</tr>
<tr>
<td>Process knowledge</td>
<td>3.9</td>
<td>Process knowledge</td>
<td>3.9</td>
</tr>
<tr>
<td>Communication</td>
<td>2.3</td>
<td>Communication</td>
<td>2.3</td>
</tr>
<tr>
<td>Data change management</td>
<td>2.3</td>
<td>Change management</td>
<td>2.5</td>
</tr>
<tr>
<td>Data change history storage management</td>
<td>2.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centralised change management</td>
<td>2.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change process clarity</td>
<td>3.1</td>
<td>Change process</td>
<td>2.7</td>
</tr>
<tr>
<td>Change process standardisation</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change process simplicity</td>
<td>2.6</td>
<td>Controllability</td>
<td>2.2</td>
</tr>
<tr>
<td>Controllability simplicity</td>
<td>2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controllability frequency</td>
<td>2.3</td>
<td>Controllability</td>
<td>2.2</td>
</tr>
<tr>
<td>Authorisation level</td>
<td>2</td>
<td>Authorisation level</td>
<td>2</td>
</tr>
<tr>
<td>Knowledge sharing</td>
<td>2.4</td>
<td>Knowledge sharing</td>
<td>2.4</td>
</tr>
<tr>
<td>Accessibility</td>
<td>3.4</td>
<td>Accessibility</td>
<td>2.8</td>
</tr>
<tr>
<td>Accessibility knowledge</td>
<td>2.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In contrast to what was expected, process knowledge was found to be high in the departments of the SCE and the MHE. As Table 4 illustrates, the worst scoring element was the authorisation. Stakeholders believe that there is far too little authorisation, which results in data quality problems. Data awareness and process knowledge both scored between ‘enough’ and ‘good’. Communication, knowledge sharing, change management, change process, controllability, and accessibility all scored between ‘weak’ and ‘enough’.

5.1.2 Unload PSAs

It was not possible to involve the WHF employees with documentation of the data regarding errors and how long it takes to solve these errors. Therefore, the employees of the MHF department were asked to provide an estimate of the time they spent solving errors in the current situation. In Figure 6, the most used unload PSAs are illustrated with the help of a Pareto chart in SPSS, derived from the 431 list.
As can be seen in the chart, most materials are unloaded at the unload PSAs M-XX. Therefore, a sample was taken of these unload PSAs. The other unload PSAs have such little material in comparison with the sample of the M-XX that they were omitted from the analysis. This sample was set by the MHE expert to ensure that the most important unload PSAs that best represent the factory were selected.

Table 5 Data quality error correction times

<table>
<thead>
<tr>
<th># gates</th>
<th>M-XX</th>
<th>Time</th>
<th>Unit of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M10</td>
<td>45</td>
<td>minutes a day</td>
</tr>
<tr>
<td>2</td>
<td>M05</td>
<td>60</td>
<td>minutes a day</td>
</tr>
<tr>
<td>3</td>
<td>M75</td>
<td>20</td>
<td>minutes a day</td>
</tr>
<tr>
<td>4</td>
<td>M70</td>
<td>20</td>
<td>minutes a day</td>
</tr>
<tr>
<td>5</td>
<td>M55</td>
<td>6</td>
<td>minutes a day</td>
</tr>
<tr>
<td>6</td>
<td>M25</td>
<td>12</td>
<td>minutes a day</td>
</tr>
<tr>
<td>7</td>
<td>M85</td>
<td>6</td>
<td>minutes a day</td>
</tr>
<tr>
<td>8</td>
<td>M65</td>
<td>6</td>
<td>minutes a day</td>
</tr>
<tr>
<td>9</td>
<td>M60</td>
<td>9</td>
<td>minutes a day</td>
</tr>
<tr>
<td>10</td>
<td>M18</td>
<td>6</td>
<td>minutes a day</td>
</tr>
<tr>
<td>11</td>
<td>M36</td>
<td>6</td>
<td>minutes a day</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>196</td>
<td>minutes a day</td>
</tr>
</tbody>
</table>

The sample shows the total time spent changing the data is 196 minutes a day, or 3.27 hours. The costs of an MHE can be calculated with the following formula:

\[
\frac{\text{Labour tariff per year} + \text{surcharge}}{\text{Individual working hours per year}} = \text{costs made per hour}
\]

An MHE employee costs on average 36.15 euro/hour, which means that every day the data quality problem costs:

\[
\text{average cost MHE} \times \frac{\text{Euro}}{\text{hour}} \times \frac{\text{duration of work}}{\text{hours}} = \text{costs made per day}
\]

\[
36.15 \times 3.27 \times \text{hours per day} = 117.50 \text{ euro per day}
\]
Below the state of the art process flow is explained, demonstrating which steps are taken to detect and solve data quality problems.

### 5.1.3 Process flow

The process flow was analysed and a process flow diagram was developed together with the experts. In this section the state of the art process flow is displayed, illustrating the process of materials from entry to storage and indicating which steps are taken to detect and solve data quality problems. Figure 7 presents the flow chart, which illustrates the steps that are taken when the department solves data quality problems when they occur. Because this study focuses on the data in SAP within the SCE department, the flow chart of the SCE department includes the activities done regarding the data in SAP. As can be seen, the SCE inserts data into the master data of SAP. After this step, four scenarios can occur. In one of the scenarios, nothing happens and the SCE has finished filling the data in SAP. In the other three scenarios, the SCE receives an email from the MHE department or the LSP department with a request to change data in SAP because an error has occurred. The SCE has to change these data and, if no further request is received, the task is then finished.

At the aforementioned three points in the process, employees may detect the data quality problems because a bottleneck is encountered in the process flows. These points are displayed in Figure 7. In the process flow of Material Handling, which is presented in Appendix VI, there are two points at which a bottleneck may be experienced. If this occurs and the bottleneck cannot be solved by the Material Handling department, the SCE department is contacted to adjust the data in order to solve the problem. In the LSP department, which is also presented in Appendix VI, there is one bottleneck. Here the same steps are taken as in the Material Handling department. At this point, as shown in Section Unload PSAs, MHF employees considerable time with correcting and contacting SCE about data quality problems. This also leads to a higher rate of corrections that have to be made by the SCE expert.

Up to three departments may be necessary to solve the data quality problem, which means that up to three departments may change the data. This costs a great deal of time and can cause more data quality problems because of the communication and the authorisation aspects that occur in this process.
5.2 Empirical analysis

As mentioned in Section 3.3.2, the analysis phase started with the detection of the symptoms of the data quality problems. This was accomplished by defining the potential causes and gathering evidence to support these causes via interviews with stakeholders. These two steps were achieved by interviewing several stakeholders and by observing the process in the factory.

5.2.1 Stakeholders interviews

The interviews identified 18 potential causes of the data quality problem. All causes have been selected that were mentioned by at least three stakeholders as a main cause. This results in five main causes of the data quality problem. These causes are listed below, along with the number of times they were mentioned by the stakeholders between parentheses.

- Different people from different departments adjust the data. (5)
- The data adjustment process is a long process that is not explained properly, and the systems are often slow. (3)
- There is too little process knowledge among stakeholders. (3)
- Communication between stakeholders is not correct. (4)
- If changes are communicated, there is lack of feedback regarding whether they are successful; this is taken for granted. (3)

Next to the potential causes, the potential consequences were also gathered in the interviews. These interviews provided 12 potential consequences of the data quality problem. All consequences have been selected that were mentioned by at least three stakeholders as a main consequence. This results in four main consequences of the data quality problem. These consequences are listed below, with the number of times they were mentioned by the stakeholders between parentheses.

- It is difficult to see what is changed and why it is changed. Sometimes decisions are reversed because stakeholders ‘think’ they could be better. (3)
- There is damage to material due to there being too much material in a packing. (3)
- Products are in a process flow other than the process indicated in the system. (5)
- MH has problems with process flow of product due to wrong data. (4)
- This costs time and therefore money. (5)

5.2.2 Problem statement analysis

Having finished the empirical analysis, this section will present the validation of the problem statement with factual information. This step was done using the Master Data list and the process knowledge of the two experts. In addition, the cause-and-effect tree of the research proposal was further analysed with the information gained from the empirical analysis.

5.2.2.1 Master data error analysis

Using the reference list collected with the experts, the downloaded data lists could be validated and analysed. The validation phase will first be explained, followed by the analysis phase.

5.2.2.1.1 Validation

The fixed combinations provided by the experts were validated using the pivot table method, which ensured that all of the fixed combinations were included. With this method, a number of abnormal values were found that were not defined by the experts as fixed combinations. These combinations were fed back to the experts and were confirmed as fixed combinations that they had forgotten to define. Eventually, all of the fixed combinations were included with this method. An example of the pivot table method is shown in Table 6. The values of Table 7 are filled into the pivot table in Table 6.
Table 6 Pivot table example

<table>
<thead>
<tr>
<th>Relation</th>
<th>If</th>
<th>Then</th>
<th>mistakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO:001 + IPC: E1 + FLC: SILS</td>
<td>Blank V INC</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

Table 7 Fixed combination of the experts check

<table>
<thead>
<tr>
<th>Relation</th>
<th>If</th>
<th>Then</th>
<th>mistakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO_IPC_FLC_VS_STOCRML</td>
<td>LO:001 + IPC: E1 + FLC: SILS</td>
<td>Blank V INC</td>
<td>12</td>
</tr>
</tbody>
</table>

The example demonstrates that if all the columns are filled in with the values in Table 6, then the STOCRML values are Blank (″) and ‘6TB’. The ‘Then’ column in Table 7 indicates that there have to be Blanks or INC; therefore, there are 12 mistakes (the 12 6TBs) (Table 7).

During the validation, not only were new combinations found, but garbling errors were also identified. For example, there were unloading PSAs with the notation M75 and M-75. When these explored errors were presented to the expert, it was concluded that these notations had the same meaning, and thus there was a meaningless state error. Furthermore, correctness errors were also found. No unambiguousness or completeness errors were found because of the lack of detailed information on the process.

An attempt was made to validate the fixed combinations with SPSS as well, but nominal values cannot be converted to variable values on the interval or ratio ordinal level. The most appropriate method using SPSS would also have been the pivot table method for validating this type of data, and this had already been done.

5.2.2.1.2 Analysis

The reference list collected during the interviews with the experts could be used to analyse the data lists downloaded in combination with the validation list. This validation list of correct data provided all mismatches in the data. These mismatches were counted, which resulted in an overview of how many mismatches there were in the current database, and provided a percentage of the mismatches found.

This was done by calculating the cumulative errors, and then comparing this cumulative errors to the total number of data to obtain the percentage of errors in the data set.

Furthermore, a categorisation was made of the different kinds of errors, and which errors could be treated and considered to fall within the scope of the project. The errors/mismatches were categorised into blanks and wrong entries because if the data were blank, it could also mean that the information of the material was not yet available.

In Table 8, the same example is given as before in the validation section.
Table 8 Calculation table of errors

<table>
<thead>
<tr>
<th>Relation</th>
<th>Columns are</th>
<th>Other column restrictions</th>
<th>Wrong entries</th>
<th>Blanks</th>
<th>Mistakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO_IPC_FLC_VS_ST</td>
<td>LO:001 + IPC: E1 + FLC: SILS</td>
<td>Blank V INC</td>
<td>12</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cumulative wrong entries</th>
<th>Cumulative percentage wrong entries</th>
<th>Cumulative blanks</th>
<th>Cumulative percentage blanks</th>
<th>Cumulative mistakes</th>
<th>Cumulative percentage mistakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0,1%</td>
<td>0</td>
<td>0%</td>
<td>12</td>
<td>0,1%</td>
</tr>
</tbody>
</table>

Because it was not possible to determine with certainty whether the blanks were wrong entries or not filled yet, they were omitted, and only the wrong entries were used for analysis and optimisation. Summing all of the wrong entries and dividing them by the total amount of running data resulted in a 5.5% error rate in the BOM that consists of 12255 lines. These data provided an opportunity to decrease this percentage to the target value set by the SCE expert, which is <0.5% of the lines of the BOM. A Pareto chart could be made with the wrong entries to show which data combinations result in the most wrong data entries. The Pareto chart is shown in Figure 8.

![Pareto chart of wrong entries](image)

Figure 8 Pareto chart of wrong entries

As can been seen in Figure 8, the 80-20 heuristic cannot be applied to the results of the wrong entries analysis. This heuristic suggests that 20% of possible mistakes will lead to 80% of all wrong entries in the data. It appears that the first four mistakes influenced 80% of the wrong entries in the system, and they accounted for only approximately 9% of the number of possible mistakes.

However the 80-20 heuristic can apply if the specific combinations are categorised depending on their relationship (given in Table 3). For example, the relationship LO_VS_IPC has three different possibilities: the LO codes 001, 003, and 008. During categorisation, these three possibilities were combined into one possibility, so that all the errors were summed. This categorisation resulted in 11 wrong entry groups.

The Pareto chart of the categorised wrong entry data is shown in Figure 9.
As can be seen, two major groups were the main causes of the wrong entries in the system: the LO_IPC_FLC_vs_STOCRML combination and the IPC_VS_UnloadPSA_AND_Sloc combination. This categorisation resulted in a better fit of the 80-20 heuristic, because now approximately 18% of the possible mistakes led to 93% of the wrong entries in the data.

To validate this argument a second time, the MHF employees were asked to document all errors that occurred during the day for four weeks. However, because of their busy schedule during the day, they were unable to do this. Observation was also not possible in a significant way because of the many unload stations, which all differ too much from each other to be authorised to take a sample of them. Therefore, the available data had to be trusted and assumed as reality.

5.2.2.1.4 Ishikawa diagram
With the gathered potential causes of the interviews, the cause-and-effect tree, and the knowledge of the findings of the categorised Pareto chart, the preliminary cause-and-effect tree could be specified in more detail. This new diagram will be called the Ishikawa diagram. In the cause-and-effect tree, the causes related more to the people, and the relations between different causes were not yet determined. The Ishikawa diagram zooms in on the scope of the project (the dashed line in the first cause-and-effect diagram). With the Ishikawa diagram, the causes were categorised based on people, environment, method, system, and measurement. This was made possible with the knowledge obtained from the expert interviews and the analysis of the data lists. By looking at the problem from multiple perspectives, the problem was illustrated in greater detail. With the help of the experts, the relationships between the different causes were also determined.

The causes in this Ishikawa diagram were analysed and relationships between the elements were determined with the help of the experts. If the method of changing the data differs between the two Supply Chain departments, the data quality problems increase. If the system is slow and the open space where the SCE works causes a high amount of distraction, this increases the wrong entries rate, which leads to an increase in data quality problems. The time-consuming and complex method of changing the data is related to the slow system. The fact that the measurement is rarely done and is done by hand, and that the fail rate is not documented, is related to the data quality problem, because if the measurements are done too infrequently, the problem will continue to increase and will never be solved. Working in two different systems leads to forgotten changes, which increase the number of wrong entries. Hence, all of the elements in the Ishikawa diagram are related to each other. In this Ishikawa diagram, some causes are drawn in dashed lines connected to causes drawn in solid lines. The dashed-line causes have the same meaning as the causes with the solid lines to which they are linked.
In comparison with the cause-and-effect diagram, where the focus was on the persons, two extra causes were added to the Ishikawa diagram. In the first cause-and-effect diagram, incorrect/meaningless entries or incorrect/meaningless adjustments were the causes. However, adjustments can also be forgotten. Furthermore, from the analysis it could be concluded that not only the stakeholders can adjust the data; the MHF and the MHE employees can also do so. Thus, an extra cause was added to the new cause-and-effect diagram, the Ishikawa diagram: many changes by different employees from different departments.
Figure 10 Ishikawa diagram
5.3 Conclusions from the analysis
The data quality problem was analysed in several ways, each of which provided information that confirmed the problem and specified it in more detail. In the Interviews with the stakeholders, the most important statements were noted. From these statements, it can be concluded that data quality errors comprise 5.5% of the data for the following reasons: employees adjust/enter incorrect/meaningless data and different people from different departments adjust data; the employees do not communicate correctly; they do not have enough process knowledge; they give no feedback; and they have to adjust the data in a long process. This results in products in wrong process flows, bottlenecks in the process, damage to material, and uncertainties in the data concerning what has been changed and what has not. This all costs time and therefore money.

Poor communication, the adjustment of data by different people in different departments, and the lack of process knowledge were validated by observing communication between MHF employees and one MHE employee. All of these causes lead to wrong entries in the data, which were analysed with the combinations gathered by interviewing the experts. The conclusion of this analysis is that the 80-20 heuristic can be applied if all of the combinations that have to be accounted for in the solution design are categorised.

In addition, the cause-and-effect tree was examined in more detail than before, when together with the experts the relationships were determined and the causes were validated. Here it can be concluded that all of the causes can be related to wrong entries, which is a highly important cause of the data quality problem. In the process analysis, several conclusions were also drawn, which strengthened the empirical analysis. The conclusion based on the questionnaires is that the stakeholders think that authorisation, communication, and controllability are the most important aspects of the data quality problems of the process. Furthermore, a conclusion was formulated based on the Unload PSAs, where it was shown that the data quality problems cost on average 117.50 euro per day. Therefore, the solution design has to develop a concept that costs less than this amount of money. Based on the process flow analysis, it is concluded that there are three points in the process where it is possible that employees detect the data quality problems because of process flow blocks. The solution design should decrease the amount of times that the data are changed at these different locations in order to improve the data quality. Thus, the solution design has to improve and use the issues presented in Table 9 to increase data quality.

Table 9 Data quality issues

<table>
<thead>
<tr>
<th>Data quality issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Different people from different departments adjust the data.</td>
</tr>
<tr>
<td>2  The data adjustment process is a long process that is not properly explained, and the systems are often slow.</td>
</tr>
<tr>
<td>3  There is too little process knowledge among stakeholders.</td>
</tr>
<tr>
<td>4  Communication between stakeholders is not correct.</td>
</tr>
<tr>
<td>5  If changes are communicated, there is lack of feedback regarding whether the changes are successful, and this is taken for granted.</td>
</tr>
<tr>
<td>6  The 80-20 heuristic can be used.</td>
</tr>
<tr>
<td>7  The main focus is on the wrong entries.</td>
</tr>
<tr>
<td>8  Authorisation, communication, and controllability must improve</td>
</tr>
<tr>
<td>9  Costs must be decreased.</td>
</tr>
<tr>
<td>10 Time must be decreased.</td>
</tr>
</tbody>
</table>
6 Solution design
Having discussed the analysis phase, this chapter will present the solution design. First, the requirements are set. Second, a data quality classification model is made. Third, the conclusions of the analysis are summarised and solutions to these conclusions are presented. Fourth, a solution design is presented. Fifth, designing the tool is explained and finally the user manual and implementation plan are briefly discussed.

6.1 Requirements
The relationships between the columns in the master data could be set as the rules for the data improvement tool. With the gathered knowledge, the actual designing could begin by identifying the design requirements: the demands that the tool has to meet. These requirements were identified in the interviews with the SCE expert, SCE stakeholders, and an information management expert.

With the SCE expert, mainly the Functional and the Boundary requirements were set. The User requirements were set with the SCE stakeholders. The Design restrictions were given by the Information management expert.

The requirements are stated below.

Table 10 Requirements

<table>
<thead>
<tr>
<th>Functional requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Must have</strong></td>
</tr>
<tr>
<td>1 The data improvement tool (DIT) helps to improve data quality.</td>
</tr>
<tr>
<td>2 The DIT checks the daily state of the art data in SAP for errors.</td>
</tr>
<tr>
<td>3 The DIT provides a list with all data quality errors that must be corrected by SCE stakeholders.</td>
</tr>
<tr>
<td>4 The DIT must provide a selection of the types of errors.</td>
</tr>
<tr>
<td>5 The DIT must give a clear representation of results for different user groups.</td>
</tr>
<tr>
<td>6 The DIT must provide data that make it easy to repair the errors in SAP.</td>
</tr>
<tr>
<td><strong>Should have</strong></td>
</tr>
<tr>
<td>7 A manual has to be developed to ensure future adaptation and understanding.</td>
</tr>
<tr>
<td>8 A manual for users has to be developed to solve problems during usage.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-functional requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Must have</strong></td>
</tr>
<tr>
<td>9 The response time of the tool may not be longer than five minutes.</td>
</tr>
<tr>
<td>10 The list of errors that is shown must be understandable for the SCE stakeholders.</td>
</tr>
<tr>
<td>11 The DIT must be easy to use.</td>
</tr>
<tr>
<td><strong>Won’t have</strong></td>
</tr>
<tr>
<td>12 The DIT cannot be developed in SAP because of the costs.</td>
</tr>
<tr>
<td>13 No connection is allowed between SAP and VBA to put data in SAP.</td>
</tr>
</tbody>
</table>

6.2 Data quality classification model
From the data quality classification model by Anders Haug et al. (2009), which was discussed in Section 2.1 Data quality, a data quality classification for improvement could be developed. The purpose was to ensure that all of the aspects of this design were treated using Haug et al.’s model. The authors state that if the intrinsic data quality dimensions and the data accessibility dimensions increase, the usefulness will increase automatically. They further state the following: ‘In addition, the first two categories are more critical than the usefulness category, since poor data accessibility and
intrinsic data quality are factors that make some daily operations impossible, while ERP system data of little usefulness can largely be ignored’ (Anders Haug et al., 2009)

Haug et al.’s (2009) model consists of three data quality categories that in turn comprise their own subcategories:

1. Intrinsic data quality dimensions: completeness, unambiguousness, meaningfulness, and correctness (based on Wand and Wang, 1996).
2. Data accessibility dimensions: access rights, storage in ERP system, understandability, and interpretability.
3. Data usefulness dimensions: relevance, value-adding, level of detail, and timeliness.

6.3 Analysis conclusions and design
Haug et al.’s (2009) model could be compared with the causes of the problem identified in the analysis; they are presented in Conclusions from the analysis

The data quality problem was analysed in several ways, each of which provided information that confirmed the problem and specified it in more detail. In the Interviews with the stakeholders, the most important statements were noted. From these statements, it can be concluded that data quality errors comprise 5.5% of the data for the following reasons: employees adjust/enter incorrect/meaningless data and different people from different departments adjust data; the employees do not communicate correctly; they do not have enough process knowledge; they give no feedback; and they have to adjust the data in a long process. This results in products in wrong process flows, bottlenecks in the process, damage to material, and uncertainties in the data concerning what has been changed and what has not. This all costs time and therefore money.

Poor communication, the adjustment of data by different people in different departments, and the lack of process knowledge were validated by observing communication between MHF employees and one MHE employee. All of these causes lead to wrong entries in the data, which were analysed with the combinations gathered by interviewing the experts. The conclusion of this analysis is that the 80-20 heuristic can be applied if all of the combinations that have to be accounted for in the solution design are categorised.

In addition, the cause-and-effect tree was examined in more detail than before, when together with the experts the relationships were determined and the causes were validated. Here it can be concluded that all of the causes can be related to wrong entries, which is a highly important cause of the data quality problem. In the process analysis, several conclusions were also drawn, which strengthened the empirical analysis. The conclusion based on the questionnaires is that the stakeholders think that authorisation, communication, and controllability are the most important aspects of the data quality problems of the process. Furthermore, a conclusion was formulated based on the Unload PSAs, where it was shown that the data quality problems cost on average 117.50 euro per day. Therefore, the solution design has to develop a concept that costs less than this amount of money. Based on the process flow analysis, it is concluded that there are three points in the process where it is possible that employees detect the data quality problems because of process flow blocks. The solution design should decrease the amount of times that the data are changed at these different locations in order to improve the data quality. Thus, the solution design has to improve and use the issues presented in Table 9 to increase data quality. These problems were divided into changes that had to be made regarding intrinsic data quality dimensions, data accessibility dimensions, and data usefulness dimensions. This is demonstrated in the table below.

Table 11 Framework of the data classification model by Haug et al. (2009)

<table>
<thead>
<tr>
<th>Intrinsic data quality</th>
<th>Data accessibility</th>
<th>Data usefulness dimensions</th>
</tr>
</thead>
</table>

37
<table>
<thead>
<tr>
<th>dimensions</th>
<th>dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of the 80-20 heuristic</td>
<td>Different people from different departments adjust the data</td>
</tr>
<tr>
<td></td>
<td>The data adjustment process is a long process that is not explained properly, and the systems are often slow.</td>
</tr>
<tr>
<td>Main focus on the wrong entries</td>
<td>Improve authorisation</td>
</tr>
<tr>
<td></td>
<td>There is too little process knowledge among stakeholders.</td>
</tr>
<tr>
<td>Controllability</td>
<td>If changes are communicated, there is a lack of feedback regarding whether the changes are successful, and this is taken for granted.</td>
</tr>
<tr>
<td></td>
<td>Costs are decreased.</td>
</tr>
<tr>
<td></td>
<td>Improve communication</td>
</tr>
<tr>
<td></td>
<td>Time is decreased.</td>
</tr>
<tr>
<td>Communication between stakeholders is not correct.</td>
<td></td>
</tr>
</tbody>
</table>

A categorisation can be made of all of the aforementioned problems. For example, within the data accessibility dimensions, two problem causes are mentioned: ‘Different people from different departments adjust the data’ and ‘improve authorisation’. Both of these problem causes mean the same and are merged. This also can be done with communication and increased process knowledge, as demonstrated in Table 12.

Table 12 Classified framework of the data classification model by Haug et al. (2009)

<table>
<thead>
<tr>
<th>Intrinsic data quality dimensions</th>
<th>Data accessibility dimensions</th>
<th>Data usefulness dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controllability</td>
<td>Authorisation</td>
<td>Increasing process knowledge</td>
</tr>
<tr>
<td>Main focus on the wrong entries</td>
<td>Communication</td>
<td>Decreasing costs</td>
</tr>
<tr>
<td>Use of the 80-20 heuristic</td>
<td></td>
<td>Decreasing time</td>
</tr>
</tbody>
</table>

With knowledge of the design, it was ensured that every important cause would be treated. Furthermore, it was also ensured that these causes were also considered important in the literature (Haug et. al, 2009). The next sub-sections will explain every problem cause in comparison with the design of a solution design for each cause. For each change proposal a solution concept was developed to improve the data quality on that particular aspect. All of the change proposals are elaborated below.

6.3.1 Decreasing time
As mentioned in the analysis, data are checked three different times in the product flow. If the data are wrong, they have to be changed. The solution is for the department that also engineers this data to check/change it – the Supply Chain Engineering department. With this solution, the three steps of checking and changing the data are merged into one. This increases usefulness in the sense that steps do not have to be repeated as many times as in the current situation. Furthermore, this adds value and improves timeliness.
This is accomplished by adding one step at the beginning of the process, as can be seen in Figure 12. By adding this step, the other departments have to contact the SCE department less often. After the insertion of the data in SAP, the data are checked.

This leads to a decrease in the frequency with which the data are wrong at a checkpoint in the process. The number of departments that adapt the information can also be decreased. However, this does not happen as a result of adding this extra step; it will be explained in the authorization solution concept.

### 6.3.2 Decreasing costs

The added step in the beginning is executed by a SCE. The time spent by an SCE per day has to be short enough to cost less money than the costs that the MHF makes every day on average. With the fixed costs price of a SCE the previously defined formula could be used. Using these calculations, the cost of an SCE is 62.50 euro/hour. In addition, it was necessary to calculate how long an SCE may work compared to a WHF employee, as calculated in Section Unload PSAs. The maximum amount of time that an SCE can spend checking and solving the data after insertion must be as follows:

\[
\frac{117.50 \text{ euro/day}}{62.50 \text{ euro/hour}} = 1.88 \text{ hour/day}
\]

This is 1 hour and 52 minutes per day. The solution to reach this target is to plan two fixed time spans of 45 minutes per day to check and solve the data errors. The data quality will not be solved in one day in this way, but in the future the time that an SCE will spend checking and solving the data quality will be much less than 1 hour and 52 minutes per day. Two sets of 45 minutes have been
chosen to ensure that the SCE’s concentration level stays high enough. This adds relevance to the solution design.

6.3.3 Increasing process knowledge
The process knowledge is a data quality problem cause that was identified during the interviews in the empirical analysis. The process knowledge is less important according to the responses to the questionnaires administered in the process analysis. However, the process knowledge of the SCE stakeholders should be improved as much as possible, because this increases the level of detail of the data. The process flow will be provided to the stakeholders to increase their knowledge of the blockages in the process that can be caused by their data input.

6.3.4 Authorisation
Based on the stakeholder interviews, authorisation was found to be the most important problem. In addition, the results of the questionnaire about the data quality process problem causes also indicate that the authorisation is the most important cause of the data quality problems. The stakeholders stated that too many employees are authorised to change data in the system, which causes data quality problems. This will be solved by changing procedures in the process so that not all departments can and may change the data. This is part of the implementation of the solution design.

6.3.5 Communication
The communication is also part of the implementation of the solution design. Whereas authorisation has a direct influence on the accessibility of the ERP system, communication has an influence on the employees of the different departments. Communication protocols are advised to ensure that the accessibility of all the employees to all information can be provided. The interpretability of the data will increase if all information is communicated correctly.

6.3.6 Controllability
Controllability is an important cause of the data quality problem; it was rated as the second most important cause in the questionnaires about the process. Controllability currently almost never happens because of the time it costs, as this check is done manually. The solution to this data quality problem is to develop a tool to check data errors in recent SAP data. With this tool the data will be checked for errors, and an overview with changes that have to be made will be given to the SCE. Furthermore, this tool will save time in comparison with the manual check, making it possible to work more efficiently. In Haug et al.’s (2009) data quality classification model, controllability falls under the intrinsic data quality dimensions. The intrinsic quality dimensions based on Wand and Wang (1996) can be divided into four dimensions: Completeness, Unambiguousness, Meaningfulness, and Correctness. These dimensions will be discussed in relation to the tool.

6.3.6.1 Completeness and unambiguousness
By keeping the process flow up-to-date, the completeness and unambiguousness of the data can partly be monitored. If the process changes, the information system data also have to change. Therefore, the tool also has to be adapted to these changes to be the most effective.

6.3.6.2 Meaningfulness and Correctness
Meaningfulness and Correctness can also been seen as garbling and will be checked by the tool. Garbling is the biggest of the data quality problems within SAP, and thus reducing it will decrease the percentage of mistakes the most. If real world states are documented in the incorrect information states or in meaningfulness states, the tool will detect these and produce an overview of these garbling errors.
6.3.7 Main focus on the wrong entries
The main focus has to be on the wrong entries. As described above with regard to meaningfulness and correctness, the tool checks the garbling errors, which are wrong entries. Thus the main focus lies on these wrong entries, as concluded in the analysis.

6.3.8 Use of the 80-20 heuristic
The 80-20 heuristic identified in the analysis can be of high value in developing the solution. To obtain a value of 0.5% wrong entries, the 80-20 heuristic is not enough. However, to do a quick check where time is saved and the most important causes are treated, the 80-20 heuristic can be of great help in finding the greatest amount of wrong entries in the shortest possible time.

6.3.9 Conclusion
It can be concluded that the main focus of the data improvement tool is on the garbling errors, together with all the accessibility dimensions and the relevancy and value-adding within the usefulness dimensions for the garbling layers. The interpretability and the understandability of the completeness and ambiguousness are considered, which will also be relevant and value-adding. These results are shown in Table 13, which presents all solutions versus the data quality classification model. Table 13 shows the concerned data quality subcategories that are influenced for each solution.

Table 13 Solutions versus data quality classification model

<table>
<thead>
<tr>
<th>Intrinsic data quality dimensions</th>
<th>Data accessibility dimensions</th>
<th>Data usefulness dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td>Unambiguousness</td>
<td>Meaningfulness</td>
</tr>
<tr>
<td>Controllability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main focus on the wrong entries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of the 80-20 heuristic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authorisation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increasing process knowledge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreasing costs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreasing time</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.4 Data improvement tool solution design
From the requirements and the design, the solution design could be developed with the COMET design approach (Berre et al., 2006). First a context statement is designed to set all stakeholders of the data improvement tool for development. Second, a goal hierarchy model is developed. Third, a business resource model is made. Fourth, a process model is developed. Finally, a WARM model is provided. With these models, the design of the data improvement tool is specified.
6.4.1 Context statement
In the context statement, the stakeholders who make use of the data improvement tool are identified. The model is provided in Figure 13.

![Context statement model](image)

**Figure 13 Context statement model**
Six SCE stakeholders will work with the data improvement tool (DIT). These stakeholders will use the DIT and cannot adapt or change the DIT. One SCE expert will perform the maintenance on the DIT. He will ensure that the tool remains up-to-date, and he will also have the password to the algorithm in the two systems in the DIT. Two systems provide information to the DIT for it to work properly. From the ERP system SAP, the master data are gathered and two lists in the repository of VDL NedCar are used as a reference. The gathered master data and the two lists provide the data to perform the check.

Within the DIT, two systems will retrieve and check the data. These systems run an algorithm that ensures that the data are gathered and checked with the help of the master data lists and the reference lists.

6.4.2 Goal hierarchy model
With a goal hierarchy model, the most important goals are stated. These were gathered from Chapter 4. The goals are:

- To reduce the number of errors by decreasing the percentage of wrong entries from 5.5% to 0.5% or less of the data.
- To increase satisfaction by reducing the stakeholders’ biggest problems, the authorisation and the controllability, from 2 and 2.2 respectively, which were from the scores of the questionnaire, to above 2.5.
- To reduce the number of times that data quality problems are changed from 3.27 hours/day to less than 0.5 hour/day.
- To increase the flexibility of checking the data quality, from only manually to also using a tool.
- To reduce costs from 117.5 euro/day to less than 57.5 euro/day.

**Figure 14 Goal hierarchy model**
The model is shown in Figure 14. The number of times that changes have to be made reduction can have positive and negative influences on efficiency. In this case, the number of times that changes have to be made reduction will decrease, which will have a positive impact on efficiency.

6.4.3 Business resource model
The main concepts of the domain that are relevant to the IS are identified in the business resource model. The resource model contains classes for assets and actors of the library, as well as classes for resources that the actors must use when performing the business processes. Figure 15 depicts the resource model of the data improvement tool.

Figure 15 Business resource model
6.4.4 Process model
Figure 16 presents the high level process model of the steps taken when using the DIT.

![Process model diagram]

Figure 16 Process model

6.4.5 Work Analysis Refinement Model (WARM) diagrams
The WARM activity diagrams illustrate the steps taken by the SCE employee and the system. The three business processes mentioned in the business process and role model are elaborated in Figure 17.

![WARM activity diagram]

Figure 17 WARM activity diagram
6.5 Designing the DIT
As mentioned in Section 3.2.1.2, the designing of the DIT consists of synthesis-evaluation iterations. The main goal of the tool is to check the daily data for errors and to provide a list of these errors presented in a clear and understandable manner. The synthesis-evaluation iterations can be explained by discussing three main concept versions of the tool. The first version served to create a basis: to obtain the data, check the data, and present the results the check in one list. The second concept version served to expand all of the rules of one list. The third concept version aimed to add the other list and combine these into one tool. All of the automation codes are written in Visual Basic for Applications (VBA).

6.5.1 VBA
VBA is a programming language that allows the development of user-defined functions, as well as the automation of certain processes and calculations. It is a quasi-object oriented programming language, where the system is composed of objects. VBA is built into most Microsoft Office applications, such as Word, PowerPoint, Access, and Excel. Other applications have also implemented VBA, such as AutoCAD, SolidWorks, and CATIA. In this study, the DIT was programmed in VBA using the Excel application. The codes created in the versions were too long to attach in the appendix; therefore, they have been given in a separate document to VDL Nedcar.

6.5.2 Version 1
As mentioned in the beginning of this section, version 1 consisted of three phases: to retrieve the data, to check the data, and to present the check results.

6.5.2.1 Retrieve data
To retrieve the daily data, several steps were needed in SAP and Excel, which cost a great deal of time. Furthermore, to ensure a successful automation, a standard set up of the lists had to be made, so that the steps will be always the same. Considering these requirements for obtaining the data for this design, the decision was made to automate the data retrieval action. This assures that the data will always be in the same setting to be checked. To check the data automatically, they need to be gathered in SAP and in VDL Nedcar’s repository. The 431 list that was gathered (as discussed in Section 4.3) was used to design the first concept tool.

To retrieve the data from SAP, first a connection had to be made with the ERP system. The first evaluation of the code to retrieve the data gave an error, so the process was iterated. Here the code was adapted and the concept tool was evaluated again. After gaining access to SAP, a connection with the SapGuiApp was made to retrieve the lists. This was accomplished by recording the steps manually taken in SAP with Script recording and playback, and using these with the VBA codes. Eventually, codes were created that worked properly during the evaluation.

6.5.2.2 Checking data and presenting results
The first concept version was created with a consideration for the logistic processes with the easiest relationship: the LO_vs_IPC relationship. As was discussed in Section 4.4.3.2, the advanced knowledge gathered from the experts was transformed to VBA if-then rules, which checked the data lists. The difference between the if-then rules used in Table 3 to explain the knowledge gathered with the if-then rules used in VBA is that the then in Table 3 has to be implemented in the if in VBA. This has to be done because the then rule in VBA has to transfer the explored error to the results list. An example of this abduction of the LO_vs_IPC relation code is given below.

Code 1

As described earlier, the then rule in the synthesis is used to present the data check results. This is done with the help of the requirements. With the evaluation of the presentation of the results, feedback was gathered from the stakeholders regarding requirement 6 in Table 10. After iteration of the process with the added the requirement that was requested, which was to add the GsPSA column, which increases the ease of data insertion, the tool worked properly during the evaluation.

6.5.3 Version 2
After version 1 was finished, the rules could be expanded to all of the rules in the 431 list, which were essentially the same as the first designed rule. After each rule was added, an evaluation took place and small iterations were executed. Regarding the rule of IPC_VS_UNLOADPSA_SLOC, which is also listed in Table 3, an abduction had to be done. This rule differs from the other rules because there are more columns that can have variable values. First, these kind of rules were written as the rule presented in code 1. This resulted in a presentation of the error only if both values of the columns were true. After evaluation, the synthesis was rewritten, as can be seen in code 2, which resulted in a working code. When all rules were added, version 2 was also ready.

If .Cells(i, "K") = "E7" And Not .Cells(i, "AE") = "M-75" And Not .Cells(i, "AE") = "M10" And Not .Cells(i, "AE") = "M-70" And Not .Cells(i, "AE") = "" Or .Cells(i, "K") = "E7" And Not .Cells(i, "AD") = "60LC" Or .Cells(i, "AD") = "" Then

Code 2

6.5.4 Version 3
In version 3, the other lists, which needed a different synthesis for checking, were also added. When working with these lists, checking with other lists was necessary and calculations had to be done from the reference list to obtain the right values. In this version the synthesis was also designed, evaluated, and iterated. This was continued until the third and final version was ready. An example of a synthesis in this version is one from the 921 list compared with the reference Handlings unit list. In the 921 list, the volumes of the Total volume of the Packing Instruction are given. In the reference list, these volumes are not presented, but the length, width, and the height are given. Hence, the tool first needs to find the needed reference material with the help of the Handling units list. Second, it then has to calculate the volume with the reference length, width, and height. Finally, it needs to check this volume with the volume of the 921 list.

Finally, all of the synthesis had to be combined to create one tool. This was accomplished by adding an option in the menu that runs a new synthesis. In this synthesis, the earlier written syntheses were combined. After the evaluation, the presentation needed to be adjusted to present all data in a clear and understandable manner. After this step, the tool was completely designed.

6.6 User manual
The tool was developed in VBA to ensure its accessibility for the stakeholders who work with Excel every day. The DIT consists of the following sections:

- Activation
- Data retrieval process
- Data checking process
  - Checking logistic process
o Checking packing instructions
o Checking logistics process and packing instructions

These sections are explained in the user manual, which is in Appendix VII.

6.7 Implementation plan

In this study, a data improvement tool was developed for improving the data quality of SAP in an automotive environment. The main aim of this tool was to decrease the percentage of garbling errors in the SAP data to below 0.5% and increase the data quality. In order to successfully reach this goal by fully taking advantage of the DIT, the tool should be implemented with care. This will be accomplished by taking care of the following aspects:

- Assigning a data improvement tool manager;
- Providing SCE stakeholder training;
- Developing an authorisation protocol; and
- Developing a communication protocol.

These aspects are elaborated in Appendix VIII.

This chapter has provided a solution design and explained how the tool is developed. The next chapter will check if the tool will improve the data quality.
7 Verification and Validation of the tool

After all of the previous stages, the DIT had to be validated. This validation was done by following the steps discussed in Section 3.2.1.3.

This chapter will check and validate the tool by first using the devil’s quadrangle which evaluates the tool. After the evaluation the tool will be validated by the pivot table method. Subsequently, the tool will be validated by experts and stakeholders. Finally, conclusions will be made.

7.1 Devil’s quadrangle

As mentioned in Section 3.2.1.3, Brand and van der Kolk (1995) provide four dimensions with which to evaluate the DIT. These dimensions are time, costs, quality, and flexibility.

7.1.1 Time

The DIT has a positive influence on the time it takes to correct the data. As calculated in Section 3.2.2, the time that the MHF department spends correcting data quality issues is 3.27 hours/day in the as-is situation. The target is set to an error rate of less than 0.5%. Hence, the time it will cost the MHF department will decrease by 90%. This means that the data improvement tool will decrease the time it costs the MHF department sample from 3.27 hours/day to 0.327 hours/day, which is approximately 20 minutes. Therefore, the time saved is 2.943 hours.

In the SCE department, the time will increase because an extra step will be added. Section 4.3.2 indicates the maximum amount of time that the department should spend on data quality. This time is 1.88 hours/day, which in the future will decrease to 0.188 hour/day.

The total decrease in time will be -1.88 hours/day at the beginning. Then, the future decrease in time will be:

\[
3.27 \text{ hour/day} - 0.188 \text{ hour/day} - 0.327 \text{ hour/day} = 2.755 \text{ hour/day}
\]

7.1.2 Costs

The data improvement tool has a positive influence on costs. These costs can be categorised as hidden production costs in the model by Haug et al. (2011). If the time that an employee spends correcting data errors decreases, the costs of this employee will also decrease. In Section 3.2.2., the costs were calculated for the as-is situation. Assuming that in the future the data quality target error rate of 0.5% will be achieved, and knowing that the data quality error rate in the as-is situation is 5.5%, the costs will optimally decrease by 90%. This is based on the assumptions that only the MHF employees’ proceedings will be considered, that there will be no more data quality errors than 0.5% of all the data, and that no costs will be incurred anywhere else.

However, the SCE department will gain an extra task, and will therefore cost more than in the as-is situation. In the beginning, as mentioned in Section 4.3.2, the time that an SCE will spend on the data improvement tool will be 90 minutes. An SCE costs 62.50 euros/hour, so the following calculation can be made:

\[
62.50 \text{ euro/hour} \times 1.5 \text{ hour} = 93.75 \text{ euro}
\]

This means that the profit will be:

\[
117.5 \text{ euro/day} - 93.75 \text{ euro/day} = 23.75 \text{ euro/day}
\]

In the future this will increase due the increasing data quality in the ERP system. Eventually the data quality check will be assumed to occur once a week, and will take 30 minutes. This means that it will
take approximately six minutes a day and will cost 6.25 a day. Hence, in the long run, the tool provides a profit of:

\[ 117.5 \text{ euro/day} - 6.25 \text{ euro/day} = 111.25 \text{ euro/day} \]

### 7.1.3 Quality
As mentioned in the previous sub-sections 6.2.1 and 2.2.2, the quality of the data will increase. The tool will reduce the error rate from 5.5\% to a maximum of 0.5\%. This is an increase of data quality of:

\[ 100\% - \left( \frac{0.5\%}{5.5\%} \times 100\% \right) = 90\% \]

### 7.1.4 Flexibility
In the as-is situation, the data quality is checked manually. With the DIT, the Supply Chain Engineers have the choice to check the data automatically or manually. In an automatic check, the tool is flexible regarding the direction in which the data quality should be checked. For example, specific errors can be searched or a quick scan can be executed. Hence, the flexibility is increased.

### 7.2 Pivot table method
The tool is validated with the pivot table method. This method was also used earlier to validate the knowledge gathered from the experts with the data in SAP. Hence, the logistic data were gathered from SAP, and with the pivot table the relation between the Lab Office (LO) and the Internal Process Code (IPC) was tested. As an example, one relation from Table 3 is used; this is shown in Table 14

<table>
<thead>
<tr>
<th>Relation</th>
<th>LO_IPC</th>
<th>IF</th>
<th>Then</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO_IPC</td>
<td>1</td>
<td>3A, D3, all E’s</td>
<td></td>
</tr>
</tbody>
</table>

As can been seen from the pivot table, almost all of the values are in the acceptable range. However, the value C1 is not allowed and thus four errors are detected. If the relation between LO and IPC is tested with the DIT, the results are given in Table 16 and Table 17.

### Table 14 Extraction of Table 3

<table>
<thead>
<tr>
<th>Relation</th>
<th>LO_IPC</th>
<th>IF</th>
<th>Then</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO_IPC</td>
<td>1</td>
<td>3A, D3, all E’s</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LO</th>
<th>001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row Labels</td>
<td>Count of Material</td>
</tr>
<tr>
<td>3A</td>
<td>35</td>
</tr>
<tr>
<td>C1</td>
<td>4</td>
</tr>
<tr>
<td>D3</td>
<td>13</td>
</tr>
<tr>
<td>E1</td>
<td>209</td>
</tr>
<tr>
<td>E2</td>
<td>158</td>
</tr>
<tr>
<td>E3</td>
<td>165</td>
</tr>
<tr>
<td>E4</td>
<td>9</td>
</tr>
<tr>
<td>E5</td>
<td>612</td>
</tr>
<tr>
<td>E7</td>
<td>4,747</td>
</tr>
<tr>
<td>E8</td>
<td>3,464</td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td><strong>9,418</strong></td>
</tr>
</tbody>
</table>

### Table 15 Pivot table of values from Table 14

<table>
<thead>
<tr>
<th>Material</th>
<th>MatDescMM</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>FLR</th>
<th>IPC</th>
<th>LO</th>
<th>GsPSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>475227301</td>
<td>ASSY BRACKET DDE8 UKL</td>
<td>X</td>
<td>WH</td>
<td>C1</td>
<td>001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>744148001</td>
<td>ASSY CONSOLE STRIKER</td>
<td>X</td>
<td>WH</td>
<td>C1</td>
<td>001</td>
<td></td>
<td>13989610</td>
<td></td>
</tr>
<tr>
<td>746981501</td>
<td>LH REINFORCEMENT DOOR INR PNL DOOR RR</td>
<td>X</td>
<td>WH</td>
<td>C1</td>
<td>001</td>
<td></td>
<td>13930510</td>
<td></td>
</tr>
<tr>
<td>746981601</td>
<td>RH REINFORCEMENT DOOR INR PNL DOOR RR</td>
<td>X</td>
<td>WH</td>
<td>C1</td>
<td>001</td>
<td></td>
<td>13930510</td>
<td></td>
</tr>
</tbody>
</table>

### Table 16 Error percentage summary

<table>
<thead>
<tr>
<th></th>
<th># errors</th>
<th>total data</th>
<th>error percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>4</td>
<td>9,418</td>
<td>0.0%</td>
</tr>
<tr>
<td>003</td>
<td>0</td>
<td>1,008</td>
<td>0.0%</td>
</tr>
<tr>
<td>008</td>
<td>0</td>
<td>99</td>
<td>0.0%</td>
</tr>
<tr>
<td>total</td>
<td>4</td>
<td>1,0525</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

### Table 17 Example of the results presentation list

<table>
<thead>
<tr>
<th>Material</th>
<th>MatDescMM</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>FLR</th>
<th>IPC</th>
<th>LO</th>
<th>GsPSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>475227301</td>
<td>ASSY BRACKET DDE8 UKL</td>
<td>X</td>
<td>WH</td>
<td>C1</td>
<td>001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>744148001</td>
<td>ASSY CONSOLE STRIKER</td>
<td>X</td>
<td>WH</td>
<td>C1</td>
<td>001</td>
<td></td>
<td>13989610</td>
<td></td>
</tr>
<tr>
<td>746981501</td>
<td>LH REINFORCEMENT DOOR INR PNL DOOR RR</td>
<td>X</td>
<td>WH</td>
<td>C1</td>
<td>001</td>
<td></td>
<td>13930510</td>
<td></td>
</tr>
<tr>
<td>746981601</td>
<td>RH REINFORCEMENT DOOR INR PNL DOOR RR</td>
<td>X</td>
<td>WH</td>
<td>C1</td>
<td>001</td>
<td></td>
<td>13930510</td>
<td></td>
</tr>
</tbody>
</table>
The tool provides a summary table in which the number of errors is given, as well as the error percentage of that amount of errors. This is presented in Table 17. In Table 17, the tool has found four errors. In the results list, which is shortened, the green indicates which value is the reference (LO column), while the red indicates which value is an error (IPC column). In the IPC column, the value C1 is given, which is the wrong value. With this information, the SCE knows that this value is wrong and has to be changed in SAP. With the information given in the total results list, these wrong values can more easily be corrected than without the tool. The pivot table and the DIT yield the same results, which validates this part of the tool. This validation method is used for all the relationships. Thus, the entire tool is validated.

7.3 Expert and stakeholder validation
The tool was also validated by the experts and the stakeholders who will make use of the tool.

7.3.1 Experts
The experts are of the opinion that the tool is a solution to the problem, because every time it will be used and the provided data errors will partially be solved, the data quality will increase. At first, they believe that the data quality improvement will take a considerable amount of time and that this will be achieved in steps. Later, however, it will only be a matter of keeping the data quality up-to-date, and the tool will be used weekly/monthly for maintenance. The experts’ opinion is that the tool represents the data in such a way that it will be easier than it currently is to change the data. Furthermore, Supply Chain Engineers’ awareness of data errors will increase.

The two experts were asked to use the tool for a period of time and to give feedback with regard to the requirements and the goals set. The task that each expert had was to use the tool to detect errors in SAP. The feedback given by the experts is presented below.

Expert 1: ‘Today, we once again sat together and discussed, viewed, and tested the tool that you made. I have carried out various checks with the tool and the output of these tests was satisfactory. I expect a big improvement in data quality after the introduction of this tool.’

Expert 2: ‘I have used the data improvement tool for several errors and found the errors with ease. The data improvement tool has found the same results as when I do the check manually. The data improvement tool will save time in checking the data for SCE.’

Based on these citations, it can be concluded that the tool works properly, and that the requirements and goals that were set have been accomplished. Furthermore, both experts predicted that the tool would accomplish the stated target of <0.5% data quality error percentage.

7.3.2 Stakeholders
The tool was presented to a group of stakeholders to show them how the tool works and to obtain their feedback. These stakeholders were convinced that the tool is clear and easy to work with. Furthermore, they thought that the tool could be of significant importance in improving the data quality within SAP.

7.4 Conclusion
From the above verification and validation sections, it can be concluded that the tool works properly and that significant results will be accomplished with it. All of the errors that the tool provided were assigned as errors by the experts, the stakeholders, and the pivot table method. With the pivot table method, it was also confirmed that the tool did not miss any errors. In terms of Brand and van der Kolk’s (1995) Devil’s quadrangle, the tool is expected to provide improvements regarding the time, the flexibility, the quality, and the costs in the long run. The next chapter will conclude all findings and will answer the research questions.
8 Conclusion, discussion, and recommendations

In previous chapters, the research questions were answered. These questions arose of the following problem statement: ‘The current percentage of data quality errors in SAP with respect to the BOM is above 5%, which results in increased costs that are categorised into delay, extra logistical operations, and dissatisfaction of the employees, and does not meet Nedcar’s expectations of data quality with a desired error percentage below 0.5%.’ The main research question, which is derived from the problem statement, is the following:

‘How can data quality in an ERP system (SAP R/3) be significantly improved in a production environment by implementing an automation tool to accomplish in the long run a percentage of inconsistencies below 0.5% for the BOM?’

This chapter presents the conclusion of the study and the recommendations for VDL Nedcar. Furthermore, it highlights the theoretical contribution of this study, and discusses future research directions.

8.1 Conclusions

The solution design is a data improvement tool whose main focus lies on detecting wrong entries (garbling) and that uses the 80-20 heuristic. The tool can be combined with an authorisation protocol and communication protocol to come to the most important measurable data quality dimensions that cannot be ignored in daily operations. Next to these, it is important to increase the process knowledge of the SCE stakeholders; the tool has to cost less than 117.50 euro/day to labour otherwise it is cheaper in the as-is situation, and the tool has to decrease the number of times that data are adapted at different locations. The benefits of the tool are elaborated in the following sections.

8.1.1 Operational level

The tool will help to increase the data quality of SAP, as was validated in Chapter 7. In the validation, the tool was evaluated using N. Brand and H. van der Kolk (1995)’s Devil’s quadrangle. The result of this evaluation is that the tool will decrease the time by 2.75 hours/day in the long run, and the cost by 111.25 euro/day. Furthermore, in the long run the tool will increase the quality of the data by 90%, as well as the flexibility and controllability of the data. These aspects are on the operational level, which is the most important level in the automotive industry because vehicles have to be produced. However, since VDL Nedcar is an independent vehicle contract manufacturer, other levels are becoming increasingly important besides the operational level. VDL Nedcar will have to ensure that the tool performs with high importance not only on the operational level, but also on the tactical and strategic levels, as mentioned in the literature by Thomas C. Redman (1998). With the help of the tool, several benefits, next to the mentioned benefits on the operational level, can be gained by performing on the desired level.

8.1.2 Tactical level

When new projects are analysed, it takes longer to make decisions if the data quality is poor. Furthermore, the decisions that are made will be poor if the data quality is poor. It is clear that decisions made with the use of data of high quality have a better chance of advancing the enterprise’s goals. The tool will also decrease the mistrust that internal organisations may have of one another. If in the future a reengineering project or a new data warehouse concept is put forward, the tool would help to smoothly incorporate such changes within VDL Nedcar.

8.1.3 Strategic level

As a direct consequence of the operational and tactical levels, the tool will also have an influence on the strategic level. The tool will decrease the difficulty for managers to set and execute a strategy. Furthermore, the tool will help managers to focus their attention on the customer and the
competition, because their attention will no longer be diverted by the data quality issues raised in each department.

8.2 Problem statement conclusion
The data quality in SAP R/3 can be increased significantly in the long run, as was validated in Chapter 7. This can be done by means of Haug et al.’s (2009) model, which is integrated in the tool, and the implementation of the tool, discussed in Chapter 6. The tool extracts and checks the SAP data with the use of process knowledge, and provides an overview of data errors that have to be adapted. In combination with the authorisation protocol, the communication protocol, and the workflow of the process, this will cover such a large part of Haug et al.’s model, which can be seen in Chapter 2, that the data quality will significantly increase, and the target data error rate will become even lower than 0.5%, as was stated by the experts in Section 7.33.

8.3 Theoretical contribution
This research contributes to the literature by developing a data improvement tool that implements the concept of Haug et al.’s (2009) model. To the best of the author’s knowledge, the designing, realisation, and implementation of such a tool in an automotive environment has not yet been investigated. Furthermore, Haug et al.’s (2009) model is not currently used to improve data quality; therefore, this study has expanded the model. This research adds value to the theory by connecting the theoretical models and concepts to the practice. The research shows the steps that are taken to design, realise, and implement the theory into a data improvement tool, in this case.

8.4 Limitations
Even though this research adds valuable practical and academic insight into the data quality problem in SAP within VDL Nedcar, it is also restricted by boundaries and limitations. These are described below.

- The tool has been checked based on true values and has been proven to work. However, to gain significant feedback, the tool has to be implemented with the implementation requirements and then be tested on the data quality percentage to yield a significant outcome.
- The analysed results are only relevant for the SCE department. If the tool is expanded to other departments, a new analysis will be necessary to ensure that the right data quality issues are improved.
- The 80-20 heuristic is useful in the as-is situation but can change in the future when using the tool. This 80-20 heuristic has to be kept up-to-date.
- Expansion of the tool can influence its response time in a negative way.
- Data cannot be put into SAP automatically when corrections are made with the tool, because of the design restrictions stated by VDL Nedcar.

8.5 Recommendations
Based on the findings presented above and in previous sections of the thesis, some recommendations are made to VDL Nedcar regarding the tool and Haug et al.’s (2009) model. The tool can be made more useful by expanding its reach to other databases within VDL Nedcar. Therefore, it is assumed that the intention is to implement the tool concept throughout all databases. To this end, the following recommendations are made.

- The tool should be used in the beginning once a day, and the data should be adapted for a maximum amount of time of 1.88 hours a day. When the data errors decrease in such a way that all of the errors can be adapted in 1.88 hours a day, the frequency of using the tool can be decreased to once a week.
- If the tool is expanded, the requirements should still be satisfied.
- The workflow given by the tool has to be kept up-to-date to ensure the added value of the completeness and the prevention of ambiguousness.
- If it is proven to add value to VDL Nedcar, the tool should be implemented in SAP. Hence, two process steps should be merged into one process step, which will result in higher efficiency.
- The tool will improve the data quality of VDL Nedcar the most if the tool is expanded to check all different databases at once.
- The SCE expert should manage the tool.
- The SCE expert should be the authorised person what has to be contacted in case of a process stop. The SCE expert can appoint someone else to assist him with this task.
- The SCE stakeholders should ask the SCE expert for help if problems occur with the tool.
- The tool should also be expanded to other departments to improve their data quality and to increase the process flow.
- The data quality analysis should be conducted twice a year to set a new 80-20 heuristic. In this way, the quick check will remain significantly important.

8.6 Future research directions
For future research, the following directions are recommended.

- Expansion to other databases could be researched to improve the data in more databases with the same method.
- The option to form one central database that provides all of the data to the other databases should be research. This would significantly decrease data inconsistency.
- The tools run-time should be constantly researched regarding optimisation to ensure employees satisfaction and to react to future expansion.
- More research should be conducted regarding ‘text entry errors’ in the business environment.
Bibliography

Alan Hevner, Salvatore March, Jinsoo Park, & Sudha Ram. (2010). Design Science Research in Information Systems, 42.


Charles Sanders Peirce. (1923). *Chance, love, and logic: philosophical essays*. University of Nebraska Press.


Appendix
I. Organograms

VDL Group

![Organogram VDL Group](image)

Figure 18 Organogram VDL Group
Figure 19 Organogram VDL Nedcar
II. History

History VDL Nedcar

VDL Nedcar is one company that consists of two main histories, the history of VDL and the history of NedCar B.V.. First the history of Nedcar until the take-over by VDL will be elaborated and next the history of VDL. Finally the company’s contemporaneity will be elaborated.

Nedcar

The history of Nedcar starts at the company of DAF. This company wants to develop a person vehicle at the begin 50’s. In 1958 it finally happens, DAF shows their first person vehicle to the public in the RAI - car exhibition. On 17th of December in 1965 plans are announce for the built of a factory in Born. To gain enough scale and to keep costs affordable a cooperation with international partners is inevitable. This is the reason why DAF divides into a company vehicle and a person vehicle business.

At the person vehicle business the first car was produced in 1967. Here has Volvo since end 1972 for 33% in shares. On 31th of January 1975 AB Volvo increased these shares in DAF Car till 75%. The remaining 25% stays in Dutch hands; 10% at DSM and 15% at the DAF Holding. On may the first 1975 the name of the enterprise is changed to Volvo Car B.V.. In 1980 it became clear that Volvo not will manage to produce more than 100.000-120.000 per year. Therefor the Nedcar-project raised: build cars in one company for two brands, each one in quantities of 90.000-120.000 per year. On this moment in time Japanese brands are looking for production-opportunities in West-Europe and Nedcar is for them a great opportunity. In 1991 this goal is realized and Volvo Car Corp. (VCC), Mitsubishi Motors Corp. (MMC) and the Dutch State transform Volvo Car B.V. into a joint venture, Netherlands Car B.V. also known as Nedcar. Each of the parties owns a third in the company. The production line is fully installed by Japanese insights, employees went to Japan to produce the first test series. In 1995, Nedcar is officially opened by Queen Beatrix with the hope that Nedcar will produce 163.000 cars in 1996. Unfortunately these quantities are not achieved through startup problems. The production quantities raised in 1997 and Nedcar produced almost 200.000 cars, thereby Nedcar receives the ISO 9002 certificate. A production record was gained in 1999; 262.196.

In 2000, Daimler Chrysler takes over 34% of the shares and they want to increase these shares till 50%. 30th of March 2001, Mitsubishi becomes 100% owner of Nedcar and Daimler Chrysler has 37,5% of the shares from Mitsubishi on that moment. On April 16th 2002 Nedcar signs the so-called “Contract Manufacturing Agreements” with Mitsubishi Motor Sales Europe B.V. (MMSE) en MCC smart. One year later Nedcar starts the first choice program and in 2004 starts the production of the new Mitsubishi Colt, who won some prices.

On july 12th 2005 Nedcar correspond with social partners that 368 job cuts are necessary due to decreased production volumes. In 2006 Daimler Chrysler announced the production stop of the Smart Forfour. Mitsubishi motors confirms production of the Colt at Nedcar and invests in a facelift program. Nedcar have to search for solutions to survive, 2007 and 2008 were dominated by finding solutions and one of these possible solutions was the Mitsubishi Outlander. In 2008 Nedcar delivers the first Mitsubishi Outlander produced on the secondary lines. End 2008 Nedcar has to adjust its production program due to the financial crisis which leads to the termination of 300 temporary contracts. In 2012 Mitsubishi Motors that it will stop produce the Colt after 2012. In the same year VDL group takes over shares from Mitsubishi Motors Corporation and Nedcar will be renamed to VDL Nedcar. VDL Nedcar starts transformation program towards independent vehicle contract manufacturer. BMW becomes the first customer of VDL Nedcar and VDL Nedcar will produce the MINI.
VDL Group

The history of VDL Group starts in 1953 with the establishment of “Metaalindustrie en Constructiewerkplaats P. van der Leegte”. The first customers of the company founded by the father of Wim van de Leegte, were Philips, the former employer of Piet van der Leegte, and DAF Trucks. The company’s first core services were drilling, turning and milling as well as welding, punching and soldering in series production, which were conducted by five employees.

In the beginning of the 1960’s, the first own made products were made by P. VDL, these were washing machines and oil heaters. 1962 the company received a large order from Honda gain guards, which made them decide to move to Hapert. However, the order was not as large as expected and they brought new finished to the market; stainless steel household articles such as bed warmers, egg cups, butter dishes etc.

In 1966 Pieter van der Leegte had a burnout for various reasons, in this year Wim van der Leegte became involved in the company. He was 19 years old, a student on work placement via the technical college. Within three months he was running the company and in 1972 he assumed ownership of all his father’s shares.

Producing stainless steel household articles became increasingly difficult. Therefore the company decided to concentrate again on serving as supplier in 1976. In 1977 an open-management model with a non-hierarchical structure was introduced. It went so well with the new Business model that Wim van der Leegte was seeking another challenge. He found this in assuming control of poorly operating companies. In 1979, a first step was made towards dynamic growth and the name of the parent company was changes to VD Leegte Metaal. This was the first takeover of the many takeovers that will come.

In 1984 the group grew to five companies, therefore changes would be made on managerial level. A general management and a board of commissioners was formed. Pieter van der Leegte was chairman of this board of commissioners. On 22 January 1991, Pieter van der Leegte, the founder of VDL Group passed away. Two important takeovers in the history of VDL group were the takeover of DAF Bus International and the takeover of production automation company Steelweld in 1993. In 1996 the group consisted of 20 companies, with 2,000 employees, and attained a turnover of 200 million euro’s. Also the headquarters opened in Eindhoven this year. The group continues to grow and again two large takeovers are done in the bus and coach sector; Berkhof Jonckheere Groep in 1998 and BOVA in 2003. In the same year, 2003, VDL celebrates his 50th anniversary. In 2006, the largest takeover in VDL groups history was made with the takeover of Philips Enabling Technologie Group. In 2012, VDL Group wins the prestigious King Willem I Prize in the category of “large enterprise”. The VDL group becomes in 2012 the only major car producer in the Netherlands wit the acquisition of NedCar. VDL Nedcar starts transformation program towards independent vehicle contract manufacturer. BMW becomes the first customer of VDL Nedcar and VDL Nedcar will produce the MINI.
III. Production stream

Press shop
The press shop within Nedcar is seen as an external supplier of parts. The press shop ranges over 26,000 m² and is equipped with three coil lines, six tandem press lines, and two crossbar cup feed transfer presses. Here sheet metal parts of the car can be pressed, however this is happening to a lesser extent through the production of the MINI in Oxford, England and in Born, the Netherlands. For the cabrio, that is exclusively built in Born, the Netherlands, parts are pressed in the press shop. If side panels are pressed, they are stored in the warehouse and automatically departed to the body shop.

Body shop
The body shops covers 60,000 m² with a minimum of 1100 handling and/or welding robots. In the body shop the OEM parts, e.g. side panels and underbody, arrive and these are welded together. The underbody with all its parts is welded together and parallel the complete side-section is constructed and these are merged to each other afterwards. All these steps are handled in work cells, till the skeleton of the car is complete. Then the rear door, hood, the hang-on parts, and small sub-assemblies are assembled.

Paint shop
The complete skeleton arrives from the body shop in the paint shop. The paint shop has several steps to paint the body, control loops ensures that the paint job will be done correctly. Supply chain engineers ensures that the right lacquer used by the paint shop are delivered at the right place.

Final Assembly shop
In the final assembly all the parts that are brought with the in-house logistic process are assembled on the car. The problem which could arise here is that the amount of parts in an packing is less than what is stated in the system. Due to the two bin system, the employee has two times the amount and will not run out before the empty packing will be replaced. If the two bin principle is not possible anymore, cause there are too many models that has to be built, the employee has a big problem if the packing runs out of parts. If the data in SAP, which is used to do the line side planning, is not the same as the real world situation, the packing can have too little parts as stated and delay can occur. Consequently, this delay leads to costs.

Test track & Yard
If the car is finished and tested in-house, the car is ready for the test track(fig. ). Here the car is driven on several subsurface to test the build quality of the car. If the car is tested and is defined as good, it is placed on the yard (fig. ). Here the cars are picked up by trucks to be transported to the dealers.
IV. Additional information

IV.I Design research guidelines

By using the design research guidelines provided by Alan Hevner et al. (2010), the knowledge problems are defined as well as the academic relevance. These guidelines ensure that the knowledge questions are answered who are nested in the practical problems, that will occur during the research. These guidelines are presented below in Table 1.

<table>
<thead>
<tr>
<th>Guideline #</th>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guideline 1</td>
<td>Problem Relevance</td>
<td>The objective of design science research in information systems is to develop technology-based solutions to important and relevant business problems.</td>
</tr>
<tr>
<td>Guideline 2</td>
<td>Research Rigor</td>
<td>Design science research requires the application of rigorous methods in both the construction and evaluation of the design artifact. Often empirical methods are needed to evaluate the artifact as part of a complete human-machine system.</td>
</tr>
<tr>
<td>Guideline 3</td>
<td>Design as a Search Process</td>
<td>The search for an optimal design is often intractable for realistic information systems problems. Heuristic search strategies produce feasible, good designs that can be implemented in the business environment. Decomposition of complex problems is an effective heuristic in the search for effective designs.</td>
</tr>
<tr>
<td>Guideline 4</td>
<td>Design as an Artifact</td>
<td>The designed artifact must be effectively represented, enabling implementation and application in an appropriate environment.</td>
</tr>
<tr>
<td>Guideline 5</td>
<td>Design Evaluation</td>
<td>The quality and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods. Good designs embody a style that is aesthetically pleasing to both the designer and the user.</td>
</tr>
</tbody>
</table>
Effective design science research must provide clear contributions in the areas of the design artifact, design construction knowledge, and/or design evaluation knowledge.

This research has to satisfy these guidelines and thus we will argue why the research does this.

- **Guideline 1:** Since the research is executed within the VDL Nedcar and the research answers questions posed by a certain department of VDL Nedcar, it is certainly relevant for the business.
- **Guideline 2:** The research uses data quality improvement categorizations, which are founded and validated in the literature. Due to the fact that this research is a very specific research which develops a data improvement tool for a certain part of data, this not as rigorous as some other studies.
- **Guideline 3:** The problem solution cycle is followed, so there is designed as a search process.
- **Guideline 4:** The artifacts are the resulting answers to the research questions.
- **Guideline 5:** The data improvement tool will be validated and then been evaluated by the four dimensions of N. Brand & H. van der Kolk (1995) which consists of time, costs, quality, and flexibility.
- **Guideline 6:** This research makes various contributions to the literature. Data quality, the categorization of data quality in ERP, and pitfalls of ERP implementation at companies are conducted.
IV.II Operational research plan

For executing the research properly, an operational research plan is designed. The operational research plan shows the steps that can be performed each week. Furthermore, the important dates are defined, so the deadlines can be made. This is shown in Figure 20.

<table>
<thead>
<tr>
<th>Activity</th>
<th>November</th>
<th>December</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holiday</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>elaborate Literature Review</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>elaborate Research proposal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pilot test and review questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offer questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fill in data in SPSS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overview errors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>determine combinations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>set up Rules-set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test Rules-set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter findings and elaborate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Update readings literature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finish all chapters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concept paper to mentor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concept paper review, layout design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>print paper en binding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hand-in paper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Important data</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. 28 nov: Hand in research proposal
2. 4 jan: Start of field research
3. 1 feb: finish field research
4. 21 mrt: concept paper to mentor
5. 11 apr: hand in final paper

Figure 20 Operational research planning
IV.III The cost of the project
Except for the wages for the intern, which is provided by VDL Nedcar, and the time that is spent for supervision, by VDL Nedcar and the TU/e, few relevant costs are made to support this project. Some time is spent by employees of VDL Nedcar during interviews and feedback sessions with the student, which might be considered costs. The purpose of the research project is to come up with improvements in the data quality, resulting in cost savings that surpass the costs accompanied with the project.

IV.IV The organization of the project
The student will be working at VDL Nedcar Five days a week and. The student started the 24th of August at VDL Nedcar to perform his literature review and develop the research proposal. From the first of January until approximately the end of April the research will be conducted and the master thesis will be written. The scope of the research and content of the thesis are as specified and agreed on in this research proposal. Furthermore, weekly meetings are scheduled with the company supervisors, and biweekly meetings are set with the first supervisor of TU/e. All the meetings are arranged by the student in consultation with his supervisors. Frequency of the meetings could be changed whenever required. The supervisors involved in the project are presented below.

Rik Eshuis
  Universitair Docent
  First supervisor TU/e

Phillipe Calseyde
  Universitair Docent
  second supervisor TU/e

Marcel Janssen
  Data engineer
  Company supervisor VDL Nedcar

Leon Heiligers
  Teamleader Supply chain Engineering
  Company supervisor VDL Nedcar
V. Data collection files

V.I Questionnaire

1. Data awareness (In what level are you aware of all the consequences the data has if it is changed?)
2. Process knowledge to fill all data in the right way (In what level is it clear on which point in the process, which data is needed?)
3. Communication (What is the quality of communication between different departments, considering the data?)
4. Knowledge-sharing (What is the quality of the knowledge that is shared?)
5. Authorization level (In what level are there limitations that not all the employees can adjust data that is set by the engineers? 1-star means that many people can adjust data, 5-star means that few people can adjust the data)
6. Centralized change-management (In what level are changes managed central?)
7. Change process clarity (In what level is it clear how to change the data?)
8. Change process standardized (In what level is the change process standardized? Is there one way everyone changes the data?)
9. Change process simplicity (In what level is it simple to change the data? Are there many steps and resources necessary to check before change can be possible?)
10. Controllability simplicity (In what level is there a possibility to easily check the data on mismatches?)
11. Controllability frequency (In what level are data quality checks performed?)
12. Accessibility (In what level can everyone check the data?)
13. Data change History storage management (In what level are changes kept in a history database)
14. Data change management (In what level is it clear who changed what and why the employee has changed the data?)
15. Accessibility knowledge (In what level knows everyone where they can find the data they need?)
VI. Process models

Figure 21 Process flow LSP

Figure 22 Product flow
Figure 23 Process flow Material Handling
VII. Data Improvement Tool (DIT) user manual

The tool is developed in Visual Basic for Applications to ensure the accessibility for the stakeholders who work every day with Excel. The development of the knowledge based can be explained the best with help of the menu’s created. The tool is named as Data Improvement Tool (DIT).

V.I Visual Basic for Applications

Visual Basic for Applications (VBA) is a programming language which allows the development of user-defined functions. VBA also allows the automation of certain processes and calculations. VBA is a quasi-object oriented programming language, where the system is composed of objects. VBA is built into most Microsoft Office applications, such as Word, PowerPoint, Access, and Excel. Also other application has implemented VBA, such as AutoCAD, SolidWorks, and CATIA. The Data improvement tool is programmed in VBA of the Excel application.

V.II Activation

First the activation command button was developed to call the Data Improvement Tool (DIT). This command button (Figure 24) has to be put in Excel and will than call the data improvement tool macro from the shared workspace.

V.III Main menu

Then the main menu was developed, where the main purposes of the tool are given. The get data and data check command buttons are there to decrease the number of errors in the SAP data. The workflow and the manual are there to provide the needed process knowledge to use the tool in the correct way. With the Exit command button the tool will be closed.

V.IV Get data Process

This process extracts data from two different databases, SAP and a network of VDL. To ensure that the tool automatically extract the data out of SAP, the employee needs to log on to SAP before the
algorithm can be activated. Therefore, the message box given below is showed before the menu to get data will be shown.

![Microsoft Excel](image)

**Figure 26 DIT tool notification**

The menu of getting the data exists of two options, the logistic process and the Packing instruction. The reason for this is that the information necessary for the logistic process is given in different data lists than the information necessary for the Packing instruction. Also both of these selections can be made, which lead to the activation to extract all the data lists out of SAP and the network.

![Getting Data](image)

**Figure 27 DIT tool getting data menu**

The macro in VBA will make a connection with SAP and will automatically extract the data lists into excel, here is each worksheet mentioned to the data list downloaded. From this list the all rows with project number 999 will be erased because these rows are not in use anymore. If the data lists of SAP are ready, the data lists of the network are extracted into different worksheets what are labeled. Finally the file will be saved in the daily network of VDL Nedcar. The daily network deletes every day the network drive, which means that these network will never been replete. If there is already a file with the same name, the file will be overwritten.

**V.V Check data process**

This process will check the SAP data lists on errors which are derived from knowledge of the experts. If this option is chosen in the menu, the following pre-menu will pop up.

![Checking Data](image)

**Figure 28 DIT tool checking data menu**

In this menu the logistic process and the packing instruction can be chosen, which mean the same as in the getting data menu. However, after one of the three choices is made, three different menus will be shown. These are:

- Checking logistic process
- Checking Packing instructions
- Checking Packing instructions & Logistics process
These three menus are explained below.

V.VI Checking logistic process
In this menu the logistic process is checked. All data about the logistic process of the materials is checked on errors. The menu exists of different pages where the stakeholder can make their choices. Every page has a label to make it more understandable. At the page category checks, the categories are shown which are discussed in the analysis phase. These check categories have a page of their own where several more in detail options can be chosen. For example at the relation LO vs IPC the more in detail options are; FAS, Bodyshop, Paintshop. There is also a total check page, where there are the options to choose or to do a total check, or to do a quick check. The total check treats all the categories mentioned at the category check page and the quick check uses the 80-20 heuristic which was necessary concluded in the analysis phase.

Figure 29 DIT tool checking data logistic process menu

V.VI.I A check category
To stay in the example of the LO versus IPC relationship, I will explain what I have done to create this category. I created a check which can be selected, if this is selected then the sub of this relationship will run. A new worksheet will be made and all necessary headings, which are defined by the stakeholders, are created in this file. Then a loop is created that will check the rules defined in the If-then reference list out the analysis phase. The loop will check every row if the row is correct with the given if-then rules.

Table 19 Relationship IF-Then example

<table>
<thead>
<tr>
<th>Relation</th>
<th>IF</th>
<th>Then</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO_IPC</td>
<td>1</td>
<td>3A,D3, all E’s</td>
</tr>
<tr>
<td>LO_IPC</td>
<td>8</td>
<td>D1,D2,D3</td>
</tr>
<tr>
<td>LO_IPC</td>
<td>3</td>
<td>C1,C5, D3</td>
</tr>
</tbody>
</table>

If a row has an error, the necessary headings will be copied and inserted in the new worksheet which is created in the beginning. At the end the founded data will be sorted in ascending order by the Lab Office. Furthermore, the percentages of the errors against the total amount of data will be calculated, to provide parameters for getting an impression of the impact the stakeholder could have on the data. These calculations will be stated in a table next to the results and will also be given in a pop up screen. This will happen for every check category.
V.VI.II Checking packing instructions

In this menu the packing instructions are set. The packing dimensions are checked here. Here the menu exists of two pages, the total page and the dimensions page. On the total page the total check can be done. One the dimensions page, the Volume, length, width, and the height check can be selected. At the packing instruction other headings are necessary then with the Logistic process. These headings are set in the worksheet. These headings are copied from the 921-list if the, for example, length dimension is not the same as the length dimension of the reference list.

![Figure 30 DIT tool checking data checking packing instructions menu](image)

V.VI.III Checking Packing instructions & Logistics process

In this menu both the packing instructions and the logistic process are combined in to one menu. The choices are the same as if the menus are open separately, but now they can be selected together. Both subs are copied to this menu and the only change that has to be made was to adapt the select box numbering.

![Figure 31 DIT tool checking data checking packing instructions & logistics process menu](image)
VIII. Implementation plan

VIII.I Data improvement tool expert
The first step is to assign a data improvement tool expert, who gets the password and only access to the algorithms in the data improvement tool. The expert will get an enhanced manual, which provide the necessary information to set up a new relationship rule or change an existing relationship rule. With changes in the process or in the data-lists in SAP, changes or adding data can be possible proceedings for the manager. This manager will document also the extracted data error lists of each time and will document an insight in the trend what is set by the data improvement tool. This trend can be used for future validation of the data improvement tool and the goals set.

VIII.II SCE Stakeholder training
The second step the stakeholders at the department of Supply Chain Engineering should be prepared for the usage of the tool. The SCE will get a training of the insertion of the data improvement tool and the use of the data improvement tool. Because 6 stakeholders are going to use this tool and the tool is managed by one data improvement tool manager, the tool has to be installed at the network and be activated by the activation button at the quick access toolbar. The training will present how this activation button is installed. In the usage training the SCE learns, how to act in several scenarios, how to use the tool to increase the data quality in SAP, and how to handle with problems in the tool.

VIII.III Authorization protocol
The third and very important step is to set up an authorization protocol, which defines which stakeholder can change data in SAP. Further this authorization has to be realized, so the stakeholders are obligated to communicate the errors in the process to the authorized Supply chain engineers. The Authorization levels can be divided in two sections:

- High level logistic process adaption
- Low level process adaption

VIII.III .I High level logistic process adaption
High level logistic process adaption are adaption what are directly related on the process. For example, the Internal Process Code can change and this is engineered by the SCE. The MHE should not be able to change this IPC because the SCE thinks that the material has another internal process. The engineered IPC may be set up on purpose, because there will be a change in this IPC. That is why the MHE always has to ask for permission to the SCE if they want to adapt the high level process data. This ensures that the department with the most knowledge of changes in the process will set up the process data and will indirectly ensure the increase and maintainability of the data quality.

VIII.III .II Low level logistic process adaption
Low level process adaption are adaption that tell more about the perceptible data errors. For example, the length of the packaging is not correct. The MHE observes this and will have the state of the art information. The MHE should document these changes and sent these changes at the end of the day to the SCE, which can check at the supplier where the error is made. For the MHE counts, no feedback on this adjustments update is good feedback. This ensures that the materials have the correct dimensions in the ERP system as in the real world and that the SCE know of the changes made and can validate this with the suppliers.
VIII.IV Communication protocol
Together with an Authorization protocol, a communication protocol has to be developed. This will ensure an increase in communication between the departments. In this communication protocol, agreements will be set between departments. It is important that the MHE and MHF employees will always inform SCE because these engineers are developing new process and make changes in existing processes. This counts also for the communication of SCE to MHF and MHE. If there is changed, SCE has to inform MHF and MHE to ensure the efficiency of the process on the floor. Otherwise, MHE or MHF will inform SCE something is wrong and SCE has to check this, and still have to inform the MHE and MHF the changed data was correct. This is very inefficient and must be prevented by the communication protocol.
IX. Sequence diagrams

Figure 32 Sequence diagram menu
Figure 33: Sequence diagram check
Figure 34 Sequence diagram manual