THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Automation of Input Data Management

Increasing Efficiency in Simulation of Production Flows

ANDERS SKOOGH

Department of Product and Production Development

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2011

Automation of Input Data Management
Increasing Efficiency in Simulation of Production Flows
ANDERS SKOOGH
ISBN 978-91-7385-604-1

© ANDERS SKOOGH, 2011.

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3285
ISSN 0346-718X

Department of Product and Production Development
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover:
Figure illustrating the need for efficient supply of shop floor data to simulation models. The figure includes original pictures from Volvo Cars Newsroom (available from www.volvocars.com) and Johansson and Zachrisson (2006).

Chalmers Reproservice
Gothenburg, Sweden 2011

Automation of Input Data Management
Increasing Efficiency in Simulation of Production Flows
ANDERS SKOOGH
Department of Product and Production Development
Chalmers University of Technology

ABSTRACT

Production is of significant importance for the social welfare and economic growth in societies worldwide. In Europe, more than 30% of all job opportunities are related to the manufacturing industry. Improvements of material flows in production are of extra importance for reducing the system losses and increasing the robustness of production systems. Unfortunately, the most powerful tools for analyzing dynamic aspects are associated with extensive data requirements and, thus, inefficient procedures for keeping models up-to-date. This thesis addresses the input data management procedure for one such tool, namely discrete event simulation (DES).

The purpose of the thesis is to enable daily use of DES to support production engineers in their work with increasing efficiency, sustainability and robustness of production systems. The aim is to reduce the time-consumption for input data management and thereby facilitate the supply of recent production data to DES models. The thesis is divided into two parts, treated as interrelated studies, addressing one research question (RQ) each.

Part One (RQ1), mapping the industrial state-of-the-art of input data management, is mainly based on qualitative methods including interviews and questionnaires with DES practitioners. The results show that collection of raw data, identification of available data sources, and data analysis and preparation are the three most time-consuming activities. There is still limited use of automatic support systems and the data are often manually collected, processed and supplied to models by means of spreadsheet interfaces. Findings in Part One also show that automated connections to external databases are important for future sustainability analyses using DES.

Part Two (RQ2), proposing and evaluating an approach for automated input data management, is mainly based on the analysis of existing industrial data sources (archive analysis). This review aims to identify the functionalities necessary to automatically transform production data (raw data) to information for a DES model. A demonstrator, called the GDM-Tool, is developed and tested in three independent case studies. The results show that the proposed automated approach reduces the time-consumption for input data management by approximately 75%.

There are still difficulties in input data management for DES, partly due to the limited access to detailed production data. Therefore, the author recommends that industrial and academic partners increase efforts necessary to facilitate continuous raw data collection and, by extension, also automated data processing. In cases where enough data are available, the proposed solution (RQ2) enables more frequent updates of DES models and provides production engineers with a powerful tool for increasing efficiency of production systems on a daily basis.

Keywords: Input data management, Discrete Event Simulation, Sustainable Production.

ACKNOWLEDGEMENTS

It is a common misapprehension that the work and efforts behind a PhD thesis are rather a one-man show. Sometimes, during times with lots of writing, I would definitely agree on that statement. The fact is, though, that even the smallest touch of distance is enough to recall that this work would be impossible without the support provided by helpful people and organizations with whom I have associated.

Firstly, I would like to thank my two supervisors Professor Johan Stahre and Assistant Professor Björn Johansson. I am very grateful for you offering me the possibility to specialize within the very interesting area of production. Today, production is more important and recognized than none of us believed just a few years ago. Of course, you are also a substantial part of the entire process behind this thesis!

I would also like to recognize the organizations supporting my research financially. VINNOVA (Swedish Governmental Agency for Innovation Systems) has funded two projects serving as a basis for this thesis: FACTS (Conceptual Factory Development) and DFBB (Digital Factory Building Blocks). Furthermore, these projects would not have been initiated without the support and participation from our industrial project partners: AB Volvo, Faurecia, Haldex, Scania, VBG Group, and Volvo Car Corporation. In addition, ProViking financed a scholarship making it possible to perform parts of my research at NIST (National Institute of Standards and Technology, USA), which resulted in one of the papers appended to this thesis. ProViking has also substantially contributed to my professional network of Swedish PhD students through their excellent research school.

I strongly appreciate the help and support provided by Edward Williams (University of Michigan and PMC, Dearborn, Michigan, USA). He is always interested in our research and consistently shares his excellent skills when reviewing and proofreading my publications. My only problem is to figure out how to return all favors.

Björn has already been mentioned but deserves some extra appreciation because of his enthusiasm and always positive attitude. We have spent lots of time together during several business trips. Despite an extensive number of work hours, I always get back home with many memorable experiences. Björn is, however, by no means the only colleague at the Department of Product and Production Development contributing to our excellent and inspiring work environment. I do not want to list all names, but I hope to have the privilege of working with you for many years to come. Many thanks also to the colleagues at KTH, Skövde University and Swerea IVF for great cooperation in the research projects listed above.

Last but not least, my family and friends deserve all my gratitude. You make me enjoy life and have fun, which is important in order to refuel with energy. Furthermore, my parents and parents-in-law have supported with baby-sitting et cetera, making it possible to combine the completion of this thesis with parental leave. Finally, to my beloved wife Sofia and my wonderful son Victor: Thanks for sharing your lives, I love you!

Anders Skoogh
Göteborg, October 2011

APPENDED PUBLICATIONS

Publication I	Skoogh, A. and B. Johansson. 2008. A Methodology for Input Data Management in Discrete Event Simulation Projects. In: *Proceedings of the 2008 Winter Simulation Conference*, eds. S.J. Mason, R. Hill, L. Moench, O. Rose, T. Jefferson, and J.W. Fowler, 1727-1735.
	A. Skoogh initiated the paper, performed the data collection and the data analysis. He wrote the paper with assistance of B. Johansson and presented the paper at the Winter Simulation Conference in 2008.
Publication II	Skoogh, A and B. Johansson. 2009. Mapping of Time-Consumption During Input Data Management Activities. *Simulation News Europe*, 19(2):39-46.
	A. Skoogh initiated the paper, performed the data collection and the data analysis. He wrote the paper with assistance of B. Johansson.
Publication III	Skoogh, A., T. Perera, and B. Johansson. Submitted. Input Data Management for Simulation - Industrial Practice and Future Trends, *Simulation Modelling Practice and Theory.* (Submitted for publication.)
	A. Skoogh initiated the paper, designed the survey and participated in the data collection. He wrote the paper together with T. Perera and B. Johansson.
Publication IV	Skoogh, A., B. Johansson, and L. Hansson. 2011. Data Requirements and Representation for Simulation of Energy Consumption in Production Systems. In: *Proceedings of CIRP Manufacturing Systems 2011*.
	A. Skoogh initiated the paper, participated in the data collection and performed the data analysis. He wrote the paper with assistance of B. Johansson and L. Hanson. Further, he presented the paper at CIRP Manufacturing Systems in 2011.
Publication V	Skoogh, A., B. Johansson, and J. Stahre. Submitted. Automated Input Data Management: Evaluation of a Concept for Reduced Time-Consumption in Discrete Event Simulation. *Simulation: Transactions of the Society for Modeling and Simulation International*. (Under 2nd revision.)
	A. Skoogh initiated the paper, designed the presented approach and the demonstrator together with M. Johansson*, J. Balderud** and A. Olofsson**. He wrote the paper together with B. Johansson and J. Stahre. * guest researcher at NIST, USA ** two very competent software engineers
Publication VI	Skoogh, A., J. Michaloski, and N. Bengtsson. 2010. Towards Continuously Updated Simulation Models: Combining Automated Raw Data Collection and Automated Data Processing. In: *Proceedings of the 2010 Winter Simulation Conference*, eds. B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1678-1689.
	A. Skoogh initiated the paper. He designed and evaluated the approach together with J. Michaloski and N. Bengtsson. All three participated in writing the paper and A. Skoogh presented it at the Winter Simulation Conference in 2010.

ADDITIONAL PUBLICATIONS BY ANDERS SKOOGH

Skoogh, A. and B. Johansson. 2007. Time-Consumption Analysis of Input Data Activities in Discrete Event Simulation Projects. In: *Proceedings of the Swedish Production Symposium 2007*.

Ng. A., M. Urenda Moris, J. Svensson, A. Skoogh and B. Johansson. 2007. FACTS Analyser: An innovative tool for factory conceptual design using simulation. In: *Proceedings of the Swedish Production Symposium 2007*.

Skoogh, A., J.-P. André, C. Dudas, J. Svensson, M. Urenda Moris and B. Johansson. 2007. An Approach to Input Data Management in Discrete Event Simulation Projects: A Proof of Concept Demonstrator. In: *Proceedings of the 6th EUROSIM Congress on Modelling and Simulation*, eds. B. Zupančič, R. Karba and S. Blažič.

Johansson, M., B. Johansson, A. Skoogh, S. Leong, F. Riddick, Y.T. Lee, G. Shao, P. Klingstam. 2007. A Test Implementation of the Core Manufacturing Simulation Data Specification. In: *Proceedings of the 2007 Winter Simulation Conference*, eds. S.G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J.D. Tew, and R.R. Barton, 1673-1681.

Skoogh, A. 2009. *Methods for Input Data Management – Reducing the Time-Consumption in Discrete Event Simulation*. Thesis for the Degree of Licentiate of Engineering, Chalmers University of Technology.

Alin, D., J. Andersson, M. Andersson, A. Isaksson, A. Skoogh and E. Helander. 2009. Examining the Relation Between EPEI-Time and Productivity Using Discrete Event Simulation. In: *Proceedings of the 2009 Swedish Production Symposium*, (2):295-301.

Bengtsson, N., G. Shao, B. Johansson, Y.T. Lee, S. Leong, A. Skoogh and C. McLean. 2009. Input Data Management Methodology for Discrete Event Simulation. In: *Proceedings of the 2009 Winter Simulation Conference*, eds. M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin and R.G. Ingalls, 1335-1344.

Johansson, B., M. Mani, A. Skoogh and S. Leong. 2009. Discrete Event Simulation to Generate Requirements Specification for Sustainable Manufacturing Systems Design. In: *Proceedings of PerMIS'09*.

Boulonne, A., B. Johansson, A. Skoogh and M. Aufenanger. 2010. Simulation Data Architecture for Sustainable Development. In: *Proceedings of the 2010 Winter Simulation Conference*, eds. B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 3435-3446.

Michaloski, J., B. Raverdy, B.E. Lee, F. Proctor, S. Venkatesh, N. Bengtsson and A. Skoogh. 2010. Push–Button Discrete Event Simulation For Analysis of Factory Floor Operations. In: *Proceedings of ASME 2010 International Mechanical Engineering Congress & Exposition*.

Gustafsson, B. and A. Skoogh. 2011. Design and Problem Oriented Education Based on the Application of Knowledge – developing Chalmers' Master's Programme in Production Engineering. In: *Proceedings of the Swedish Production Symposium 2011*.

Andersson, J., A. Skoogh, B. Johansson and S. Leong. 2011. Environmental Activity Based Cost Using Discrete Event Simulation. In: *Proceedings of the 2011 Winter Simulation Conference*, eds. S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu. (Accepted for publication.)

Lindskog, E., L. Lundh, J. Berglund, Y.T. Lee, A. Skoogh and B. Johansson. 2011. A Method for Determining the Environmental Footprint of Industrial Products Using Simulation. In: *Proceedings of the 2011 Winter Simulation Conference*, eds. S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu. (Accepted for publication.)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ACS = Automated Collection System
CA = Controllability Analysis
CAD = Computer Aided Design
CAEX = Computer Aided Engineering Exchange
CBS = Corporate Business System
CMSD = Core Manufacturing Simulation Data
CNC = Computer Numerical Control
DES = Discrete Event Simulation
DFBB = Digital Factory Building Blocks
ED = Enterprise Dynamics
ELCD = European reference Life Cycle Database
ERP =Enterprise Resource Planning
GDM = Generic Data Management
IDEF = Integrated computer aided manufacturing DEFinition
ISO = International Organization for Standardization
IT = Information Technology
KS = Kolmogorov-Smirnov
LCA = Life Cycle Assessment
MDA = Manufacturing Data Acquisition
MES = Manufacturing Execution System
MLE = Maximum Likelihood Estimation
MRP = Material Requirement Planning
MTBF = Mean Time Between Failures
MTTR = Mean Time To Repair
NIST = National Institute of Standards and Technology
OEE = Overall Equipment Efficiency
OLE = Object Linking and Embedding
OPC = OLE for Process Control
PLC = Programmable Logic Controller
PLM = Product Lifecycle Management
RQ = Research Question
SCADA = Supervisory Control And Data Acquisition
SISO = Simulation Interoperability and Standards Organization
SME = Small and Medium Enterprises
STEP = Standard for the Exchange of Product model data
SQL = Structured Query Language
TBF = Time Between Failures
TTR = Time To Repair
UML = Unified Modeling Language
UPLCI = Unit Process Life Cycle Inventories
WSC = Winter Simulation Conference
XML = Extensible Markup Language

1 INTRODUCTION

Production is of major importance for the welfare and development of societies. Compared to other lines of business, production offers more employment possibilities and contributes more to economic turnover than any other area. This is exemplified by the statistics in Figure 1, showing that 30% of the jobs and 42% of the economic value added to the European Union in 2006 stem from the production area.

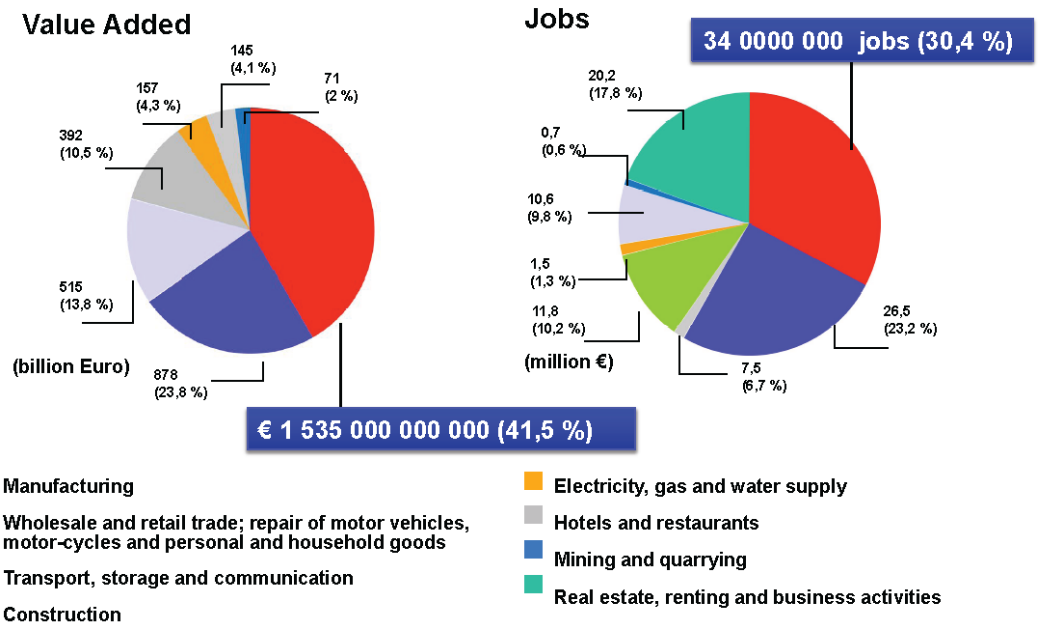

Figure 1: Production is important! The production area contributes major parts of the employment opportunities and economic turnover in the European Union (MANUFUTURE Eurostat 2006).

A central factor for competitive companies contributing to a sustainable society is to rapidly adapt to changes in consumption patterns. When it comes to capacity issues, this means responding to increased demands with short ramp-up times. Unfortunately, this adaptation is often handled by significant investments instead of increasing utilization of existing equipment. Consequently, the Overall Equipment Efficiency (OEE), as an indicator of production efficiency, is only around 50% on average in Swedish industry (Ingemansson 2004).

Naturally, to increase competitiveness, it is very important to focus on reducing the inherent losses by complementing investment projects with continuous improvement of production flows. Discrete Event Simulation (DES) is a powerful tool for such improvements, especially when regarding dynamic aspects of production systems (Ingalls 2002, McLean and Leong 2001). This is necessary in order to effectively reduce balancing losses and system losses (Wild 1975), the latter caused by varying processing times and unplanned disturbances.

However, despite its potential, DES is often renounced in favor of static production analysis tools or qualified guesses, which is due to the extensive time-consumption of dynamic simulation studies (McNally and Heavey 2004). This leads to less detailed analyses, and by extension, to production systems designed for ideal circumstances, disregarding variation and disturbances. The reason for the extensive time-consumption is arguably the input data management process, which takes time due to its importance for model quality and the specifically detailed data requirements DES (Moon and Phatak 2005, Robertson and Perera 2002).

Therefore, strategic work towards more efficient input data management is required, including structured guidelines for data collection and IT solutions to automate parts of the process. Automation would enable continuous update of the data required in simulation models and facilitate DES to be used on a daily basis, as a desktop resource for production engineers. Such a desktop resource provides possibilities to reduce balancing and system losses by continuous and fact-based action, e.g. in production planning, management of systems constraints, and re-balancing. In addition, recent research shows that an increased use of DES helps companies to reduce the environmental impact of production and, thus, improves their sustainability performance (Solding, Petku, and Mardan 2009; Heilala et al. 2008).

1.1 Input Data Management in Simulation

Here, input data management is defined as all activities required to obtain quality-assured, and simulation-adapted, representations of all relevant input parameters for a simulation model - in other words, the process of managing the input data required in simulation models. Input data management starts with the collection of raw data and ends with providing processed data as information to a simulation model, typically in a standardized file or a customized spreadsheet. Thus, the actual process of supplying the final information into the simulation model, often realized by an automated connection between the data interface and the simulation model, is excluded. The entire process is visualized in Figure 2. The three areas of input data management all consist of one or several activities, e.g. data correction as a part of data processing. The activities can in turn be broken down in tasks, for example to remove irrelevant data points. Tasks specifically performed in data processing use data operations, for instance filtering of data points based on date and/or time.

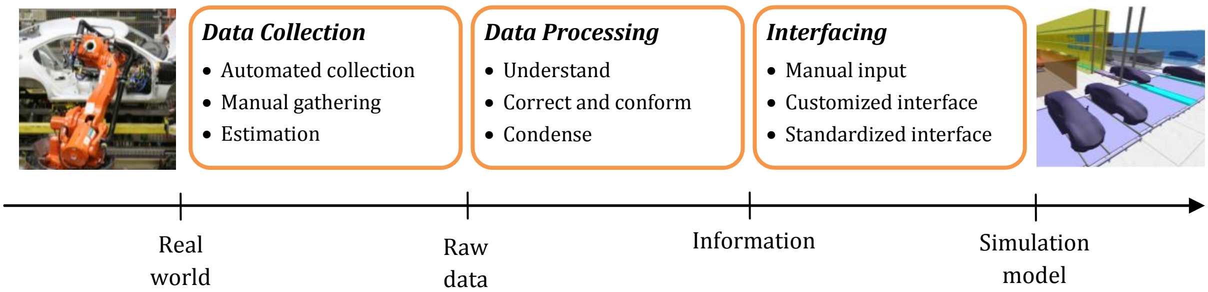

Figure 2: The different parts of the input data management process.

As indicated in the background, input data management is a time-consuming step contributing as much as 40% of the project time in simulation studies (Trybula 1994). The reasons are that many different aspects and parameters of production resources are included in detailed models, and that stochastic representation of simulation parameters requires lots of raw data samples. For example, it is desirable to collect more than 200 real-world measurements when representing machine breakdown patterns, such as Mean Time Between Failures (MTBF) and Mean Time To Repair (MTTR), in dynamic simulations (Perrica et al. 2008). In addition, substantial care is required due to the importance of data for model quality (McNally and Heavey 2004). Thus, the expression *Garbage in – Garbage out (GIGO)* is often used among simulation specialists (Robertson and Perera 2002).

The positive aspect of this problem is that improvements in the input data management process hold great potential to reduce the time-consumption of entire simulation studies.

At present, the input data management process almost exclusively implies some manual participation. This includes manual methods for data gathering as well as human involvement in data processing, e.g. data cleansing using formulas in MS Excel® and data condensation supported by distribution-fitting software or similar functionality provided by simulation applications. Robertson and Perera (2002) identified only one case of a completely automated connection to computer applications within the Corporate Business System (CBS) for the bulk of data. The reason is arguably the lack of simulation data in these applications, especially data describing the dynamics of production systems (Moon and Phatak 2005). Moreover, finding all necessary input data in a condensed form, suitable for simulation, is very unlikely.

In current research and development on input data management, the main focus includes elevating the level of automation by integration of simulation models and major data sources within the CBS (Robertson and Perera 2002, Randell and Bolmsjö 2001). These data sources are typically exemplified by Enterprise Resource Planning (ERP) systems and Product Lifecycle Management (PLM) systems. Such integration certainly holds potential but faces several challenges due to the diversity of simulation tools (Semini, Fauske, and Strandhagen 2006) and the lack of detailed data mentioned above.

1.2 PURPOSE

The purpose of this thesis is to enable continuous use of DES as a desktop resource for production engineers. By extension, this will lead to more efficient production flows, which are well balanced with regard to varying processing times and robustly designed to reduce the negative effects of production disturbances. Better utilization of existing resources will of course reduce the need for major investments in new capacity. The challenge is to efficiently manage the extensive demand for data and information for dynamic simulations.

1.3 AIM

The aim of this thesis is to enable reduction of the time-consumption for input data management in simulation of material flows in production, i.e. the time required for the process to manage the input data. This work is mainly focused on the processing of raw data to information, with the assumption that automation of data input activities is an appropriate alternative. Standards for raw data collection and data interfaces, enabling efficient supply of updated information to simulation models, are briefly evaluated. Throughout the input data management process, the quality of data must be similar to the result of a common industrial approach.

One additional aim is to contribute to the development of industrial tools effecting automated input data management for DES. A demonstrator will serve as an aid to convey the thoughts behind the presented concepts and exemplify how the research can be realized in industry. However, the presented demonstrator, called the GDM-Tool (Generic Data Management), is not intended to be a commercial product itself; it should rather be used as inspiration for further development by major DES users or other software vendors in the area of production data management.

1.4 RESEARCH QUESTIONS

Two research questions (RQ) have been identified as specifically important for satisfying the purpose and aim of this thesis. The answer to the first one describes industrial practice in the process of preparing input data for simulation models, and the subsequent question concerns how the efficiency of activities in this data input process can be increased.

RQ1. What is the current industrial state-of-the-art for the input data management process?

Describing the industrial state-of-the-art is an important reference and starting point when developing new methods and tools for the input data management process, as in RQ2. Important activities have to be identified and specifically those having highest impact on the total time-consumption. Further, since an automated solution is presumed, the current level of automation for the data input activities has to be identified.

RQ2. How can efficient and automated input data management, for simulation of material flows in production, be realized?

In the design of a concept for increased efficiency in input data management, it is crucial to identify important functionalities for automating the data input activities. The proposed concept is demonstrated as a software solution and its general applicability is tested in three case studies. The thesis also evaluates whether and how much the time-consumption can be reduced compared to a traditional industrial data input procedure (a reference procedure).

1.5 FOCUS

The results of this thesis are based on research within manufacturing companies, mainly in the automotive and aerospace industries. Production data typically come from production lines with automated and semi-automated work stations, performing machining or assembly operations. This includes for example milling, drilling, turning and material handling operations as well as assembly tasks supported by hand-held tools. However, despite the focus on manufacturing companies, and the listed operations, it is supposed that the designed and evaluated methods can easily be applied also in other types of industries.

Another aspect to mention is that the research is strongly related to DES for analysis of material flows in production. This is shown throughout the aim, research questions, and case descriptions and, consequently, it is impossible to claim and promise a wider application area. However, it is likely that the outlined and evaluated solutions to data management also work for other tools with similar data requirements, such as line balancing, scheduling, and production monitoring applications.

1.6 DELIMITATIONS

A central word in this thesis is *data* and throughout the text it refers to quantitative data, e.g. processing times and repair times. Thus, logical relations, such as routing rules and priority decisions on the shop floor, are disregarded. That kind of information is assumed to be handled during model building. Consequently, in order to use DES as a desktop resource, the automated solution for input data management has to be complemented with routines for keeping the simulation model itself updated when changes in the production system occur.

Data quality is not specifically addressed in this research. The aim of all described methods and tools is to reduce the time-consumption during data input activities. Nevertheless, it is clearly stated that the data quality must be similar to a traditional approach to input data management in industry.

During the design of automated input data management (RQ2), the focus has been on increasing efficiency of the processing of raw data to information, i.e. data processing (section 1.1). The measured time-consumption (Δt) is therefore delimited to start from available sets of raw data and stop when the simulation information is exported to a simulation interface; see Figure 3. The actual collection of raw data (e.g. sensors and databases) and the development of standards for data interfacing towards the simulation models are, thus, delimited. However, existing solutions have been included in the evaluation of the concept and tested together with the proposed demonstrator for automated input data management.

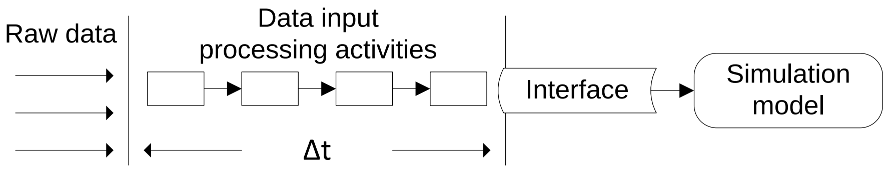

Figure 3: Delimitations for research and time measurements in the input data management process.

Because of the limited time, the evaluation of automated input data management (for RQ2) is delimited by three case studies from two types of industries, automotive and aerospace. Specifications of necessary data operations as well as the circumstances for validation are therefore dependent on conditions from these cases. Moreover, when comparing the time-consumption for automated input data management and the traditional industrial approach, the influence and possible variation due to different users are delimited. Each specific data input procedure is measured once, with a single user, because simulation engineers with enough experience in input data management are few in the case study companies.

1.7 Thesis Structure

This thesis follows the structure of interdependent studies, described by Wilkinson (1991). The first part includes the research performed to answer research question 1 (RQ1) and contains the Publications I, II, III, and IV. After an interim discussion, the results lead to the initialization of a second "study". Naturally, this part contains the research related to research question 2 (RQ2), which includes Publications V, VI, and an additional previously unpublished case study. The complete outline is visualized in Figure 4. Notice that the word "study" here is not equal to "case study" used later in the thesis.

Chapter	**Content**
Introduction	Includes background, definition of input data management, aim, and research questions.
Method	Describes the work procedure and research methods applied in both parts (for answering RQ1 and RQ2).
Frame of reference	A joint review of literature used for answering both research questions.
Results and interim discussions	The results chapter combines the presentation of all appended publications with interim discussions, including reflections on the methods used. The chapter is organized as follows: Results RQ1, discussion RQ1, results RQ2, and discussion RQ2.
General discussion	Synthesizes the results of both parts with literature, and reflects back on the thesis' purpose and aim.
Conclusions	Condensed answers to RQ1 and RQ2.

Figure 4: Outline of the thesis, following the structure of interdependent studies (Wilkinson 1991).

2 METHOD

This thesis originates in systems theory and the overall objective is to describe how system entities such as people, machines and IT solutions interact as one unit. This implies that the system as one unit might have different qualities and characteristics than the sum of the individual objects (Wallén 1996). For example, disturbances on individual machines can occur without affecting the performance of an entire production system, especially on machines which are not current bottlenecks. The research questions aim to map and improve the interaction and division of tasks between engineers and their support tools in order to increase efficiency in the analysis of production systems.

An empirical approach (Flynn et al. 1990) is applied to study the behavior of production systems, their support functions and the effects of methods and tools proposed in this thesis. Hence, information for state-of-the-art descriptions (RQ1) and design criteria for the demonstrator of automated input data management (RQ2) are collected from industrial simulation projects. Additionally, test implementations and evaluation of the demonstrator are performed in real-world industrial case studies. The selection of research methods is strongly inspired by Flynn et al. (1990) and their thorough description of empirical research in operations management.

A fundamental objective of this research is to ensure that proposed methods and tools are applicable and value-adding for industry. It is considered very important to really understand the industrial work procedures regarding input data management and to closely study the effects of the suggested improvements. Therefore, the author has cooperated with industrial partners (FACTS and DFBB projects (Chalmers PPU 2011)) and participated in the mapping of current practice as well as development, implementation and validation of the automated support systems. Such an approach is called Action Research (Coughlan and Coughlan 2002). The combination of empirical and action research is often criticized for being close to industrial development, which should be performed by the companies themselves. A key aspect is that academic research has to ensure that generated knowledge is generic, which requires the use of recognized academic methods, transparent descriptions enabling repeated tests, close and iterative connection to theory, and validation in multiple case studies (Flynn et al. 1990, Gummesson 2000).

Applied research based on empirical findings often incorporates methods originating from social science. Therefore, in contrast to classical research on a technical faculty (closer to positivism), qualitative elements are increasingly applied in operations research due to the importance of organizational aspects. This fact has of course had its impact on the methods presented later in this chapter. The qualitative methods are used to collect experiences and opinions from the human resources in the studied organizations in order to ensure correct understanding of current work procedure and desired efficiency of proposed improvements. The mix of traditional quantitative methods and the qualitative elements described above is encouraged by several authors stating that research methods should be selected on the basis of the research purpose and aim, instead of old research traditions (Bryman and Bell 2007, Danemark et al. 1997). Table 1 contains a compilation of the selected methods.

Table 1: A mapping of the research methods used in the appended publications.

	Design			Data collection					Data analysis		
Part1 – RQ1	Survey	Case study	Multiple Case Studies	Questionnaires	Interviews	Physical measurements	Archive analysis	Participant observations	Descriptive statistics	Interview coding	DES validation approaches
Publication I	X				X				X	X	
Publication II	X			X	X				X	X	
Publication III	X			X					X		
Publication IV		X				X			X		
Part2 – RQ2											
Publication V			X				X	X	X		X
Publication VI			X				X	X	X		X
Unpublished Case			X				X	X	X		X

Part 1, corresponding to RQ1, has the aim of *theory development* describing the industrial state-of-the-art for input data management. An inductive approach (Wallén 1996) using qualitative methods dominates the research design. Interviews are used to define the data input activities as well as the tools and techniques supporting the process. Questionnaires are applied to assess the time-consumption for each activity and for identifying the levels of automation in industrial input data management. Data analyses are based on theoretical coding and descriptive statistics; see further information below.

Part 2, corresponding to RQ2, combines *theory development* and *theory verification* using a deductive approach (Starrin and Svensson 1994). A demonstrator of a concept for automated input data management is developed, based on previous theory. Historical archive analysis is used to specify the requirements on the software, based on the design of raw data sources in three industrial case studies. Participant observations are used to evaluate the performance of the demonstrator. As far as possible, within the given timeframe, the results are validated using the multiple case study design and space triangulation; see below.

2.1 WORK PROCEDURE

The following figure (Figure 5) illustrates the research procedure. Theoretical inputs to the research are displayed above the time-line and the empirical contributions, such as case studies in research projects, below. Note that the publications are numbered to support the line of arguments in this thesis instead of in a sequential order. Further, the publications' position on the time-line illustrates when the original work was performed rather than their publication year. This means, for example, that journal articles are dated by when the study was completed and documented, thus not including the review process.

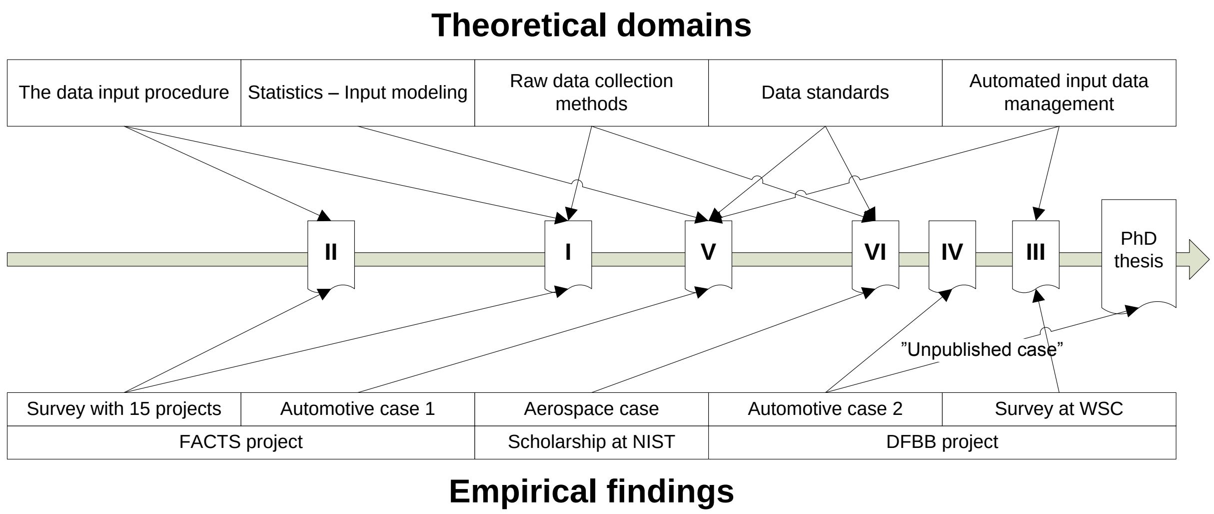

Figure 5: Research procedure: publications mapped against theoretical domains and empirical case studies.

The theoretical parts are divided into six main domains where most of the literature review can be categorized. However, some additional theory is provided in Chapter 3. The empirical findings are obtained within two different projects: FACTS (Conceptual Factory Development) and DFBB (Digital Factory Building Blocks) (Chalmers PPU 2011). The FACTS project developed a simulation tool and an input data management application intended for analysis of production systems on a conceptual level in early project phases. The DFBB project currently realizes a database containing production equipment represented on neutral formats, enabling information sharing between various engineering tools and continuous update of operational data. Additionally, one case study was performed in cooperation with NIST (National Institute of Standards and Technology, USA) during a scholarship sponsored by the ProViking research school (ProViking 2011).

2.2 RESEARCH DESIGN

The first thing to do after establishing the research questions is to select a research design. Depending on whether the question concerns theory development or theory verification, different designs are appropriate. Flynn et al. (1990) list several types of research designs including single or multiple case studies, field studies, panel studies, focus groups and surveys. Note that the research designs are not detailed methods themselves, but serve as umbrellas for various data collection and analysis methods, often combined with each other.

In this thesis, surveys are selected for most of the theory development in RQ1. A single case study is also used as a supplement for collecting the necessary data in Publication IV. Further, a multiple case study design is applied for the development and validation of a concept for automated input data management when answering RQ2.

2.2.1 SURVEYS

Surveys, using qualitative data collection methods such as interviews and questionnaires, are suitable for studies defining state-of-the-art within a specific group or context (Flynn et al. 1990). In RQ1, the aim is to describe the industrial state-of-the-art in input data management for DES in manufacturing industry. According to the same article, it is common to select a population which is homogeneous with regard to one important characteristic. As a result, the population for the interviews and questionnaires in this survey represents DES users from a set of industrial projects performed at companies with various location, size, line of business, and previous experience of DES in order to cover a broader spectra of industrial practice.

A second survey is also performed within the scope of RQ1. This survey aims to map the level of automation in the industrial input data management process. To cover a wider range of companies and also to include possible differences around the world, the population was extended to consist of more people but still homogeneous with DES practitioners. The DES practitioners include industrial representatives, consultants and researchers at the major research conference in the area called the Winter Simulation Conference (WSC). See section 2.3.2 for further information about the data collection using questionnaires.

2.2.2 MULTIPLE CASE STUDIES

The case study design is in general good for collection of detailed data and information from real-world environments, e.g. using archive analysis and interviews. From the collected data and information, a case study is also a powerful approach in order to investigate interactions between a phenomenon and its real-world context (Dubois and Gadde 2002, Wallén 1996). This characteristic in combination with its strength to deal with a full variety of evidence, such as observations, interviews, documents and artifacts, contributed to case studies' appropriateness in this thesis. Therefore, this research design is selected for collecting the data requirements on a demonstrator for automated input data management and also to investigate the impact of automation on the time-consumption (RQ2).

Many of the negative effects related to case studies stem from the involvement of the researcher, who is always present, at least in the role of an observer. This close involvement has meant that case studies are frequently challenged with regard to objectivity and credibility (Wallén 1996), due to unavoidable interpretations by the researcher. Therefore, it is important to declare and discuss the researcher's background and also to explain the real-world context from where the results are observed. Compare the discussion on empirical and applied research provided earlier in this chapter.

Based on Weick (1979), Dubois and Gadde (2002) describe another common criticism against case studies. This is that case studies sometimes tend to be too broad, trying to describe everything, which usually ends up by describing nothing. To overcome this drawback, case studies must be preceded by the description of a solid theoretical framework (Yin 1994). Furthermore, understanding the background theory will facilitate the research design and by extension give good guidance of what relations to establish, what data to collect and so forth. Previous theory has served as a basis for designing the concept of automated input data management and for developing the demonstrator later used for validation and presentation purposes.

2.3 DATA COLLECTION

Four data collection methods have been applied in order to systematically document observations. Interviews and questionnaires are, partly in tandem, used to describe current industrial practice (RQ1). In addition, physical measurements of electrical power have been performed in order to investigate how environmental parameters should be represented in DES studies (also RQ1). Archive analysis is the major data collection method during the development and validation of the demonstrator for automated input data management (RQ2).

2.3.1 *INTERVIEWS*

In this thesis, interviews are applied to describe the industrial state-of-the-art of input data management (RQ1). More specifically, required activities and applied techniques are identified on the basis of empirical findings from previously performed DES projects. In general, interviews are appropriate for gaining insight into people's opinions and experiences (Denscombe 2007), which is valuable when identifying current industrial practice.

Interviews can be designed in different ways depending on the type of data to collect. Three common categories are: *structured*, *semi-structured* and *unstructured* interviews (Denscombe 2007). Firstly, a structured interview means that the interviewer keeps tight control over the topic, format and order of the questions. It is almost comparable to a face-to-face questionnaire and is usually applied to collect large amounts of data. In contrast, unstructured interviews aim to let the interviewees develop their thoughts more thoroughly about topics that interest themselves. The researcher just plants a topic or an issue, which results in problems with foreseeing the outcome of the interview, but hopefully gives a deep understanding of information conveyed by the interviewee.

In this research, the semi-structured approach has been applied, which combines the qualities of the structured and unstructured interviews. It was therefore chosen for keeping control and focus of the process but still being able to get in-depth answers. In practice the researcher has a clear list of issues to deal with, but is still flexible with specific topics and order of questions, letting the interviewee control parts of the process (Denscombe 2007). Moreover, another reason is that the administration of semi-structured interviews is easier thanks to increased predictability of resources and time-consumption compared to the unstructured approach.

2.3.2 *QUESTIONNAIRES*

Two different questionnaires are used in this thesis. One is face-to-face, clipboard style (Denscombe 2007), performed on the same population as the interviews described above. The aim of this questionnaire was to assess the time-consumption related to the different activities in the input data management process. The second one is a questionnaire investigating the different levels of automation in the process of data input preparation. This questionnaire was supplied to DES experts at the WSC 2010 and via a web-based format for reminders and additional answers.

A questionnaire is a good alternative to interviews when the questions are straightforward and when large amounts of data will be collected (Denscombe 2007). Furthermore, the fact that all respondents answer the same questions provides standardized answers, which are often correct and easy to analyze. Questionnaires provide either *factual information* or *opinions* (Denscombe 2007). Here, factual information about industrial practice is collected in both Publications II and III.

2.3.3 *ARCHIVE ANALYSIS*

Archive analysis is a data collection method closely tied to single or multiple case study designs (Flynn et al. 1990). Throughout all three case studies used for answering RQ2, production data have been collected from databases containing breakdown frequencies, repair times and processing times for workstations. These data are used to identify requirements on the demonstrator for automated input data management, such as data formats to support and necessary data operations (e.g. to calculate MTBF and MTTR).

The most significant advantage of archive analysis is that the data are unbiased because they are collected prior to the initiation of research (Flynn et al. 1990). Consequently, the demonstrator developed here will be designed to work with real-world industrial raw data not influenced by special needs of laboratory environments. A negative aspect is that it may be impossible to find all necessary data, since the researcher does not control the sources of data collection. An example is that the three companies included in this thesis did not have all parameters necessary for DES in their existing data sources. To cover up for this shortage, supplementary requirements on the demonstrator were identified during additional data gathering and informal interviews within the project groups.

2.3.4 ELECTRICAL POWER MEASUREMENTS

Publication IV maps the variability of electrical power used as an input parameter in DES models. This is an initial step in describing how new simulation input parameters for environmental analysis should be handled and represented. In a separate case study, the electrical power utilization of five multi-operational machines was sampled during production. The power monitoring equipment was connected to the incoming three-phase connection using Y-connections (Stevenson 1982). See Publication IV for more details about the machines and the equipment.

2.4 DATA ANALYSIS

Two major strategies have been applied for analysis of gathered data. Interview analysis, including structured coding and theory development, has naturally been used for the interviews related to RQ1. In other analyses, statistical analyses have been used for compiling questionnaire responses and electrical power measurements (also RQ1). The same method has also been applied for evaluating the effects of automated input data management with regard to time-consumption and data quality (RQ2).

2.4.1 INTERVIEW ANALYSIS

The analysis of interview results in Publications I and II was performed using an approach based on Grounded Theory (Glaser and Strauss 1967). Materials based on interview notes were initially coded with as little reference to previous knowledge as possible. This step is similar to the term *open coding* in Grounded Theory. Later, the initial codes were compared to each other in order to find synergies and key codes, which can be compared to the step of *selective coding* in Grounded Theory. The last step in Grounded Theory, *theoretical coding*, results in formulation of a theory based on the codes from steps one and two in the previous process. In this thesis, significant codes were mainly selected on the basis of frequency. In other words, the state-of-the-art description in Publication I is a compilation of the most common activities and techniques used by DES practitioners.

2.4.2 STATISTICAL ANALYSIS

Basic statistical calculations have been performed in order to describe industrial practice, mainly to analyze the results from the questionnaires in Publications II and III. In other words, such calculations are used to evaluate how frequently different levels of automation are applied and for assessing the time-consumption in different data input activities. Descriptive statistics is a common and useful analysis method for description of current work procedures in empirical research (Flynn et al. 1990, Miles and Huberman 1994). Moreover, similar calculations are used to identify the most common data input activities and techniques from the interview materials in

Publication I, and for quantifying the variance of electrical power from the measurements in Publication IV.

As a clarification, more advanced statistical analyses, such as parameter estimation using maximum-likelihood estimation and distribution fitting with the Kolmogorov-Smirnov test, were applied as functionalities within the GDM-Tool (RQ2). Hence, these types of analyses are considered as parts of the systems design for the GDM-Tool rather than research methodologies; see sections 0 and 3.3.2.4.

2.5 VALIDATION

The description of research strategies in the introduction of this chapter includes terms such as *empirical research, applied research* and *action research*. All of these terms are related to challenges regarding validity and generality (Flynn et al. 1990, Gummesson 2000). However, using several cases in a multiple case study design provides a basis for validation of the proposed concept. Flynn et al. (1990) state that multiple case studies are capable of theory validation by supporting or falsifying proposed concepts. Three case studies are included in this thesis for validation of the proposed approach to automated input data management. The case studies are performed in three different companies, representing two lines of business, which is considered enough for initial validation.

This line of argument is also supported by Denscombe (2007), in his description of *triangulation*. Triangulation often means collection of the same data using different collection methods. However, Denscombe extends the term (data triangulation) to include the collection of data from different data sources, e.g. with various cultural, social or geographical contexts. These contexts are here varied by the selection of different case study companies; see Figure 6.

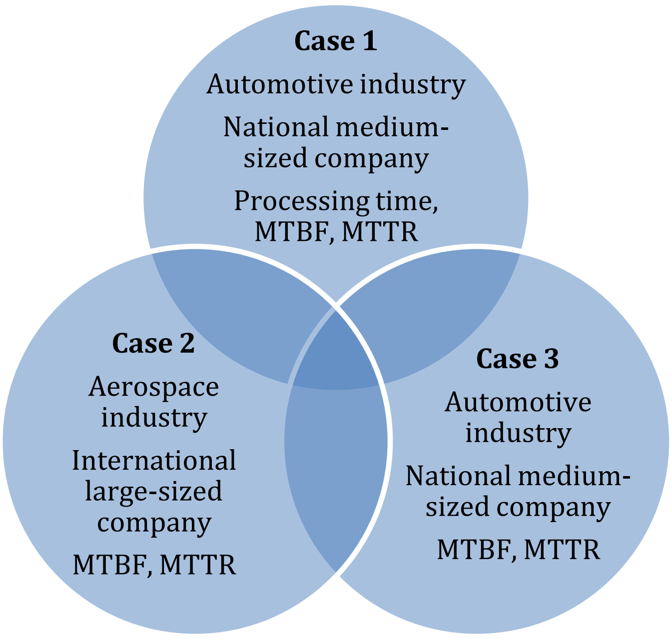

Figure 6: Data triangulation for validation of the GDM-Tool in a multiple case study design.

Moreover, to evaluate the data quality between the two approaches in Publication V, statistical hypothesis testing was applied. A test for difference in mean (Montgomery and Runger 1999) was used to compare the output of a simulation model with data prepared, first using the traditional industrial approach (reference procedure), and then using the GDM-Tool. A significance level of $\alpha=0,05$ was used. See Publication V for more information on the test procedure. In Publication VI and in the previously unpublished case, the same comparison is performed using a combination of descriptive statistics and face validation of the data and simulation models (Sargent 2005). In this way, process experts within the research projects are used to determine whether data quality can be considered similar between the two approaches.

3 Frame of Reference

This frame of reference covers several areas necessary for increasing in input data management. Firstly, the fundamental starting point is presented by defining the terms *data* and *information* and the value-adding activities required to obtain the transformation between the two of them, i.e. the chief objective of input data management. Thereafter, the different parts of input data management (collection, processing and interfacing) are described successively. There is an emphasis on data processing, which is the central part for the research questions (RQ1 and RQ2). Existing solutions to increase efficiency and elevate the level of automation during the entire input data management process are finally reviewed.

3.1 The Difference Between Data and Information

An established distinction between *data*, *information* and by extension also *knowledge* is crucial when describing and developing the input data management process. For example, since the aim of input data management is to transform data to information for simulation models, the activities involved are heavily dependent on different types of values necessary to add in order to obtain information from data. Table 2 contains some of the common definitions of data in simulation-related literature:

Table 2: Compilation of data definitions.

Definition	Publication
"Data is a set of discrete, objective facts about events"	Davenport and Prusak (1998).
"Not yet interpreted symbols"	Van der Spek and Spijkervet (1997).
"Data consists of analog or digital signals or indications (syntax) and are used for the representation of information in the purpose of further processing."	Bernhard and Wenzel (2005) based on DIN (1995).

The corresponding definitions of information are provided in Table 3:

Table 3: Compilation of information definitions.

Definition	Authors
"A flow of meaningful messages"	Nonaka and Takeuchi (1995).
"Data with meaning"	Van der Spek and Spijkervet (1997).
"Data with relevance and purpose"	Davenport (1997).
"A message meant to change the receiver's perception"	Davenport and Prusak (1998).
"Data vested with meaning"	Choo, Detlor and Turnbull (2000).

For simulation purposes, the model is the receiver of information, which means that the viewpoint of "receiver perception" (Davenport and Prusak 1998) is less applicable here. Instead, it is more relevant to conclude that the mission of data input activities is to interpret and process data in order to add meaning, relevance and purpose. This statement conforms well to the definition of data by Davenport and Prusak (1998) and the definition of information by Davenport (1997). Knowledge, on the other hand, is more connected to the mental models, skills, proficiency, know-how, and experience of people. Such knowledge about production systems is usually added after simulation modeling and, thus, is not necessarily part of the input data management process (Figure 7). It is however important to mention that experienced simulation and production engineers can deduce partial knowledge about the studied system by just looking at the information supplied to the simulation model. The creation of such knowledge, though, is not the main objective of this research.

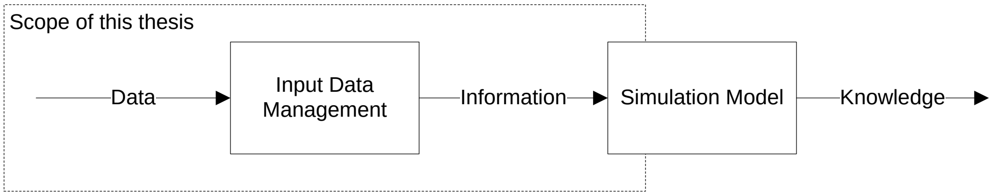

Figure 7: The scope of data, information and knowledge included in this thesis.

3.1.1 DATA REQUIREMENTS IN DES

All simulation models differ between one another because of variations in production system configurations, as well as in the model-building procedure. Therefore, simulation is often said to be an art rather than a science. This fact makes it difficult to provide general guidelines about the data requirements in DES models. For example, models aimed for conceptual analysis in early stages of factory design require less detailed data than models built for optimizing production systems in full operation (Hatami 1990). Approximations are more common in early phases in order to test as many alternatives as possible, while accuracy is more valued for decision support in on-going production. In the latter case, sample sizes around 200 raw data points or more are required to correctly mimic the dynamics of the studied system (Perrica et al. 2008). Comparing to other common analysis tools (e.g. static line-balancing and spreadsheet capacity calculations) the reader can easily understand the efforts required in input data management, but hopefully also the potential of using DES.

Moreover, it is not just the number of samples for each model parameter that contributes to the significant work load. There are also many different parameters needed to correctly represent all events in an operating production system. Processing times (e.g. assembly time and machine time) and breakdown patterns (e.g. MTBF and MTTR) are often argued to be specifically important for mimicking the dynamics (Hatami 1990, Williams 1994). As stated above, it is irrelevant to give precise guidelines but the following list contains some of the most common parameters in DES models (Hatami 1990):

- Processing times
- Set-up times
- Breakdown frequency (MTBF or Mean Jobs Between Failure)
- Repair times (MTTR or Mean Down Time)
- Product mix
- Work schedules
- Speeds on material handling equipment
- Quality-related parameters (measuring frequencies, scrap rates, etc.)
- Etc.

3.1.2 DATA CATEGORIES

Data can be divided into various categories depending on format, availability and intended area of use. Especially the two first categorizations are important for the design and selection of methods and tools during the input data management process. The most important aspect of data formats is the definition of *qualitative* and *quantitative* data. The term quantitative data usually means numbers and is, in the context of DES, often exemplified by processing times, breakdown frequencies and production schedules (Robinson 2004). A general definition of quantitative data is (Bordens and Abbot 2005):

"Data collected that are represented by numbers that can be analyzed with widely available descriptive and inferential statistics."

Qualitative data on the other hand are equivalent to "non-numeric facts and beliefs about the system" (Robinson 2004). In simulation studies, qualitative data are usually expressed in words or pictures. Examples of qualitative DES data are CAD drawings of a layout, and rules to control elements of the system (Robinson 2004). Such data are excluded from this thesis and expected to be handled by the simulation engineer during model development. It should also be mentioned that the management of qualitative data is more difficult to automate than for quantitative data. A general definition of qualitative data is (Bordens and Abbot 2005):

"Data in which the values of a variable differ in kind (quality) rather than in amount."

The second categorization is based on the degree of availability and collectability (Table 4) (Robinson and Bhatia 1995). Firstly, category A data are already available, for instance in automated data collection systems, CBS or just previously measured data stored in local sources. Of course, this type of data is very convenient, since further work is limited to data processing and validation. Secondly, category B data require additional effort because they need to be gathered during the simulation study. Finally, category C data are neither previously available nor collectable, often due to new processes or equipments in the investigated system. Estimation of category C data requires both a well-designed strategy and scrupulous care, in order to maintain model quality. A high portion of category A data is required to succeed with automated input data management (RQ2).

Table 4: Categorization of data based on availability and collectability (Robinson and Bhatia 1995).

Category A	Available
Category B	Not available but collectable
Category C	Not available and not collectable

Finally, the third categorization stems from Pidd (1996), who categorizes data with regard to their area of use. He states that they are intended for either preliminary investigation of a system (contextual data), model realization, or model validation. Naturally, input data management concerns the data necessary for model realization.

3.2 Data Collection

Data collection is a central part of input data management, even though the development of methods and tools for data collection is delimited from this research. The review below contains information about state-of-the-art procedures to serve as a starting-point for the more central step of data processing. In addition, the format and quality of raw data, generated during data collection, are highly important when selecting methods to automate the data processing step.

3.2.1 Manual Gathering

Many companies use DES for single projects instead of as an integrated part of the production engineering process (McNally and Heavey 2004, Hollocks 2001). Therefore, manual efforts are still very common in data collection to avoid the significant investment costs of the more efficient continuous collection systems (see below). Several different methods are available when gathering data points manually. The most common are (Robertson and Perera 2002; Banks, Carson, and Nelson 1996; Pegden, Shannon, and Sadowski 1995; Williams 1994):

- Video recording
- Interviews
- Collecting data using a stopwatch
- Collection of individual domain knowledge from production engineers
- Plant specification and design documents
- Recording forms
- Etc.

The choice of method depends on circumstances in each specific production system. Such circumstances can for example be: how often the events of interest occur, operators' availability, and union agreements. However, video recording is often more reliable than real-time observations and is therefore preferred if possible (Banks, Carson, and Nelson 1996). Another general recommendation is to combine different sources to increase the data credibility (Pegden, Shannon, and Sadowski 1995).

3.2.2 Continuous Data Collection

There are mainly two types of systems for continuous collection of raw data necessary for DES models: systems dependent on operator involvement, and completely automated solutions. In this case, "continuous" means that data are collected at any time and, thus, collection is not just initiated to provide data for a specific simulation project. It should also be mentioned that the continuous data collection systems are seldom implemented solely for simulation purposes; they rather originate in the needs of the maintenance organization.

Systems dependent on manual involvement typically consist of computer terminals (Kleindienst and Juricic 2007) where operators record timestamps for breakdowns, set-ups and similar events. A recent trend within this area is data and information collection using Wiki-based software solutions (Dungan and Heavey 2010). Wikis are good for gathering knowledge from many individuals within an organization, but are stronger in collecting information than the type of detailed raw data necessary for DES modeling. The bottom line is that collection systems including human involvement generally have lower investment costs than completely automated solutions, but the quality of data can be limited if people forget to record events or just incorrectly recall the timestamps before entering them at the terminal.

Automated collection systems (ACS) on the other hand are better suited to collect large amounts of data with consistent quality (Ingemansson, Ylipää, and Bolmsjö 2005). Erroneously logged data points do exist due to communication problems, but they are possible to detect and correct using clever data processing functionality (RQ2) (Zaum, Olbrich, and Barke 2008). Common data collection systems in this category are based on the timestamps registered by PLCs (Programmable Logic Controllers) or machine tool clients, for example supervised by a SCADA (Supervisory Control And Data Acquisition) system (Kühn 2006). The most common standard for communication between such devices and systems is OPC (OLE for Process Control) (OPC Foundation 2010) with MTConnect (MTConnect Institute 2010) as an alternative.

3.3 DATA PROCESSING

A crucial part of this thesis aims to describe and improve the process of transforming operational data from the shop floor (raw data) to information for simulation models. Davenport and Prusak (1998) think that this process implies adding value to the data; see for example the definitions of data and information in section 3.1. The authors specifically mention five important activities (Table 5) necessary to accomplish this increment in value. They also mention the possibility to automate parts of the transformation process by means of computerized solutions (RQ2), but state that it is difficult to completely omit human involvement. Especially the steps of contextualization and categorization are complicated to automate.

Table 5: Activities required for transforming data to information (Davenport and Prusak 1998).

Contextualization	Knowledge about what purpose the data were collected for
Categorization	Knowledge about units of analysis or key components of the data
Correction	Removal of errors from the data
Calculation	Mathematical calculations or statistical analysis of the data
Condensation	Summarizing of the data in a more concise form

In input data management for DES, the context of raw data is ideally added when the collection system or gathering procedures are designed. In these cases, the purposes of the data and the collection process are well explained. However, in numerous real-world situations, context is not added until the raw data are reviewed in connection with the establishment of a conceptual model (Van der Zee and Van der Vorst 2007). Categorization is also, in ideal cases, added before or even during actual collection of raw data. But the lack of well-structured databases, containing DES data, usually implies additional work of understanding and grouping raw data (Perera and Liyanage 2000). Quantitative DES data are often categorized with regard to specific production resources, e.g. processing times are distributed to the different machines included in the model.

Moreover, correction of data implies, for example, the removal of erroneously logged data samples due to communication problems. For instance, it is common to disregard very small time-stamps (e.g. below 30 seconds) from automatically collected breakdown logs, especially when modeling automated work stations (Alexandersson and Wirf 2001). Calculations are frequently needed in order to obtain simulation parameters such as the MTBF, which requires the subtraction of one stop time with the previous one. Finally, condensation (or input modeling (Leemis 2004)) is normally done using statistical or empirical distributions for parameters including variability (Robinson 2004). This specifically complex and time-consuming task will be further explained in sections below.

3.3.1 DATA REPRESENTATIONS

The first decision to make when supplying an input data set to a simulation model is what kind of representation to use. Statistical or empirical distributions are most common because they condense the data set to a convenient size (Robinson 2004). Empirical distributions work basically like mathematical descriptions of histograms, categorizing the samples in different intervals. Consequently, this approach requires more space for conveying information than do statistical distributions. Statistical distributions condense all data to a distribution family name and a set of parameters, usually two when dealing with continuous and univariate distributions.

Another type of representation is to keep the original data points as traces or bootstraps (Robinson 2004) (see Table 6). A trace is data samples listed in the order of collection and, thus, the simulation model reads them successively from top down. A bootstrap takes a list of empirical samples and randomly reorganizes them before presentation to the simulation model. Traces, bootstraps and empirical distributions have the advantage of keeping the data close to their original form, which guarantees a realistic model behavior. However, in this thesis statistical distributions are preferred because their ability to extend the time span of data sets to include production behavior not specifically observed during data collection (Robinson 2004). In addition, the convenient format is advantageous to include in standardized as well as customized model interfaces.

Table 6: Examples of traces and bootstraps as alternatives to statistical distributions.

Breakdown time-stamp (hh:mm)	Time between failures Trace (min)	Time between failures Bootstrap 1 (min)	Time between failures Bootstrap 2 (min)
08:41	N/A	N/A	N/A
09:33	52	99	24
09:38	5	42	215
11:17	99	5	52
14:52	215	52	42
15:16	24	24	5
15:58	42	215	99

3.3.2 INPUT MODELING

Input modeling, i.e. the condensation of raw data to suitable representations, includes several steps and relatively complex statistical calculations. This, in combination with the manual steps in data correction and calculation, is a major reason for the extensive time-consumption in input data management. There are many research contributions presenting methods to support the input modeling process but, they focus on the mathematical calculations behind, and therefore mainly aim to increase the quality of data representations; see for example Leemis (2004).

As indicated above, the input modeling procedure starts with data that are already corrected and calculated. Consequently, the data are available as a number of data points, for instance representing samples from a machine's processing time. Going from these data points to a statistical distribution is a process that includes a number of steps depending on different authors' divisions of tasks. However, all authors have the same opinion of how the process should be performed, and the most common description includes four steps (Banks, Carson, and Nelson 1996; Pegden, Shannon, and Sadowski 1995; Leemis 2004):

1. Evaluating the basic characteristics of the empirical data set.
2. Select distribution families for evaluation.
3. Select the best-fitting parameter values for all chosen distribution families.
4. Determine the "goodness-of-fit" and select the best distribution.

Depending on whether the process will be executed with or without computer calculation support, one or more distributions can be evaluated. Manual calculations are complex and, thus, only one distribution family is usually selected in step two. Using computer support, there is an opportunity to compare several distributions to each other and probably reach closer to the optimal selection.

3.3.2.1 Evaluating the Basic Characteristics

Firstly, the data set needs to be evaluated with regard to sample independence, to make sure that there is no systematic change in the data during the collection period. For example, if a learning curve effect is present for the processing time at a manual assembly station, a strictly random representation is inappropriate. Instead the learning effect has to be identified and separately modeled as a known variable. There are several methods to assess sample independence (Leemis 2004). Among the graphical solutions, the scatter-plot (Figure 8) is most widely used, which means that the data points are plotted in the order of collection. The data set can be considered independent if no tendency is identified as a function of time (cloud-shaped scatter-plot). For more detailed evaluations and increased possibilities for automation, it is also possible to test the data set mathematically, e.g. by linear regression (Leemis 2004).

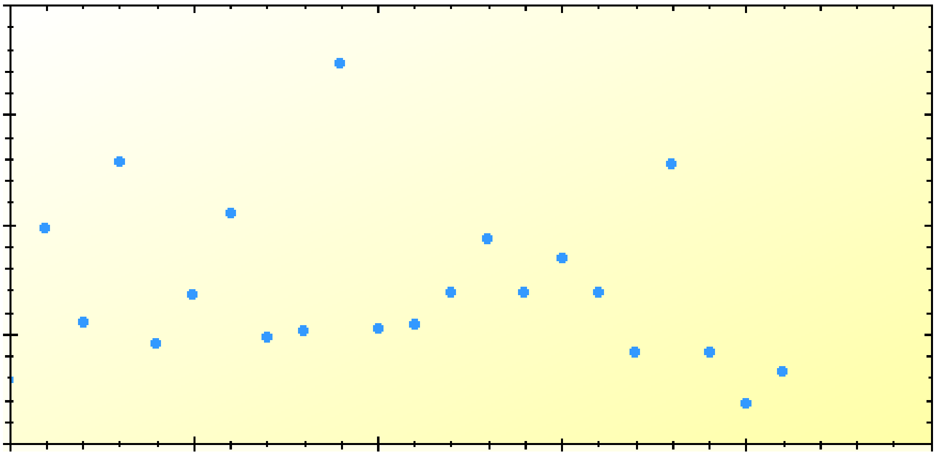

Figure 8: Example of a scatter-plot.

3.3.2.2 Distribution Families

After ensuring the data with regard to independence, a suitable distribution family should be selected for further evaluation, given that statistical representation is desired. In a manual approach, the distribution family is chosen based on the nature of sampled data. Using the calculation capacity of automated solutions, more families could be evaluated, which is advantageous since many families have similar properties. This thesis uses mostly the following four *continuous* and *univariate* distributions, which are considered sufficiently accurate for most industrial DES projects and also represented in all commercial DES software packages (Law 2007):

- *Exponential* – Usually used for the time between customer arrivals in service systems or for time between failures (TBF) of production equipment. Good assumption for TBF if only the mean value is known.
- *LogNormal* – Time to complete a task, e.g. time to repair (TTR).
- *Weibull* – Time to complete a task, e.g. processing times at a manual work station.
- *Gamma* – Time to complete a task, similar to LogNormal and Weibull.
- *Triangular* – Requires min, mean and max or min, mode and max. Good approximation for category C data (Robinson and Bahtia 1995).

3.3.2.3 Maximum Likelihood Estimation

The third step in the input modeling procedure, outlined in section 3.3.2, is to estimate the input parameters for the selected distribution families. In practice, this estimation is often done by simple calculations based on the sample mean and variance and the relationship between distribution mean and variance and its parameters. In a gamma distribution, for example:

$\mu = k\theta$

$\sigma = \text{sqrt}(k\theta^2)$

where μ is the sample mean, σ is the sample standard deviation, k is the scale parameter and θ the shape parameter.

However, solving this equation system to obtain the scale and shape values seldom identifies the best-fitting distribution, especially when the empirical data do not exactly fit a given distribution. Instead, due to desirable statistical properties, the method of maximum likelihood or maximum likelihood estimation (MLE) is preferable. The definition of the likelihood function is (Montgomery and Runger 1999):

$$L(\theta) = f(x_1;\theta) * f(x_2;\theta) * \ldots * f(x_n;\theta)$$

where: θ is the unknown parameter (or vector of parameters)

f is the probability density function

x_1, x_2, ..., x_n are the observed values

The maximum likelihood estimator of θ is the value of θ which maximizes the likelihood function $L(\theta)$.

3.3.2.4 Goodness-of-fit Tests

When an appropriate distribution family is selected and associated with a set of estimated parameters, the complete distribution must be evaluated by comparing its conformity with the empirical data. In a manual input modeling procedure, this is usually done by comparing a plot of the distribution to a histogram of the empirical data. However, such graphical comparison is problematic to quality-assure and also unsuitable for automation. Instead, there are several types of statistical goodness-of-fit tests available, and one of them is called Kolmogorov-Smirnov (KS) (Law 2007). The KS test calculates the maximum distance between the empirical and the fitted cumulative distribution functions (CDF), which is applied for automatically selecting the best-fitting distribution in the demonstrator used for answering RQ2.

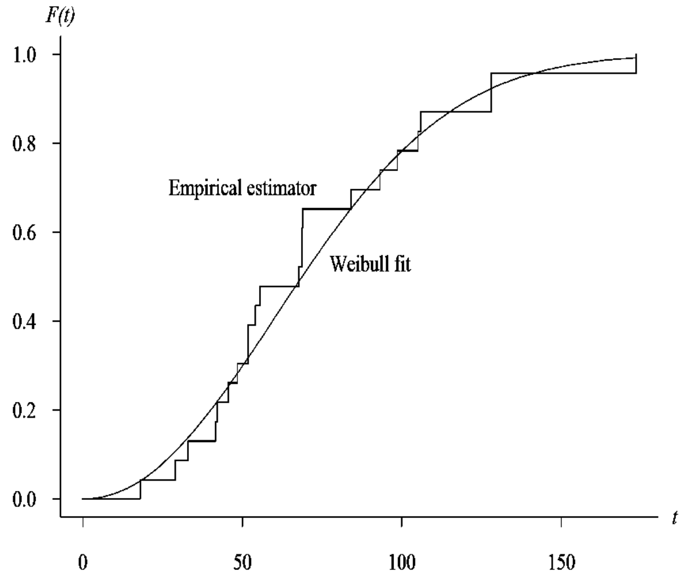

Figure 9: KS tests use the difference between the empirical and fitted distributions to evaluate their compliance.

The empirical cumulative distribution function is the likelihood of finding a sample smaller than or equal to a given value:

$$F_n(x) = P(X \leq x) = \begin{cases} \dfrac{i-1}{n} & \text{just before } x = X_i \\ \dfrac{i}{n} & \text{just after } x = X_i \end{cases}$$

where n is the number of samples and $1 \leq i \leq n$
Moreover, the fitted cumulative distribution function is:

$$\hat{F}(x) = P(X \leq x) = \int_0^x f(y)dy$$

where $f(y)$ is the density function for the selected distribution.
Further, the KS test statistic (D) is the maximum distance between the two functions, calculated at all points of X_i:

$$D = \max_{1 \leq i \leq n}(\hat{F}(X_i) - \frac{i-1}{n}, \frac{i}{n} - \hat{F}(X_i))$$

D can later be used to directly compare selected distributions to each other, simply by stating that a smaller D indicates a better fit, which is implemented as default functionality within the demonstrator (the GDM-Tool) for RQ2. However, D can also be used to statistically evaluate the fit of a specific choice by calculating the *p-value* in a hypothesis test with a null hypothesis stating that the empirical samples come from the distribution under evaluation (Montgomery and Runger 1999, Law 2007).

It is important to mention that there are also several other goodness-of-fit tests available, for example Anderson-Darling and Chi-Square tests (Law 2007). However, Chi-Square tests are troublesome when selecting the number and size of histogram intervals for comparing the empirical data with a statistical distribution. This is specifically problematic for automated input data management. Anderson-Darling tests are better in that sense but incorporate more complex calculations than Kolmogorov-Smirnov tests. They also focus more on comparing the tails of distributions rather than on the most frequent events of production resources. Therefore, Kolmogorov-Smirnov tests are considered most appropriate for automated comparisons in this thesis (RQ2).

3.3.3 Software Support in Data Processing

The reader can imagine that performing all steps described above for data processing is a complex task requiring competence, experience and time. This is a major reason for the fact that simulation practitioners often select a distribution family by guesswork and estimate the related parameters by experience; see the argumentation in the introduction to this thesis (Chapter 1). Fortunately, it is common that companies experienced in DES take advantage of available software solutions. Note, however, that manual involvement is still required to link the tools as long as the complete chain is not completely integrated as proposed in this thesis.

One common example is MS Excel®, alone or in combination with Visual Basic macros, which is used for data categorization, correction and calculations (Kumar and Nottestad 2009). Special-purpose software solutions are frequently preferred for data condensation because of their convenient interface for simulation users and their ability to provide programming code ready to use in commercial DES packages (Kumar and Nottestad 2009). ExpertFit® (Law and McComas 2003) and Stat::Fit® (Geer Mountain Software Corporation 2011) are two examples of such programs. Data condensation can also be performed using mathematics and statistics software applications with a wider area of use, e.g. Matlab®, Minitab® and SPSS®. Finally, some DES packages also provide partial support for data processing, mainly in the condensation step.

3.4 Data Interfaces and Standards

Information management problems, i.e. the supply of processed data to simulation models, affect many aspects of manufacturing operations (Gallaher, O'Connor, and Phelps 2002). They are a particular hindrance to the creation and reuse of manufacturing simulations. For example, Robertson and Perera (2002) state that current industrial procedures often include customized MS Excel® interfaces for the supply of information to simulation models. In some companies the information is even supplied by manually typing it into the model code; see further explanations in section 3.6. Both solutions have proved to be time-consuming when updating simulation models with recent data and when setting up various kinds of experimental designs. For establishing procedures including automated input data management, a standardized format for supplying the data to simulation models is highly desirable.

3.4.1 CMSD

One effort on such standardized format is titled CMSD (Core Manufacturing Simulation Data) which was developed by NIST, in collaboration with Chalmers University of Technology, other universities, and industrial partners. The outcome resulted in a standard, launched in September 2010, which follows the guidelines, policies, and procedures of the Simulation Interoperability and Standards Organization (SISO) (SISO 2010, 2011).

The CMSD specification describes a CMSD information model using the Unified Modeling Language (UML) (UML Resource Page 2009). The primary objective of this information model is to provide a data specification for efficient exchange of manufacturing life-cycle data in a simulation environment. The CMSD objective aims to: foster the development and use of simulations in manufacturing operations; facilitate data exchange between simulation and other manufacturing software applications; enable and facilitate better testing and evaluation of manufacturing software; and increase manufacturing application interoperability.

It is important to state that CMSD covers the representation of input data to simulation models and other engineering applications with similar data needs. Thus, logical relations describing model behavior and representation of output data are not covered. Such data and information have to be handled separately by the model builder, possibly using other standards or information models such as SysML (Huang, Ramamurthy, and McGinnis 2007). CMSD covers the following data categories:

- **Resource** information describes the people and equipment that perform activities.
- **Order** information specifies an external request to the manufacturing enterprise.
- **Calendar** information specifies time periods when production is and is not ongoing.
- **Skill definition** information describes the skills and proficiency levels a resource has.
- **Set-up definition** describes time to configure a resource, and to change configuration.
- **Part** information specifies materials, subcomponents, and end product.
- **Bill-of-materials** information specifies the subcomponent parts and quantities.
- **Process plan** information specifies production activities needed to make products.
- **Maintenance plan** information specifies maintenance processes for a resource.
- **Job** information specifies an internal request for production activities to take place.
- **Schedule** information specifies a time-plan for production activities.
- **Distribution** information specifies statistical distributions.
- **Layout** information specifies spatial data and relationships between resources.

Several test implementations in industrial case studies have been performed to prove that CMSD is a feasible neutral format and, thus, compatible with numerous simulation software packages (Johansson et al. 2008, Johansson et al. 2009, Kibira and Leong 2010). These studies show that CMSD is a good alternative when handling solely DES data, for example as a link between data processing applications and simulation applications. However, the standard lacks evidence for supporting data exchange between a wider range of engineering tools, e.g. to enable data exchange between manufacturing process preparation tools and simulations of material flows in production.

3.4.2 STEP AP214

Another standard which the author has been in contact with in the research project DFBB (see section 2.1) is STEP AP214 – ISO 10303-21. In contrast to CMSD, this standard does not originate from the DES application area but aims to demonstrate and extend its capabilities for dynamic simulations. This standard mainly stems from the needs of mechanical products within the automotive industry and is frequently applied for the representation of information such as geometry and kinematics (Kjellberg et al. 2009). STEP AP214 has previously been evaluated with regard to the use within production systems applications, and for the handling of process data such as production sequences and resource capabilities (Falkman et al. 2008). However, until recently, STEP AP214 has not been utilized for describing data on the detailed level as required for DES, but the DFBB project has shown initial progress in stochastic representations of breakdown patterns and processing times for DES models.

3.4.3 AUTOMATIONML

AutomationML is also a standardization effort aiming to connect different engineering applications in the design and development of production systems. Using several existing standards, AutomationML comprises information about factory topology, geometry, kinematics, and logics (sequencing, behavior and control). The main format used in AutomationML is CAEX, but COLLADA is also frequently applied for geometry and kinematics data. In addition, any other data standard may be used if the recommended formats do not suffice. For example, CMSD was used to carry DES data in a brief demonstration within the DFBB project. In such cases, a pointer to a separate document is linked via the CAEX file (AutomationML consortium 2010). Thanks to the diversity in allowed data formats, there is an extensive flexibility in using AutomationML but there is not such an established data model behind as compared to STEP AP214.

In correspondence to STEP AP214, Automation ML does not originate from the area of production flow simulation. Instead, it has the capability of storing and carrying data for a wider range of applications. However, a pre-study within the DFBB project demonstrated the possibility to carry breakdown data and processing times to a DES model built in ARENA®, but there is no demonstration of the possibility to represent the entire set of data needed in DES models. The bottom line is that both STEP AP214 and Automation ML hold the potential to serve as a link between DES and more engineering applications than for example CMSD. More research is necessary to do this, which makes CMSD more appropriate for the time being.

3.5 THE INPUT DATA MANAGEMENT PROCEDURE

The following section considers all three parts of the input data management: collection, processing and interfacing. Existing methods for increasing efficiency are reviewed and possible pitfalls are highlighted. This information is important in order to focus on the activities having highest impact on time-consumption during input data management when designing new efficient solutions as in RQ2.

3.5.1 METHODS FOR RAPID INPUT DATA MANAGEMENT

According to several previous publications (Perera and Liyanage 2000, Lehtonen and Seppälä 1997), there is an increasing need for systematic approaches and documented procedures of input data management due to the substantial number of non-experts using simulation. This is also a prerequisite for a wider dissemination of DES, especially among Small and Medium Enterprises (SME), which is one of the driving forces behind the state-of-the-art description in RQ1.

One contribution, using a systematic approach, is a methodology based on the Integrated computer-aided manufacturing DEFinition (IDEF) (Perera and Liyanage 2000). The methodology focuses mainly on reducing required time for identification of parameters to include in simulation models. After investigating the production system, a functional model is built using pre-developed IDEF constructs. The functional model can be compared to a conceptual model, which is a more common terminology; see Van der Zee and Van der Vorst (2007) for more information on conceptual modeling. Thereafter, a required entity model is generated, which can be translated into a relational database, providing the model builder with a structure to follow during data collection and for data storage. The generation of the entity model from the functional model is done using a mapping table like Table 7.

Table 7: Table mapping the relations between functional modeling elements, required input data and model entities.

Functional modeling element	Required data	Corresponding RM entity
Part	Part ID	PART
	Part description	PART
	Batch size	MACHINE_OPERATION
	Max batches	MACHINE_OPERATION
	Inter arrival time	MACHINE_OPERATION
Machine	Machine ID	MACHINE
	Machine description	MACHINE
	MTBF	MACHINE_GROUP
	MTTR	MACHINE_GROUP
	Input buffer capacity	MACHINE_GROUP
	Output buffer capacity	MACHINE_GROUP
Operator	Operator ID	OPERATOR
	Operator description	OPERATOR
	Efficiency	OPERATOR
	Skills	OPERATOR
	Learning curve effect	OPERATOR

Another approach named controllability analysis (CA) has been used to increase efficiency in problem definition and data management phases of simulation projects (Lehtonen and Seppälä 1997). CA is an iterative approach intended to focus only on relevant aspects of the problem to solve. At each aggregation level, the aspect of major relevance is focused upon and further analyzed in order to pinpoint the most important factors with regard to project objectives. This structured methodology identifies important parameters, and facilitates the data management process by minimizing collection of data irrelevant for solving the problem.

The advantage of these two methodologies is that they focus on specific and relevant problems of input data management. However, they merely address delimited steps during the data input procedure, which is insufficient when considering the entire chain of data input activities using a holistic approach. Instead, Bernhard and Wenzel (2005) propose a methodology covering more aspects of the input data management process. This publication proposes an eight-step methodology based on the knowledge and experiences from a cross-disciplinary team with background in data acquisition, statistics and visualization. The identified steps are: Goal setting, Information identification, Preparation of collection, Collection, Data recording, Data structuring, Data analysis, and Validation.

3.5.2 Crucial Activities with Regard to the Time-Consumption

All methods presented in 3.5.1 aim to reduce the impact of known difficulties in the input data management process. These difficulties considered to have highest impact on the total time-consumption are (Perera and Liyanage 2000):

1. Poor data availability
2. High-level model details
3. Difficulty in identifying available data sources
4. Complexity of the system under investigation
5. Lack of clear objective
6. Limited facilities in simulation software to organize and manipulate input data
7. Wrong problem definition

There are no other publications listing pitfalls in input data management as comprehensively as the list above, but other authors agree and point out separate objects. For example, Moon and Phatak (2005) argue that the high level of model detail requires many data parameters and samples. They also state that there is a lack of data sources containing data suitable for DES. Consequently, substantial efforts in additional data gathering are required.

3.6 Automated Input Data Management

Different approaches to input data management use various levels of automation (Robertson and Perera 2002), which is very important for the design and verification of the proposed concept for automated input data management (RQ2). Figure 10 shows four alternatives ranging from an entirely manual work procedure to a completely automated link between existing data sources and simulation models.

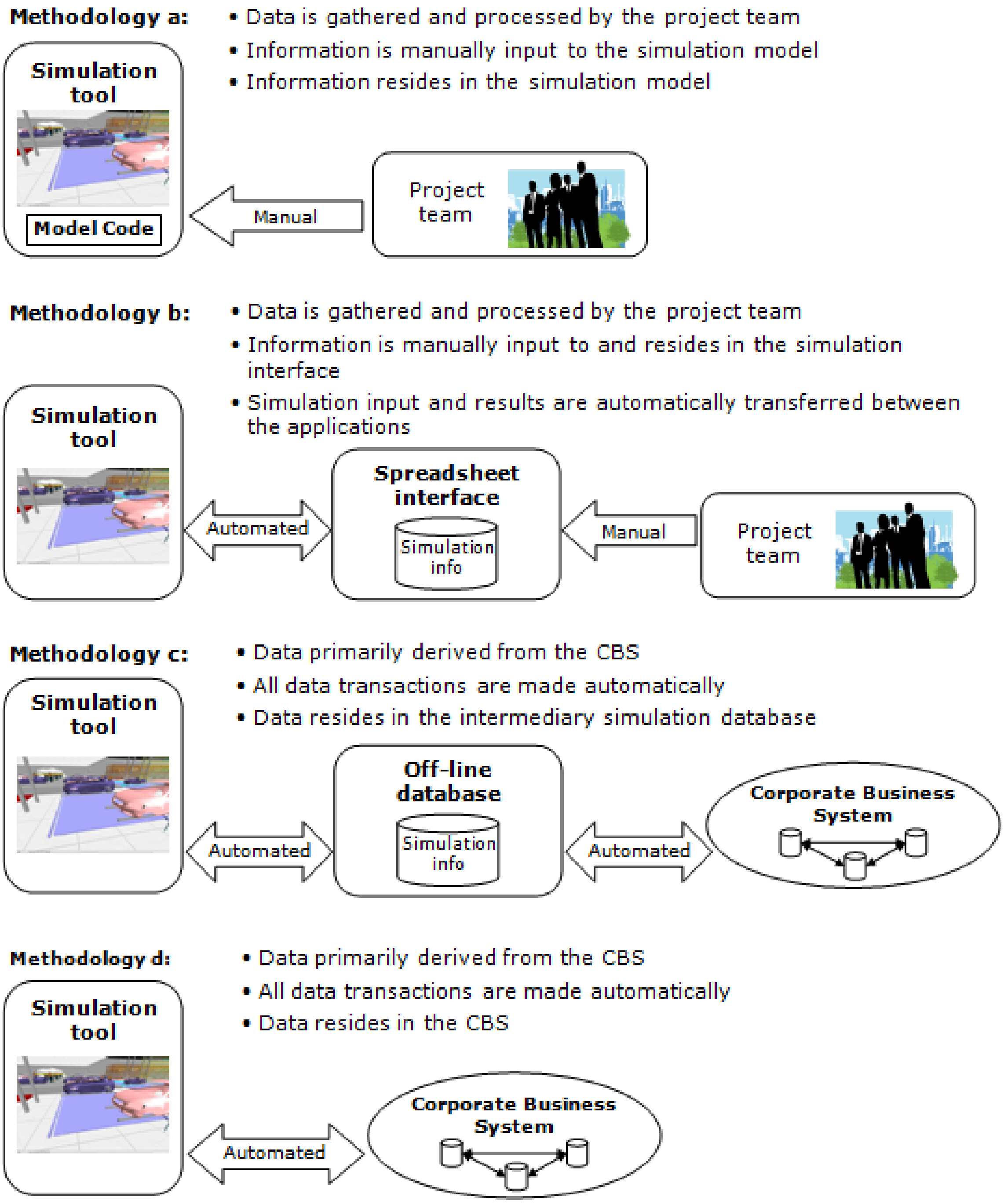

Figure 10: Four approaches to input data management using various levels of automation (Robertson and Perera 2002).

In summary, Robertson and Perera (2002) states that methodologies a and b are most present in industry but hypothesize that organizations will strive towards methodologies c and d. The main reasons for this are that it would increase the data accuracy and reliability whilst also minimizing the efforts during the entire input data management chain.

3.6.1 METHODOLOGIES A & B

Firstly, the solution with lowest level of automation (methodology a in Figure 10) implies that the model builder, or other members of the project team, manually collects the raw data needed from the appropriate data sources. This can include manual extraction of category A data as well as measurements and interviews to obtain category B and C data. After manual analysis and transformation to information, the results are manually typed into the model code where they also finally reside. This approach is easy to follow and involves a continuous validation of input data. On the other hand, it takes an extensive amount of time just to go through the process once, and due to its inflexibility, the time-consumption grows even more significant when the system changes and the data need to be updated.

The second methodology (b) is equivalent to the first one for the collection of raw data and transformation of data to information. However, there is a difference in how the results are supplied to the simulation model. In methodology b, the information is presented in an intermediary spreadsheet and automatically imported to the model. In this way the flexibility increases as it becomes easier to update, change and experiment with the information than in methodology a. Still, the disadvantage is that the collection and transformation of data to information rely on manual efforts. However, the separation of model and information, which enable use of the model among people unfamiliar with model building, makes this solution most popular in industry. Both methodologies a and b involve significant manual work since no software support is integrated in the data storage systems. Thus, data processing is often manually performed, as described in section 3.3, sometimes supported by separate data analysis software packages.

3.6.2 METHODOLOGY C

In the third alternative (c), the simulation model utilizes an off-line intermediary simulation database that is connected to the CBS and automatically retrieves and stores recent data for the simulation model. Moreover, the intermediary simulation database is connected to the simulation model and, thus, the supply of information to the model is also automated. In this way, the time-consumption for collecting and transforming data can be dramatically reduced and the flexibility for changes and updates is still present. Despite all advantages only one real-world case is known by Robertson and Perera (2002), but there are also some efforts on this methodology published as results from other research projects. One example is published by Randell and Bolmsjö (2001) where they fed a simulation model with information from an ERP-system via an intermediary simulation SQL (Structured Query Language) database.

Another concept exemplifying methodology c is named Manufacturing Data Acquisition (MDA), which incorporates both the collection and some initial processing of raw data from production resources (Aufenanger, Blecken, and Laroque 2010). These features enable combination of data samples and give a good consistency in data formats. However, the concept requires that all related technical solutions are implemented consistently, already in the collection of raw data.

3.6.3 METHODOLOGY D

Finally, methodology d implies that the transfer of simulation information from the CBS to the simulation model is fully automated. This reduces the time-consumption dramatically since human involvement is needed neither for data collection and transformation nor for importing information to the model. The major drawback is the lack of available data (Moon and Phatak 2005, Robertson and Perera 2002) already prepared for DES purposes in major CBS applications.

One example of methodology d is often referred to as the “digital factory”. This approach aims to connect data from different tools used throughout the entire product and production engineering process. There are basically two possible solutions to enable this integration:

1. Use a commercial PLM package, such as Siemens Teamcenter® (Siemens 2011) or Delmia® (Dassault Systemes 2011).
2. Connect all the individual engineering tools, and other necessary sources within the CBS, to each other using neutral formats or customized scripts.

The first solution presupposes that all engineering tools are selected from the same software vendor, which has proved difficult due to the diversity of tools with strengths in various parts of the product and production realization process (Kühn 2006). The second solution, using neutral formats, allows the use of a variety of engineering tools and is therefore more promising if the aim is to cover the entire process. However, substantial research and development is required to provide this link, without relying on expensive customized scripts (Kühn 2006).

3.7 SOFTWARE ARCHITECTURE FOR GENERIC DATA MANAGEMENT

The previous sections in this frame of reference state that the input data management process includes several activities and requires multiple data operations, e.g. for data cleansing, calculations and statistical input modeling. Major reasons are the extensive data requirements for DES (section 3.1.1), the variety in level of automation, and the diversity and evolution of data sources (section 3.2). Consequently, computer applications aimed to support the complete chain of activities have to provide a variety of features (data operations) and the possibility of customization and rapid adaption to changes in data sources.

A plug-in-based architecture (Balderud and Olofsson 2008) satisfies the demands above by adding well-delimited functionality in specific software components. The architecture enables developers to update, add or remove functionality continuously without affecting the main application or the data storage. Hence, development or configuration can be performed based on organizational context and passed on to third-party developers if desired. These qualities are highly valuable for the iterative application development in the multiple-case-study approach in this research. Plug-ins usually operate by using services provided by the main application and by communicating with a managing component keeping track of registered plug-ins and displaying them in the user interface (Balderud and Olofsson 2008); see Figure 11. The described architecture originates in the Command Pattern (Gamma et al. 1995) encapsulating requests (i.e. data operations) including user-defined inputs (not known by the software). Such input can be the specification of a data column (input 1) which will be converted to another user-defined data format (input 2). Another common application area for plug-ins is the add-ins used for customization of MS Office® functionality. For example, plug-ins are used to create shortcuts to other software packages or to extend the number of calculation operations (e.g. for data analysis in MS Excel®).

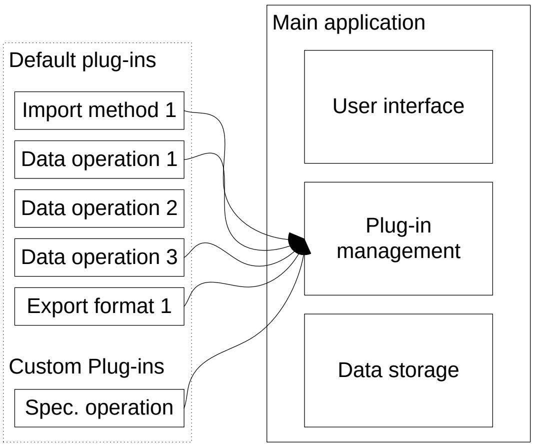

Figure 11: Each plug-in performs a user-specified request. Plug-ins can be selected to suit the desired functionality at a specific company, and the structure also allows further development of needed functionality.

4 RESULTS

This chapter contains the research results of the two interrelated parts (named "studies" in Wilkinson 1991) presented in this thesis. The first part relates to RQ1 and the second to RQ2. After the results of each part, there is an interim discussion summarizing the findings with focus on one research question at a time. The interim discussions aim to connect the publication results to each other, to literature, and to the thesis purpose and aim. They also include discussions on the methods used. A general discussion, connecting the studies to each other, is provided in the next chapter. Below is the structure of this chapter on results:

4.1 Results related to RQ1 – Publications I, II, III, and IV.
4.2 Interim discussion focusing on RQ1.
4.3 Results related to RQ2 – Publications V, VI, and an additional unpublished case study.
4.4 Interim discussion focusing on RQ2.

4.1 INDUSTRIAL STATE-OF-THE-ART – RQ1

This first part of the results chapter aims to answer RQ1: What is the industrial state-of-the-art in the input data management process? The purpose is to describe the necessary activities, how they are executed in industry, and the level of automation used to support the process. Publication I maps the necessary activities and describes best-practice work procedures for each activity. Publication II identifies the activities having highest impact on the total time-consumption and, therefore, should be considered extra interesting to automate. Publication III presents current industrial approaches to automated input data management and investigates their dissemination. The last publication in part 1 (Publication IV) contributes to the description of future requirements on input data management, derived from the need for new input parameters when DES analyses are extended to include sustainability aspects.

4.1.1 PUBLICATION I

- A Methodology for Input Data Management in Discrete Event Simulation Projects

Knowing and mastering the important activities in input data management is crucial for enabling efficient simulation studies. This applies in procedures with significant manual involvement as well as in the requirement specification of automated support systems. Automation is of course a tempting and potent solution for reduced time-consumption, but all organizations are not ready to adopt such systems, e.g. due to inferior availability of category A data (see section 3.5.2). An additional aspect is that the number of non-specialists working with DES increases and, thus, easier and more comprehensive methodologies, such as practical guidelines, are highly desired (section 3.5.1).

Objective and Contribution to RQ1

The objective of this publication is to map current practice in input data management and thereby compile a systematic best-practice methodology. Such methodology holds potential to enable more efficient and accurate input data management for simulation projects, for all levels of automation, except completely automated processes. The proposed methodology includes and describes all activities detected during an investigation of industrial DES projects, which is an important contribution to the answer of RQ1. In addition, the publication provides state-of-the-art guidelines for the tasks executed in each activity.

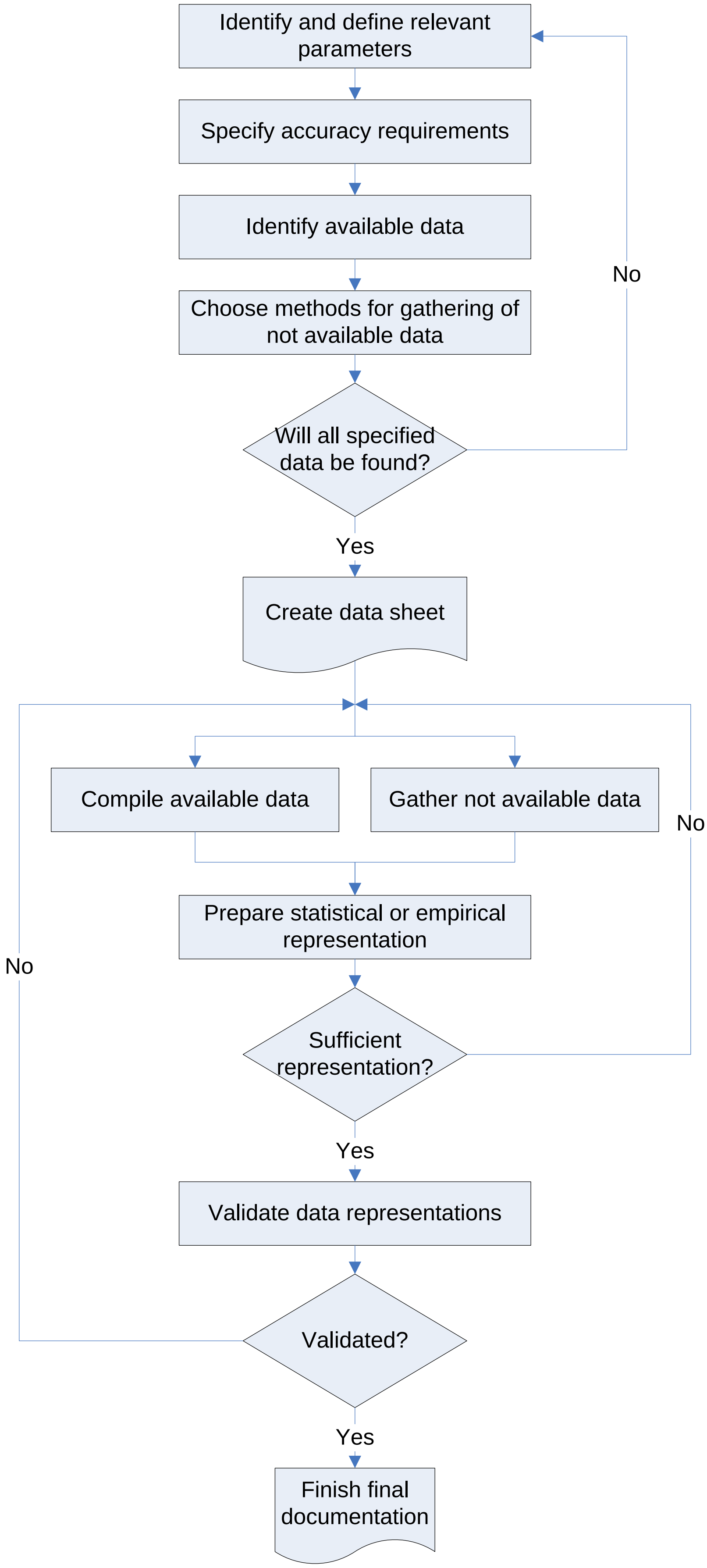

Figure 12: Activities in input data management, structured as a best-practice methodology.

Study Description
The mapping of important data input activities was performed by evaluating 15 completed industrial simulation projects. The evaluation embraced pure industrial projects as well as projects including parties from both industry and academia. All plants simulated in the projects were located in Scandinavia, mainly in Sweden. Moreover, to obtain as general results as possible, the involved companies were selected to represent a variety of contextual factors such as organizational size, line of business, and previous experience in DES. Semi-structured interviews (section 2.3.1) were used to specify common input data activities, to collect information about their internal work procedures, and to identify main problems resulting in extensive time-consumption (topics are provided in the Appendix).

Results and Conclusions
In current industrial input data management for DES, practitioners perform thirteen distinct activities; see Figure 12. Further descriptions and guidelines supporting the tasks in each activity are provided in the appended publication. Represented as such a best-practice methodology, the activities fit well into the frequently cited works of Banks, Carson, and Nelson (1996), Law (2007), and Pegden, Shannon, and Sadowski (1995), all providing comprehensive methodologies for DES projects. In these methodologies, the input data management part represents a smaller portion of an entire project. That smaller portion is more thoroughly described in this publication. The authors suppose that the profit of using the methodology is most significant in organizations with limited experience of DES. The argument is that experienced organizations and simulation engineers continuously discover and document efficient working procedures in an iterative manner. However, there are no studies aimed to quantify the methodology's impact, so the main contribution thus far is the state-of-the-art description provided as a part of RQ1.

4.1.2 PUBLICATION II
- Mapping of Time-Consumption During Input Data Management Activities

Practitioners and researchers jointly argue that high-quality input data is crucial in simulation studies. This, in combination with too few and insufficient methods and tools, contributes to the fact that input data management is one of the most time-consuming parts of simulation projects. Previous measurements and estimations claim that input data management consumes 10-40% of the total time in DES projects (see section 1.1). However, few studies have closely investigated the different data input activities for finding the primary causes and quantifying their individual time-consumption.

Objective and Contribution to RQ1
This article presents an empirical mapping of current industrial work procedures for input data management. It continues the work of Publication I by assessing the time-consumption for each activity in the input data management process. The main objective is to identify the input data management activities having highest impact on the total time-consumption as a part of the answer to RQ1. By extension, this study may serve as a guideline for the design of IT support systems and other methodologies in future research on efficient data management. Additionally, the study summarizes the most common reasons for extensive time-consumption in input data management, as supplementary support in the development of such tools and methodologies.

Study Description

The empirical mapping was performed by evaluating 15 completed industrial simulation projects, the same as in Publication I. Semi-structured (2.3.1) interviews enabled specification of common input data activities and identification of common problems resulting in extensive time-consumption. In addition, face-to-face questionnaires (section 2.3.2) were used to assess the time-consumption for each activity and to quantify the amount of available data (questions are provided in the Appendix). In order to compare the activities to each other, the time-consumption for each activity was related to the time-consumption of the entire input data management process in the specific project. Using such relative comparison compensates for the difference in scope between the projects.

Results and Conclusions

Figure 13 shows that the three activities having highest impact on the total time-consumption are: collection of raw data, mapping of available data, and data analysis and preparation. From this result, it is argued that highly time-consuming activities should recieve specific attention when increasing efficiency in the complete input data management process. A detailed summary of the time-consumption for each activity in all 15 projects is available in Table 8.

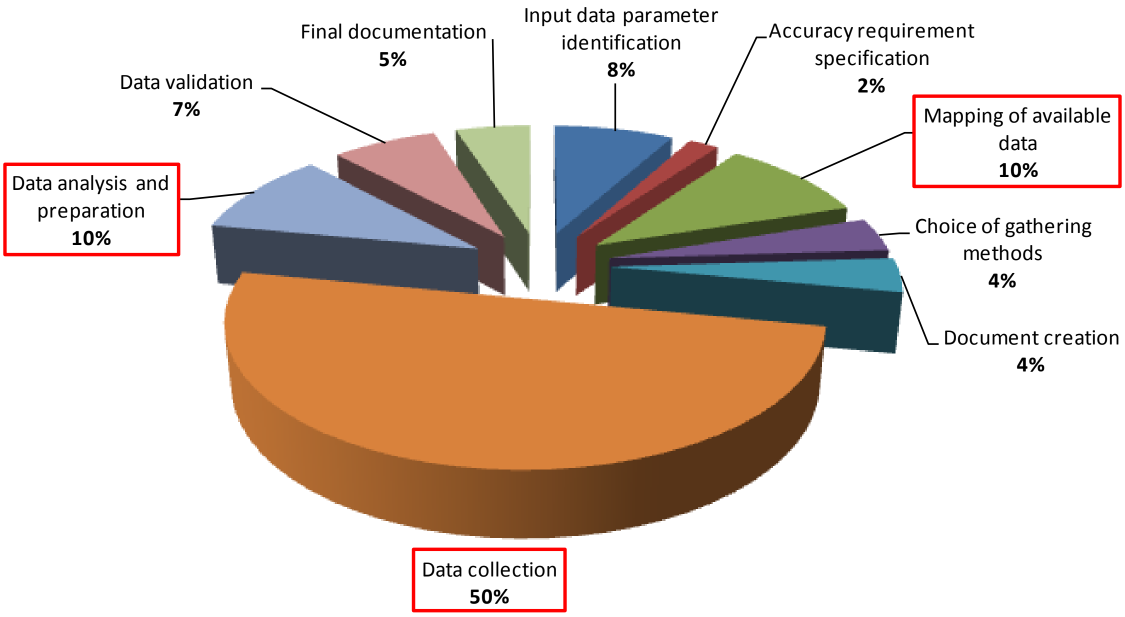

Figure 13: The time-consumption of each individual data input activity.

Furthermore, the study identifies that the two major reasons for problems which occurred during input data management were: substantial need for manual raw data gathering due to inferior data availability, and complex designs of computerized data sources. The latter slows down the identification of available data and indicates that sources containing raw data are generally not designed for simulation purposes. By extension, this is also an explanation to the limited availability of raw data. In addition, the article summarizes the time-consumption of the input data management process in all 15 projects and concludes that it accounts for on average 31% of the total project time.

Table 8: Data table displaying the time-consumption during data input activities in all projects evaluated in Publication II.

Project	Input data parameter identification	Accuracy requirement specification	Mapping of available data	Choice of gathering methods	Document creation	Data collection	Data analysis and preparation	Data validation	Final documentation
# 1	12%	2%	2%	2%	0%	60%	12%	4%	5%
# 2	3%	0%	7%	7%	1%	51%	7%	22%	0%
# 3	5%	2%	12%	1%	2%	63%	1%	6%	6%
# 4	5%	2%	4%	5%	5%	61%	7%	5%	5%
# 5	3%	3%	12%	1%	6%	57%	12%	0%	6%
# 6	3%	0%	15%	3%	5%	58%	8%	5%	5%
# 7	1%	4%	2%	2%	1%	40%	25%	12%	12%
# 8	9%	0%	9%	4%	9%	52%	9%	4%	4%
# 9	5%	0%	9%	5%	5%	45%	23%	5%	5%
# 10	4%	4%	9%	4%	7%	50%	7%	9%	7%
# 11	33%	11%	11%	7%	0%	24%	2%	11%	2%
# 12	14%	7%	14%	11%	7%	21%	14%	4%	7%
# 13	5%	0%	10%	10%	5%	50%	10%	5%	5%
# 14	5%	3%	13%	5%	8%	56%	3%	5%	3%
# 15	10%	0%	21%	0%	0%	62%	8%	0%	0%
Average	**8%**	**2%**	**10%**	**4%**	**4%**	**50%**	**10%**	**7%**	**5%**

Project Timeline

4.1.3 PUBLICATION III

- Input Data Management for Simulation - Industrial Practice and Future Trends.

Automation of data input activities is naturally one of the solutions to the extensive time-consumption described in Publication II. Reduction of the human involvement can be achieved by addressing different parts of the input data management process, i.e. the raw data collection, the data processing and the supply of information to the simulation model. In the most automated approach, the simulation model is totally integrated to all necessary data sources, either to major business systems (e.g. ERP) or to other tools used during the product and production engineering process (e.g. a PLM environment); see section 0. However, issues like interoperability problems and limited data availability make additional, less automated, solutions necessary. A compilation of different approaches is provided by Robertson and Perera (2002), which is used as a starting point for this publication.

Objective and Contribution to RQ1

The aim of this paper is to map the current industrial practice in input data management with regard to the level of automation. It is intended to be an update of a previous publication (Robertson and Perera 2002) and, therefore, uses the defined approaches therein. This article forms a part of the industrial state-of-the-art description (RQ1) by completing the findings about data input activities from Publications I and II with information about applied support systems.

Study Description

This publication presents the results of a survey performed during the WSC 2010. WSC is one of the world's major forums for DES specialists representing industry, academia and government. A questionnaire was distributed to all participants and the industrial representatives were asked to answer 12 questions (see Appendix) about the simulation procedures at their specific companies, mainly focused on input data management. Researchers with close connection to industry (a recent case study) were also asked to complete the form with information obtained at the case study company. Reminders were sent out by e-mail containing a link to a web-questionnaire (exact copy of the original form). Answers from 86 companies were collected, including different business areas such as: manufacturing (35 responses), logistics, health care and military applications. Data were analyzed using descriptive statistics to show how many companies used the different approaches to automated input data management (section 3.6).

Results and Conclusions

The questionnaire responses show that DES is used on a regular basis, with frequent reuse of models, by 65% of the participating companies. 8% have even integrated DES in their business process as a mandatory tool in major development projects. These are fairly decent figures compared to previous literature. However, the reader should keep in mind that few SME visit WSC; such enterprises are therefore most likely under-represented due to study delimitations.

Taking a closer look at the input data management procedure, there is an obvious lack of structured approaches as well as continuous collection of raw data. Among the manufacturing companies, 63% do not even use support of checklists, templates or documented guidelines to increase efficiency of the data input activities. Furthermore, many companies have computerized systems as their main source of data, but Figure 14 also shows that the diversity of sources to compile all necessary data is extensive. All these factors indicate the need of structured approaches to input data management, and also imply that suggested solutions for automated input data management should support import of data from several sources.

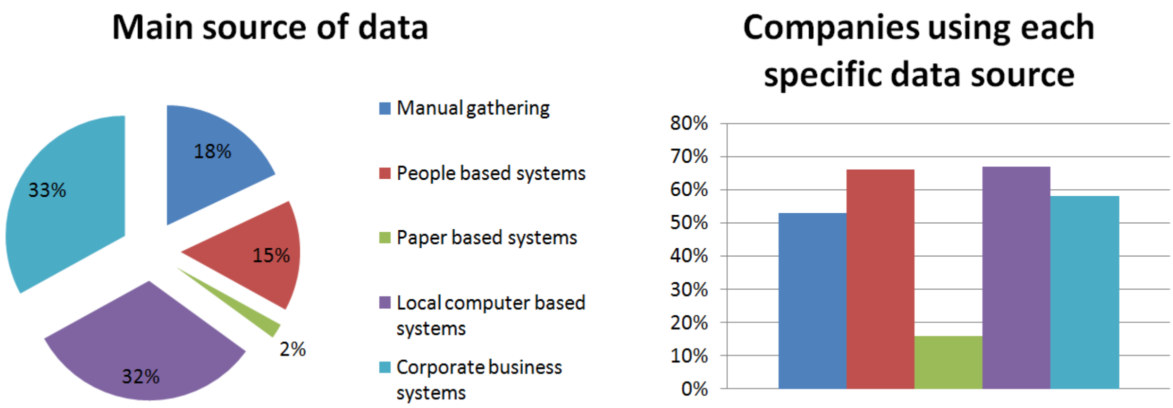

Figure 14: The use of different data sources in manufacturing industry.

Regarding the use of automated solutions during the entire input data management procedure, approximately 20% of all companies use one of the completely automated approaches (17% methodology c and 3% methodology d in the manufacturing industry); see section 3.6. The most common approach is still methodology b, using a spreadsheet interface automatically connected to the DES model but relying on significant manual work during data collection and processing; see Figure 15. The same figure also shows that many companies desire and foresee an increased use of more automated solutions in ten years.

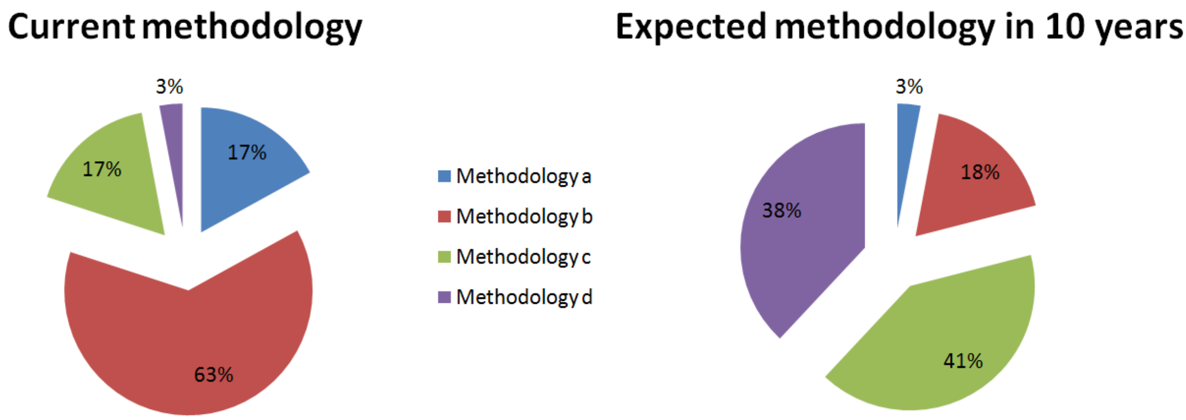

Figure 15: Current methodology and level of automation in input data management among manufacturing companies.

The results of Publication III clearly show progress in the use of automated solutions for input data management in DES during the last decade. In the original study (Robertson and Perera 2002), just a few companies reported implementations of the more automated methodologies c and d, including database connections. Today, around 20% of the companies have automated connections to their simulation data sources, most of them using an intermediary database. The rest of the companies rely mostly on manual input data management, and still strive to reach a higher level of automation. However, the lack of simulation data in their CBS and insufficient support in automated data processing are reported as two major hindrances. More research and development in this area would probably increase the use of simulation on a regular basis.

4.1.4 PUBLICATION IV

- Data Requirements and Representation for Simulation of Energy Consumption in Production Systems.

Sustainability thinking is nowadays a natural part of production systems development, and there are numerous research contributions addressing detailed technological applications as well as improvements on a system-level. However, there is often higher potential in the latter, for example by eliminating non-value-added activities in order to reduce energy consumption (Cao, Chou, and Cheng 2009). In other words, waiting times should be carefully analyzed in order to minimize the effects of balancing and system losses. DES is a powerful tool for such analyses and its application area has therefore recently been extended from focus on economic aspects to include ecological sustainability (Chapter 1). This shift introduces requirements on new input parameters in simulation models and, thus, is likely to drive changes in the input data management process. Yet few previous research contributions (Solding, Petku, and Mardan 2009; Solding, Thollander, and Moore 2009) have closely investigated the new requirements when handling data parameters such as electrical power in DES models.

Objective and Contribution to RQ1

The aim of this publication is to specify how electrical power should be represented as an input parameter in DES models. As stated above, new input parameters will probably introduce additional activities in the input data management process or, at least, changes in the conventional work procedure. This publication is therefore interesting for the description of input data management activities in RQ1, especially when looking into the future of DES modeling. However, more research is required in this field and the reason for including the publication in this thesis, at this early stage, is to stimulate further research and to prepare support tools and input data management procedure to these novel requirements.

Study Description

This case study is performed at an automotive company and includes measurements of the electrical power utilization for five multi-operational tooling machines. All five machines perform milling operations in a production line for engine components. The power utilization is measured with a frequency of 1 Hz on the incoming three-phase connection. Thus, all functions of the machines were measured as one unit, including contributions from major machine systems (e.g. machine spindle) as well as from peripheral functions such as lights, control system and pumps. In total, more than 230 000 samples were collected.

After collecting the data, all samples were assigned to one of the following four machine states: busy, idle, down, and stand-by. Descriptive statistics were calculated for all four machine states, both for the individual samples within a cycle and for the calculated average power utilization of cycles; see Figure 16. The descriptive statistics reported below are the average values and the standard deviations as a measurement of variability. Variability is a key factor for collection and representation of DES parameters, since highly variable parameters have to be stochastically represented using statistical distributions or similar approaches. Consequently, such parameters require more data samples, compared to those represented only by mean values.

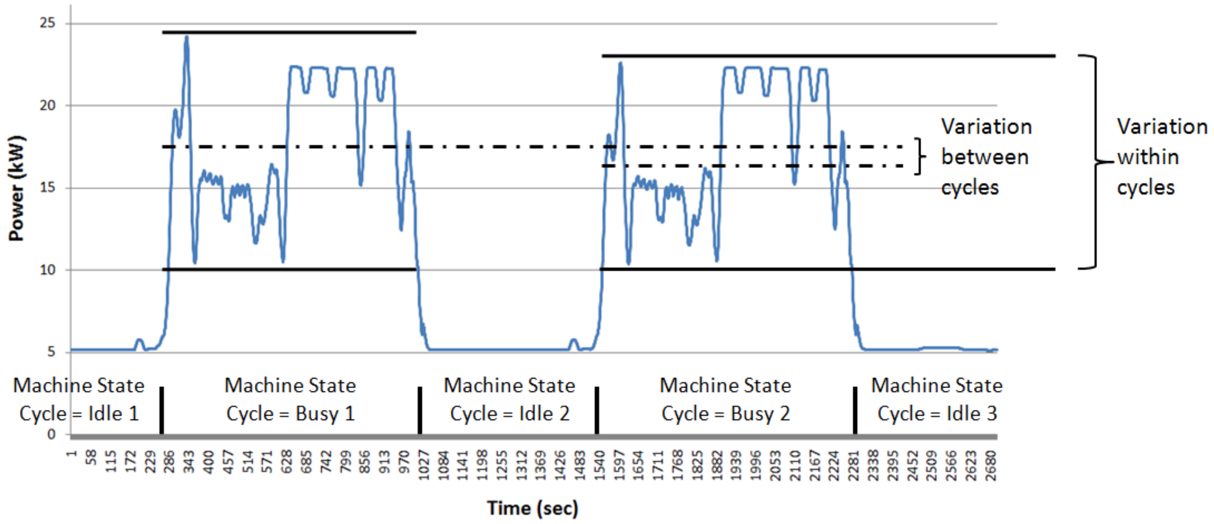

Figure 16: Illustration of machine state cycles and their corresponding variations in power levels.

Results and Conclusions

The results show that the standard deviation, for the average power utilization in busy cycles, ranges from 1 to 2% of the power levels for the five machines included in this study. For idle and down cycles, the same values are 9% and 1% respectively (average standard deviations for all five machines); see Table 9 and Table 10. Looking further into the busy state, the variability between product cycles is even smaller when also considering the product variant a known factor (1.5%). These numbers show that the variability between product cycles is limited and unnecessary to include in conventional DES modeling, even though DES models are dynamic. The reason is that the variability in processing time has considerably higher impact on the final energy consumption. This means that the collection of raw data describing power utilization can be limited to a few samples, just enough to calculate a credible mean value. However, if DES models are used or will be used to analyze the environmental footprint, using the individual product as the unit of analysis, electrical power probably needs to be considered a stochastic variable. This must be evaluated in future research.

Table 9: Average power utilization per machine, distributed between the different machine state cycles.

Machine	PiB (kW)	PiB (V1)	PiB (V2)	PiI (kW)	PiD (kW)	PiS (kW)
OP 20	10.0	9.9	10.0	6.2	5.4	1.0
OP 30_1	17.0	16.9	17.0	5.6	5.5	1.2
OP 30_2	17.8	17.6	18.0	5.4	5.4	1.1
OP 40_1	10.5	n/a	n/a	5.7	5.5	n/a
OP 40_2	10.9	10.7	11.1	6.2	5.8	n/a
Average	**13.2**	**13.8**	**14.0**	**5.8**	**5.5**	**1.1**

Table 10: Standard deviations for the average power utilization between individual machine state cycles.

Machine	SbBc (kW)	SbBc (V1)	SbBc (V2)	SbIc (kW)	SbDc (kW)	SbSc (kW)
OP 20	0.15	0.18	0.11	0.68	0.07	n/a
OP 30_1	0.15	0.21	0.08	0.35	n/a	n/a
OP 30_2	0.22	0.15	0.10	0.19	*0.02*	n/a
OP 40_1	0.11	n/a	n/a	0.37	n/a	n/a
OP 40_2	0.21	0.12	0.06	1.02	*0.00*	n/a
Average	**0.17**	**0.17**	**0.09**	**0.52**	0.03	**n/a**

In Table 9, PiB = Power in Busy; PiB (V1 & V2) = Power in Busy given one of the two product variants; PiI = Power in Idle; PiD = Power in Down; PiS = Power in Stand-by. All numbers represent the average power utilization. In Table 10, SbBc = Standard Deviation (StDev) between Busy cycles; SbBC (V1 & V2) = StDev between Busy Cycles given one of the two product variants; SbIc = StDev between Idle cycles; SbDc = StDev between Down cycles; SbSc = StDev between Stand-by cycles.

An additional result obtained in this study is that 33% of the total energy consumption for the five machines stems from non-production time. In other words, a substantial part of the energy cost, and of the related environmental impact, stems from non-value-added time explained by balancing and system losses. Note that this result is specific to the particular production system and the time of measurements for this study. However, it is a strong indication that improvement of production flows is a very important area.

4.2 Interim Discussion – RQ1

This part of the thesis investigates the industrial state-of-the-art in input data management and the current level of automation applied in this process. Publication I identifies thirteen activities performed by industrial simulation practitioners to transform raw data from the shop floor to information for simulation models. Best-practice descriptions are included in the same publication. Further, Publication II identifies the three most time-consuming data input activities, which are important to focus on in order to reduce the time-consumption during input data management. Publication III shows that the level of automation is still limited in companies worldwide and that the significant manual involvement results in extensive time-consumption for keeping models up-to-date. In fact, the input data management procedure constitutes as much as 31% of the total time in DES projects, which is still comparable to research results reported almost 20 years ago (Trybula 1994). Thus, improved support systems, both computerized and manual, are important in order to increase the applicability of DES.

Additionally, Publication IV initiates a more thorough investigation of how to represent environmental parameters, such as electrical power, in DES models. This is a crucial research area for the extended capabilities of DES towards sustainability analyses. At present, it seems that electrical power does not need stochastic representation for common-purpose DES models, but further studies are required. Future studies should include other manufacturing processes, environmental parameters, and model purposes.

4.2.1 Data Input Activities

The 13-step state-of-the-art description provided in this thesis can be applied as a valuable methodology for increased rapidity and precision in input data management, especially for the increased number of non-specialists working with DES (Perera and Liyanage 2000, Lehtonen and Seppälä 1997). At present, the major part of literature on DES data covers separate elements of the input data management procedure. For example, publications aiming to improve the identification of relevant parameters (e.g. Perera and Liyanage (2000), Lehtonen and Seppälä (1997)) only address issues directly related to 8% of the time-consumption of the complete input data management process according to findings in Publication II. The description provided here is one of the few (Bernhard Wenzel (2005) is another) contributions addressing the entire chain of data input activities.

In addition to the described state-of-the-art procedure, some common shortages were identified during the interviews and in the questionnaire responses. For example, little time and effort are spent on defining the accuracy requirements on the included parameters. This might result in too few samples collected for important parameters (quality issue) as well as too many samples for less crucial production resources (time-consumption issue). It is also obvious that simulation engineers often skip a separate data validation, which is likely to lead to late additional iterations to secure an acceptable data quality. A final shortage is the lack of sufficient tools for data processing and analysis (Perera and Liyanage 2000). There are special-purpose software solutions for statistical analysis of data (for example ExpertFit® (Law and McComas 2003) or Stat::Fit® (Geer Mountain Software Corporation 2011)) but they have limited capabilities to categorize, correct and calculate the data. Additionally, these statistics applications require some repetitive manual work in order to feed the application with raw data and to supply the results to the simulation software. This is done for every analysis, which of course adds to the time-consumption of input data management.

The time-consumption analysis provided in Publication II presents results in line with previous research regarding the problems related to input data management. Data collection, identification of available data, and data analysis and preparation are identified as the most problematic and time-consuming data input activities. They conform well to three of the major pitfalls presented in literature (Perera and Liyanage 2000): poor data availability, difficulties in identifying available data sources, and limited facilities to organize and manipulate input data. However, this study adds new knowledge about the time-consumption related to separate data input activities and facilitates for quantification of the benefits expected from possible solutions. Such solutions should focus on support tools for data processing, and systems enabling reduction of the manual work during the entire input data management chain.

It is also important to develop the computerized sources to meet the extensive raw data requirements of detailed production analysis tools, e.g. DES. At present, such systems are mainly designed for the logistic, financing and maintenance organizations (Moon and Phatak 2005). Additionally, more established data models defining simulation parameters (e.g. what is a processing time, MTBF and MTTR?) would facilitate data identification, collection, and processing as well as the interoperability between data systems and analysis tools.

4.2.2 Level of Automation in Input Data Management

This thesis shows (Publication III) that there has been progress in the use of automated solutions to input data management during the last decade. Going from single pilot implementations around year 2000 (Robertson and Perera 2002), there is now one out of five companies using automated connections between computerized data sources and their simulation models. The most common solution among these companies includes an intermediary off-line database allowing data manipulation required to create what-if scenarios. The solution is also convenient for security reasons compared to a direct link to the CBS. However, around 80% of the companies still rely on extensive manual work in data collection and processing, and the link between the processed simulation information and the model typically consists of an MS Excel® spreadsheet. This finding indicates that simulation projects are still often performed on a consultancy basis with limited use as a desk-top resource for production engineers.

A very interesting additional finding from Publication III is the increasing need for data automatically extracted from external databases. In manufacturing applications, this is probably due to the fact that environmental analyses have been increasingly combined with DES. Such studies often include LCA data collected from external databases such as the European reference Life Cycle Database (ELCD) (Institute for Environment and Sustainability 2010) or EcoInvent (Swiss Centre for Life Cycle Inventories 2011). Publication IV is a first step towards a correct representation of environmental parameters in DES models. The results, showing that deterministic representations of electrical power are enough, indicate that automated connections to databases containing such environmental data seem relevant and important. The author is also involved in a project named EcoProIT (EcoProIT research project 2011) working with automated connections to LCA databases.

4.2.3 METHODOLOGICAL DISCUSSION

The most obvious issue of a study, including interview and questionnaires, is the sample sizes and the number of respondents. In the study including 15 DES projects (Publications I and II), the number of samples is considered sufficient since the data collection was performed using face-to-face communication enabling in-depth understanding and attendant questions if necessary. In Publication III, the questionnaire was distributed to around 700 DES researchers and practitioners and 86 responses were collected. The questions were aimed to collect information about industrial business procedures, so it is likely that many people declining to submit an answer were researchers without close connections to industry.

As a comment to the reference on Grounded Theory (Glaser and Strauss 1967) for interview analysis in Publications I and II, it should be clarified that the author has previous industrial experience in DES and input data management. This is important to declare when using an inductive approach since the experience, most probably, affects the coding of empirical data. Thus, the work procedure corresponds more to the interpretation of Grounded Theory later published by Strauss. He advocates a more pragmatic use of previous theoretical and practical knowledge than does Glaser. Data coding, data analysis and other knowledge should not be seen as distinct activities.

Further, in Publication II, it is important to state that the intention is to identify time-consuming data input activities and to compare activities to each other. It is for instance inappropriate to use separate assessments and infer that exact time measurements are performed. One reason is that the times are not measured in real-time, but based on the team members' perception and memory after the project is completed. Another reason why the assessments should be used for comparison rather than as absolute numbers is that the definitions of activities were initially somewhat vaguely described. However, the respondents did not find this problematic and no questions about activity delimitations arose. The fact that some activities are renamed between Publications I and II is also because the methodology and activity definitions evolved during the research process. A specific example is the activity called "prepare statistical or empirical representation" in Publication I and "data analysis and preparation" in Publication II.

4.2.4 CONNECTION TO RQ2

Despite solid state-of-the-art descriptions and systematic guidelines, it is difficult to drastically reduce the time-consumption during input data management. This is mainly due to the manual involvement required to collect data and to carry data and information between the sources and different processing applications (e.g. MS Excel® and distribution-fitting software). Therefore, there is a significant potential in automating the data input activities identified for answering RQ1, especially the most time-consuming activities.

Empirical data related to RQ1 show that data collection is the most time-consuming activity. Therefore it might seem natural to proceed with finding technical solutions supporting the collection of raw data in production systems. However, the root cause is that companies have not adopted existing technology for ACS, rather than a complete lack of such equipment including sensors and databases (Ingemansson, Ylipää, and Bolmsjö 2005). This thesis will therefore, from now on, focus on the processing of available raw data to information and the supply of information using standardized interfaces for simulation models.

This first part (RQ1) also shows that the data input activities consume 31% of the total time in an average DES project, which is a significant reason for the relatively low dissemination of DES in manufacturing industry. Thus, user-friendly best-practice descriptions and extended automated solutions are necessary in order to reduce systems and balancing losses, and by extension to increase efficiency in production systems. Some companies with limited access to category A data will prefer the systematic approach and other companies will implement automated solutions such as methodology c or d in Robertson and Perera (2002), hopefully influenced by the design specifications presented in the next part of this thesis (RQ2).

4.3 AUTOMATED INPUT DATA MANAGEMENT – RQ2

From the discussion in section 4.2, it is obvious that manual involvement during data input activities results in extensive time-consumption. In turn, this fact hinders the dissemination of DES as a desk-top resource for production engineers trying to increase the equipment efficiency in production systems. Part 2 (RQ2) proposes, tests and validates an approach to input data management, which increases the level of automation significantly in comparison to the current industrial practice. Publication V outlines the necessary functionalities of such an approach, develops a demonstrator, and presents a first test case performed in the automotive industry. Publication VI presents a test case of the same approach and demonstrator performed in the aerospace industry and adds a few necessary functionalities. This chapter also includes an additional test case in the automotive industry, which is not yet published nor appended as a publication. The additional test case has the same purpose as Publication VI.

4.3.1 PUBLICATION V

- Automated Input Data Management: Evaluation of a Concept for Reduced Time-Consumption in Discrete Event Simulation

To increase the level of automation in input data management, previous research contributions have primarily suggested automated connections between simulation models and data sources within the CBS, such as ERP and MRP (Material Requirement Planning) systems. The problem is, though, that these major systems do not often include all required data for DES (section 0). Hence, there is a need for a solution that is also able to extract raw data from other sources of category A data. Examples of such sources are major CBS applications, legacy systems, and person-based spreadsheet solutions.

Objective and Contribution to RQ2

The aim of this paper is to develop a concept, and an associated demonstrator, of automated input data management in simulation of material flows in production. An additional objective is to perform a first comparison of the time-consumption and data quality to a traditional industrial approach to input data management (a reference procedure). This paper addresses all aspects of RQ2 by identifying necessary functionalities of the concept and providing an initial evaluation of the time-consumption in the automotive industry.

Note that the demonstrator, called the GDM-Tool, is not intended to reach the requirements of a commercial software solution. It is developed as a demonstrator for the proposed concept with the purpose of facilitating validation and presentation.

Study Description

The research approach in this paper is to develop a software solution based on design criteria identified using an actor's approach during a case study in a Swedish automotive company. These design criteria are for example: the type of data structures to import, required data format to support and conform, and necessary data operations for converting the data to simulation information. All information was collected in project meetings, workshops and informal meetings during participation in the development of a simulation model at the company. The project team included process experts and simulation engineers from the company and from NIST in the USA, together with researchers from Swedish universities and institutes.

After developing the demonstrator, the solution was tested as a first step towards validation. The test includes comparison of the time-consumption and data quality to a traditional industrial approach to input data management. This reference procedure consists of: manual raw data extraction; categorization, correction, and calculations using MS Excel®; and condensation using a commercial distribution-fitting software solution. The data quality was validated using hypothesis testing, with a level of significance equal to 95%, on the output results of the simulation model (section 2.5).

Both the design and the testing of the software solution are delimited to comprise the data sources and the production process in the actual case study. The production line modeled in this case study consists of semi-automated assembly stations, and the simulation parameters included are: processing times, MTBF and MTTR.

Results and Conclusions

Figure 17 illustrates the proposed concept of automated input data management, including three major functions: data extraction, data processing (conversion) and output preparation. A key feature is the ability to extract data from several sources with different internal structures. When all raw data are imported, a series of operations is required to convert the data from a crude form into relevant simulation input. Such operations typically provide functionalities for conformation of data types, data filtering, calculations and condensation. Finally, to enable efficient data sharing, the final information is presented in CMSD format. However, other output options are also supported to avoid hindrance to the application of customized solutions.

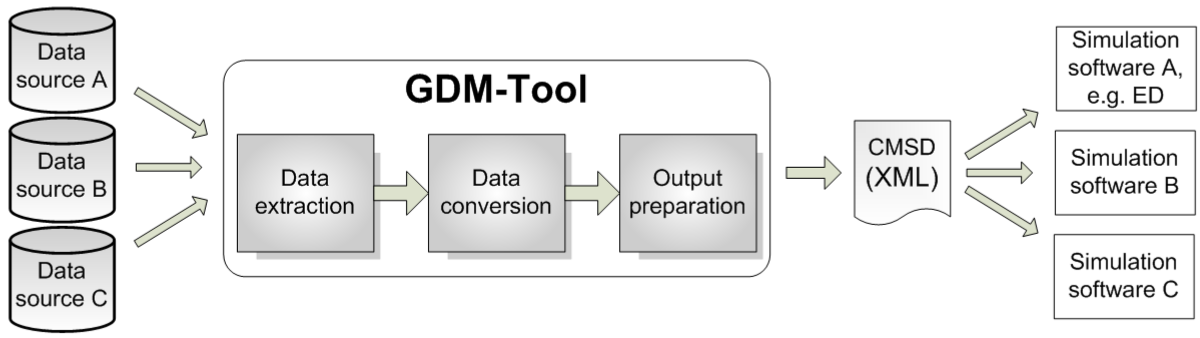

Figure 17: Overview of the proposed concept for automated input data management.

The proposed method, and consequently also its demonstrator the GDM-Tool, is divided into two very central user activities: configuration and automation (see Figure 18). Configuration is required once to specify the sequence of operations for import and processing of data, and export of information to CMSD. Once this mapping is performed, data processing can be repeated in automation mode without further efforts, as long as the modeled system remains unchanged.

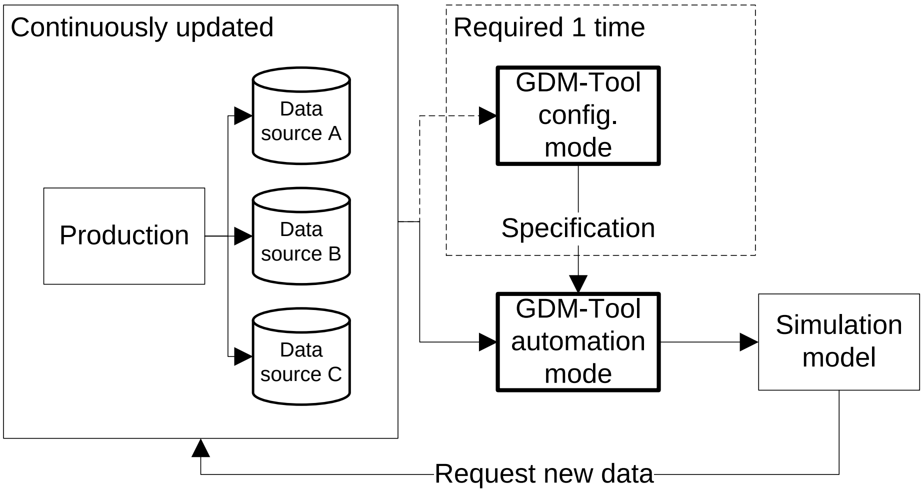

Figure 18: Illustration of the difference between configuration and automation modes in the GDM-Tool.

Configuration is performed by applying a series of tools (area A in Figure 19) and the data can continuously be reviewed in the table view (area B in the same figure). The series of tools is stored as a configuration path (area C), which can automatically be repeated for obtaining updated data sets in automation mode (Figure 20).

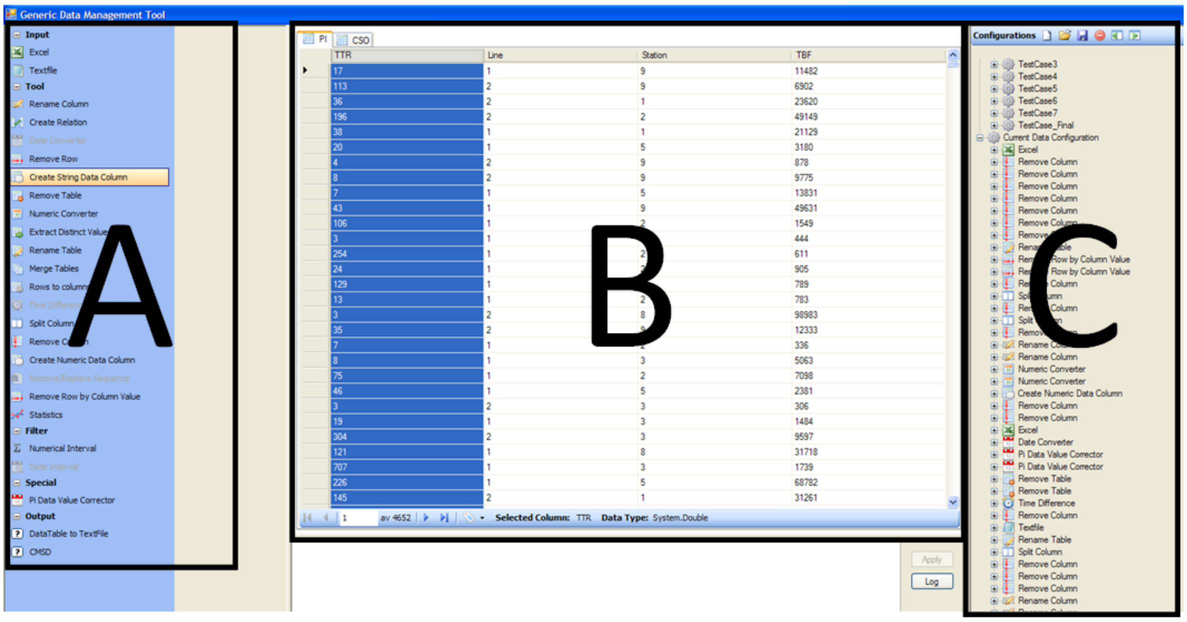

Figure 19: User interface developed to demonstrate the functionalities required for automated input data management.

Automation mode is intended to be more frequently applied than configuration mode in order to gain advantage compared to the reference procedure. Every time the simulation engineer plans to run the model, he or she loads the previously specified configuration, specifies the location of the data sources (latest version) and selects a target for the output file. All steps along the configuration path will be executed when the user clicks the “Run” button, and the CMSD file will be updated with the most recent production data.

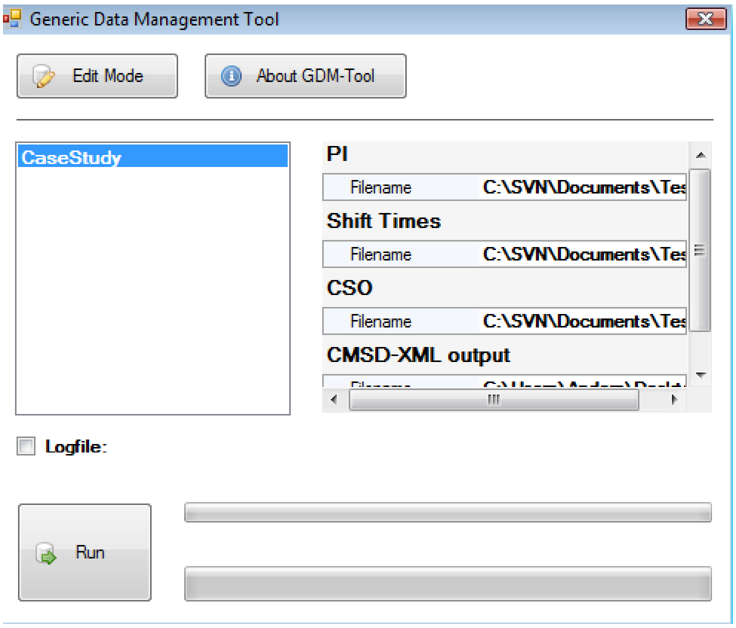

Figure 20: Dialog box for executing a data update using automation mode in the GDM-Tool.

The GDM-Tool is a Windows®-based desktop program written in C# .NET. Due to the lack of standardized data structures, it is unreasonable to strive towards a completely generic interface between data sources and simulation applications. Instead, the GDM-Tool uses a plug-in-based architecture (section 3.7) to facilitate the configuration process described in previous sections. Thus, all data operations for import, processing and export correspond to a plug-in. The plug-in structure also enables easy extension of the GDM-Tool. This is possible since plug-ins are separately developed and compiled, and the application will automatically detect new plug-ins and allow users to apply them without modifying or re-compiling the main program.

The test results, comparing the automated approach to the previously described reference procedure, are shown below. Both the manual process and the GDM-Tool are compared to throughput statistics from the real-world process. The output statistics are collected during the same weeks as the processing times were mapped. However, the breakdown times are collected during a longer period of time to obtain sufficient samples for rigorous statistical analysis.

Table 11: Comparison of the time-consumption between the traditional industrial approach and the GDM-Tool in Publication V.

Process/Activities	Tools	Time-consumption
Reference procedure		
Extraction, categorization, calculations, cleansing	MS Excel®	6 hours, 15 minutes
Condensation, documentation	Distribution-fitting tool	3 hours
Total manual		**9 hours, 15 minutes**
Automated	*The GDM-Tool*	
Configuration	Configuration mode	2 hours
Automated	Automatic mode	< 1 min
Total automated		**2 hours**
Difference		**7 hours, 15 minutes**

Table 11 shows the results with regard to time-consumption measured over the entire process, starting with extraction of raw data and ending with simulation data residing in an interface ready to use in a simulation model. The time-consumption was reduced by 78%, including the configuration steps, given that all necessary plug-ins are available. There were slightly more than 18000 rows of raw data for breakdowns and around 7200 for processing times.

Table 12: A comparison between simulation outputs from traditional and automated input data management.

	Period with known processing times			Extended simulation period		
Output	Manual	Real-world	GDM-Tool	Manual	Real-world	GDM-Tool
Mean	64.9	63.6	62.3	330	324.9	311
Std Dev	3.8	5.7	3.2	4.75	25.3	8.06

Table 12 shows the output data from the real-world process and from one and the same simulation model with input data prepared both manually and by means of the GDM-Tool. All results are given in products per time unit, but the time unit is unpublished for secrecy reasons. To the left in Table 12, the simulation results are compared to real-world data from the same period of time that the raw data for processing time were collected. To the right, the simulation period was extended to six months but still used the same data for processing times. Hence, these data are expected to be generally applicable.

The results show that the data prepared by the GDM-Tool underestimate the total output of products by 2% during the period with correct processing times. During the same time, the reference procedure overestimates the output by 2%. Using hypothesis testing, it is stated that there is no statistical basis for inferring a difference between the two approaches of input data management. For the extended simulation period, the same differences are 4% and 2% in comparison to real-world data for the GDM-Tool and the reference procedure respectively.

4.3.2 Publication VI

- Towards Continuously Updated Simulation Models: Combining Automated Raw Data Collection and Automated Data Processing

This publication originates from the same line of arguments as Publication V. The proposed automated approach, linking data collection, processing and interfacing, is assumed to enable reduction of the extensive time-consumption in input data management. Techniques for automated collection of raw data constitute a more significant part in this paper than in the previous one. Thus, the complete chain of input data management is addressed. Similar solutions, such as MDA (see Chapter 3.6.2), often include highly customized components both for raw data collection and data processing algorithms.

Objective and Contribution to RQ2

The aim of this publication is to evaluate the feasibility of combining automated raw data collection and automated data processing into a push-button solution for DES. The case study is important for validating the proposed concept of automated data management by presenting an additional test case performed in a different type of industry than in Publication V. Furthermore, it adds a measurement of the difference in time-consumption between traditional and automated input data management in order to quantify possible benefits of the concept.

Study Description

This publication combines the capabilities of two existing technologies, MTConnect for automated raw data collection (see section 3.2.2) and the GDM-Tool for automated data processing (see section 4.3.1). The combined solution is designed and tested in a case study at a manufacturing company in the aerospace industry to ensure that it is applicable in a real-world context. All parts of the study were performed in close collaboration with NIST (USA). They built the simulation model requesting the information, contributed detailed knowledge about the production data, and were responsible for the contact with process experts at the company. The study includes the management of MTBF and MTTR for CNC machines, and the information was supplied to the simulation model using CMSD in order to demonstrate how the results can be presented in a neutral format.

In addition to designing and demonstrating the solution, this study also measures the time-consumption for automatically completing the input data management process, by combining MTConnect and the GDM-Tool, and compares it to the industrial reference procedure outlined in section 4.3.1. The reference procedure was performed by a simulation engineer at NIST and the GDM-Tool was configured by the author. Note that the collection of raw data was previously performed at the case study company using MTConnect and, thus, also excluded from the time measurements.

Results and Conclusions

This publication states that the outlined approach to automated input data management, including MTConnect, the GDM-Tool and CMSD, works for the production data included in this case study. Compared to the reference procedure, the time-reduction is 75% (from 4 hours to 1 hour) just for the processing of raw data. If the time for raw data collection is included, which normally takes several days or weeks, the reduction is of course even more significant. Another option evaluated by the authors is to use MS Excel® macros instead of the GDM-Tool and this solution can be almost as efficient, given that the user has built the macros in advance or that he/she has a well-established library of suitable code sections. However, the advantage of the GDM-Tool is that it provides data processing operations applicable for any manufacturing company storing their raw data in some relational table format.

Both MTConnect and the GDM-Tool are quite new applications for input data management to DES, and it is therefore necessary to improve them using further industrial case studies. In this case study some experiences should be highlighted:

- Data provided by MTConnect are polled in intervals specified by the user and, thus, presented as a list of machine *states*. Similar systems in manufacturing industry typically store raw data as *events* containing information of both start time and duration. However, the case set-up included a pre-developed script transforming the states to events.
- Further development of the GDM-Tool, according to the finding above, would streamline the data flow even more by reading the raw data directly from the MTConnect XML-file and eliminate the need for MS Excel®.
- There are still a few issues on how to interpret data points provided by MTConnect. One example is that some down-time samples seemed to be too short to be considered as machine breakdowns in a DES model. Rather, they appeared to be logged due to communication problems between the machine and the MTConnect agent. In any case, more studies are required before using MTConnect data in sharp industrial DES studies.

4.3.3 ADDITIONAL CASE STUDY IN THE AUTOMOTIVE INDUSTRY

As a part of a current research project DFBB (section 2.1), this case study is not yet published as a scientific paper. One aim of the DFBB project is to provide various engineering tools with data throughout the complete life-cycle of production systems, e.g. including conceptual design, implementation, ramp-up and steady state production. The proposed concept for automated input data management, demonstrated by GDM-Tool, is here used for processing of operational data supplied to and stored in digital building blocks for use in DES.

Objective and Contribution to RQ2

This case study is performed at a Swedish automotive company (not the same as in either of Publications V or VI) and aims to test the concept of automated input data management proposed in this thesis. Further, the findings will serve as a basis for possible additions or changes in functionalities provided by the demonstrator (the GDM-Tool). In addition, the difference in time-consumption between traditional (the reference approach outlined in 4.3.1) and automated input data management is measured.

Study Description

In correspondence with the two other case studies used for answering RQ2, this work consists of applying the GDM-Tool for automating the data input activities to a DES model. Here, the simulation model represents parts of a production line for engine components. The machines perform milling operations and they are arranged as a serial production line including parallel machines within operation steps; see Figure 21. The input parameters automatically supplied to the DES model are MTBF and MTTR. The time-consumption for the manual and automated data input processes are measured in the same way as in Publications V and VI, using one sample for each process. The small amount of samples is due to the limited number of personnel trained in the input data management process (see further discussion in section 4.4.3).

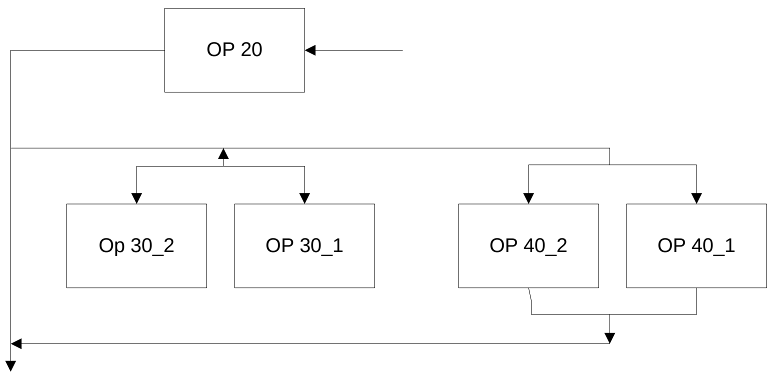

Figure 21: Flow chart of the production line.

Results and Conclusions

The test implementation showed that no further functionality was required compared to the data operations obtained in Publication V. Furthermore, the test implementation resulted in reduced time for data processing, from 2 hours to 30 minutes (75%). The main reason for the increased efficiency is the automatic link between the categorized, corrected and calculated data to the distribution-fitting function.

Table 13: Comparison of the data quality between automated and manual input data management.

Machine	Parameter	Automated processing	Traditional processing	Difference in mean
Op 20	MTTR	Weibull 244, 0.29	Weibull 410, 0.30	Too few samples
	MTBF	Gamma 741731, 0.20	Gamma 617111, 0.25	Too few samples
Op 30_1	MTTR	LogNormal 3.00, 1.87	LogNormal 2.76, 2.02	5.7%
	MTBF	Weibull 2644, 0.28	Weibull 2643, 0.28	0.0%
Op 30_2	MTTR	Weibull 36.48, 0.49	LogNormal 2.56, 1.97	16%
	MTBF	Gamma 946305, 0.15	Gamma 946305, 0.15	0.0%
Op 40_1	MTTR	Weibull 106.24, 0.54	Weibull 126.21, 0.54	16%
	MTBF	Weibull 9652, 0.27	Weibull 9642, 0.27	0.0%
Op 40_2	MTTR	Weibull 23.98, 0.50	Weibull 23.98, 0.50	0.0%
	MTBF	Weibull 6887, 0.25	Weibull 6878, 0.25	0.0%

In this case, the data sets included 5941 rows of data samples from slightly more than one month of production. The information processed by the GDM-Tool and supplied to the simulation model (see Figure 22) turned out to be similar between automated approach and the reference procedure; see Table 13. The data set contained too few stops in Op 20 for statistical data processing and is therefore left out of the comparison. MTBF corresponds well between the two approaches and the difference in mean for MTTR is also relatively low and well within an acceptable interval. For example, the error often differs more between two different distribution families, representing the same data set, because of their individual abilities to mimic short or long breakdowns. This is a proof that the data operations, such as the statistical distribution-fitting plug-in, works correspondingly to commercial stand-alone applications. Consequently, the data were not repeatedly validated using the simulation outputs.

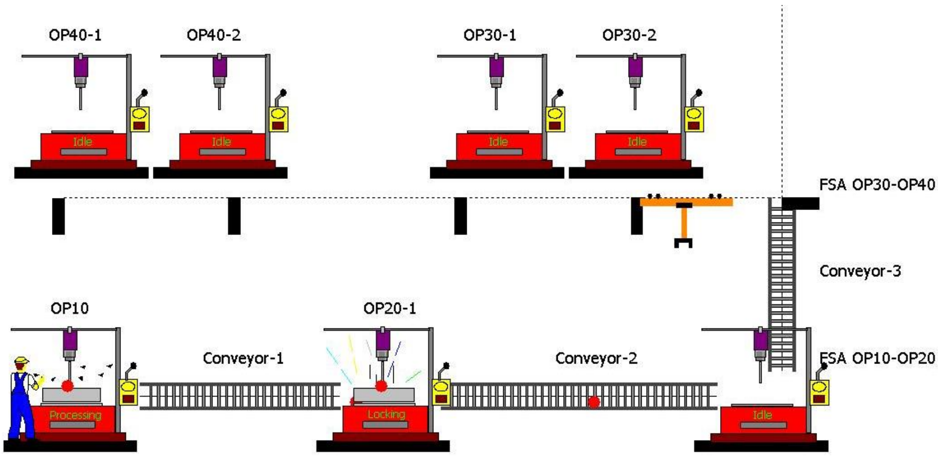

Figure 22: A simple user view of the DES model developed in the commercial simulation package ARENA®.

4.4 Interim Discussion – RQ2

This part includes the specification of required functionality for a middleware solution capable of transforming production data (raw data) to information for DES models. Further, a demonstrator (the GDM-Tool) of the proposed concept is designed, developed and evaluated through three separate and independent case studies. It is extremely important to understand that the author aims to propose a concept for automated input data management, not a commercial software solution ready for the market (despite existing inquiries). The software demonstrator should be considered as a proposed framework, inspiring researchers and companies to develop stand-alone applications or integrated modules in more comprehensive IT systems for production purposes.

The proposed concept is similar to methodology c (Figure 23) presented by Robertson and Perera (2002), which is based on an intermediary database connecting simulation models to CBS applications (e.g. ERP systems). Such solutions are also previously evaluated in research case studies (Randell and Bolmsjö 2001), providing valuable information to this thesis for validation purposes. The difference is that the GDM-Tool can extract data from several sources, more than those traditionally included in the CBS, which is necessary according to the findings in Publication III (RQ1). Further, the tool contains specific data processing functionality to meet the extensive processing requirements connected with DES data (see section 3.1.1). Its architecture also allows efficient extension and customization necessary in special purpose models.

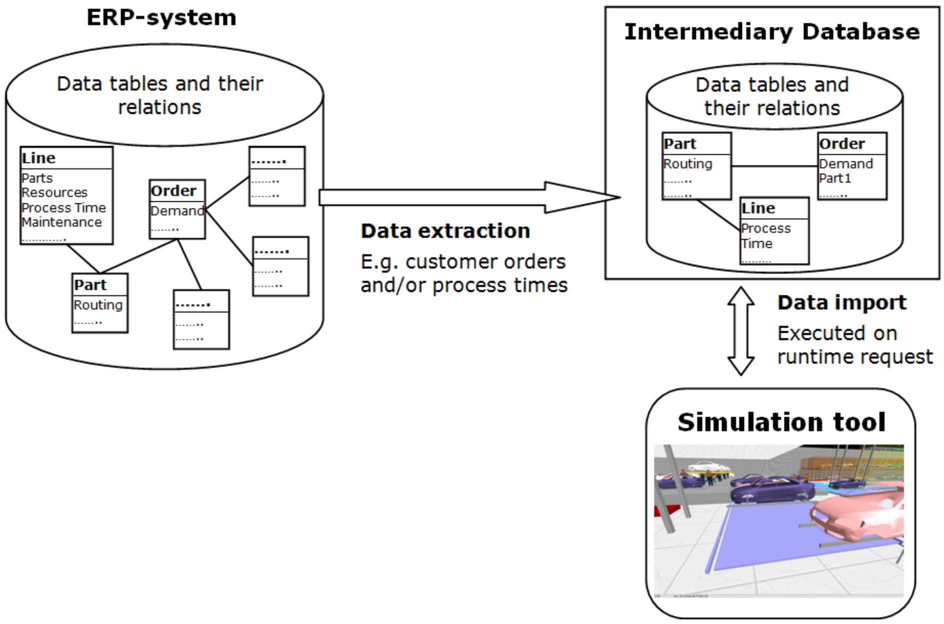

Figure 23: Example of methodology c for automated input data management (Robertson and Perera 2002).

4.4.1 REQUIRED FUNCTIONALITY

The required functionalities of an efficient solution for automated input data management are derived in this thesis by combining information from previous research with empirical findings from the case studies. Table 14 summarizes these functionalities and motivates why they are required. Note that most functionalities relate to the fundamental activities required to transform data to information (Davenport and Prusak 1998). For a more thorough list of all data operations (plug-ins) implemented in the demonstrator (the GDM-Tool); see Publication V. As a complement to the list in Publication V, Publication VI also identified the need for differentiating between *states* and *events* in a table of raw data (e.g. collected using MTConnect). Such functionality can be implemented by allowing removal of rows based on the content in previous rows. This is not yet implemented in the demonstrator due to capabilities of an external script already present in the case study.

Table 14: Required functionalities in an efficient solution for automated input data management.

Functionality	Exemplifying reasons for implementation
Data import from several sources. *Mostly realized by text-file or spreadsheet interfaces to databases or local sources.*	• Previous literature reports poor data availability and problems of identifying available data sources (Perera and Liyanage 2000). • The three case studies required data from multiple data sources. • Publication III reports requirements of several sources, usually connected with using text-file interfaces.
Data table manipulation. *E.g. splitting data columns, merging tables, removing irrelevant rows or columns.*	• Archive analysis from the case studies identified needs for several operations, e.g. to split the date and time from one original column. • Common to remove erroneously logged data (Alexandersson and Wirf 2001).
Formatting of data samples. *Changing data formats to achieve conformity between data sources and to enable necessary calculations.*	• Empirical findings from the case studies show that data from different sources, and even from columns within the same table, are of various formats and unsuited for calculations.
Calculations *All types of calculations, e.g. for finding the TBF or to switch between time units.*	• All parameters encountered during the case studies required calculations. • Limited facilities to organize and manipulate input data in current simulation software (Perera and Liyanage 2000).
Data filtering *To exclude unwanted data points, e.g. because of the desired simulation period.*	• Historical data need to be filtered in order to exclude samples from another system state, e.g. before a major improvement project.
Statistical analysis and condensation *Evaluation of sample independence and selection of the best-fitting distribution.*	• One of the most time-consuming data input activities; see Publication II. • Data need to be condensed to suit neutral formats (SISO 2011). • Limited facilities to organize and manipulate input data in current simulation software (Perera and Liyanage 2000).
Data categorization *Assigning data points to the correct production equipment and tagging the samples to conform to a selected export format.*	• Required to assign data samples to the correct system entities in DES models and data standards (SISO 2011).
Information export to several formats *Export of the results, either to neutral formats (currently CMSD) or to customized interfaces.*	• Interoperability problems are costly for industry (Gallaher, O'Connor, and Phelps 2002). • DES users still use customized formats to a significant extent; see Publication III.

The proposed concept addresses the three most time-consuming activities identified in Publication II (RQ1). Data collection is supported by allowing connections to several data sources. Of course, it presupposes that raw data are automatically, or at least previously, collected and does not fill the common need for manual gathering. Identification and understanding of data sources are needed only once, which will save time in the long run. However, the most significant contribution is in the processing of raw data (addressing the data analysis and preparation), including correction, calculation and condensation. This is a well-known problem (Perera and Liyanage 2000, Publication III), which is automated in the proposed solution.

The functionality of data processing is to a large extent handled by the statistics plug-in in the GDM-Tool. This plug-in imports data sets and starts with analyzing them with regard to sample independence using scatter plots (section 3.3.2.1). Furthermore, it automatically identifies distribution parameters using MLE (section 3.3.2.3) and selects the best-fitting statistical distribution for data condensation by means of KS tests (section 3.3.2.4). Figure 24 visualizes an example of the goodness-of-fit evaluation visualized in a P-P plot.

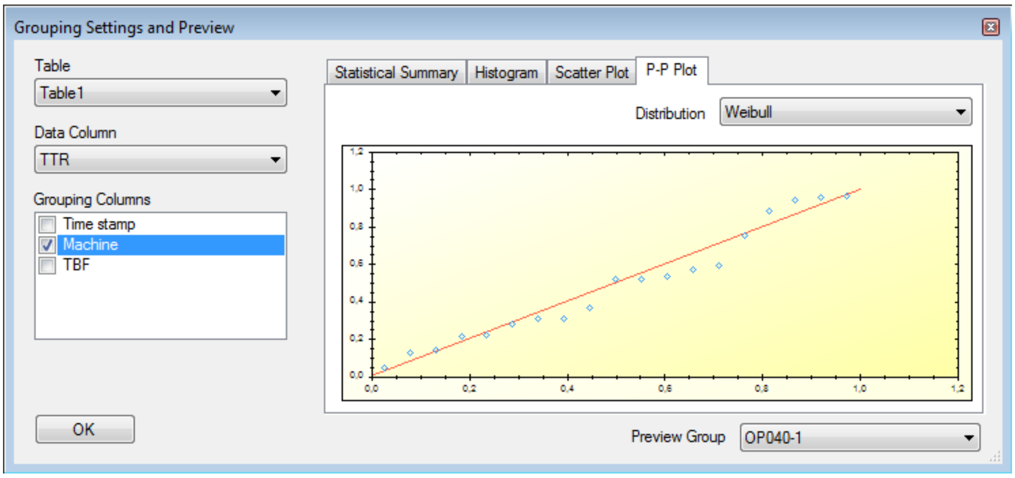

Figure 24: A P-P plot exemplifying the goodness-of-fit functionality provided by the statistics plug-in.

As mentioned in Publication V, the automated execution of data activities requires a first-time configuration. Once a configuration is set, the most recent data files can automatically be processed and the results supplied to the DES model for up-to-date decision support. An example of a configuration path using common data operation plug-ins is shown in Figure 25.

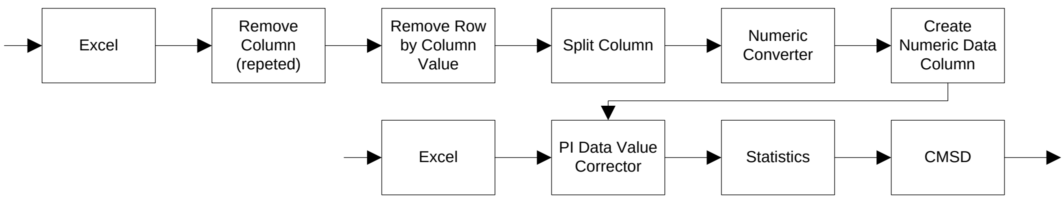

Figure 25: Example of common data operations applied in a typical configuration for obtaining the MTTR from an ACS.

There is of course potential to improve the demonstrator even further in future research projects, especially regarding additional support in statistical input modeling. For example, the present condensation of raw data sets exclusively supports statistical distributions. This representation is selected because of its capability to extend the simulation period further than the actual time for data collection (Robinson 2004). However, some simulation practitioners prefer the precision offered by more empirical representations such as empirical distributions, traces and bootstraps. An extension to include these representations is possible, but traces and bootstraps are more difficult due to the lack of support in existing standards. Additional data representations would also allow the engineers to refuse statistical representations based on the P-value, which is not possible at present. The best-fitting statistical distribution is automatically selected based on the D-statistics; see section 3.3.2.4.

Another possible extension is to increase the level of automation in the detection of outliers, e.g. data samples erroneously collected due to communication problems between data collection and storage applications (Alexandersson and Wirf 2001). Algorithms for outlier detection are already available and, thus, implementation to the GDM-Tool is just a matter of time. Similar algorithms are available for statistically evaluating the independence of samples, which is now a manual monitoring option performed by studying a scatter-plot. However, note that all possible extensions above are additions to a common industrial procedure (the reference procedure) and not required to match the quality of conventional DES models.

4.4.2 Evaluation of the Proposed Concept

Table 15 contains the findings from the three case studies used to validate the proposed concept of automated input data management for DES. The time-reduction compared to the reference procedure (section 4.3.1) is calculated including the data input activities: collection, processing, and documentation. Note that all possible data input activities are not included, just parts necessary to update when using simulation models on a continuous basis. For example, manual gathering does not exist in any of the three case studies since the presence of category A data is a prerequisite. Other activities only have to be performed on one initial occasion, e.g. the identification of relevant parameters and the sources to collect raw data from. This is of course the same for both the automated and the conventional approach.

Table 15: Compilation of the results from the validation of the proposed concept of automated input data management performed in three test cases.

Case study	Data sources	DES parameters	Time-reduction	Identified functionalities
#1	Processing times and shift times from planning system, stop logs from local ACS.	Processing times, MTBF, MTTR	78%	All necessary functionality (first case study).
#2	MTConnect alarm time stamps.	MTBF, MTTR	75%	Transformation between states and events.
#3	Stop logs from local ACS, shift times from local spreadsheet.	MTBF, MTTR	75%	
Average			**76%**	

The evaluation shows that the proposed concept for automated input data management is capable of processing input data from several different sources more efficiently than traditional industrial approaches. The main reasons are that the GDM-Tool demonstrates more comprehensive functionality regarding necessary data operations, and that all steps in the input data management process can be completed in one procedure without manual involvement. The traditional approach relies on manual handling and special-purpose tools for data categorization, correction, calculations and condensation. Such tools have limited capabilities to categorize, correct and calculate the data (Perera and Liyanage 2000), and the use of special-purpose solutions also results in manual handling of data between applications, for example to supply data from one application to another.

Regarding the supply of data to DES models, the approach is validated using the neutral format CMSD (SISO 2011) in all three case studies. The demonstrator also supports export to customized interfaces such as MS Excel® spreadsheets. Moreover, initial data mappings of AutomationML (AutomationML consortium 2010) and STEP AP214 (Kjellberg et al. 2009, Falkman et al. 2008), performed in the research project DFBB (Chalmers PPU 2011), also shows that interoperability with the GDM-Tool should be possible to implement. This is, however, not yet tested in a real-world environment.

A detailed quantification of data quality obtained in the three case studies is delimited in this thesis; see section 1.6. However, it is important to highlight that all three case studies partly assessed the data quality in order to ensure that it is maintained compared to the conventional industrial approach. One case compared the approaches using hypothesis testing (Montgomery and Runger 1999) and the other two by comparing the input parameters using descriptive statistics and face validation (Sargent 2005). In the additional case study, the difference in mean between the automated and manual approaches might seem extensive (Table 15) for some operations. However, such differences are normal also between manually selected distributions and the original data samples, due to the statistical distributions' way of representing short and long breakdowns. When using univariate distributions, long breakdowns are often unrepresented in the data supplied to the simulation, which is normally desired by the simulation engineer who wants to model production systems under "normal" circumstances.

The validation is mainly performed using two DES parameters: MTBF and MTTR. However, the solution can easily be applied to other quantitative parameters, such as set-up times and processing times. The latter is demonstrated by being part of the validation in Publication V. The selection of MTBF and MTTR is strategic, since the handling of these two parameters is very extensive and covers the needs of most other (quantitative) simulation parameters. The handling of MTBF and MTTR includes (Williams 1994):

- An extensive number of samples due to the parameters' importance for model dynamics.
- The raw data formats are often very crude and require corrections and conformation.
- Calculations are required, e.g. for subtracting a start time of a breakdown with the corresponding value for the previous stop (MTBF).
- Data from several sources are often required. For example, a combination of stop data and work schedules is required to obtain MTBF and MTTR cleansed from non-production time.
- Condensation is preferred to facilitate the supply of all data points as information to the simulation model.
- "Outliers" are commonly encountered, e.g. due to communication problems between the logging equipment and the database.

4.4.3 METHODOLOGICAL DISCUSSION

Development and validation of the proposed approach to automated input data management, and its demonstrator the GDM-Tool, are based on three case studies. This is generally considered satisfactory for initial validation (Flynn et al. 1990). Furthermore, the three cases represent different companies, countries and lines of business, which is good for varying the contextual factors in a data triangulation approach (Denscombe 2007). However, the fewer cases the more influence of company-specific circumstances, so it is of course desirable to continue with further tests and evaluations in future research projects. The limitation to three samples also makes the statistical calculations somewhat vulnerable. 76% time-reduction should therefore be considered more as an indication than as an absolute value at this stage. The fact that the automated approach reduces the time-consumption (RQ2) is nonetheless considered to be proven.

A second consideration in the validation cases is the possible variation between users of the demonstrator (the GDM-Tool). Thus, the user aspect is delimited here and all configurations and test implementations have been performed by people from the research team, mainly by the author himself. The chief reason is that there are not enough employees in the partner companies with the necessary education and experience in input data management to set up a complete test including user variations. The author has experience of several industrial DES projects including sole responsibility of all steps of input data management, so it would be interesting to make future tests including engineers with less experience in DES.

A final comment is that no case study shows a situation where automated input data management is inappropriate. It would have been interesting to obtain empirical findings from such circumstances, but the appropriateness of automation is now discussed solely on a theoretical basis; see section 5.1. The selection of case studies is dependent on the companies participating in the three projects during the PhD studies (Figure 5).

5 DISCUSSION

This chapter aims to connect the two parts of this thesis (RQ1 and RQ2) and relate the results to the thesis' purpose and aim. Additional discussions of findings associated with each part can be found in the interim discussions located in sections 4.2 and 4.4. Note that the methodological discussions are also provided in the interim discussions.

Input data management is still one of the most critical and time-consuming phases of a DES project. Ten years ago Robertson and Perera (2001) stated that "It is strongly argued that data collection is the most crucial and time-consuming stage in the model building process" and there has been no significant improvement of the efficiency since then. This thesis shows that input data management still consumes on average 31% of the total project time. Thus, further efforts in this area are needed to increase the relatively low dissemination of DES in industry (McNally and Heavey 2004).

Results from Publication III (RQ1) show that many companies work hard to implement more efficient solutions (methodologies c and d in Robertson and Perera (2002)), but need further support from researchers, which is provided in this thesis. The desire and spirit to move forward mean that successful solutions to the difficulties in input data management will increase the use of DES on a daily basis. Models will be continuously updated without major efforts and, thus, production engineers have access to an analysis tool capable of including dynamic aspects of production systems. Balancing and system losses can be reduced in a more effective way compared to the present situation. This will in turn lead to more robust and efficient production systems.

The main solution proposed in this thesis (RQ2) relies on automated input data management. In order to be as effective as possible, it is mainly designed to support engineers in the work with the three most time-consuming activities identified in RQ1: data collection, identification of available data, and data analysis and preparation. A demonstrator is developed and three independent case studies have shown the concept to be feasible and that the time-consumption was reduced by 76% on average. The key features of the proposed solution are that the intermediary application allows import of data from several sources and provides necessary functionality for automated transformation of data to information for DES models. The first feature facilitates data collection and reduces the need for repetitive identification of available data. The second feature naturally increases efficiency in data analysis and preparation.

This concept can be categorized as a methodology c solution (section 3.6), which might be perceived as somewhat defensive for being front-end research. However, findings from Publication III indicate that industry prefers such an approach, and the publication also pinpoints several problems with the even more integrated methodology d. These problems are:

- Data in sources within the CBS are not detailed enough for simulation (Moon and Phatak 2005).
- Simulation projects still rely on data from several sources; significant parts of the data reside in local systems (Publication III).
- There is a need for additional data processing and possibilities for creating what-if scenarios.
- There is a diversity of simulation tools and they are selected on the basis of their different strengths as well as the employees' experience (Semini, Fauske, and

Strandhagen 2006). This fact makes totally integrated PLM packages (example of methodology d) challenging to implement throughout organizations.

As an alternative to exclusively relying on single PLM packages, the latter problem above could instead be addressed by increased interoperability using neutral data formats; see examples in section 4.4.2. Neutral formats would also facilitate the distribution of data between other production engineering applications, e.g. ergonomic simulations, layout planning tools and line balancing software.

5.1 Is Automation Always Feasible?

Although automation of data input activities holds tremendous potential for reducing the time-consumption in DES projects, it is possible to identify situations where other approaches are still more applicable. The most obvious is in SME not working with simulation on a regular basis. Such companies usually have no specialists dedicated to DES analyses, too much unavailable (category C) data, and limited possibilities for investing in the necessary equipment. Therefore, systematic approaches outlined as best-practice guidelines for manual input data management are also necessary. The state-of-the-art descriptions provided in Publication I can be used for this purpose, but there are also other supporting publications in the area (see for example Bernhard and Wenzel 2005, Perera and Liyanage 2000, Lehtonen and Seppälä 1997).

Furthermore, in addition to the reliance on available data, the confidence in automated solutions is dependent on the competence and experience of the user. A user who is familiar with the modeled production system and common data input activities can follow the process and interpret the output of a software solution automating certain steps (e.g. the GDM-Tool). However, less experienced users might perceive the automated process as a black box and, thus, question the validity and credibility of the information submitted to the simulation models. This is a situation where the more systematic approaches also would be preferable.

An additional prerequisite for successful automation is that the modeled system must stay unchanged with regard to production equipment, parts routing, etc. Such major changes require modifications in the model logic, which are not handled by an automated input data management application. Consequently, automated input data management is most beneficial when connected to models used on a frequent basis for continuous improvement of production performance. Frameworks for facilitating updated model logics belong to another research area, for example described in publications about SysML (Huang, Ramamurthy and McGinnis 2007).

5.2 Research Contribution and Possible Industrial Applications

One contribution to the research community is the mapping of current industrial state-of-the-art in the input data management process. The identification of the most time-consuming activities is specifically important for prioritizing future efforts related to research on more efficient data handling. Focus on supporting the activities having highest impact on the total time-consumption holds higher potential for significant increases in efficiency. Moreover, the outline of a solution for automated input data management adds value to previous contributions by providing suggestions on detailed functionality, i.e. necessary data operations. Previous contributions in the area are more targeted on higher-level system architecture.

For industry, one contribution is the hands-on guideline on input data management, including best practice-descriptions of the different activities. Such systematic guidelines are useful for increasing efficiency of input data management in companies with a limited amount of category

A data, mainly SME. The main contribution, however, is the description of functionalities and software architecture for automated input data management, which is presented using the demonstrator called the GDM-Tool. This demonstrator realizes the middleware solution requested and identified as a gap in section 1.1. The GDM-Tool is not yet robust enough to be launched as a commercial product, but it may well serve as a prototype for commercial software solutions realized by major DES users, consultancy firms and vendors of production data management systems.

5.3 Future Research

Many interesting ideas for further research have been identified in the previous discussion sections (0 and 4.4). One that is highly prioritized for the author and the Virtual Production research group at Chalmers is to incorporate environmental analyses in DES studies. Significant progress is already obtained but the acquisition of accurate data is frequently mentioned as a major problem. Firstly, further studies are needed to evaluate whether the type of data representation described in Publication IV is valid for other parameters than electrical power and for different types of production equipment. Secondly, research is required on automated extraction and processing of data from external databases, e.g. EcoInvent, ELCD and UPLCI (Overcash, Twomey and Kalla 2009).

As discussed in section 4.4.3, three test cases are used for the development and validation of the proposed concept for automated input data management. This limited number of cases results in difficulties of presenting exact quantifications of the concept potential. Consequently, more case studies are needed for better precision of the obtained reduction in time-consumption, and for increased confidence in the necessary data processing functionality. Additionally, current work focuses on the technical aspects of automation, and future cases should therefore also include human aspects of the engineers working together with the application. What education and experience are required to reduce the time-consumption using the GDM-Tool, and how should the user interface be improved to support the usage among non-experts?

An additional recommendation, for both industry and researchers, is to intensify the work on increasing the amount of category A data. At present, significant efforts are required for manual data collection. This fact contributes to the extensive time-consumption and hinders automated solutions. One solution is further implementation of continuous collection systems, preferably systems enabling data collection on a temporary basis. Such a solution would also increase the possibilities for efficient data collection in SME. Another alternative is to include the data requirements of DES when implementing major CBS applications, for example Manufacturing Execution Systems (MES) or ERP systems.

A great option for a future research project is therefore to develop and test a flexible data collection system and automatically connect it to a sharp implementation of the GDM-Tool for further data processing. A DES model used for continuous improvements should be the receiver of the final result. Full implementations running at three companies for six months each would be satisfactory for studying the precision in data collection, the reduction in time-consumption including the user influence, and the effects of using DES as a desktop resource. The author is fully convinced that the participating companies would experience an increased utilization of existing resources as well as a more robust and reliable production output. By extension, increased capacity, higher service levels, and possibilities to reduce stock levels throughout the value chain would be obtained.

6 CONCLUSIONS

> RQ1: What is the current industrial state-of-the-art for the input data management process?
>
> RQ2: How can efficient and automated input data management, for simulation of material flows in production, be realized?

The aim of this thesis is to reduce the time-consumption for input data management in simulation of production flows. Increased efficiency in this process will enable more frequent use of dynamic simulations and, thus, support production engineers in improving performance and robustness of production systems. Results in one of the appended publications indicate limited progress during the last decade and states that companies need further assistance in finding appropriate solutions for automating the data input activities. One such solution is presented and evaluated in this thesis.

Today, the industrial state-of-the-art for the input data management process (RQ1) includes a significant amount of manual work. This is a major reason for the fact that input data management consumes on average 31% of the entire time during Discrete Event Simulation projects. The level of automation is generally low in the gathering of raw data as well as in data processing. Furthermore, the supply of information to simulation models is also heavily dependent on human involvement, either using customized spreadsheet interfaces or even typing the values directly in the model code. Thirteen distinct data input activities are identified and described in section 4.1.1. This description may well be used as a best-practice guide-line for increasing efficiency and precision in input data management. An evaluation of 15 industrial DES projects also shows that the three most time-consuming activities are: data collection, identification of available data, and data analysis and preparation.

The proposed solution for automated input data management (RQ2) is presented using a software demonstrator, called the GDM-Tool. The most significant difference between this solution and the common industrial approach is that the GDM-Tool enables automation and complete integration of data collection, data processing, and supply of information to simulation models. The GDM-Tool extracts data from several sources, demonstrates the necessary operations for data processing, and provides an automated link to simulation models using neutral formats. The current set of data operations has shown sufficient capabilities in three independent case studies performed in the automotive and aerospace industries. However, the plug-in-based architecture of the GDM-Tool also provides the possibility to add functionality if future case studies or real-world implementations require more customized solutions.

Three case studies have shown that the time-consumption for input data management can be reduced by approximately 75% compared to the common industrial approach, given that the necessary raw data are available. This number includes a one-time configuration, so the potential is even higher when simulation models are repeatedly used. The main reason for the increased efficiency is improvement in the data processing step. Several manual activities for data categorization, correction, calculations, and condensation are removed thanks to the unbroken chain of automated activities. The integration of data operations also eliminates the need for manual handling of data between different special-purpose applications.

7 REFERENCES

Alexandersson, T. and L. Wirf. 2001. *Improving the Input Data Treatment for Increased Simulation Output Quality*. M.Sc. Thesis, Department of Production Engineering, Chalmers University of Technology, Gothenburg, Sweden.

Aufenanger, M., A. Blecken, and C. Laroque. 2010. Design and Implementation of an MDA Interface for Flexible Data Capturing. *Journal of Simulation* 4(4):232-241.

AutomationML consortium. 2010. Whitepaper AutomationML Part 1 – AutomationML Architecture (State May 29 2010). Available from: http://www.automationml.org/. Accessed on July 24 2011.

Balderud, J., and A. Olofsson. 2008. *A Plug-in Based Software Architecture for Generic Data Management.* M.Sc. Thesis, Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden.

Banks, J., J.S. Carson and B.L. Nelson. 1996. *Discrete-Event System Simulation* (2nd ed.), Prentice-Hall, Upper Saddle River, New Jersey.

Bernhard, J. and S. Wenzel. 2005. Information Acquisition for Model Based Analysis of Large Logistics Networks. In: *Proceedings of the 19th European Conference on Modelling and Simulation,* eds. Y. Merkuryev, R. Zobel and E. Kerckhoffs, 37-42.

Bordens, K.S. and B.B. Abbot. 2005. *Research Design and Methods: A Process Approach*. McGraw-Hill, New York.

Bryman, A., and E. Bell. 2007. *Business research methods* (2nd ed.), Oxford University Press, New York.

Cao, H.-J., Y.-C. Chou, H. H. Cheng. 2009. Mobile Agent Based Integration Framework for Flexible Dynamic Job Shop Scheduling, In: *ASME Conference Proceedings IDETC/CIE2009*.

Chalmers PPU - Department of Product and Production Development. 2011. http://www.chalmers.se/ppd/EN/reserach/production-system [accessed on May 28, 2011].

Choo, C. W., B. Detlor and D. Turnbull. 2000. *Web work: information seeking and knowledge work on the World Wide Web*, Kluwer Academic, Dordrecht.

Coughlan, P., and D. Coughlan. 2002. Action Research for Operations Management. *International Journal of Operations & Production Management* 22(2):220-240.

Danemark, B., M. Ekström, L. Jakobsen, and J.Ch. Karlsson. 1997. *Att förklara samhället*, Studentlitteratur, Lund (in Swedish).

Dassault Systemes. 2011. Delmia Digital Manufacturing & Production. Available from http://www.3ds.com/products/delmia/welcome/. Accessed on August 3, 2011.

Davenport, T. H. 1997. *Information Ecology*, Oxford University Press, New York.

Davenport, T. H. and L. Prusak. 1998. *Working knowledge: How organizations manage what they know*. Harvard Business School Press, Boston, Massachusetts.

Denscombe, M. 2007. *The good research guide: for small-scale social research projects* (3rd ed.), Open University Press, Maidenhead.

DIN Deutsches Institut für Normung e.V. 1995. Informationsverarbeitung – Part 1. DIN44300-1. Beuth, Berlin (in German).

Dubois, A. and L.-E. Gadde. 2002. Systematic combining: an abductive approach to case research. *Journal of Business Research*, 55:553-560.

Dungan, P. and C. Heavey. 2010. Proposed Visual Wiki System for Gathering Knowledge About Discrete Event Systems. In: Proceedings of the 2010 Winter Simulation Conference, eds. B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 513-521.

EcoProIT research project. 2011. Chalmers University of Technology. Available from: http://www.EcoProIT.com. Accessed on August 11, 2011.

Falkman, P., J. Nielsen, B. Lennartsson and A. von Euler-Chelpin. 2008. Generation of STEP AP214 models from discrete event systems for process planning and control. *IEEE Transactions on Automation Science and Engineering*, 5(1):113-126.

Flynn, B.B., S. Sakakibara, R.G. Schroeder, K.A. Bates, and E.J. Flynn. 1990. Empirical Research Methods in Operations Management. *Journal of Operations Management* 9(2):250-284.

Gallaher, M., A. O'Connor and T. Phelps. 2002. Economic Impact Assessment of the Standard for the Exchange of Product Model Data (STEP) in Transportation Equipment Industries. *Planning report 02-5*, Prepared for the National Institute for Standards and Technology.

Gamma, E., R. Helm, R. Johnson and J. Vlissides. 1995. *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley Longman Publishing Co. Inc., Boston, MA, USA.

Geer Mountain Software Corporation, Stat::Fit Commercial Webpage. Available from http://www.geerms.com/. Accessed on August 4, 2011.

Glaser, B.G. and A.L. Strauss. 1967. *The discovery of grounded theory: strategies for qualitative research*, Aldine de Gruyter, New York.

Gummesson, E. 2000. *Qualitative Methods in Management Research* (2nd ed.). SAGE, Thousand Oaks.

Hatami, S. 1990. Data Requirements for Analysis of Manufacturing Systems Using Computer Simulation. In: *Proceedings of the 1990 Winter Simulation Conference*, eds. O. Balci, R.P. Sadowski and R.E. Nance, 632-635.

Heilala, J., S. Vatanen, J. Montonen, H. Tonteri, B. Johansson, J. Stahre, and S. Lind. 2008. Simulation-Based Sustainable Manufacturing System Design. In *Proceedings of the 2008 Winter Simulation Conference*, eds. S.J. Mason, R.R. Hill, L. Mönch, O. Rose, T. Jefferson, and J.W. Fowler, 1922-1930.

Hollocks, B.W. 2001. Discrete-event simulation: an inquiry into user practice. *Simulation Practice and Theory*, 8:451-471.

Huang, E., R. Ramamurthy and L.F. McGinnis. 2007. System and Simulation Modeling using SysML. In: *Proceedings of the 2007 Winter Simulation Conference*, eds. S.G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J.D. Tew and R.R. Barton, 796-803.

Ingalls, R.G. 2002. Introduction to Simulation. In *Proceedings of the 2002 Winter Simulation Conference*, eds. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 7-16.

Ingemansson, A. 2004. *On Reduction of Production Disturbances in Manufacturing Systems Based on Discrete-Event Simulation*. Doctorial dissertation, Department of Mechanical Engineering, Lund University, Lund, Sweden.

Ingemansson, A., T. Ylipää and G. S. Bolmsjö. 2005. Reducing bottle-necks in a manufacturing system with automatic data collection and discrete event simulation. *Journal of Manufacturing Technology Management*, 16:615-628.

Institute for Environment and Sustainability. 2010. European Reference Life Cycle Database. Available from: http://lca.jrc.ec.europa.eu/lcainfohub/datasetArea.vm. Accessed on August 11, 2011.

Johansson, B., Å. Fasth, J. Stahre, J. Heilala, S.K. Leong, Y.T. Lee, and F.H. Riddick. 2009. Enabling Flexible Manufacturing Systems by Using Level of Automation as Design Parameter. In: *Proceedings of the 2009 Winter Simulation Conference*, eds. M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin and R.G. Ingalls, 2176-2184.

Johansson, M., B. Johansson, S.K. Leong, F.H. Riddick, Y.T. Lee. 2008. A Real World Pilot Implementation of the Core Manufacturing Simulation Data Model. In: P*roceedings of the Summer Computer Simulation Conference*. Edinburgh, Scotland.

Johansson, M., and R. Zachrisson. 2006. *Modeling automotive manufacturing process*. Master's thesis. Department of Product and Production Development, Chalmers University of Technology, Gothenburg, Sweden.

Kibira, D. and S. K. Leong. 2010. Test of Core Manufacturing Simulation Data Specification in Automotive Assembly. In: *Proceedings of the Simulation Interoperability Standards Organization (SISO) and Society for Modeling and Simulation (SCS) International European Multi Conference*, Orlando, Florida, USA.

Kjellberg, T., A. von Euler-Chelpin, M. Hedlind, M. Lundgren, G. Sivard, and D. Chen. 2009. "The machine tool model - A core part of the digital factory". In: *CIRP Annals - Manufacturing Technology*, 58(1): 425-428.

Kleindienst, J. and D. Juricic. 2007. Optimal Selection of Information Terminals for Data Acquisition in Manufacturing Processes. In: *Proceedings of the 6th EUROSIM Congress on Modelling and Simulation*, eds. B. Zupančič, R. Karba and S. Blažič.

Kumar, S. and D.A. Nottestad. 2009. Flexible capacity design for the Focus Factory – a case study. *International Journal of Production Research*, 47(5): 1269–1286.

Kühn, W. 2006. Digital Factory – Simulation Enhancing the Product and Production Engineering Process. In *Proceedings of the 2006 Winter Simulation Conference*, eds. L.F. Perrone, F.P. Wieland, J. Liu, B.G.Lawson, D.M. Nicol, and R.M. Fujimoto, 1899-1906.

Law, A. M. and M. G. McComas. 2003. How the ExpertFit Distribution-Fitting Software can make your Simulation Models more Valid. In: *Proceedings of the 2003 Winter Simulation Conference*, eds. S. Chick, P.J. Sanchez, D. Ferrin and D.J. Morris, 169-174.

Law, A.M. 2007. *Simulation modeling & analysis* (4th ed.), McGraw-Hill, New York.

Leemis, L. 2004. Building credible input models. In: *Proceedings of the 2004 Winter Simulation Conference,* eds. R.G. Ingalls, M.D. Rossetti, J.S. Smith and B.A. Peters, 29-40.

Lehtonen, J.-M. and U. Seppälä. 1997. A methodology for data gathering and analysis in a logistics simulation project. *Integrated Manufacturing Systems*, 8:351-358.

MANUFUTURE EuroStat 2006

McLean, C., and S. Leong. 2001. The expanding role of simulation in future manufacturing. In *Proceedings of the 2001 Winter Simulation Conference*, eds. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 1478-1486.

McNally, P., and C. Heavey. 2004. Developing simulation as a desktop resource. *International Journal of Computer Integrated Manufacturing* 17:435-450.

Miles, M.B. and A.M. Huberman. 1994. *Qualitative data analysis: an expanded sourcebook* (2nd ed.), SAGE Publications, Thousand Oaks.

Montgomery, D.C. and G.C. Runger. 1999. *Applied statistics and probability for engineers* (2nd ed.), Wiley & Sons, New York.

Moon, Y. B., and D. Phatak. 2005. Enhancing ERP system's functionality with discrete event simulation. *Industrial Management & Data Systems* 105:1206-1224.

MTConnect Institute. 2010. Available from http://www.mtconnect.org. Accessed on August 7, 2011.

Nonaka, I. and H. Takeuchi. 1995. *The knowledge-creating company: how Japanese companies create the dynamics of innovation*, Oxford University Press, New York.

OPC Foundation. 2010. Available from http://www.opcfoundation.org. Accessed on August 7, 2011.

Overcash, M., J. Twomey, and D. Kalla. 2009. Unit Process Life Cycle Inventory for Product Manufacturing Operations, *ASME Conference Proceedings MSEC2009*, West Lafayette, IN, USA.

Pegden, C.D., R.E. Shannon and R.P. Sadowski. 1995. *Introduction to simulation using SIMAN* (2nd ed.), McGraw-Hill, New York.

Perera, T. and K. Liyanage. 2000. Methodology for rapid identification of input data in the simulation of manufacturing systems. *Simulation Practice and Theory*, 7:645-656.

Perrica, G., C. Fantuzzi, A. Grassi, G. Goldoni, and F. Raimondi. 2008. Time to Failure and Time to Repair Profiles Identification. In *Proceedings of the 5th FOODSIM conference*. Dublin, Ireland.

Pidd, M. 1996. *Tools for Thinking: Modelling in Management Science*, John Wiley & Sons, Chichester.

ProViking. 2011. http://www.chalmers.se/hosted/proviking-en/ [accessed on May 29, 2011].

Randell, L.G., and G.S. Bolmsjö. 2001. Database driven factory simulation: a proof-of-concept demonstrator. In: Proceedings of the 2001 Winter Simulation Conference, eds. B.A. Peters, J.S. Smith, D.J. Medeiros, and M.W. Rohrer, 977-983.

Robertson, N. and T. Perera. 2001. Feasibility for automatic data collection. In: *Proceedings of the 2001 Winter Simulation Conference,* eds. B.A. Peters, J.S. Smith, D.J. Medeiros, and M.W. Rohrer, 984-990.

Robertson, N., and T. Perera. 2002. Automated data collection for simulation?. *Simulation Practice and Theory* 9:349-364.

Robinson, S. and V. Bhatia. 1995. Secrets of successful simulation projects. In: *Proceedings of the 1995 Winter Simulation Conference,* eds. C. Alexopoulos, K. Kang, W.R. Lilegdon and D. Goldsman, 61-67.

Robinson, S. 2004. *Simulation: The Practice of Model Development and Use*. John Wiley & Sons Ltd, Chichester.

Sargent, R.G. 2005. Verification and validation of simulation models. In: *Proceedings of the 2005 Winter Simulation Conference,* eds. M.E. Kuhl, N.M. Steiger, F.B. Armstrong, and J.A. Joines, 130-143.

Semini, M., H. Fauske, and J.O. Strandhagen. 2006. Applications of Discrete-Event-Simulation to Support Manufacturing Logistics Decision-Making: A Survey. In *Proceedings of the 2006 Winter Simulation Conference*, eds. L.F. Perrone, F.P. Wieland, J. Liu, B.G.Lawson, D.M. Nicol, and R.M. Fujimoto, 1946-1953.

Siemens. 2011. Siemens PLM Software, Teamcenter. Available from http://www.plm.automation.siemens.com/en_us/products/teamcenter/. Accessed on August 3, 2011.

SISO. 2010. Simulation Interoperability Standards Organization: SISO Policies and Procedures. Available from http://www.sisostds.org/. Accessed on November 16, 2010.

SISO. 2011. Simulation Interoperability Standards Organization, CMSD Product Development Group. *SISO-STD-010-2011*, April 12 2011.

Solding. P., D. Petku, and N. Mardan. 2009. Using simulation for more sustainable production systems – methodologies and case studies. *International Journal of Sustainable Engineering* 2:111-122.

Solding, P., P. Thollander, P.R. Moore. 2009. Improved energy-efficient production using discrete event simulation. *Journal of Simulation,* 3:191-201.

Starrin, B. and P.G. Svensson. 1994. *Kvalitativ metod och vetenskapsteori*, Studentlitteratur, Lund (in Swedish).

Stevenson, W.D. 1982. *Elements of Power System Analysis* (4rd ed.), McGraw Hill, New York.

Swiss Centre for Life Cycle Inventories. 2011. The EcoInvent Database. Available from: http://www.ecoinvent.org/database/. Accessed on August 11, 2011.

Trybula, W. 1994. Building simulation models without data. IEEE International Conference on Systems, Man, and Cybernetics. *Humans, Information and Technology*, 1:209-214.

UML Resource Page. 2009. Unified Modeling Language. Available from http://www.uml.org/. Accessed on January 3, 2009].

Van der Spek, R. and A. Spijkervet. 1997. Knowledge Management: Dealing Intelligently with Knowledge. In: *Knowledge Management and its Integrative Elements*, eds. J. Liebowitz and L. Wilcox, 31-58, CRC Press.

Van der Zee, D.-J. and J.G A.J. Van der Vorst. 2007. Guiding principles for conceptual model creation in manufacturing simulation. In: *Proceedings of the 2007 Winter Simulation Conference*, eds. S.G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J.D. Tew and R.R. Barton, 776-784.

Wallén, G. 1996. Vetenskapsteori och forskningsmetodik (2nd ed.), Sudentlitteratur, Lund (in Swedish).

Weick, K.E. 1979. *The social psychology of organizing* (2nd ed.), Random House, New York.

Wild, R. 1975. On the selection of mass production systems. *International Journal of Production Research* 13:443-461.

Wilkinson, A.M. 1991. *The Scientist's Handbook of Writing Papers and Dissertations*, Prentice-Hall, New Jersey.

Williams, E.J. 1994. Downtime Data -- its Collection, Analysis, and Importance. In: *Proceedings of the 1994 Winter Simulation Conference,* eds. J.D. Tew, M.S. Manivannan, D.A. Sadowski, and A.F. Seila, 1040-1043.

Williams, E.J. 1996. Making Simulation a Corporate Norm. In: *Proceedings of the 1996 Summer Computer Simulation Conference*, eds. V.W. Ingalls, J. Cynamon and A.V. Saylor, 627-632.

Yin, R.K. 1994. *Case study research: design and methods* (2nd ed.), SAGE Publications, Thousand Oaks.

Zaum, D., M. Olbrich and E. Barke. 2008. Automatic data extraction: A prerequisite for productivity measurement. In: *Proceedings of the IEEE International Engineering Management Conference, Europe: Managing Engineering, Technology and Innovation for Growth*.

APPENDED MATERIALS

This part of the thesis includes questions and topics used in interviews and questionnaires for data collection in Publications I, II and III. In some cases, the materials are translated from the original format (in Swedish) to English.

Topics used for the semi-structured interviews in Publications I and II

1. Please draw an outline of your applied work procedure during input data management in your DES project (using white-board or pen and paper depending on the meeting location).
    a. Please describe the data input activities performed in your DES project?
    b. Try to use as general terminology as possible.
    c. In what order did you perform the activities?
    d. If possible, please draw the relations between activities to form a flow diagram.
    e. Was there a need for iterations of activities during your work-procedure?
2. Can you identify possible improvements of your applied work-procedure?
3. What would you change in the work procedure if it were possible to do the project over again?
4. Based on the experience from your project, what is most important in order to increase efficiency in input data management?
5. Do you think that the input data management phase in your project would have been more rapid if a structured methodology was applied? Please motivate.

Questions included in the face-to-face questionnaire for Publication II

1. How many people are employed at the company where the DES project was completed?
2. How frequently does the company use DES for improvements of production flows?
    a. Never before, this was the first project.
    b. Sporadically, in few improvement projects.
    c. In major change projects.
    d. In almost all improvement projects.
    e. Sporadically in order to identify improvement possibilities.
    f. Continuously in order to identify improvement possibilities.
    g. On a daily basis for planning and control purposes.
3. How much time did the entire project consume?
4. How much time was planned for input data management according to the project plan?
5. How much time did each data input activity consume (respond in man-hours)?
6. Which input parameters were included in your model, from what source did you collect them, and how much time did you spend on collecting each specific parameter?
7. If any, please list the software solutions or other tools used for data processing?
8. Were you able to separately validate the collected data?
9. Were you able to validate the simulation model according to the real-world system?
10. Did the input data management step follow the time-consumption estimated in the project plan?
11. Did the complete project finish on time according to the project plan?
12. Was the project considered successful based on the objectives in the project plan?

Questions included in the questionnaire for Publication III

1. Please specify your major areas of application for DES, e.g. manufacturing, logistics, health care, and military.
2. What makes you to use simulation in your business? (please select the most appropriate answer)
    a. Simulation is used to address a specific business need such as design of a new factory. Model is not re-used once the project is completed.
    b. Simulation is regularly used to improve business operations. Models are often re-used.
    c. Use of simulation is mandatory within the business in every improvement project.
3. Do you apply a structured approach to input data management (including raw data collection and data processing), such as data collection templates, guidelines and/or checklists?
    a. Yes, please specify.
    b. No.
4. Which is the main source of input data to DES models? (please select one answer)
    a. Manual gathering (e.g. stop watch, movie recording)
    b. People-based systems (e.g. interviews, expert knowledge)
    c. Paper-based systems (brochures etc.)
    d. Local computer-based systems (e.g. spreadsheets)
    e. Computer-based corporate business systems (e.g. ERP, MES, PLM)
    f. Other, please specify.
5. Which sources of input data are commonly used? (several alternatives are allowed)
    a. Manual gathering (e.g. stop watch, movie recording)
    b. People-based systems (e.g. interviews, expert knowledge)
    c. Paper-based systems (brochures etc.)
    d. Local computer-based systems (e.g. spreadsheets)
    e. Computer-based corporate business systems (e.g. ERP, MES, PLM)
    f. Other, please specify.
6. What is your major approach for selection between duplicate data sources (if you have multiple sources for the same data item)? (please select the most appropriate answer)
    a. Data duplication is never encountered
    b. Select the most recent data
    c. Base the selection on personal experience
    d. Combination of data sources
    e. Base the selection on team knowledge
    f. Select data most local to the source/origin
    g. Other, please specify.
7. How are data accuracy, reliability and validity mainly assured? (please select the most appropriate answer)
    a. Interviewing area experts
    b. Basic “sanity” checks
    c. Personal experience
    d. The internal or external customer’s responsibility
    e. Model validation runs

   f. Other, please specify.
8. Models develop and evolve; how is data validity maintained? (please select the most appropriate answer)
   a. Continuous manual efforts for data collection
   b. Manual efforts for data collection, only initiated when the model will be used
   c. Automated collection for parts of the data
   d. Continuous automated collection of all necessary data
   e. Models are not maintained and reused
   f. Other, please specify.
9. How are data (information) supplied to the simulation model? (please select the most appropriate answer)
   a. Manually written in the model code
   b. Via an external spreadsheet (automatically connected to the model) or similar
   c. An off-line database automatically connected to the model
   d. Direct link between corporate business systems and simulation model
   e. Other, please specify.
10. Where is the majority of data (information) held, i.e. where does the processed data reside? (please select the most appropriate answer)
   a. In the simulation model
   b. In a paper-based system
   c. In a local computer-based system (e.g. a spreadsheet)
   d. A computer-based corporate business system (e.g. ERP, MES)
   e. Other, please specify.
11. Considering the entire input data management process, which is the most common methodology (Figure 10 was appended)? (please select the most appropriate answer)
   a. Methodology A
   b. Methodology B
   c. Methodology C
   d. Methodology D
   e. Why? Please describe benefits and problems.
12. Which methodology do you think will be used in ten years? (please select the most appropriate answer)
   a. Methodology A
   b. Methodology B
   c. Methodology C
   d. Methodology D
   e. Please explain your choice.