

University of Alberta

**An Integrated Framework for Managing Labour Resources Data in
Industrial Construction Projects:
A Knowledge Discovery in Data (KDD) Approach**

by

Ahmed Mohamed Hammad

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Construction Engineering and Management

Department of Civil and Environmental Engineering

©Ahmed Mohamed Hammad

Fall, 2009

Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

ABSTRACT

Improper management of labour resources causes major problems for companies working on multiple industrial construction projects. To address these problems, an integrated framework is developed based on a five-step knowledge discovery in data model. The framework transfers existing multidimensional historical data from completed projects into useful knowledge for future projects.

First, a synthesis of previous research is presented. Second, an inclusive analysis of the industrial construction domain is performed. Third, the concept of predefined progressable work packages is introduced to address issues in current data management practices. Fourth, a prototype data warehouse is built using the snowflake schema to centrally store the data, produce dynamic reports and exchange knowledge. Fifth, data mining techniques are applied to extract useful knowledge from three sets of real projects data.

Results show that the developed framework is capable of transferring previously unanalyzed data to valuable knowledge that significantly improves current resources management practices.

ACKNOWLEDGEMENTS

My utmost grace goes to Almighty God for all the gifts and blessings and for enlightening my path to complete this dissertation. My sincere appreciation goes to my supervisor Dr. Simaan AbouRizk for his continuous support, encouragement and inspiration. He has been an academic mentor and a friend who never ceased to provide me with the right advice since I first met him twelve years ago. He has been a great scholar to learn from in all aspects of life. I'd like also to thank my committee members: Dr. Yasser Mohamed, Dr. SangHyun Lee and Dr. Hooman Askari-Nasab for their valuable feedback; special thanks go to Dr. Peter Flynn and Dr. George Jergeas, whose wealth of experience in the industry brought new insights into this work. I'd like to extend my appreciation to Dr. Osmar Zaiane, who introduced me to the thrilling field of data mining and knowledge discovery in data and Dr. Ivan Ordev for helping me with programming and statistical analysis.

WorleyParsons Canada – Edmonton Division, Waiward Steel Fabricators and the Natural Sciences and Engineering Research Council (NSERC) - Alberta Construction Industry Research Chair provided the necessary data and funding. I thank all of them for supporting this research. Special thanks to Mr. Edim Cemic of WorleyParsons Canada, my career mentor and friend, for his guidance, encouragement and trust in my visions, abilities and ideas. My sincere thanks go to the industry experts who shared their knowledge with me from Bantrel, Chevron, Jacobs Engineering, KBR, PCL Industrial, SNC Lavalin, and Suncor.

I'd like also to thank all my friends who have been there for me when I needed them specially Dr. Ashraf ElAssaly and Dr. Abbas Khani-Hanjani. My thankfulness goes to Dr. Sallama Shaker, the ex Egyptian Ambassador to Canada and my father, both of them told me that it is never too late to fulfil my dreams.

My love and gratitude go to our daughter Selma who always cheered me up with her smile and to our parents, sisters and brothers, in Egypt, Morocco and Canada for their prayers, love and continuous enthusiasm. Special thanks to my mother Dr. Nabila Gabr for teaching me the love of learning even before I go to school and to my brother Khalid for his continuous encouragement and support.

There are no words that can express my true love and deep appreciation to my wife Dr. Samira ElAtia for her unconditional love, endless understanding and infinite support through our life journey. She is the one who has always been there for me, even before I asked for her help or needed her, the one who accepted to sacrifice a lot of our family time to make this dissertation a reality and the one who spent many hours with me editing the language.

TABLE OF CONTENTS

Chapter 1: Introduction.....	1
1.1 Background.....	1
1.2 Problem Identification	3
1.3 Research Objectives.....	6
1.4 Knowledge Discovery in Data (KDD).....	8
1.5 Research Methodology	11
1.6 Thesis Organization	14
Chapter 2: Literature Review	16
2.1 Introduction.....	16
2.2 Resources Management in Project Environment.....	17
2.2.1 Heuristic Approaches.....	18
2.2.2 Numerical Optimization Approaches	21
2.2.3 Genetic Algorithms (GA) Approaches	23
2.3 Forecasting Techniques in Construction Projects.....	27
2.4 Transferring Projects Data to Useful Knowledge.....	29
2.5 Data Warehousing and On Line Analytical Processing (OLAP) Techniques	36
2.5.1 Multidimensional Data Warehousing	36
2.5.2 On Line Analytical Processing (OLAP) Techniques.....	40
2.6 Data Mining Methods and Techniques.....	42

2.6.1	Data Mining Techniques Based on Unsupervised Learning.....	44
2.6.2	Data Mining Techniques Based on Supervised Learning.....	48
2.6.3	Outliers Detection	53
2.6.4	Previous Applications of KDD and Data Mining Techniques in Construction.....	55
2.7	Conclusion	59
Chapter 3:	Industrial Construction Projects Domain	61
3.1	Industrial Construction Projects.....	61
3.2	The Vertical (Within-Project) Analysis	69
3.3	Horizontal (Cross-projects) Analysis.....	72
3.3.1	The Cross-projects Elements	72
3.3.2	Resource Management Structure in Contracting Companies	78
3.4	The Process Modeling.....	86
3.4.1	The Feasibility Study Process.....	86
3.4.2	Processes within the Engineering Phase	88
3.4.3	Processes within the Procurement Phase	101
3.4.4	Processes within the Construction Phase.....	107
Chapter 4:	Resources Management Data.....	111
4.1	Multidimensionality of Resources Data.....	111
4.2	Current Resources Management Practices.....	112
4.2.1	Scope Management Practices	114

4.2.2	Schedule Management Practices.....	117
4.2.3	Cost Management Practices.....	122
4.2.4	Performance Management Practices.....	127
4.2.5	Responsibility Management Practices	129
4.2.6	Summary analysis	131
4.3	Proposed Integrated Data Management Approach	134
4.3.1	The Concept of Predefined Progressable Work Packages.....	136
4.3.2	Benefits of using Predefined Progressable Work Packages	140
4.3.3	The Proposed Data Management Flow Chart.....	143
4.4	Implementation of the Proposed framework.....	146
Chapter 5: The Labour Resources Data Warehouse		156
5.1	Benefits of WarehouseING Resources Data.....	156
5.2	Building the Data Warehouse	158
5.2.1	The Multidimensional Data Model.....	158
5.3	The Snowflake Schema.....	161
5.4	The Data Warehouse Unput (Backend)	163
5.5	The Data Warehouse Output (Frontend).....	173
5.5.1	Populating the Data Warehouse.....	173
5.5.2	OLAP Reports.....	175
5.5.3	Utilizing the Slice and Dice OLAP Technique.....	177
5.5.4	Utilizing the Roll-up and Drill-down OLAP Technique	181
5.5.5	Utilizing the Pivoting OLAP Technique.....	183

5.6	The Knowledge Exchange Tool.....	190
5.7	System Implementation	Error! Bookmark not defined.
Chapter 6: Case Studies on Knowledge Discovery in Data		196
6.1	discovering Knowledge in the First dataset.....	196
6.1.1	Data Cleaning and Preprocessing	196
6.1.2	The Initial Investigation.....	202
6.1.3	The Outliers Detection Procedure.....	206
6.1.4	Clustering of Unit Cost using Statistical Methods.....	210
6.2	discovering Knowledge in the Second dataset.....	217
6.2.1	Data Gathering, Cleaning and Preprocessing	217
6.2.2	Clustering of the Cost and Duration Units.....	220
6.2.3	Case Study Results Validation.....	229
6.3	Discovering Knowledge in the Third Dataset.....	231
6.3.1	Data Gathering, Cleaning and Preprocessing	231
Chapter 7: Conclusion		241
7.1	Research Summary	241
7.2	Research Contributions.....	246
7.2.1	Academic Contribution.....	246
7.2.2	Industry Contribution.....	249
7.3	Recommendations for Future Research.....	250
Bibliography		251

LIST OF TABLES

Table 5.1: Example of the Data Set	174
Table 6.1: The Raw Data for the First Data Set.....	197
Table 6.2: The Dataset after Cleaning and Pre-Processing.....	199
Table 6.3: The Descriptive Data Test	204
Table 6.4: The Output from the Outlier Detection Tool.....	207
Table 6.5: Univariate ANOVA Test Results for the Three Main Attributes	211
Table 6.6: Post Hoc Test Results for the Three Main Attributes.....	212
Table 6.7: Duncan Test Results	213
Table 6.8: Statistical Analysis for the Eight Data Clusters.....	213
Table 6.9: Raw Dataset for the Second Analysis.....	219
Table 6.10: Calculating the Total Fabrication Duration	219
Table 6.11: An Excerpt of the CSV Data File for the Fabrication Phase	222
Table 6.11: Clustering of Fabrication Hourly Unit Cost	225
Table 6.12: Clustering of Fabrication Weekly Unit Duration	226
Table 6.13: Clustering of Shop Drawings Hourly Unit Cost.....	226
Table 6.14: Clustering of Shop Drawings Weekly Unit Duration.....	227
Table 6.15: An Excerpt of the Validation Tool for the Fabrication Phase	230
Table 6.16: List of Completed Projects between 2004 and 2007	232
Table 6.17: Weekly Actual Working Hours per Resource	233
Table 6.18: The Normalized Dataset	235
Table 6.19: The Normalized Dataset after Interpolation	236

Table 6.20: Clustering of Total Resource Hours	237
Table 6.21: Clustering of Total Duration Weeks.....	237
Table 6.22: An Example of the Coefficients Output	239

LIST OF FIGURES

Figure 1.1: Example of the Variance between Planned versus Actual Resource ...	5
Figure 1.2: Current Resource Management Data Practices	7
Figure 1.3: Proposed Resource Management Data Cycle	7
Figure 1.4: Research Methodology Using the Modified Hybrid KDD Model	14
Figure 2.1: The Knowledge Pyramid.....	31
Figure 2.2: The Knowledge Cycle	32
Figure 3.1: The Vertical Hierarchy of the Industrial Construction Market	69
Figure 3.2: A Typical Project Lifecycle broken into Stages.....	72
Figure 3.3: Industrial Construction Project Lifecycle broken into Phases	74
Figure 3.4: The Predefined Resources Breakdown Structure (RBS).....	76
Figure 3.5: Project States in Multiple-Project Environment.....	78
Figure 3.6: Workload vs. Capacity in a Multiple-Project Environment	81
Figure 3.7: Matrix Organization Structure Contracting Companies.....	84
Figure 3.8: The Initiation Process	87
Figure 3.9: Uncertainty and Ability to Influence Projects vs. Project Expenses ..	90
Figure 3.10: The FEL I Process	91
Figure 3.11: The FEL II Process.....	92
Figure 3.12: The FEL III Process	94
Figure 3.13: The Detailed Engineering & Design (DED) Process	95
Figure 3.14: The Shop Drawings Process.....	97
Figure 3.15: The Procurement Support (PS) Process	98
Figure 3.16: The Construction Support Process	99

Figure 3.17: The As-Building Process.....	100
Figure 3.18: The Engineering Support Process in Procurement Phase	103
Figure 3.19: The Requisitioning, Bidding & Awarding Process	104
Figure 3.20: The Contract Administration Process.....	105
Figure 3.21: The Materials Management Process.....	106
Figure 3.22: The Engineering Support Process in Construction Phase	107
Figure 3.23: The Fabrication Process	108
Figure 3.24: The Assembly Process.....	109
Figure 3.25: The Site Installation Process	110
Figure 4.1: The Five Dimensions of Labour Resources Data.....	137
Figure 4.2: Types of Progressable Work Packages	139
Figure 4.3: Flowchart of the Pre-planning Stage	144
Figure 4.4: Flowchart of the Planning and Execution Stages.....	145
Figure 4.5: Transferring the Schedule into Crosstab Report	149
Figure 4.6: The Actual Cost Collection System	150
Figure 4.7: Resource Baseline Histograms.....	152
Figure 4.8: Plotting Resource Values	152
Figure 4.9: Plotting Cumulative Values per Weeks.....	153
Figure 4.10: The Enhanced Workload vs. Capacity Graph	154
Figure 5.1: The Multidimensional Data Model	159
Figure 5.2-a: The Industrial Owners and Construction Projects Schema.....	161
Figure 5.2-b: The Contractors and Internal Projects Schema	162
Figure 5.2-c: The Work Package Schema.....	163

Figure 5.3: Industrial Owners (IO) Definition Screen	165
Figure 5.4: Contractors (PPF) Definition Screen.....	165
Figure 5.5: Industrial Construction Project Definition Screen	166
Figure 5.6-a: The Internal Project Definition Screen.....	167
Figure 5.6-b: The Definition Screen for Internal Project Phases.....	167
Figure 5.7-a: Scope Definition Screen for Work Packages	168
Figure 5.7-b: Resource Data Entry Screen per Work Package	168
Figure 5.7-c: Weekly Progress per Work Package	169
Figure 5.7-d: Progress Activities per Work Package.....	169
Figure 5.7-e: Weekly Hours per Individual	169
Figure 5.8: Definition Screen for Locations	170
Figure 5.9: Definition Screen for Industries	170
Figure 5.10: Definition Screen for Project Phases.....	171
Figure 5.11-a: Data Entry Screen for the Contractor’s Staff	172
Figure 5.11-b: Responsibility Data Entry Screen for the Contractors’ Staff.....	172
Figure 5.12: Example of the Summary Internal Project Hours at Completion...	174
Figure 5.13: Example of the Detailed Internal Project Hours at Completion.....	175
Figure 5.14: Detailed Single-Project Multiple-Resources Report	178
Figure 5.15: Summarized Single-Project Multiple-Resource Report	179
Figure 5.16: Detailed Single-Resource Multiple-Projects Report	180
Figure 5.17: Summarized Single-Resource Multiple-Projects Report	180
Figure 5.18: Summarized Report Grouped by Internal Program.....	182
Figure 5.19: Summarized Report Grouped by Industrial Portfolio	182

Figure 5.20: The Pivoting Structure for the Three Main Data Elements.....	184
Figure 5.21: Pivoting Report Grouped by Year, Quarter and Month.....	184
Figure 5.22: Dynamic Pivot Graph for Current Planned Values (CPV).....	185
Figure 5.23: Dynamic Pivot Graph for Actual Earned Values (AEV).....	185
Figure 5.24: Summary of the Three Main Data Elements over Years.....	186
Figure 5.25: The Three Main Data Elements Grouped by Phase and Resource.	187
Figure 5.26: Resource Contribution to ASV per Year in Column Format.....	187
Figure 5.27: Resource Contribution to ASV per Year in Pie-chart Format.....	188
Figure 5.28: Weekly Resources Utilization per Internal Project.....	189
Figure 5.29: Weekly Resource Utilization per Reporting Period.....	189
Figure 5.30: The Knowledge Definition Screen.....	191
Figure 5.31: The Screen for Exchanging Tacit Knowledge.....	192
Figure 5.32: The Screen for Exchanging Explicit Knowledge.....	192
Figure 5.33: The Output of Finding Knowledge Elements.....	193
Figure 6.1: Frequency of Data Points within the Three Phases.....	202
Figure 6.2: Frequency of Data Points within the Five Resources.....	203
Figure 6.3: Boxplot of the Unit Cost Showing Outliers per Package.....	205
Figure 6.4: Boxplot of the Unit Cost Showing Outliers per Phase.....	205
Figure 6.5: The Decrease in Standard Deviations of the Data Classes.....	208
Figure 6.6: The Decrease in Ranges of the Data Classes.....	208
Figure 6.7: The Output Summary Tree.....	209
Figure 6.8: Fitting Distribution to a Class of Data.....	210
Figure 6.9: BoxPlots for the Eight Data Clusters.....	214

Figure 6.10: The Final Dataset with the Select and Cluster Variables	215
Figure 6.11: The Descriptive Analysis Screen of Weka.....	223
Figure 6.12: The Visualizing Capabilities of Weka.....	224
Figure 6.13: Weka Results Viewer	228
Figure 6.14: The Frequency Histogram for the Obtained Clusters.....	228
Figure 6.15: Example of Plotting the Polynomial Function vs. the Original Data	239
Figure 6.16: Connecting the Average Points to Represent the Data Class.....	240

LIST OF ABBREVIATIONS

AAC	Approved Project Changes
AEV	Actual Earned Values
AFE	Appropriation For Expenditure
AHP	Analytical Hierarchy Process
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ASV	Actual Spent Values
ATR	Auto Regression Tree
AV	Actual Values
BI	Business Intelligence
BU	Business Unit
C&SU	Commissioning & Start-up
CBR	Case Based Reasoning
CBS	Cost Breakdown Structure
CF	Current Float
CII	Construction Industry Institute
CoA	Code of Accounts
CPCA	Cluster Principle Component Analysis
CPI	Cost Performance Index
CPM	Critical Path Method
CPV	Current Planned Values
CWP	Contractual Work Package
DBM	Design Basis Memorandum
DED	Detailed Engineering & Design
DoD	Department of Defense
DSS	Decision Support Systems
EAC	Estimate At Completion
EBAC	Enumerative Branch-And-Cut
EDS	Engineering Design Specifications
EM	Expectation Maximization
ENR	Engineering News Records
EPC	Engineering, Procurement and Construction Engineering, Procurement and Construction
EPCM	Management
ETC	Estimate To Complete
EV	Earned Values
EVM	Earned Value Management

EWP	Engineering Work Packages
FAT	Factory Acceptance Tests
FEL	Front End Loading
FF	Finish-to-Finish
FWP	Fabrication Work Packages
GA	Genetic Algorithms
GC	General Contractor
GDP	Gross Domestic Product
GIS	Geographical Information System
GO	Governmental Organizations
HAZOP	Hazardous Operations
HLA	High Level Architecture
HMB	Heat and Material Balances
HRMS	Human Resources Management System
IC	Industrial Components
ICP	Industrial Construction Projects
IFF	Issue-For-Fabrication
IGP	Integer Goal Programming
IO	Industrial Owners
IP	Internal Project
IPV	Initial Planned Values
JV	Joint Venture
KDD	Knowledge Discovery in Data
LCVA	Life Cycle Value Analysis
LDT	Line Designation Tables
MCD	Minimum Covariance Determinant
ML	Most Likely
MMS	Material Management System
MVE	Minimum Volume Ellipsoid
NASA	National Aeronautics and Space Administration
NGO	Non Governmental Organizations
OBS	Organization Breakdown structure
OLAP	On Line Analytical Processing
OLTP	On Line Transactional Processing
OPV	Original Planned Values
OSS	Operation Support Systems
P&ID	Process and Instrumentation Diagrams
PCM	Project Controls Manager
PDW	Project-oriented Data Warehouse
PEP	Project Execution Plans
PFD	Process Flow Diagrams

PKM	Project Knowledge Manager
PM	Project Manager
PMI	Project Management Institute
PMT	Project Management Teams
PO	Purchase Order
PPWP	Predefined Progressable Work Packages
PV	Planned Values
PWP	Purchasing Work Packages
QA	Quality Assurance
QC	Quality Control
RAS	Required At Site
RBS	Resources Breakdown structure
RFI	Requests for Information
ROI	Return on Investment
RSE	Relative Squared Error
SAT	Site Acceptance Tests
SLD	Single Line Diagrams
SME	Semantic Modeling Engine
SoC	Suppliers of Choice
SPI	Schedule Performance Index
SPS	Special Purpose Simulation
SS	Scoping Study
SS	Start-to-Start
SVM	Support Vector Machine
SWP	Site Work Packages
TIC	Total Installed Cost
VE	Value Engineering
VSM	Value Stream Map
WBS	Work Breakdown Structure
WP	Work Packages

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND

Construction sector is fundamental to any national economy. It is a major contributor to any country's Gross Domestic Product (GDP) and an indicator of its economy's prosperity. Construction sector includes: commercial, residential, infrastructure, and industrial type projects. Manufacturing, chemical processing, oil production, refineries and electrical power plants are examples of industrial construction projects. In Canada, more than \$230 billion was invested in industrial construction in 2007 (Statistics Canada, 2008) and a rapid growth of 34% in the capital spending was recorded (Industrial Reports Inc., 2008).

Industrial construction projects share many similarities with other construction projects; however, they also have characteristics that are specific to them. Due to the specific nature of the final product, this type of construction projects is known for being more complicated, utilizing more sophisticated management tools and paying more attention to safety and environmental concerns.

Industrial construction projects involve large number of stakeholders with different, sometimes conflicting, interests. Stakeholders include owners, Project Management Teams (PMT) engineers, suppliers, fabricators, constructors, environmental and other governmental agencies, plant operators and maintainers and general public.

These projects typically start without a complete scope definition and the “rolling wave” planning technique is used to develop the project’s scope and detailed plans as an iterative process while projects progress. When the economy is fast-paced, most of these projects are fast-tracked, i.e. activities within the project take place concurrently not sequentially.

Nearly all industrial construction projects are performed as a set of smaller projects, each of which is performed by a contractor. These smaller projects are referred to as “internal projects” in this research. These contractors include Engineering, Procurement and Construction Management (EPCM) offices, fabrication shops and module assembly yards. Contrary to construction sites, which are temporary set-ups, these contractors try to maintain their workforce in order to be able to compete for new projects.

These contractors are producing various services and rely solely on a continuous supply of projects to generate their revenues. Contractors utilize different types of resources to produce their final products. These resources can be classified as capital, materials, equipment, labour, facilities and information (Kerzner, 2006). In this research, the term “resources” is used to refer only to labour resources in contracting companies. Construction Industry Institute (CII, 1987) identified labour resources issues as one of the major reasons for cost overruns in industrial construction.

1.2 PROBLEM IDENTIFICATION

As a result of the previously mentioned challenges, many of the recently completed mega industrial projects faced considerable scope creep, significant schedule delays, and severe budget overruns. In Alberta, mega oil sands projects reported \$7.3 billion dollars overruns within three years and none of them was completed on schedule (Alberta Economic Development Authority, 2004). Jergeas and Ruwanpura (2008) found out that unrealistic cost and schedule baselines accompanied with lack of complete scope definition are some of the main factors driving the cost overruns in Alberta oil sands mega projects. With the cost of lost production, each day in schedule delay represents a huge hidden overrun and a major loss to the project owner.

One of the major causes of schedule delays and budget overruns in industrial construction is the improper management of labour resources (Jergeas, 2008). Labour resources is the most difficult to manage due to the human factor. It also leads to loss of profit for industrial owners and contractors, decreased client satisfaction, inability to compete in the market and damaged reputation of the industry. It also generates intolerable levels of stress for team members who always feel incompetent and incapable of achieving success in their projects. Contrary to materials, labour resources cannot be bought instantly when needed. The required hours to complete a task are uncertain because of issues with productivity. Labour productivity is impacted by the learning curve, fatigue, boredom, team harmony, team leadership skills, weather and many other factors.

Contractors manage multiple projects in a changing environment using the same pool of resources (Tharachai, 2004). According to Huemann et al. (2007), the number and the sizes of the projects are constantly changing in this environment. The supply of projects is dependent on market conditions, which are difficult to predict. This dependency causes significant uncertainty and makes it very difficult to estimate the required amount of hours to complete the expected projects (workload) and the necessary resources to perform this workload (capacity).

Some contractors try to utilize commercially available software such as Primavera, Excel, Access or MS Project to forecast their expected workload and future capacity. These applications usually don't consider the high degree of uncertainty in projects, are not originally designed to manage multiple projects and do not utilize historical records from previous projects. An example from a real dataset shows that the difference between estimated and actual resource hours in a single project exceeded 260 hours per week as shown in Figure 1.1. These variances when aggregated have severe impacts on any contractor.

A large amount of resources data is generated, collected, and stored in different formats during planning and executing projects. It is a huge loss for any contractor that the collected data is hardly analyzed and is not transferred to useful knowledge to improve resource management practices.

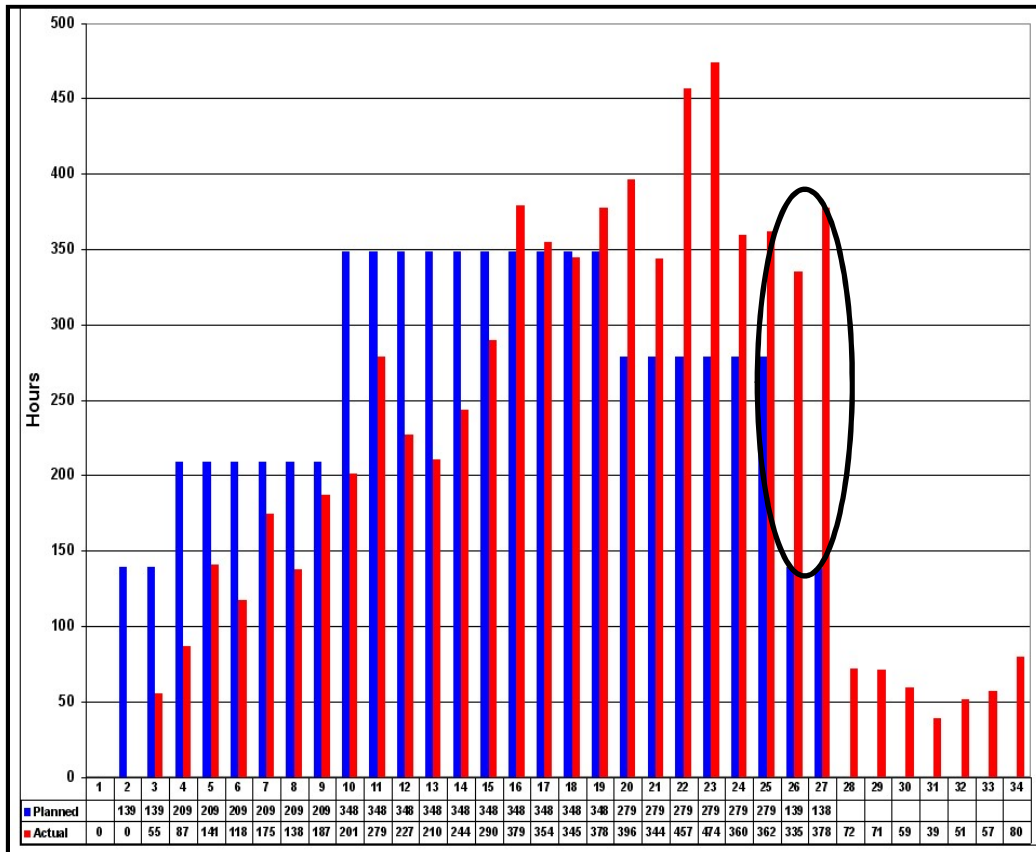


Figure 1.1: Example of the Variance between Planned versus Actual Resource

Given the complexity of this context, a major resource management problem stands out. Previous research has not addressed this problem in an integrated and applicable approach. Most research on labour resources management focused on resources leveling and allocation not on transferring knowledge from previously completed projects to future projects. They also focused on studying projects independently from each other, not only managing multiple projects with one common pool of resources. With the volatile market condition, it is necessary that the issue of improving labour resources management practices by learning from previous data be addressed.

1.3 RESEARCH OBJECTIVES

Two research questions are raised in order to address the problems identified above:

1. Regardless of the uniqueness of each project, would it be feasible to develop an integrated data acquisition system for collecting and storing resource management data from all projects?
2. Can this collected data be transferred to useful knowledge for providing more realistic estimates of future resource requirements?

The main objective of this research is to develop an integrated framework for managing labour resources data in the multiple-project environment of industrial construction projects. The main purpose of the framework is to develop a closed knowledge cycle where resource management data from completed projects is generated, collected and then utilized for better estimating of resource requirements in new projects. Better forecasting of resource requirements in the future enables contractors to run different scenarios to predict their optimum capacity.

In the current practices, cost and schedule baselines are generated during the planning stage of projects. These two baselines are combined to form the resources baseline that represents the planning portion of labour resources data for any project. During the execution stage, data regarding project changes, actual durations and resource utilization is also obtained.

When projects are completed, almost all this data is stored away without being looked at, as shown in Figure 1.2, and few efforts are spent to analyze this data and

transfer it to useful knowledge. This research aims to introduce a framework that closes the cycle and allows for the proper generation, collection and storage of labour resources data and transfer this data to useful knowledge that is fed back into future projects as shown in Figure 1.3.

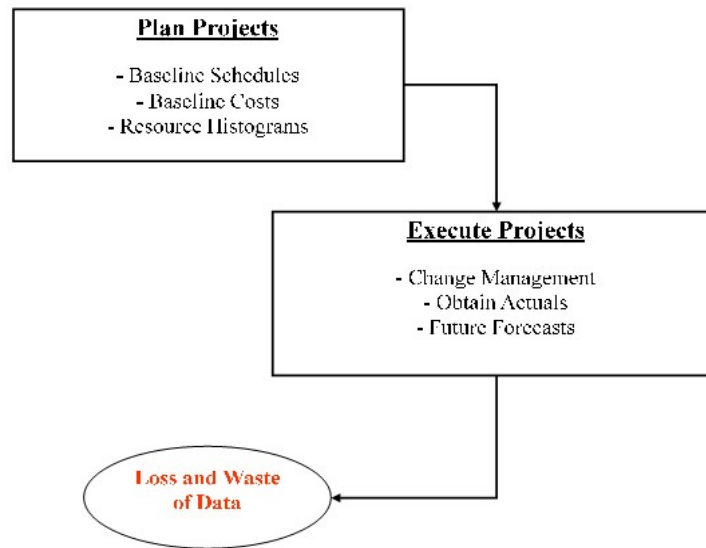


Figure 1.2: Current Resource Management Data Practices

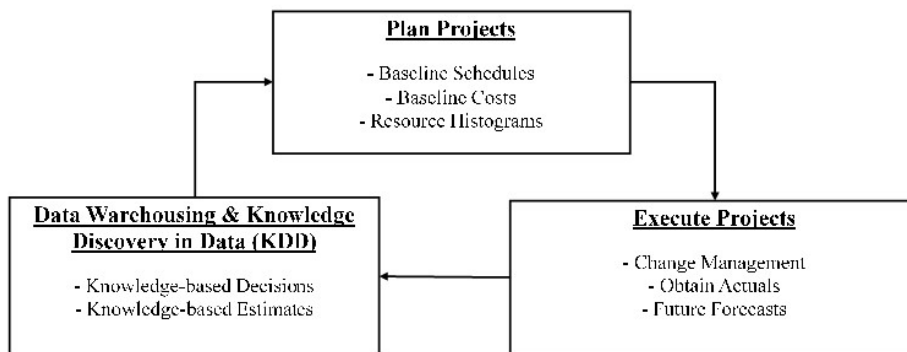


Figure 1.3: Proposed Resource Management Data Cycle

1.4 KNOWLEDGE DISCOVERY IN DATA (KDD)

The methodology to solve the research problem and achieve the research objective is to implement a Knowledge Discovery in Data (KDD) model to develop the integrated framework for managing labour resources data in industrial construction projects. The KDD approach is selected because it provides a complete, integrated, and self-learning solution to solve problems. The approach is data-oriented and provides powerful tools to learning from historical data. The KDD procedure combines knowledge from machine learning, statistical analysis and artificial intelligence fields to find hidden knowledge in large sets of data. It is an iterative process that utilises data warehousing and data mining in a complete procedure to ensure proper discovery and presentation of useful knowledge to decision-makers.

KDD combines the quantitative and qualitative research approaches together. In quantitative research, the focus is on the systematic collection and quantification of research data into understandable paradigms (Olson, 1995). The qualitative research focuses on observing, underlying and finding theory in the problem under investigation (Creswell, 2007). Jergeas (2008) mentioned that the main difference between the two approaches is that the quantitative research work with more data and few variables, in the meantime qualitative research rely on less data and lot more variables. KDD allows working with large amounts of data and large number of variables as well.

There are different models that can be implemented to define the steps of the KDD procedure. These models can be categorized as: academic, industrial and hybrid (Cios, 2007). Fayyad et al. (1996) introduced an academic model that consists of nine steps. These steps are:

1. Developing and understanding the application domain
2. Creating a target data set
3. Data cleaning and pre-processing
4. Data reduction and projection
5. Choosing the data mining task
6. Selecting the data mining algorithm
7. Data mining
8. Interpreting the results
9. Consolidating the discovered knowledge

Han and Kamber (2006) introduced another academic model consisting of seven steps. These steps are:

1. Data cleaning (to remove noise),
2. Integration from multiple sources,
3. Selection (only data required for the analysis)
4. Transformation (perform summarizing or aggregating operations) and
5. Mining (extract data patterns),
6. Pattern evaluation, and
7. Knowledge presentation to decision makers.

A large group of European companies developed CRISP-DM, which has become a primary industrial model (Cios, 2007). The model consists of six steps. These steps are:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Cios et al. (2007) developed a hybrid model by modifying the CRISP-DM to fit for academic purposes. Their model is more research-oriented, replaces the modeling step with a data mining step and consists of six steps. These steps are:

1. Understanding of the problem domain
2. Understanding of the target data
3. Preparation of the dataset
4. Data mining
5. Evaluation of the discovered knowledge
6. Use of the discovered knowledge

Each of the model steps is interactive; and cycles can take place between every two steps until satisfying results are achieved. The model is also capable of fulfilling both industrial and academic requirements.

1.5 RESEARCH METHODOLOGY

The hybrid model of Cios et al. (2007) addresses both academic and industrial requirements; and as such, it represents the most fitting KDD model for the research problem. This KDD model was adapted in this research to start with a comprehensive literature review. The literature review covers the five main topics that are covered in this research. These topics are: resource management practices focusing on multiple-project environment, forecasting techniques in construction projects, transferring projects' data to useful knowledge, data warehousing techniques and data mining methods.

The second step in the methodology is the understanding of the problem domain. The problem domain in this research is industrial construction. To fully understand industrial construction, three analyses were performed. The first is a within project analysis to detect main elements that impact resources management. The second is a cross-project analysis to determine the elements that can be utilized by all projects. The third is an analysis of all processes that take place during the procedure of managing industrial construction projects.

These analyses were performed through monitoring a set of various industrial projects over a long period of time and performing structured interviews with a group of industry experts from both owners and contractors. The output of these analyses includes a complete definition of the seven main objects that have to be modelled. For each of these objects, a set of control attributes is defined to be used

for data mining and knowledge discovery. Also another set of cross-project elements are defined such as project phases, stages and resources.

The third step in the methodology is the understanding of the problem data. In this research, resources data represents the target dataset. This dataset is multidimensional with five main dimensions. These dimensions are: scope, responsibility, schedule, cost and performance. Within each of these dimensions, a complete analysis of current industry practices and issues with these practices is performed. Based on this analysis, a data management concept is developed to overcome the issues with current practices and introduce consistency and integrity to data management practices. The proposed concept relies on using predefined progressable work packages in order to plan and execute industrial projects. In order to estimate resources needs for each package, data mining techniques are used to obtain cost and duration units. Resource utilization graphs are also obtained using data mining to be used for estimating weekly resource needs. The data elements within each dimension are also clearly defined in order to be collected and stored for mining purposes.

The next step in the research methodology is the development of a prototype data warehouse. The data warehouse stores collected data and produces dynamic On Line Analytical Processing (OLAP) reports. Due to the complexity and multidimensionality of the research problem, several attempts took place in order to obtain the proper design of the data warehouse that is capable of addressing the

needs of the various end-users. The sophisticated snowflake schema is found to be the most suitable to design the prototype and is used to represent the hierarchical nature of the data dimensions.

In order to validate the applicability of the research model and to test its ability to extract useful knowledge from real projects data, three case studies were conducted three different sets of real projects' data. The first case study applied supervised data mining technique to analyze the cost units' data in a large EPCM firm. The output of the study showed the value of obtaining cost units from previously complete projects, highlighted the problems with exiting data and provided recommendations to solve these problems.

The second case study applied unsupervised data mining technique to analyze the duration units' data in a large structural steel contractor. The output of the study also showed the value of obtaining duration units from previously complete projects, highlighted the problems with exiting data and provided recommendations to solve these problems.

The third case study applied supervised data mining technique to classify resource utilization graphs from a large EPCM firm into groups . The output of the study also showed the value of obtaining duration units from previously complete projects, highlighted the problems with exiting data and provided recommendations to solve these problems.

The modified model is illustrated in Figure 1.4.

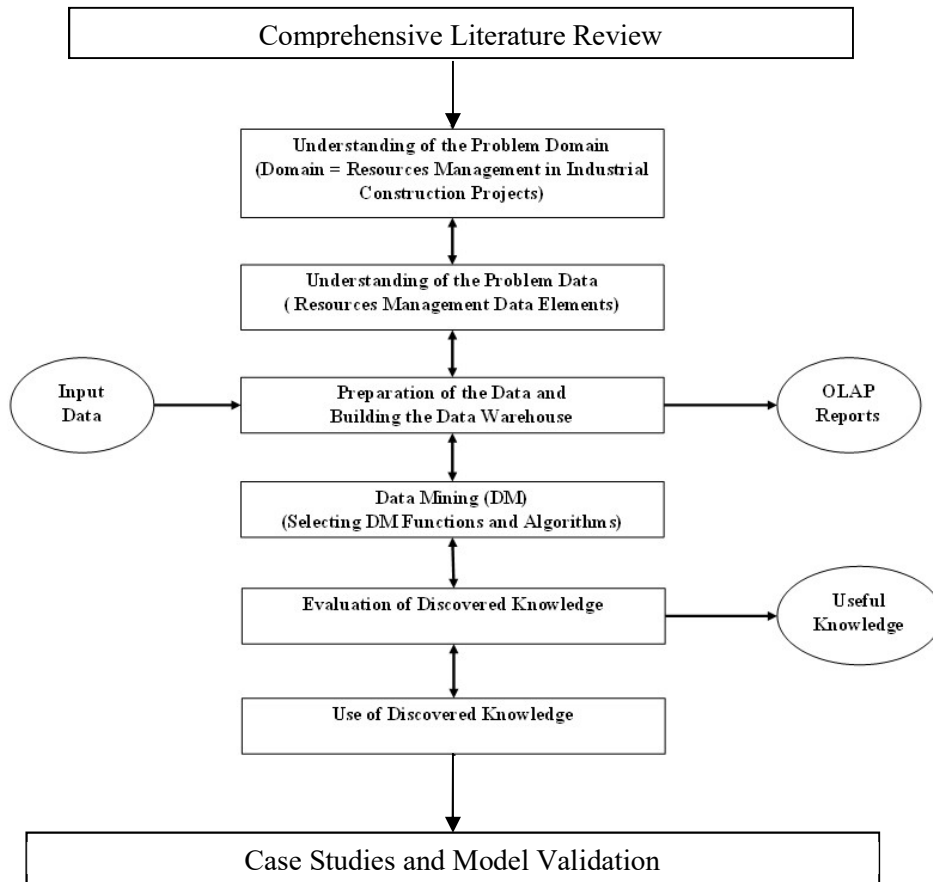


Figure 1.4: Research Methodology Using the Modified Hybrid KDD Model

1.6 THESIS ORGANIZATION

Chapter one of this thesis is an introduction that presents the problem identification, research objectives and the research methodology.

Chapter two of this thesis represents step number one of the hybrid KDD model. The chapter is a comprehensive literature review of the related topics to management of labour resources in multiple-project environment.

Chapter three of this thesis represents step number two of the hybrid KDD model. The chapter includes an inclusive analysis of the industrial construction processes, their input and output. The chapter also introduces an analysis of the industrial construction projects domain to define its main objects.

Chapter four of this thesis represents step number three of the hybrid KDD model. The chapter contains a comprehensive analysis of the current resource management data generation practices, the current issues with these practices and an integrated approach for managing resources data.

Chapter five of this thesis represents step number four of the hybrid KDD model. The chapter presents the data warehouse, its snowflake schema and examples of using the OLAP reports in presenting the stored data. The chapter also introduces the use of data warehouse as a knowledge exchange tool.

Chapter six of this thesis represents three case studies that cover steps number: five, six and seven of the hybrid model. The chapter shows how to implement data mining techniques to extract useful knowledge from three data sets that are obtained from real projects.

Chapter seven is a conclusion that presents a summary of the research, research contribution and set of recommendations for future research.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

The methodology, to achieve the objective stated above, is a hybrid KDD model adapted to fit this research. As shown in the previous chapter, the first step of this model, and covered by this chapter, is a comprehensive literature review. Given the sophistication and complex nature of the research problem, five fields are covered in the literature review. These five fields complement each other in this research and are essential in building the framework.

In this review the following research areas are covered. First, findings from previous research in the area of management of project labour resources are reviewed. Second, resource forecasting techniques in construction projects are studied. Third, previous studies that attempted to transfer collected projects' data to useful knowledge are investigated. Fourth, the concept of data warehousing and its differences from the traditional relational databases is discussed. Fifth, a summary of data mining methods and techniques and how they are used in discovering useful knowledge in the construction domain is presented.

This chapter represents a synopsis that synthesises findings from these five areas. With the multi-facets nature of this research, this synthesis is essential in defining the mutual contribution of each area to this research.

2.2 RESOURCES MANAGEMENT IN PROJECT ENVIRONMENT

The process of resources management starts during the planning stage of any project. The output of the planning stage includes: the scope of the project, the amounts of each required resource to complete this scope and their cost (baseline budget), the expected durations, start and finish dates of all the necessary activities to perform that scope (baseline schedule) and the loading of these resources to the required activities (baseline resource histograms). The baseline resource histograms are needed to know the required amount of each resource per time unit in order to prepare staffing plans to meet these requirements.

Traditional planning techniques such as Critical Path Method (CPM) assume unlimited availability of resources when needed. However, this is not realistic assumption, and consequently, several resource-constrained planning techniques were developed. In general, these techniques apply one of two methods: resource leveling (also called time-constrained scheduling) or resource allocation (also called resource-constrained scheduling). Resource leveling techniques assume unconstrained amount of resources in order to maintain the original estimated duration of projects and try to minimize the fluctuation of resource requirements between time periods. Resource allocation techniques assume constrained availability of resources, allocate these resources to project activities according to pre-defined rules and then calculate the modified duration of projects. In commercially available software applications, the two terms are used exchangeably.

It is estimated that more than 90% of projects take place in a multiple-projects environment (Payne, 1995). Payne also stated that one of the major problems in this environment is that the balance between resource requirements and availability is hardly achieved. It is very important to understand the previous research in the area of resources management in multiple-projects environment, find out the problems that are not solved yet and attempt to provide solutions to these problems. Some of the previous research in this area treated multiple projects as independent; others combined all activities from all projects in one single project by adding one artificial start and another artificial finish activity. Resource leveling and allocation techniques can be grouped into three main categories: heuristic rules, numerical optimization, and genetic algorithms.

2.2.1 Heuristic Approaches

Heuristic approaches primarily utilize pre-defined rules to find an acceptable solution to a problem. Heuristics approaches were used in the first attempts to address resource management problems and are still in use in most commercially available software applications. Most of these techniques try to solve the resource allocation problem by prioritizing the activities from all projects and then allocate the constrained resources to activities with highest priorities. Some of the developed applications prioritize the activities only once, others stochastically reevaluate the priority of each activity with the changes in network logic and resource availability.

The use of heuristics to allocate constrained resources to multiple projects started in the late sixties. Fendley (1968) identified eight rules to prioritize activities from multiple projects for resource allocation. These eight rules are: Most Available Resources (MAR), Most Critical Activities (MCA) Most Succeeding Activities (MSA), Modified Most Succeeding Activities (MMSA), Shortest Operation First (SOF), Minimum Slack First (MSF), Modified Minimum Slack First (MMSF) and First In First Out (FIFO). Fendley stated that the use of MSF rule provides best results to minimize project durations.

An attempt was made to categorize projects using summary measures to determine which heuristic rules perform better within each category of projects (Kurtulus and Davis, 1982). The first summary measure categorizes projects based on the peak of total resource requirements meanwhile the second measure relies on determining the utilization rate of each resource. They also suggested six new rules that can be utilized to solve the resource allocation problem in multiple projects. These rules are: Minimum Total Work Content (MINTWK), Maximum Operation First (MOF), Maximum Slack First, Maximum Total Work Content (MAXTWK), Shortest Activity from Shortest Project (SASP) and Longest Activity from Longest Project (LALP). They applied this approach to a set of test projects and concluded that the MSF rule seems to work best for certain categories of projects while the SASP rule provided better results for other categories according to their classification of projects.

Dumond (1992) analyzed the impact of different resource availability levels and resource allocation rules on project completion times and performance in a multiple-projects environment. The research found that when availability levels exceed 130%, the tested heuristic rules provide similar results. He also stated that when availability levels exceed 160%, activity durations are not shortened.

Another study stated that a dual-level resource management structure is typically evident in multiple-project environment in matrix organizations (Yang and Sum, 1993). The higher level decisions of assigning resources from the pool to projects are made by functional managers; meanwhile the lower level decisions of assigning resources to activities are made by project managers and their team leaders. The study also found that First in System First Served (FIFS) rule performed better than other tested rules in reducing project durations.

Another heuristic approach that combined resource and time constraints on a CPM network was called Resource Activity Critical Path Method - RACPM (Lu and Li, 2003). This approach considered three different states for each activity: TO-DO, CAN-DO and DONE. The work content, which is the total amount of required resources for an activity, was used to prioritize project activities in the schedule. Each activity was broken down to a set of single-resource parallel activities, which were plotted graphically in a resource/activity interaction scheme. Afterwards, both forward and backward calculations were performed under resource constraints to obtain minimum project duration.

2.2.2 Numerical Optimization Approaches

Mohanty and Siddiq (1989) stated that the management of multiple-resources in a multiple-project environment is a problem of combinatorial explosion nature. That means the number of alternative solutions increases factorially with the size of projects represented in number of activities and number of resources. They also stated that there is a conflict between completing projects in shortest possible duration and maintaining a fully utilized pool of resources. This conflict is significant as a result of the somewhat fixed number of resources in a contracting company and the random supply of workload in the form of projects. To solve these problems, they developed a multiple-objective Integer Goal Programming (IGP) model to simultaneously minimize the delay of projects, overutilization of resources and total cost of projects.

The model assumed that activities duration and resource requirements are known, fixed and can't be interrupted. The model was tested on three small projects with only three common resources. Mohanty and Siddiq's approach is not feasible in large projects because of the combinatorial problem. However, this approach introduced multiple-objective IGP to solve the resource management problem.

A dynamic programming application was developed to optimize resource allocation in repetitive construction projects (El-Rayes and Moselhi, 2001). The dynamic programming model utilized a scheduling algorithm to meet the constraints of project logic, crew availability and continuity of work flow to the crews. Another algorithm was used to generate a feasible set of activity interruption

vectors for each crew combination instead of obtaining the set from the users. The dynamic programming model was used to minimize the duration of a repetitive construction project through finding the optimum crew formation and interruption vector. The application provided better solution to a problem from the literature but was not tested on real projects.

Jiang and Shi (2005) introduced the Enumerative Branch-And-Cut procedure (EBAC) to minimize the total project duration under multiple resource constraints. The EBAC approach relies on starting from a root node and building a tree where better solutions to the problem are added and worse solutions are terminated. In order to reduce the number of branches, the branch-and-cut technique was used to eliminate the solutions that are possible but not worthy of keeping. A schedule calendar, starting at time zero and clicks each time an activity is ready to start, was used to determine which activities can start. Activities can only start if they meet the logical sequence needs and resource availability constraints. The approach was tested on 110 projects from the literature; each has between seven to fifty activities. Jiang and Shi were able to find the optimal solution, but the number of generated nodes and computing time was exponentially increasing as the number of project activities increased.

Vaziri et al. (2007) proposed a combination of simulation and optimization techniques to minimize project duration under constrained resource. Their model introduced uncertainty around activity durations due to resource availability and

expected productivity. The approach started by assigning a random set of resource multipliers to each task and used Monte Carlo simulation to calculate the total project duration and criticality index of each activity. A neighbour solution was then produced based on transferring more resources from the least critical activities to the most critical ones. The approach was tested on a 21 activities real project and was able to provide expected average duration and costs with less variance around the mean by optimizing resource utilization. Their approach, however, was implemented on a single project and its applicability and reliability were not tested in multiple-projects environment.

2.2.3 Genetic Algorithms (GA) Approaches

In the late 1990s, the concept of Genetic Algorithms (GA) was used to solve resource management problems in construction. GA is a computing technique that imitates the real-life evolution process in order to find approximate solutions to optimization problems. The procedure starts with generating a set of random solutions where each solution is a single string called chromosome. Each chromosome consists of a set of linear boxes that are called genes. Each gene is defined by its value and position in the chromosome. These solutions go through a cycle of generation, evaluation, selection and recombination based on the principle of “*survival of the fittest*” until the termination condition is reached. Survival of the fittest meant that the genes with more fitness to the evaluation criteria have higher chances of being selected for recombination. The most common techniques for recombination are crossover and mutation. In crossover, two parent genes are

selected randomly to exchange part of their genes to form two new chromosomes. The mutation procedure randomly changes the gene values in a chromosome and is used to introduce unexpected or random changes to ensure the diversity of the generated genes.

A GA model was adapted by Chan et al. (1996) to minimize the difference between needed and available resources to solve both resource allocation and leveling problems. They used a schedule-builder to represent different possible alternatives as chromosomes. These chromosomes utilized hard constraints (relationships between activities) and soft constraints (project duration and resource availability) to prevent the GA model from generating illogical schedules. They used the concept of Current Float (CF) to prioritize schedule activities. Although the model was able to generate multiple solutions with different resource profiles, it failed to single out the optimum solution. Moreover, the application was tested only using a low number of activities, fifty in total.

Another application of GA was developed by Hegazy (1999) to perform both resource leveling and allocation simultaneously. His approach assigned random priorities to project activities prior to utilizing commercial resource leveling application in order to find the shortest project duration. He used an objective function in order to minimize both resource periodical changes and total usage in the project. This use of the GA model reduced the estimated project duration and minimized the moment of the generated resource histograms. However, the process

was only implemented to a single critical resource and limited number of project activities.

Leu and Hung (2002) combined GA and Monte Carlo Simulation to introduce uncertainty of activity durations to the resource leveling problem. The model assumed that resource supply is unlimited. In this study, activity durations were generated from probability distributions. This model minimized the averaged resource leveling index, which represented the sum of absolute differences between actual and average resource utilization. Their GA model used the Roulette-Wheel principle to select the chromosomes with higher fitness values to be regenerated. Even though this approach introduced uncertainty to the project durations, it did not consider uncertainty in resource usages.

In Kandil and El-Rayes (2006), a parallel multiple-computer multiple-objective GA framework was introduced to optimize resource utilization in large construction projects in order to simultaneously minimize both project duration and cost. The process started by randomly generating acceptable resource usage options for each activity. Then, the total project cost and duration for each of the generated options was calculated to be used for fitness evaluation. The third step was the implementation of the typical GA operations of crossover and mutations to generate the next set of options. The final two steps were repeated until the termination criteria were met. Kandil and El-Rayes used the global and coarse-grained parallel computing methods to distribute the GA calculations between multiple computers

in order to solve a 720 activities project. After several trials, the computing time was reduced from 137 hours using 50 connected computers to only 7 hours using 10 computers. This approach tried to solve real-life problem, but it requires a sophisticated knowledge of GA optimization and parallel computing. Unfortunately, this knowledge is not usually available among typical project management teams.

GA was also used for resource allocation and leveling in linear projects (Georgy, 2007). The objective functions included the minimization of the day-to-day fluctuation of resource utilization and the absolute from the average total resource availability. The model was tested on a nine-activity single-resource project from the literature using an initial set of 10,000 random-generated schedules. After that a corresponding resource profile is generated for each solution. The GA module is then used to find the optimum resource profile using the Roulette-wheel selection and single-point crossover. The model outperformed the previous solution and provided a more-leveled resource profile. However, the model is limited to the leveling of a single resource and assumes constant resource utilization during the execution of each activity.

2.3 FORECASTING TECHNIQUES IN CONSTRUCTION PROJECTS

In proper project management, it is very important to compare the baseline values with the most likely forecasted values in order to detect variances as early as possible (Nassar, 2004). This also applies to comparing the properly estimated baseline resource requirements to more realistic forecasts of the future resource requirements. Forecasting project performance is one of the most challenging tasks in project management according to Nasira and Abd.Majid (2006). According to Nassar (2004), the forecasting technique has to be able to accept judgemental feedback, unbiased, timely, stable, simple enough to be used by the project team and sophisticated enough to provide reliable results. Most of the applications in construction focused on forecasting the final cost and duration of projects not the resource requirements.

The concept of sliding moving averages was used to forecast the final cost and budget of construction projects (Teicholz, 1993). This system calculates cost at completion by adding cost to date to the multiplication of remaining unit cost by the remaining percent complete. The remaining unit cost was calculated as the linear projection of the average unit cost for a previous period of the project.

This period is selected so that it is not too large or too small to represent the current trends in the project performance.

Stochastic S-curves (SS-Curves) were used to replace the traditional deterministic S-curves in order to obtain more accuracy in forecasting project final cost (Barraza,

2000). The SS-curves were generated by introducing uncertainty to the cost and duration of every activity in a project using simulation. These distributions are then presented graphically as one envelop of S-curves with time as independent variable and cost as a dependant variable. The project actual performance at a point in time is then added to the graph as actual cost and percent complete. The graph is then used to calculate the budget and duration distribution for actual progress. This probabilistic approach provides a range of outcomes rather than deterministic results that have a low probability of occurrence.

A dynamic Markov Chain approach was developed for forecasting project performance (Nassar, 2004). He used a combined index I to identify five project states. These states are; outstanding ($I > 1.15$), exceeds target ($1.05 < I \leq 1.15$), within target ($0.95 < I \leq 1.05$), below target ($0.85 < I \leq 0.95$) and poor ($I \leq 0.85$). Since Markov Chains are “memoryless”, which means the future state of a project depends only on the current state, he used cumulative performance measures to include the project history in the forecasting process.

A transition probability matrix $P_{(5,5)}$ was suggested with $p_{(i,j)}$ represents the probability of the project going from state i at time = t to state j at time = $t+1$. The system needed a set of projects’ data to calculate $p_{(i,j)}$ by dividing the number of projects moved from state i to state j by the total number of projects in the set.

These last two approaches require a powerful structured approach towards data generation, collection and storage in order to be able to calculate the necessary

probabilities for the forecast of resource requirements. The development of this data management framework is the main objective of this research.

2.4 TRANSFERRING PROJECTS DATA TO USEFUL KNOWLEDGE

There are so many definitions of the word knowledge, which shows the richness, complexity and interpretability of the topic. Knowledge in philosophy is defined by Plato as “justified true belief” or “those propositions or sets of propositions individuals believe with good reason to be true that, in actuality, are true (plusroot.com, 2007). Knowledge can be also described as the product of learning, which is personal to an individual or an intangible economic resource from which an organization can draw future revenues (Orange et al., 2000). This definition distinguishes between individual’s knowledge and organization’s knowledge. Organizational knowledge can be broken down into four main categories; human, market, technology and procedural (Fu et al., 2006).

Another definition of knowledge is that it is reasoning about information and data to actively enable performance evaluation, problem solving, and decision-making, learning and teaching (Beckman, 1999).

In organizations, internal knowledge assets accumulate in the firm to form knowledge stocks (Wang et al., 2007). Meanwhile knowledge loss is also taking place in the firm. The amount of knowledge stocks is the cumulative result of incoming and outgoing flows of knowledge into the firm. Hari et al., 2005 affirmed that the world is moving to the knowledge era since early 1990’s due to

globalization, internalization of markets, the liberalization of trade and deregulation. They also noted the difference between knowledge, which expands and grows when utilized versus natural resources and other physical capital, which are depleted when used. A recent study reported that more than 75% of surveyed companies believe that knowledge is a strategic asset and they are losing about 6% of business opportunities due to the lack of proper management of available knowledge (KPMG, 2003).

The knowledge pyramid consists of data at its base, information, knowledge and wisdom at the tip of it as shown in Figure 2.1 (Liebowitz et al., 2003). Data represents raw elements such as readings of process flow pressures during plant operations or weekly actual resource utilization during project's planning and execution. When these elements are patterned in a certain way, data is transformed to information, which can be analyzed. Once certain rules or heuristics are applied to this information, knowledge is then generated as actionable information for producing value-added benefits. Knowledge is more integrated and includes the human expertise, reasoning behind decisions, ideas, improvements, innovations and lessons learned from the structured information. Wisdom in the knowledge pyramid represents the ability to make right decisions using the available knowledge to maximize the value-added benefits.

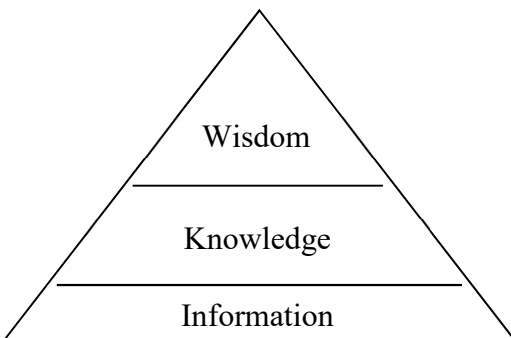


Figure 2.1: The Knowledge Pyramid

Knowledge is typically classified into two categories, tacit and explicit knowledge (Polanyi, 1966). Explicit knowledge can be presented in the form of books, research papers, reports, graphs, memos, engineering drawings, etc. Tacit knowledge, on the other hand, is mainly stored in the human minds and represents the experience they acquire trying to solve daily work problems. According to Davenport and Prusak (1998), knowledge is always generated in organizations while running the business. This knowledge needs to be codified so it becomes accessible to those who need it. Once the codification process is completed, knowledge can then be transferred to others who would also generate new knowledge. These three steps form the complete knowledge cycle as shown in Figure 2.2.

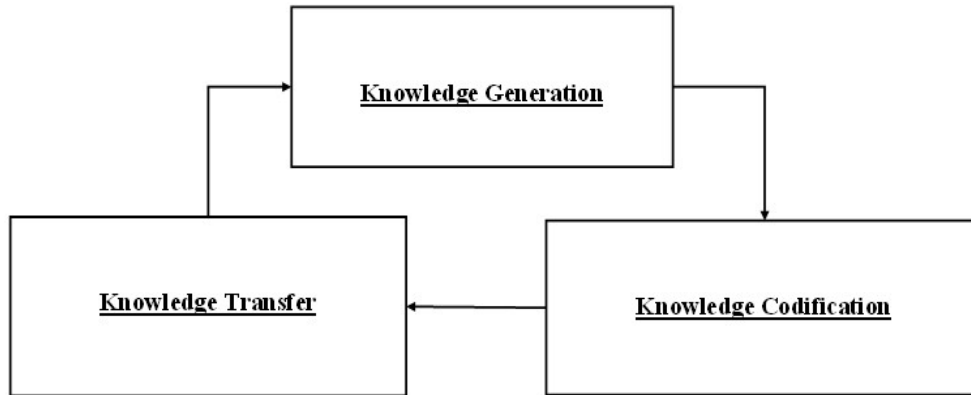


Figure 2.2: The Knowledge Cycle

Very few attempts were made to capture and transfer knowledge between projects. Case Based Reasoning (CBR) technique was utilized to transfer previous engineering design data to useful knowledge (Leake & Wilson, 2001). Their model relies on retrieving relevant previous design cases and using them as guidelines for new designs. It combined CBR with *Knowledge Mapping* techniques to be able to transfer design knowledge from senior to junior engineers. They developed a system called DRAMA (Design Retrieval and Adaptation Mechanisms for Aerospace) that included interactive user interface to define a set of standard attributes, which are used to organize previous cases.

These attributes are also used to retrieve the nearest matching previous design based on the users answers to predefined questions. The recommended results by the system were compared to real designs selected by senior engineers to validate the model. This study was a good step towards trying to capture the rational behind

design decisions; however, it required extensive efforts from the users to define the design cases; thus, making it very difficult to implement in real life applications.

Another application was called KTfD (Knowledge Tools for Designers) to help transfer previous design knowledge to new designs (Crowder et al., 2003). They developed a library of design issues, which was structured using the Cambridge UTP/EDC Design Knowledge Model (DKM) ontology. The application also provided a knowledge map of colleagues who worked on similar design issues before. They also suggested a design development scenario to be implemented by all designers in the company, which encourages knowledge capturing and sharing during the development steps. The system was only a suggestion and was not really implemented and tested in reality.

In construction, Al-Jibouri, and Mawdesley (2002) developed a knowledge-based system to support construction project managers in decision making activities. They established an initial information model to link the 55 tasks performed by project managers. The tasks were obtained from previous research and a set of structured interviews with a group of experts. The model considered construction projects as an information flow over time. It broke down these projects into tasks, and then it defined the necessary input and output information for each of these tasks. As a result, project managers could sift through the knowledge-based system task by task, providing the status of available information for each one, and assessing the impact of missing information on other tasks using simplified

decision trees. This model considered information as attributes of each task, and allowed that information to be linked to other information and not tasks to tasks.

CoMem (Corporate Memory) was developed by Stanford University as a prototype design knowledge management tool (Fruchter & Demian, 2002). CoMem consists of a Semantic Modeling Engine (SME) and project memories combined to formulate corporate memory. SME is a framework that enables designers to map objects from an AutoCAD file to multiple semantic representations. The SME was broken into Graphics objects contain drawing files, discipline objects containing a list of component classes (ontology to describe the semantic meaning of the graphics) and component objects (instances from a particular graphic model that are relevant to interpretation), Component objects (capture the link between graphic entities and symbolic entities) person objects (serve as a record of the project participants and their roles), Note objects (to capture the design rationale), W-Doc objects (for linking a component object to sources of information) and Change notification objects (to communicate design changes between team members). An SME corporate memory is a hierarchical data structure in which a corporation contains multiple projects, a project consists of multiple disciplines, and a discipline contributes multiple components. The system relies mainly on the AutoCAD model and it was very complicated and required continuous maintenance.

The CAPRIKON (capture and reuse of project knowledge in construction) research project developed a methodology for life management of construction projects'

knowledge (Tan, 2007). The methodology adopted the live methodology for knowledge capture proposed by Kamara et al., 2000. The proposed methodology included 7 blocks. The first block is defining an Integrated Work-Flow System and Project Knowledge Manager (PKM) who is responsible for ensuring that all users are reporting the new knowledge in the system. The second block is Capture Knowledge from Group. Block 3 is Capture Knowledge from individuals. Block 4 is Capture of Rationale for Making Changes to Documents. Block 5 is knowledge Validation. Block 6 is Project Knowledge File. Block 7 is Dissemination and Reuse.

It is very difficult to implement a complete knowledge management approach in construction projects environment due to lack of incentive to contribute, issues around ownership of generated knowledge and difficulty to form a complete knowledge cycle (Leseure & Brookes, 2004). Al-Ghassani et al., 2004 added these factors: lack of enough time and resources to capture knowledge, proper organizational culture and standard work processes.

2.5 DATA WAREHOUSING AND ON LINE ANALYTICAL PROCESSING (OLAP) TECHNIQUES

According to Mohamed (2008), there is a flood of data in all aspects of human knowledge. The data growth rate is exponentially increasing with more than 30% annual raise. For example, the size of the largest database in the Winter Corp survey has tripled between the years 2003 and 2005 (Winter Corporation, 2008). Measurement of data has moved from Kilobytes (10^3 bytes of data) to Megabytes (10^6 bytes of data), to Gigabytes (10^9 bytes of data), to Terabytes (10^{12} bytes of data), to Petabytes (10^{15} bytes of data), to Exabytes (10^{18} bytes of data), to Zettabytes (10^{21} bytes of data) all the way up to Yottabytes (10^{24} bytes of data). KDD, data mining, data warehousing and OLAP techniques have become crucial in order to efficiently manage this data flood, to analyze it properly and to transfer it to useful knowledge for end users.

2.5.1 Multidimensional Data Warehousing

The need for data warehouses is significantly increasing in today's knowledge era especially with the enormous amount of collected data in every domain. Data warehouses are used to turn collected data to useful subject-oriented knowledge (Sumathi and Sivanandam, 2008). Data warehouses are dedicated, read-only and non-volatile databases that are used for discovering useful knowledge in large sets of data (Inmon, 2005). Inmon, the first to coin the term 'data warehouse' back in 1990, explains that data warehouse focuses on centrally storing validated historical data and utilizing this data for Decision Support Systems (DSS) not Operation

Support Systems (OSS). Data warehouses are not used for storing daily transaction data in a generic way; but rather, they are used to store specific subject-oriented data according to the needs of end-users.

Most organizations collect and store large quantities of data using different tools and formats. This data is generally stored in relational databases such as MS Access, MS SQL Server, Primavera, SAP or Oracle databases. None of these systems is capable, solely, of providing all the needs of an organization. As a solution, many large system providers, such as Oracle, SAP and IBM, acquire additional add-on subsystem to fill the gaps in their original system claiming that they can provide a one-stop solution to meet all the needs of their clients. In spite of these efforts, most organizations are still obliged to use several systems to fulfil their needs and requirements. Since these multiple systems do not directly communicate with each other, it becomes difficult to extract necessary data to be used for timely decision-making. And thus, it prevents the organizations from performing proper data analysis and mining to transfer data to useful knowledge.

A data warehouse solves this problem by storing all required data for a specific decision-making problem(s) into one central location. Chau et al. (2002) emphasized the importance of storing all the needed data for decision-making in one central location. They stated that, unless a data warehouse is established, most of the analysis time is wasted trying to collect the necessary data from different sources. Wrembel (2007) supports that the first step into proper data analysis is to

extract, clean-up, validate and integrate data from multiple sources in one central location typically referred to as data warehouse.

Data warehouses use multidimensional datasets to allow decision-makers to analyze the data around certain subject from different points of view and various hierarchical levels of detail (Han and Kamber, 2006). Each dimension represents an attribute of the stored data in the cells of the multidimensional dataset. The structure of data warehouses relies mostly on the star schema for simple datasets and on snowflake schema for complicated datasets. Star schema consists of a fact table that contains data and dimension tables that contain the attributes of this data. The snowflake schema is used either when multiple fact tables are needed or when dimension tables are hierarchical in nature (Giovinazzo, 2000). Using the snowflake schema makes the queries more complicated but capable of producing the reports according to the user specific requirements. The multidimensional snowflake structure is highly efficient in meeting the needs of a clearly defined domain application (Inmon, 2005).

According to Ahmad et al. (2004), a data warehouse typically consists of three main components: the data acquisition systems also known as backend, the central database and the knowledge extraction tools, known as the frontend. In the backend of a data warehouse, the data can be extracted from text files, spreadsheets or On Line Transactional Processing (OLTP) operational databases. The central component is a strong relational database that is designed to store historical data. It

is not used to run daily business transactions. The frontend consists of a combination of data viewing and data mining techniques that are used to extract the required knowledge and present it to the end-user in their preferred format. According to Chau (2002), the contents of the data warehouse can be either a replica of data from operation systems, results of queries on joint tables or both.

Bain et al. (2001) defined six main differences between multidimensional data warehouses and traditional relational databases. First of all, data warehouse is subject-oriented meanwhile traditional databases are application-oriented. Second, data warehouses include pre-processed and summarized data to expedite viewing and querying and are not limited only to operational transactions data. Third, data warehouses target decision-makers while databases target daily users. Fourth, data warehouses focus on historical data to be used for future forecast not on current data. Fifth, data in data warehouses is cleaned, validated and not subject to change after storing. Sixth data warehouse schema is dynamic in nature to allow addition of new dimensions to the data whereas database schema is usually static.

Data warehouses focus on obtaining related data around one subject from multiple data sources such as spreadsheets, relational databases and any other data source and store it in one central location. They have been used in several operation-based industries such as financial institutions, retail stores e-commerce and home-land security. However their use in the construction industry is still very limited.

2.5.2 On Line Analytical Processing (OLAP) Techniques

The first use of the term On Line Analytical Processing (OLAP) was by Codd (1993) who introduced the concept of multidimensional data analysis and its principles for analyzing enterprises' datasets. The focus is on consolidating and aggregating datasets using variable paths in order to enable the dynamic data analysis and meet the various needs of business decision-makers. Contrary to static data analysis, dynamic data analysis requires the users' interaction to produce output according to their decision-making needs. Codd also defined the twelve principles of OLAP tools as:

1. Multidimensional Conceptual View
2. Transparency
3. Accessibility
4. Consistent Reporting Performance
5. Client-Server Architecture
6. Generic Dimensionality
7. Dynamic Sparse Matrix Handling (dealing with blank data points)
8. Multi-User Support
9. Unrestricted Cross-dimensional Operations
10. Intuitive Data Manipulation
11. Flexible Reporting
12. Unlimited Dimensions and Aggregation Levels

All these aspects were taken into consideration while utilizing OLAP reporting techniques to extract useful information out of the resources data warehouse.

OLAP techniques and data warehouses complement each other. A data warehouse stores and manages data meanwhile OLAP techniques transform data in the data warehouse into useful information that could be viewed and analyzed properly by decision-makers (OLAP Council, 1997). According to Fan (2007) OLAP techniques for data manipulation and reporting include: roll-up and drill-down, slice and dice, and data pivoting. These techniques are used to provide the end-users with customized reports and graphs to meet their decision-making needs.

Roll-up and drill-down techniques are used to view data at different levels of details according to user's needs. For example resources data can be viewed by internal project, program of internal projects or portfolio of internal programs. Similarly the data can be viewed by industrial project, program of industrial projects or portfolio of industrial programs. On time dimension, the data can be viewed by week, month, quarter or year. Slice and dice are used to view the data either from one dimension or multiple dimensions. Data pivoting shows the data on a two dimension matrix with multiple row and column headings. This technique also provides users with powerful graphical illustrations and filtering capabilities to customize the output reports precisely towards users' needs.

2.6 DATA MINING METHODS AND TECHNIQUES

Data mining is the main step for extracting knowledge from datasets in any KDD model. Hand et al. (2001) defined data mining as “The analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owners”. Fayyad et al. (1996) emphasized that the discovered knowledge after the data mining procedure has to be previously unknown, non-trivial and really useful to the data owners.

According to Tan et al. (2006), data mining can be either descriptive to drive patterns that summarizes the datasets or predictive to forecast the values of certain attributes based on historical performance. In this research, OLAP reporting is used as a descriptive tool to present the stored data in the resources data warehouse; meanwhile, data mining is used to provide predictive capabilities for better forecasting of future resource needs. Data mining methods for discovering knowledge can also be categorized in two main categories; unsupervised learning and supervised learning (Cios et al., 2007).

Han and Kamber (2006) grouped Data mining techniques into five categories: Characterization and Discrimination tools, Association and Correlation analysis, Classification and Prediction techniques, Clustering methods and Outlier analysis. Data characterization and discrimination tools focus on defining the general features of a target set of data. Association and correlation analysis focuses on finding frequent patterns in data meanwhile, classification and prediction

techniques try to build a model to represent data and use it for assigning new data points to the most appropriate class. Clustering methods try to classify data in groups of similar behaviour. Outlier analysis locates data points that do not follow the general behaviour of a data set.

Data mining combines ideas such as sampling, estimation, hypothesis testing, search algorithms, modeling techniques from the sciences of statistics, Artificial Intelligence (AI) and machine learning (Tan, 2006). Data mining techniques are implemented in many fields such as marketing and sales, stocks, credit cards, sports, health care, web-based and e-commerce data. Data warehousing and mining techniques are essential for companies that want to increase their Business Intelligence (BI). BI is defined by The Data Warehouse Institute (TDWI) as “the processes, tools, and technologies that are required to turn data into information and to turn information into knowledge and effective business plans” (TDWI, 2008). Both, data mining and BI are little used and known in the project-based construction industry, mostly because of erroneous perceptions that unlike operational data, project data is not suitable for data-mining techniques and BI.

Traditional data analysis techniques such as stochastic models and time-series analysis have major limitations in finding useful knowledge in datasets. They rely heavily on building mathematical models to represent relationships between already established variables. Yet, in many cases, neither the variables nor the

relationships between them are easily located. To overcome these limitations, data mining techniques are used.

Sumathi and Sivanandam (2008) stated that data mining focuses on finding hidden relationships in business data to allow decision-makers to predict future performance. Data mining techniques rely on data, which represents recorded facts, with little input from domain and subject matter experts (Fan et al., 2007). Many of these techniques can be visualized as understandable patterns for the decision-makers who do not have extensive computer knowledge. Data mining models are self-learning models in a sense that they are easily updated with the availability of new data.

2.6.1 Data Mining Techniques Based on Unsupervised Learning

In unsupervised learning, data is analyzed and structures in data are discovered without user interference. In this type of learning, it is required to find the structure of a dataset:

$$X_{(i=1:n)} = (x_1, x_2, x_3, \dots, x_n) \quad [2.1]$$

Where each data point ($x_{(i)}$) has a value and is defined by a number of attributes. The number of attributes represents the number of dimensions in the dataset. In this type of learning, it is required to find a classifier:

$$\Phi(x_{(i)}) = \omega_i \quad [2.2]$$

Where $\omega_i \in (1, 2, 3, \dots, n)$ represent the class labels. The unsupervised learning techniques include mainly clustering and mining association rules.

2.6.1.1 Clustering Techniques

Clustering methods rely mainly on the concept of minimizing the distances between data points in the same cluster and maximizing the distances between data points in different clusters (Zaiane, 2006). Different distances can be used such as Hamming distance, Euclidean distance and Tchebyshev distance (Tan et al., 2006). These distances can be represented by the Minkowski general formula:

$$d(x,y) = (\sum_{i=1:n} |x(i) - y(i)|^r)^{1/r} \quad [2.3]$$

Where $r = 1$ provides the Hamming distance and $r = 2$ provides the Euclidean distance.

The clustering methods are typically grouped into five main categories (Zaiane, 2006). These categories are:

- Partitioning algorithms
- Hierarchy algorithms
- Density-based methods
- Grid-based methods
- Model-based methods

Partitioning algorithms such as *K-means* clustering rely on optimizing an objective function to discover the structure of the dataset (Cios et al., 2007). Centroid is the statistical mean of the data points in each cluster. The number of clusters (k) is usually provided by the data miner. The method starts with partitioning the dataset into k clusters at random, calculate the centroid of each cluster and then assign each data point to the nearest cluster based on the distance between the data point and

the centroid of the cluster. The process is repeated until the objective function is met.

Hierarchy algorithms can be either agglomerative (bottom-up) or divisive (top-down). These applications provide their output in a tree structure format that is usually called dendrogram (Hand et al., 2001). Several methods are used to represent distance between clusters such as single link (shortest distance between any two data points from two different clusters), complete link (longest distance between any two data points from two different clusters) or average link (the average distance between all data points from two different clusters).

In density-based methods, clusters are treated as dense sub-sets of data points in the data space (Han, 2006). Input parameters usually include the maximum radius and minimum number of data points in each cluster. Examples of these methods include Density Based Spatial Clustering of Applications with Noise DBSCAN (Ester et al., 1996) and TURN (Foss and Zaiane, 2002).

Grid-based methods starts by defining a grid in the data space and then draw geometric constructs called hyperboxes around data clusters (Cios et al., 2007). Examples of this method are STatistical INformation Grid (STING) and WaveCluster.

Model-based clustering methods optimize the fit between the dataset and a mathematical model (Han et al., 2006). Expectation Maximization (EM) is one of these methods and is the one selected to be used in this research as an example of unsupervised learning. This method is based on the *finite mixture* model, which assumes that the dataset consists of a k number of probability distributions equal to the number of clusters (Witten and Frank, 2005). It assumes that each data point belongs to one cluster only and clusters are not equally likely. To simplify the problem, all clusters are assumed to have Normal distribution but with different means and standard deviations. The EM method finds the mean and standard deviation for each cluster and the prior probability, which reflects the relative population of each cluster.

2.6.1.2 Association Rules Mining

The second category of unsupervised techniques is association analysis or mining association rules. It is used mostly in finding the shopping patterns in department stores and credit card transaction databases. For a dataset D with a total number of transactions = T , we need to calculate the support and confidence of the rule that the combination of A and B exists. Support measures the number of times the rule existed in the dataset; meanwhile the confidence measures the strength of the rule (Han, 2006).

Support and confidence of the rule are calculated using the following formulas:

$$\text{Support (A\&B)} = P(A \cup B) = \#(A \cup B) / T \quad [2.4]$$

$$\text{Confidence (A\&B)} = P(B|A) = \#(A \cup B) / X \quad [2.5]$$

Where X = number of transactions that contained both A and B. Rules that have low support and confidence are usually rejected. The main problem of this approach is the increasing number of rules with the increasing number of attribute values. Several algorithms are available to solve the problem such as Naïve, Apriori and the Frequent Pattern Tree.

2.6.2 Data Mining Techniques Based on Supervised Learning

Contrary to unsupervised learning techniques, supervised learning techniques rely on the user provide class labels (represented as a small set of integers) and the technique has to assign the most appropriate class label to each data point. The basic principle of supervised learning is to build a model using a training dataset to define data classes, evaluate the model and then use the developed model to classify each new data point into the appropriate class. Most classification techniques need a labeled dataset that are divided to a training set and test data. Once the model is generated using the training data, it is tested and evaluated using the test data and then used to classify the unlabeled new data. Models are evaluated against their predictive accuracy, speed, scalability, robustness and interpretability (Zaiane, 2006).

The supervised learning techniques are classified into two main categories: statistical methods and classification. Statistical techniques include Bayesian methods, regression analysis and other data stratification techniques. Classification techniques include Decision Trees, Rule-Based Algorithms, Artificial Neural Networks (ANN), k-Nearest Neighbours (k-NN or lazy learning), Support Vector Machine (SVM) and many other data classification techniques (Cios, 2007).

2.6.2.1 Statistical Techniques

Bayesian methods apply the Bayes theorem to calculate the probability that a data point belongs to a certain class. The Bayes theorem states that the posterior or conditional probability of A given B can be calculated using the formula:

$$P(A|B) = P(B|A) * (P(A) / P(B)) \quad [2.11]$$

Where $P(B|A)$ is the conditional probability of B given A, $P(A)$ is the prior or marginal probability of A and $P(B)$ is the prior or marginal probability of B. This theory is used in classification by utilizing the dataset for calculating the probability ($P(A|B)$) that a data point (B) belongs to class (A), repeating this process with all classes and assigning the data point to the class with highest probability. The Naïve Bayes classifier assumes that all data attributes are all equally important and completely independent from each other. Bayesian Belief Networks are used to find class conditional dependencies.

Regression analysis is used to apply curve fitting to existing data in order to predict future values of a data variable (Fayyad et al., 1996). Regression models can have multiple shapes and equations, some of these models are shown below:

Linear regressions (straight line) $Y = \beta_0 + \beta_1 X$ [2.6]

Quadratic regressions (parabola) $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ [2.7]

Third degree polynomial (S curve) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$ [2.8]

Exponential $Y = \beta_0 + \beta_1 * \text{Exp}^{(X)}$ [2.9]

Another stratification technique is based on comparing the means of data sets using The Univariate Analysis of Variance (ANOVA) and the Post Hoc tests.

Typically, when comparing the means of two independent samples, the student's t -test is used (Gamst et al., 2008). The test assumes that the two samples are random and independent and the two populations are normally distributed or sample sizes are large enough (>30). The null hypothesis H_0 is always that μ_1 (of the first sample) = μ_2 (of the second sample). While the H_1 hypothesis can be either:

- $\mu_1 \neq \mu_2$ for the two-tailed hypothesis
- $\mu_1 < \mu_2$ for the left-tailed hypothesis
- $\mu_1 > \mu_2$ for the right-tailed hypothesis

The t_0 value is calculated and compared to the critical value from the table of the t -distribution and the null hypothesis is accepted or rejected based on the results. A

similar approach is used to compare two standard deviations using F -test, where $F_o = (\text{Variance}_{(1)} / \text{Variance}_{(2)})$ and the table of the F -distribution.

While comparing three or more means, the one-way-ANOVA test is mostly used (Sullivan, 2007). The test assumes that samples have equal variances and checks the null hypothesis that all means are equal. The test is an extension to the two-tailed t -test where the null hypothesis is rejected if the P -value is smaller than the level of significance (α). The level of significance (α) is usually set at 0.05 within a 95% confidence interval where the P -value = the sum of the area under the two tails.

To find out which sample means are different; multiple comparison methods (Post Hoc tests) are performed. These tests check for the null hypothesis H_o where:

$$H_o: \mu_i = \mu_j \quad \text{for all cases where } i \neq j \quad [2.10]$$

Tukey and Duncan tests are the most commonly used Post Hoc tests. Both tests use the studentized range distribution to compare the means of all possible pairs. Duncan test is powerful and effective in detecting differences between means, while Tukey test is more conservative and less powerful (Salem, 1998).

2.6.2.2 Classification Techniques

Decision trees classify data in a flow-chart-like tree structure where internal nodes represent test on a data attribute and branches represent an outcome of the test. Decision trees are built using the recursive (top-down) process that starts by setting all data points at the root of the tree, start the partitioning process using an attribute and then prune the tree to eliminate the unneeded branches (Zaiane, 2006). Different algorithms or goodness functions are used to select the partitioning attribute.

Artificial Neural Networks (ANN) is a data structure that mimics the behaviour of neuron cells in the human brain. It consists of an input layer, a calculation layer of interconnected nodes and an output layer. Mathematical functions and the training dataset are used to calculate the weights of each connection, $w_{(i,j)}$ between node_(i) to node_(j), and use these weights to classify the testing dataset.

The k-Nearest Neighbours (k-NN) is called the lazy learning because it doesn't generate a model but directly use the classes of training dataset (Cios, 2007). When there is a need to classify a new data point, this point is compared to the training dataset to find its closest k neighbours and then assign the class of the majority of these neighbours to the new data point. The user has to define the k number and the distance function to be used for finding the nearest neighbours.

Support Vector Machines (SVM) consider data as sets of vectors in an (n) multidimensional space and generate separating hyperplane (multidimensional plane) to split classes in that space (Taniar, 2008). The vectors that have minimum distances to the maximum separating hyperplane are called support vectors. The support vectors and the maximum separating hyperplane are found using sequential minimal optimization.

In summary, data mining is a promising field of human knowledge that is growing rapidly with new techniques and application of these techniques are introduced daily. The next section of this chapter discusses the application of KDD and data mining techniques in the construction industry.

2.6.3 Outliers Detection

In data mining, outliers' detection is also referred to as deviation detection, anomaly detection, intrusion detection or exception mining (Zaiane, 2006). The objective of this process is to find data points that are significantly different from most other data points.

An outlier can be defined as “Given a set of observations X , an outlier is an observation that is an element of this set X but which is inconsistent with the majority of the data or inconsistent with a sub-group of X to which the element is meant to be similar” (Fan, 2006). Another definition is “Outliers are those data records that do not follow any pattern in an application” (Chen, 2003). A third

definition of an outlier is “an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism” (Hawkins, 1980).

There are three main causes for generating outliers. These causes are either an error in data measurement, the data point belongs to a different class of data, or the data point is a rare extreme case. Most datasets can be modeled as statistical distributions; however modeling data prior to detecting outliers may lead to distribution characteristics that do not optimally represent the datasets. Outliers can be also a subject of further analysis, if needed, to determine their causes and their impact on the data distribution and statistics. Outliers can be either global, where a data point doesn't belong to the whole dataset, or local, where an outlier doesn't belong to a data subset (certain cluster) of data. This is a major issue especially when the dataset has multiple dimensions similar to the dataset under investigation. This means defining data points to be outliers is subjective to the user judgment on which data subset to use. In this research, a score is calculated for each data point based on the number of cases it becomes an outlier in a subset.

Outlier detection techniques can be either supervised, if a training set of data with class labels (Outlier vs. Not-outlier) is available, or unsupervised where the technique has to detect the outliers without previous knowledge. Outlier detection techniques can be grouped into three categories (Tan, 2006). These categories are:

- Model-based techniques. In these techniques, a model is built to fit the data and the points that do not fit the model are considered outliers.
- Proximity-based (distance-based) techniques. In these techniques, a proximity measure (usually distance between data points) is used and the data points that are remote from most other points are considered outliers.
- Density-based techniques. In these techniques, the density of data distribution is calculated and the data points falling into low density zones are considered outliers.

These techniques handle the outlier detection as univariate problem. However, in some datasets, the problem has to be treated as multivariate. Multivariate outliers detection is a very complicated process and require different techniques such as Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD) and Cluster Principle Component Analysis (CPCA) (Stefatos, 2007).

2.6.4 Previous Applications of KDD and Data Mining Techniques in Construction

The first introduction of data warehousing and mining techniques to the construction industry took place at the beginning of the third millennium. Data mining and warehousing techniques have only been recently introduced to the construction industry. As a result, there are few applications of these techniques in the literature. Some of these applications focused on operational data not on project data.

Soibelman and Kim (2002) used KDD and data mining to analyze the problem of schedule delays. Their KDD approach consisted of five steps: problem identification, data preparation, data mining, data analysis and knowledge refinement. They also confirmed that data preparation was the most important, difficult and time-consuming step in the KDD approach.

Chau et al. (2002) combined the concepts of data warehousing, Decision Support Systems (DSS) and OLAP to develop the Construction Management Decision Support System (CMDSS). They used star schema to build their data warehouse. Their fact tables included material inventory and use, machine cost and use, project progress and noncompliance. Multiple data cubes were developed to be used by the DSS. The DSS was basically an interface that allows both experienced and junior users to print reports and graphs out of the data warehouse. However, their DSS lacked any tools to analyze the collected data or find hidden knowledge.

Ahmad et al. (2004) developed a DSS for selecting residential-housing development sites using data warehousing concepts. They merged data from different sources including a Geographical Information System (GIS) in a data mart, which is a sub-set of a data warehouse that focuses only on one business process. In addition, they used the Analytical Hierarchy Process (AHP) to rank the available sites in the data mart and recommend the most suitable for development. A questionnaire survey was sent to a group of experts to define the five factors affecting the decision for the AHP analysis. The results from the DSS were

validated by the same group of experts and found to be consistent with their expectations.

Zhiliang et al. (2005) developed a prototype system to utilize electronically exchanged documents for decision support in construction projects. Their schema consisted of four fact tables: quality, material, payment and schedule. The output interface was only able to provide pivot tables and charts for users to analyze the contents of the data warehouse.

Rujirayanyong and Shi (2006) developed a Project-oriented Data Warehouse (PDW) for contractors. The PDW consisted of 10 fact tables and 16 dimension tables that were directly populated from other applications such as Primavera, MS Access and MS Excel. They used the snowflake schema to represent the required hierarchical nature for dimension tables. Their output was also limited to querying the warehouse without any knowledge finding or data mining.

Moon et al. (2007) introduced probability analysis to the historical cost data in their application Cost Data Management System (CDMS). Their cost data cube had four dimensions: time, size, region and Work Breakdown Structure (WBS). Their probability model calculated only the mean and variance of unit costs for different construction activities. They considered a correlation coefficient between dependent construction activities and a degree of dispersion, which represented the data scattering around the mean, to adjust the obtained estimates. The measure of

dispersion and correlation coefficients were calculated using the OLAP Analysis Services of MS SQL Server. The system provided the users with better understanding of cost uncertainties and provided them with more reliable estimates of construction costs.

Fan et al. (2008) used the Auto Regression Tree (ATR) data mining technique to predict the residual value of construction equipment. This technique represents an easily interpreted non-linear regression model. It uses the Bayesian updating technique, which treats the model parameters as statistical distributions. It finds the tree topology that best fits the training data set. Their model included only one type of equipment and found out that equipment age, make, horsepower and conditions are the most important features in dividing the tree and predicting the residual value. The model was validated using the Relative Squared Error (RSE), which is the total squared difference between the predicted values and actual values divided by total squared difference between the predicted values and the average value of the data subset at this tree leaf.

Since the introduction of computers to the construction industry, more and more data is becoming available for researchers. KDD and data mining provide tools to extract useful knowledge from this data. As seen from previous research, these advanced techniques have significant potential for improving productivity and increasing efficiency of construction operations in the future.

2.7 CONCLUSION

Most of the previous research assumed that the durations of project activities as well as the resource needs were pre-set and deterministic. Moreover, previous studies worked on theoretical projects that have few activities and do not mirror real projects. Based on these assumptions, they tried to solve the problem of resource allocation or leveling. However, these assumptions are not realistic since real-life projects typically have large number of activities with high amount of uncertainty around their durations and resources' needs.

The most common approaches for resources leveling and allocation were heuristic rules, numerical optimization and Genetic Algorithms (GA). With heuristic approaches, there is no way of ensuring that the obtained solutions are truly the optimal solutions. There are two problems with the optimization techniques. First, they are not able to work on real projects because of the combinatorial explosive nature of the problem. Second, optimization techniques focus on one objective of the project; meanwhile projects typically have multiple objectives. Whereas GA approaches are very complicated and require enormous computing capacity and the required knowledge to use GA is not currently available in the industry.

There are three main findings from the literature on labour resource management practices and transfer of projects data to useful knowledge. First, none of the techniques presented earlier was able to singlehandedly address with success issues related to management of one common pool of labour resources in a multiple

project environment. Second, these techniques were no able to fully use data from previously completed projects to better estimate resource needs for future projects. Third, research on transferring data to useful knowledge shows lots of promises and able to add valuable insights in ways to improve project management practices. As a result, KDD was selected as a theoretical model that the integrated framework would be built upon.

Findings from previous research on data warehousing, OLAP and data mining showed that these techniques were used to successfully solve problems of similar nature in complexity and sophistication in construction and in other fields as well. In previous studies, they represent one of the most practical tools to collect, store, codify and analyse data in order to extract useful knowledge. Thus, these techniques are most suitable for translating the KDD model into an applicable framework for developing an integrated solution.

CHAPTER 3: INDUSTRIAL CONSTRUCTION PROJECTS DOMAIN

3.1 INDUSTRIAL CONSTRUCTION PROJECTS

Construction of industrial plants is one of the most sophisticated types of projects. There are different categories of industrial plants: such as chemical processing, manufacturing, energy generation, oil and gas production facilities, etc. Industrial construction projects vary in size from few thousand dollars to multi-billion dollar projects. These projects can have small foot print or they can occupy several acres of land almost like a small town. Furthermore, industrial projects are either constructed in land (On-shore projects) or in water (Off-shore projects). The scope of an industrial project can be the construction of a new plant (known as green field projects), the expansion, fixing or modification of existing plants (known as brown field or debottlenecking projects), routine major maintenance (known as shutdown projects) or the dismantling and de-commissioning of existing plans (known as demolition projects).

Industrial projects are known for their complexity due to several factors. First, the product of an industrial project is highly complicated. A typical industrial plant looks like a steel maze that includes general items such as processing units, tanks, vessels, pumps, heat exchangers, pipe-racks, connecting pipes, valves, measurement instrumentations, electrical and instrumentation cables, transformers, administration buildings, control rooms, special purpose items, etc....

A second source of complexity in industrial construction is the lack of clearly defined scope at the beginning of a project. The scope of an industrial project is typically defined through the procedure of Front End Loading (FEL) planning. During the FEL planning procedure, Value Engineering (VE) or Life Cycle Value Analysis (LCVA) practices take place in order to ensure that the plant is going to operate successfully and produce according to the required design capacity. Constructability and maintainability reviews also take place during FEL planning. Constructability reviews focus on maximizing construction efficiency by selecting the most fitting construction materials and methods and finding the optimum purchasing and contracting strategies. Successful constructability reviews have to involve the owner(s), engineer(s), main fabricators and constructors if possible. Maintainability reviews involve the operation and maintenance teams to ensure that all their requirements are met during the design processes.

A third source of complexity is that industrial projects are exposed to higher degrees of managerial and technical risks. Managerial risks include scope creep, schedule delays, budget overruns, etc. Industrial projects technical risks are much higher than the technical risks in other construction types. Mistakes in engineering, procurement or construction may lead to explosions, leak of extremely hazardous materials or severe environmental damages. These mistakes, known as industrial disasters, receive significant amount of public and media exposure and may lead to irrevocable damages to the reputation of the involved companies.

Because of this high level of technical risks, the Hazardous and Operability Analysis known as HAZOP has to take place during the FEL planning. This operation involves experts from all aspects of the process plant, including safety and fire-protection experts, and a facilitator who try to identify all the potential problems that might generate hazardous situations. The experts use the plant layouts and Process and Instrumentation Diagrams (P&ID's) to look at the flow levels, temperatures and pressures.

A fourth source of complexity is that industrial projects require substantial amounts of coordination and sophisticated project management due to the specific complicated nature of this type of construction. In traditional residential and commercial construction, the architect prepares the engineering and bid documents, and a General Contractor (GC) is assigned to execute the project. The GC farms out most of the work to specialized subcontractors who perform the work under the supervision of the architect and the GC. Lump sum type of contracts is a very common form of contracting in this type of construction. Different from this practice, an industrial project starts as a request from an owner to build, modify or demolish a plant. The owner hires an engineering only, Engineering, Procurement and Construction Management (EPCM) or Engineering, Procurement and Construction (EPC) firm(s) to perform the FEL planning and define the scope of the project.

The owners try to involve fabricators and constructors as early as possible in the FEL planning to learn from their experience. Many contracts are cost-reimbursable due to the lack of complete scope definition. This type of contracts requires high quality project controls to closely monitor expenditures, working hours, and work performance in order to keep the project on the right track. It is a very complicated procedure to ensure the proper and timely flow of the design documents and construction materials within all participants in an industrial project. Current studies show that productive tool time on job sites is mostly around 50% of spent time mostly due to coordination issues (Choy and Ruwanpura, 2007).

All the sources mentioned above show the complexity of an industrial project. In addition to this complexity, industrial projects can be way larger than average residential or commercial construction projects. In order to illustrate the magnitude of a large industrial project, according to Jergeas (2008) performed a study using 2008 price rates. He stated that an average mega tar-sand project would cost between \$7 and \$8 billion Canadian dollars of TIC. Such a project would require up to 3.5 million engineering work hours at a cost of \$490 million Canadian dollars, 50.000 detailed engineering drawing, 20.000 shop drawings, 15 million construction work hours at a cost of \$2.25 billion Canadian dollars, a labour force of 8,000 workers with a turnover rate of 300% annually, 800 staff personnel and 80 million items of material.

In addition to the sophisticated multi-discipline engineering and the global procurement, many industrial projects adopt modular construction as a part of their project execution plans. Modular construction (sometimes called construction modularization) is introduced to industrial construction in order to increase efficiency and decrease costs, safety issues and on-site construction hours. It is a method of constructing small modules of an industrial plant at fabrication shops, assembling these modules in assembly yards and later shipping and installing them at the final construction site (Gupta et al., 1997).

Gupta et al. also defined assembly as “The process by which various materials, pre-fabricated components, and process equipment are joined together at a location remote from the construction site for subsequent installation as one unit.” Fabrication shops typically include: structural steel fabrication shops, piping spools shops and specialty shops such as pump and exchanger manufacturers. Structural steel fabrication shops obtain the detailed shop drawings from engineering and raw steel sections through procurement and apply the processes of cutting, drilling, fitting, welding, inspection, painting and fire-proofing to build the structural skeleton of the modules. The piping spool shops apply the processes of cutting, roll fitting and welding, position fitting and welding, checking, stress relief, inspection and painting to connect a pipe to its fittings (Wang, 2006).

The module structural steel, pipe spools, mechanical equipment, electrical and instrumentation cables and all other instruments are all shipped to the module yard,

where they are assembled to form a complete module. Engineers have to take into consideration the transportation rules and regulations along the distance from the module yard to the construction site. They also have to consider the carnage plan for the empty and loaded modules to ensure the safe handling of the module until final installation on the construction site.

Transportation of specialty modules require several authority permits, a detailed logistics plan and sometimes get media exposure and even resistance from environmentalists. In a global economy, modules may be fabricated and assembled between several countries and are transported by trains, trucks, ships and planes. This procedure is complex, involves lots of risks and safety issues, and requires massive logistics planning to meet all the laws and regulations of the involved countries. In some projects, modules are even re-assembled in mega-modules that are shipped to site using specialty transportation methods. The purpose of using mega-modules is to reduce construction camp costs, on-site hours and safety incidence during construction. However, the assembly and transportation of mega-modules is a risky, complicated and sophisticated procedure that requires utmost accuracy.

On industrial construction sites, very tight safety, quality and environmental measures are applied to deliver the product according to the project specifications, minimize near-misses and safety incidents and avoid any spills or environmental hazards. Multi-discipline contractors must work together with high level of

cooperation and coordination to complete the project on schedule. Construction site layouts have to be designed intelligently to allow for the proper flow of workers, materials, cranes and construction machinery during the construction period. Sophisticated project management and controls is essential in any industrial project site.

O'Neill (1989) classifies industrial construction projects as *Team* projects to emphasize the role of multiple teams of highly-skilled individuals in successful completion of this type of projects. People or human capital represents the number one area of focus in industrial project management accompanied with processes and technologies (Badiru, 2008). Similar to other types of projects, three major constraints impact the management of human capital in industrial projects. These constraints are time, cost and performance (Badiru, 2008). Human capital is also responsible for integrating project and resource management techniques, information systems, corporate goals and latest technology advancements to ensure the success of a project. Very few attempts were found in the literature to model processes related to industrial construction. A summary of the findings from the previous research is introduced below.

Song (2004) developed a model for determining the productivity of both structural steel drafting and fabrication. His model was based on Special Purpose Simulation (SPS) and provided a virtual steel fabrication workshop for decision-makers to analyze productivity. His conceptual model for the steel fabrication process

consisted of detailing, fitting, welding, surface preparation, surface protection and shipping. Each of these processes was modeled using simulation. He also modeled internal projects of steel fabrication using a Work Breakdown Structure (WBS) that consists of five levels. These levels are division, load-list, drawing, piece and component. A schematic model of the actual fabrication facility was also used to analyze the impact of different shop layouts on productivity. His work focused only on the sub-processes of structural steel drafting and fabrication. These sub-processes start after receiving the Issue-For-Fabrication (IFF) drawings from the engineering contractor.

Wang (2006) developed a model of the pipe spool fabrication shop to facilitate implementing lean construction concepts. The main objective was to transfer the shop from the traditional batch-and-queue layout to the more efficient flow production system. He used the Value Stream Map (VSM) technique to compare the results of both the old and the new system. The VSM technique was not capable of handling the specific characteristics and the amount of uncertainty in the pipe spool fabrication shop. A SPS model was developed to overcome the disadvantages of the VSM and was found to provide better modeling and representation of the problem. Wang (2006) also developed a small model to represent material flow in pipe spool fabrication shops. His model starts with receiving the isometric drawings from the engineers, goes through the drafting of shop drawings, spool fabrication and then shipping to either module yard or construction site.

The Vertical (Within-Project) Analysis

The vertical analysis starts with the industrial construction market as shown in Figure 3.1. The industrial construction market contains two main players: (x) number of Industrial Owners that are represented with the variable ($IO_{(1:x)}$); and (z) number of Contractors that are represented with the variable ($Con_{(1:z)}$). Each of these industrial owners initiates a ($y_{(IO)}$) number of industrial projects. These Industrial Construction Projects are represented with the variable ($ICP_{(IO)}$). Each of these industrial construction projects is planned and executed as ($m_{(ICP)}$) number of internal projects.

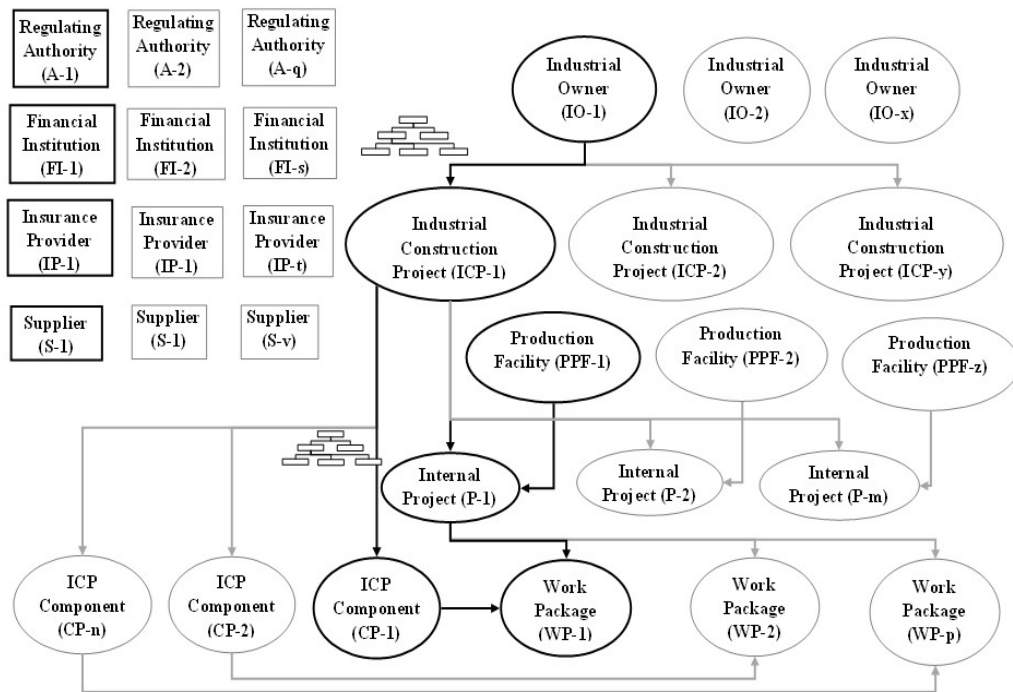


Figure 3.1: The Vertical Hierarchy of the Industrial Construction Market

Each of these internal projects is performed by one of the contractors. The internal projects are represented with the variable ($P_{(ICP,PPF)}$). Each of these internal projects are broken down into ($p_{(IP)}$) number of work packages that are represented with the variable ($WP_{(IP)}$). These work packages are planned and executed by individuals that represent the contractor's common pool of resources.

Industrial construction projects are also broken down to Industrial Components ($IC_{(ICP)}$). Pipe racks, tanks, turbines, generators, control rooms, heat exchangers, pump skids, electrical transformers and processing units are examples of these industrial components. These components are typically grouped in units, plants or areas to form the hierarchical Work Breakdown Structure (WBS) of any industrial project.

Industrial components represent the building blocks of any industrial project. In this research, each industrial component is assigned to a predefined industrial component type for data mining purposes. Industrial components are composed of a set of work packages depending on the nature of the component. Work packages are the building blocks of internal projects. In this research, each work package is assigned to one contractor and is also assigned to a predefined work package type for data mining purposes.

The vertical analysis provides seven main objects to be modeled in the data warehouse. These objects are Industrial Owners (IO), Industrial Construction

Projects (ICP), Industrial Projects Components (IC), Contractors (Con), Internal Projects (P) and Work Packages (WP) and Individuals.

Each of these objects is defined with a set of open data fields and data fields that read from other lookup tables. Some of these tables are also hierarchical in nature. For example, the location attribute is driven from the hierarchy: continent, country, province/state, and city. The data fields that are limited to reading from lookup tables are called control attributes in this research. These control attributes are used for OLAP reports and data mining experiments as explained in the next two chapters.

There are other players in the industrial construction market, which are not part of the scope of this research. Some of those players are: Governmental Organizations ($GO_{(1:q)}$), Financial Institutions ($FI_{(1:s)}$), Non Governmental Organizations ($NGO_{(1:t)}$) and off-shelf Suppliers ($S_{(1:u)}$). Examples of these players are: environmental and energy ministries, labour unions, banks, insurance providers and public groups. Since these players have little to no direct impact on the resource management data collection, they are not modeled in the data warehouse.

3.2 HORIZONTAL (CROSS-PROJECTS) ANALYSIS

3.2.1 The Cross-projects Elements

The horizontal analysis represents the standard elements utilized by all industrial and internal projects to introduce consistency to the data collection procedure. The first element in the horizontal analysis is the project stage. According to the Project Management Institute (PMI), any project life cycle, regardless of the project type, is broken down into 5 standard stages (PMI, 2004). These stages are: Initiation, Planning, Execution, Control and Closeout as shown in Figure 3.2. The first and last stages are usually short; however it is of utmost importance that projects are initiated and closed out properly and consistently for appropriate data collection. The planning and execution stages are lengthier and last for the majority of the project duration. The control stage is a continuous stage happening in parallel with the planning and execution stages as shown in Figure 3.2. The feasibility study or bid/proposal stage takes place prior to initiating of any project.

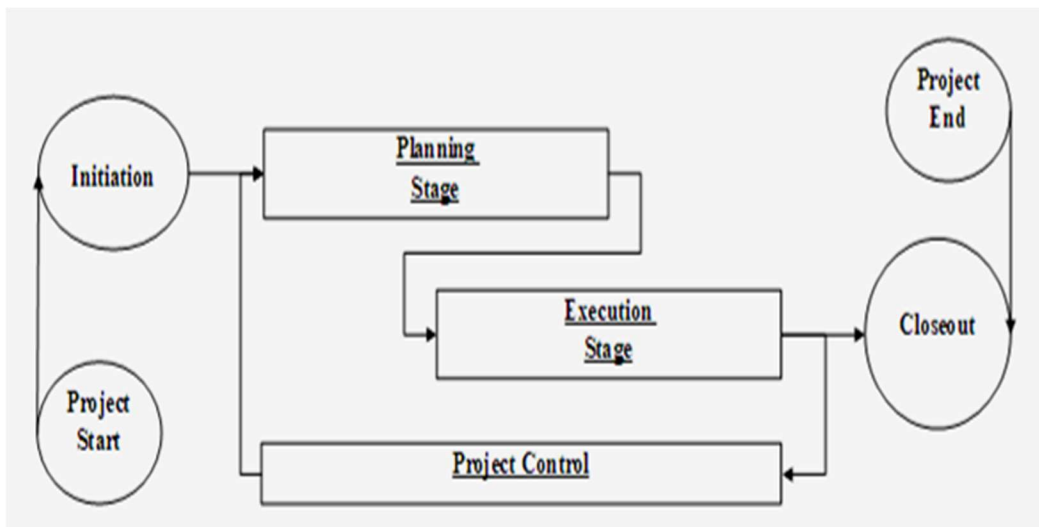


Figure 3.2: A Typical Project Lifecycle broken into Stages

The second element of the horizontal analysis is the project phase. The horizontal analysis shows that industrial projects life cycle could be also divided into five main phases that take place during each project. Within the domain of industrial construction, these phases are classified as: Pre-Engineering, Engineering, Procurement, Construction and Commissioning & Start-up (C&SU).

The pre-engineering phase takes place during the initiation stage of any industrial project; meanwhile the C&SU phase takes place as part of the closeout stage leading to transitioning the project to the owner(s). The focus of this research is on the engineering, procurement and construction phases that occur during the planning and execution stages of any industrial project as shown in Figure 3.3.

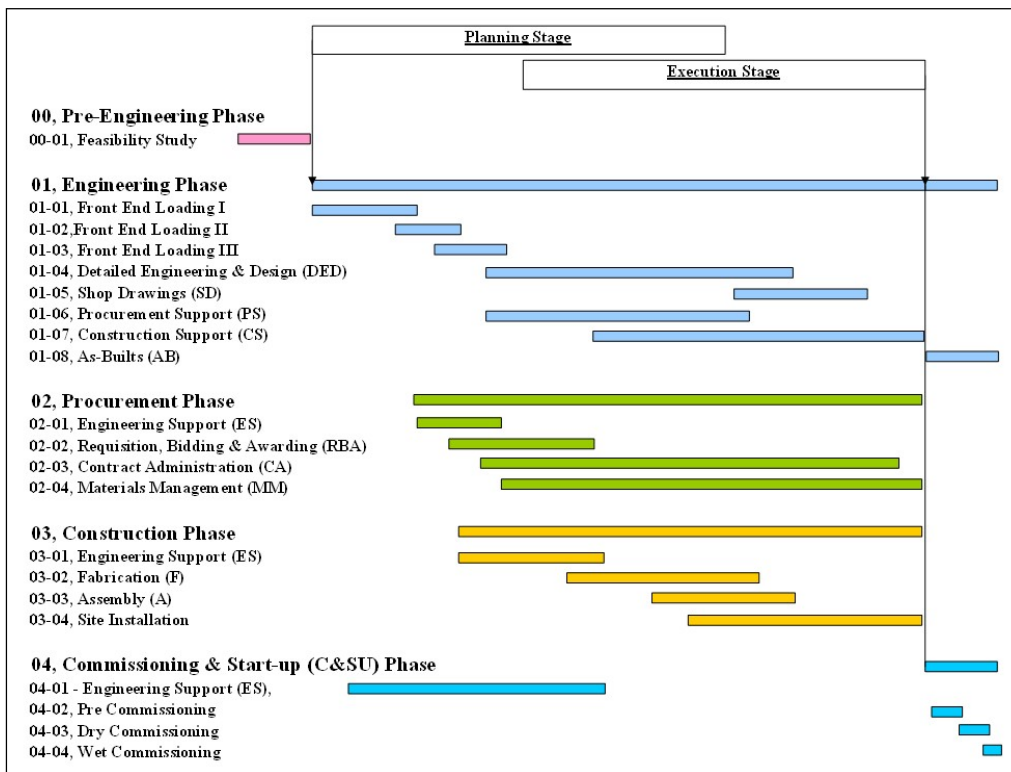


Figure 3.3: Industrial Construction Project Lifecycle broken into Phases

Each of these standard phases is also broken down into a set of sub-phases, each of which represents a process that will be modeled in this chapter. The pre-engineering phase includes only one process, which is the 00-01 - Feasibility Study. The engineering phase includes eight main processes. These processes are:

- 01-01 - Front End Loading I (FEL1),
- 01-02 - Front End Loading II (FEL2),
- 01-03 - Front End Loading III (FEL3),
- 01-04 - Detailed Engineering & Design (DED),
- 01-05 - Shop Drawings (SD),
- 01-06 - Procurement Support (PS),
- 01-07 - Construction Support (CS) and
- 01-08 - As-Building (AB).

In some small projects, the three processes of the FEL planning are combined in one process, which is called: 01-09 – Front End Loading planning (FEL) in this research. Also in very small projects, all engineering processes are combined in one process, which is called: 01-00 – All Engineering (AE) in this research.

The procurement phase includes four main processes. These processes are:

- 02-01 - Engineering Support (ES),
- 02-02 - Requisition, Bidding & Awarding (RBA),
- 02-03 - Contract Administration (CA) and

02-04 - Materials Management (MM).

The construction phase includes four main processes. These processes are:

03-01 - Engineering Support (ES),

03-02 - Fabrication (F),

03-03 - Assembly (A), and

03-04 - Site Installation (S).

The Commissioning & Start-up (C&SU) phase includes four main processes. These processes are not going to be modeled in this chapter, since they are out of the scope of this research. These processes are:

04-01 - Engineering Support (ES),

02 - Pre-commissioning,

03 - Dry-commissioning and

04 - Wet-commissioning.

The third element is the predefined set of industrial resources. This set of industrial resources can be grouped in a hierarchical Resources Breakdown structure (RBS) to be used in all projects. The structure proposed in this research consists of 5 levels as shown in Figure 3.4. The first level of this hierarchy is the resource category, which contains labour, material, equipment and other resources. In this research and in order to introduce full consistency to the resource management procedure, the labour branch of this predefined RBS forms also the Organization Breakdown Structure (OBS) for contracting company.

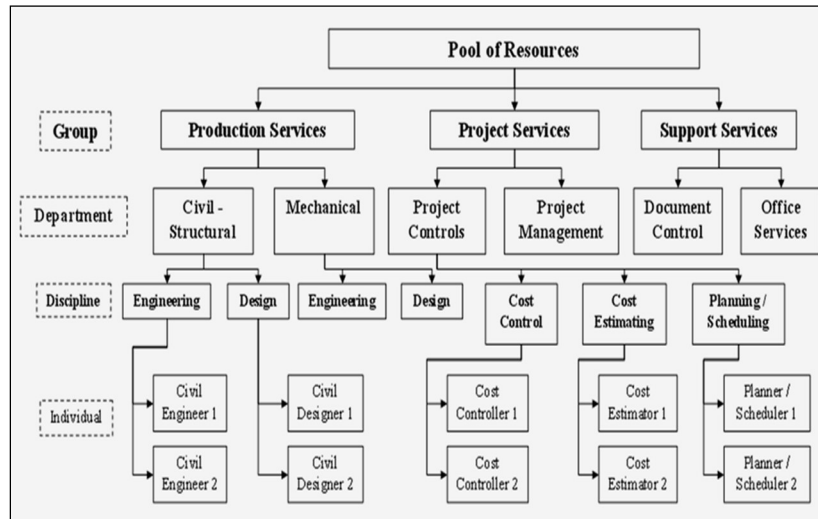


Figure 3.4: The Predefined Resources Breakdown Structure (RBS)

The second level consists of the resource groups. There are five resource groups in this level. These groups are:

- Corporate Services
- Project Services
- Engineering Services
- Procurement Services
- Construction Services

The corporate services branch include departments such as corporate management, legal, accounting, office services, and human resources management. The project services branch include project management, quality management, project controls, and document controls. Planning and scheduling, cost estimating, and cost controls are disciplines under the project controls department. Within the engineering

services, the lower level represents the engineering departments: process, mechanical, electrical, civil, structural, and instrumentation. Engineering and design disciplines are found under each of these departments, and thus representing the lowest level of this RBS branch. Procurement services include departments such as purchasing, contracting, expediting, and material management. Finally, construction services include departments such as construction management, carpentry, concrete placement, pipe fitting, welding, etc.

The fourth element of the horizontal analysis is the industrial component. These components can be extracted from the Cost Breakdown Structure (CBS), also known as Code of Accounts (CoA). The purpose of using this component is to be able to run analysis on the required resources, cost and time per component type. Examples of these components include pipe racks, tanks, pump houses, vessels, transformers, electrical substations, and all other components that are common in industrial construction projects. The industrial components are clustered on a three-level hierarchical structure in this research.

The fifth element is the production packages, which resemble the types of work packages used by a contractor. For instance, a pipe rack component requires several work packages such as foundation, structure steel, piping, electrical cables and instrumentations. Production packages are used to enable data analysis of all work packages, from multiple projects, that belong to a specific production package.

3.2.2 Resource Management Structure in Contracting Companies

Both the vertical and the horizontal analyses show that most of the resource management take place in a contracting firm and hence require a closer look at its practices. A complete analysis of the resource management process in multiple-project environment was preformed. The analysis started by describing the flow of multiple projects from commencement till conclusion. In this environment, projects go through five different states as identified in Figure 3.5. The first state is bid/proposal, where a contractor tries to secure a business opportunity. Direct requests from clients, bidding for projects, joint-ventures and sub-contracts are examples of sources of business opportunities.

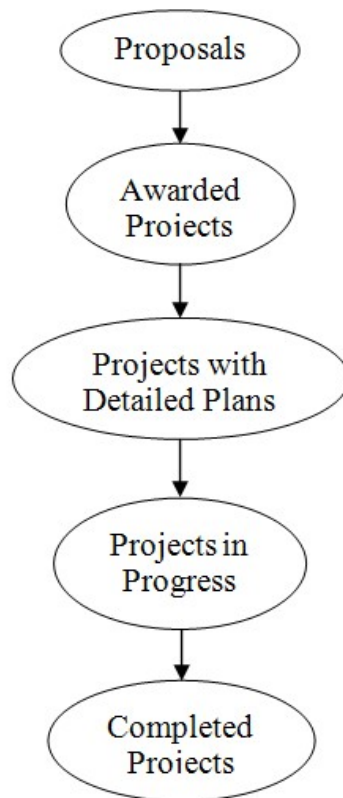


Figure 3.5: Project States in Multiple-Project Environment

Proposed projects are not yet awarded to the contractor. They typically require a rough schedule and estimate of cost of services to prepare initial resource histograms (staffing plans) and verify that the contractor has the capacity capability to successfully perform the proposed project. In most cases, the contractor has to carry the cost of preparing proposals as part of their overheads.

The second state is when a proposal is accepted and is transferred to a revenue-generating internal project. Some contractors initiate an internal project in their management systems as a proposal and then transfer it to a chargeable project, others wait until receiving the Purchase Order (PO) or Contract and then initiate the project in their systems and start charging the project expenses to the client. After that, the internal project goes through the planning stage, where a complete set of schedule, cost and resource baselines are developed. These baselines get reviewed and accepted by the internal project manager and is then submitted to client project management team for approval.

The third state takes place when the project baselines are approved by the client and the internal project is ready to move to the execution stage. The fourth state starts when internal projects are in-progress and the project controls procedure is used to measure progress and evaluate performance against the approved baselines during the execution stage. The final state is when projects are completed and the lessons-learned from them is documented and stored as part of the closeout procedure.

As shown in Figure 3.6, the total hours from all revenue-generating projects distributed over time units is referred to as Workload. The time unit used is work week, since nearly all contractors collect, approve and process timekeeping data on a weekly basis. The term Capacity represents the total weekly hours that can be provided by all personnel to revenue-generating projects. The normal capacity is calculated using normal work weeks (typically 40 hours per week per individual) and maximum capacity is calculated using the maximum hours per week that each individual can provide including overtime.

All workload prior to time = $t_{(now)}$ is called actual workload, which is the aggregation of work hours from completed projects and the completed portions of in-progress projects. All workload after time = $t_{(now)}$ is called expected workload and it is composed of the forecast of required work hours per time-unit for all production resources from the incomplete portion of in-progress projects, projects that have detailed plans, awarded projects and proposed projects.

Nearly every contractor tries to maintain a graph that represents their forecast workload versus existing or planned capacities. In current practices, these graphs are based on best guesses from resource or project managers. Spreadsheets or other simple tools are used to deterministically combine the guesses from multiple projects. These current practices do not reflect the uncertainty or level of confidence around these future forecasts.

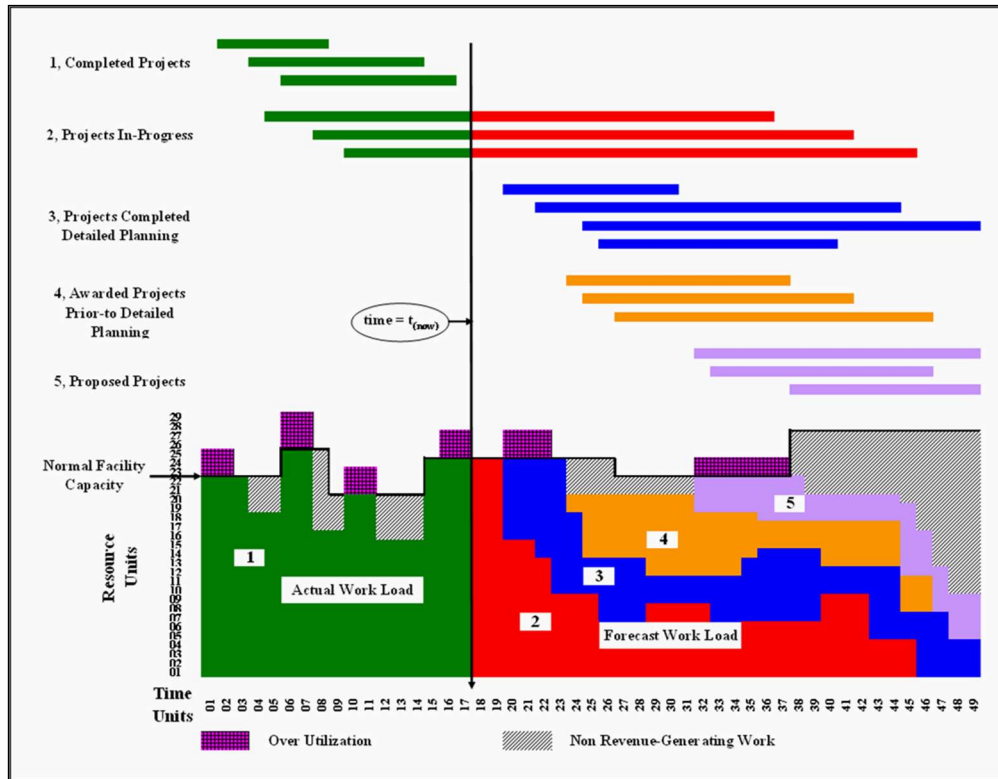


Figure 3.6: Workload vs. Capacity in a Multiple-Project Environment

Contractors are challenged to forecast their workload and optimum capacity to minimize the resources overutilization and idleness as shown in Figure 3.6. Overutilization takes place when individuals have to work extra hours over their normal availability. Working overtime typically increases projects costs and decreases productivity. Idle resources are resources available for utilization but are not assigned to any revenue-generating project (Lova, 2000). Idle resources decrease the profits and increase overhead costs. Contractors have to retain a set of core resources and cannot just release resources, lose their knowledge, and then struggle to hire them back. Also releasing resources impacts the ability to compete for new projects in the future.

There are five main types of contractors in industrial construction: engineering, procurement, fabrication, assembly and site installation contractors. The first four work on their project within a physically fixed location whereas for the site installation work is done in different construction locations. However, all five share the same structure for labour resources management. Contractors are mainly structured as Matrix organizations (Kerzner, 2006). Within this type of organizations, labour resources are grouped in departments of similar functions. Process, mechanical, civil-structural, electrical and instrumentations-controls are examples of these departments for an engineering contractor; while, drafting, welding, cutting, fitting, inspection, painting and fireproofing are examples of these departments for a fabrication contractor.

Functional managers are responsible for the overall management of resources in a matrix organization. They are also responsible for obtaining new resources, providing training and upgrading skills of existing resources. Meanwhile, project managers are responsible for the success of the internal projects within the organization. Typically functional managers are referred to as department heads in a matrix organization.

When an industrial project is initiated by an industrial owner, it is broken to a set of internal projects, each of which is performed by a contractor. Each of these internal projects needs to go through all the five project stages starting with initiation to closeout.

During the initiation stage of any internal project, a Project Manager (PM) is selected to start planning for that project. The PM specifies the required resources that are needed to complete the project. The term resource in this research refers to a certain specific skill of labour resources such as process engineering, electrical drafting, welding or pipe fitting. In some projects, all the resources are needed to complete the project; which means all functional managers must provide teams to work on that project. In other projects, only few resources are required and less functional managers have to provide teams to complete that project. Functional managers assign team members to each project and a team leader to act as resource manager. Typically, these assignments require the approval of the PM. Resource managers may be given different names; in some cases they are called foremen, superintendents, chief engineers, etc. However, regardless of the name change, they still play the same role, which is a resource manager in an internal project.

Internal projects may be grouped in programs based on pre-defined characteristics: it could be the type of the project, the location of the project or the client requesting the work. Typically, in the case of grouping projects in programs, resource coordinators are appointed to manage a pool of resources within each program. In some case, programs are also grouped into portfolios of projects. In large companies, the general manager appoints both a manager of production who supervises the work of functional managers and a manager of projects who supervises the work of project managers as show in Figure 3.7.

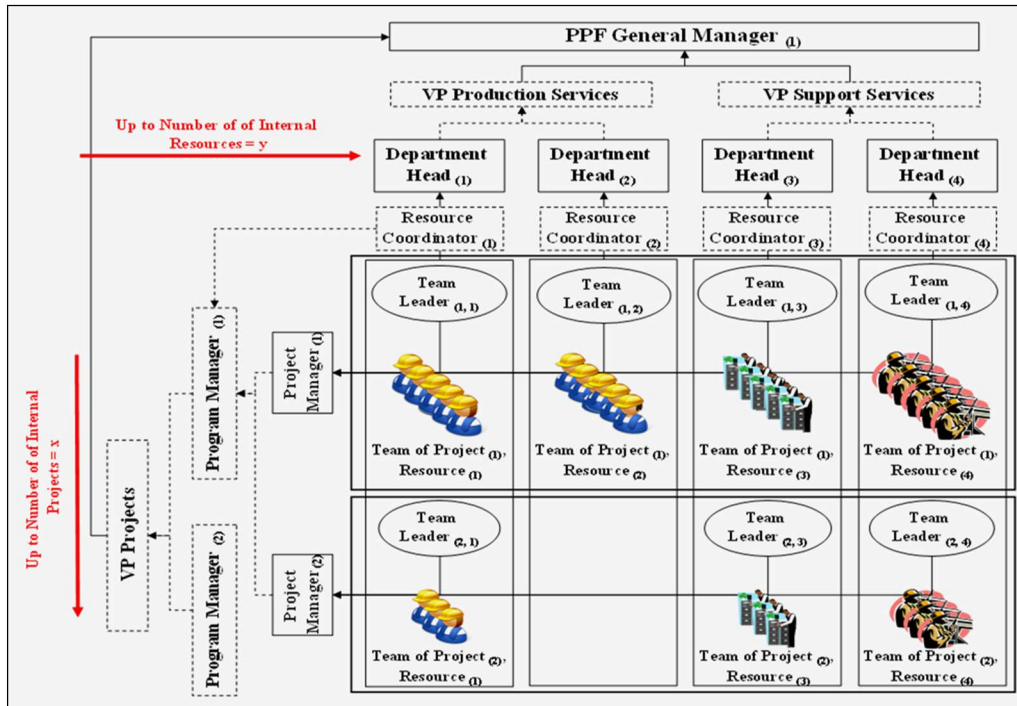


Figure 3.7: Matrix Organization Structure Contracting Companies

Within this hierarchy, resource managers report to both project managers and functional managers. Program managers report to portfolio managers or manager of projects; while functional managers typically report to manager of production. In smaller contracting companies, both project and resource managers report directly to the general manager of the company as shown in Figure 3.7.

All team members have to report to both their functional and project managers, which causes concerns within the matrix organization structure. It may lead to power struggles, lack of proper communications and delays in decision-making (Davis, 1977). In addition, this structure may cause difficulties for functional managers to develop long-term staffing plans due to the uncertainty around

expected work load. Moreover, training and personal development plans are hindered due to projects' commitments always taking precedence over functional needs.

According to this model of matrix organization, several major managerial roles can be assigned to any individual in the common pool of labour resources. These managerial roles include: team member, team leader (resource manager in a project), project manager, program manager, portfolio manager, functional manager, manager of projects, manager of production and general manager. Each of these managerial levels has different responsibilities and requires different views of the resources data reports and graphs.

3.3 THE PROCESS MODELING

Each process is modeled separately for clear illustration of inputs and outputs. Nevertheless, some of these processes occur concurrently in industrial construction projects particularly in nowadays fast-track environment. Each of these processes is modeled through a detailed description, a graphical display, and a list of inputs and outputs.

3.3.1 The Feasibility Study Process

Any industrial construction project starts as a project idea that needs evaluation. These ideas can be either for *stay-in-business* or for revenue generating projects. *Stay-in-business* projects are mandatory ones and have to be performed in order to comply with external regulations such as environmental laws or safety standards. *Stay-in-business* projects have to be performed regardless of their cost or expected revenue. Because of that, these projects move directly to the planning stage to start their engineering. Meanwhile revenue generating projects have to undergo initial feasibility study to assess their impact on the owner's business goals, objectives and plans. Badiru (2008) defined feasibility study as "A study conducted to ascertain the practicality of the proposed product. The practicality is considered in terms of available technology, cost constraints, production process, labour skills availability, organizational goals and market structure". Feasibility studies are sometimes called business-cases. The major outcome of feasibility study is to ensure that the expected Return on Investment (ROI) meets the threshold set by the owner or a joint-venture of owners.

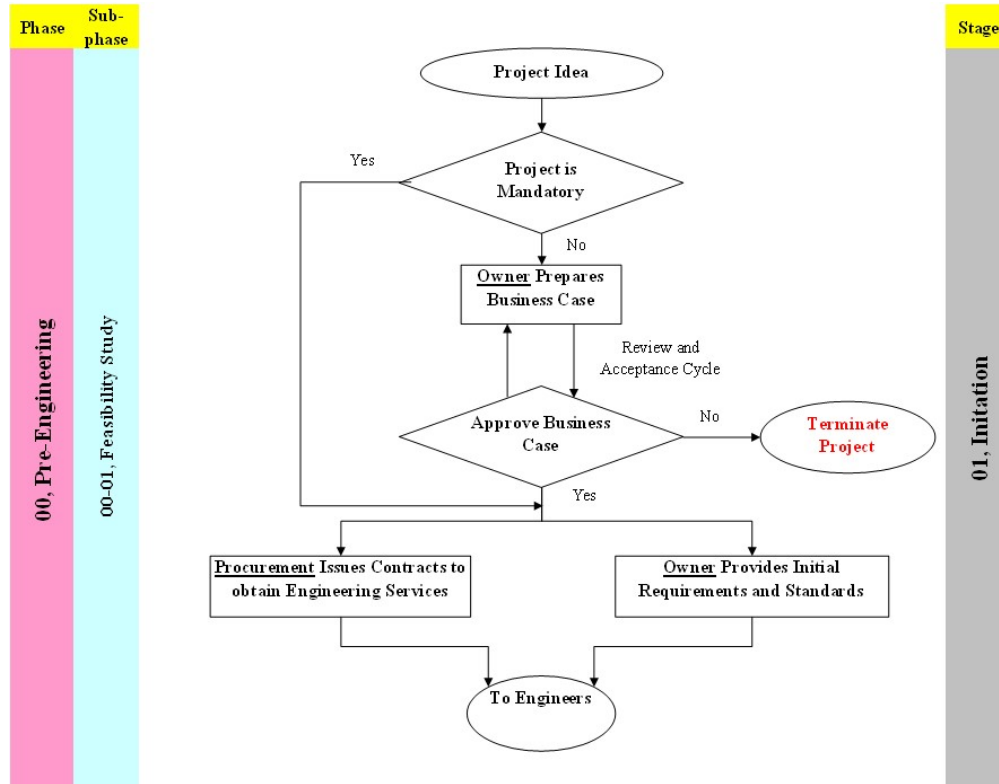


Figure 3.8: The Initiation Process

If an industrial project is mandatory or if its ROI meets or exceeds the required threshold, the owner(s) make a decision to proceed to the planning stage and engineering phase of the project as shown in Figure 3.8. Once this decision is made, the planning stage starts with the procurement unit within the owner(s) organization awards the necessary contracts to an engineer or a group of engineers to start the engineering phase of the project. In some cases the engineering unit with the owner’s organization would perform the engineering for the project if they have the necessary capacity to do so. Contracts are awarded to engineers through bidding processes, long-term agreements or direct orders. Afterwards, the owner has to

provide the engineers with the project initial requirements typically through the feasibility study. The owner(s) also provide a list of standards and specifications to be applied in the project design. These standards and specifications are either owner-specific or developed by other institutions such as the American National Standards Institute (ANSI). Completing this process marks the end of initiation stage and pre-engineering phase and the start of the planning stage and engineering phase of an industrial project.

3.3.2 Processes within the Engineering Phase

Once engineers receive contracts from the owner(s), they initiate internal projects within their organizations in order to provide the required services. Internal projects not only take place in engineering organizations, but also in procurement, fabrication, assembly and construction organizations. Multiple internal projects have to be completed by a number of contractors in order to complete a single industrial construction project. Each internal project goes through the 5 stages of standard life cycle from initiation to closeout. Each of these internal projects utilizes a group of labour resources that are available in the performing organization. Consequently, a large set of resource management data is generated during planning and executing every internal project. The differentiation between internal projects and industrial construction projects is crucial for the proper collection of the generated data. The objective is to collect and store the resources management data from all internal projects in a structured and consistent format that can be

transferred to useful knowledge. This knowledge can then be fed to the planning of new projects.

Each construction project needs to start with a complete scope definition, accurate engineering documents, and realistic Project Execution Plans (PEP) in order to increase the probability of its success. For industrial construction projects, these deliverables are developed during the planning stage and engineering phase through Front End Loading (FEL) planning (Lavingia, 2007). FEL planning is a widely used term in industrial construction and represents one of the major differences between industrial and non-industrial construction projects. FEL planning applies the concept of “rolling wave planning” turning the planning into progressive elaborating process (PMI, 2004). This approach is used to overcome the lack of information at the beginning of any industrial project. It decreases the uncertainty around execution plans with the availability of more engineering and design information. It also maximizes the possibility of optimizing project outcomes by spending extra time, cost and efforts during the engineering phase to avoid massive spending to fix mistakes on the construction site. As shown in Figure 3.9, the ability to influence projects and project uncertainty decrease with time meanwhile project expenses increases specially during the execution stage (Kerzner, 2006). Due to schedule constraints, some projects skip one or more of the FEL processes, which often leads to problems during the execution stage.

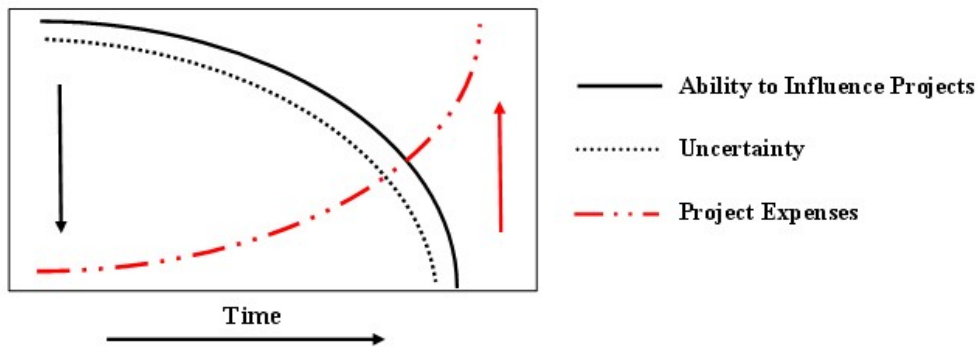


Figure 3.9: Uncertainty and Ability to Influence Projects vs. Project Expenses

FEL planning usually consists of three processes,

- FEL I (sometimes called Identify Alternatives or Scoping Study (SS)),
- FEL II (sometimes called Select Alternative, Evaluate or Design Basis Memorandum (DBM)), and
- FEL III (sometimes called Develop Selected Alternative or Engineering Design Specifications (EDS)).

Each of these processes is modeled in the following sections of this chapter.

3.3.2.1 The FEL I Process

In FEL1, all alternatives to achieve the desired requirements of a project are identified. To identify these alternatives, a complete investigation of latest available technologies (sometimes new technologies are even developed), various chemical processes to obtain the project products and different site layouts are investigated. After that, Value Engineering (VE) or Life Cycle Value Analysis (LCVA),

Constructability Reviews and initial Hazardous Operations (HAZOP) analysis take place. The procurers, fabricators, constructors, operators and maintainers are invited to provide their input to this process as part of their engineering support process. The idea is to gain from their vast experience as early as possible in the project to minimize constructability, safety, maintainability and operability issues.

A high level Total Installed Cost (TIC) estimate, project schedule and Project Execution (or Implementation) Plans (PEP or PIP) are also prepared. The engineers also provide a detailed schedule, cost of engineering services estimate and staffing plans for the next engineering process. In most projects, this output goes through a structured review procedure (Gate or Peer Review) to determine whether to proceed or not to next process as shown in Figure 3.10.

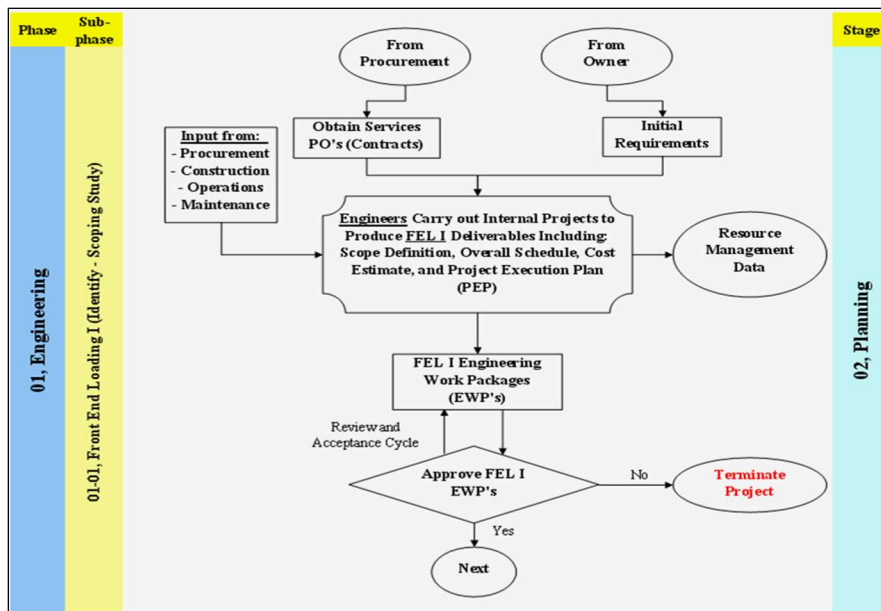


Figure 3.10: The FEL I Process

3.3.2.2 The FEL II Process

In FEL II, as shown in Figure 3.11, all alternatives, which were previously identified during FEL I, are evaluated, compared and one alternative is selected to be developed to a complete design basis. This design basis typically includes the Process Flow Diagrams (PFD), Heat & Material Balances (HMB) and a set of updated engineering documents and detailed plans for the next engineering process. In fast track projects, engineers also prepare Procurement Work Packages (PWP) for some critical items that require long time prior to delivery (typically known as very long lead items). The preparation of the PWP is part of Procurement Support process, which is explained in detail in this chapter. Similar to the FEL I process, a complete evaluation and economical analyses of the FEL II output takes place to ensure that the project is still feasible after obtaining the updated output from FEL II.

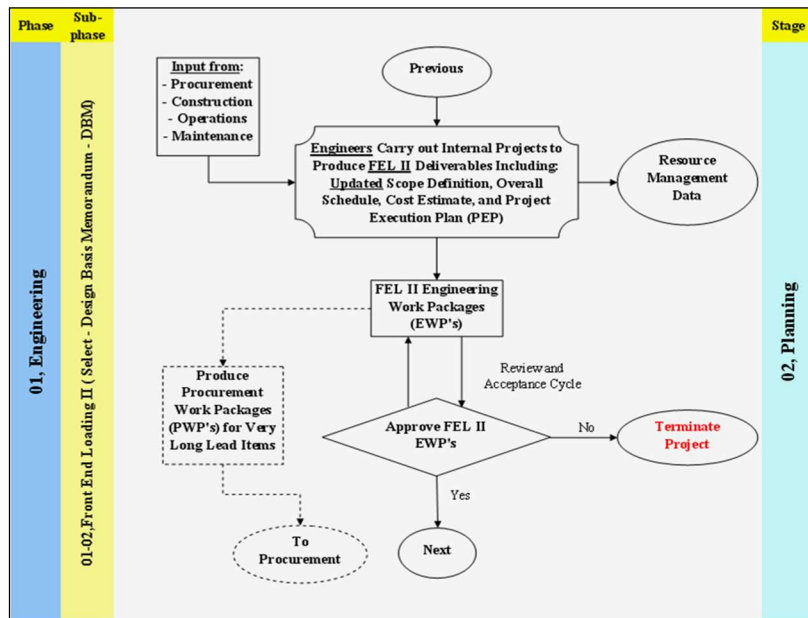


Figure 3.11: The FEL II Process

3.3.2.3 The FEL III Process

In the FEL III process, as shown in Figure 3.12, the previously developed design basis is advanced to a complete set of design specifications. The most important deliverable of this process is the Process & Instrumentation Diagrams (P&ID), which shows all the actual tanks, pumps, pipelines, etc... in the plant accompanied with all the instrumentations required to control this equipment. If a three-dimension (3D) computer model is required for the industrial project, the preparation of that 3D model typically starts in this process. The final TIC estimate and level III overall project schedule are also developed in this process to form the baselines of the project. A formal change management procedure takes place after obtaining the baselines to monitor any changes to the approved scope, budget and schedule. The output also includes detailed plans, schedule and budget for the next engineering process, which is the Detailed Engineering & Design (DED). In fast track projects, engineers prepare PWP for long lead items to meet Required At Site (RAS) dates. This procedure takes place as part of the procurement support process. They also prepare Site Work Packages (SWP) for some critical components such as site clearing or piling to meet specific construction windows. This procedure takes place as part of the Detailed Engineering & Design (DED), which is explained in detail in the next section of this chapter. Similar to FEL I and FEL II processes, FEL III output goes through the last gate review to determine if the project is still feasible, may be deferred or it needs to be terminated. The project fund is also obtained through the Appropriation For Expenditure (AFE), which forms the cost baseline of the project.

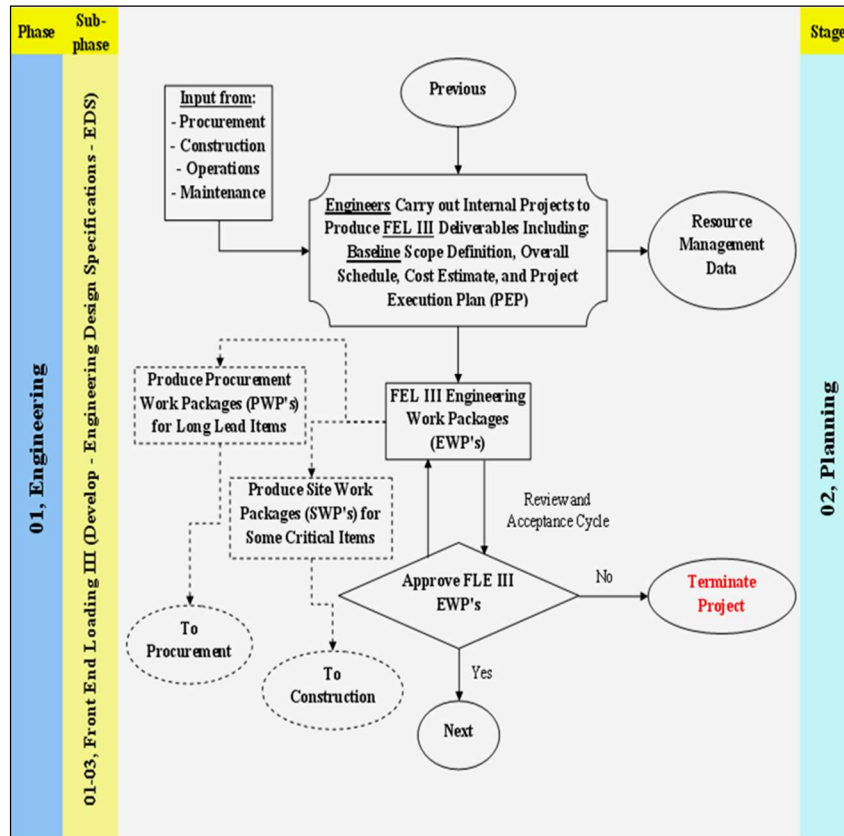


Figure 3.12: The FEL III Process

Obtaining the output of the FLE III marks completion of the planning and beginning of the execution stage. However, as shown with the dashed lines in Figures 3.10, 3.11 and 3.12, some execution packages may be issued during the planning stage.

3.3.2.4 The Detailed Engineering & Design (DED) Process

The Detailed Engineering & Design (DED) process is the lengthiest, most expensive and most resource consuming process in the engineering phase. It can

cost between 10% ~ 15% of the total installed cost of an industrial project, as shown in Figure 3.13. In this process, the previously approved design specifications are transferred to detailed drawings that are grouped in Engineering Work Packages (EWP). These EWP are issued to either, fabrication shops (EWP-F), assembly yards (EWP-A), or construction sites (EWP-S). The main deliverable of this process is the Isometric drawings (ISO's) that are used are either generated from a 3D model or drafted in 2D software. These EWP are used by the fabricators, assemblers and site installers to prepare the necessary detailed drawings during the shop drawings process.

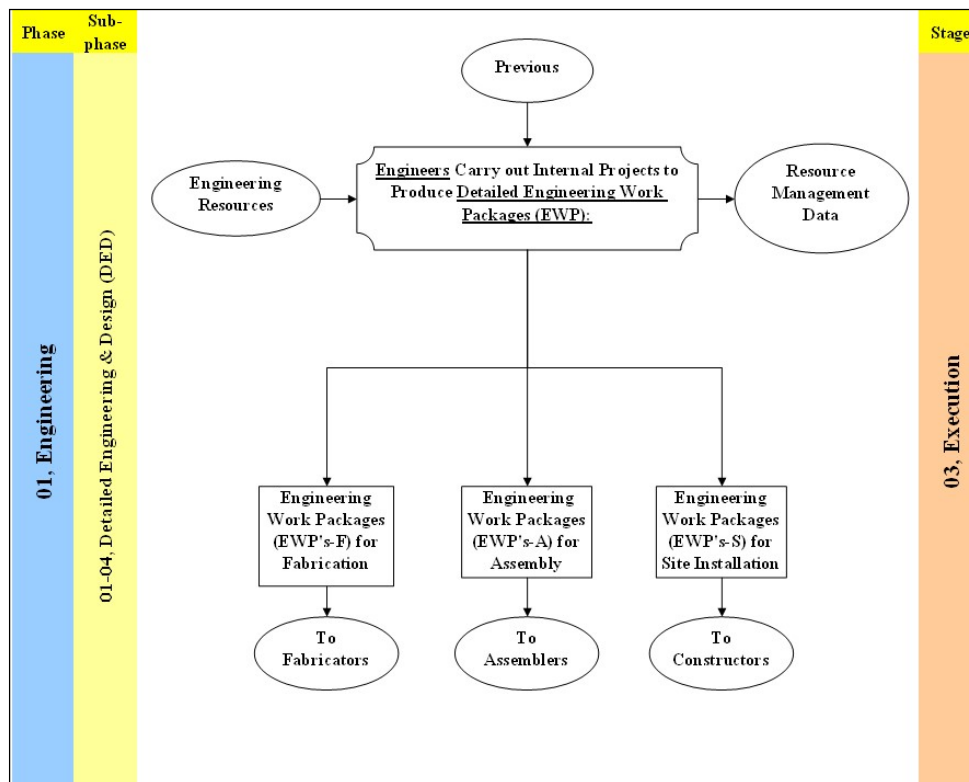


Figure 3.13: The Detailed Engineering & Design (DED) Process

In large and mega projects the engineering scope is frequently performed by multiple engineers working separately or as a joint-venture. In this case, a lot of communication, scope and interface management is required to ensure the quality, completeness and consistency of engineering work packages are maintained by all engineers. This also requires a clear definition of Battery Limits for each project element and the interfaces between them.

3.3.2.5 The Shop Drawings Process

In this process, shown in Figure 3.14, the previously approved engineering work packages (EWP) are transferred to detailed shop drawings that can be used by the fabricators, assemblers and site installers to construct the plants. These drawings are typically prepared by the engineering division of the constructors' organizations. For the structural steel fabrication, the shop drawings include each steel piece and the details of their connections. For the pipe fabrication, the shop drawings include each pipe spool, elbow and valve. For the reinforced concrete, the shop drawings include rebar arrangements.

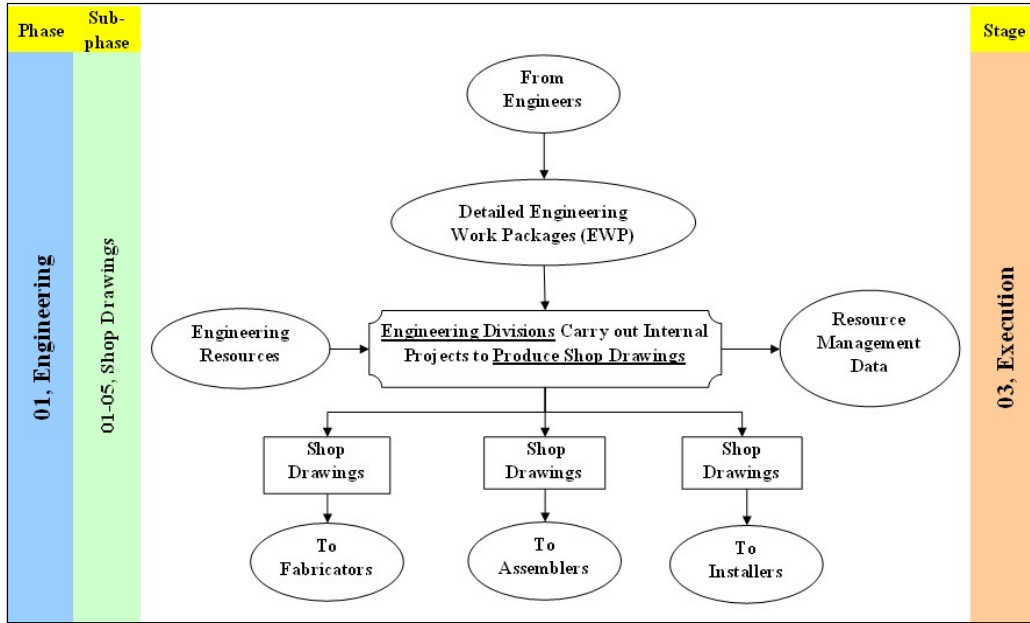


Figure 3.14: The Shop Drawings Process

3.3.2.6 The Procurement Support Process

In this process, shown in Figure 3.15, the engineers prepare EWP for purchasing (EWP-P) and EWP for contracting (EWP-C) to start the procedure of obtaining materials and services. These packages are sometimes called Material Requests (MR), which is not an accurate term since they also include requests for services. These packages should be complete with all the technical specifications, delivery schedules and progress payment plans. After a PO or a contract has been awarded, the engineers also have to answer Requests For Information (RFI) from vendors, approve the vendor shop drawings, perform shop and site inspection and approve

the results of the Factory Acceptance Tests (FAT) and the Site Acceptance Tests (SAT).

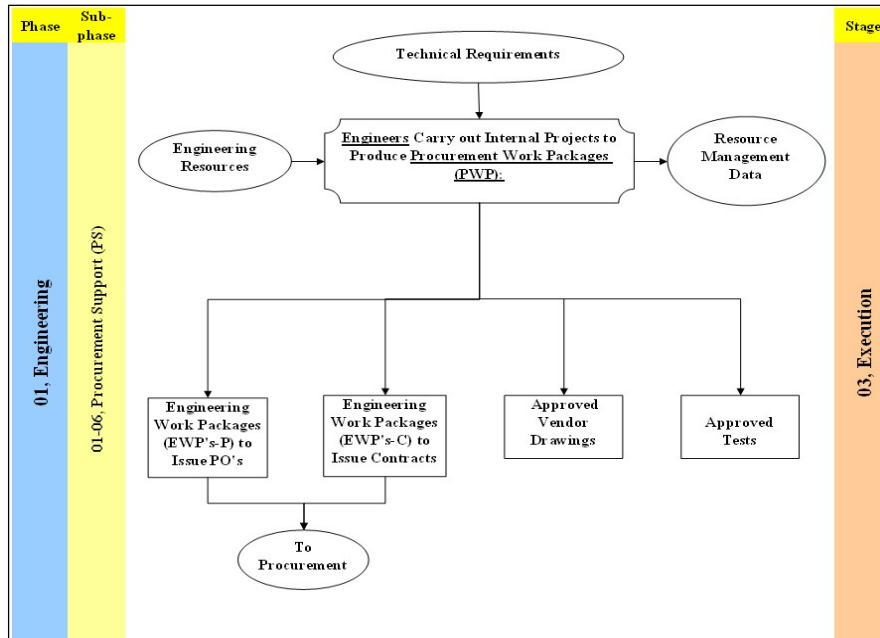


Figure 3.15: The Procurement Support (PS) Process

3.3.2.7 The Construction Support Process

During the construction support process, engineers provide answers to questions or Requests for Information (RFI) from fabricators, assemblers and site installers (Figure 3.16). It is an on going process that can not be scheduled or budgeted in advance. Hence the length and amount of effort spent during this process depends on the amount and type of questions raised. Some engineers may also move to site to solve problems as they occur or work with constructors to modify or understand the engineering drawings.

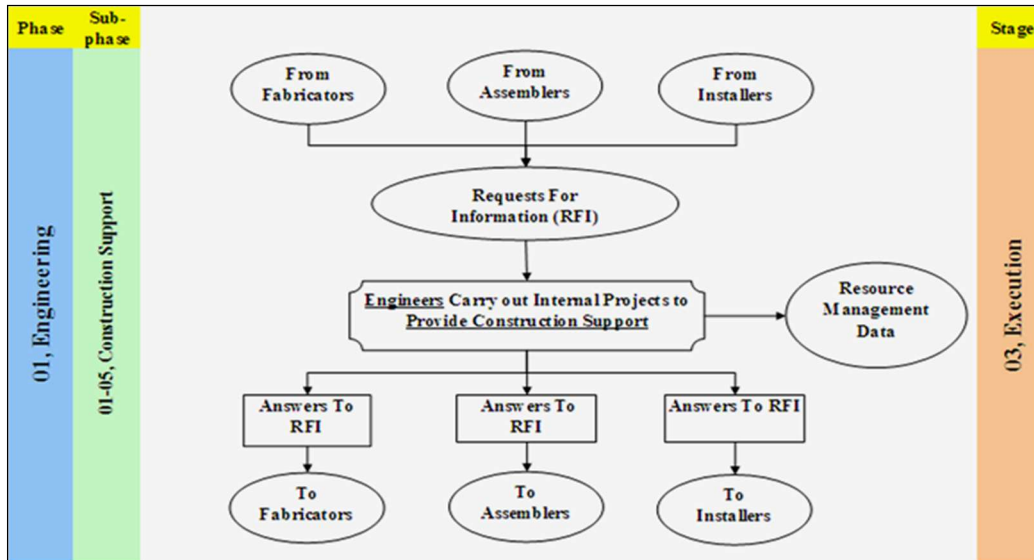


Figure 3.16: The Construction Support Process

3.3.2.8 The As-building Process

After completing the construction of all work packages of an industrial project, it is a common practice to ask the engineers to prepare as-built drawings. The As-Building process, as shown in Figure 3.17, takes place the closeout stage of any industrial project and marks the end of the engineering phase of that project. During this process, engineers produce new or marked-up (red-lined) drawings and documents that represent the final status of all completed products. The plant operators and maintainers refer to these drawings in their work in the plants. These drawings provide very important input to the engineering phase if a decision is made to initiate projects to expand or modify existing plants.

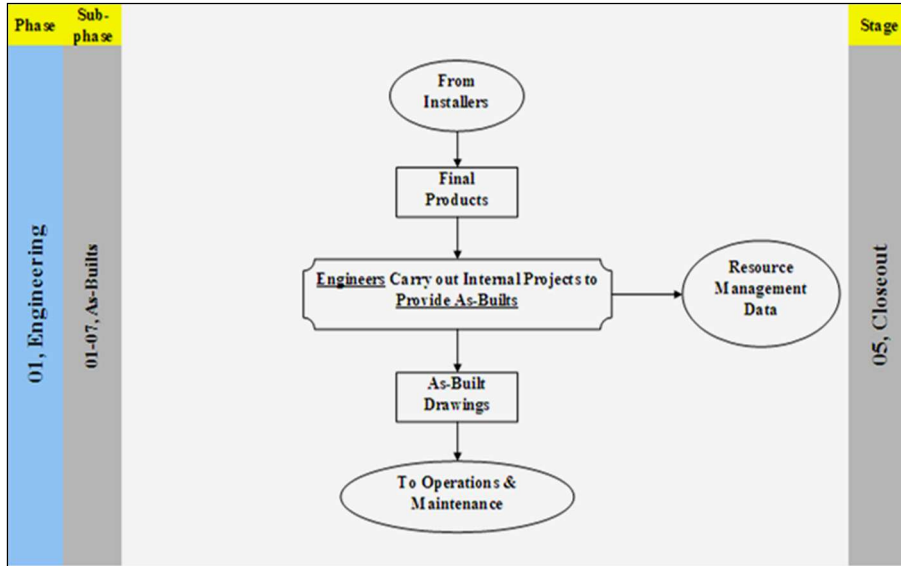


Figure 3.17: The As-Building Process

3.3.3 Processes within the Procurement Phase

Procurement is defined as the “processes required to acquire goods and services from outside the performing organization” (PMI, 2004). According to industry practices, the term *contract* is used for orders to provide services or services and materials by a contractor. Orders to provide materials only are called Purchase Orders (PO). In industrial construction, procurement is mostly handled by the procurement divisions in an Engineering Procurement and Construction (EPC) firm, which acts on behalf of the owner(s), the construction firms or by the industrial owner(s) themselves. In industrial construction, procurement is handled by the three parties (EPC firms, constructors and owners). This type of arrangement requires a tremendous amount of efforts, integration and interface management to avoid problems such as:

- Ordering materials that do not meet the specifications
- Materials or services are not ordered (it is not known who is supposed to order it)
- Delays in receiving materials and services
- Double ordering of the same materials (one party is not aware that another party ordered it)
- Double or even triple handling of materials, which increases project costs
- Shipping to wrong locations
- Lost materials that have to be ordered again

In this research, procurement processes are classified different from the typical PMI classification. The main reasons for this deviation from PMI classification are to fit for the specific needs and nature of industrial construction projects and to match the typical industry practices.

3.3.3.1 The Engineering Support Process in Procurement Phase

In this process (Figure 3.18), the procurement team supports engineers by providing budgetary quotes for equipment and bulk materials to the TIC estimate. The team also provides delivery times to help with establishing the activity durations in the overall project schedule. In addition to that, the procurement team prepares and updates the procurement management plan as part of the overall project execution plan (PEP). They also contribute to the risk management procedure by providing expected risks and their impact, probability and mitigation plans.

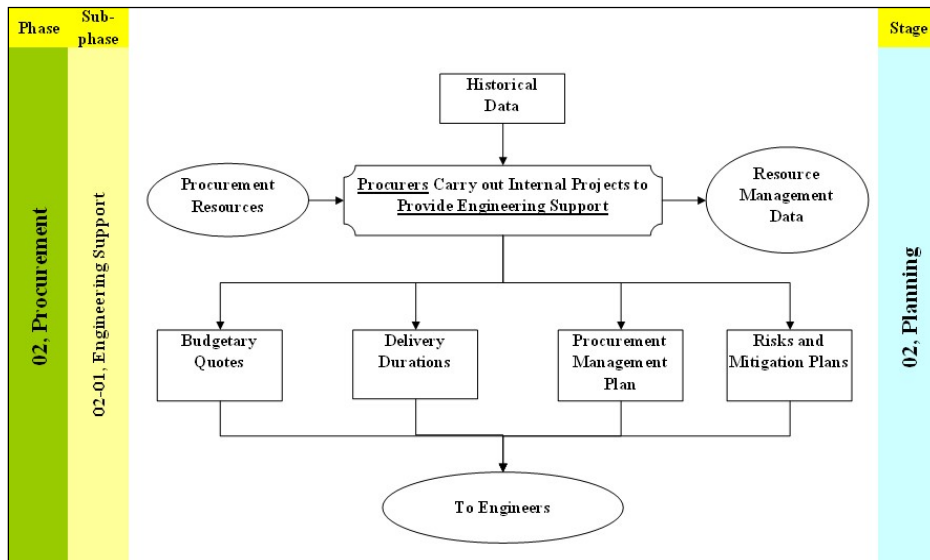


Figure 3.18: The Engineering Support Process in Procurement Phase

3.3.3.2 The Requisition, Bidding & Awarding Process

When a EWP-P is issued to procurement, it is referred to as Purchasing Work Package (PWP). Meanwhile, when a EWP-C is received it is referred to as Contractual Work Package (CWP). The term CWP is sometimes used in the industry to refer to construction work package performed on the construction sites. To avoid confusion, work packages performed on site are called Site Work Packages (SWP) in this research. PWP and CWP are scheduled, budgeted and progressed during the procurement phase in the same way as EWP in the engineering phase. After receiving the packages from the engineers, the procurement team prepares a qualified bidders list for each package. These lists are typically prepared after a pre-qualification procedure or through long-term arrangements such as Suppliers of Choice (SoC) or preferred suppliers agreements. All qualified bidders receive a bidding package (known as Request for Bid -RFB) to provide their estimated prices. Once the bids are received, technical and commercial evaluations are performed. The winning bid gets awarded a PO or a Contract to start delivering the requested materials and/or services (Figure 3.19).

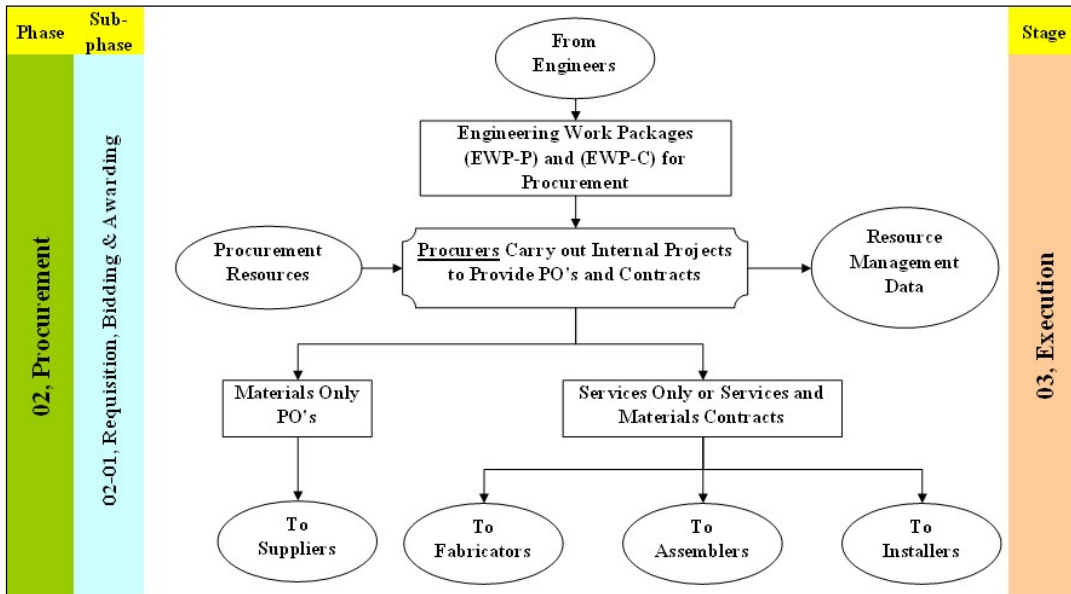


Figure 3.19: The Requisitioning, Bidding & Awarding Process

3.3.3.3 The Contract Administration Process

The contract administration process, as shown in Figure 3.20, includes all activities performed by the procurement division to manage the relationship with the contractors who were awarded contracts from the previous process. Contractors include engineers, fabricators, assemblers and site installers. These activities include expediting, logistics, change management, Quality Control (QC), Quality Assurance (QA), shop inspection and approve invoice payments. Expediting refers to the activities performed to ensure that the required services are progressing according to the approved baseline plans and will be delivered according to the agreed-upon schedule.

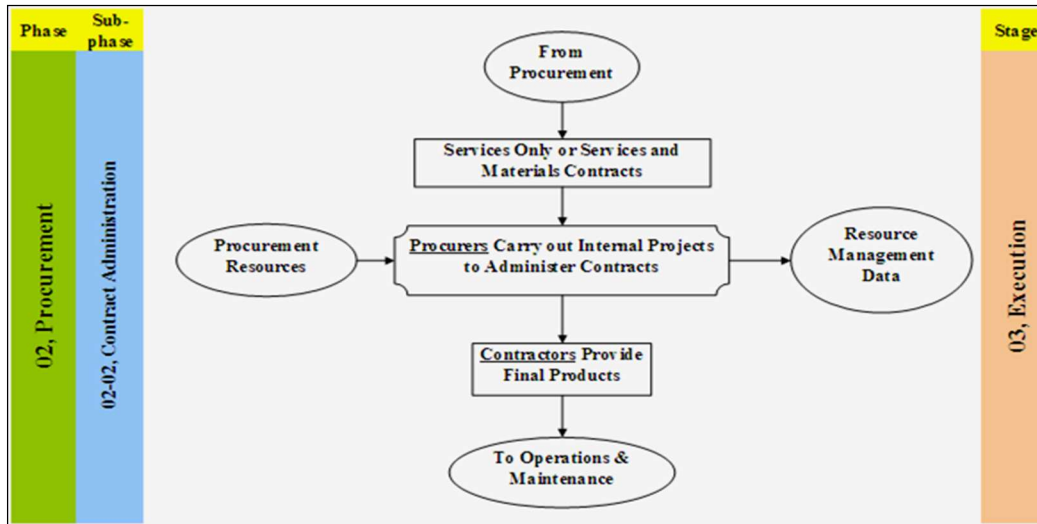


Figure 3.20: The Contract Administration Process

3.3.3.4 The Materials Management Process

The materials management process, as shown in Figure 3.21, includes all activities required to manage the relationship with the suppliers who were awarded PO's during the requisition, bidding and awarding process. These activities include expediting, change management, and invoice payments. The main difficulty in this process is to ensure that materials are shipped to the right destination. A lot of procurers use Material Management System (MMS) software packages to help distribute the materials between fabricators, assemblers and the construction site. Some of these applications can communicate with the 3D modeling application to directly read the materials list.

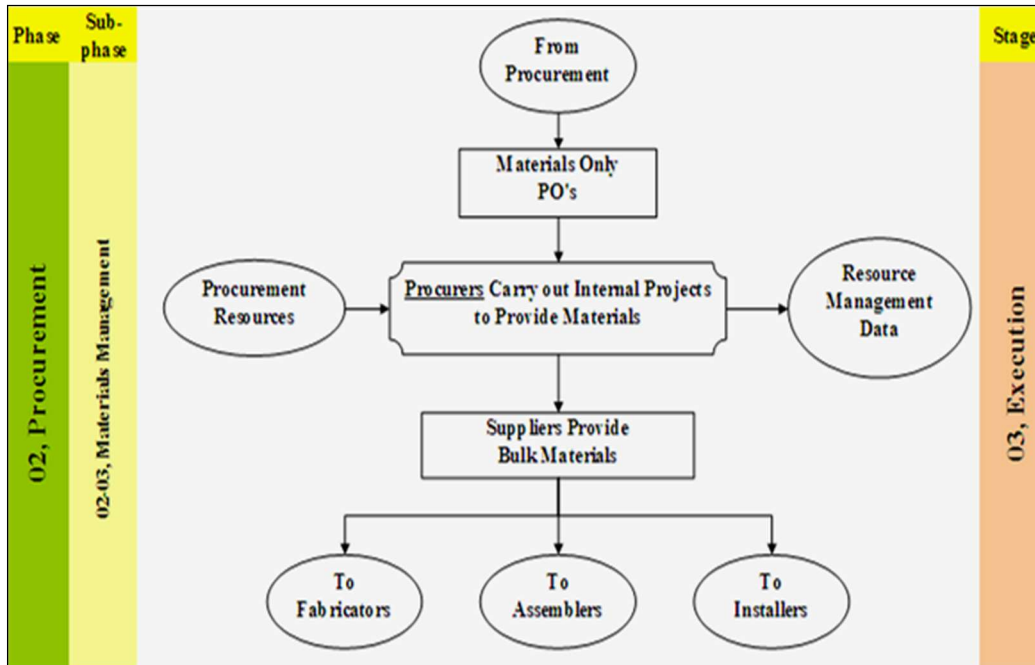


Figure 3.21: The Materials Management Process

This research suggests adding the EWP description as an attribute into the 3D model to help identify the correct shipping location for each piece of materials. As explained before, EWP are categorized according to their receiver. Therefore, having EWP description as an attribute in the model is going to significantly help resolving the problems of shipping materials to the wrong location. Erroneous shipping is one of the main contributors to delays, budget overruns and efficiency reduction in industrial construction projects.

3.3.4 Processes within the Construction Phase

3.3.4.1 The Engineering Support Process in Construction Phase

In this process, shown in Figure 3.22, the constructors (fabricators, assemblers and site installers) provide their input to the engineers to help with the design optimization, site layout, LCVA, constructability reviews and provide expected risks and their probability, impact and mitigation plans. The constructors provide their estimates for the duration and cost of work packages to help with producing the overall project schedule and TIC estimate. They also provide their input to the construction management plan, which is a major component of the overall Project Execution Plan (PEP).

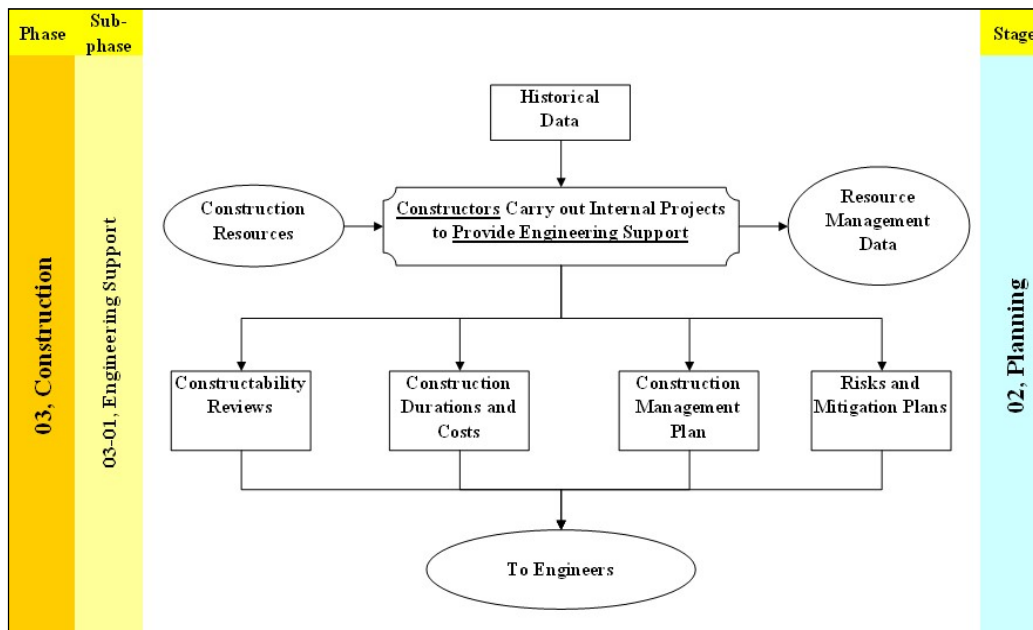


Figure 3.22: The Engineering Support Process in Construction Phase

3.3.4.2 The Fabrication Process

In a typical industrial project, multiple contractors are involved in the fabrication process (Figure 3.23). These contractors include structural steel, pipe spooling, pumps, tanks, vessels, compressors, exchangers and other equipment fabricators. After the contracts are awarded to the fabricators, the process starts by receiving the shop drawings from the engineers and drafters and the required materials from the procurers. After the products are fabricated, they get shipped either to the assembly yards or directly to the construction site. Shipping the right material to the right location on the right time is a major challenge in any industrial project and is handled during the material management process. Some fabricators use barcodes to ensure that all pieces that belong to the same package are shipped, stored and delivered together to the right location (Hajjar, 1999).

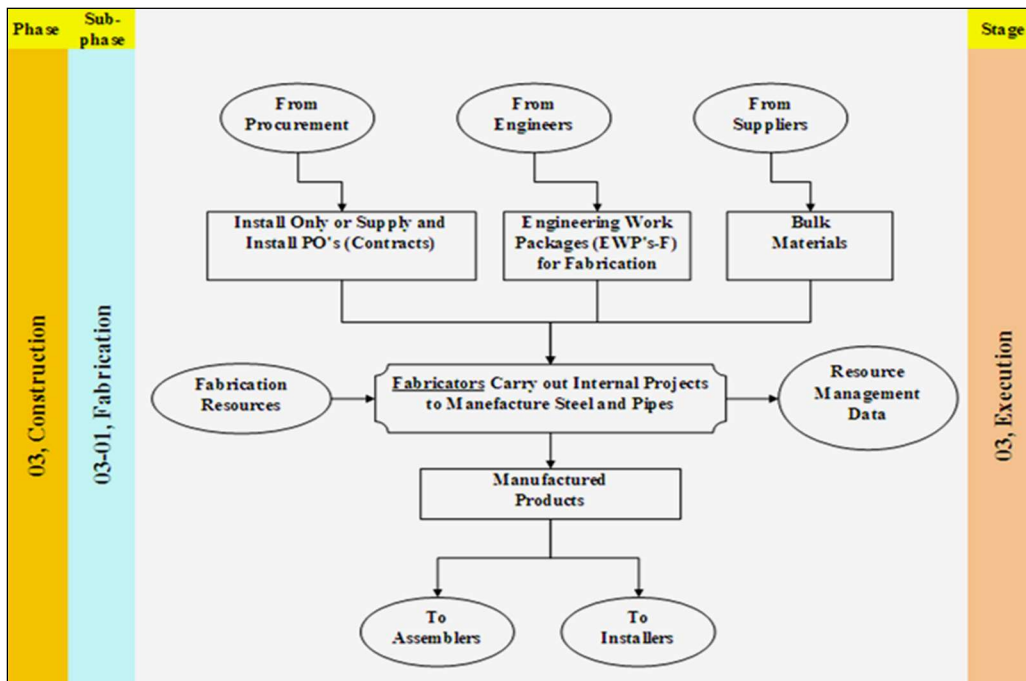


Figure 3.23: The Fabrication Process

3.3.4.3 The Assembly Process

The assembly process, as shown in Figure 3.24 takes place in module yards outside of the construction location. Modular construction is used to maximize product quality and labour productivity and minimize environmental impacts on construction, risks, costs, safety concerns, rework on site and projects' duration (Mandel, 2007). During this process the module's structure steel is received from the fabricators and the maximum possible amount of pipes, electrical cables, insulation, fireproofing and control instruments are installed prior to shipping to the construction site.

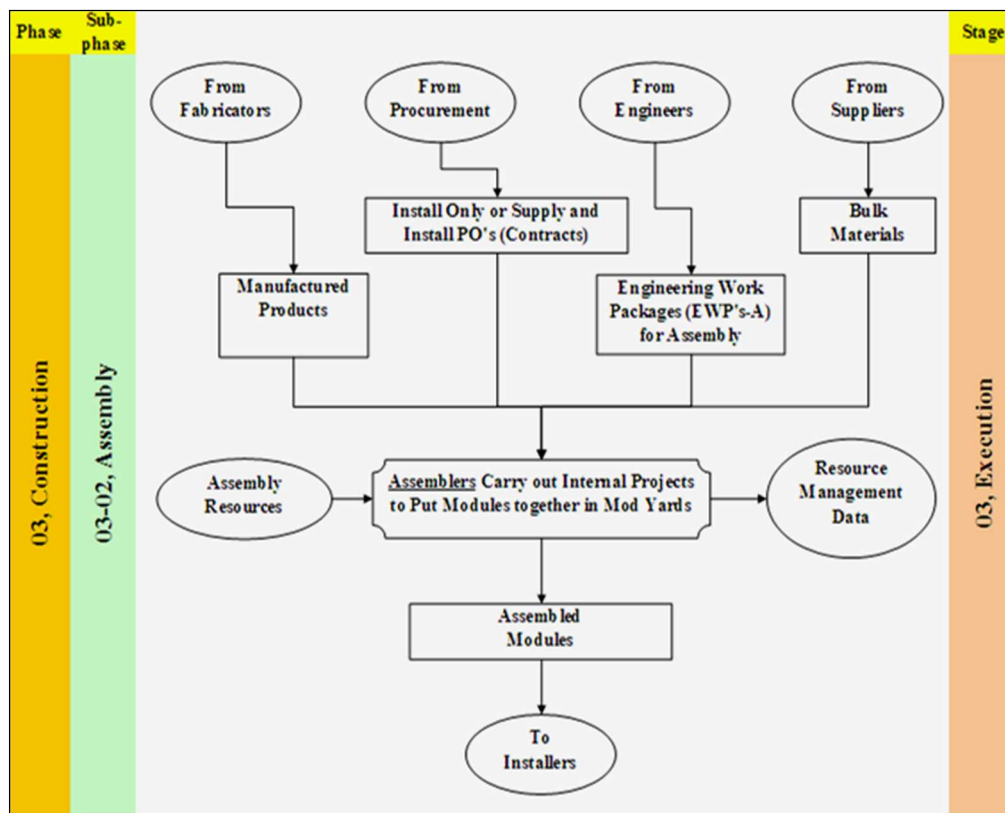


Figure 3.24: The Assembly Process

3.3.4.4 The Site Installation Process

Site installation (as shown in Figure 3.25) refers to the activities of building the final product of an industrial project on the construction location. These activities include site preparation, rough and final grading, pilling, foundations, modules installation, electrical and instrumentation cable wiring, etc.

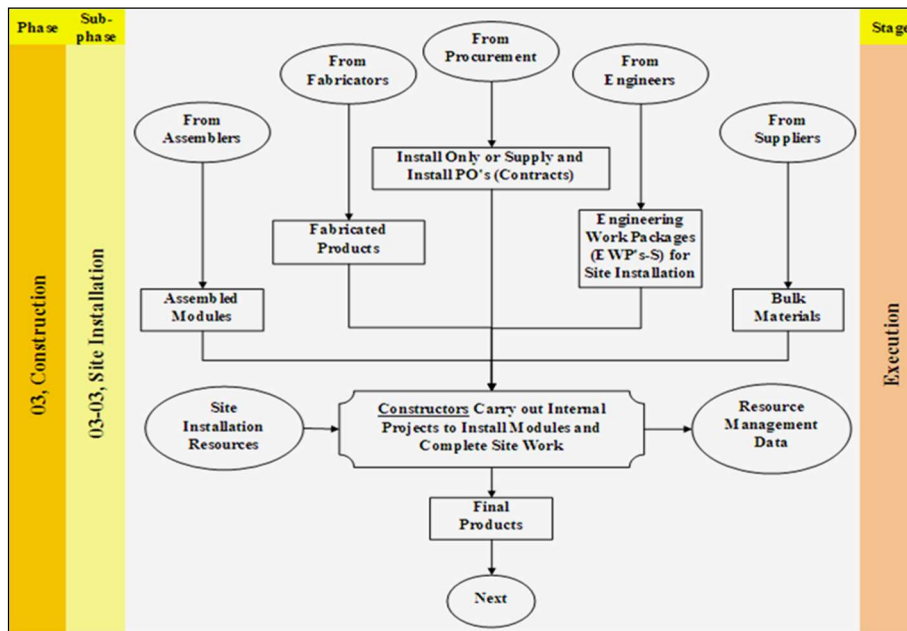


Figure 3.25: The Site Installation Process

CHAPTER 4: RESOURCES MANAGEMENT DATA

4.1 MULTIDIMENSIONALITY OF RESOURCES DATA

Resource management data is generated throughout the different processes of industrial construction projects. Any industrial construction project is planned and executed as a set of internal projects, each of which is performed by a contractor. In this environment, each contractor manages multiple internal projects using one common pool of resources. Managing this one common pool of resources in a multiple-project environment is a very complex procedure. Each contractor needs to monitor the baseline plan, current performance and actual utilization of resources for every undertaken project in order to properly manage their resources. The term *resource management* is a broad term that touches on a variety of aspects in the overall project management practices. Resources cannot be managed without proper management of the scope, time and cost of projects.

First of all, for a contractor to obtain the resource baselines during the planning stage, it has to obtain both schedule and cost baselines. These two baselines are necessary in distributing the required amount of resources over the planned durations of a given project. However, the original scope of each project must be clearly defined prior to obtaining both the cost and schedule baselines.

When projects proceed to the execution stage, actual resource utilization data is collected as part of the project controls tasks. Both the baseline plans and collected actual utilization data is used to calculate the current performance of each resource

in every internal project. Future prediction of resources requirement is determined by the up-to-date actual performance: if performance is below expectation, more resources are needed to complete the project, and vice versa. Furthermore, the current performance of a resource is impacted by the technical and managerial capabilities of all the managers responsible for resource management procedure.

In summary, resource management data has five dimensions: scope, schedule, cost, performance and responsibility. These five dimensions of resource management data are intertwined; each of them has an impact on the other four. When scope is complicated, it is expected that the project would require more time, costs and resources. When the project has time constraint, it is expected to cost more and use more resources. When a project has cost constraint, it is expected to take longer and consume fewer resources to avoid extra costs.

4.2 CURRENT RESOURCES MANAGEMENT PRACTICES

In order to analyze current resource management and data collection practices, a set of formal interviews were conducted with a group of industry experts from different owners and contractors. The purpose of these interviews was to develop better understanding of the current resources' data generation, collection, and utilization practices in these organizations. The experts were asked to provide feedback on the existing systems and recommendation for improvements to these systems. The questions targeted the five dimensions of resources management data, scope, schedule, cost, performance and responsibility. The questions covered the

project life cycle from initiation to closeout and the three main groups of phases: engineering, procurement and construction. The interviews focused on data generation practices, data storage systems and data utilization after storage if any.

The following is a list of some of the questions that were used in the interviews:

- 1) What information do you record when initiating projects?
- 2) What is your project initiation software?
- 3) How do you define projects' scope?
- 4) Do you use Work Breakdown Structure (WBS)? If yes, how? If no, why?
- 5) Do you use formal definition for scope of services? How?
- 6) What is your labour resources structure?
- 7) Who assigns resources to projects?
- 8) Is there a structured method to assign managers of project resources?
- 9) How do you define schedule activities for your projects and your client's projects?
- 10) What is your approach to calculating activity durations?
- 11) Is there a standard method to organizing (Codes, Phases, WBS) schedules?
- 12) Is there a standard set of resources to load the schedule?
- 13) Do you utilize actual duration data from completed projects into new projects' schedules?
- 14) Is there a standard practice to estimating hourly budgets?
- 15) Do you use a formal structure for your TIC estimates? How?
- 16) Is there a standard Code of Account (CoA) to be used in all projects?

- 17) Do you need to recast your estimates into budgets? How?
- 18) Do you utilize actual cost data from completed projects into estimating budgets for new projects?
- 19) Do you develop baseline resource histograms and staffing plans?
- 20) What method do you use for measuring performance?
- 21) Are you able to obtain multiple-project reports? How ?
- 22) Does every project in your company have frozen baselines?
- 23) When do you update project frozen baselines and how?
- 24) Is there a formal procedure for change management?
- 25) How do you record reasons for change?
- 26) How do you implement project changes into budgets, schedules and performance measures?

The literature review, formal interviews and author's experience provided the foundation for analyzing the current practices of resources management. The analysis is used for identifying the data elements to be collected and stored in the data warehouse.

4.2.1 Scope Management Practices

When a construction project is initiated by an industrial owner, either the owner organization or an EPCM firm, acting on behalf of the owner, starts defining the project overall scope and WBS during the Front End Loading (FEL) phases. The main objective of using the WBS is defined by Chehayeb (1996) as to divide the

scope of a project into discrete and definable components by forming a hierarchical division of the end product of the project. However, the current scope management practices misuse the WBS and do not produce components that are discrete and fully definable.

Many projects do not focus on dividing the end product, but they add levels such as responsible discipline, cost categories and project phases to the WBS leading to problems and confusion when managing the projects. After defining the original baseline scope of a project, it is supposed to be frozen and can only be changed through a formal change management procedure. However, in reality this procedure mostly fails because the original baseline scope was not clearly defined and never frozen leading to extreme difficulty in defining current project scope.

The overall scope gets typically broken down to n number of internal projects; each of them is performed by a contractor. Developing the overall scope is an iterative process and takes place through the three FEL processes as explained previously in Chapter 3 of this thesis. When a contractor receives a request to perform part of the scope of an industrial construction project, the contractor has to initiate an internal project to perform this scope. After that, the contractor has to work with the owner to clearly define the scope of services for the internal project. Different methods are used to document internal projects' scope of services. These methods include: contracts, PO's, scope statements, emails, deliverables lists, etc. It is still a common

practice in the industry that scope of services is not even documented anywhere and it remains an ambiguous verbal agreement.

Several issues arise from the analysis of the current scope management practices in industrial construction projects:

- Many projects progress to the execution stage without having spent enough time in the planning stage and FEL planning to ensure that the project has a clearly defined overall project scope and a documented description of the scope of services for all internal projects. Jergeas (2008) stated that incomplete scope definition and inadequate FEL planning is one of the main causes of overruns in industrial construction projects.
- Lack of a consistent method in defining overall project scope and internal projects' scope of services, which leads to grey areas that create conflict.
- Lack of a clear strategy to ensure that project's WBS include only the physical breakdown of the scope of work and do not mix phases, responsibilities and cost categories with the WBS.
- When scope is not clearly defined, team members waste valuable time, cost and resources trying to figure out who is supposed to perform project tasks.
- High levels of stress and frustration amongst team members due to ambiguity.
- Difficulty in managing project changes and defining impact of these changes on project baselines due to the lack of a clearly defined original scope.

- Lack of proper scope definition can lead to claims and conflicts that lead to major delays and extra costs due to litigation fees, time and efforts spent to manage the claims. Projects that end in litigation bring severe damage to organizations' image and reputation.
- Problems with scope definition would lead to unsatisfied clients and thus loss of business opportunities for contractors.

4.2.2 Schedule Management Practices

The development of an industrial construction project overall schedule typically starts in parallel with the process of scope definition during the FEL phases. The schedule is prepared as baseline during the planning stage and is then used to control the project during the execution stage.

The overall schedule starts as the level I summary schedule during the FEL I phase of any industrial project. The level I summary schedule is prepared for executive managers and contains key milestones and major summary activities of the project. The level II schedule is prepared during the FEL II phase and contains the milestones and major activities for the engineering, procurement, construction and commissioning of the project.

The level III detailed schedule is prepared by the end of FEL III phase and is used to control the execution phases of the project. It includes a complete set of logical, sequential and properly linked activities in a network format that are used to

determine the project completion date and its critical path. Finding the optimum level of detail for level III schedules is a major challenge for all project planners. If the schedule doesn't include enough details, it can't be used to properly control the project. If the schedule is too detailed, it becomes very lengthy, difficult to be updated and loses its value as project controls tool.

Since any industrial project is performed as a set of internal projects, the detailed level III schedule is supposed to include all major activities from the detailed schedules of each internal project. Each of the participating contractors is asked to provide a detailed level III schedule and submit it to be included in the overall project schedule. These schedules vary in the tool used, format, level of detail, coding structure, amount of constraints, open-ended activities and ability to apply proper logic to link dependent activities.

Overall project schedulers face a major challenge trying to integrate these inconsistent schedules into one overall baseline schedule. These schedules tend to be too lengthy, (it can go up to 50,000 activities or even more), which makes it nearly impossible to use them for proper management and control of the industrial project. After obtaining the baseline schedules, they have to be updated with actual dates of completed activities during project execution.

There are no defined standards to control the level of detail in the overall project schedule and the schedules of internal projects. Major industrial owners and EPCM firms are trying to introduce some consistency to the project scheduling procedure.

However, these procedures vary between each firm and many contractors struggle with adherence to these procedures. More and more industrial projects are performed using the Joint Venture (JV) approach due to lack of a single firm that is capable of performing the complete scope of work. This approach adds more pressure on the project overall schedulers due to the conflicts between various procedures and not knowing which one should take precedence.

Managing schedules of multiple internal projects is another very challenging task, especially when using the traditional and commercially available CPM-based scheduling applications. These applications provide acceptable results only when projects are not constrained by time or resources (Hegazy, 1999). However, this is not a realistic situation, in most cases projects are time-constrained, resources-constrained or both. Chehayeb (1996) also stated that CPM-based applications do not reflect the proper resource utilization and interaction. He also stated that these applications force schedules to be too lengthy because all reporting is performed at the activity level. Consequently, these applications don't produce acceptable results especially in scheduling multiple projects.

Moreover, the CPM applications produce unrealistic critical path(s) that do not include the activities that require management attention. These critical paths are heavily impacted by the amount of constraints in the schedules. The start-to-start (SS) and finish-to-finish (FF) relationships between activities cause sometimes illogical results. The produced schedules are mainly driven by dates and durations

and not by resource availability. In addition, resource leveling can be done using only one resource at a time. Meanwhile, these applications allow the users to load activities with multiple resources making it nearly impossible to optimize project schedules based on resources availability constraints.

When these applications were first introduced to the project management industry, they were originally designed to handle one project at a time not a set of multiple projects in a contracting organization. As the demand grew for tools to simultaneously schedule multiple projects, these applications introduced portfolio management in their latest versions. These applications do not provide a systematic methodology to manage multiple project schedules; it is left to independent schedulers to define their own rules. There are some contractors that try to enforce a standardized scheduling procedure. However, in most cases, each scheduler is still able to define the layout of the project schedule, activity codes and resources according to his/her liking. Even though this autonomy allows for maximum flexibility, it hinders the ability to open multiple projects simultaneously in the same layout. It also makes it difficult to obtain multiple-project reports, graphs and resource histograms. This lack of coding standardization makes it nearly impossible to extract the knowledge gained regarding the duration of completed activities and utilize this knowledge for better estimating of new activities' durations.

Given the complexity of managing multiple project schedules with common pool of resources, several issues arise from the analysis of the current scheduling management practices in industrial construction projects:

- Listing thousands of project unnecessary activities on the overall project schedule makes it extremely lengthy and very difficult to be updated. Users, such as schedulers, resource, project and program managers, struggle to extract needed information out of these lengthy schedules.
- In these very lengthy schedules, critical path(s) are not realistic and are impacted by constraints, illogical relations and can be misleading.
- Focusing on critical path and not on critical portions of work that have to be completed all together lead to project delays and falling behind schedule.
- In many cases, project schedules are built from scratch without using templates or completed schedules from previous projects, which is a major loss of time, effort and money.
- It is very difficult to link the overall project schedule to all the schedules of its internal projects.
- Contractors struggle trying to manage multiple internal project schedules using the existing tools that are originally designed to manage a single project at a time.
- There is no consistent procedure on how to develop internal project schedules.
- Industrial owners also struggle trying to manage portfolios of multiple industrial projects schedules.

- In most cases, there is no consistency among schedulers when using activity names, activity codes and project resources.
- In many projects, activities are added after freezing the baseline leading to activities that do not have a baseline.
- Since the codes and names of activities are not consistent between projects, it becomes very difficult to extract knowledge regarding actual durations from previously completed projects.
- When analyzing schedule risks, project leaders from engineering, procurement and construction are asked to provide their expectations of the optimistic, pessimistic and most likely durations of critical activities. Most team members provide these expectations without utilizing historical records since these records are very difficult to obtain. Because of that, it is nearly impossible to obtain consensus from the team members, which decreases the quality and reliability of the obtained probabilistic schedule. These schedules are hardly used in the industry and are only developed to meet the FEL planning requirements not to manage projects.

4.2.3 Cost Management Practices

The current cost management practices for any industrial project consist of three main components. These are cost estimating, cost baselining, and cost control. Cost estimating is the procedure of forecasting the cost of completing the approved baseline project scope within the timeframe defined by the baseline schedule. The

level of accuracy of the cost estimate improves with the availability of design information during the FEL planning phases of the project.

Once the planning procedure is completed, the cost estimates is frozen and an Approved For Expenditure (AFE) budget is obtained to fund the project. Cost baselining or recasting is the procedure of distributing the approved baseline cost estimate over the project elements to obtain the baseline budget for each of these elements. Combining the baseline budget and schedule provides the project cash flow. The procedure of cost control starts when project proceed to the execution stage. Cost control involves updating the original cost baseline to reflect the impact of project changes and maintain the current baseline. It also involves obtaining the actual cost of each project element and comparing it to the current baseline. In addition, cost controls also involve forecasting cost at completion of each project element if it is expected to deviate from the current cost baseline. Although the three cost management procedures take place at the industrial project level, a similar procedure has to be applied to ensure proper cost management for the scope of services of every internal project.

Cost of labour resources is a major component in the project cost baseline. Cost of labour resources is always estimated by predicting the number of hours and the multiplication of this number by the average cost of each hour. The main challenge in estimating the cost of labour resources is the uncertainty around the required number of hours to perform a given task. The estimated hours are impacted by the

human factor, i.e. labour resources cannot be bought off-shelf like materials, as well as the productivity of the team is dependent on the harmony between team members and the productivity of each individual. Labour cost is obtained through recording the spent hours by individuals and then multiply these hours by the cost per hour for each individual. When planning, an average hourly rate is assumed by resource type. These average rates are also very difficult to be estimated and they are impacted by current market conditions, structure of the team (% of senior vs. junior team members), currency issues in international projects and length of project duration.

Since there is no consistent way in the industry to define project elements and the standard resources required to complete the project, estimating the cost of labour resources becomes a difficult task. The estimates for labour resources cost are usually prepared on different levels of detail. Some estimates are minutely detailed while others are general and not detailed enough.

Cost baselining is also very challenging task and cost controllers struggle with recasting estimates due to the difficulty of defining project elements. The recasting procedure requires experienced cost controllers who have to work with the cost estimators and the project management team, and end up spending extra time and effort trying to come up with a an acceptable cost baseline. When estimates are not recasted properly, the obtained cost baselines are not accurate. Inaccurate cost baselines lead to major issues with measuring project performance. When project

performance measures are not reliable, projects suffer from budget overruns and schedule delay due to the postponement of taking corrective actions when they are needed on time (Nassar, 2004).

After obtaining the cost and schedule baselines for internal projects, planned schedule activities are loaded with approved budgets to obtain hourly resource histograms or resources baseline. Project resource managers have to staff these resource histograms with qualified individuals who are capable of performing the tasks on time, budget and according to the predefined quality specifications. These staffing plans typically require the approval of the contractor's project manager and resource managers and the owner's project manager. Many projects proceed to the execution stage without resource baselines causing staffing and performance measurement problems. Many projects focus on obtaining cash flow graphs instead of the hourly resources baseline, which is not enough for proper management of labour resources in any project.

When a contractor is managing multiple projects at a time with once common pool of resources, the cost management procedure becomes very complicated. Given the uncertainty of forecasting project cost and the challenges posed by managing labour resources in multiple-project environment, several issues arise from the analysis of the current cost management practices in industrial construction projects:

- It is very difficult to combine estimates and cost records from multiple internal projects that form a single industrial construction project.

- Industrial owners struggle to combine estimates and cost records from multiple industrial projects to analyze the cost of labour resources between these projects.
- Contractors struggle to combine estimates and cost records from multiple internal projects to analyze the cost of labour resources between these projects.
- Project management teams do not spend enough efforts to clearly and completely define the Project Execution Plan (PEP) and rely on the recasting procedure to transfer cost estimates to baseline budgets.
- The recasting procedure is inaccurate, subjective and consumes lots of time, efforts and expensive resources.
- Performance measurements are not accurate due to the lack or inaccuracy of the resource baselines.
- Inaccurate performance measurements delays corrective actions leading to completing projects over budget.
- It is very difficult to obtain multiple-project cost reports.
- The inconsistency in generating, collecting, and storing cost management data makes it very difficult to utilize this data for data mining and better estimating of new projects.
- When performing cost risk analysis to determine the contingency amounts, project leaders from engineering, procurement and construction are asked to provide their expectations of the optimistic, pessimistic and most likely costs of critical cost accounts. It is also difficult to define the critical cost

accounts due to the lack of historical records. Most of team members provide these expectations without utilizing historical records, which are very difficult to go through. Because of that, it is nearly impossible to obtain consensus from the team members, which decreases the quality and reliability of the obtained contingency amounts.

4.2.4 Performance Management Practices

The most common practice for evaluating labour resources performance and is the Earned Value Management (EVM) technique. EVM integrates scope, budget, schedule and resources to objectively evaluate project performance. (PMI, 2008). The use of EVM started in industrial manufacturing as a financial analysis tool and later was adopted by the United States Department of Defense (DoD) as a project management tool in the 1960s. It's capable of representing both cost and schedule using hours or currency amounts and provides various performance measures that can be implemented to forecast “Estimate To Complete” (ETC) and “Estimate At Completion” (EAC).

The current practices of EVM in industrial projects utilize *work hours* to measure performance of resources using three variables. These variables are Planned Values (PV), Actual Values (AV) and Earned Values (EV) all represented in work hours. These values are used to measure, calculate and summarize project performance at any required level.

As mentioned previously, resource baselines are developed during the planning stage of any project. These baselines represent the planned values for each resource. EVM is typically implemented at the resource level using planned, actual and earned values. Planned values are obtained by distributing baseline hourly values over the baseline schedules. The actual values are obtained from the time keeping system. Earned values are obtained by multiplying the current hourly budget by actual physical percent complete. Obtaining the actual physical percent complete from the resource managers is the most challenging task in EVM and several methods are used to increase the accuracy of the obtained values. Some contractors measure progress on a weekly basis, others measure every two weeks, twice-a-month or monthly.

There are specific issues and problems that arise when it comes to combining project schedules and costs to obtain resource histograms, which represent the baseline for EVM. Some of these issues are:

- Even though, EVM is the selected performance measurement method for almost all industrial projects, there is no consistent method to define the level of detail and how to apply EVM to all components of a project.
- The lack of using pre-defined agreed-upon project attributes for analyzing performance data, for a specific resource, from all internal projects.
- There is no consistent method to combine performance data from all resources to measure the overall performance of a contractor.

- Difficulty to summarize performance data from all internal projects that form one industrial project to measure the overall performance of that industrial project.
- Difficulty to combine multiple-project reports to satisfy the needs of portfolio and resource managers from both contractors and owners.
- Collected actual performance data is not used for improving resource estimating practices in new projects and forecasting capabilities for in-progress projects.
- EVM data is not always collected by phase, resource and work package making it very difficult to summarize the data according to the required reporting level.
- Most currently used prediction techniques assume that current performance would remain the same until project completion, which is an unrealistic assumption (Nassar, 2004). There is a need for better forecasting tools that consider performance fluctuation as projects progress based on historical records.

4.2.5 Responsibility Management Practices

When an industrial construction project is initiated, the owner(s) assign a Project Management Team (PMT) from its/their own organization to oversee the project progress. This PMT breaks down the initiated project into a set of internal projects that are then handed over to a group of contractors. After receiving the work, each assigns an internal PMT from their own organization to handle the internal project. Since contractors are mostly matrix organization, functional managers are

responsible for assigning resources to internal projects. Each of these streams of managers has a reporting hierarchy that varies according to the size of the performed project and the organization. In some instances, projects are grouped in programs or even portfolios, where management teams are also assigned. The level of technical and managerial experience of each of the members of management teams has an impact on projects' resource utilization and performance. Projects are seldom managed by the same team without any changes from beginning to end. Hence, it is not only important to track who is managing projects and resources at the beginning of a project, but it is also important to track changes to the management teams during the project progress.

Most companies involved in industrial construction store data about their staff using different Human Resource Management Systems (HRMS) such as SAP or Oracle. These systems are not designed primarily for project management, and hence the stored data is only helpful for employee payments and benefit plans. These systems do not provide a tool to track who was responsible for managing projects and resources on a timely basis.

Given the complexity of the managerial teams and the large number of individual who impact management of project resources, several issues arise from the analysis of the current responsibility management practices in industrial construction projects:

- The existing HRMS systems do not store the vital data regarding who is responsible for managing projects and resources in a timely basis.
- Project responsibility hierarchical structure is not stored in one central location for easy access to this information.
- The lack of the historical data makes it almost impossible to analyze the impacts of changes in management teams on projects and resources' performance.
- Very difficult to find who was responsible for which task on a project making it a difficult task, sometimes unfair, to assign the right person to perform the right task in new projects. It is also very hard for knowledge seekers to find out who have the right knowledge they need.
- A lack for an objective performance evaluation tool at the individual level making it a subjective task that relies on perceptions not facts.

4.2.6 Summary analysis

After analyzing the processes in the industrial construction projects domain and the current practices of managing labour resource data, several issues are noted. Every contractor collects resource management data in different formats using a different suite of tools and software applications. That means resource management data is scattered between different applications in both electronic and hard-copy formats. Since each industrial construction project is planned and executed as a set of internal projects between multiple contractors, this makes it very difficult for industrial owners to obtain a complete set of resource management data on any of

their projects. In the current practices, it is also very difficult for a single contractor to combine, analyze and discover knowledge from the collected projects' data.

There are major problems with the current management practices on all five dimensions of resources data. For example, there is no consistent methodology to manage the scope of both internal and industrial projects. Project schedules lack the definition of standard activity types and activity attributes making it very difficult to analyze schedules of multiple projects simultaneously. Detailed cost estimates require a difficult and inaccurate recasting procedure before it can be transferred to cost baselines. Progress measurement and performance evaluation practices are not consistent in all internal projects making it very difficult to find the overall performance of a contracting company or an industrial project. The history of who was responsible for managing which task in a project is not recorded or stored in one central location. And the most difficult problem is the lack of integration between the five dimensions of managing project resources.

All these problems make it nearly impossible to transfer the collected data to useful knowledge. This analysis is confirmed by the findings of Chau (2002) who stated that data from completed projects cannot be used by new projects because collection of project's data relies mostly on temporary and specific activities to obtain project schedules and cost estimates. This practice makes it very difficult to compile and analyze data from completed project in a systematic way. There is a need for an integrated data generation and collection approach that is capable of

solving the previously mentioned problems. This integrated approach is explained in the next section of this chapter.

4.3 PROPOSED INTEGRATED DATA MANAGEMENT APPROACH

This research introduces a new integrated and consistent approach to define project elements, collect resources management data and store it in a structured format in a data warehouse ready for data mining and knowledge discovery. Collecting data using current practices means users have to spend a lot of time and efforts to obtain missing data, cleanup, preprocess and validate existing data prior to analyzing it. Useful knowledge is lost due to omitting the collection of very important data. To overcome most of the existing issues with current resource management practices, this integrated data acquisition approach could be easily implemented in industrial construction projects and contracting companies. This approach would make the stored data in the data warehouse ready for data mining and knowledge discovery saving all the time, costs and efforts spent on preparing the data.

By providing more data and prohibiting data loss, the approach also increases the accuracy of the knowledge discovery procedure and the value of the discovered knowledge. This discovered knowledge is used to improve the estimates of new projects, increase productivity and improve efficiency leading to higher profits, customer satisfaction and ability to compete for new project. Most importantly, the approach generates a continuous cycle of proper data generation, collection and storage, knowledge discovery and knowledge utilization.

As there is continuous supply of projects, there will be a continuous supply of new resources management data from completed projects. Thus, the cyclic approach is

self-learning, and the more data collected in the warehouse, the more analysis can be applied and the more useful knowledge is discovered. This continuous-cycle self-learning approach saves the time, cost, and efforts and transfers worthless data to valuable wealth of knowledge.

The proposed approach is based on improving the concept of Work Packages (WP) to act as building blocks and knowledge carrier while planning, executing and controlling of both internal and industrial projects. In this model, the work package acts as the core element for collecting resource management data. Work packages can also act as common denominators to collect non labour resources data such as risks, safety and quality issues between all projects.

Two main problems exist within the current practices of work packaging in industrial construction. First, these work packages are linked only to engineering or construction rather than being assigned to a specific production type such as foundations or structural steel. Second, the use of these packages start late in projects and there is no clear link to monitor the development of a work package and all its elements through the lifecycle of any project. To overcome these problems, the research approach introduces the concept of Predefined Progressable Work Packages (PPWP). This concept is clearly illustrated with a practical example in the next section of this chapter.

4.3.1 The Concept of Predefined Progressable Work Packages

A work package represents a manageable component of a project and the concept of work packaging was developed by the National Aeronautics and Space Administration (NASA) and the US Department of Defense (DoD) in the late sixties (Chehayeb, 1996). The Construction Industry Institute (CII) emphasizes that in order to manage a complex operation; this operation has to be broken down into well-defined components in hierarchical levels of detail where responsibility is clearly assigned to each level of this hierarchy (CII, 1988).

The work packages represent the lowest level of any project WBS (PMI, 2008) and WBS elements are seen as aggregating levels of the work packages. Each work package is composed of a set of deliverables that can be budgeted, scheduled and progressed as one package. In this research, the concept of work packaging is enhanced to be predefined progressable work packages. This enhanced concept is used for managing the five dimensions of labour resource data (scope, responsibility, schedule, cost and performance) at the work package level as shown in Figure 4.1.

Instead of starting planning each project from scratch, predefined work packages collected from previous projects are adapted to fit for new projects. Combining these customized work packages together, similar to building blocks, formulates the complete scope, schedule and budget for new projects.

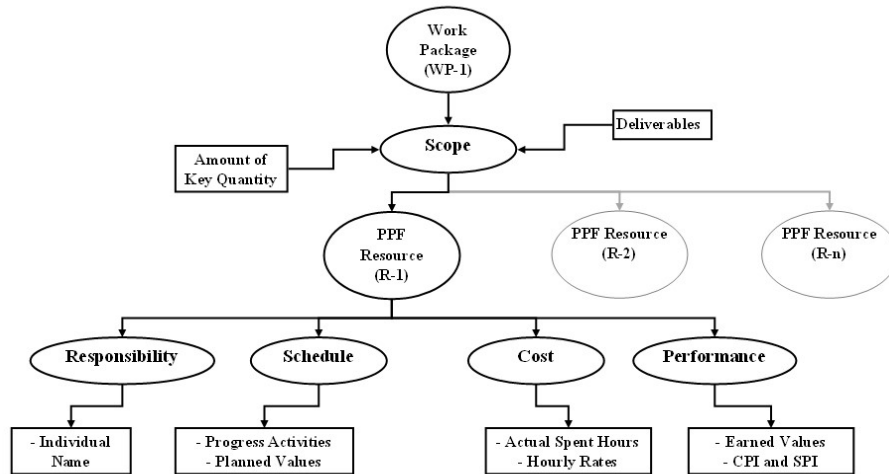


Figure 4.1: The Five Dimensions of Labour Resources Data

According to CII (1988), the concept of work packaging can be implemented to engineering, procurement and construction. The development of these progressable work packages starts during the FEL planning phases as Engineering Work Packages (EWP). EWP is a commonly known terminology in industrial construction. The fact that EWP can be issued to procurement, fabrication, assembly or site installation might cause confusion and difficulty. To avoid this confusion, this research suggests to use the term (EWP-P) for packages issued for purchasing, (EWP-C) for packages issued for contracting, (EWP-F) for packages issued for fabrication, (EWP-A) for packages issued for assembly and (EWP-S) for packages issued for site installation.

Work packages can be either planning packages or EPC packages. Planning packages are completed during the planning stage and do not require execution. Project charters, Project Execution Plans (PEP), TIC estimates and project

schedules are examples of planning packages. EPC progressable work packages are prepared and progressed from one pre-defined phase to the subsequent phase through the execution stage of any internal project.

There are different types of progressable work packages as shown in Figure 4.2. Type I represents packages that require engineering only such as Process Flow Diagrams (PFD's), Heat and Material Balances (HMB's) and Process & Instrumentation Diagrams (P&ID's). Type II represents packages that require engineering and procurement such supply-only Purchase Orders (PO's). Type III represents packages that require engineering, procurement and site installation such as foundations. Type IV represents packages that require engineering, procurement, fabrication, and site installation such as pump skids. Type V represents packages that require engineering, procurement, fabrication, assembly, and site installation such as pipe rack modules, and vessels.

The research approach is based on collecting the resource management data from all internal projects consistently at the work package level, which represents the optimum level of detail. By doing so, it allows straightforward summarizing and analyzing of resource data in both vertical and horizontal directions. On the one hand, vertical summarizations mean adding resource data from all internal projects that are performed by different contractors but form one industrial construction project.

Once the data for each industrial project is collected in the same format, it becomes easy for an industrial owner to summarize and compare the resource data of all their industrial projects. On the other hand horizontal summarization means summarizing and comparing all resource data from all internal projects that are performed by a single contractor. These internal projects belong to multiple industrial projects which are owned by different industrial owners.

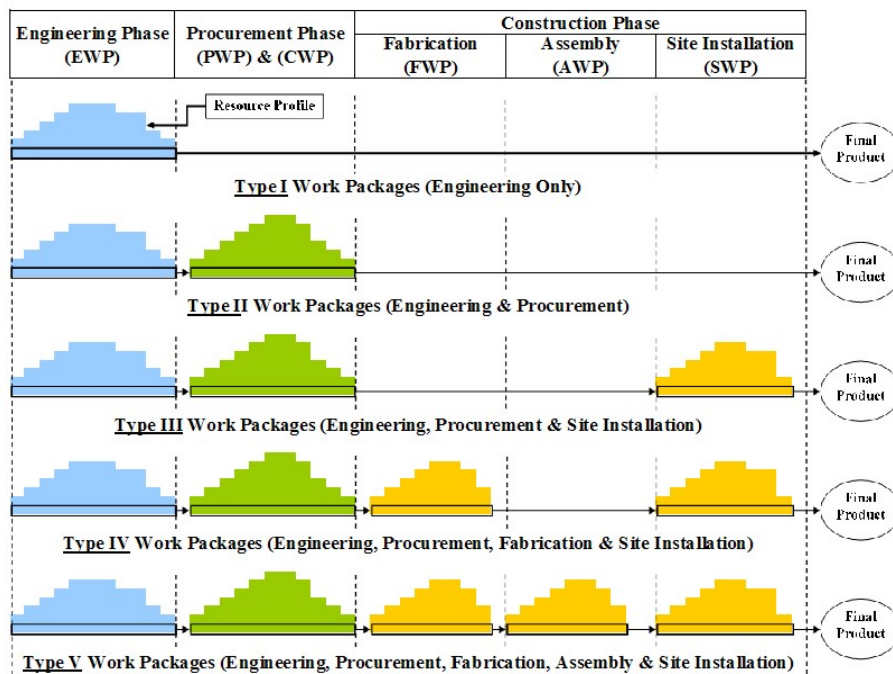


Figure 4.2: Types of Progressable Work Packages

Each work package is assigned to a production package that is predefined by the contractor for all internal projects. Process Flow Diagrams (PFD's), Heat and Material Balances (HMB's), Process and Instrumentation Diagrams (P&ID's), Line Designation Tables (LDT's) piling, foundations, structural steel, Single Line

Diagrams (SLD's) and Instrumentation Indices are examples of engineering production packages. Pipe-racks, plate work and handrails are examples of fabrication production packages.

Each contractor maintains a library of standard work packages that is easily used to define the scope of new projects. Some of these standard work packages can even be grouped together to form standard project templates.

4.3.2 Benefits of using Predefined Progressable Work Packages

Using the concept of predefined progressable work packages accompanied with predefined objects and attributes, provides integrity and consistency to the resource management practices as follows:

- The industry is familiar with the concept of work packages since the late 1960s. The proposed improvements to the concept are easily applied with minimum initial costs and using the existing tools and systems.
- Once implemented, the improved concept is expected to significantly increase consistency, efficiency and productivity in managing industrial construction projects. It is also expected to minimize the probability of schedule delays and budget overruns.
- The expected ROI from the increased productivity is really significant due to the difference between the low initial investment and the high returns.
- This concept presents an optimum level of detail for managing the five dimensions of labour resources. Managing resources at the activity level is

proven to be impractical and inapplicable, meanwhile managing at the project level doesn't provide sufficient detail for future use of the collected data.

- This concept introduces a consistent methodology to manage internal projects' scope as a set of work packages and industrial projects scope as a set of internal projects.
- This concept facilitates the procedure of projects change management and the distribution of the impact of these changes on the affected work packages.
- Work packages can be tracked from their definition at the FEL planning phases all the way to site installation.
- The concept facilitates tracking down individuals who worked on each package type. This is beneficial for finding team members who have experience on certain types of work packages and for transferring tacit knowledge between these individuals.
- Using predefined work packages saves time, costs and efforts spent to develop project baselines from scratch.
- The concept introduces consistency to multiple projects scheduling regardless of their type, duration or complexity.
- The concept allows focusing on critical packages not critical activities.
- The concept enables the seamless generation of multiple-project reports and graphs at any required level of detail.

The concept also:

- Provides the ability to use crosstab and pivot tables to present project schedules in a user-friendly format that maximizes the utilization of these schedules in managing industrial projects.
- Minimizes the need for the costly and inaccurate cost recasting procedure by estimating projects using the work packages that would be used for project execution.
- Initiates a reliable and consistent methodology to measure progress and evaluate performance in industrial construction projects. Results from this approach can be detailed or summarized to meet the necessities of various users at different management levels.
- Provides risk management facilitators with a consistent approach to build risk models based on predefined work packages for industrial projects.
- Supplies team members during quantitative risk assessment workshops with reliable historical data that really reflect uncertainties around projects' schedules, costs and resource requirements.
- Offers contractors a consistent approach for performing different scenarios to forecast their workload and use these scenarios to determine the optimum staffing capacity of any contractor.
- Supports data mining and knowledge extraction practices in order to transfer knowledge gained and lessons learned from completed projects to future and in-progress projects.

4.3.3 The Proposed Data Management Flow Chart

The procedure starts with a contractor identifying an opportunity to pursue part of the scope of an industrial project as shown in Figure 4.3. Opportunity is presented through direct request, bidding procedure or long term alliance. After identifying an opportunity, the contractor initiates an internal project in the bid/proposal stage in order to start collecting charges for preparing a bid or a proposal to the industrial owner. A Project Manager (PM) is assigned to manage this newly initiated internal project. Prior to performing any detailed planning of the project, the project manager identifies the required resources and develops an initial schedule, budget and resource histograms using the available information. Initial Planned Values (IPV) is obtained at the project level based on adjusted historical data from previously completed projects.

These IPV are added to the overall resource profile to verify the availability of adequate resources for the new project. It is of essential importance to ensure that the contractor has the required resources available when needed or at least a feasible plan to obtain these resources. If the project fits well within the overall contracting company workload, the company would go ahead trying to secure the new project; otherwise, the project will be terminated and the costs of this pre-planning stage are charged to overheads. If the project is secured, it is progressed to the planning stage as shown in Figure 4.4.

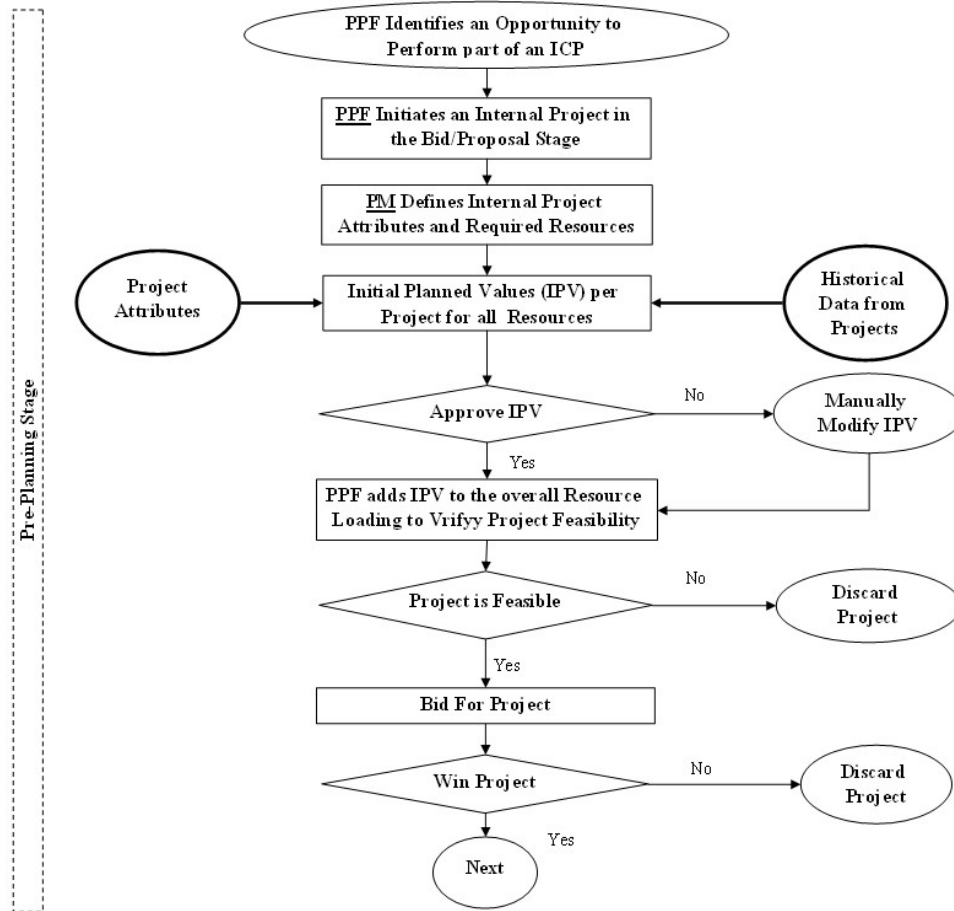


Figure 4.3: Flowchart of the Pre-planning Stage

The main purpose of the planning stage is to obtain frozen scope, schedule, cost and resources baselines to be used during the execution stage of the project. Freezing doesn't mean the baselines are not going to change, it means the change has to take place through the formal change management procedures to avoid problems in execution. According to the research approach, the planning stage starts with developing the detailed baseline scope using the concept of progressable work packages as explained later in this chapter. Based on the defined baseline scope, both the schedule and cost baselines are obtained utilizing the discovered

knowledge from previous projects. Both the cost and schedule baselines have to be combined together to form the resources baselines, which is called Original Planned Values (OPV). Obtaining the OPV marks the completion of the planning stage and the beginning of the execution stage of any project.

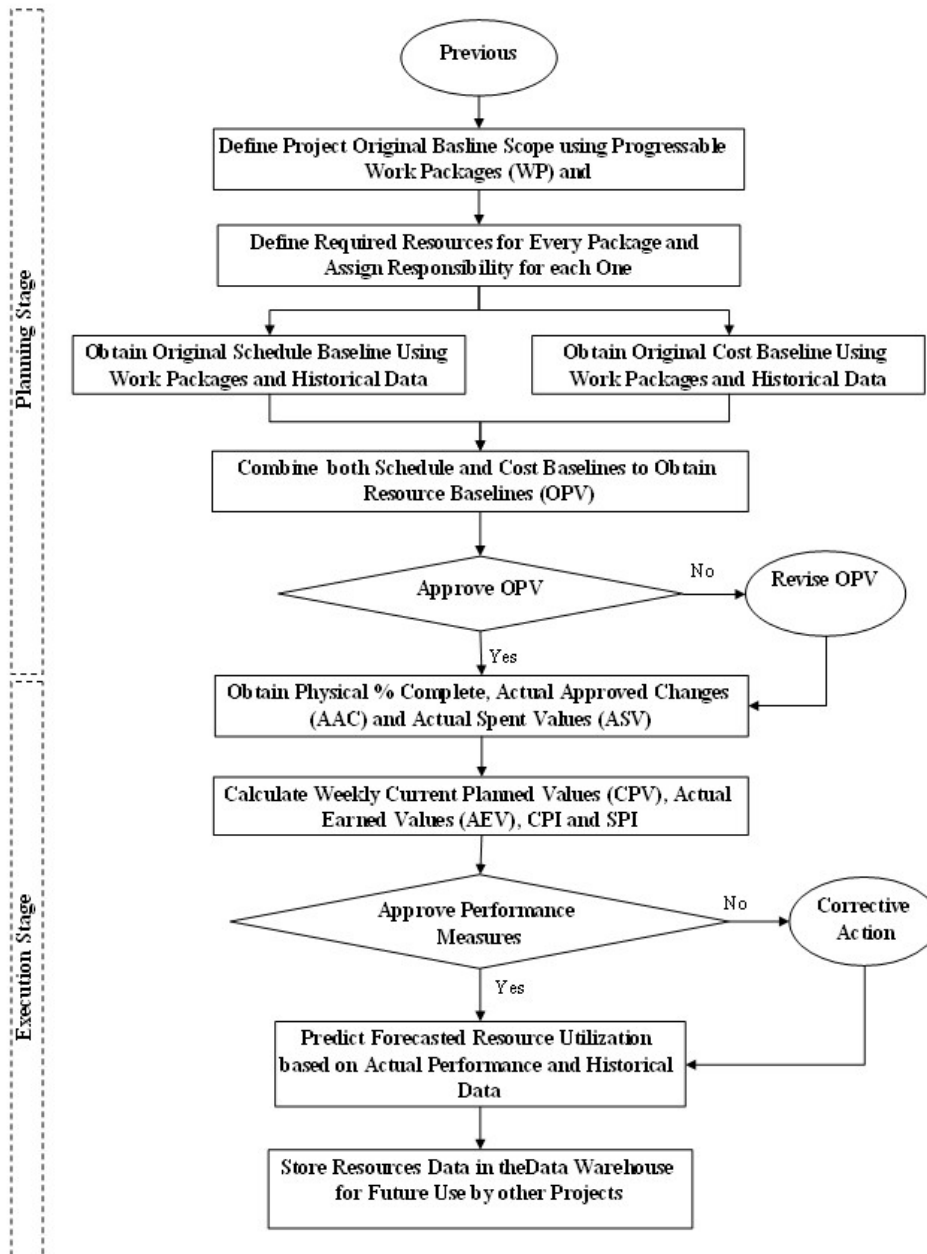


Figure 4.4: Flowchart of the Planning and Execution Stages

During the execution stage, the actual physical percent complete, Approved Project Changes (AAC) and Actual Spent Values (ASV) are obtained periodically. All these values have to be distributed on a weekly basis regardless of the length of the reporting period. These values are used to calculate the weekly Actual Earned Values (AEV), Cost Performance Index (CPI) and Schedule Performance Index (SPI). If these performance measures are not acceptable, timely corrective actions have to take place to bring the project back on the right track. These performance measures are also used to predict the forecast completion date, cost at completion and resource requirements.

4.4 IMPLEMENTATION OF THE PROPOSED FRAMEWORK

This chapter started with an analysis of the current practices for labour resources' data generation, collection and storing in industrial construction projects. This analysis revealed that labour resources data has several dimensions. These dimensions are scope, schedule, cost, responsibility and performance. The analysis also showed that there are major problems with the current practices in each of the five dimensions. The existing problems hinder the proper application of data mining and knowledge discovery techniques in this domain.

In order to overcome the existing problems, the concept of predefined progressable work packages is used. This concept is an enhancement to the existing practices of work packaging that addresses most of the problems with current practices. With

the proposed framework, the five dimensions of labour resources data are managed in an integrated manner.

First, the scope of each industrial construction project is completely defined as the scope of a set of internal projects. The scope of each of these internal projects is completely defined as a set of predefined progressable work packages. The scope of each of these work packages is fully defined as determinate amount of a specified key quantity. Examples of key quantities are: number of drawings for engineering work packages, tonnes of steel for structural steel fabrication work packages, and cubic meters of concrete for foundations site installation work packages. Five data elements are collected on these quantities during the planning stage and another three data elements are collected during the execution stage of any project. The planning stage data elements are: FEL I estimated quantity, FEL II estimated quantity, FEL III estimated quantity, Bid estimated quantity and original baseline quantity. The execution stage data elements are: approved quantity changes, current baseline quantity and actual quantity. Current baseline quantity is calculated by adding the original baseline to the approved quantity changes. Not every package will have values for all these data elements; however original baseline, current baseline and actual quantities are mandatory fields.

Second, the history of individuals who are responsible for managing all the required resources to perform the scope of each work package is stored in the data warehouse. The data elements are: the individual name, resource name and the start

and finish dates for assuming the responsibility. These data elements represent the responsibility dimension of the labour resources data. Storing these data elements in the data warehouse enables analyzing the impact of changing resource managers on projects.

Third, the schedule of each work package is represented as a group of interlinked pre-defined progress activities. These groups can be stored in the time management software as templates for different types of work packages, internal projects or even industrial projects. By using these templates as a start point to develop schedules, the level of consistency between projects schedules increases significantly. By using pre-defined progress activities, data mining techniques are easily implemented to extract knowledge from multiple project schedules. Each progress activity has a predefined weight to be used for progress measurement and performance evaluation. The total weight of all progress activities has to equal one. Detailed analysis studies can be applied to a sample for each production package to determine these progress weights.

Furthermore, using pre-defined progress activities enables schedulers to present lengthy schedules as crosstab reports as shown in Figure 4.5. This presentation is truly user friendly and summarizes lengthy schedules in an understandable format. Programming is used to highlight activities that have to be completed in two weeks, delinquent and completed activities.

Deliverable	Issue For External Review			Return from Review			Issue For Tender (IFT)					
	Start	Wt	%	Start	Wt	%	Start	Wt	%			
Utilities & Offsites (U)	04/18/03	✓	0.75	100	04/25/03	✓	0	100	05/02/03	✓	0.25	100
Battery Limits Table	04/21/03	✓	0.75	100	04/28/03	✓	0	100	05/05/03	✓	0.25	100
Utilities & Offsites Int	04/21/03	✓	0.75	100	04/28/03	✓	0	100	05/05/03	✓	0.25	100
	04/21/03	✓	0.75	100	04/28/03	✓	0	100	05/05/03	✓	0.25	100
Line Resizing	04/21/03	✓	0.75	100	04/28/03	✓	0	100	05/05/03	✓	0.25	100
Line List	04/21/03	✓	0.75	100	04/28/03	✓	0	100	05/05/03	✓	0.25	100
	04/28/03	✓	0.75	100	05/05/03	✓	0	100	05/12/03	✓	0.25	100
3D Modeling	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	✓	0.25	100
Electrical Layouts	04/28/03	✓	0.75	100	05/05/03	✓	0	100	05/12/03	✓	0.25	100
General Arrangemen	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	✓	0.25	100
Plot Plans	04/11/03	✓	0.75	100	04/14/03	✓	0	100	04/21/03	✓	0.25	100
Key Plan	04/11/03	✓	0.75	100	04/18/03	✓	0	100	04/25/03	✓	0.25	100
Process Input for Plo	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	✓	0.25	100
	04/10/03	✓	0.65	100	04/17/03	✓	0	100	04/24/03	✓	0.25	100
Electrical Load List	04/10/03	✓	0.65	100	04/17/03	✓	0	100	04/24/03	✓	0.25	100
	04/14/03	✓	0.75	100	04/21/03	☐	0	0	04/28/03	☐	0.25	50
Electrical Single Line	04/14/03	✓	0.75	100	04/21/03	☐	0	0	04/28/03	☐	0.25	50
	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	☐	0.25	0
Control Synopsis	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	☐	0.25	0
	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	☐	0.25	0
Grading Plans	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	☐	0.25	0
Firewater Plans	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	☐	0.25	0
Roadway/Surfacing	04/14/03	✓	0.75	100	04/21/03	✓	0	100	04/28/03	☐	0.25	0

Figure 4.5: Transferring the Schedule into Crosstab Report

The data elements are start and finish dates for original baseline, current baseline and actual execution for each work package. This data can be automatically obtained from the proposed timekeeping system that is also used to manage the cost dimension of the resources data.

Fourth, the cost data is estimated and controlled at the work package level as the cost of all the required resources to complete the package scope. Using the predefined set of labour resources, hourly budget for every required resource estimated and the average rates are used to transfer these hours to control budget. Cost data elements include the minimum, maximum and most-likely original baseline, current baseline and actual labour hours and costs of every resource in all work packages. Pre-planning data such FEL I, FEL II, FEL III and Bid estimated hours and costs can be stored as well if needed.

The proposed time collection system is shown in Figure 4.6. This timekeeping system collects actual hours by work package, phase, stage, resource, individual, type and location where the work is performed. Any timekeeping system that is currently used can be easily modified to collect actual data in the integrated structured format. The type field is used to distinguish between productive vs. non-productive time to maximize the value of the collected data.

Project / Work Package	Phase	Stage	Resource	Type	Location	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Total
						0	0	5	7	10	8	10	40
R(1), Foundation for Piprack # 1	Detailed Engineering & Design	Planning	Structural Engineering	Productive	Edmonton Office			5	7	10	8	5	10
R(1), Structural Steel for Piprack # 1	Detailed Engineering & Design	Execution	Structural Engineering	Waiting	Edmonton Office				7	10	8	5	30
													0
													0
													0

Project / Work Package	Phase	Stage	Resource	Type	Location	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Total
						0	0	8	8	10	8	8	42
R(1), Structural Steel for Piprack # 1	Fabrication	Execution	Cutting	Productive	Edmonton Shop			8	8	4			20
R(1), Structural Steel for Piprack # 2	Fabrication	Execution	Cutting	Training	Edmonton Shop					6	8	8	22
													0
													0
													0

Project / Work Package	Phase	Stage	Resource	Type	Location	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Total
						0	0	10	8	8	10	8	44
R(1), Structural Steel for Piprack # 1	Assembly	Execution	Welding	Rework	Edmonton Yard			10	8	8	10	8	44
													0
													0
													0

Project / Work Package	Phase	Stage	Resource	Type	Location	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Total
						0	0	10	8	12	10	6	46
R(1), Foundation for Piprack # 1	Site Installation	Execution	Welding	Productive	Fort McMurray Site			5					5
R(1), Structural Steel for Piprack # 1	Site Installation	Execution	Welding	Preparing	Fort McMurray Site			5	8	12	10	6	41
													0
													0
													0

Figure 4.6: The Actual Cost Collection System

Fifth, the proposed concept provides a consistent methodology to measure progress and evaluate performance at the work package level. By the end of the planning stage of any project, weekly Original Planned Values (OPV) is obtained for every resource. During the execution stage, approved changes are added to the Original Planned Values (OPV) to obtain Current Planned values (CPV). Moreover, progress is measured using percent complete of progress activities. This progress is represented as weekly Actual Earned values (AEV).

Actual spent hours and costs are obtained on a weekly basis and are used to calculate the performance measures; CPI and SPI. Seven data elements are collected and stored in the data warehouse. The first three data elements are the minimum, maximum, and most likely Original Planned Values ($OPV_{(1:z(ICP))}$), where $z_{(ICP)}$ equals to the duration of the work package. The fourth data element is the Actual Approved Changes $AAC_{(1:z)}$. The fifth data element is the Current Planned Values ($CPV_{(1:z)}$). CPV represents the modified baseline resource-profiles after considering the impacts of all approved project changes. The sixth element is the Actual Earned Values ($AEV_{(1:z)}$). $AEV_{(1:z)}$ is calculated by multiplying $CPV_{(1:z)}$ by Physical % complete $_{(1:z)}$. Physical % complete is obtained from the resource manager of each resource at the work package level. The seventh element is the Actual Spent Values ($ASV_{(1:z)}$) obtained from the timekeeping system.

These data elements are presented graphically as shown in Figures 4.7, 8 and 9 to form the resource baselines. Figure 4.7 shows an example of a resource baseline histogram that shows the Min, ML and Max estimated resource hours plotted over the Min, ML and Max estimated durations. These histograms are obtained after completing the planning stage and are frozen prior to execution. Figure 4.8 shows an example of plotting the current planned, actual earned and actual spent hours over the actual duration. These graphs are obtained after the completion of the execution stage. Figure 4.9 shows an example of plotting the cumulative values per week.

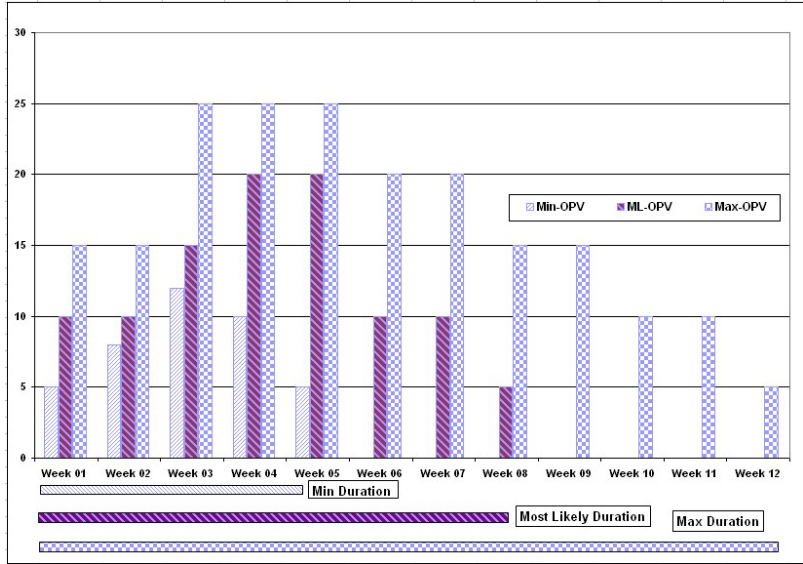


Figure 4.7: Resource Baseline Histograms

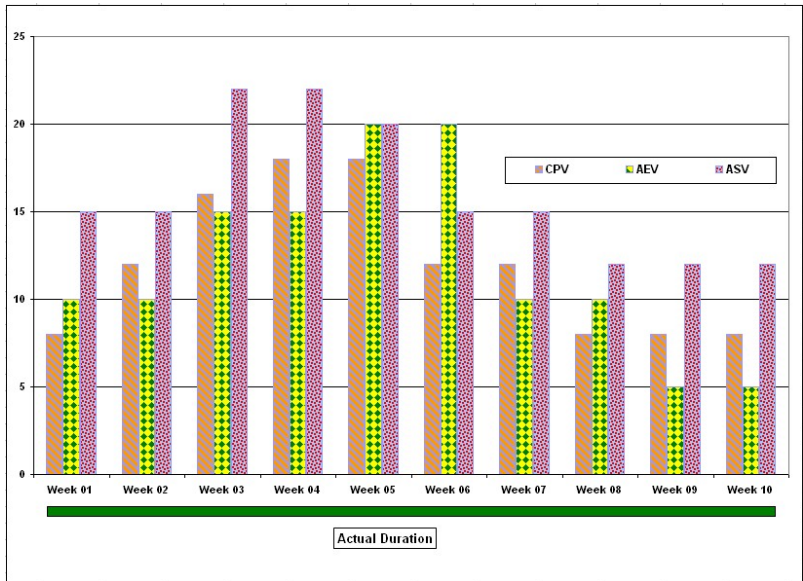


Figure 4.8: Plotting Resource Values

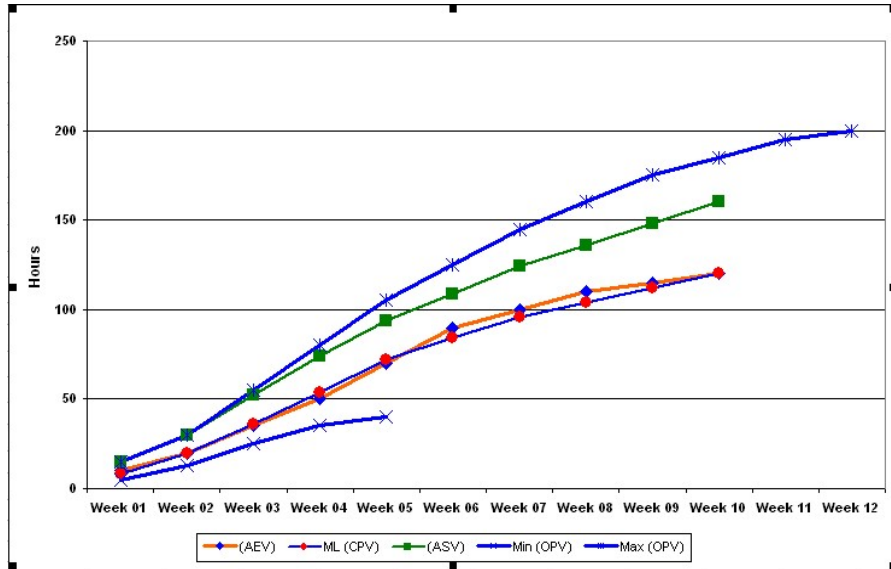


Figure 4.9: Plotting Cumulative Values per Weeks

With the introduction of the proposed concept, contractors can obtain sophisticated forecast of their workload based on aggregating consistent data from all work packages. This data aggregation can be graphically illustrated for each of in-progress, planned, awarded and proposed projects.

Figure 4.10 shows an example of the proposed graphs for in-progress projects. The graph shows actual hourly workload from the completed portion of in-progress projects and the Min, ML and Max forecast values distributed over the Min, ML and Max durations for these projects. These graphs can be prepared using the mathematical summation of values from all work packages or using probabilistic resource management software such as PertMaster.

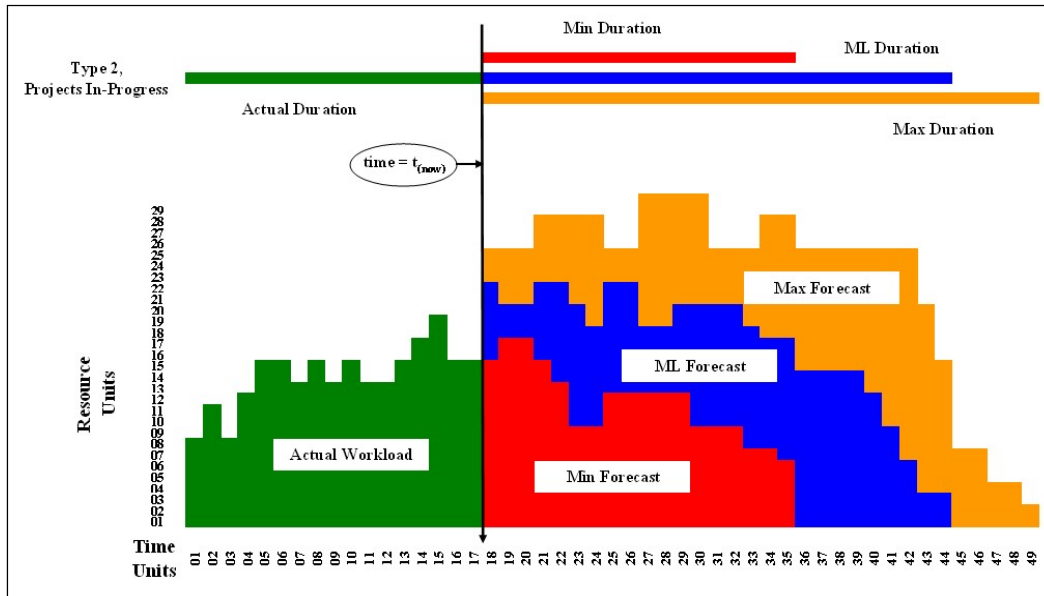


Figure 4.10: The Enhanced Workload vs. Capacity Graph

Since all graphs use hours as measuring units, graphs from all resource and all types of projects are combined to obtain the overall company workload. These graphs are based on facts and historical data and not on best guesses. The company management can apply different scenarios based on their knowledge and experience and utilizing the obtained graphs to determine the optimum capacity. The objective of determining the optimum capacity is to minimize the resources idle time and overutilization in order to maximize the profit and efficiency.

The same approach can be used to forecast the resource requirements for an industrial project by combining data from all its internal projects. Since project phases and resources are predefined, the summation and summarization process is automatic, straight forward and doesn't require user interference.

Other resources data such as materials and equipment need to be added to the proposed approach to maximize its value. Collecting this data using the concept of predefined progressable work packages is expected to be straight forward exercise. It requires completing the materials branch of the predefined RBS in order to introduce consistency to the data collection process. Adding this other resources data to the data warehouse, would make it complete with actual TIC costs of any work package. Aggregating these costs up to the internal project and industrial project level provides really useful information that can be used for data mining.

At the beginning of any project, owner's executives are very interested in obtaining a rough estimate of project costs to decide if they should spend money on the FEL planning of the project. Utilizing the stored data in the data warehouse accompanied with data mining techniques, cost per square meter for any specified type of projects. These costs can be easily escalated to provide the executives with these rough estimates for any of their proposed projects. Hence, this makes the proposed system more appealing for executives to implement in their companies.

Executives rely on project managers to look at the details of the resource planning and they always look for information at a very high level. Adding other resources data to the proposed approach add executives to the range of end-users of the similar to functional and project managers.

CHAPTER 5: THE LABOUR RESOURCES DATA WAREHOUSE

5.1 BENEFITS OF WAREHOUSEING RESOURCES DATA

With the introduction of computers, construction industry is inundated with data. Industrial construction projects are among the most sophisticated types of construction that presently faces many challenges. One major challenge is the proper handling of the data generated while managing labour resources. The objective of this research is to transfer this data to useful knowledge that would be used to improve and to increase efficiency and productivity of labour resource management practices.

In order to achieve this objective, Chapter 3 of this thesis provided an analysis of the industrial construction projects domain and all the labour resources data generation processes in that domain. Chapter 4 of this thesis analyzed the current practices of resources data management and the existing issues with these practices. Based on this analysis, an enhanced approach to introduce consistency in data generation and collection at all five dimensions was developed.

To close the data utilization cycle, there is a need for a central, powerful and structured location for storing the data in a format ready for data mining and knowledge extraction. A data warehouse provides an optimum solution to satisfy these needs.

Data warehouses are different from traditional databases. They are subject-oriented, contain only non-volatile, clean and validated historical data; and they are designed for decision-support not operation-support. A data warehouse saves the time, cost and efforts required to clean, preprocess and validate data. It prevents manual data entry typos and mistakes. Storing clean, preprocessed and validated data in a central location shortens the processing time required to respond to complicated queries. The initial time, costs and efforts that are necessary to setup a data warehouse are reasonable. To build a warehouse, there is a variety of commercially available software in the market.

Data warehouses are designed and built to handle multidimensional data. Therefore, by opting to use a data warehouse, contractors must adhere to a structured way of data collection where all dimensions of resources data are captured properly. Moreover, this would also encourage storing both planning and execution data in the same central location. Existing practices store them in different locations and format.

A data warehouse structures the stored data in one consistent format to support decision-making and dynamic interactive data viewing from different angles. By labeling each data point properly, decision makers can directly analyze the stored data instead of getting lost between hardcopy documents and different format electronic files. A data warehouse automatically exchange data between multiple types of electronic files and data mining applications.

5.2 BUILDING THE DATA WAREHOUSE

5.2.1 The Multidimensional Data Model

In order to design the data warehouse for resources data, all the data objects and dimensions of each object have to be defined first. According to the resource management data analysis that was illustrated in the previous chapter, seven objects have to be modeled in the data warehouse. These objects are Industrial Owners (IO), Industrial Construction Projects (ICP), Industrial Components (IC), Contractors (CON), Internal Projects (IP), Work Packages (WP) and Individuals.

The multidimensional data model utilizes the work package as the main object for data collection and storage. The six other objects can be seen as dimensions of the work package, but they were modeled as objects to increase the efficiency and generic applicability and integrity of the developed warehouse. Many dimensions are organized in a hierarchical, aggregated structure, parent-child relationships. Data entities at each level of the hierarchy can have different attributes. Examples of these hierarchies are:

- Continent, country, province/state, city
- Year, quarter, month, week
- Group of companies, company, business unit
- Portfolio, program, project
- Category of resources, group of resources, resource, individual
- Category of phases, group of phase, project phase

As shown in Figure 5.1, the work package is the centre of the multidimensional data model. Each work package is connected to time dimension (representing as calendar week), project progress over time (represented as reporting period number) project phase, required resources, production package and industrial component. The internal project is connected to an industrial project and a contractor. The industrial component is connected to an industrial project, which is linked to an industrial owner.

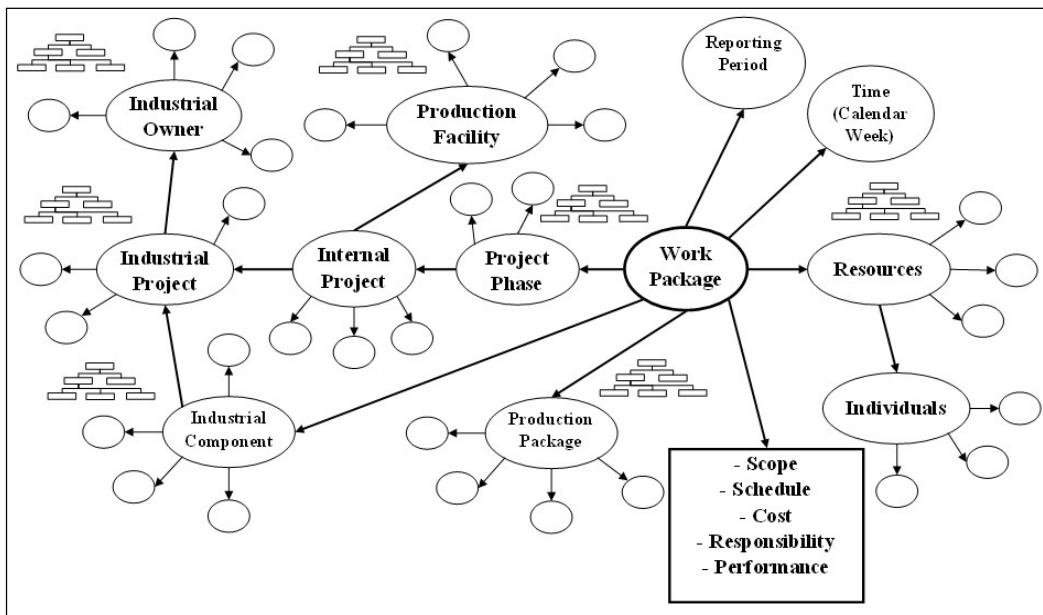


Figure 5.1: The Multidimensional Data Model

Dynamic reports, graphs and data mining techniques make use of any of the available attributes of all these objects to extract the knowledge from the data and present it to the user in a useful and user-friendly format. By utilizing this structure, all data points are now forming a well connected network that provides endless options for data analysis.

The work package also stores data elements the five dimensions of resources data. The data elements include data from the planning stage represented as minimum (Min), maximum (Max) and most likely (ML) original baselines. Storing three values for original baselines allows for probabilistic planning and uncertainty modeling.

Data elements from the execution stage include current baselines, earned and actual values and responsibility assignments. Storing the current baselines reflects the changeable nature of projects and represents the fact that very few projects are executed without any changes to the original plans. Storing both actual and earned values allows for using Earned Value Management (EVM) techniques to connect both costs and schedule performance measures. Storing responsibility assignment data provided a tool for exchanging tacit knowledge between individuals.

It is important to understand how the multidimensional data model is translated into the powerful snowflake schema while developing the labour resources data warehouse prototype. The prototype was developed in MS Access for testing and validating the research concepts. The prototype performed very well while being tested using a large dataset from a large number of real industrial projects. The dataset was provided from two partner companies and was modified for confidentiality purposes as would be explained later in this chapter.

5.3 THE SNOWFLAKE SCHEMA

Figure 5.2 shows the snowflake schema for the first two objects; the industrial owners and industrial construction projects. Two fact tables represent both objects and several dimension tables represent their attributes. The hierarchical dimension table for the location, industry and WBS are shown in the graph. The owners' staff information is also stored in a third fact table. Industrial construction projects are broken down to internal projects and industrial components as shown in the diagram.

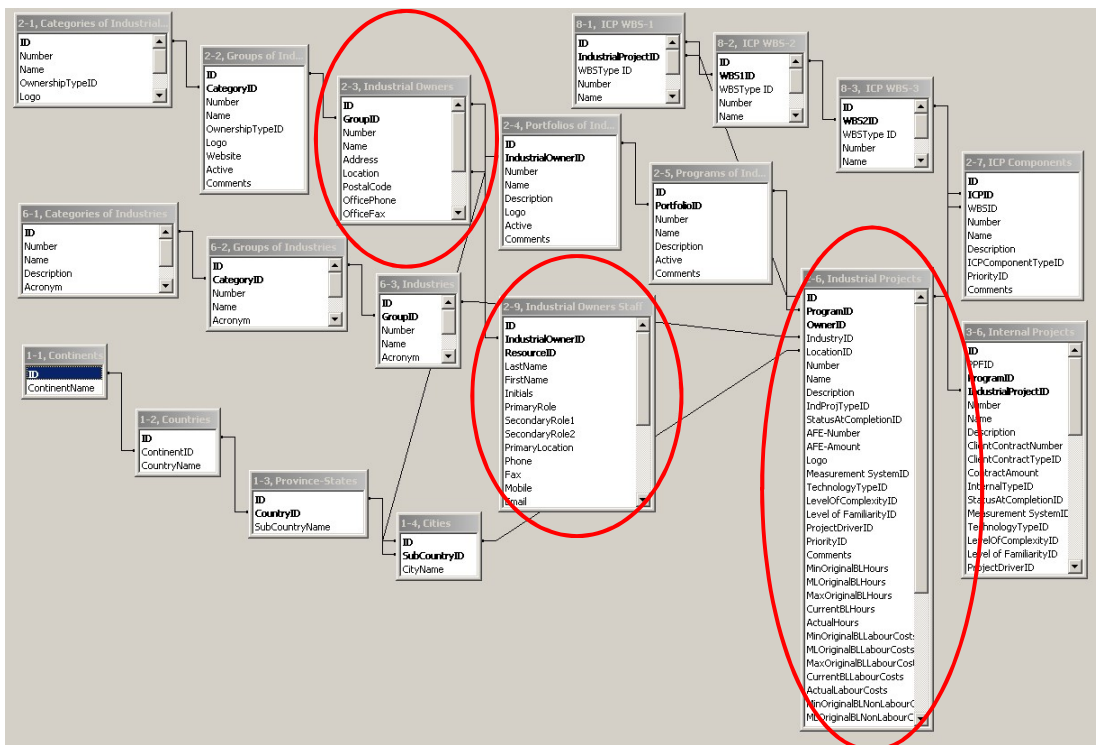


Figure 5.2-a: The Industrial Owners and Construction Projects Schema

Figure 5.2-b presents the snowflake diagram for another two objects; contractors and internal projects. The individual staff information is stored in a two fact tables, one for the industrial owners' staff and the other for contractors staff.

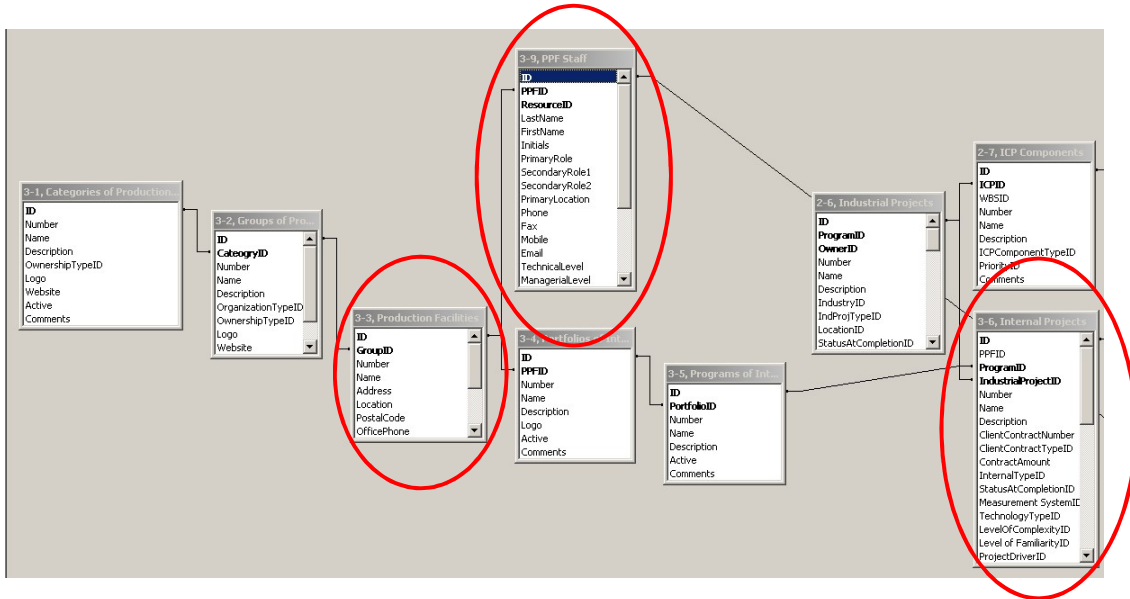


Figure 5.2-b: The Contractors and Internal Projects Schema

Figure 5.2-c presents the snowflake diagram for the work package object and its fact tables for storing resource management data. Every work package is linked to an industrial component and a production package. After that, each package is broken into a set of progress phases, each of which is linked to an internal project. Resource management data is collected at that level in four fact tables. Each fact table is linked to its dimensional tables as shown in the figures.

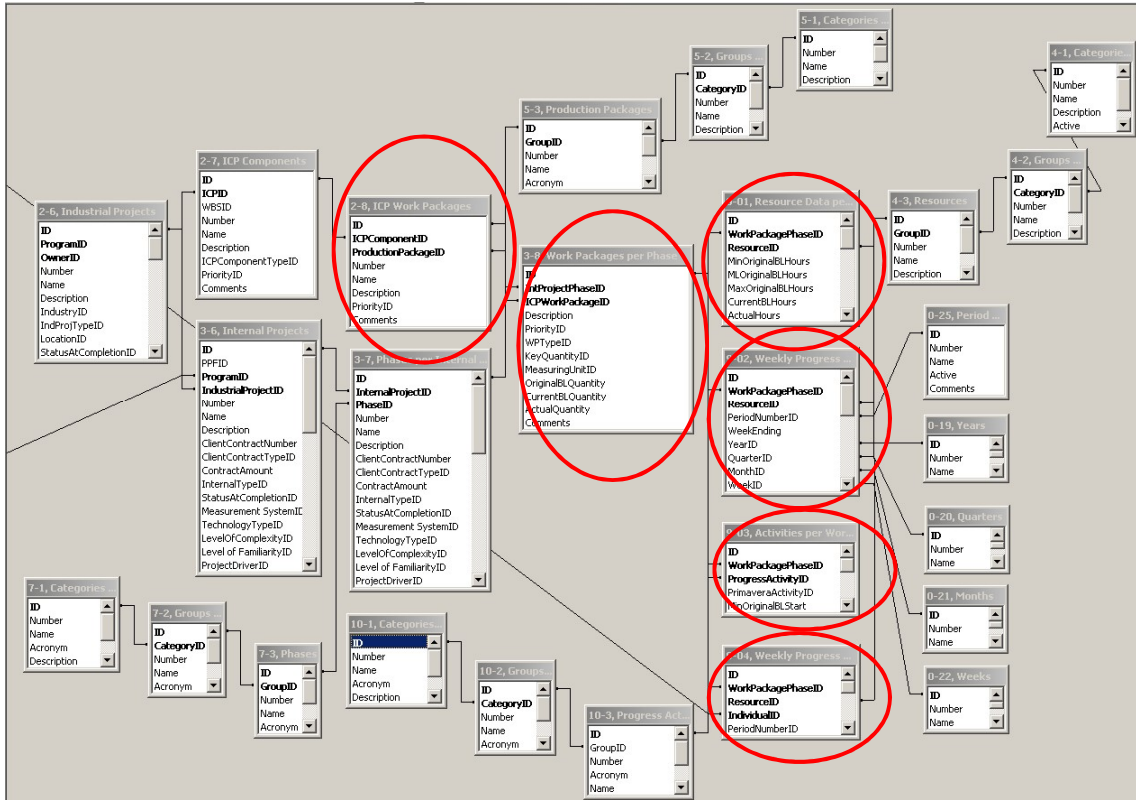


Figure 5.2-c: The Work Package Schema

5.4 THE DATA WAREHOUSE UNPUT (BACKEND)

The backend of any data warehouse represents the user interface for data entry. The prototype data warehouse is designed to allow manual data entry or automatic data transfer for other Online Transactional Processing (OLTP) systems. These systems include Primavera or MS Project for scheduling data, cost estimating timekeeping, and accounting systems for cost data and Human Resources Management Systems (HRMS) for individuals' data. The manual data entry required the design of a set user interfaces to enable the hierarchical data entry that fit for the snowflake schema of the data warehouse.

Each of the seven objects is modeled as a fact table with a set of control attributes, which will be used for dynamic reporting and data mining. The first of these seven objects is the Industrial Owners ($IO_{(1:x)}$), which is modeled as hierarchy of three levels with the highest level to be group of industrial owners and the lowest level to be Business Unit (BU) and individuals. This hierarchical modeling is very important to enable dynamic reporting using all levels of the hierarchy.

The definition screen for the first object is shown in Figure 5.3 and it reflects the 3-level hierarchical nature of industrial owners ($IO_{(1:x)}$) as defined in the data warehouse. The object is modeled as fact table with several dimension tables linked to it. Examples of these dimension tables are: office locations and ownership types. These dimension tables also represent the control attributes for dynamic reporting and data mining. More control attributes can be easily added to the data warehouse if needed.


Similar to industrial owners, contractors ($CON_{(1:z)}$) are also represented through a three-level hierarchical structure using the same control attributes as shown in Figure 5.4. One more control attribute is added to represent the type of contractor for filtering and grouping purposes. The values for this attribute include: engineering only, procurement only, construction only, Engineering, Procurement and Construction (EPC) and Engineering, Procurement and Construction Management (EPCM).

Industrial Owners

Group Name: Suncor Group of Companies Website: <http://www.suncor.com/start.aspx>

Group Number: Website Active:

Ownership Type: Public Company

Logo: 

Comments:

Company Name	Number	Ownership Type	Logo	Website	Act	Comments
- Suncor Energy Inc.		Public Company	Bitmap Image	http://www.suncor.com/start.aspx	<input checked="" type="checkbox"/>	
Business Unit Name	Address	Location	PostalCode	Phone	Fax	
▶ Head Office	112 - 4 Avenue S.W.	Calgary, Alberta, Canada	T2P 2V5	(403) 269-8100	(403) 269-6200	
Major Projects		Fort MacMurray, Alberta,				
Operations						
Maintenance						
*					<input checked="" type="checkbox"/>	

Record: 1 of 4

Record: 1 of 3


Figure 5.3: Industrial Owners (IO) Definition Screen

Project-based Production Facilities (PPF)

Group Name: WorleyParsons Ltd. Website: <http://www.worleyparsons.com/Pages/default.aspx>

Group Number: Website Active:

Ownership Type: Public Company

Logo: 

Comments:

Company Name	Number	Type	Ownership Type	Logo	Website	Ac	Comments
- Colt WorleyParsons		EPCM	Public Company	Bitmap Image	http://www.colteng.com/	<input checked="" type="checkbox"/>	
Business Unit Name	Business Unit Address	City	Postal Code	Phone	Fax	Act	
▶ Edmonton Office	120, 5008 - 86th Street	Edmonton, Alberta,	T6E 5S2	(780) 440-5300	(780) 440-5555	<input checked="" type="checkbox"/>	
Calgary Office						<input checked="" type="checkbox"/>	
Sarnia Office						<input checked="" type="checkbox"/>	
Toronto Office						<input checked="" type="checkbox"/>	
Corporate Office						<input checked="" type="checkbox"/>	
*						<input checked="" type="checkbox"/>	

Record: 1 of 5

Record: 1 of 3

Figure 5.4: Contractors (PPF) Definition Screen

The third object is the Industrial Construction Projects (ICP_(IO)). The definition screen for industrial construction projects is shown in Figure 5.5.

These projects are defined as three-level hierarchy; portfolio of programs, program of projects and industrial construction project. Some of these attributes are hierarchical such as locations and industrial sector. From that screen, the user can define all the internal projects for any industrial construction project.

Owner	Major Projects, Suncor Energy Inc., Suncor Group of Companies		
Program	Steepbank Projects, Debottlenecking Programs		
Industry	Ore Preparation Plants, Oil Sands, Hydrocarbons		
Construction Location	Fort MacMurray, Alberta, Canada, North America		
Project Number	82-003-011	AFE-Number	
Name	Dry Surge (Initiative # 01)		
Description			
Project Type	Greenfield	Status At Completion	Completed
Measurement System	Metric	Level of Familiarity	High
Technology Type	Proven	Project Driver	Schedule
Level of Complexity	Medium	Priority	Medium
Comments	All numbers are distorted for confidentially purposes		Logo

Add Internal Projects

Figure 5.5: Industrial Construction Project Definition Screen

The fourth object is the Internal Project ($IP_{(ICP,PPF)}$), which are linked to both industrial construction projects and contractors. The definition screen for internal projects is shown in Figure 5.6-a. From that screen, the project-phase definition screen can be accessed as shown in Figure 5.6-b.

Figure 5.6-a: The Internal Project Definition Screen

Figure 5.6-b: The Definition Screen for Internal Project Phases

The fifth object is the Work Package ($WP_{(IP,Phase)}$), which is used to collect all the resource management data. The data entry screen for the dimensions of resources data are shown in Figures 5.7, a, b, c, d and d. Figures 5.8, 5.9 and 5.10 show examples of the definition screens for some of the hierarchical dimensions.

Work Packages for a Phase of an Internal Project

Project - Phase: 04E2665-04, Dry Surge - Detailed Engineering & Design

ICP Work Package: 005, Foundations for Piperack # 001

Description:

Priority: High WP Type: EWP, Engineering Work Package

Key Quantity: Number of Drawings Unit of Measure: Each

Scope		
Original Baseline	Current Baseline	Actual
Quantity: 75.00	Quantity: 80.00	Quantity: 83.00

Comments:

Record: 1 of 176

Figure 5.7-a: Scope Definition Screen for Work Packages

Resources Data per Work Package

Work Package: 04E2665-04, Dry Surge - Detailed Engineering & Design Phase, 005, Foundations for Piperack # 001

Resource: Engineering Services, Civil - Structural - Engineering

	Original Baseline			Current Baseline	Actual
	Min	Max	ML		
Hours	50	80	65	72	113
Labour Costs	\$2,500	\$4,000	\$3,250	\$3,600	\$3,753
Non-Labour Costs					
Start Date			05-May-07		12-May-07
Finish Date			05-Jun-07		22-Jun-07

Record: 1 of 1

Figure 5.7-b: Resource Data Entry Screen per Work Package

Weekly Progress per Work Package

Work Package: 04E2665-04, Dry Surge Project - Detailed Engineering & Design Phase, 005, Foundations for Piperack # 001
 Resource: Engineering Services, Civil - Structural - Engineering
 Period Number: Period Number 02
 Week Ending: 15-May-07
 Year: 2007
 Quarter: Second Quarter
 Month: May
 Week: Week Number 22

	Original Baseline			Current Baseline
	Min	Max	ML	
Hours				150
Labour Costs				\$7,500.00

	Earned	Actual	CPI	SPI
	Hours	130.00	170	0.76
Labour Costs	\$5,670.00	\$8,052.00	0.70	0.76

Record: 1 of 1

Figure 5.7-c: Weekly Progress per Work Package

Progress Activities per Work Package

Work Package: 04E2665-04, Dry Surge Project - Detailed Engineering & Design Phase, 005, Foundations for Piperack # 001
 Progress Activity: Engineering, Front End Loading, Issue for Review
 Activity Number: 54266

	Original Baseline			Current Baseline	Actual
	Min	Max	ML		
Start Date			05-May-07	12-May-07	19-May-07
Finish Date			19-May-07	26-May-07	26-May-07

Record: 1 of 1

Figure 5.7-d: Progress Activities per Work Package

Weekly Progress per Individual

Work Package: 04E2665-04, Dry Surge Project - Detailed Engineering & Design Phase, 005, Foundations for Piperack # 001
 Resource: Project Services, Project Controls - Cost Control
 Individual: Project Services, Project Controls - Management, (Raj, Hans)
 Period Number: Period Number 02
 Week Ending: 22-May-07
 Year: 2007
 Quarter: Second Quarter
 Month: May
 Week: Week Number 22

	ML Original Baseline	Current Baseline	Actual
	Hours	10	10
Labour Costs	\$500.00	\$500.00	\$600.00

Record: 1 of 1

Figure 5.7-e: Weekly Hours per Individual



Figure 5.8: Definition Screen for Locations

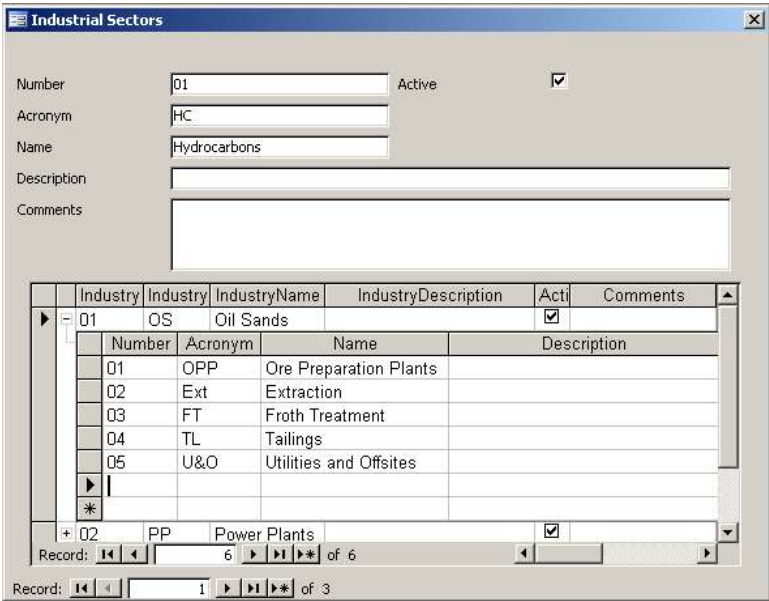


Figure 5.9: Definition Screen for Industries

Project Phases

Category Number: 01 Active

Category Acronym: ICP

Category Name: Industrial Construction Projects

Category Description:

Comments:

Phase	Phase	PhaseName	PhaseDescription	Act	Comments
02	E	Engineering		<input checked="" type="checkbox"/>	
01	SS	Front End Loading (FEL) I	Identify	<input checked="" type="checkbox"/>	Scoping Studies
02	DBM	Front End Loading (FEL) II	Evaluate	<input checked="" type="checkbox"/>	Design Basis Memorand
03	EDS	Front End Loading (FEL) III	Define	<input checked="" type="checkbox"/>	Engineering Design Spec
04	DED	Detailed Engineering & Design	Execute	<input checked="" type="checkbox"/>	
05	SD	Shop Drawings		<input checked="" type="checkbox"/>	
06	PS	Procurement Support		<input checked="" type="checkbox"/>	
07	CS	Construction Support		<input checked="" type="checkbox"/>	
08	AB	As-Building		<input checked="" type="checkbox"/>	

Record: 1 of 8

Record: 1 of 4

Figure 5.10: Definition Screen for Project Phases

The individuals object is represented in the data warehouse as two fact tables to store the data regarding all who are involved in industrial construction projects from both the industrial owners and contractors. Some of the fields in these tables can be automatically populated from the HRMS, others such as level of technical and managerial experience would require a supervisor to assess them.

The data in the tables is limited to what the data warehouse needs to produce the reports and provide a tool to exchange knowledge. The individuals' data is linked to an industrial owner or a contractor. After that, they are linked to the Resources Breakdown Structure (RBS), which consists of three levels; category of resources, group of resources, and resource. The position of the individual on this tree hierarchy represents that individual's functional reporting relationships in the organization.

Figure 5.11-a shows the data entry screen for the contractors staff. The history of all positions held and responsibility assigned to each individual is stored in another two fact tables in the data warehouse. The responsibility assignment screen for internal projects is shown in Figure 5.11-b.

Project-based Production Facilities' Staff

Group of Organizations	WorleyParsons Ltd.	Group	Project Services
Organization	Colt WorleyParsons	Department	Project Controls
Business Unit	Edmonton Office	Discipline	Management
Last Name	Raj	First Name	Hans
Primary Role	Functional Manager	Initials	RH
Secondary Role1	Team Leader	Secondary Role2	
Office Location		Phone	
Email		Fax	
NetworkID		Mobile	
TechnicalLevel	Senior	Access Permission	
ManagerialLevel	Senior	Active	<input checked="" type="checkbox"/>
Normal Hours / Week	40	Max Hours / Week	50
Comments			

Record: 1 of 9

Figure 5.11-a: Data Entry Screen for the Contractor's Staff

Project-based Production Facility Responsibility History

Organization		Resources	
Group	WorleyParsons Ltd.	Group	Project Services
Organization	Colt WorleyParsons	Department	Project Management
Business Unit	Edmonton Office	Discipline	
Internal Program	SUNCOR SoC	Manager	Srivastava, Ajay
Group of Projects	Steepbank Extraction Plant	Start Date	01-May-04
Internal Project		Finish Date	31-May-07
Comments			

Record: 1 of 1

Figure 5.11-b: Responsibility Data Entry Screen for the Contractors' Staff

5.5 THE DATA WAREHOUSE OUTPUT (FRONTEND)

5.5.1 Populating the Data Warehouse

In order to demonstrate the applicability of the integrated data collection structure, a data set was obtained from one of the contractors that are partners in the NSERC - Alberta Construction Industry Chair. The partner company is a global EPCM firm that ranks amongst the top 10 in engineering firms in the world (Engineering News Records - ENR, 2007). This firm specializes in all fields of industrial construction such as hydrocarbons, minerals and metals, and power generation. The data contained basic attributes for a set of industrial owners, industrial construction projects and a collection of internal projects that are grouped in four internal programs. For each internal project, the resource data was obtained and missing data was obtained using a random number generating function. All numbers were multiplied by random numbers for confidentiality purposes. The dataset has thousands of records and the prototype data warehouse is capable of handling the data and producing the reports without any noticeable delays. Table 5.1 shows an example of the data set. All the required dimension tables were also populated.

The dataset is used to generate both detailed and summary project reports for performance evaluation and benchmarking at the project level. Examples of these reports are shown in Figures 5.12 and 5.13. Moreover, the dataset is used to generate detailed OLAP reports as explained in the following section of this chapter.

ID	InternalProject	Resource	Phase	Period Number	Week Ending	YearID	QuarterID	MonthID	WeekID	Current BL Hours	Earned Hours	Actual Hours	Min O
1	1	2	8	1	28-May-04	2004	2	5	22	64.00	23.22	19.00	
2	1	2	8	2	04-Jun-04	2004	2	6	23	4.00	4.61	7.00	
3	1	2	8	3	11-Jun-04	2004	2	6	24	35.00	41.10	32.00	
4	1	2	8	4	18-Jun-04	2004	2	6	25	28.00	53.70	60.00	
5	1	2	8	5	25-Jun-04	2004	2	6	26	16.00	26.70	56.00	
6	1	2	8	6	02-Jul-04	2004	3	7	27	89.00	65.22	53.00	
7	1	2	8	7	09-Jul-04	2004	3	7	28	117.00	21.09	27.00	
8	1	2	8	8	16-Jul-04	2004	3	7	29	24.00	40.88	30.00	
9	1	2	8	9	23-Jul-04	2004	3	7	30	36.00	22.76	38.00	
10	1	2	8	10	30-Jul-04	2004	3	7	31	70.00	71.68	42.00	
11	1	2	8	11	06-Aug-04	2004	3	8	32	49.00	55.18	34.00	
12	1	2	8	12	13-Aug-04	2004	3	8	33	25.00	37.60	30.00	
13	1	2	8	13	27-Aug-04	2004	3	8	35	11.00	7.17	18.00	
14	1	2	8	14	03-Sep-04	2004	3	9	36	88.00	82.72	58.00	
15	1	2	8	15	10-Sep-04	2004	3	9	37	49.00	76.12	44.00	
16	1	2	8	16	17-Sep-04	2004	3	9	38	197.00	71.20	40.00	
17	1	4	8	1	21-May-04	2004	2	5	21	2.00	2.99	2.00	
18	1	4	8	2	28-May-04	2004	2	5	22	3.00	4.99	5.00	
19	1	4	8	3	04-Jun-04	2004	2	6	23	4.00	2.90	8.00	
20	1	4	8	4	11-Jun-04	2004	2	6	24	6.00	2.64	2.00	
21	1	4	8	5	18-Jun-04	2004	2	6	25	4.00	2.19	6.00	
22	1	4	8	6	25-Jun-04	2004	2	6	26	2.00	2.04	2.00	
23	1	4	8	7	02-Jul-04	2004	3	7	27	3.00	1.49	3.00	
24	1	4	8	8	09-Jul-04	2004	3	7	28	5.00	2.76	3.00	
25	1	4	8	9	16-Jul-04	2004	3	7	29	20.00	20.82	17.00	
26	1	4	8	10	23-Jul-04	2004	3	7	30	14.00	14.60	13.00	
27	1	4	8	11	30-Jul-04	2004	3	7	31	4.00	6.42	12.00	
28	1	4	8	12	13-Aug-04	2004	3	8	33	8.00	5.90	23.00	
29	1	4	8	13	20-Aug-04	2004	3	8	34	45.00	24.52	19.00	
30	1	4	8	14	27-Aug-04	2004	3	8	35	45.00	39.28	32.00	
31	1	4	8	15	03-Sep-04	2004	3	9	36	19.00	4.66	3.00	
32	1	4	8	16	10-Sep-04	2004	3	9	37	1.00	0.41	2.00	
33	1	5	8	1	28-May-04	2004	2	5	22	13.00	7.75	8.00	
34	1	5	8	2	04-Jun-04	2004	2	6	23	9.00	16.22	10.00	
35	1	5	8	3	11-Jun-04	2004	2	6	24	9.00	10.11	6.00	
36	1	5	8	4	18-Jun-04	2004	2	6	25	5.00	5.91	7.00	
37	1	5	8	5	25-Jun-04	2004	2	6	26	4.00	4.21	2.00	
38	1	5	8	6	02-Jul-04	2004	3	7	27	76.00	41.53	23.00	
39	1	5	8	7	09-Jul-04	2004	3	7	28	5.00	7.24	34.00	
40	1	5	8	8	16-Jul-04	2004	3	7	29	31.00	13.23	18.00	
41	1	5	8	9	23-Jul-04	2004	3	7	30	8.00	1.68	2.00	
42	1	5	8	10	30-Jul-04	2004	3	7	31	4.00	2.89	20.00	
43	1	5	8	11	06-Aug-04	2004	3	8	32	21.00	12.51	17.00	
44	1	5	8	12	13-Aug-04	2004	3	8	33	202.00	42.64	24.00	
45	1	5	8	13	20-Aug-04	2004	3	8	34	32.00	48.09	26.00	
46	1	5	8	14	27-Aug-04	2004	3	8	35	195.00	55.26	63.00	
47	1	5	8	15	03-Sep-04	2004	3	9	36	7.00	5.02	5.00	
48	1	5	8	16	10-Sep-04	2004	3	9	37	1.00	1.31	1.00	

Table 5.1: Example of the Data Set

<i>Summary Internal Project Hours At Completion Report</i>					
	<i>Original Baseline</i>	<i>Current Baseline</i>	<i>% Variance</i>	<i>Actual Hours</i>	<i>% Variance</i>
04E2665, Dry Surge Project					
Industrial Construction Projects, Engineering Phases, Front End Loading (FEL) II	1,388.00	2,326.00	67.58%	1,238.00	-0.47
Industrial Construction Projects, Engineering Phases, Front End Loading (FEL) III	2,794.00	4,368.00	56.34%	2,822.00	-0.35
Industrial Construction Projects, Engineering Phases, Detailed Engineering & Design	16,076.00	30,074.00	87.07%	17,214.00	-0.43
Subtotal Per Internal Project	20,258.00	36,768.00	81.50%	21,274.00	-0.42
04E2662, MFT Transfer Pond 6 To Pond 7 Project					
Industrial Construction Projects, Engineering Phases, Front End Loading (FEL) II	1,675.00	2,412.00	44.00%	1,718.00	-0.29
Industrial Construction Projects, Engineering Phases, Front End Loading (FEL) III	6,140.00	9,462.00	54.10%	6,366.00	-0.33
Industrial Construction Projects, Engineering Phases, Detailed Engineering & Design	9,344.00	15,904.00	70.21%	10,103.00	-0.36
Subtotal Per Internal Project	17,159.00	27,778.00	61.89%	18,187.00	-0.35

Figure 5.12: Example of the Summary Internal Project Hours at Completion

Detailed Internal Project Hours At Completion Report

<i>Resource</i>	<i>Original Baseline</i>	<i>Current Baseline</i>	<i>% Variance</i>	<i>Actual Hours</i>	<i>% Variance</i>
04E2665, Dry Surge Project					
Industrial Construction Projects, Engineering Phases, Front End Loading (FEL) II					
Engineering Services, Piping - Design/CAD	702.00	882.00	25.64%	588.00	-0.33
Project Services, Project Controls - Cost Estimating	139.00	185.00	33.09%	152.00	-0.18
Project Services, Project Controls - Cost Control	279.00	624.00	123.66%	268.00	-0.57
Construction Services, Construction General Mana	268.00	635.00	136.94%	230.00	-0.64
Subtotal Per Project Phase	1,388.00	2,326.00	67.58%	1,238.00	-0.47
Industrial Construction Projects, Engineering Phases, Front End Loading (FEL) III					
Engineering Services, Piping - Design/CAD	199.00	286.00	43.72%	235.00	-0.18
Project Services, Project Controls - Cost Estimating	153.00	389.00	154.25%	169.00	-0.57
Project Services, Project Controls - Cost Control	523.00	796.00	52.20%	521.00	-0.35
Construction Services, Construction General Mana	1,919.00	2,897.00	50.96%	1,897.00	-0.35
Subtotal Per Project Phase	2,794.00	4,368.00	56.34%	2,822.00	-0.35

Wednesday, February 11, 2009

Page 1 of 113

Figure 5.13: Example of the Detailed Internal Project Hours at Completion

5.5.2 OLAP Reports

The difference between OLAP reporting and traditional reporting is that OLAP focuses on analyzing and exploring historical data, meanwhile traditional reporting focuses on accessing data for daily business needs (Howson, 2008). Howson also stated that OLAP reports do not have to be generated from a specific OLAP reporting tool as long as they maintain a set of characteristics. These characteristics are:

1. Multidimensionality, which provide the users with ability to analyze data from different angles.
2. Highly interactive, where the users are provided with the ability to select different grouping and filtering variables for their reports.

3. Variability of aggregating levels, where the users are able to view the data at a highest possible level or drill down to any required level of detail.
4. Cross-dimensional calculations, which requires the development of complex queries to enable the users to analyze the data using a single dimension or multiple dimensions.
5. Speed, which means pre-prepared query results are stored in the data warehouse to shorten the report processing time to a minimum possible.

The developed prototype considered all these aspects for producing the reports. The multidimensionality that was established through the snowflake schema provided vast options for data viewing from different angles according to users' needs. The users are able to dynamically customize the reports by selecting the grouping and filtering attributes. The produced reports can be aggregated to any required level starting from the work package all the way up to a contracting company or industrial owner. The prototype allows the use of multiple dimensions to combine in complex queries such as combining time, phases, resources and grouping levels in one report. All the generated reports from the prototype took less than a fraction of a second to be produced due to the preprocessing and summarization of data.

The design of snowflake schema is one of the main challenges in this research and consumed months and efforts to achieve the optimum design that properly represent the data structure and is capable of producing all the necessary reports.

The snowflake schema and the sophisticated structured queries are used to enable the users to view the stored data in the data warehouse from any required view to meet the business needs. The users include both axes of the matrix organizations, functional and project managers. OLAP techniques include slice and dice, drill-down and roll-up and pivoting. Slice and dice provides the users with filters and grouping capability to view a specific sub-set of the stored data. Drill-down and roll-up provides the users with aggregating capability to view a specific sub-set of the stored data according to the required level of detail. Pivoting enables the users to view the data from different angles. All these techniques are applied to the data warehouse using the simulated set of data as shown in this section.

5.5.3 Utilizing the Slice and Dice OLAP Technique

The slice and dice technique is ideal for viewing resource data in both vertical and horizontal directions. Slice technique is used when dealing with one dimension of the data, while dice is used when dealing with multiple dimensions. Vertical grouping summarizes resource data for a single internal or industrial project to enable project managers from either contractor or owner side to perform their analysis. The analysis includes package durations, resource amounts and performance measures. This analysis accompanied with lessons-learned from previously completed projects can help in improving performance of new projects, avoiding repeating the same mistakes and minimizing any false perceptions that may exist. An example of a detailed report showing every resource per work package for an internal project is shown in Figure 5.14.

Detailed Internal Project Hours Grouped by Phase and Work Package

	Original			Current			Actual		
	Quantity	Budget	Hours/Item	Quantity	Budget	Hours/Item	Quantity	Budget	Hours/Item
Internal Project: Breaker Pump Box									
Project Phase: Front End Loading (FEL) I									
Work Package: Block Diagrams for BPB Scope of Work									
Engineering Services, Electrical - Design/CAD		482.00	4.73		701.00	8.45		598.00	3.11
Engineering Services, Instrumentation - Design		413.00	4.05		381.00	4.59		366.00	1.91
Engineering Services, Civil - Structural - Engin		259.00	2.54		161.00	1.94		659.00	3.43
Work Package Sub-total:	102	1,154	11.31	83	1,243	14.98	192	1,623	8.45
Work Package: Buildings for BPB Scope of Work									
Engineering Services, Electrical - Design/CAD		390.00	2.75		20.00	0.11		612.00	6.58
Engineering Services, Civil - Structural - Desig		419.00	2.95		562.00	2.96		46.00	0.49
Engineering Services, Instrumentation - Engine		423.00	2.98		541.00	2.85		190.00	2.04
Work Package Sub-total:	142	1,232	8.68	190	1,123	5.91	93	848	9.12

Figure 5.14: Detailed Single-Project Multiple-Resources Report

The report is grouped by project phase and work package, however it can be grouped or filtered by any of the control attributes such as internal program, etc. A summarized version of the report without resource details is shown in Figure 5.15. These reports are also produced for a single industrial project showing the resource data from all the internal projects that were performed. All numbers shown in the report were modified for confidentiality purposes.

Detailed Internal Project Hours Grouped by Phase and Work Package

	Original			Current			Actual		
	Quantity	Budget	Hours/Item	Quantity	Budget	Hours/Item	Quantity	Budget	Hours/Item
Internal Project: Breaker Pump Box									
Project Phase: Front End Loading (FEL) I									
Block Diagrams for BPB Scope of Work	102	1,154	11.31	83	1,243	14.98	192	1,623	8.45
Buildings for BPB Scope of Work	142	1,232	8.68	190	1,123	5.91	93	848	9.12
Foundations for BPB Scope of Work	172	766	4.45	46	1,276	27.74	96	1,097	11.43
Installation Details for BPB Scope of Work	211	793	3.76	113	1,027	9.09	240	829	3.45
Instrumentation Dwgs for BPB Scope of Work	11	814	74.00	201	913	4.54	150	1,467	9.78
Layouts for BPB Scope of Work	31	1,065	34.35	138	1,163	8.43	246	1,175	4.78
Lists for BPB Scope of Work	185	924	4.99	154	987	6.41	50	1,393	27.86
Mgmt. & Coord. for BPB Scope of Work	24	1,445	60.21	50	975	19.50	60	779	12.98
Misc. Dwgs. for BPB Scope of Work	113	1,352	11.96	122	1,028	8.43	112	1,256	11.21

Saturday, February 07, 2009

Page 1 of 18

Figure 5.15: Summarized Single-Project Multiple-Resource Report

Horizontal summarization is applied to view single resource data from multiple internal or industrial projects. These reports are designed for resource managers and benchmarking purposes. An example of a detailed report for a single resource grouped by internal project and project phase is shown in Figure 5.16. A summarized version of the report that doesn't demonstrate the project phases' data is shown in Figure 5.17. These report examples illustrate the endless powerful capabilities of the slice and dice OLAP techniques once the data is generated, collected and stored properly in the data warehouse. The snowflake schema combined with the slice and dice technique provides the user with ability to develop very powerful dynamic queries to meet their exact specific needs regardless of their managerial level in the matrix organization structure.

Resource Hours Grouped by Project and Phase

	Original			Current			Actual		
	Quantity	Budget	Hours/Item	Quantity	Budget	Hours/Item	Quantity	Budget	Hours/Item
PPF Resource: Engineering Services, Civil - Structural - Design/CAD									
Internal Project: Breaker Pump Box									
Front End Loading (FEL) I			3,725			3,824			3,192
Front End Loading (FEL) II			3,589			5,139			1,910
Front End Loading (FEL) III			3,895			4,279			3,106
Internal Project Sub-total			11,209			13,242			8,208
Internal Project: Dry Surge									
Front End Loading (FEL) II			4,466			3,939			3,224
Front End Loading (FEL) III			2,299			3,031			3,876
Internal Project Sub-total			6,765			6,970			7,100

Saturday, February 07, 2009

Page 1 of 11

Figure 5.16: Detailed Single-Resource Multiple-Projects Report

Resource Hours Grouped by Program and Project

	Original Baseline			Current Baseline			Actual		
	Quantity	Budget	Hours/Item	Quantity	Budget	Hours/Item	Quantity	Budget	Hours/Item
PPF Resource: Engineering Services, Civil - Structural - Design/CAD									
Breaker Pump Box			11,209			13,242			8,208
Dry Surge			6,765			6,970			7,100
MFT Transfer Pond 6 To Pond 7			6,351			9,130			7,256
South Booster Pump House (SBPH)			8,545			9,772			9,189
Internal Program Sub-total:			32,870			39,114			31,753
PPF Resource Sub-tot			32,870			39,114			31,753
PPF Resource: Engineering Services, Civil - Structural - Engineering									
Breaker Pump Box			7,745			8,618			8,159
Dry Surge			5,954			4,891			4,094
MFT Transfer Pond 6 To Pond 7			11,350			12,201			12,289

Saturday, February 07, 2009

Page 1 of 4

Figure 5.17: Summarized Single-Resource Multiple-Projects Report

The same set of reports is also produced for industrial projects to enable industrial owners to compare performance of the same resource between various internal projects that are performed by different contractors.

5.5.4 Utilizing the Roll-up and Drill-down OLAP Technique

The developed data structure and snowflake schema in the data warehouse enables the full utilization of the roll-up and drill-down OLAP technique; also known in the industry as the-peel-the-onion technique. The available resource data is viewed either in a very detailed format at the resource per work package level or in different levels of summarized format up to the industrial portfolio levels. Because of the way the data collection was structured and the way the metadata was assigned, the output is seamlessly obtained. Figure 5.18 shows an example of a report grouped by internal program, meanwhile Figure 5.19 shows an example of a report grouped by industrial portfolio.

Internal Program Hours Grouped by Project and Phase

	Original Baseline	Current Baseline	Actual
Internal Program: SUNCOR SoC			
Internal Project: Breaker Pump Box			
Engineering Services, Civil - Structural - Design/CAD	3,725	3,824	3,192
Engineering Services, Civil - Structural - Engineering	3,332	2,771	2,436
Engineering Services, Electrical - Design/CAD	3,307	4,090	4,600
Engineering Services, Electrical - Engineering	1,817	1,347	2,572
Engineering Services, Instrumentation - Design/CAD	1,833	1,595	2,037
Engineering Services, Instrumentation - Engineering	4,087	4,695	3,367
Front End Loading (FEL) I	18,101	18,322	18,204
Engineering Services, Civil - Structural - Design/CAD	3,589	5,139	1,910
Engineering Services, Civil - Structural - Engineering	2,031	1,992	2,666

Saturday, February 07, 2009

Page 1 of 3

Figure 5.18: Summarized Report Grouped by Internal Program

Resource Hours Grouped by Program and Project

	Original Baseline	Current Baseline	Actual
Industrial Portfolio: Debottlenecking Programs			
Breaker Pump Box (Initiative # 02)	54,850	60,956	53,434
Dry Surge (Initiative # 01)	35,968	37,288	37,728
Industrial Program Sub-total:	90,818	98,244	91,162
Industrial Portfolio Sub-tota	90,818	98,244	91,162
Industrial Portfolio: Tailings Programs			
MFT Transfer Pond 6 To Pond 7	53,058	58,433	57,371
South Booster Pump House (SBPH)	57,565	58,079	55,378
Industrial Program Sub-total:	110,623	116,512	112,749
Industrial Portfolio Sub-tota	110,623	116,512	12,749

Saturday, February 07, 2009

Page 1 of 1

Figure 5.19: Summarized Report Grouped by Industrial Portfolio

5.5.5 Utilizing the Pivoting OLAP Technique

Pivoting is one of the most useful and powerful OLAP techniques. It provides users with infinite possibilities to view and analyze stored data. The key to successful data pivoting is the consistency in storing data and assigning the right metadata to it. Pivot reports present the data in a two-dimensional matrix that has grouping filter(s) and multiple row and column headings. In addition, pivoting provides users with the ability to utilize control attributes for filtering the data to view only the required data subset. Figure 5.20 presents the pivot structure for a report showing the three main resource data elements (current baseline, earned and actual spent values) for a single contractor represented in hours. The values are grouped by year, quarter and month and can be filtered by internal portfolio, program, project phase or resource. The pivot table report is shown in Figure 5.21. Whereas, Figure 5.22 is a graphical representation of the Current Planned Values (CPV), and 5.23 is a graphical representation of Actual Earned Values (AEV). Both these graphs are generated automatically from the pivot table report. The pivot table and graphs allow contractors to analyze their performance over the years as shown in Figure 5.24.

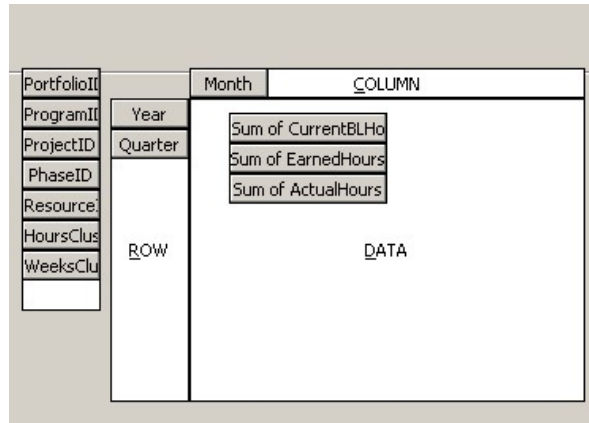


Figure 5.20: The Pivoting Structure for the Three Main Data Elements

PortfolioID	(All)		Month												Grand Total	
ProgramID	(All)		1	2	3	4	5	6	7	8	9	10	11	12		
ProjectID	(All)															
PhaseID	(All)															
ResourceID	(All)															
Year	Quarter	Data														
2004	1	Sum of CurrentBLHours	622.85	1,116.90	2,465.71											4,205.47
		Sum of EarnedHours	343.92	866.58	1,367.94											2,578.44
		Sum of ActualHours	320.25	995.50	1,469.75											2,785.50
	2	Sum of CurrentBLHours				3,119.14	3,189.73	4,266.27								10,575.14
		Sum of EarnedHours				2,105.55	1,803.45	2,548.05								6,457.05
		Sum of ActualHours				2,048.50	1,697.00	2,603.75								6,349.25
	3	Sum of CurrentBLHours							6,492.17	5,641.33	9,772.38					21,905.88
		Sum of EarnedHours							3,573.26	3,549.10	5,867.44					12,989.81
		Sum of ActualHours							3,883.00	3,681.50	5,655.50					13,220.00
	4	Sum of CurrentBLHours										9,461.66	9,112.16	13,446.50		32,020.32
		Sum of EarnedHours										6,550.52	6,577.51	7,386.12		20,514.15
		Sum of ActualHours										6,747.00	6,346.75	7,571.50		20,665.25
2004 Sum of CurrentBLHours			622.85	1,116.90	2,465.71	3,119.14	3,189.73	4,266.27	6,492.17	5,641.33	9,772.38	9,461.66	9,112.16	13,446.50	68,706.81	
2004 Sum of EarnedHours			343.92	866.58	1,367.94	2,105.55	1,803.45	2,548.05	3,573.26	3,549.10	5,867.44	6,550.52	6,577.51	7,386.12	42,539.55	
2004 Sum of ActualHours			320.25	995.50	1,469.75	2,048.50	1,697.00	2,603.75	3,883.00	3,681.50	5,655.50	6,747.00	6,346.75	7,571.50	43,020.00	
2005	1	Sum of CurrentBLHours	16,130.95	17,150.69	19,165.88										52,447.53	
		Sum of EarnedHours	8,520.27	10,434.89	10,840.07										29,795.23	
		Sum of ActualHours	7,902.50	10,076.75	10,249.50										28,228.75	
	2	Sum of CurrentBLHours				21,241.00	20,417.93	15,374.62							57,033.56	
		Sum of EarnedHours				13,084.20	11,890.11	9,744.94							34,719.24	
		Sum of ActualHours				13,798.75	11,550.25	11,251.75							36,600.75	
	3	Sum of CurrentBLHours							22,575.58	16,012.55	18,275.68				56,863.82	
		Sum of EarnedHours							12,335.74	10,095.47	13,845.82				36,277.02	
		Sum of ActualHours							13,143.50	10,606.25	13,406.00				37,155.75	
	4	Sum of CurrentBLHours										17,004.82	14,145.73	10,535.39	41,685.94	
		Sum of EarnedHours										9,540.71	8,392.45	6,579.96	24,513.13	
		Sum of ActualHours										9,764.50	8,915.00	6,688.50	25,368.00	
2005 Sum of CurrentBLHours			16,130.95	17,150.69	19,165.88	21,241.00	20,417.93	15,374.62	22,575.58	16,012.55	18,275.68	17,004.82	14,145.73	10,535.39	208,030.85	
2005 Sum of EarnedHours			8,520.27	10,434.89	10,840.07	13,084.20	11,890.11	9,744.94	12,335.74	10,095.47	13,845.82	9,540.71	8,392.45	6,579.96	125,304.62	
2005 Sum of ActualHours			7,902.50	10,076.75	10,249.50	13,798.75	11,550.25	11,251.75	13,143.50	10,606.25	13,406.00	9,764.50	8,915.00	6,688.50	127,353.25	
2006	1	Sum of CurrentBLHours	9,232.19	9,925.96	13,898.15										33,056.30	
		Sum of EarnedHours	5,678.15	5,992.84	7,213.16										18,884.16	
		Sum of ActualHours	5,347.50	5,744.50	7,059.25										18,151.25	
	2	Sum of CurrentBLHours				9,199.65	9,002.32	8,397.22							26,599.19	
		Sum of EarnedHours				5,770.22	5,112.62	5,492.65							16,375.49	
		Sum of ActualHours				5,874.75	5,366.00	5,795.50							17,036.25	
	3	Sum of CurrentBLHours							5,892.69	4,112.97	5,420.16				15,425.81	
		Sum of EarnedHours							3,652.19	2,797.46	3,223.13				9,672.78	
		Sum of ActualHours							3,452.25	2,752.75	3,272.25				9,477.25	
	4	Sum of CurrentBLHours										4,943.57	2,578.02	2,678.64	10,200.23	
		Sum of EarnedHours										2,542.38	1,965.19	1,961.88	6,469.45	
		Sum of ActualHours										2,460.25	2,103.00	1,930.25	6,493.50	

Figure 5.21: Pivoting Report Grouped by Year, Quarter and Month

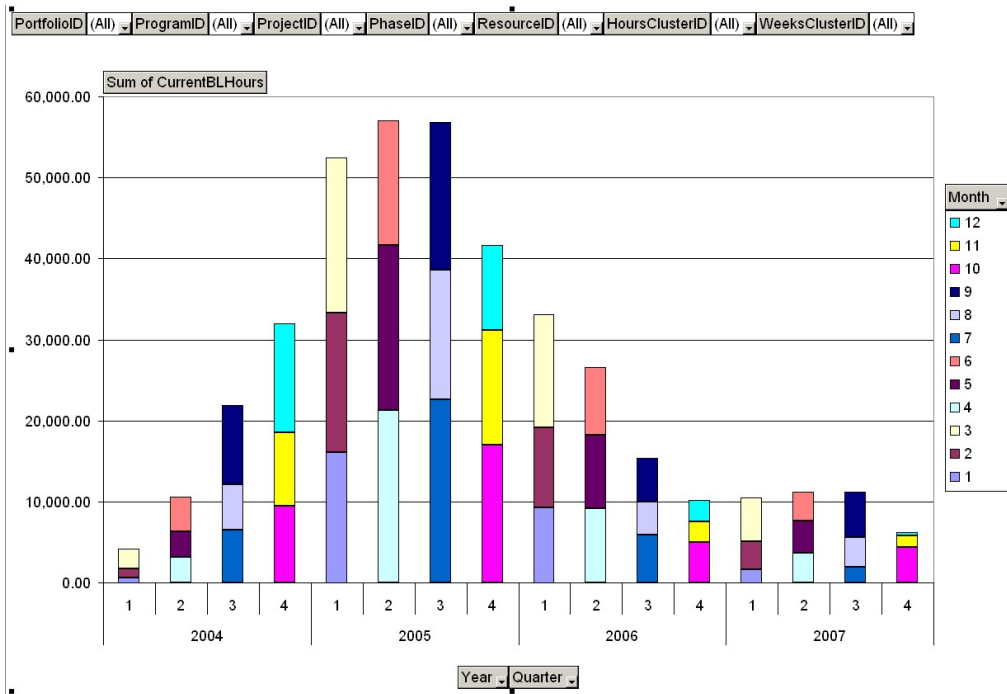


Figure 5.22: Dynamic Pivot Graph for Current Planned Values (CPV)

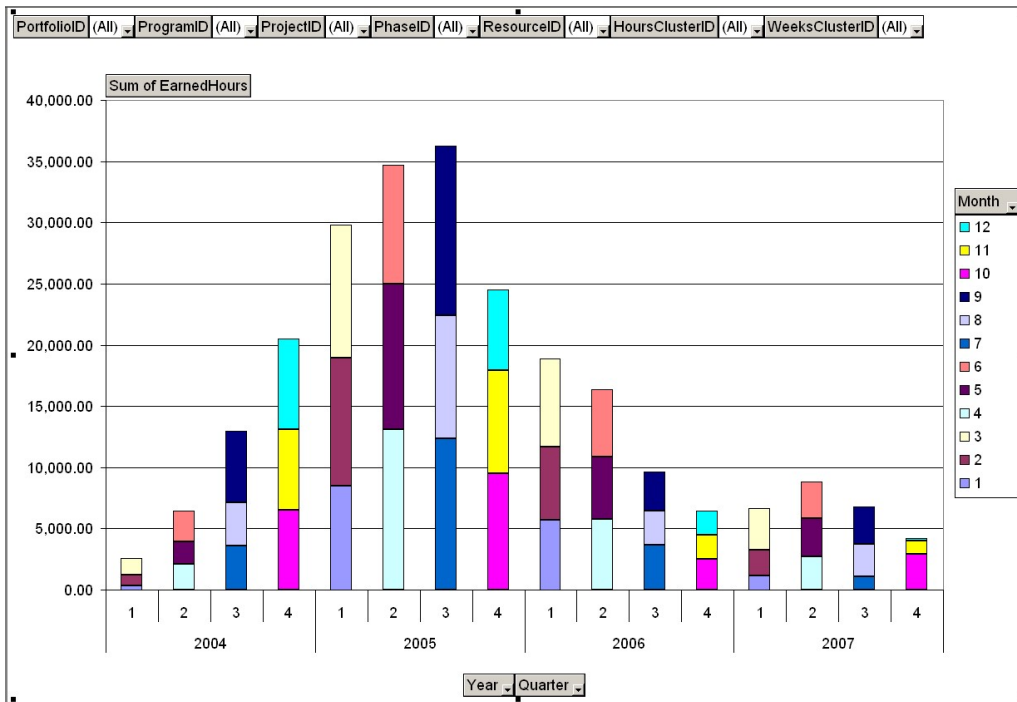


Figure 5.23: Dynamic Pivot Graph for Actual Earned Values (AEV)

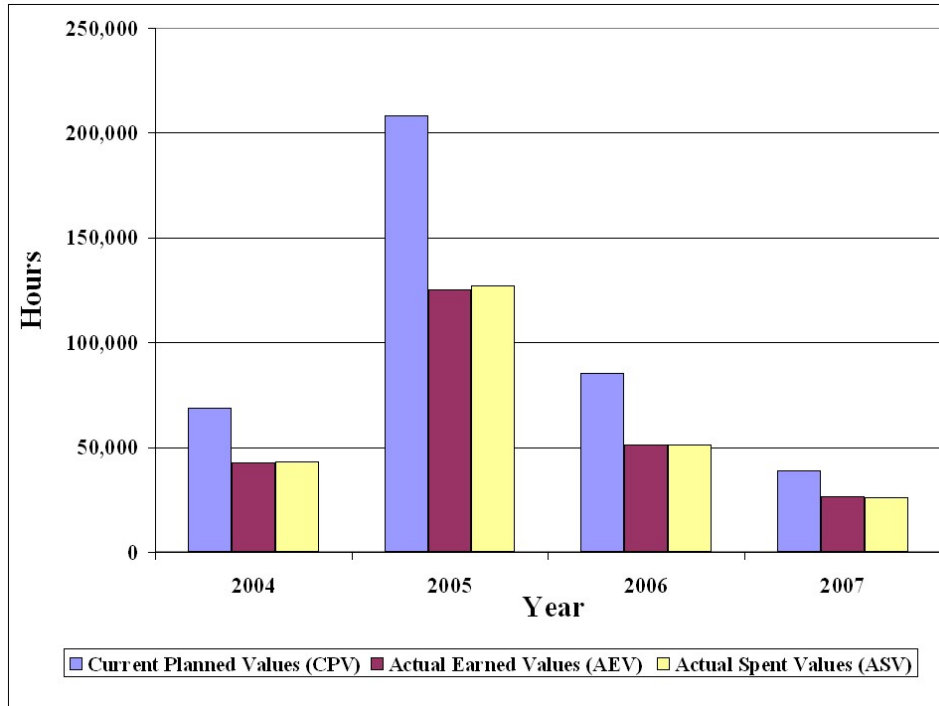


Figure 5.24: Summary of the Three Main Data Elements over Years

Another variation of analyzing the same dataset is by using pivoting by resource and phase as shown in Figure 5.25. This report is summarized in Figures 5.26 and 5.27 as columnar and pie-chart formats in order to show the variation in the contribution of each resource to the ASV between different years.

		2004 Total											2005											2006 Total											2007							
Phase	Resource	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	2007				
6	1	Sum of CurrentBLHours	10.38	33.54	893.69	64.98	14.71	6.00	1.39																	87.07													87.07			
		Sum of EarnedHours	10.11	9.52	571.19	39.42	14.66	9.63	0.19																		63.90													63.90		
		Sum of ActualHours	12.75	6.50	583.50	36.25	15.25	6.50	1.00																		59.00													59.00		
	2	Sum of CurrentBLHours	176.20	377.83	2,459.51	191.23	240.96	190.34	37.52	80.71	66.34	57.40	1.21	56.08	97.97	129.55	326.82										1,476.13	215.25												215.25		
		Sum of EarnedHours	160.93	222.50	1,518.91	149.13	193.58	83.22	29.19	47.76	54.01	55.07	1.66	39.35	67.70	84.32	97.21										902.19	145.61												145.61		
		Sum of ActualHours	192.00	216.00	1,619.00	127.00	169.25	92.00	30.50	55.00	60.50	68.00	1.00	28.25	64.00	65.75	133.75										895.00	124.65												124.65		
4	1	Sum of CurrentBLHours	221.85	277.61	3,316.34	125.61	24.76	29.12	29.50	22.08	74.45	15.09	27.00	27.00	104.88	79.97											575.19	119.53												119.53		
		Sum of EarnedHours	226.35	173.27	1,779.73	80.04	20.18	30.05	20.88	10.75	32.38	13.34	24.98	23.05	132.37	89.36											492.76	102.25												102.25		
		Sum of ActualHours	200.00	136.50	1,729.75	98.75	34.50	44.00	17.50	11.25	25.50	19.00	25.00	37.25	124.00	84.00											538.50	102.25												102.25		
	5	Sum of CurrentBLHours			206.48																							141.07												141.07		
		Sum of EarnedHours			145.89																								131.44												131.44	
		Sum of ActualHours			194.75																								86.25												86.25	
6	Sum of CurrentBLHours	29.05	20.04	1,282.90	35.61	132.76	272.62	134.42	209.43	95.92	175.91	155.45	135.84	122.12	222.32	301.13											1,993.54	103.34												103.34		
	Sum of EarnedHours	17.11	8.48	831.01	37.60	102.29	95.15	101.57	174.95	85.82	91.70	99.58	101.71	88.44	104.21	146.69											1,229.70	64.83												64.83		
	Sum of ActualHours	41.50	12.50	828.00	58.50	129.00	103.50	108.25	160.50	135.00	85.50	100.50	98.00	132.25	126.00	163.00											1,400.00	55.75												55.75		
6	Sum of CurrentBLHours	437.48	709.02	8,158.93	417.43	413.19	498.09	202.82	312.22	236.72	311.38	242.34	216.92	344.38	431.84	643.68											4,273.90	438.19												438.19		
	Sum of EarnedHours	414.51	413.27	4,846.73	306.20	330.70	218.04	151.82	233.46	172.21	238.80	172.23	164.10	295.25	277.68	259.30											2,819.99	312.24												312.24		
	Sum of ActualHours	446.25	371.50	4,955.00	320.50	348.00	246.00	157.25	226.75	221.00	221.50	156.25	163.50	327.75	257.75	314.50											2,978.75	282.00												282.00		
7	1	Sum of CurrentBLHours	736.87	1,055.84	9,476.21	1,030.18	300.64	680.79	758.95	51.89	29.76	176.14	219.24	420.94	5.09	151.48	234.09										4,050.19	161.01												161.01		
		Sum of EarnedHours	306.83	362.16	5,612.86	434.14	239.92	457.64	402.58	63.61	9.01	159.30	111.34	191.60	9.42	126.03	199.22										2,403.81	182.08												182.08		
		Sum of ActualHours	372.75	341.00	5,518.50	359.00	259.25	320.75	404.50	44.25	14.25	176.25	107.25	200.75	10.50	112.00	195.75										2,204.50	151.00												151.00		
	2	Sum of CurrentBLHours			446.31																							1,540.90	572.19												572.19	
		Sum of EarnedHours			470.94																								1,320.04	243.94												243.94
		Sum of ActualHours			541.50																								1,258.50	231.50												231.50
4	1	Sum of CurrentBLHours	13.73		35.16	139.16	9.42																				930.91	264.93												264.93		
		Sum of EarnedHours	10.80		15.67	53.96	10.30																				527.91	179.14												179.14		
		Sum of ActualHours	18.75		30.50	37.00	12.50																				489.50	152.25												152.25		
	5	Sum of CurrentBLHours																										474.32	94.24												94.24	
		Sum of EarnedHours																										130.66	130.11												130.11	
		Sum of ActualHours																										234.25	78.25												78.25	
6	Sum of CurrentBLHours	29.67	6.25	560.06																							606.72	229.35												229.35		
	Sum of EarnedHours	27.05	10.24	269.66																							370.51	260.53												260.53		
	Sum of ActualHours	36.25	8.25	271.25																							320.00	187.75												187.75		
7	Sum of CurrentBLHours	700.26	1,062.09	10,517.74	1,169.35	310.05	722.49	868.69	293.19	214.21	570.50	546.57	956.42	537.09	609.56	1,204.91											8,012.03	1,311.70												1,311.70		
	Sum of EarnedHours	344.68	372.39	6,369.13	488.10	250.22	476.14	511.80	173.39	177.77	370.11	372.19	507.83	323.97	337.09	897.64											4,889.24	995.88												995.88		
	Sum of ActualHours	427.75	349.25	6,361.75	396.00	271.75	343.50	489.50	158.00	163.75	390.00	343.00	547.75	293.00	341.50	831.00											5,468.75	800.75												800.75		
8	1	Sum of CurrentBLHours	416.23	425.81	3,264.10	416.29	284.09	165.05	202.75	1,442.52	475.92	1,891.49	459.21	6.37	1.12											5,344.81																
		Sum of EarnedHours	387.93	144.59	1,845.24	285.08	270.80	156.47	142.78	828.21	667.51	773.65	278.83	7.08	2.03											3,412.44																
		Sum of ActualHours	314.00	130.50	1,574.50	287.50	241.00	140.75	162.50	624.75	751.50	806.00	363.50	12.00	1.75											3,391.25																
	2	Sum of CurrentBLHours	397.95	42.66	2,213.37	13.53																					1,113.71															
		Sum of EarnedHours	381.04	58.27	1,857.49	9.09																					648.18															
		Sum of ActualHours	332.50	54.00	1,733.75	20.25																					635.25															
4	Sum of CurrentBLHours	190.23	140.23	1,093.71	42.11																					256.03	47.44												47.44			
	Sum of EarnedHours	228.99	126.19	974.21	21.02																					177.17	11.38												11.38			
	Sum of ActualHours	254.50	86.50	1,049.50	17.25																					149.25	15.23												15.23			
5	Sum of CurrentBLHours	130.56		1,255.69	40.47	20.14	1.74	10.71	24.08	2.53	5.01															411.58	145.66												145.66			
	Sum of EarnedHours	145.17		871.92	20.13	5.86</																																				

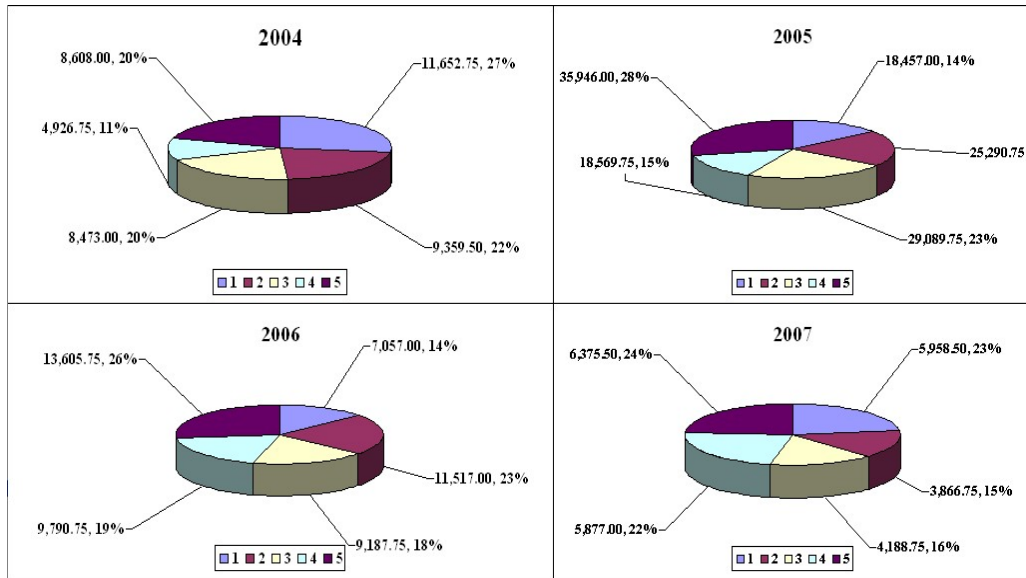


Figure 5.27: Resource Contribution to ASV per Year in Pie-chart Format

In addition to yearly and monthly viewing, it is possible to view the data on a detailed weekly level. Figure 5.28 shows the three main data elements distributed over a weekly calendar basis. Such a representation allows users to analyze resource utilization over detailed time periods.

On the other hand, Figure 5.29 shows the same data subset distributed over reporting period for comparing resource utilization as projects progress. The purpose of showing this sample of reports is to illustrate the powerful capabilities of dynamic reporting techniques in helping various users to view and analyze historical data, make sound and timely decisions, maintain competitive edge and maximize the business profit.

PortfolioID	(All)									
ProgramID	(All)									
PhaseID	(All)									
ResourceID	(All)									
ProjectID	Data	30-Jan-04	06-Feb-04	13-Feb-04	20-Feb-04	27-Feb-04	05-Mar-04	12-Mar-04	19-Mar-04	26-Mar-04
134	Sum of CurrentBLHours	14.84	18.41	12.93	8.55	6.54	10.14	24.70	1.90	17.16
	Sum of EarnedHours	19.79	17.14	4.90	3.37	12.19	5.99	35.54	2.94	25.91
	Sum of ActualHours	20.00	11.25	9.75	8.25	17.25	12.00	24.75	16.00	16.00
135	Sum of CurrentBLHours	21.13	37.85	24.73	6.23	18.90	11.10	17.38	22.89	28.10
	Sum of EarnedHours	34.17	55.42	40.75	9.04	25.40	8.91	16.78	26.16	11.43
	Sum of ActualHours	25.25	43.00	42.00	47.00	42.00	39.75	26.25	14.00	26.00
136	Sum of CurrentBLHours	8.44	10.19	16.97	2.47	13.23	7.69	2.80	22.43	67.33
	Sum of EarnedHours	6.30	12.70	30.21	3.94	17.90	9.19	4.69	31.66	17.05
	Sum of ActualHours	44.75	36.25	30.00	28.00	10.00	11.25	9.75	41.00	44.50
137	Sum of CurrentBLHours									
	Sum of EarnedHours									
	Sum of ActualHours									
138	Sum of CurrentBLHours	132.42	15.62	28.00	3.07	103.96	65.44	46.71	49.08	20.04
	Sum of EarnedHours	132.93	28.69	30.55	1.34	118.67	63.64	7.03	25.66	8.08
	Sum of ActualHours	90.75	47.50	33.00	2.00	67.00	67.00	54.00	60.25	46.00
139	Sum of CurrentBLHours									
	Sum of EarnedHours									
	Sum of ActualHours									
140	Sum of CurrentBLHours									
	Sum of EarnedHours									
	Sum of ActualHours									
141	Sum of CurrentBLHours		2.05			61.61	37.66	36.75	71.07	23.32
	Sum of EarnedHours		1.72			8.63	8.38	57.02	36.40	4.75
	Sum of ActualHours		6.75			33.75	5.00	30.25	20.00	11.00
142	Sum of CurrentBLHours	1.18	0.34	191.98	49.07	38.88	0.13			4.69
	Sum of EarnedHours	1.55	0.59	21.52	32.83	9.34	0.15			6.93
	Sum of ActualHours	2.25	1.50	11.50	32.75	11.00	0.25			5.00
143	Sum of CurrentBLHours									
	Sum of EarnedHours									
	Sum of ActualHours									
144	Sum of CurrentBLHours		36.68	37.04	67.23	92.82	40.66	34.79	54.90	26.36
	Sum of EarnedHours		42.60	30.99	47.03	35.93	33.05	48.95	53.85	40.29
	Sum of ActualHours		39.00	35.25	32.25	54.50	32.00	37.50	38.50	40.00
145	Sum of CurrentBLHours		42.44	3.21	50.03	25.72	14.64	22.51	41.94	375.49
	Sum of EarnedHours		50.08	4.73	14.70	32.73	22.48	41.10	35.70	47.02
	Sum of ActualHours		29.00	8.00	38.25	20.00	19.00	28.75	27.00	55.00
146	Sum of CurrentBLHours			6.18	7.31	29.11	58.77	25.76	2.51	12.14
	Sum of EarnedHours			8.42	10.40	34.28	40.08	17.81	2.15	11.39
	Sum of ActualHours			13.25	26.75	36.25	31.75	38.00	1.50	24.25

Figure 5.28: Weekly Resources Utilization per Internal Project

PortfolioID	(All)										
ProgramID	(All)										
PhaseID	(All)										
ResourceID	(All)										
ProjectID	Data	PeriodNumber	1	2	3	4	5	6	7	8	9
24	Sum of CurrentBLHours		256.76	418.92	1,017.29	486.41	994.97	395.43	1,907.85	519.40	471.06
	Sum of EarnedHours		234.45	349.85	405.70	323.38	474.29	257.87	465.15	262.21	416.41
	Sum of ActualHours		218.00	348.75	374.00	322.75	430.00	307.25	385.25	298.00	342.75
39	Sum of CurrentBLHours		287.95	179.22	77.27	164.27	218.58	160.12	207.04	288.65	270.63
	Sum of EarnedHours		73.16	135.45	104.92	104.00	128.97	169.17	100.14	217.31	196.69
	Sum of ActualHours		76.75	145.50	197.00	183.50	180.75	125.00	110.50	156.75	192.50
70	Sum of CurrentBLHours		16.65	43.44	30.94	24.95	35.10	133.50	1.68	8.81	10.39
	Sum of EarnedHours		16.08	35.41	16.22	33.56	46.78	45.19	1.56	14.01	10.62
	Sum of ActualHours		16.00	26.50	50.75	47.50	30.75	31.00	2.75	16.75	7.75
71	Sum of CurrentBLHours		4.42	48.80	31.93	17.90	16.66	5.91	8.18	7.00	7.99
	Sum of EarnedHours		3.40	44.58	51.71	11.89	12.56	8.55	13.23	9.28	9.87
	Sum of ActualHours		6.00	35.00	27.50	20.50	11.75	12.00	15.75	8.00	6.00
72	Sum of CurrentBLHours		21.81	13.19	45.67	38.71	43.04	17.68	37.73	17.03	55.75
	Sum of EarnedHours		15.80	20.41	38.98	27.23	16.44	14.51	56.96	25.84	29.36
	Sum of ActualHours		19.75	20.00	21.75	25.50	20.75	25.50	67.50	34.25	28.25
118	Sum of CurrentBLHours		145.68	214.55	86.10	362.99	306.92	552.78	280.68	268.17	240.24
	Sum of EarnedHours		245.83	196.80	99.73	271.05	195.31	410.52	376.81	254.19	159.70
	Sum of ActualHours		176.50	252.75	209.50	256.75	288.75	306.00	275.50	271.75	255.50
134	Sum of CurrentBLHours		6.85	2.99	14.84	18.41	12.93	6.55	6.54	10.14	24.70
	Sum of EarnedHours		3.25	3.09	19.79	17.14	4.90	3.37	12.19	5.99	35.54
	Sum of ActualHours		5.00	2.00	20.00	11.25	9.75	8.25	17.25	12.00	24.75
135	Sum of CurrentBLHours		115.64	21.13	37.85	24.73	6.23	18.90	11.10	17.38	22.89
	Sum of EarnedHours		32.95	34.17	55.42	40.75	9.04	25.40	8.91	16.78	26.16
	Sum of ActualHours		25.75	25.25	43.00	42.00	47.00	42.00	39.75	26.25	14.00
136	Sum of CurrentBLHours		33.14	8.44	10.19	16.97	2.47	13.23	7.69	2.80	22.43
	Sum of EarnedHours		56.38	6.30	12.70	30.21	3.94	17.90	9.19	4.69	31.66
	Sum of ActualHours		31.00	44.75	36.25	30.00	28.00	10.00	11.25	9.75	41.00
137	Sum of CurrentBLHours		100.89	17.73	21.44	4.72	74.08	0.29	258.40	9.94	0.67
	Sum of EarnedHours		33.25	26.23	24.68	3.53	10.17	0.05	32.43	7.67	0.96
	Sum of ActualHours		19.75	16.50	15.00	4.25	7.00	0.25	34.00	4.75	7.75
138	Sum of CurrentBLHours		286.23	132.42	82.75	57.54	36.58	51.97	45.94	40.21	40.90
	Sum of EarnedHours		53.51	132.93	118.60	72.77	6.71	43.55	27.15	34.81	18.46
	Sum of ActualHours		73.50	90.75	98.50	83.25	52.25	66.75	61.75	31.00	27.25
139	Sum of CurrentBLHours		349.30	18.89	6.31	1.18	14.97	12.68	2.11		
	Sum of EarnedHours		46.78	22.64	1.87	1.40	8.22	4.68	1.79		
	Sum of ActualHours		36.75	18.00	4.00	1.50	5.00	19.00	2.00		
140	Sum of CurrentBLHours		29.91	5.70	10.84	2.83	12.08	2.98	33.92	35.21	51.43
	Sum of EarnedHours		17.15	3.15	14.00	4.53	5.91	0.64	23.08	13.90	21.36
	Sum of ActualHours		15.75	2.00	10.75	7.00	7.00	2.75	15.25	15.00	15.00
141	Sum of CurrentBLHours		2.05	61.61	37.56	36.75	71.07	23.32	2.49		
	Sum of EarnedHours		1.72	8.63	8.38	57.02	36.40	4.75	3.63		
	Sum of ActualHours		6.75	33.75	6.00	33.75	20.00	11.00	2.00		

Figure 5.29: Weekly Resource Utilization per Reporting Period

5.6 THE KNOWLEDGE EXCHANGE TOOL

The developed data warehouse is easily used in helping users exchange tacit and explicit knowledge elements. To achieve this objective, a simple knowledge definition screen is added to the data warehouse as shown in Figure 5.30. The knowledge definition screen allows the user to define knowledge elements by work package and resource.

By linking every knowledge element to a work package, the knowledge element inherits all the attributes of the parent work package. These attributes include the production package type, project phase, internal project, program or portfolio, contractor, industrial project, program or portfolio and industrial owner.

The process of linking knowledge elements to a specific work package allows users to search the knowledge base using any of the package predefined attributes. The existing knowledge portals, which are rarely used in the industry, rely mostly on searching by key words returning lots of irrelevant results to the users.

Knowledge elements are classified into groups and categories as well. Lessons learned, risks and issues are examples of these knowledge categories. The risk category includes two main groups; technical and managerial. The issues categories include several groups such as quality, safety, communications, material management, software, etc.

The screenshot shows a window titled "Knowledge Elements" with the following fields:

- Work Package:** 04E2665-04, Dry Surge Project - Detailed Engineering & Design Phase, 005, Foundations for Piperack # 001
- Resource:** Engineering Services, Civil - Structural - Engineering
- Knowledge Element:** Risks, Technical, Soil Condition
- Description:** Problem with Type A foundations due to the soil conditions
- Solution:** Use type B foundation for this soil conditions
- Related Files:** C:\Project Files\04E2665-04\Photos\Foundations

At the bottom, there is a record navigation bar showing "Record: 1 of 1" with navigation icons.

Figure 5.30: The Knowledge Definition Screen

To exchange explicit knowledge, the user can easily query the data warehouse to find the required knowledge elements using the dynamic knowledge finding screen as shown in Figure 5.31. The dynamic knowledge finding screen allows the user to filter the knowledge base by any of the control attributes and create any necessary combinations.

To exchange tacit knowledge, the user can easily query the data warehouse to find the individuals who worked on the areas related to the required knowledge using the dynamic personnel finding screen as shown in Figure 5.32. The dynamic personnel finding screen enables the users to find individuals who have worked on certain production package types, project phases, internal projects, programs or portfolios, contractors, industrial projects, programs or portfolios. All the fields on the form are showing the hierarchical nature of the attributes and allow the user to select values at any level of the hierarchy. This means the user can select all internal

projects, a specific internal project, a program or a portfolio of internal projects. This provides the user with maximum flexibility to narrow down their search and find only the targeted answers.

Exchange Tacit Knowledge

Find individuals who worked on:

Industrial Owner: All Owners

Industrial Project: All Industrial Projects

Components: All Components

Industry: Ore Preparation Plants, Oil Sands, Hydrocarbons

PPF: Edmonton Office, Colt WorleyParsons, WorleyParsons Ltd.

Internal Project: All Internal Projects

Project Phase: Industrial Construction Projects, Engineering Phases, Detailed Engineering & Design

Production Pckge: EPCM Packages, Civil Work, FOUNDATIONS

Resource: Engineering Services, Civil - Structural - Engineering

Between Dates:
From: 01-Jan-00 To: 31-Dec-08

Find Individuals

Figure 5.31: The Screen for Exchanging Tacit Knowledge

Exchange Explicit Knowledge

Find the Knowledge You Need in:

Knowledge Item: Risks, Technical, Soil Condition

Industrial Owner: All Owners

Industrial Project: All Industrial Projects

Components: All Components

Industry: Ore Preparation Plants, Oil Sands, Hydrocarbons

PPF: Edmonton Office, Colt WorleyParsons, WorleyParsons Ltd.

Internal Project: All Internal Projects

Project Phase: Industrial Construction Projects, Engineering Phases, Detailed Engineering & Design

Production Pckge: EPCM Packages, Civil Work, FOUNDATIONS

Resource: Engineering Services, Civil - Structural - Engineering

Between Dates:
From: 01-Jan-00 To: 31-Dec-08

Find Individuals

Figure 5.32: The Screen for Exchanging Explicit Knowledge

An example of the output for exchanging explicit knowledge is shown in Figure 5.33. This seemingly simple procedure saves tremendous amounts of time, effort and frustration that stems from trying to locate knowledge elements or individuals who possess a specific needed knowledge. Once enough knowledge elements are obtained from projects, a generic library can be built and shared with all project management teams prior to starting any work package. This practice is expected to minimize repeating the same mistakes to a significant extent.

Knowledge Elements Report

<i>Work Package</i>	<i>Knowledge Element</i>	<i>Solution</i>	<i>Hyperlink</i>
Risks, Technical, Soil Condition			
<i>Engineering Services, Civil - Structural - Engineering</i>			
04E2665-04, Dry Surge Project - Detailed Engineering & Design Phase, 005, Foundations for Piperack # 001	Soil may be contaminated	Do enviromental check	
04E2663-04, South Booster Pump House (SBPH) Project - Detailed Engineering & Design Phase, 005, Foundations for SBPH Scope of Work	Rocks Might be found on site	Do more soil investigation	
04E2662-04, MFT Transfer Pond 6 To Pond 7 Project - Detailed Engineering & Design Phase, 005, Foundations for MFT Scope of Work	Problem with Type A foundations due to the soil conditions	Use type B foundation for this soil conditions	C:\Project Files\04E2665-04\Photos\Foundations

Tuesday, February 16, 2009 *Page 1 of 1*

Figure 5.33: The Output of Finding Knowledge Elements

5.7 SYSTEM IMPLEMENTATION

In order to implement the proposed framework, several steps have to be taken in three main streams. The first stream is to build the data warehouse according to the proposed design. This task would require the development of both input and output interfaces that can dynamically interact with various end-users according to their specific needs. After that, the data warehouse needs to be populated with the existing historical data that matches the predefined data elements in the system. This procedure would require a large amount of cleaning, validating and preprocessing because the existing data was not generated nor collected for data mining purposes. The company would need to perform a cost-benefit analysis to decide whether the obtained knowledge from existing data worth the required efforts or it is more beneficial to start collecting data from new projects.

The second main stream is modifying the existing project controls systems in the company to be able to export the weekly data directly to the data warehouse. These modifications would apply to the project initiation, scheduling, accounting, human resources, timekeeping and progress measurement systems. These modifications accompanied with the direct export would save the time, effort and possible errors in populating the data warehouse.

The third and most important stream before the company can start mining the knowledge from the data warehouse is the organization culture change. Organization culture forms in any company and it reflects the values, norms,

attitudes, beliefs and comfort zones of the current management team. Managers typically hire people who share similar values to themselves. This practice strengthens and sustains existing organization culture. Labour resources are usually comfortable with existing organization culture and are not willing to accept change. The suggested approach requires a culture of knowledge sharing not hiding and learning from the past through accurate recording of actual data, decisions taken and the reasoning behind these decisions.

In many contracting companies, staff members are not familiar with the concepts of knowledge sharing and data mining. There are many issues around knowledge sharing such as ownership of generated knowledge, pride of individuals and fear of getting fired after recording the knowledge. Staff members are used to record their time only to get paid and they don't pay enough attention to allocate the charged hours accurately to the correct work packages and to distinguish between productive vs. non-productive time. The successful organization culture change starts with first assessing the existing culture, plan the change, and monitoring the progress during implementation. It requires full support of executive management who has to lead by example to encourage everyone else to follow. The system implementation by itself would help with changing the organization culture because knowing that the data would be mined would encourage individuals to pay more attention to recording data properly. Once they start to receive the benefits of the shared knowledge, they would be more willing to share their own knowledge.

CHAPTER 6: CASE STUDIES ON KNOWLEDGE DISCOVERY IN DATA

6.1 DISCOVERING KNOWLEDGE IN THE FIRST DATASET

6.1.1 Data Cleaning and Preprocessing

The purpose of this case study is to validate that data mining can be used to improve and increase the efficiency of labour estimating practices in contracting companies. Most of these companies rely on cost estimating units (norms) that are not based on historical data and are not updated to reflect changes in the industry. Applying the proposed approach that relies on data mining is expected to provide companies with knowledge-based probabilistic dynamic estimating units that always reflect the latest changes.

The first dataset contains data regarding the scope of a set of engineering work packages. This scope is represented as determinate amounts of key quantities per work package. The key quantity for engineering packages is the number of engineering deliverables. The data set was obtained from the estimating system of this contractor. This estimating system is based on an old version of MS Access. The dataset contains the original and current baseline hours, for five of the involved resource in this group of work packages. The current baseline values reflect the project scope after implementing all approved changes. The selected data set to be analyzed in this case study contained data for more than one hundred projects, four project phases and five different resources.

The contractor did not track actual spent hours per work package, however the same analysis can be easily applied if the data exists. The analysis was used to check the consistency of the estimating practices in this contracting company. The data was directly exported from the estimating system to MS Excel, where the cleaning and preprocessing took place prior to exporting the data to the data to the warehouse. Table 6.1 shows an example of the raw data. Not only data was missing, but also metadata (data about the data) was also missing. The data lacked the values for two important control attributes: the internal program and the project phase and had to be assigned manually.

This manual procedure required going back to the archived project documents to find the appropriate values to be assigned to each data point. The procedure consumed a lot of time and effort until the dataset was completed and verified.

Project No: 06E1150 - Firewater Main Replacement - Phase 1 - Execute				Original Budget			9592			Current Budget			9697		
Discipline	L1Code	TOLCode	TOLDescription	ENGR			DESIGN & CAD			ENGR			DESIGN & CAD		
				QTY	HRS	UNIT HRS	QTY	HRS	UNIT HRS	QTY	HRS	UNIT HRS	QTY	HRS	UNIT HRS
PROCESS	1130	90.00	PROCESS OVERALL			0					0				0
MECHANICAL	1130	01.000	PROJECT INITIATION	1	2				1	2					
MECHANICAL	1130	02.000	DESIGN COORDINATION	1	194				1	194					
MECHANICAL	1130	03.000	ESTIMATING	1	20				1	20					
MECHANICAL	1130	04.000	STUDIES												
MECHANICAL	1130	05.100	SYSTEM DEFINITION / DESIGN CRITERIA												
MECHANICAL	1130	05.200	REPORTS & WRITE-UPS												
MECHANICAL	1130	05.300	EQUIPMENT & MAT'L SPECS / DG	1	442				1	442					
MECHANICAL	1130	05.400	CONTRACTS (CMPs / EMPs)												
MECHANICAL	1130	05.800	OTHER (INCL. CUSTOM LINES)												
MECHANICAL	1130	06.200	CALCULATIONS & SIZING												
MECHANICAL	1130	06.300	STRESS ANALYSIS												
MECHANICAL	1130	06.800	OTHER (INCL. CUSTOM LINES)												
MECHANICAL	1130	07.100	PFDs												
MECHANICAL	1130	07.200	P&IDs												
MECHANICAL	1130	07.320	PLOT PLANS	10	50	5			10	50	5				
MECHANICAL	1130	07.330	GRG / ORTHOS / DETAILS												
MECHANICAL	1130	07.399	MISC. DRAWINGS / SKETCHES												
MECHANICAL	1130	07.400	SCHEDULES / LISTS / INDICES / SPREADSHEETS												
MECHANICAL	1130	07.900	OTHER (INCL. CUSTOM LINES)	1	195				1	195					
MECHANICAL	1130	08.100	MODEL DEVELOPMENT												
MECHANICAL	1130	08.200	"SMART" P&ID / "NITOOLES"												
MECHANICAL	1130	08.300	DRAWINGS PRODUCTION												
MECHANICAL	1130	08.400	SCHEDULES / LISTS / INDEXES / SPREADSHEETS												
MECHANICAL	1130	08.500	MODEL ADMINISTRATION												
MECHANICAL	1130	08.800	OTHER (INCL. CUSTOM LINES)	5	25	5			5	25	5				
MECHANICAL	1130	09.100	REGULATORY / PERMITTING												
MECHANICAL	1130	09.200	MANUALS / OTHER DOCUMENTATION	1	2				1	2					
MECHANICAL	1130	09.300	AS-BUILT DRAWINGS	1	60				1	60					
MECHANICAL	1130	09.900	OTHER (INCL. CUSTOM LINES)	1	4				1	4					
MECHANICAL	1130	10.100	INTERDISCIPLINE CHECKING	1	40				1	40					
MECHANICAL	1130	10.200	QA / QC AND FIELD CHECKING	1	46				1	46					
MECHANICAL	1130	10.900	OTHER (INCL. CUSTOM LINES)	90	90	1			90	90	1				
MECHANICAL	1130	11.000	PROCUREMENT SUPPORT	1	46				1	46					
MECHANICAL	1130	12.000	CONSTRUCTION SUPPORT	1	100				1	100					
MECHANICAL	1130	13.000	PROJECT CLOSEOUT	1	21				1	21					
MECHANICAL	1130	90.00	MECHANICAL OVERALL			1283					1340				
PIPING	1132	01.000	PROJECT INITIATION				1	2				1	2		
PIPING	1132	02.000	DESIGN COORDINATION				1	116			1	40		1	116
PIPING	1132	03.000	ESTIMATING				1	10				1	10		
PIPING	1132	04.000	STUDIES				1	20				1	20		
PIPING	1132	05.100	SYSTEM DEFINITION / DESIGN CRITERIA												
PIPING	1132	05.200	REPORTS & WRITE-UPS												
PIPING	1132	05.300	EQUIPMENT & MAT'L SPECS / DG												
PIPING	1132	05.400	CONTRACTS (CMPs / EMPs)												
PIPING	1132	05.900	OTHER (INCL. CUSTOM LINES)												
PIPING	1132	06.200	CALCULATIONS & SIZING	1	200						1	200			
PIPING	1132	06.300	STRESS ANALYSIS												

Table 6.1: The Raw Data for the First Data Set

Furthermore, while working with the archived documents, some documents were not clear enough and assumptions have to be made to compensate for the missing data. Many of the staff members who were involved in these projects could not be located. And even if they could be contacted, they could not provide meaningful input on the data as time has passed. All the effort and time spent searching, sorting, and cleaning in the archive would have been easily avoided if the data and its metadata were collected in the proposed integrated format.

Table 6.2 shows the data from Table 6.1 after it was cleaned, pre-processed and is ready for storage in the data warehouse. The objective of this analysis is to test if the resource unit cost per production package type could be extracted for future estimating of resource requirements in upcoming projects. For this analysis, fifteen standard production packages, three engineering phases and five engineering resources were selected. The data was modified by random numbers for confidentiality issues. This modified data was used in the analysis. Hence, all numbers shown here are not actual ones and are used only for illustration.

Program	Project	Package	Phase	Resource	OriginalUnitCost	CurrentUnitCost
2	34	13	9	5	0.50	0.50
2	43	13	10	2	2.00	2.00
2	47	13	10	2	2.00	2.00
2	67	13	9	1	4.00	
2	118	13	9	1	4.00	4.00
2	39	13	10	1	4.00	4.00
2	69	13	9	2	5.00	
1	9	13	9	1	6.00	
1	13	13	9	1	6.00	
1	11	13	9	1	10.00	
1	27	13	9	1	10.00	
1	3	13	10	1	10.00	
1	6	13	10	1	10.00	
1	24	13	10	1	10.00	
2	36	13	9	6	15.00	15.00
1	18	13	9	1	20.00	
1	21	13	9	1	20.00	
2	36	13	9	1	20.00	20.00
1	17	13	8	1	25.00	
1	20	13	8	1	25.00	
2	63	13	9	1	40.00	40.00
1	26	13	8	1	330.00	
2	43	14	10	2	1.00	1.00
2	69	14	9	5	1.00	
2	34	14	9	6	1.00	1.00
2	67	14	9	6	1.00	
2	32	14	10	5	1.50	1.50
1	11	14	9	1	2.00	
2	120	14	9	1	2.00	2.00
2	47	14	10	1	2.00	2.00
1	11	14	9	2	2.00	
2	118	14	9	2	2.00	2.00
1	22	14	10	2	2.00	
2	39	14	10	2	2.00	2.00
2	45	14	9	5	2.00	2.00
2	67	14	9	5	2.00	

Table 6.2: The Dataset after Cleaning and Pre-Processing

Three control attributes are selected for this analysis. These control attributes are represented in this analysis with the independent variables: $\text{Package}_{(1:15)}$, $\text{Phase}_{(1:3)}$ and $\text{Resource}_{(1:5)}$. These are nominal variables with values assigned to them as discrete integers. These discrete integers are equal to the ID's used in the data warehouse for direct referencing. To test the significance of adding more attributes to the analysis, the internal program control attribute is selected. This attribute is represented in the analysis with the independent variable $\text{Program}_{(1:3)}$.

In this study, the term 'class' refers to a unique combination of values of the three variables: Package, Phase and Resource. For example, class 1 contains all the data points that have the value $\text{Pk}_{(1)}$ for the variable Package, $\text{Ph}_{(1)}$ for the variable Phase

and $R_{(1)}$ for the variable Resource. A class 2 contains all the data points have the value $Pk_{(1)}$ for the variable Package, $Ph_{(1)}$ for the variable Phase and $R_{(2)}$ for the variable Resource, etc. The number of classes resulting from all the possible combinations is calculated using the formula:

$$\begin{aligned} \text{Number of Classes} &= \text{Number of Packages} * \text{Number of Phases} * \text{Number of} \\ &\text{Resources} \\ &= 15 * 3 * 5 = 225 \quad \text{Classes} \quad [6.1] \end{aligned}$$

It is important to note that the dataset may not include data points for all the classes. Certain classes of the three main attributes do not exist in reality. For examples, some packages are not needed in every phase or some packages do not utilize all the five resources under investigation.

The key quantity for all the packages is the number of engineering deliverables. This analysis is implemented to the hourly portion of the collected data; since estimating of labour resource requirements rely on hourly units and not on cost. The dataset was normalized to eliminate the differences in project sizes by calculating three dependent variables: “Original Hourly Unit Cost”, “Current Hourly Unit Cost,” and “Actual Hourly Unit Cost.” These variables are calculated using the following formulas:

$$\text{Original hourly unit cost} = \text{original baseline hours} / \text{original quantity} \quad [6.2]$$

Current hourly unit cost = current baseline hours / current quantity [6.3]

Actual hourly unit cost = actual hours / actual quantity [6.4]

Because of the multidimensionality of this dataset, four new variables were formulated to represent the possible combinations of the three main attributes. These new variables are Package/Phase, Package/Resource, Phase/Resource and Package/Phase/Resource. These variables are assigned unique values by combining ID's from the three main attributes.

After defining all the necessary variables, the dataset was then exported to the first analysis tool, SPSS-16 for Windows. SPSS was selected because of its ability to perform a wide range of statistical analysis tests, its ability to easily import and export data from databases and its user friendliness. It can be easily obtained by any contractor or industrial owner who needs to perform statistical analysis of the collected data in the data warehouse.

The objective of this analysis is to develop an estimating methodology that can be implemented using unit costs and key quantities. First, the dataset is divided into clusters using stratification. Significant differences of means are used to establish these clusters. Within each cluster, unit cost and the characteristics of the most fitting distribution are obtained. Therefore, instead of relying solely on their intuitions, the estimators are presented with mined values for the unit costs that can

be multiplied by the known determinate key quantities in order for these estimators to predict the resources requirements more accurately.

6.1.2 The Initial Investigation

Data mining models suggest starting any exercise with visual presentation of the available dataset. First, the frequency of data points within each independent variable is plotted. Figure 6.1 graphically shows that phase Ph-03 has more data points than the other two phases. Figure 6.2 shows that not every package utilizes the five resources and that some packages only utilize single resource.

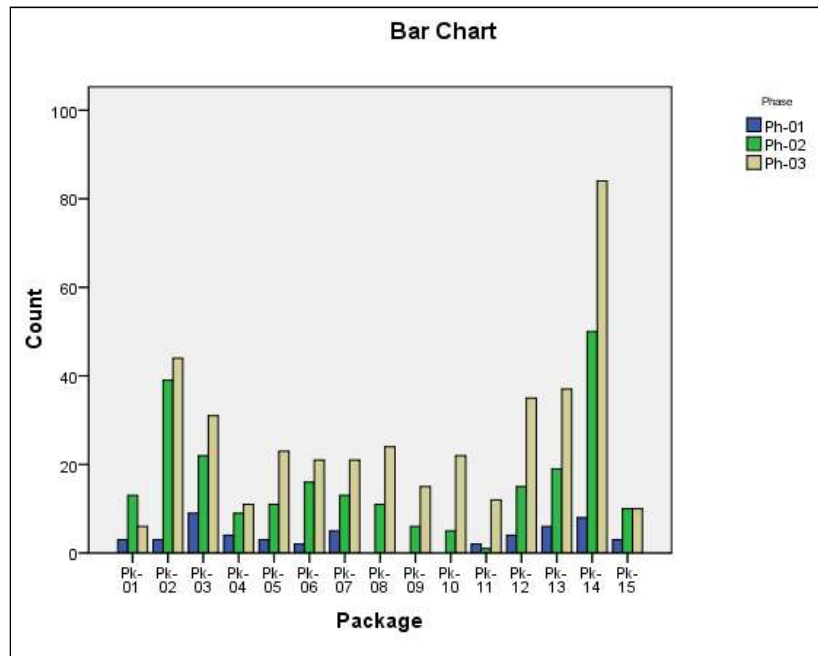


Figure 6.1: Frequency of Data Points within the Three Phases

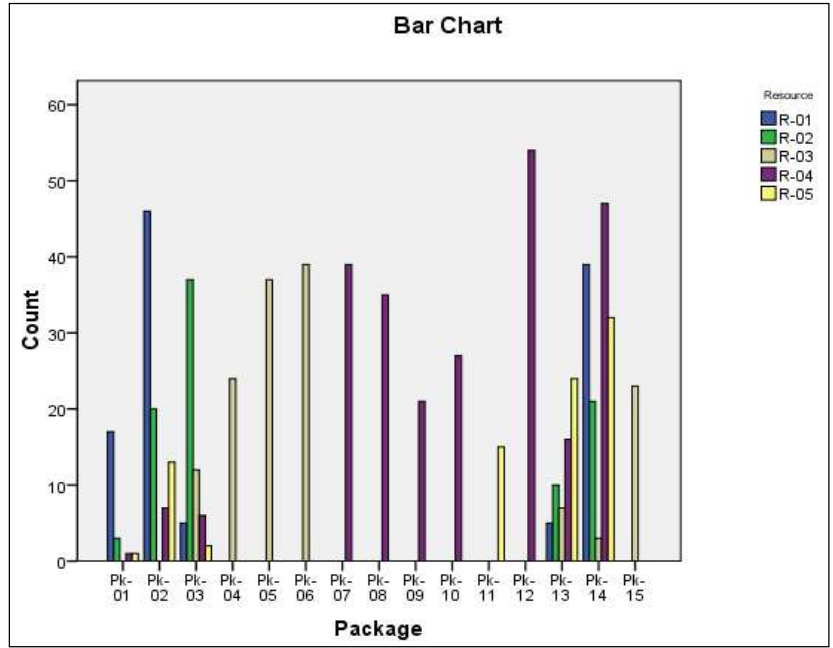


Figure 6.2: Frequency of Data Points within the Five Resources

Second, the data descriptive “case summaries” test is performed to collect statistics on each class or data subset. Since the data is multidimensional, subsets can be generated using one attribute, combination of any two attribute, or the combined all together three attributes. The following statistics are obtained: mean, standard deviation, number of data points, minimum value, maximum value and data range.

Table 6.3 shows an excerpt of the test results.

Case Summaries								
Pa...	Ph...	Re...	Mean	Std. Deviation	N	Minimum	Maximum	Range
Pk-01	Ph-01	R-01	126.6667	176.09183	3	25.00	330.00	305.00
		Total	126.6667	176.09183	3	25.00	330.00	305.00
	Ph-02	R-01	14.0000	11.27436	10	4.00	40.00	36.00
		R-02	5.0000	.	1	5.00	5.00	.00
		R-04	.5000	.	1	.50	.50	.00
		R-05	15.0000	.	1	15.00	15.00	.00
		Total	12.3462	10.69537	13	.50	40.00	39.50
	Ph-03	R-01	8.5000	3.00000	4	4.00	10.00	6.00
		R-02	2.0000	.00000	2	2.00	2.00	.00
		Total	6.3333	4.08248	6	2.00	10.00	8.00
	Total	R-01	32.5882	77.26420	17	4.00	330.00	326.00
		R-02	3.0000	1.73205	3	2.00	5.00	3.00
		R-04	.5000	.	1	.50	.50	.00
		R-05	15.0000	.	1	15.00	15.00	.00
		Total	26.2955	68.52748	22	.50	330.00	329.50
Pk-02	Ph-01	R-01	16.6667	5.77350	3	10.00	20.00	10.00
		Total	16.6667	5.77350	3	10.00	20.00	10.00
	Ph-02	R-01	50.3039	106.83125	18	2.00	465.00	463.00
		R-02	9.7000	8.52513	10	2.00	25.00	23.00
		R-04	2.8000	1.64317	5	1.00	5.00	4.00
		R-05	9.3333	12.75408	6	1.00	33.00	32.00
		Total	27.4992	74.87835	39	1.00	465.00	464.00

Table 6.3: The Descriptive Data Test

Subsequent to that, statistical dispersion is measured using Boxplots that are obtained for each of the data subsets. Boxplots show the Inter Quartile Range - IQR (the 25th percentile, the median, 75th percentile) minimum, maximum and extreme values. SPSS points out to the raw-number that contains data points that are out of the normal range.

The descriptive statistics as well as the boxplots show very wide ranges and variances (Figures 6.3 and 6.4). They also show that the dataset contains extreme outliers. As a result of this situation, it becomes necessary to implement an outlier detection procedure.

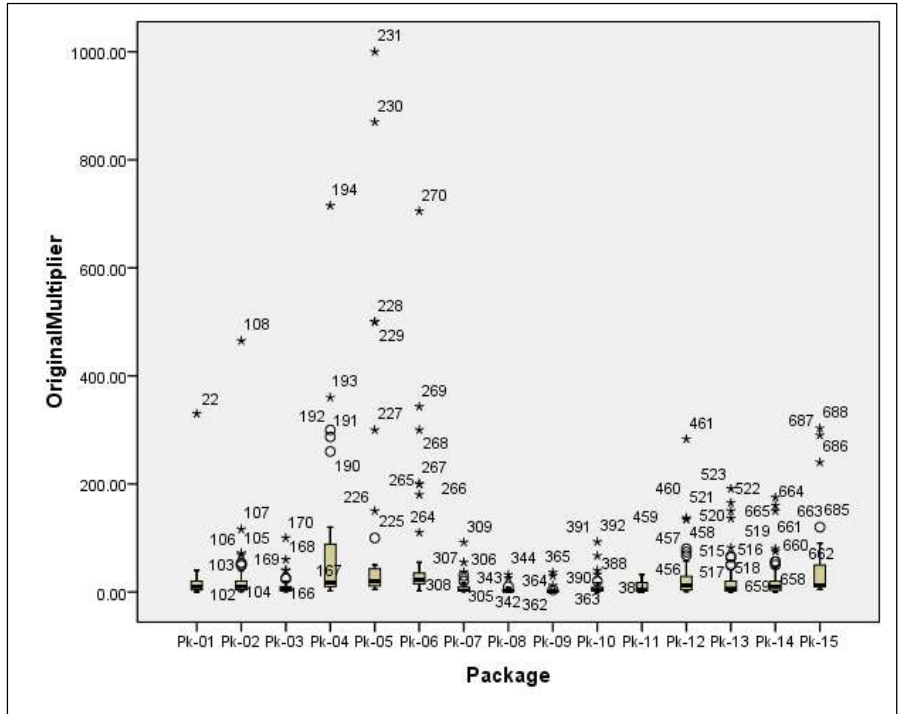


Figure 6.3: Boxplot of the Unit Cost Showing Outliers per Package

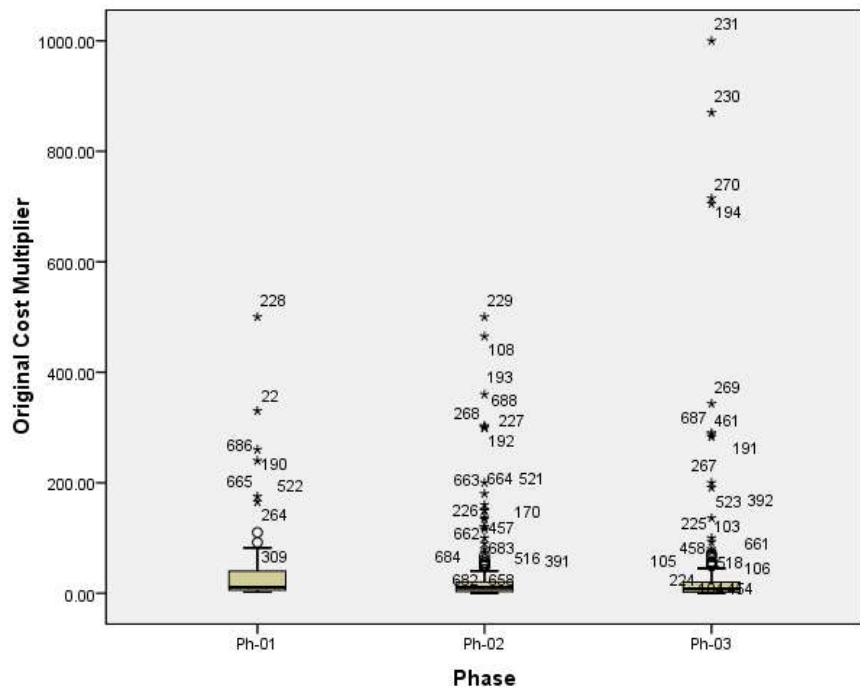


Figure 6.4: Boxplot of the Unit Cost Showing Outliers per Phase

6.1.3 The Outliers Detection Procedure

Given that the bloxplot results showing outliers in the dataset, detecting these becomes necessary. In this research, the technique to be implemented is based on Chebyshev Theorem (Zaiane, 2006). This theorem can be used for single dimension (univariate) outliers analysis. Assuming the dataset follows a normal distribution, the mean and standard deviation of the distribution can be defined by calculating the mean (μ) and standard deviation (σ) of the dataset. Chebyshev stated that since most data points falls between ($\mu + 3\sigma$) and ($\mu - 3\sigma$), those that fall outside of this range can be considered outliers.

A four layer outlier analysis tool is developed based on the three-dimensional dataset.

- First layer = all data
- Second layer = each attribute
- Third layer = three possible combination of paired attributes represented as three new category variable. (Package*phase provides 45 combinations, package*resource provides 75 combinations and phase*resource provides 15 combinations).
- Fourth layer = all attributes combined (provides 225 combinations) represented as new category variable.

Total of eight possible cases of outliers are calculated using the obtained means and standard deviations obtained from SPSS. Each data point was tested against the

eight cases and was assigned a value of 1 if found to be outlier in any case. A total outlier score is calculated by adding the number of cases where a data point was an outlier. An example of the output is shown in Table 6.4. It is up to the user to go back and verify the outliers or eliminate them and perform the analysis. The procedure was repeated three times until the obtained standard deviations and ranges were found to be acceptable as shown in Figures 6.5 and 6.6.

Package PhaseRe source	All	Package	Phase	Resource	Package / Phase	Package / Resource	Phase / Resource	Package / Phase / Resource	Score	Outlier
1381	1	1		1		1	1		5	1
1491	1	1	1	1	1	1	1	1	8	1
2014	1								1	1
2094	1		1						2	1
2094	1		1						2	1
2014	1	1	1	1		1	1		6	1
2194	1		1						2	1
2184	1		1				1		3	1
2194	1		1				1		3	1
2114	1	1	1	1		1	1		6	1
2114	1	1	1	1	1	1	1	1	8	1
2294	1		1						2	1
2214	1		1						2	1
2214	1	1	1	1	1	1	1	1	8	1
2415		1			1	1		1	4	1
2515					1			1	2	1
2615		1		1	1	1	1	1	6	1
2915	1	1		1	1	1	1	1	7	1
3015						1		1	2	1
3116					1			1	2	1
3191		1			1	1		1	4	1
3214	1								1	1
3294	1								2	1
	17	20	13	16	16	18	19	10		23

Table 6.4: The Output from the Outlier Detection Tool

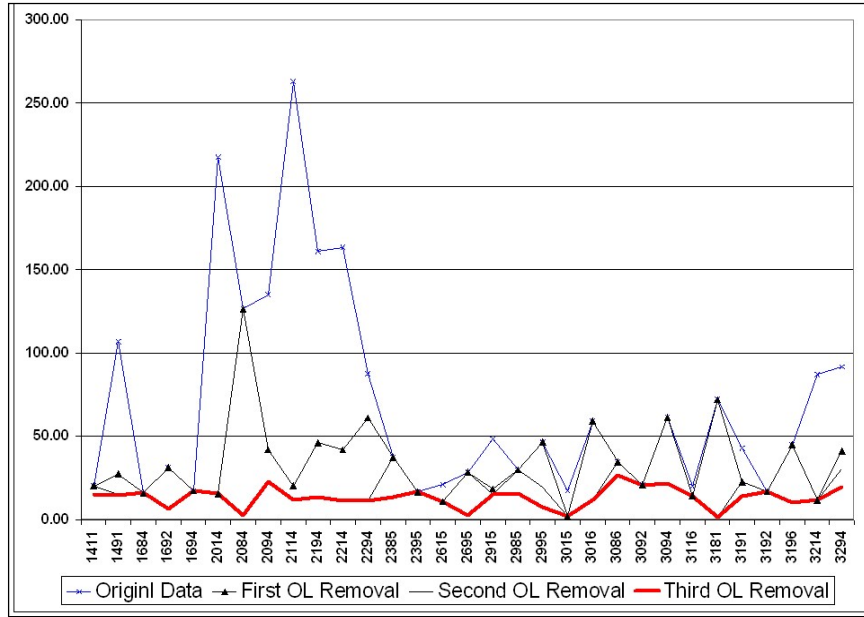


Figure 6.5: The Decrease in Standard Deviations of the Data Classes

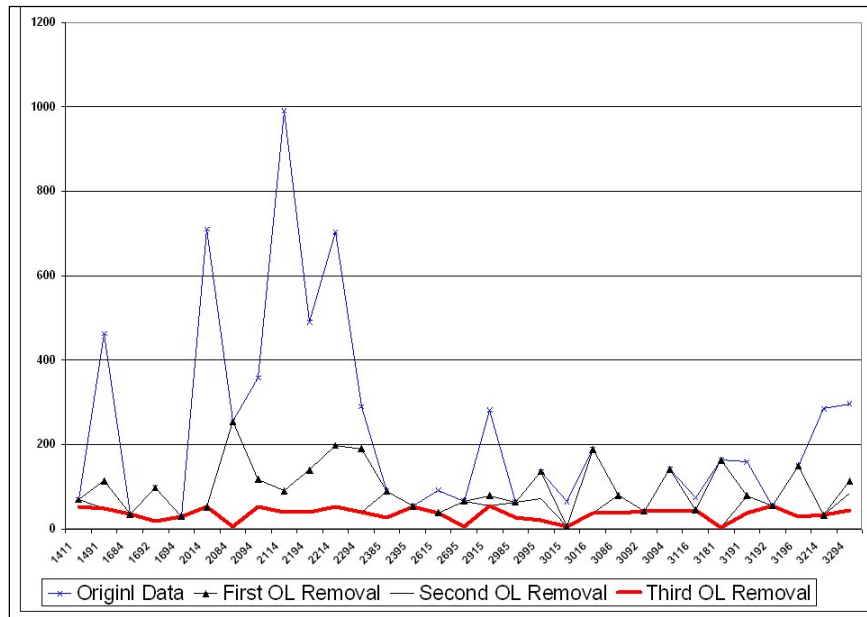


Figure 6.6: The Decrease in Ranges of the Data Classes

Cases with less than three data points were eliminated from the analysis. The mean and standard deviation of every class is calculated and summarized, as shown in Figure 6.7, graphically on a tree. The user can now use the summary tree to find out the unit cost multiplier distributions to be used for estimating new projects in the future. For each layer, a new variable Select (K) is assigned to each data point, where k = layer number.

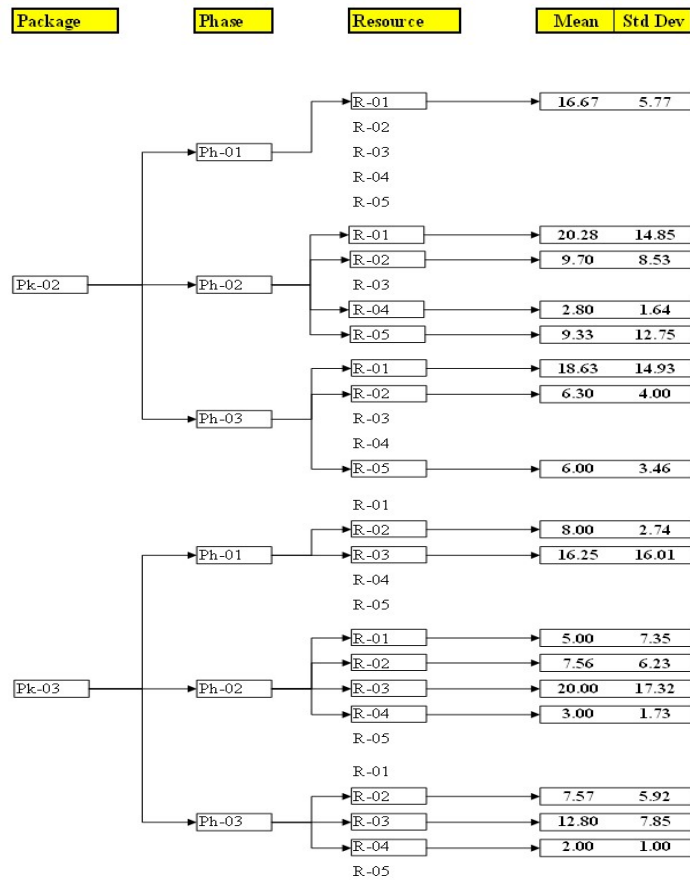


Figure 6.7: The Output Summary Tree

Instead of using the mean and standard deviation of the normal distribution, the user can also use fitting-distribution software such as @Risk to find the most fitting distribution for the data in a class. Figure 6.8 shows an example of finding the most fitting distribution for one of the classes.

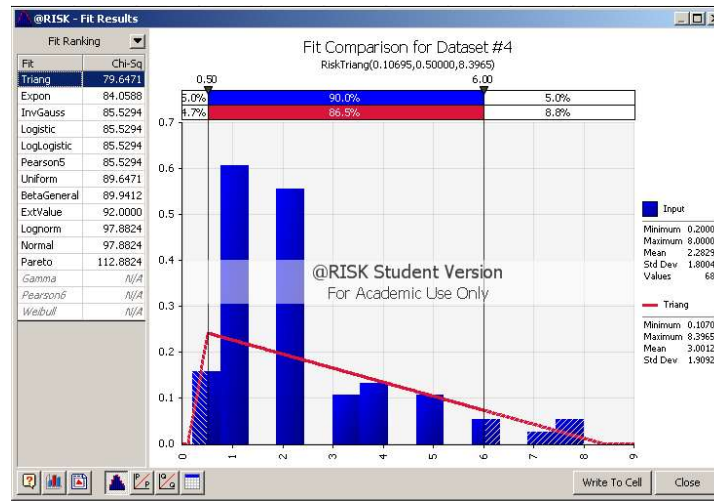


Figure 6.8: Fitting Distribution to a Class of Data

6.1.4 Clustering of Unit Cost using Statistical Methods

Building the unit cost tree shows large number of classes, which can drastically increase if more variable are added to the dataset. To simplify the estimating procedure, classes that are not significantly different from each other are combined together in summary groups (clusters) with one distribution representing each cluster. The ANOVA test was implemented to the dataset to check the significance of mean differences within the seven data attributes and the results are shown in Table 6.5.

Tests of Between-Subjects Effects

Dependent Variable: OriginalMultiplier

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	683820.684 ^a	79	8655.958	1.309	.045
Intercept	200551.553	1	200551.553	30.336	.000
Package	106060.040	14	7575.717	1.146	.314
Phase	9222.261	2	4611.130	.698	.498
Resource	6486.429	4	1621.607	.245	.913
Package * Phase	54561.285	24	2273.387	.344	.999
Package * Resource	14832.718	14	1059.480	.160	1.000
Phase * Resource	3390.225	7	484.318	.073	.999
Package * Phase * Resource	4903.522	13	377.194	.057	1.000
Error	4019430.959	608	6610.906		
Total	5278751.469	688			
Corrected Total	4703251.644	687			

a. R Squared = .145 (Adjusted R Squared = .034)

Table 6.5: Univariate ANOVA Test Results for the Three Main Attributes

The results for the Post Hoc tests for the three main attributes with $\alpha = 0.05$ are shown in Table 6.6.

		OriginalMultiplier				
Package	N	Subset				
		1	2	3	4	
Tukey B ^a	PK-08	35	4.2642			
	PK-09	21	4.9286			
	PK-07	39	9.7103			
	PK-11	15	10.6167			
	PK-03	62	10.7258			
	PK-10	27	12.4575			
	PK-14	142	16.8041			
	PK-02	86	20.9814			
	PK-13	62	22.3144			
	PK-01	22	26.2955			
	PK-12	54	27.0093			
	PK-15	23	59.5676	59.5676		
	PK-06	39	69.6022	69.6022		
	PK-04	24		99.1034		
PK-05	37		109.4454			
Duncan ^b	PK-08	35	4.2642			
	PK-09	21	4.9286			
	PK-07	39	9.7103			
	PK-11	15	10.6167			
	PK-03	62	10.7258			
	PK-10	27	12.4575			
	PK-14	142	16.8041	16.8041		
	PK-02	86	20.9814	20.9814		
	PK-13	62	22.3144	22.3144		
	PK-01	22	26.2955	26.2955		
	PK-12	54	27.0093	27.0093		
	PK-15	23		59.5676	59.5676	
	PK-06	39		69.6022	69.6022	
	PK-04	24		99.1034	99.1034	
PK-05	37			109.4454		
Sig.			.361	.062	.063	.061

	Phase	N	Subset	
			1	2
Tukey B ^a	Ph-03	396	25.0976	
	Ph-02	240	30.6041	30.6041
	Ph-01	52		50.6154
Duncan ^a	Ph-03	396	25.0976	
	Ph-02	240	30.6041	30.6041
	Ph-01	52		50.6154
	Sig.		.607	.062

	Resource	N	Subset	
			1	2
Tukey B ^a	R-04	253	12.0119	
	R-02	91	12.2527	
	R-05	87	18.1609	
	R-01	112	26.3850	
	R-03	145		77.4241
Duncan ^a	R-04	253	12.0119	
	R-02	91	12.2527	
	R-05	87	18.1609	
	R-01	112	26.3850	
	R-03	145		77.4241
	Sig.		.220	1.000

Table 6.6: Post Hoc Test Results for the Three Main Attributes

If the user decides to use only one attribute for dividing the dataset, test results in Table 6.6 show that packages can be grouped into four classes; phases can be grouped into two classes; and resources can be grouped into two classes.

If the user decides to use the combination of the three main attributes (Package*Phase*Resource), Table 6.7 shows the Post Hoc test results for this combination. The test results are used to group the classes into eight clusters and a new variable Cluster_(1:8) is assigned to each data point. The case summary and BoxPlot tests were repeated and the results are shown in Table 6.8 and Figure 6.9.

Package Case Res Source	N	Subset										Group	
		1	2	3	4	5	6	7	8	9	10		
2515	13	1.5769											1
2415	21	1.8213											1
1615	3	2.0000											1
3011	4	2.2050											1
1405	5	2.8000											1
3015	12	2.9392											1
1695	3	3.0000	3.0000										1
3095	3	3.1667	3.1667										1
2095	4	3.3750	3.3750										1
2315	21	4.2238	4.2238										2
1091	4	5.0000	5.0000										2
3195	13	5.0385	5.0385										2
2495	11	5.7273	5.7273										2
1416	7	6.0000	6.0000										2
3096	7	6.2857	6.2857										2
1412	10	6.3000	6.3000										2
2595	6	6.8333	6.8333										2
2084	3	7.0000	7.0000										3
1092	9	7.5556	7.5556										3
1012	21	7.5714	7.5714										3
2615	21	7.7549	7.7549										3
1682	5	8.0000	8.0000	8.0000									3
1311	4	8.5000	8.5000	8.5000									3
2995	12	8.7917	8.7917	8.7917									3
3012	6	8.8333	8.8333	8.8333									3
1406	6	9.3333	9.3333	9.3333									3
1492	10	9.7000	9.7000	9.7000									3
3111	19	10.9211	10.9211	10.9211									4
2716	12	11.3542	11.3542	11.3542									4
2395	13	11.3846	11.3846	11.3846									4
3196	9	11.3889	11.3889	11.3889									4
3116	19	11.5395	11.5395	11.5395									4
3115	34	11.7479	11.7479	11.7479									4
3018	11	11.8182	11.8182	11.8182									4
2385	4	12.5000	12.5000	12.5000	12.5000								4
3191	13	12.7692	12.7692	12.7692	12.7692								4
1614	5	12.8000	12.8000	12.8000	12.8000								4
1391	10	14.0000	14.0000	14.0000	14.0000	14.0000							5
2985	3	14.0000	14.0000	14.0000	14.0000	14.0000							5
3181	4	14.0000	14.0000	14.0000	14.0000	14.0000							5
3112	10	15.1000	15.1000	15.1000	15.1000	15.1000							5
2915	33	16.2121	16.2121	16.2121	16.2121	16.2121							5
1684	4	16.2500	16.2500	16.2500	16.2500	16.2500							5
1481	3	16.6667	16.6667	16.6667	16.6667	16.6667							5
3214	9	17.0556	17.0556	17.0556	17.0556	17.0556							6
2194	8	17.0953	17.0953	17.0953	17.0953	17.0953							6
1411	23	18.6273	18.6273	18.6273	18.6273	18.6273	18.6273	18.6273	18.6273				6
3092	4	19.0000	19.0000	19.0000	19.0000	19.0000	19.0000	19.0000	19.0000				6
2294	13	19.1606	19.1606	19.1606	19.1606	19.1606	19.1606	19.1606	19.1606				6
2014	9	19.3194	19.3194	19.3194	19.3194	19.3194	19.3194	19.3194	19.3194				6
1094	3	20.0000	20.0000	20.0000	20.0000	20.0000	20.0000	20.0000	20.0000				7
1491	16	20.2794	20.2794	20.2794	20.2794	20.2794	20.2794	20.2794	20.2794				7
3086	2	21.0000	21.0000	21.0000	21.0000	21.0000	21.0000	21.0000	21.0000	21.0000			7
3192	10	21.9000	21.9000	21.9000	21.9000	21.9000	21.9000	21.9000	21.9000	21.9000			7
3294	7	21.9364	21.9364	21.9364	21.9364	21.9364	21.9364	21.9364	21.9364	21.9364			7
2214	18	23.4109	23.4109	23.4109	23.4109	23.4109	23.4109	23.4109	23.4109	23.4109			7
2094	6	23.6009	23.6009	23.6009	23.6009	23.6009	23.6009	23.6009	23.6009	23.6009			7
2114	20	23.9359	23.9359	23.9359	23.9359	23.9359	23.9359	23.9359	23.9359	23.9359			7
3094	3	32.6667	32.6667	32.6667	32.6667	32.6667	32.6667	32.6667	32.6667	32.6667			8
Sig.		.055	.052	.051	.054	.051	.051	.054	.052	.052	.056		

Table 6.7: Duncan Test Results

Original Cost Multiplier										
Duncan Group	N	Mean	Median	Std. Deviation	Minimum	Maximum	Range	Skewness	Kurtosis	
G-01	68	2.28	2.00	1.81	0.20	8.00	7.80	1.56	2.00	
G-02	79	5.41	3.00	6.08	0.50	36.00	35.50	2.50	8.53	
G-03	97	8.21	6.00	7.75	1.00	40.00	39.00	1.74	3.44	
G-04	139	11.68	7.00	11.91	0.25	55.00	54.75	1.28	0.98	
G-05	67	15.51	12.00	12.64	1.00	57.00	56.00	1.17	1.10	
G-06	66	18.45	14.01	13.64	2.00	56.88	54.88	1.41	1.16	
G-07	82	22.45	20.00	14.53	1.00	56.00	55.00	0.60	-0.42	
G-08	3	32.67	40.00	21.94	8.00	50.00	42.00	-1.34	0.00	
Total	601	11.98	8.00	12.53	0.20	57.00	56.80	1.51	1.83	

Table 6.8: Statistical Analysis for the Eight Data Clusters

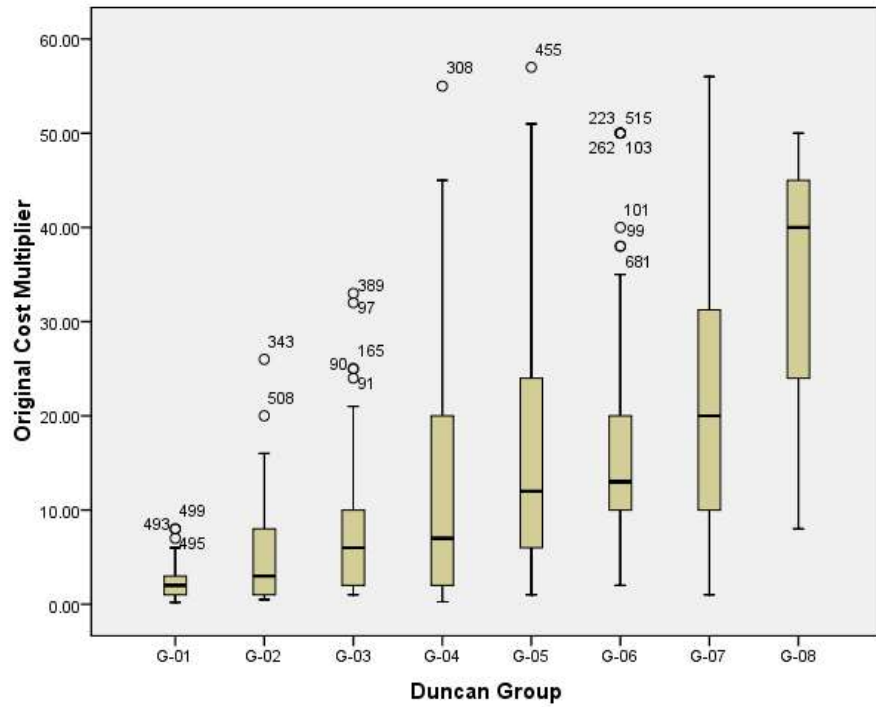


Figure 6.9: BoxPlots for the Eight Data Clusters

Figure 6.10 shows the dataset in SPSS after assigning all the analysis variables. That dataset can be used for lot more tests if more data and attributes are available. The simplicity of the analysis and the techniques used in it opens the door for the end user to continue searching for more patterns and hidden knowledge in the collected data.

Package	Phase	Resource	OriginalMultiplier	PackagePhaseResource	PackagePhaseResource	PackagePhaseResource	Program	Project	Select1	Select2	Select3	TSC_2408	Duncan Group	Summary Group	CurrentMultiplier	
1	Pk-01	Ph-02	R-04	0.50	139	135	95	1395	2	34					0.50	
2	Pk-01	Ph-03	R-02	2.00	131	132	12	1312	2	43					2.00	
3	Pk-01	Ph-03	R-02	2.00	131	132	12	1312	2	47					2.00	
4	Pk-01	Ph-02	R-01	4.00	139	131	91	1391	2	67	1	1	1	G-05	SG-02	4.00
5	Pk-01	Ph-02	R-01	4.00	139	131	91	1391	2	118	1	1	1	G-05	SG-02	4.00
6	Pk-01	Ph-03	R-01	4.00	131	131	11	1311	2	39	1	1	1	G-03	SG-01	4.00
7	Pk-01	Ph-02	R-02	5.00	139	132	92	1392	2	69						
8	Pk-01	Ph-02	R-01	6.00	139	131	91	1391	1	9	1	1	1	G-05	SG-02	
9	Pk-01	Ph-02	R-01	6.00	139	131	91	1391	1	13	1	1	1	G-05	SG-02	
10	Pk-01	Ph-02	R-01	10.00	139	131	91	1391	1	11	1	1	1	G-05	SG-02	
11	Pk-01	Ph-02	R-01	10.00	139	131	91	1391	1	27	1	1	1	G-05	SG-02	
12	Pk-01	Ph-03	R-01	10.00	131	131	11	1311	1	3	1	1	1	G-03	SG-01	
13	Pk-01	Ph-03	R-01	10.00	131	131	11	1311	1	6	1	1	1	G-03	SG-01	
14	Pk-01	Ph-03	R-01	10.00	131	131	11	1311	1	24	1	1	1	G-03	SG-01	
15	Pk-01	Ph-02	R-05	15.00	139	136	96	1396	2	36						15.00
16	Pk-01	Ph-02	R-01	20.00	139	131	91	1391	1	18	1	1	1	G-05	SG-02	
17	Pk-01	Ph-02	R-01	20.00	139	131	91	1391	1	21	1	1	1	G-05	SG-02	
18	Pk-01	Ph-02	R-01	20.00	139	131	91	1391	2	36	1	1	1	G-05	SG-02	20.00
19	Pk-01	Ph-01	R-01	25.00	138	131	81	1381	1	17	1					
20	Pk-01	Ph-01	R-01	25.00	138	131	81	1381	1	20	1					
21	Pk-01	Ph-02	R-01	40.00	139	131	91	1391	2	63	1	1	1	G-05	SG-02	40.00
22	Pk-01	Ph-01	R-01	330.00	138	131	81	1381	1	26						
23	Pk-02	Ph-03	R-02	1.00	141	142	12	1412	2	43	1	1	1	G-02	SG-01	1.00
24	Pk-02	Ph-02	R-04	1.00	149	145	95	1495	2	69	1	1	1	G-01	SG-01	
25	Pk-02	Ph-02	R-05	1.00	149	146	96	1496	2	34	1	1	1	G-03	SG-01	1.00
26	Pk-02	Ph-02	R-05	1.00	149	146	96	1496	2	67	1	1	1	G-03	SG-01	
27	Pk-02	Ph-03	R-04	1.50	141	145	15	1415	2	32						1.50
28	Pk-02	Ph-02	R-01	2.00	149	141	91	1491	1	11	1	1	1	G-07	SG-03	
29	Pk-02	Ph-02	R-01	2.00	149	141	91	1491	2	120	1	1	1	G-07	SG-03	2.00
30	Pk-02	Ph-03	R-01	2.00	141	141	11	1411	2	47	1	1	1	G-06	SG-03	2.00
31	Pk-02	Ph-02	R-02	2.00	149	142	92	1492	1	11	1	1	1	G-03	SG-01	
32	Pk-02	Ph-02	R-02	2.00	149	142	92	1492	2	118	1	1	1	G-03	SG-01	2.00
33	Pk-02	Ph-03	R-02	2.00	141	142	12	1412	1	22	1	1	1	G-02	SG-01	
34	Pk-02	Ph-03	R-02	2.00	141	142	12	1412	2	39	1	1	1	G-02	SG-01	2.00
35	Pk-02	Ph-02	R-04	2.00	149	145	95	1495	2	45	1	1	1	G-01	SG-01	2.00
36	Pk-02	Ph-02	R-04	2.00	149	145	95	1495	2	67	1	1	1	G-01	SG-01	

Figure 6.10: The Final Dataset with the Select and Cluster Variables

This case study presents the value of obtaining the unit costs from historical data using data mining. It shows that extracting useful knowledge from data can be maximized if all data elements are collected properly. Two major problems pertinent to the dataset were found. First, discrepancies were found among the different estimators entries. Estimators are supposed to enter both the estimated quantity of a deliverable and the estimated amount of unit hours per quantity item. The system would then calculate the total estimated hours for a package. However, this was not the case for all data points. Some estimators did not provide estimated quantity; they only put the number '1' in the quantity field. This practice, hence, led to erroneous hourly unit estimation.

Another source of discrepancy was found in the estimating of hours required to complete work packages. Some estimators included all the support activities, such as meetings, site visits and quality inspections, in their production package estimates. Others estimated the required for the support activities independently from the production packages. Again, this led to erroneous hourly unit estimates of production packages.

In addition to the discrepancies found in the estimating entries, discrepancies were found in recording actual entries. The actual hours spent were collected at the project level, as opposed to the planning hours that were estimated at the work package level. Given the levels where the data was collected, there was no possibility to compare or analyze the variance between the estimated and the actual hours spent. Similar to the estimated dataset, actual dataset should have been collected at the work package level.

These discrepancies caused inconsistencies in the data. When the dataset was analyzed, large amount of outliers caused significant disparity in the results. These outliers were highlighted using the outlier detection tool developed in this research and were presented to the data owner for corrective action. Two recommendations were made to the company about these issues, and were approved to be implemented: first, to issue estimating guidelines to ensure consistency among different estimators; second, to modify the timekeeping system in a way to collect actual hours spent at the work package level.

6.2 DISCOVERING KNOWLEDGE IN THE SECOND DATASET

6.2.1 Data Gathering, Cleaning and Preprocessing

The purpose of this case study is to validate the concept that mining historical data enables contactors to better estimate the duration of their work packages. Current practices rely mostly on estimating the duration by dividing the total work hours by the daily number of hours or the scheduler experience. Both practices struggle to provide reliable estimates of package durations that utilize prior experience and current project conditions.

The second dataset, used in this research, contains actual duration and working hours for a large group of fabrication work packages. This dataset included 13,498 data points and is obtained from the second partner company. This company is a large EPC firm that specializes in fabricating structural steel for industrial construction projects. The data was obtained from the scheduling information system of this company, which is a SQL-Server database that was originally designed by the author and developed by the NSERC - Alberta Construction Industry Chair. The data was automatically extracted out of the SQL-Server data tables to MS Excel for cleaning and preprocessing.

The researcher helped the contractor to develop a predefined set of progress activities for their fabrication packages. The start and finish date for each one of these progress activities were collected over a long period of time. The actual steel weight and working hours to complete each fabrication package were also stored in

the information system. The steel weight represents the key quantity for each of these work packages. However, the production package (work package type) was not assigned to the obtained dataset.

The cleaning procedure started by selecting the data point that represents the completed work packages, which means start-date and end-date were marked actual. After that the obvious data entry errors, such as negative values, were also eliminated.

The data for handrails and miscellaneous very small fabrication packages were eliminated as well because they are handled by a separate facility, and is not in the scope of this data-mining exercise. After the cleaning procedure, a large data set with more 5,590 data points is still available to analyze.

The duration $(D)_{(n)}$ in work weeks was calculated using the formula:

$$D_{(n)} = \text{NETWORKDAYS}(\text{FinishDate}, \text{StartDate}) / 5 \quad (6.5)$$

Table 6.10 shows the data from Table 6.9 after it was cleaned, pre-processed and was ready for storage in the data warehouse.

proj_id	job_id	div_id	sub_id	description	det_mhrs	fab_mhrs	weight	fab_start_date	actual_fab_end_date	actual_fab_end_date	actual_ship_start_date	actual_ship_end_date	actual
1	512	1059	1059	25E-0054 Support Structure	2.50	20.00	10.00	07-Aug-01	TRUE	28-Aug-01	TRUE	28-Aug-01	TRUE
1	512	1101	1101	Plant 25 Temp Pipe supports	2.50	20.00	1.00	07-Aug-01	TRUE	28-Aug-01	TRUE	28-Aug-01	TRUE
1	514	1115	1115	Dow-Blk 250 T396 Platform	12.00	22.00	1.50	27-Aug-01	TRUE	07-Sep-01	TRUE	07-Sep-01	TRUE
1	518	1184	1184	Husky Oil	2.00	16.00	26.00	09-Oct-01	TRUE	15-Oct-01	TRUE	15-Oct-01	TRUE
1	518	1185	1185	Husky Oil	2.00	16.00	18.00	09-Oct-01	TRUE	23-Oct-01	TRUE	23-Oct-01	TRUE
1	518	1186	1186	Husky Oil	2.00	16.00	9.00	21-Oct-01	TRUE	02-Nov-01	TRUE	02-Nov-01	TRUE
1	521	1197	1197	ROM Bldg Hopper	5.00	65.00	36.80	09-Nov-01	TRUE	21-Dec-01	TRUE	23-Feb-02	TRUE
1	521	1198	1198	Course Products Bin	5.00	65.00	20.90	10-Dec-01	TRUE	12-Feb-02	TRUE	12-Feb-02	TRUE
1	521	1199	1199	Rejects Bin	5.00	65.00	10.60	10-Dec-01	TRUE	08-Feb-02	TRUE	12-Feb-02	TRUE
1	532	1269	1269	Suncor Breaker Beam	1.00	62.00	4.20	09-Oct-01	TRUE	24-Oct-01	TRUE	24-Oct-01	TRUE
1	535	1338	1338	Job Beams	0.00	12.00	3.00	16-Oct-01	TRUE	20-Oct-01	TRUE	20-Oct-01	TRUE
1	536	1337	1337	Dow Chemical	0.00	54.00	2.70	15-Oct-01	TRUE	29-Oct-01	TRUE	29-Oct-01	TRUE
1	537	1336	1336	Dow Chemical	0.00	30.00	4.00	15-Oct-01	TRUE	29-Oct-01	TRUE	29-Oct-01	TRUE
1	538	1361	1361	Gunther Construction - Exterior	8.00	36.00	1.10	31-Oct-01	TRUE	07-Nov-01	TRUE	07-Nov-01	TRUE
1	538	1405	1405	Gunther Construction - Exterior	8.00	36.00	0.20	05-Nov-01	TRUE	07-Nov-01	TRUE	07-Nov-01	TRUE
1	539	1357	1357	Slave lake Pulp	0.00	0.00	0.80	31-Oct-01	TRUE	06-Nov-01	TRUE	06-Nov-01	TRUE
1	539	1358	1358	Slave lake Pulp	0.00	0.00	34.00	31-Oct-01	TRUE	06-Nov-01	TRUE	06-Nov-01	TRUE
1	540	1359	1359	Slave lake Pulp Flash Dryer	0.00	33.00	0.80	27-Nov-01	TRUE	05-Dec-01	TRUE	05-Dec-01	TRUE
1	541	1360	1360	Slave lake Pulp Star Landing	0.00	36.00	0.80	11-Jan-02	TRUE	16-Jan-02	TRUE	21-Jan-02	TRUE
1	542	1363	1363	Blanchett Neon Limited	0.00	0.00	0.75	24-Oct-01	TRUE	30-Oct-01	TRUE	30-Oct-01	TRUE
1	544	1364	1364	(5) 2000mm Dp Plate Girders	1.50	8.00	302.70	14-Jan-02	TRUE	12-Mar-02	TRUE	12-Mar-02	TRUE
1	544	1365	1365	Floor Beams c/w Stiffners	1.50	24.00	171.50	25-Jan-02	TRUE	20-Mar-02	TRUE	20-Mar-02	TRUE
1	544	1366	1366	(2) End Wall Assemblies	1.50	40.00	59.30	24-Jan-02	TRUE	20-Feb-02	TRUE	15-Mar-02	TRUE
1	544	1367	1367	Bridge Rails /Grating (WT of	1.50	22.00	15.40	25-Feb-02	TRUE	08-Mar-02	TRUE	15-Mar-02	TRUE
1	545	1371	1371	Dow Chemical	0.00	33.00	3.00	29-Oct-01	TRUE	05-Nov-01	TRUE	05-Nov-01	TRUE
1	546	1406	1406	Blanchett Neon	0.00	5.00	0.59	29-Oct-01	TRUE	31-Oct-01	TRUE	31-Oct-01	TRUE
1	547	1407	1407	6 x 3-1/2 angles	0.00	4.00	4.40	29-Oct-01	TRUE	05-Nov-01	TRUE	05-Nov-01	TRUE
1	547	7561	7701		0.00	0.00	0.00		FALSE	FALSE	FALSE	FALSE	FALSE
1	548	1408	1408	2 x 2 angles	0.00	1.50	2.98		FALSE	31-Oct-01	TRUE	31-Oct-01	TRUE
1	549	1427	1427	Dow Site Outfall Upgrade	10.40	28.60	1.20	19-Nov-01	TRUE	04-Dec-01	TRUE	04-Dec-01	TRUE
1	549	1436	1436	Dow Site Outfall Upgrade	10.40	28.60	2.50	07-Jan-02	TRUE	14-Jan-02	TRUE	14-Jan-02	TRUE
1	550	1428	1428	10 Channels for MRC	0.00	15.00	1.30	13-Nov-01	TRUE	15-Nov-01	TRUE	15-Nov-01	TRUE
1	550	1433	1433	Deaerator Pipeway 240 FP-2-	0.00	20.00	1.30	28-Nov-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE
1	550	1434	1434	Platform for Turbidity Probes	0.00	30.00	0.70	03-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE
1	550	1435	1435	Tailings Pumphouse S/S Unity	0.00	15.00	6.50	03-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE
1	550	1450	1450	Unity Water Heater Area FP-	0.00	15.00	0.20	28-Nov-01	TRUE	30-Nov-01	TRUE	30-Nov-01	TRUE
1	550	1470	1470	Impact Cushions FP-2C-5370	0.00	34.00	0.90	07-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE
1	550	1471	1471	HVAC Supports FP-2-5371	0.00	12.00	4.80	12-Dec-01	TRUE	18-Dec-01	TRUE	18-Dec-01	TRUE
1	550	1472	1472	TSRU Train Pipe Supports FP-3-	0.00	18.00	24.00	07-Jan-02	TRUE	14-Jan-02	TRUE	14-Jan-02	TRUE
1	550	1510	1510	Tailings Pumphouse platforms	0.00	15.00	2.20	27-Dec-01	TRUE	16-Jan-02	TRUE	18-Jan-02	TRUE
1	551	1431	1431	Skid and Loading Frames	0.00	110.00	4.40	19-Dec-01	TRUE	04-Feb-02	TRUE	31-Jan-02	TRUE
1	552	1442	1442	Tower Crane Base For Potam	0.00	15.00	0.50	16-Nov-01	TRUE	26-Nov-01	TRUE	26-Nov-01	TRUE
1	553	1443	1443	(2) HSS Cols	0.00	15.00	0.63	26-Nov-01	TRUE	27-Nov-01	TRUE	27-Nov-01	TRUE
1	554	1448	1448	33 Low W-Beam Brackets	0.00	57.00	0.89	03-Dec-01	TRUE	07-Dec-01	TRUE	12-Dec-01	TRUE
1	555	1451	1451	Atco - Ruth Lake power station	0.00	14.00	9.80	29-Nov-01	TRUE	10-Dec-01	TRUE	07-Dec-01	TRUE
1	556	1462	1462	Rotolift Protect	17.00	35.00	4.00	18-Dec-01	TRUE	03-Jan-02	TRUE	07-Jan-02	TRUE

Table 6.9: Raw Dataset for the Second Analysis

ID	Program	Project	Package	Type	Weight	FabHours	FabCostMultiplier	ShopDuration	PaintDuration	FPDuration	ShipDuration	FabDuration	FabDurationMultiplier
640	1	1569	8640	2	1.10	3.60	3.27	1.40	0.20		0.20	1.80	1.64
680	1	1597	7932	1	8.87	30.00	3.38	1.40			0.20	1.60	0.18
692	1	1607	8003	1	1.67	11.00	6.59	1.20			0.40	1.60	0.96
695	1	1609	8022	2	21.86	48.00	2.20	6.00	0.80		0.20	7.00	0.32
696	1	1609	8023	2	26.08	48.00	1.84	5.00	0.60		0.20	5.80	0.22
697	1	1609	8024	2	35.84	48.00	1.34	4.40	0.60		0.20	5.20	0.15
698	1	1609	8025	2	7.45	48.00	6.44	4.80	0.40		0.20	5.40	0.72
703	1	1611	8035	2	1.44	52.80	36.74	1.00	0.60		0.20	1.80	1.25
715	1	1626	8137	2	13.80	33.00	2.39	5.40	1.20		1.20	7.80	0.57
732	1	1632	8197	2	4.20	38.00	9.05	1.40	0.40		0.20	2.00	0.48
739	1	1640	8239	1	1.49	20.82	13.97	1.00			0.20	1.20	0.81
741	1	1640	8239	1	11.97	20.80	1.74	2.60			0.20	2.80	0.23
742	1	1640	8240	1	35.30	16.27	0.46	2.00			0.20	2.20	0.06
746	1	1640	8243	1	5.31	36.40	6.85	1.40			0.60	2.00	0.38
747	1	1640	8244	1	6.20	53.22	8.58	1.60			0.80	2.40	0.39
749	1	1640	8245	1	4.80	35.32	7.36	2.20			0.20	2.40	0.50
754	1	1641	8253	2	1.50	29.60	19.73	1.40	0.60		0.20	2.20	1.47
756	1	1642	8252	1	35.37	18.00	0.51	6.20			0.20	6.40	0.18
776	1	1679	8476	2	2.00	14.83	7.42	7.80	1.40		0.20	9.40	4.70
777	1	1679	8476	1	1.30	14.80	11.38	1.20			1.80	3.00	2.31
778	1	1679	8477	2	18.00	14.80	0.82	5.20	1.40		0.20	6.80	0.38
779	1	1679	8478	2	24.10	14.80	0.61	4.80	1.40		0.40	6.60	0.27
780	1	1679	8479	2	15.70	14.80	0.94	7.40	1.40		0.40	9.20	0.59
781	1	1679	8482	2	2.46	14.80	6.02	5.80	1.40		0.40	7.60	3.09
790	1	1683	8530	4	63.33	18.00	0.28	2.00	0.20	0.20	1.20	3.60	0.06
791	1	1683	8531	4	69.00	14.68	0.21	2.80	0.20	0.20	4.20	7.40	0.11
793	1	1683	8532	4	32.04	20.40	0.64	4.00	0.20	0.20	0.20	4.60	0.14
794	1	1683	8533	4	50.74	20.00	0.39	3.00	0.20	0.20	1.00	4.40	0.09
795	1	1683	8534	4	15.00	21.22	1.41	3.80	0.20	0.20	4.00	8.20	0.55
797	1	1683	8535	4	9.30	57.15	6.15	5.20	0.20	0.20	0.20	5.80	0.62
822	1	1691	8551	1	15.60	28.80	1.85	1.60			5.00	6.60	0.42
829	1	1696	8571	1	14.38	24.00	1.67	8.40			0.20	8.60	0.60
831	1	1696	8714	1	5.13	24.00	4.68	8.40			0.20	8.60	1.68
840	1	1699	8591	1	1.50	20.66	13.77	3.20			0.20	3.40	2.27
863	1	1729	8729	1	1.45	47.05	32.45	1.40			0.20	1.60	1.10
875	1	1752	8787	2	57.78	20.00	0.35	3.00	1.20		0.20	4.40	0.08

Table 6.10: Calculating the Total Fabrication Duration

6.2.2 Clustering of the Cost and Duration Units

The second dataset is obtained from a large EPC firm that is also a member of the NSERC-Alberta Construction Industry Chair. The dataset contained more than five thousand work packages for two standard phases: shop drawings and fabrication. The actual quantities of deliverables, hours and weeks spent on each package is recorded. The fabrication and shop drawings hourly unit cost and weekly unit duration are calculated for every work package in the dataset. This data has been collected over a long period of time. This data has not been analyzed or used before for data mining or knowledge discovery.

The purpose of the analysis of this data is to use historical actual data to develop realistic, reliable and more accurate estimating units for both resource requirement and expected duration. These estimating units are then multiplied by the known quantities to estimate the total duration and resource requirement of a work package.

Since this data is based on actual values, the dataset has been used to validate the developed estimating methodology in this research. The dataset was divided into two parts. The first part, consisting of 85% of the data points, was selected randomly and used for calculating the estimating units. The second part, remaining %15 of the data points, was used for testing purposes.

The software selected to perform the analysis is called *Weka* (Waikato Environment for Knowledge Analysis), which is a free, very powerful and user-friendly data mining and machine learning tool. Weka is developed by University of Waikato in Hamilton, New Zealand (Witten and Frank, 2005). The software is selected because of its powerful data mining capabilities. The software is also easy to obtain, and doesn't require any special hardware; and therefore, it would be accessible to any contractor seeking to perform data mining without incurring major cost.

Minimizing the cost of implementing data mining in industrial construction makes it more appealing to decision makers and also maximizes the return on investment of the increased efficiency.

Weka is able to read data from different types of data files. The first 85% of the dataset is exported from the data warehouse to a Comma Separated Values (CSV) file. Then, it was transferred to Weka in order to perform the analysis. The data contained a unique ID for each data point, two control variables: program and project, the actual amount of key quantity, and total hours and weeks for two resources. One resource is utilized during the fabrication phase and the other one is utilized during the shop drawings phase. The unit cost was calculated by dividing total hours by the key quantity. The unit duration was calculated by dividing the total number of weeks by the key quantity. An excerpt of the CSV data file for the fabrication resource is shown in Table 6.11.

ID	Program	Project	Package	Weight	FabHours	FabHourPerUnit	FabWeeks	FabWeeksPerUnit
1	1	512	1059	10	20	2	3.2	0.32
3	1	514	1115	1.5	22	14.67	2	1.33
4	1	518	1184	26	16	0.62	1	0.04
5	1	518	1185	18	16	0.89	2.2	0.12
6	1	518	1186	9	16	1.78	2	0.22
7	1	521	1197	36.8	65	1.77	6.2	0.17
8	1	521	1198	20.9	65	3.11	9.4	0.45
9	1	521	1199	10.6	65	6.13	9	0.85
10	1	532	1269	4.2	62	14.76	2.4	0.57
12	1	536	1337	2.7	54	20	2.2	0.81
13	1	537	1336	4	30	7.5	2.2	0.55
14	1	538	1361	1.1	36	32.73	1.2	1.09
22	1	544	1365	171.5	24	0.14	7.8	0.05
27	1	547	1407	4.4	4	0.91	1.2	0.27
30	1	549	1427	1.2	28.6	23.83	2.4	2
33	1	550	1433	1.3	20	15.38	2.6	2
35	1	550	1435	6.5	15	2.31	2	0.31
38	1	550	1471	4.8	12	2.5	1	0.21
39	1	550	1472	24	18	0.75	1.2	0.05
40	1	550	1510	2.2	15	6.82	3	1.36
41	1	551	1431	4.4	110	25	6.8	1.55
46	1	556	1462	4	35	8.75	2.6	0.65
47	1	557	1467	2.7	15	5.56	1.6	0.59
55	1	565	1507	1.5	15	10	2.2	1.47
56	1	566	1506	3.75	15	4	2.2	0.59
63	1	573	1548	5.32	51	9.59	2	0.38
65	1	574	1551	11.7	51	4.36	9.4	0.8
70	1	579	1582	18.9	16	0.85	1.6	0.08
71	1	579	1646	24	16	0.67	2	0.08

Table 6.11: An Excerpt of the CSV Data File for the Fabrication Phase

The first element Weka provides when the data is first loaded is complete descriptive analysis for each numeric data variable as shown in Figure 6.11. This descriptive analysis includes the following statistics: the minimum, maximum, mean, standard deviation and a histogram of frequencies.

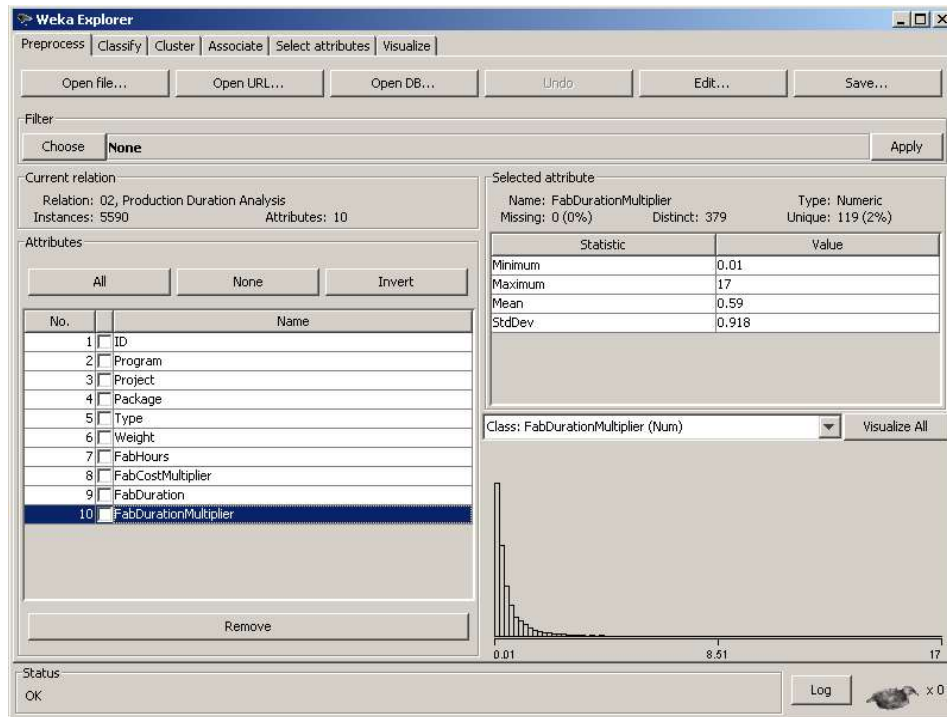


Figure 6.11: The Descriptive Analysis Screen of Weka

Along with this first output, Weka provides the user with a suite of data mining techniques grouped under four categories: classification, clustering, rule association and selecting the most influential attributes. Weka also provides a visual tool that enables users to plot the data using any combination of attributes. Figure 6.12 shows an example of this visualizing technique in Weka.

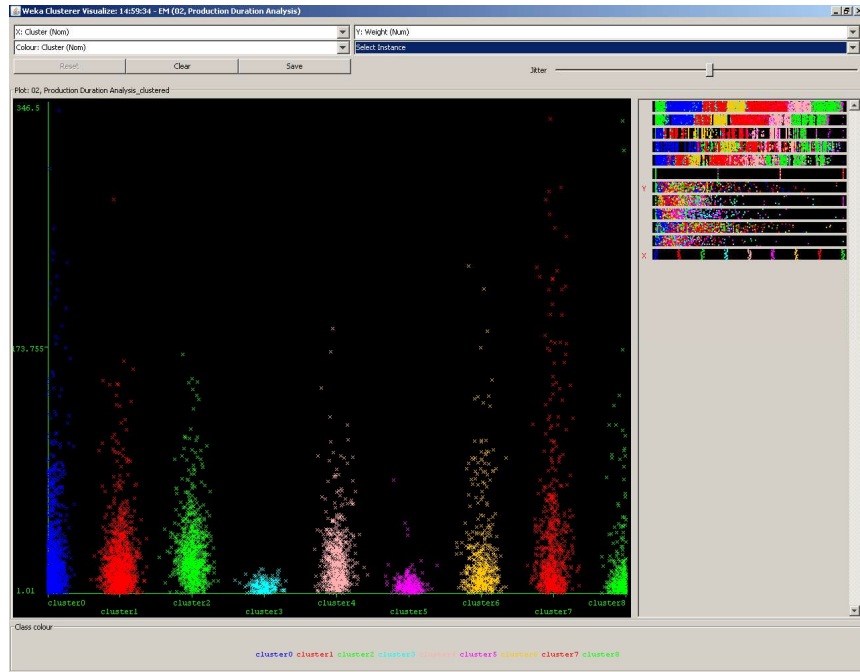


Figure 6.12: The Visualizing Capabilities of Weka

Unlike the first dataset where several resources in multiple phases with different package type were analyzed simultaneously, for this dataset, the analysis is carried on one single resource per phase. And since there is no data collected regarding production package type, the data was analyzed with the assumption that it is all under one production package type. For this analysis, clustering, which is an unsupervised learning technique, is selected. Among the several clustering techniques available in Weka that were tested, the Expectation Maximization (EM) technique was found to be the most efficient one. The software developers highly recommend this technique for clustering large sets of data and it shows as the default technique to be used.

The EM clustering technique is implemented to the dataset and the results are summarized in Tables 6.11, 6.12, 6.13 and 6.14. In order to ensure the stability of the clustering results, each clustering analysis was repeated three times, with each run taking about two and half hours of processing time on an Intel Pentium(R) personal computer. The results were as follow: nineteen clusters were obtained for the fabrication hourly unit cost (Table 6.11), thirteen clusters for the fabrication weekly unit duration (Table.12), five clusters for the shop drawings hourly unit cost (Table 6.13) and six clusters for the shop drawings weekly unit duration (Table 6.14). For each cluster, the number of data points, mean, standard deviation, and prior probability are obtained from Weka.

Fabrication Hours per Unit					
Cluster	N	%	Mean	StdDev	Prior Probability
0	233	4.94%	3.92	0.5080	0.0520
1	39	0.83%	46.12	13.2537	0.0123
2	73	1.55%	28.58	3.9670	0.0159
3	265	5.62%	2.41	0.2544	0.0562
4	142	3.01%	8.81	0.9488	0.0305
5	77	1.63%	19.48	1.7999	0.0170
6	227	4.81%	3.08	0.2724	0.0433
7	2	0.04%	112.09	2.0850	0.0004
8	30	0.64%	37.96	1.2437	0.0044
9	62	1.31%	13.23	0.8030	0.0129
10	32	0.68%	23.05	0.6121	0.0047
11	376	7.97%	1.83	0.2362	0.0814
12	95	2.01%	15.53	0.9313	0.0190
13	1151	24.39%	0.64	0.2078	0.2432
14	171	3.62%	6.81	0.7811	0.0362
15	593	12.57%	0.26	0.1254	0.1210
16	123	2.61%	11.12	0.9338	0.0274
17	212	4.49%	5.27	0.7573	0.0468
18	816	17.29%	1.17	0.2478	0.1753
Total	4,719.00	100.00%	4.22	7.8950	1.00

Log likelihood: -2.17993

Table 6.11: Clustering of Fabrication Hourly Unit Cost

Fabrication Weeks per Unit					
Cluster	N	%	Mean	StdDev	Prior Probability
0	184	3.90%	1.12	0.1774	0.0376
1	396	8.39%	0.33	0.0589	0.0933
2	38	0.81%	3.06	1.0114	0.0113
3	275	5.83%	0.65	0.1006	0.0600
4	4	0.08%	5.55	2.5024	0.0015
5	143	3.03%	1.61	0.2966	0.0287
6	754	15.98%	0.14	0.0305	0.1411
7	70	1.48%	2.19	0.4809	0.0184
8	202	4.28%	0.88	0.1452	0.0441
9	774	16.40%	0.04	0.0165	0.1409
10	436	9.24%	0.47	0.0754	0.0862
11	622	13.18%	0.22	0.0458	0.1407
12	821	17.40%	0.09	0.0252	0.1962
Total	4,719.00	100.00%	0.40	0.5820	1.00

Log likelihood: 0.05683

Table 6.12: Clustering of Fabrication Weekly Unit Duration

Shop Drawings Hours per Unit					
Cluster	N	%	Mean	StdDev	Prior Probability
0	17	0.69%	10.07	8.3876	0.0087
1	360	14.64%	0.82	0.3455	0.1615
2	641	26.07%	0.32	0.1335	0.2887
3	186	7.56%	2.39	1.1511	0.0888
4	1,255	51.04%	0.12	0.0595	0.4524
Total	2,459.00	100.00%			1.00

Log likelihood: -0.24503

Table 6.13: Clustering of Shop Drawings Hourly Unit Cost

Shop Drawings Weeks per Unit					
Cluster	N	%	Mean	StdDev	Prior Probability
0	1,293	31.24%	0.42	0.1781	0.329
1	621	15.00%	1.05	0.4051	0.1673
2	199	4.81%	2.45	0.8836	0.0555
3	14	0.34%	12.64	5.2183	0.0041
4	1,966	47.50%	0.15	0.0742	0.4296
5	46	1.11%	4.78	1.5437	0.0145
Total	4,139.00	100.00%			1.00

Log likelihood: -0.41406

Table 6.14: Clustering of Shop Drawings Weekly Unit Duration

Initial results of the clustering exercise demonstrate trends that would benefit the contractor. Cluster with large number of data points are expected to represent common cases of packages in the contracting company. While clusters with small number of data points represent either rare types of work packages or outliers that have to be further investigated.

For instance, results in Table 6.11 show that almost a quarter of the work packages fall in cluster 13, with a mean of 0.6 hours per unit. In the same table, packages in cluster 7 represent a case of outliers that should be investigated. When a contractor needs to investigate the clustering analysis results, they can easily find out which data point belongs to which cluster, since Weka assigns the results of the clustering to every data point in the dataset and automatically draws the frequency histograms as shown in Figures 6.13 and 6.14. Assigning clusters to every data point makes it easy for contractors to go back to their files and find out the reasons behind the variation in actual package costs and durations.

Viewer
Relation: 02, Production Duration Analysis_clustered

No.	Instance_number Numeric	ID Numeric	Program Numeric	Project Numeric	Package Numeric	Type Numeric	Weight Numeric	FabHours Numeric	FabCostMultiplier Numeric	FabDuration Numeric	FabDurationMultiplier Numeric	Cluster Nominal
1	0.0	1.0	1.0	512.0	1059.0	1.0	10.0	20.0	2.0	3.4	0.34	cluster12
2	1.0	3.0	1.0	514.0	1115.0	1.0	1.5	22.0	14.67	2.2	1.47	cluster10
3	2.0	4.0	1.0	518.0	1184.0	1.0	26.0	16.0	0.62	3.2	0.12	cluster14
4	3.0	5.0	1.0	518.0	1185.0	1.0	18.0	16.0	0.89	3.2	0.18	cluster15
5	4.0	6.0	1.0	518.0	1186.0	1.0	9.0	16.0	1.78	3.0	0.33	cluster12
6	5.0	7.0	1.0	521.0	1197.0	1.0	36.8	65.0	1.77	6.4	0.17	cluster8
7	6.0	8.0	1.0	521.0	1198.0	1.0	20.9	65.0	3.11	10.6	0.51	cluster17
8	7.0	9.0	1.0	521.0	1199.0	1.0	10.6	65.0	6.13	9.2	0.87	cluster11
9	8.0	10.0	1.0	532.0	1269.0	1.0	4.2	62.0	14.76	3.2	0.76	cluster5
10	9.0	12.0	1.0	536.0	1337.0	1.0	2.7	54.0	20.0	2.8	1.04	cluster16
11	10.0	13.0	1.0	537.0	1336.0	1.0	4.0	30.0	7.5	2.8	0.7	cluster11
12	11.0	14.0	1.0	538.0	1361.0	1.0	1.1	36.0	32.73	2.4	2.18	cluster9
13	12.0	21.0	1.0	544.0	1364.0	1.0	302.7	8.0	0.03	9.0	0.03	cluster14
14	13.0	22.0	1.0	544.0	1365.0	1.0	171.5	24.0	0.14	8.0	0.05	cluster14
15	14.0	23.0	1.0	544.0	1366.0	1.0	59.3	40.0	0.67	5.2	0.09	cluster14
16	15.0	25.0	1.0	545.0	1371.0	1.0	3.0	33.0	11.0	1.4	0.47	cluster11
17	16.0	27.0	1.0	547.0	1407.0	1.0	4.4	4.0	0.91	1.4	0.32	cluster15
18	17.0	30.0	1.0	549.0	1427.0	1.0	1.2	28.6	23.83	2.6	2.17	cluster10
19	18.0	31.0	1.0	549.0	1436.0	1.0	2.5	28.6	11.44	1.6	0.64	cluster11
20	19.0	33.0	1.0	550.0	1433.0	1.0	1.3	20.0	15.38	2.8	2.15	cluster10
21	20.0	35.0	1.0	550.0	1435.0	1.0	6.5	15.0	2.31	2.2	0.34	cluster17
22	21.0	38.0	1.0	550.0	1471.0	1.0	4.8	12.0	2.5	1.2	0.25	cluster8
23	22.0	39.0	1.0	550.0	1472.0	1.0	24.0	18.0	0.75	1.4	0.06	cluster14
24	23.0	40.0	1.0	550.0	1510.0	1.0	2.2	15.0	6.82	3.2	1.45	cluster0

Undo OK Cancel

Figure 6.13: Weka Results Viewer

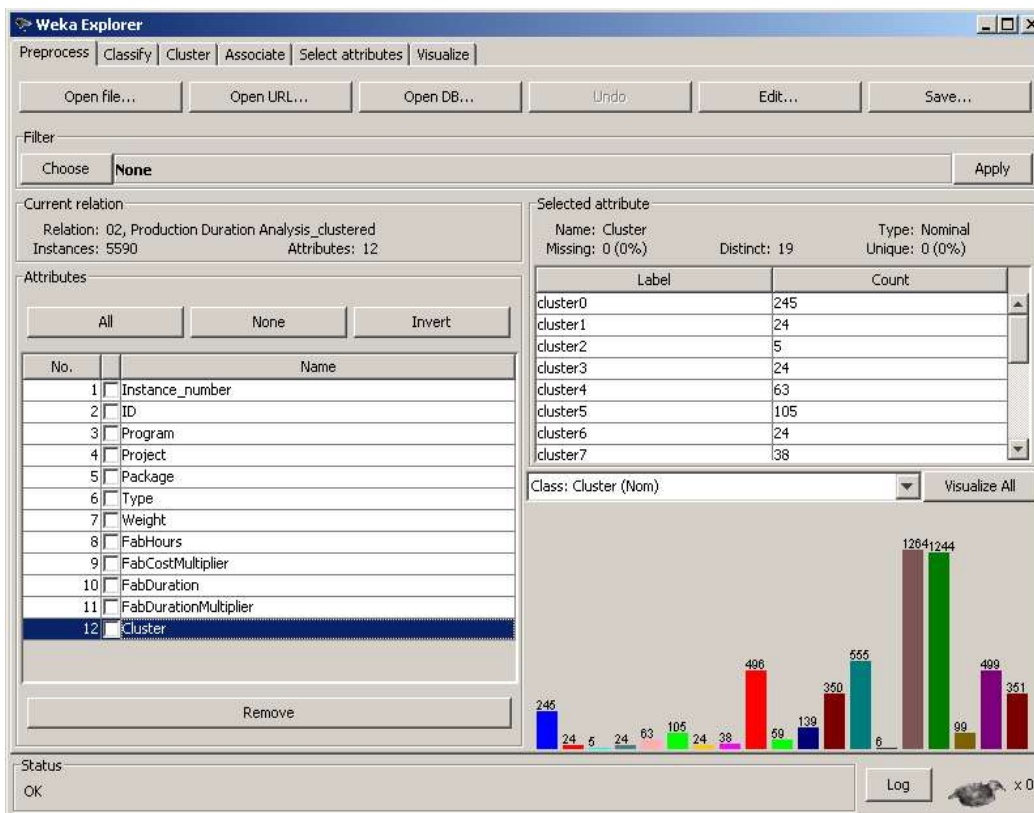


Figure 6.14: The Frequency Histogram for the Obtained Clusters

The Weka analysis supports the claim of this research model that data, which up to now was not used, can be transferred into useful knowledge that ultimately provides meaningful insights into the working of contractors. When data is collected, stored and pre-processed in a proper way as proposed in this research, an endless wealth of knowledge can be harvested from this data. After assigning the clusters, a fitting distribution can be found for each cluster.

6.2.3 Case Study Results Validation

The second part of the data, the remaining 15 % was used for validation, as mentioned earlier. The obtained unit costs and durations from the clustering analysis are used to estimate the resource requirement and duration of each work package in the validation dataset. Each package was assigned a duration unit cluster and a cost unit cluster (Table 6.15). The means of these two clusters were used to estimate the total resource requirement and duration for each package. Both the cost and duration variances accompanied with errors percentage were calculated for each package as well.

The validation test showed that, when comparing the estimated values using the obtained unit based on historical data with the actual values that were recorded for these packages, more than 80% of the tested data points had an estimating error of only below 25%. These results demonstrate a significant increase in the accuracy of estimating practices when relying on historical data that existed already in the contractor's management systems.

ID	Weight	Actual Fab Weeks	FabWeeks Per Unit Cluster	Estimated Fab Weeks	Weeks Variance	Weeks Error
25	3	1.2	cluster13	1.4	0.21	14.77%
31	2.5	1.2	cluster13	1.2	-0.03	2.28%
57	4	1.2	cluster7	1.3	0.11	8.31%
75	9.7	5.6	cluster9	6.3	0.72	11.43%
91	6	3.6	cluster9	3.9	0.31	7.95%
109	7.9	3.2	cluster13	3.7	0.51	13.69%
161	1.5	1.6	cluster6	1.7	0.08	5.00%
171	1.6	4.6	cluster2	4.9	0.30	6.12%
177	5.3	9	cluster5	8.5	-0.49	5.72%
185	2.2	1	cluster13	1.0	0.03	3.14%
201	5.6	2	cluster7	1.8	-0.17	9.15%
235	49.9	3.8	cluster10	4.3	0.52	11.96%
247	2.6	1.8	cluster1	1.7	-0.11	6.21%
251	2.1	3.6	cluster5	3.4	-0.23	6.72%
266	11.4	2.8	cluster8	2.5	-0.27	10.54%
279	2.8	1.6	cluster9	1.8	0.23	12.33%
294	5.4	3.6	cluster1	3.5	-0.08	2.28%
316	2.7	1.6	cluster9	1.8	0.16	9.08%
336	117	5.4	cluster12	5.1	-0.25	4.90%
353	3.4	2.4	cluster1	2.2	-0.18	8.30%
364	2.2	1	cluster13	1.0	0.03	3.14%
366	13.2	1.8	cluster15	1.9	0.07	3.63%
422	5.3	2.2	cluster13	2.5	0.29	11.55%
459	30.3	14	cluster13	14.2	0.22	1.55%
505	2.8	14.2	cluster14	15.5	1.33	8.54%
559	1.7	1	cluster9	1.1	0.11	9.75%
583	8.5	3	cluster7	2.8	-0.22	7.87%

Table 6.15: An Excerpt of the Validation Tool for the Fabrication Phase

The work package types were not identified when its data was recorded. When the data mining analysis was conducted, data clusters were identified. Consequently, it was left to the estimator to decide which cluster to use for estimating future projects. The partner company did not record its planned data in a structured way as it did with the actual data. Thus, performance evaluation using EVM was not possible.

6.3 DISCOVERING KNOWLEDGE IN THE THIRD DATASET

6.3.1 Data Gathering, Cleaning and Preprocessing

The purpose of this case study is to validate the concept that data mining can be used to provide reliable probabilistic resource utilization graphs (resource baseline histograms) that can be used for proper staffing of projects. These graphs show the required weekly hours of a specific resource during the duration of a project or work package. Data mining provide a set of various graphs based on different combinations of control attributes; hence provide contractors with the ability to utilize the most suitable graph. The current practices mostly rely on using uniform or predefined distribution graphs that do not rely on historical data and are not customized to reflect current conditions.

The third dataset to be used in this research is obtained from the same partner company that provided the first data set. This third dataset contains the actual weekly hours for a set of resources per project phase. The current practice in the company is to collect actual hours by project phase instead of work packages. Although, this data is not collected at the work package level as proposed in this research, this data is still very useful for providing analysis on the project level for providing Initial Planned Values (IPV) of project resource requirements. The same methodology can be applied to obtain resource utilization curves per work package for estimating resource requirements during the detailed planning stage of any project.

The procedure of obtaining the third data set started with getting a list of all completed projects between the years 2004 and 2007 as shown in Table 6.16. This list was obtained from the timekeeping system of the company, which is an in-house developed SQL server application. The list contained more than 1,500 projects that vary in duration, cost and complexity. The data was automatically extracted out of the SQL-Server data tables to MS Excel for cleaning and preprocessing.

JobNum	JobGroup	Company	Description	EIC	ProjSponsor	ProjManager	Comments
00E1276	PACER Alliance	PETRO-CANADA PACER	Motor Protective Relay Enhancements		GAM	F NOLTE	
00E1299	PACER Alliance	PETRO-CANADA PACER	2000 / 2001 Heavy Oils Platforms		GAM	F NOLTE	
00E1450	PACER Alliance	PETRO-CANADA PACER	Replacement of HF Detection / H2O Monitor PLC				
00E1450c	PACER Alliance	PETRO-CANADA PACER	Replacement of HF Detection / H2O Monitor PLC	CKS	G MACMILLAN	A LENUJK	
01E1465		COLT ENGINEERING CORP - EDMONTON					
01E1476	CORE PROJECTS	BP CANADA CHEMICAL	Set Up Project Management Files & System	RJT	B Bowhay	B TURCOT	
01E1505	CORE PROJECTS	BP CANADA CHEMICAL	MOC #62 - Rail Loading Pumps Shutdown Control	RJT	B Bowhay	B TURCOT	
01E1506	CORE PROJECTS	BP CANADA CHEMICAL	MOC #79 - T-5906 A/B Valve for Rail Loading	RJT	B Bowhay	B TURCOT	
01E1510	CORE PROJECTS	BP CANADA CHEMICAL	MOC #89 Portable Nitrogen Heater Cart Construction	RJT	B Bowhay	B TURCOT	
01E1511	CORE PROJECTS	BP CANADA CHEMICAL	MOC #49 K-5201 A/B Suction Line Drain	RJT	B Bowhay	B TURCOT	
01E1512	CORE PROJECTS	BP CANADA CHEMICAL	MOC #10 VP-5452 Vapour Bypass	RJT	B Bowhay	B TURCOT	
01E1514	CORE PROJECTS	BP CANADA CHEMICAL	MOC #104 T-5802 Rework Tank Transfer to D-5615	RJT	B Bowhay	B TURCOT	
01E1534	CORE PROJECTS	BP CANADA CHEMICAL	MOC #15 D-5475, D-5480 Hotwell Sample Port	RJT	B Bowhay	B TURCOT	
01E1535	CORE PROJECTS	BP CANADA CHEMICAL	MOC #102 C-5430 Butane Sample Points	RJT	B Bowhay	B TURCOT	
01E1537	CORE PROJECTS	BP CANADA CHEMICAL	MOC #082 Hot Oil Pump Isolation Valve Controls	RJT	B Bowhay	B TURCOT	
01E1560	CORE PROJECTS	SHELL CHEMICALS CANADA	Alliance Procedures		GAM / ASN	W MATTER	
01E1562		SUNCOR ENERGY INC 200	Millennium Extraction Wood Removals	RA	B Bowhay	M EWANCHUK	
01E1571	CORE PROJECTS	BP CANADA CHEMICAL	MOC #02 Zone Store in LAO Plant	RJT	R Karren	T Kucher	
01E1572	CORE PROJECTS	BP CANADA CHEMICAL	MOC #072 Control Building Lab. Modifications	RJT	B Bowhay	B TURCOT	
01E1572A	CORE PROJECTS	BP CANADA CHEMICAL	MOC 72 - Control Building Lab. Modifications	RJT	B BOWHAY	B TURCOT	
01E1578	CORE PROJECTS	BP CANADA CHEMICAL	MOC #132 Decene for Seal Liquid D-5982 and D-5984	RJT	B Bowhay	B TURCOT	
01E1582	CORE PROJECTS	BP CANADA CHEMICAL	Redlines and As-Builts	RJT	B Bowhay	B TURCOT	
01E1582A	CORE PROJECTS	BP CANADA CHEMICAL	Redlines & As-Builts - NON COLT MOC's REV 6B	DGL	B Bowhay	B TURCOT	
01E1582B	CORE PROJECTS	BP CANADA CHEMICAL	Redlines & As-Builts - NON COLT MOC's after April	DGL	R Karren	T Kucher	
01E1582D	CORE PROJECTS	BP CANADA CHEMICAL	Structural As-Builts	Don Li	B BOWHAY	B TURCOT	
01E1597	CORE PROJECTS	BP CANADA CHEMICAL	MOC #135 - Installation of Maintenance Access Door	RJT	B Bowhay	B TURCOT	
01E1599	CORE PROJECTS	BP CANADA CHEMICAL	Document Control	RJT	B Bowhay	B TURCOT	
01E1608	CORE PROJECTS	BP CANADA CHEMICAL	MOC #150 Drainage for Control Valves Outside	RJT	B Bowhay	B TURCOT	
01E1614	CORE PROJECTS	BP CANADA CHEMICAL	MOC #164 - Install tie-ins to reroute SF-5210A/B O	RJT	B Bowhay	B TURCOT	
01E1617	CORE PROJECTS	LAO CANADA CHEMICAL P	MOC 086 - Maintenance Small Equipment Decontami	RJT	R Karren	B Turcot	
01E1643	CORE PROJECTS	BP CANADA CHEMICAL	AA - Line Design Pressure Change for Start-Up Mode				
01E1657	CORE PROJECTS	BP CANADA CHEMICAL	Inst. of Pad and Utilities for Skid Fuel Tanks	DM	B Bowhay	B TURCOT	
01E1662	CORE PROJECTS	BP CANADA CHEMICAL	Betz BFW Treatment Skid Shelter (MOC #84)	DL	B Bowhay	B TURCOT	
01E1663	CORE PROJECTS	BP CANADA CHEMICAL	Cylinder / Packing Lubrication Reservoir Upgrade (DL	B Bowhay	B TURCOT	

Table 6.16: List of Completed Projects between 2004 and 2007

Project phase is an important control attribute for the data mining exercise. However, the company did not clearly assign project phases to the data points in their timekeeping system. As a result, it was necessary in this research to go back to the archives in order to assign the proper phase to each project. This process again consumed lots of time and efforts.

Since the construction support phase is mostly responding to requests from sites and is not performed based on clearly defined scope, projects that were assigned to the “construction support” phase were eliminated from the dataset. Projects that were cancelled or put on hold prior to delivering their scope were eliminated from the list as well. At the end, there were more than 350 projects in the dataset. For each of these projects, a SQL statement was run to query the weekly working hours per resource type as shown in Table 6.17.

JOBNUM	SUBJOB	LICODE	JobGroup	SUM(Hours)	Period End
04E2583	891	1130	CORE PROJECTS	1.00	07-Jan-05
04E2583	891	1130	CORE PROJECTS	5.50	21-Jan-05
04E2583	891	1130	CORE PROJECTS	16.50	11-Feb-05
04E2583	891	1130	CORE PROJECTS	0.50	25-Feb-05
04E2583	896	1130	CORE PROJECTS	2.50	03-Sep-04
04E2583	903	1130	CORE PROJECTS	0.50	25-Feb-05
04E2583	904	1130	CORE PROJECTS	0.50	25-Feb-05
04E2583	904	1130	CORE PROJECTS	1.00	04-Mar-05
04E2583	905	1130	CORE PROJECTS	7.50	10-Dec-04
04E2583	908	1130	CORE PROJECTS	1.00	26-Nov-04
04E2583	908	1130	CORE PROJECTS	0.50	10-Dec-04
04E2583	908	1130	CORE PROJECTS	4.50	21-Jan-05
04E2583	908	1130	CORE PROJECTS	7.00	04-Feb-05
04E2583	908	1130	CORE PROJECTS	16.25	25-Feb-05
04E2583	914	1130	CORE PROJECTS	10.00	26-Nov-04
04E2583	914	1130	CORE PROJECTS	6.50	17-Dec-04
04E2583	914	1130	CORE PROJECTS	1.00	07-Jan-05
04E2583	914	1130	CORE PROJECTS	0.50	21-Jan-05
04E2583	920	1130	CORE PROJECTS	1.00	12-Nov-04
04E2583	920	1130	CORE PROJECTS	1.50	26-Nov-04
04E2583	920	1130	CORE PROJECTS	1.00	03-Dec-04
04E2583	920	1130	CORE PROJECTS	0.50	17-Dec-04
04E2583	920	1130	CORE PROJECTS	1.00	07-Jan-05
04E2583	920	1130	CORE PROJECTS	2.50	18-Feb-05
04E2583	920	1130	CORE PROJECTS	23.25	04-Mar-05
04E2583	921	1130	CORE PROJECTS	1.00	03-Dec-04
04E2583	949	1130	CORE PROJECTS	1.00	12-Nov-04
04E2553	0	1130	SHELL ALLIANCE	5.00	26-Mar-04

Table 6.17: Weekly Actual Working Hours per Resource

The company did not store the original planned, current planned, earned hours on a weekly basis. Therefore, the missing data was simulated using random numbers

in order to populate the data warehouse according to the proposed structure. The complete dataset is used to calculate the performance measures (CPI and SPI) on a weekly basis for all the data points.

The period end-date was used to calculate the week, month, quarter, and year numbers for each data point to expedite the procedure of running OLAP reports and queries. The formula used to calculate the year is:

$$\text{YearNumber} = \text{Year}(\text{PeriodEndDate}). \quad [6.6]$$

The formula used to calculate the month number is:

$$\text{MonthNumber} = \text{Month}(\text{PeriodEndDate}). \quad [6.7]$$

The formula to calculate the week number is:

$$\text{WeekNumber} = \text{Weeknum}(\text{PeriodEndDate}). \quad [6.8]$$

The three-point sliding moving average was used to reduce the noise in the dataset (Teicholz, 1993). After that, the duration data was normalized by dividing the week number by the total number of weeks. The cost data was also normalized by dividing the weekly hours by the total number of hours. The normalized data is shown in Table: 6.18.

Period	Hours				Normalized	
Number	Raw		Smoothe	Project	Duration	Hours
4	33.00	27.50	28.86	04E2510	0.36364	0.12401
5	2.00	17.00	18.36	04E2510	0.45455	0.07890
6	16.00	11.58	12.95	04E2510	0.54545	0.05563
7	16.75	12.17	13.53	04E2510	0.63636	0.05813
8	3.75	10.00	11.36	04E2510	0.72727	0.04882
9	9.50	5.42	6.78	04E2510	0.81818	0.02913
10	3.00	5.00	6.36	04E2510	0.90909	0.02734
11	2.50	1.83	3.20	04E2510	1.00000	0.01374
1	1.00	4.17	4.37	04E2518	0.05556	0.03776
2	11.50	10.00	10.20	04E2518	0.11111	0.08815
3	17.50	13.17	13.37	04E2518	0.16667	0.11551
4	10.50	11.00	11.20	04E2518	0.22222	0.09679
5	5.00	7.33	7.54	04E2518	0.27778	0.06511
6	6.50	7.42	7.62	04E2518	0.33333	0.06583
7	10.75	6.67	6.87	04E2518	0.38889	0.05936
8	2.75	8.75	8.95	04E2518	0.44444	0.07735
9	12.75	6.83	7.04	04E2518	0.50000	0.06080
10	5.00	8.33	8.54	04E2518	0.55556	0.07375
11	7.25	4.75	4.95	04E2518	0.61111	0.04280
12	2.00	4.58	4.79	04E2518	0.66667	0.04136
13	4.50	2.33	2.54	04E2518	0.72222	0.02192
14	0.50	1.83	2.04	04E2518	0.77778	0.01760
15	0.50	0.50	0.70	04E2518	0.83333	0.00608
16	0.50	2.75	2.95	04E2518	0.88889	0.02552
17	7.25	5.92	6.12	04E2518	0.94444	0.05288
18	10.00	5.75	5.95	04E2518	1.00000	0.05144

Table 6.18: The Normalized Dataset

Nassar, 2004 stated that dividing project progress to twenty equal periods with 5% increments is a very good method to measure projects performance. Based on that, the dataset was normalized using the interpolation function of the R software. An example of the output of the code is shown in Table 6.25.

Group	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
61	0.01859	0.02289	0.02719	0.03903	0.05181	0.07207	0.09446	0.12672	0.16392	0.1924	0.2139	0.21933	0.20465	0.18332	0.14865	0.11441	0.08164	0.05089	0.03634	0.02175
61	0.01995	0.07725	0.11094	0.11779	0.10512	0.10697	0.09117	0.05608	0.04095	0.0595	0.06176	0.03778	0.0117	0.00559	0.00567	0.00588	0.00588	0.00668	0.01049	0.00975
61	0.02356	0.06549	0.06856	0.08981	0.06422	0.05598	0.03317	0.03209	0.04361	0.08195	0.09916	0.07918	0.04058	0.02804	0.03306	0.03437	0.03302	0.02071	0.01593	0.0066
61	0.04786	0.06744	0.07782	0.08053	0.07686	0.07584	0.07539	0.07488	0.07311	0.07044	0.06952	0.07117	0.07674	0.08202	0.08556	0.08553	0.06789	0.04823	0.02819	0.00888
62	0.10235	0.11396	0.06453	0.03201	0.04584	0.03713	0.03239	0.02761	0.02808	0.02949	0.02898	0.028	0.02385	0.01907	0.01997	0.02202	0.03247	0.03811	0.04391	0.03504
62	0.01623	0.06654	0.07814	0.05078	0.03472	0.03697	0.03313	0.01664	0.01469	0.01502	0.01675	0.0188	0.01989	0.01255	0.01099	0.01726	0.04354	0.05835	0.02588	0.01266
62	0.07212	0.09202	0.10451	0.11304	0.12194	0.11577	0.09677	0.06834	0.03744	0.06714	0.08914	0.09273	0.09273	0.05487	0.02689	0.02543	0.01863	0.01817	0.01784	0.01311
62	0.0326	0.04334	0.04427	0.04209	0.0394	0.03789	0.04062	0.05405	0.06983	0.07717	0.07885	0.07507	0.06484	0.06752	0.07524	0.07625	0.07566	0.07138	0.0601	0.0493
62	0.01968	0.01574	0.03836	0.08971	0.07553	0.05062	0.03374	0.04031	0.02409	0.01349	0.01261	0.00909	0.00853	0.00915	0.01303	0.0228	0.01869	0.01617	0.01285	0.0060
62	0.06871	0.06296	0.02233	0.01239	0.01293	0.01333	0.00713	0.0063	0.01412	0.02501	0.02642	0.02136	0.01937	0.02623	0.03702	0.02638	0.08154	0.04518	0.0367	0.0273
62	0.14623	0.1613	0.15377	0.05207	0.05395	0.03983	0.02571	0.01911	0.03889	0.04171	0.04548	0.02571	0.02665	0.02006	0.02006	0.01629	0.01629	0.02476	0.03983	0.032
62	0.10074	0.11945	0.10345	0.07748	0.05526	0.04021	0.02757	0.01228	0.0232	0.03914	0.06518	0.08846	0.08332	0.06835	0.06536	0.06201	0.06019	0.04842	0.02146	0.01144
62	0.04551	0.066	0.05578	0.05076	0.07815	0.08364	0.05865	0.04745	0.03553	0.02654	0.01848	0.01245	0.00983	0.01153	0.0107	0.00647	0.00708	0.00514	0.00384	0.0043
62	0.09602	0.12268	0.11389	0.09043	0.08726	0.06083	0.05065	0.03169	0.02296	0.01256	0.01223	0.01035	0.00955	0.01431	0.01907	0.02601	0.03076	0.05158	0.05158	0.04011
62	0.0492	0.03532	0.02403	0.05303	0.07449	0.06198	0.04523	0.02948	0.01003	0.01645	0.01979	0.01419	0.01455	0.03837	0.05535	0.03507	0.01796	0.01776	0.03841	0.0371
62	0.00473	0.03866	0.02601	0.01466	0.02131	0.02185	0.03203	0.0367	0.03206	0.03272	0.02357	0.02541	0.03922	0.03245	0.02915	0.01669	0.01443	0.04119	0.03652	0.0200
62	0.02429	0.0349	0.03324	0.02859	0.02243	0.02896	0.03083	0.02716	0.02486	0.05468	0.08358	0.07939	0.04651	0.03138	0.05715	0.07279	0.0388	0.02375	0.02856	0.0259
62	0.06597	0.09953	0.0936	0.09454	0.08653	0.07754	0.06537	0.04497	0.03625	0.03894	0.03988	0.04168	0.04201	0.03938	0.03675	0.03445	0.03264	0.0379	0.05632	0.0451
62	0.09371	0.11381	0.1339	0.14769	0.14567	0.14366	0.14331	0.14421	0.1451	0.14064	0.13527	0.13046	0.12889	0.12733	0.12644	0.12644	0.12644	0.11591	0.10118	0.0864
62	0.01884	0.03198	0.0421	0.04837	0.05125	0.05312	0.05708	0.06101	0.0714	0.09075	0.10459	0.11497	0.11498	0.10548	0.09016	0.07685	0.05582	0.03824	0.02461	0.00888
64	0.09938	0.12642	0.12385	0.09674	0.07962	0.05666	0.04242	0.09605	0.10975	0.10975	0.12327	0.09068	0.05715	0.05264	0.04432	0.0353	0.02629	0.02311	0.01936	0.014
64	0.15785	0.25616	0.20546	0.09575	0.02066	0.02937	0.03789	0.03059	0.0195	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.0071
64	0.06106	0.06313	0.04826	0.02068	0.01489	0.02183	0.0352	0.04017	0.09357	0.1575	0.18482	0.19526	0.1776	0.10806	0.06828	0.04412	0.03625	0.03144	0.02441	0.01444
64	0.02371	0.02109	0.0128	0.05771	0.08689	0.0733	0.02612	0.04898	0.06482	0.07795	0.05974	0.07428	0.04957	0.02614	0.0239	0.0216	0.03612	0.03064	0.01184	0.0033
64	0.00284	0.01235	0.0785	0.14162	0.1128	0.04629	0.02198	0.0403	0.051	0.05442	0.03735	0.01442	0.0109	0.00521	0.00413	0.00454	0.00874	0.00928	0.01131	0.0118
64	0.00221	0.0031	0.00399	0.00965	0.03999	0.0831	0.12392	0.14896	0.14574	0.1485	0.13848	0.10853	0.07038	0.03671	0.01968	0.01297	0.00775	0.00562	0.00688	0.004
64	0.1053	0.14783	0.18566	0.18123	0.17007	0.13197	0.09919	0.0788	0.06817	0.07216	0.07637	0.0808	0.07725	0.06839	0.06635	0.06724	0.05891	0.04827	0.03525	0.02194
64	0.08889	0.13292	0.14024	0.12296	0.09349	0.06728	0.06154	0.06965	0.07092	0.05089	0.03279	0.02792	0.02824	0.03288	0.03317	0.03447	0.03001	0.02422	0.02178	0.0087
64	0.0702	0.08209	0.08414	0.07985	0.0734	0.07649	0.07836	0.07872	0.07892	0.07809	0.06651	0.04627	0.03761	0.03591	0.04353	0.06043	0.06909	0.07359	0.06856	0.0494
64	0.05332	0.09461	0.09371	0.08929	0.05732	0.02126	0.01253	0.01015	0.01344	0.01168	0.01963	0.01951	0.00939	0.00814	0.01833	0.01932	0.0083	0.00738	0.01706	0.01648
64	0.02551	0.02743	0.01311	0.01157	0.01701	0.03446	0.0715	0.09031	0.04855	0.04007	0.0665	0.03428	0.01832	0.00877	0.00934	0.01527	0.01256	0.01404	0.01259	0.00621
64	0.07831	0.109	0.13969	0.1518	0.16159	0.13885	0.10683	0.10327	0.11394	0.12338	0.13184	0.1389	0.14423	0.1346	0.09502	0.06117	0.04738	0.03413	0.02524	0.0163
64	0.00232	0.00582	0.02755	0.05636	0.06334	0.04084	0.01961	0.01852	0.04414	0.03799	0.02813	0.05846	0.08723	0.08978	0.08671	0.04851	0.00767	0.00277	0.00452	0.0039

Table 6.19: The Normalized Dataset after Interpolation

As shown in Table 6.19, each resource is now presents as a array $R_{(1, 20)}$. Each array is assigned to a single class. Each class represents a unique combination of a project phase, resource, size cluster and duration cluster. To obtain the size and duration clusters, the M-means clustering technique from Weka is used to classify the total resource of hours and project durations into groups. The clustering results are shown in Tables 6.20 and 6.21.

Hours - All Data					
Cluster	N	%	Mean	StdDev	Prior Probability
Cluster0	221	40.48%	128.58	36.3162	0.3554
Cluster2	220	40.29%	291.38	111.7130	0.4201
Cluster1	89	16.30%	968.03	484.1862	0.1880
Cluster3	16	2.93%	2841.21	1,774.2157	0.0365
Total	546.00	100.00%	453.93	687.3120	1.00

Table 6.20: Clustering of Total Resource Hours

Weeks - All Data					
Cluster	N	%	Mean	StdDev	Prior Probability
Cluster0	360	65.93%	15.04	5.6524	0.5951
Cluster2	148	27.11%	29.85	9.6972	0.3083
Cluster1	38	6.96%	56.24	20.3446	0.0966
Total	546.00	100.00%	25.58	15.7160	1.00

Table 6.21: Clustering of Total Duration Weeks

A dynamic program that allows using the polynomial regression to develop a function that represents the variation of resource utilization per week is developed in R. Polynomial regression is used when a relation between a dependent variable Y and independent variable X cannot be fit to a linear or curvilinear such as logarithmic ($\text{Log}(X)$), power (X^b) or exponential (b^X) relationships, where b is a constant. As shown in the code below, the program reads the data from a Comma Separated Values (CSV) file and checks for the number of classes in the file. After that, a cycle is used to transpose the data of each group and assign it in an array that can be recognized by the R software. For each array, the “Fit” function is used to

obtain a polynomial regression function of the third degree that represents the data in each group. The function is in the format:

$$Y = b_0 + b_1 * (X) + b_2 * (X^2) + b_3 * (X^3) \quad [6.9]$$

The third degree polynomial, or sometimes referred to as cubic function, provides an S-curve, which fits reasonably well to the distribution of resource utilization over the project % complete. The output of the developed code is a list of the coefficients: b_0 , b_1 , b_2 and b_3 . The user can easily change the degree of the polynomial to any other degree using the function “PolDgr”. The goodness of fit is measured using the least square errors (R^2) and the user can try different functions to find the one that fits best for the dataset under investigation.

The output of the code is written to another CSV file and an example of it is shown in Table 6.22. The goodness of fit is tested using the R^2 function and graphically. The output for each class is plotted accompanied with the original values of any class to visually test the goodness of fit as shown in Figure 6.15.

At the beginning of any internal project, the project can decide on the size and duration class for each resource, use the characteristics of these classes accompanied with the polynomial function for the distribution of these resource utilization over project % complete to predict the initial planned values for each resource. These predicted values are based on PM judgment and historical data.

Group	b0	b1	b2	b3	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
					P00	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10
61	0.03646	0.44515	-0.9392	0.47088	0	0.05642	0.07205	0.08369	0.09168	0.0964	0.09819	0.09739	0.09438	0.08949	0.08309
62	0.05595	-0.0827	-0.0071	0.07006	0	0.05181	0.04768	0.04362	0.03969	0.03593	0.0324	0.02915	0.02622	0.02369	0.02159
64	0.22147	-0.9332	1.33796	-0.6206	0	0.17808	0.14091	0.1095	0.08338	0.0621	0.04517	0.03214	0.02254	0.01591	0.01178
65	0.12167	0.0576	-0.0232	-0.157	0	0.12447	0.12704	0.12925	0.131	0.13216	0.13262	0.13225	0.13095	0.12858	0.12505
66	0.00571	0.3515	-0.6502	0.30994	0	0.0217	0.03466	0.04485	0.05248	0.05779	0.06101	0.06237	0.06211	0.06046	0.05765
71	0.03005	0.02464	-0.0711	0.03333	0	0.0311	0.03183	0.03225	0.0324	0.03228	0.03194	0.03139	0.03066	0.02978	0.02876
72	0.12914	0.23537	-1.0044	0.67833	0	0.13848	0.14331	0.14414	0.14146	0.1358	0.12767	0.11756	0.10599	0.09347	0.08051
74	0.013	0.10858	0.33491	-0.2978	0	0.01923	0.02691	0.03582	0.04573	0.05643	0.06768	0.07926	0.09096	0.10255	0.1138
75	0.02104	0.00439	0.20323	-0.1888	0	0.02174	0.02332	0.02563	0.02854	0.03189	0.03555	0.03938	0.04323	0.04697	0.05044
76	0.0367	0.216	-0.2163	0.0314	0	0.04697	0.05617	0.06434	0.0715	0.07768	0.08289	0.08716	0.09051	0.09297	0.09456
81	0.00592	0.39438	-0.6569	0.29545	0	0.02404	0.03909	0.0513	0.06089	0.06808	0.0731	0.07616	0.07748	0.0773	0.07582
82	0.03564	-0.0609	0.54407	-0.4564	0	0.0339	0.03453	0.0372	0.04156	0.04728	0.054	0.06139	0.06911	0.0768	0.08414
84	0.03982	-0.1486	0.65913	-0.5018	0	0.03397	0.03104	0.03066	0.03244	0.03601	0.041	0.04703	0.05371	0.06068	0.06756
85	0.01738	0.1342	0.09617	-0.2218	0	0.02431	0.03154	0.03893	0.0463	0.05348	0.06031	0.06663	0.07226	0.07704	0.0808
86	0.03438	-0.0103	0.43175	-0.454	0	0.03489	0.03721	0.04102	0.04596	0.05169	0.05789	0.06419	0.07028	0.0758	0.08041
91	0.02108	0.15613	-0.2557	0.09549	0	0.02826	0.03423	0.03907	0.04284	0.04562	0.04748	0.04849	0.04872	0.04825	0.04714
92	0.0212	0.04282	0.17614	-0.2099	0	0.02376	0.02704	0.03088	0.03513	0.03964	0.04423	0.04877	0.05308	0.05701	0.06041
94	0.01992	0.26625	-0.3084	0.05304	0	0.03247	0.04351	0.0531	0.06126	0.06804	0.07347	0.07761	0.08048	0.08212	0.08258
95	0.0244	0.19691	-0.2891	0.07442	0	0.03353	0.04127	0.04768	0.05281	0.05672	0.05947	0.0611	0.06167	0.06125	0.05989
96	0.02095	0.0163	0.23396	-0.2686	0	0.02232	0.02465	0.02775	0.03142	0.03545	0.03964	0.0438	0.04771	0.05118	0.05401
101	0.01586	0.52732	-1.3385	0.8201	0	0.03898	0.05603	0.06761	0.07434	0.07685	0.07573	0.07161	0.06511	0.05683	0.0474
102	0.0205	0.19671	-0.4563	0.24289	0	0.02922	0.03585	0.04056	0.04353	0.04495	0.045	0.04386	0.04172	0.03875	0.03514
104	-0.0002	0.56418	-1.3467	0.79482	0	0.02472	0.04352	0.05679	0.06511	0.06908	0.06929	0.06635	0.06085	0.05339	0.04455
105	0.02886	-0.0122	0.22436	-0.2233	0	0.02878	0.02966	0.03133	0.03361	0.03635	0.03937	0.04251	0.0456	0.04847	0.05095
106	-0.0261	0.72905	-1.2376	0.54602	0	0.00735	0.035	0.05728	0.0746	0.08737	0.096	0.10089	0.10247	0.10113	0.09729
221	0.04911	0.20791	-0.4086	0.19002	0	0.05851	0.066	0.07174	0.07587	0.07852	0.07984	0.07997	0.07906	0.07724	0.07467
222	0.15758	-0.0957	-0.3868	0.37041	0	0.15187	0.14451	0.13577	0.12593	0.11526	0.10405	0.09257	0.08111	0.06993	0.05932

Table 6.22: An Example of the Coefficients Output

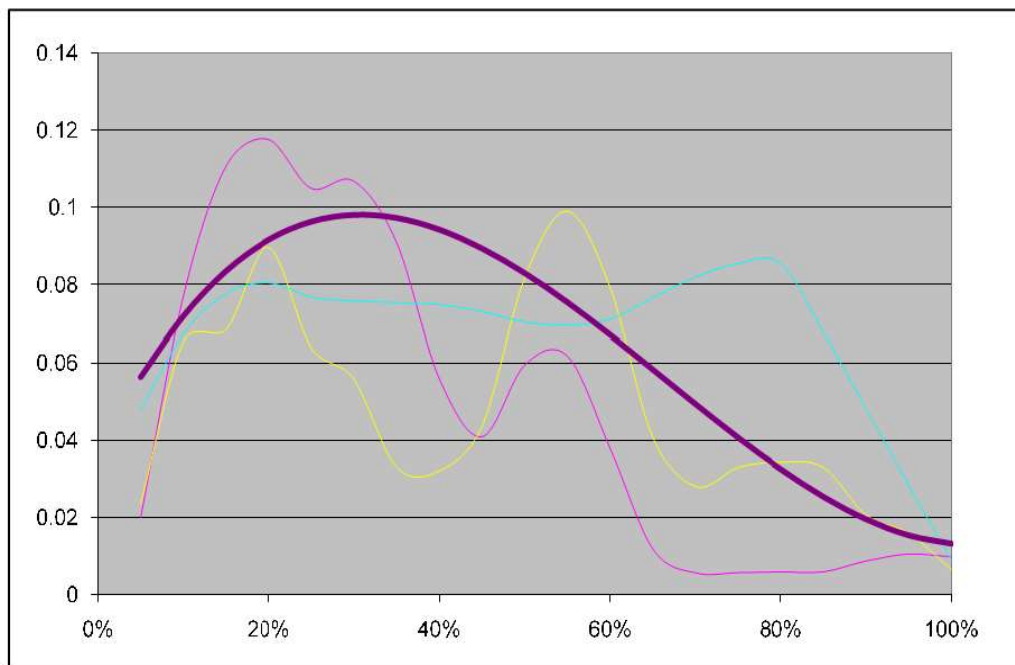


Figure 6.15: Example of Plotting the Polynomial Function vs. the Original Data

Another approach is to connect the averages of each % complete (P1, P2 to P20) as shown in Figure 6.16 below. It is up to the user to decide on which methodology fits better for the existing data. This case study is used to provide the user with the Initial Planned Values (IPV) that is needed prior to the detailed planning of any project.

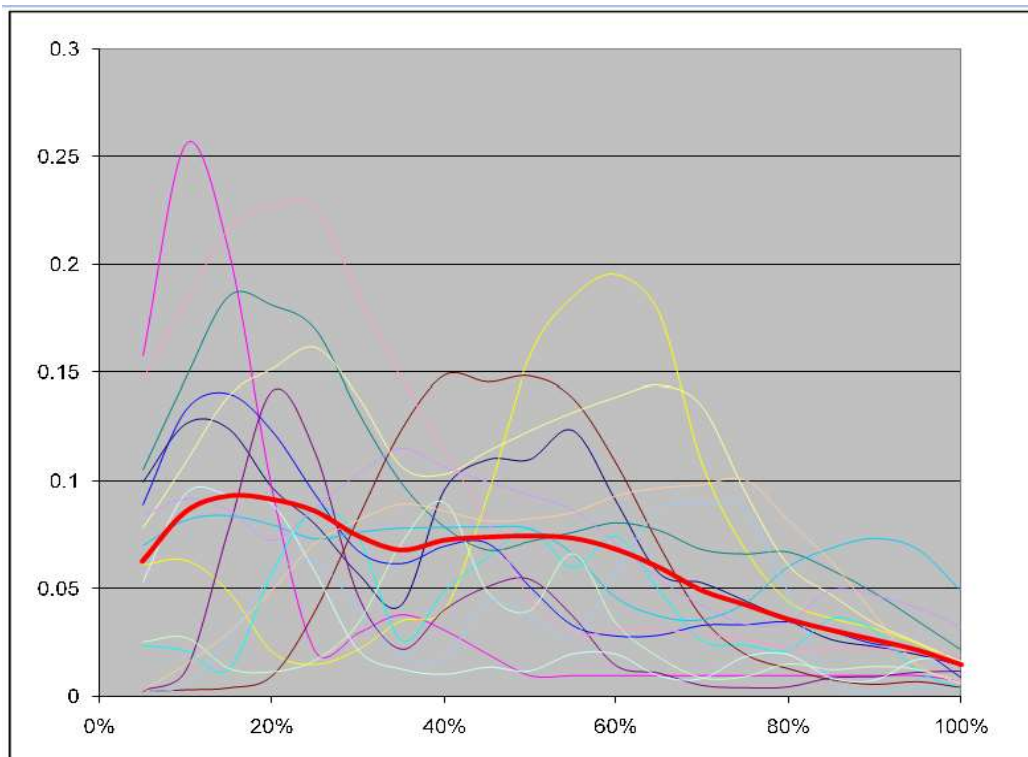


Figure 6.16: Connecting the Average Points to Represent the Data Class

In the third dataset, project attributes were not clearly identified when data was collected. Moreover, some of the projects were not broken into clearly defined phases as proposed in this research. When data was analyzed, discrepancies were found among the resource utilization graphs. These discrepancies were highlighted and recommendations were made to the partner company.

CHAPTER 7: CONCLUSION

7.1 RESEARCH SUMMARY

Industrial construction projects are known for their complexity and sophistication. A typical industrial construction project is planned and executed as a set of internal projects, each of which is performed by a contractor. Each of these contractors manages multiple internal projects simultaneously using one common pool of labour and non-labour resources. The very nature of working with one common pool of resources results in many challenges that face the management of projects and contracting companies. These contractors face the dilemma of completing projects with minimum cost and duration, while maintaining a steady flow of workload to sustain or increase resources capacity.

In addition to the difficulties of working with one common pool, managing labour resources is further complicated by economic instability and human factors. The former leads to drastic changes in the supply and demand for skilled labour; and the latter leads to difficulty in predicting labour productivity. Improper management of labour resources causes severe budget overruns, significant schedule delays, damaged reputation, intolerable level of stress and decreased productivity, profit and client satisfaction.

To address these issues, the aim of this research is to improve resources management practices by using existing historical data from completed projects to forecast needs of future projects.

During the process of managing labour resources in multiple-project environment, a large amount of multidimensional data is generated, collected and stored in scattered formats. Currently, there is no consistent methodology to manage this wealth of data. Most of this data gets lost and is never viewed, analyzed or transferred to useful knowledge that could be an asset in improving resource management practices.

This research developed an integrated framework for managing resources data in multiple-project environment. The framework is built on a KDD model to transfer the collected multidimensional historical data from completed projects to useful knowledge for new projects.

The integrated framework would fulfill many purposes. First, it includes a consistent methodology for generating, collecting and storing labour resources data in a structured format ready for data mining. Second, it prevents this data from being lost. Third, it saves the time and efforts that are required to clean, prepare, and pre-process unstructured data. Fourth, it uses data mining to transfer the collected data into useful knowledge. This accumulated knowledge is expected to improve resource management practices in industrial construction projects. It also increases the ability to forecast work load and optimum staffing capacity. As a result of this improved resource management practices, projects would be less prone to schedule delays and budget overruns, and the contractors would be more efficient and profitable, and hence could complete strongly.

First, a comprehensive literature review that covered previous research in the areas of project resource management, transferring projects' data to useful knowledge, resource forecasting techniques in projects, data warehousing and OLAP reporting and data mining techniques. This step is explained in chapter two of this thesis.

Second, a comprehensive analysis of the industrial construction projects domain was performed. The analysis was performed in both vertical (within an industrial project) and horizontal (across multiple projects) to obtain full understanding of the domain. The output of the vertical analysis included the identification of the seven main objects that have to be modeled in the data warehouse. The output of the horizontal analysis included the identification of the control attributes that have to be used by all projects for proper data mining. The output also included a set of illustrated models for the main processes of industrial construction. This analysis is explained in chapter three of this thesis.

Third, a comprehensive analysis of the issues with current practices of labour resources data management in industrial construction projects was performed. To overcome these issues, the concept of predefined progressable work packages is presented. These packages are used as building blocks, common denominator and knowledge carrier to manage the multiple dimensions of resources data. This analysis is explained in chapter four of this thesis.

Fourth, a prototype data warehouse was built to centrally store the data. OLAP reporting techniques were used to present the data according to the various needs of end-users. A knowledge exchange tool was also developed to exchange knowledge elements between projects and individuals. The data warehouse is presented in chapter five of this thesis.

Fifth, three case studies were performed to validate the applicability of the developed framework to real projects data. The first dataset was obtained from a large EPCM firm and was utilized to define the distribution parameters of estimating unit costs. An anomaly detection methodology was developed to highlight the inconsistent data points for the end-user. A unit cost tree with 225 branches was obtained. PostHoc tests and the One-way ANOVA were used to classify the cost units into smaller number of groups.

The second dataset was obtained from a large EPC firm and was used to define the distribution parameters of estimating unit durations within different data clusters. The dataset was randomly divided into training set and testing set for validation purposes. More than 85% of the testing data points had an estimating error of less than 25%.

The third dataset was used to analyze various resources utilization patterns over time units and to find the most fitting resources utilization curve per cluster.

The data used in this research already existed in scattered format between planning and timekeeping systems. It was collected mainly for three main purposes: making payment for individuals, producing schedules and keeping track of projects progress. Once projects were finished, this data was stored away without any further analysis. This research aimed to learn from this data.

By studying the original dataset, several problems were identified. These problems are mainly pertaining to the lack of a proper definition of data dimensions, objects and attributes and to the lack of a systematic consistent integrated approach to data collection and storage. There is a perception in the industry that each project is unique and its data is unique as well, and therefore, data from projects are not easily aggregated nor transferred to useful knowledge.

By implementing the data collection integrated framework to the original data set, this research demonstrated that data can be collected in a systematic and consistent manner, which then could be analysed in a variety of ways, and then leads to extracting useful knowledge that would improve labour resources management practices and forecasts. As a result of this framework, productivity and efficiency would increase. As well, a continuous knowledge cycle and a self-learning loop would be established between completed and future projects.

7.2 RESEARCH CONTRIBUTIONS

7.2.1 Academic Contribution

Academically, this research pioneers the use of a complete KDD model to labour resources management in industrial construction projects. KDD has been widely used in other research areas mainly in the fields of computer sciences, business administration and finance. By applying KDD in these fields, researchers were able to extract useful knowledge out of collected unutilized data. However, KDD was not fully introduced to the construction management research because of the projectized nature of the available data. Most previous researchers believed that since projects are temporary and unique, collected data from projects cannot be transferred to useful knowledge. Therefore, the general assumption was that KDD would not work with such data. This research paved the way for modeling projects data, regardless of project type, size or location, in a structured consistent format that can be easily used for data mining and KDD. The extracted knowledge is expected to open the door for further research trying to analyze and understand the findings of the KDD procedure.

Through this KDD model, new insights into labour resource management were brought to light. This model allowed discovering a wealth of knowledge from already existing historical data that has not been used nor analyzed before. By standardizing the generation, collection, and codification of data, the model introduced consistency to the data management procedure regardless of the project. This consistency is essential in learning from data.

The structured format is generic in a way that it does not enforce a certain data mining technique. Clustering, classification, finding association rules and other data mining techniques could all be applied to the collected data; and hence allowing new window for future research. Furthermore, the structured data represents a platform that can be also used by many other research areas such as Special Purpose Simulation (SPS), Fuzzy Sets, High Level Architecture (HLA) simulation or Artificial Neural Networks (ANN).

The developed approach also provided a single integrated probabilistic solution to the sophisticated, complex and multi-dimensional problem of forecasting multiple-resources needs in a multiple-project environment. The proposed solution deals with project's scope, schedule, costs, resources, historical data, project changes and uncertainties. Most previous research focused only on one of these aspects of projects. Few studies tried introducing an integrated solution.

The proposed approach is also a multidisciplinary solution that uses knowledge from various fields of research to provide an integrated solution. It feeds from computer sciences, machine learning, construction and project management, data mining and knowledge management sciences. With the increased complexity of construction research problems, this multidisciplinary approach is becoming essential for researchers to apply knowledge from other research fields into the field of construction management.

The proposed framework is not only a theoretical model that can be applied in academic research, but it can also be applied by the industry. Any industrial owner or contractor can adopt the proposed framework and utilize it to mine for useful knowledge in their collected data.

This research used data collected from large number of real projects to test its applicability. Most previous research relied on using data from theoretical projects that do not mimic real projects. Many previous researches were limited in their ability to deal with real projects data due to calculations complexity and long computing times. Meanwhile the proposed framework is capable of handling data from large scale projects and the more data stored in the system, the better results are obtained.

The results show, with high significance, improved resource estimates that are realistic and representative of future resource needs. By transferring the discovered knowledge from completed projects to future projects, the model acts as a closed-cycle self-learning approach that improves automatically as more data becomes available and added to the model. Thus, whenever a project is completed, its data is added to the data warehouse and it serves to enrich the outcome of future project forecasts.

7.2.2 Industry Contribution

This research was an eye opener for partner companies that showed them the value of their unutilized data. The implementation of this KDD model in the industry is expected to improve estimating practices and enable contractors to use more realistic, customized and updatable estimating units that are based on historical data not gut feelings. Therefore, it would increase efficiency, productivity and profitability.

This KDD model transfers existing data in timekeeping and scheduling systems to a wealth of knowledge. The extracted knowledge provides forecasts of resource needs and future workload from all projects in a contracting company. Having aggregated forecasts of expected workload, contractors are provided with a probabilistic tool to run different scenarios to determine the optimum staff capacity and to maximize their efficiency and profitability.

Applying the proposed framework is expected to make positive changes to the organization culture. First, all team members would pay more attention collecting projects data knowing that this data would be analyzed and mined. Second, schedulers, cost estimators and cost controllers are expected to work together in a more integrated fashion. Third, it would foster a culture of knowledge sharing not hiding and learning from past experience not repeating the same mistakes over and over again.

7.3 RECOMMENDATIONS FOR FUTURE RESEARCH

The developed KDD model was implemented into the management of labour resources data in industrial construction projects domain. Further research can be carried out to investigate the feasibility of applying this model to other non-labour resources types. In addition, other researchers can investigate extending the application of this model to other domains such as infrastructure or commercial construction.

Clustering and anomaly detection data mining techniques were used to extract knowledge from the available datasets. Future research can apply other data mining techniques or knowledge discovery techniques such as classification, finding association rules, simulation, artificial neural networks, and fuzzy sets. The data warehouse would provide a systematic methodology to model projects, their objects and projects' data for analysis by these sophisticated research methods.

Once populated with enough data, the data warehouse along with advanced research techniques can be used to identify the main factors impacting labour resources performance and overall project performance.

BIBLIOGRAPHY

Abeyasinghe, M. C. L., Greenwood, D. J., and Johansen, D. E. (2001). "Efficient method for scheduling construction projects with resource constraints." *Int.J.Project Manage.*, 19(1), 29-45.

AbouRizk, S., and Mohamed, Y. (2002). "Optimal construction project planning." *Proceedings of the 2002 Winter Simulation Conference*, IEEE, San Diego, CA, USA, 1704-8.

AbouRizk, S. 1., and Mohamed, Y. 1. (2000). "Symphony-an integrated environment for construction simulation." *Proceedings of WSC 2000, Winter Simulation Conference*, IEEE, Orlando, FL, USA, 1907-14.

AbouRizk, S. M. (1993). "Stochastic simulation of construction bidding and project management." *Second Canadian Conference on Computing in Civil Engineering*, Ottawa, Ont., Canada, 343-53.

AbouRizk, S. M., and Van Tol, A. A. (2006). "Simulation modeling decision support through belief networks." *Simulation Modelling Practice and Theory*, 14(5), 614-40.

Abourizk, S. M., and Sawhney, A. (1993). "Subjective and interactive duration estimation." *Canadian Journal of Civil Engineering*, 20(3), 457-470.

Aguilar-Saven, R. S. (2004). "Business process modelling: review and framework." *Int J Prod Econ*, 90(2), 129-49.

Ahmad, I., Azhar, S., and Lukauskis, P. (2004). "Development of a decision support system using data warehousing to assist builders/developers in site selection." *Automation in Construction*, 13(4), 525-42.

Al-Ghassani, A. M., Kamara, J. M., Anumba, C. J., and Carrillo, P. M. (2004). "An innovative approach to identifying knowledge management problems." *Engineering, Construction and Architectural Management*, 11(5), 349-357.

Al-Jibouri, S., Mawdesley, M., Scott, D., and Gribble, S. (2005). "The use of a simulation model as a game for teaching management of projects in construction." *International Journal of Engineering Education*, 21(6), 1195-202.

Alberta Economic Development Authority (2004). "Mega Project Excellence: Preparing for Alberta's Legacy, An Action Plan"

Amor, R., and Faraj, I. (2001). "Misconceptions about integrated project databases." *Electron.J.Inf.Technol.Constr.*, 6.

Anumba, C. J., Egbu, C. O., and Carrillo, P. M. (2005). "Knowledge management in construction." 226.

Aouad, G., Child, T., Marir, F., and Brandon, P. (1997). "Developing a virtual reality interface for an integrated project database environment." *97TB100165*), IEEE Comput. Soc, London, UK, 192-7.

Ayyub, B. M., Gupta, M. M., International Symposium on Uncertainty Modelling and Analysis (2nd : 1993 : University of Maryland at College Park), and International Symposium on Uncertainty Modelling and Analysis. (1994). *Uncertainty modelling and analysis : theory and applications*. New York ; Elsevier, Amsterdam.

Baccarini, D. (1999). "The logical framework method for defining project success." *Project Management Journal*, 30(4), 25.

Badiru, A. B., Badiru, A., and Badiru, A. (2008). *Industrial project management : concepts, tools, and techniques*. CRC Press, Boca Raton.

Baragona, R., and Battaglia, F. (2007). "Outliers detection in multivariate time series by independent component analysis." *Neural Comput.*, 19(7), 1962-84.

Barkley, B. (2006). *Integrated project management*. McGraw-Hill, New York.

Barnett, V., and Lewis, T. (1995). *Outliers in statistical data*. Wiley, Chichester; New York.

Barraza, G. A., Back, W. E., and Mata, F. (2000). "Probabilistic monitoring of project performance using SS-curves." *J.Constr.Eng.Manage.*, 126(2), 142-148.

Bassioni, H. A., Price, A. D. F., and Hassan, T. M. (2004). "Performance measurement in construction." *J.Manage.Eng.*, 20(2), 42-50.

Ben-David, A., Dvir, D., Sadeh, A., and Shenhar, A. J. (2006). "Critical managerial factors affecting defense projects success: A comparison between neural network and regression analysis." *Eng Appl Artif Intell*, 19(5), 535-43.

Benjamins, V. R., Fensel, D., and Perez, A. G. (1998). "Knowledge management through ontologies." *Second International Conference on Practical Aspects of Knowledge Management*, Swiss Life, Basel, Switzerland, 5-1.

Bhatt, G. D. (2002). "Management strategies for individual knowledge and organizational knowledge." *Journal of Knowledge Management*, 6(1), 31.

Boskers, N. D., and Abourizk, S. M. (2005). "Modeling scheduling uncertainty in capital construction projects." *2005 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers Inc., New York, NY 10016-5997, United States, Orlando, FL, United States, 1500-1507.

Bouchlaghem, D., Kimmance, A. G., and Anumba, C. J. (2004). "Integrating product and process information in the construction sector." *Industrial Management and Data Systems*, 104(3), 218-233.

- Bresnen, M., Edelman, L., Newell, S., Scarbrough, H., and Swan, J. (2003). "Social practices and the management of knowledge in project environments." *Int.J.Project Manage.*, 21(3), 157-66.
- Carlile, P. R., and Reberich, E. S. (2003). "Into the black box: the knowledge transformation cycle." *Management Science*, 49(9), 1180-95.
- Carrillo, J. E., and Gaimon, C. (2004). "Managing knowledge-based resource capabilities under uncertainty." *Management Science*, 50(11), 1504-1518.
- Carrillo, P. (2005). "Lessons learned practices in the engineering, procurement and construction sector." *Engineering Construction and Architectural Management*, 12(3), 236-50.
- Carrillo, P., and Chinowsky, P. (2006). "Exploiting knowledge management: The engineering and construction perspective." *J.Manage.Eng.*, 22(1), 2-10.
- Chan, A. P. C., Scott, D., and Chan, A. P. L. (2004). "Factors affecting the success of a construction project." *J.Constr.Eng.Manage.*, 130(1), 153-155.
- Chan, C. W. (2004). "From knowledge modeling to ontology construction." *International Journal of Software Engineering and Knowledge Engineering*, 14(6), 603-624.
- Chan, W., Chua, D., and Kannan, D. (1996). "Construction Resource Scheduling with Genetic Algorithms" *Journal of Construction Engineering and Management*, 122(2), 125.
- Chang, S., and Ahn, J. (2005). "Product and process knowledge in the performance-oriented knowledge management approach." *Journal of Knowledge Management*, 9(4), 114.

Chau, K. W., Cao, Y., Anson, M., and Zhang, J. (2002). "Application of data warehouse and decision support system in construction management." *Autom. Constr.*, 12(2), 213-24.

Chehayeb, N. N. (1996). "Simulation-based project control." PhD thesis, University of Alberta (Canada), Canada.

Chehayeb, N. N., and AbouRizk, S. M. (1995). "Applications of simulation in progress reporting and control." *Proceedings of the 1995 Winter Simulation Conference, WSC'95*, IEEE, Piscataway, NJ, USA, Arlington, VA, USA, 1009-1016.

Chen, P. (2008). "Integration of cost and schedule using extensive matrix method and spreadsheets." *Autom. Constr.*, 18(1), 32-41.

Cheng, M., Tsai, M., and Xiao, Z. (2006). "Construction management process reengineering: Organizational human resource planning for multiple projects." *Autom. Constr.*, 15(6), 785-799.

Chira, O., Chira, C., Roche, T., Tormey, D., and Brennan, A. (2006). "An agent-based approach to knowledge management in distributed design." *J.Intell.Manuf.*, 17(6), 737-50.

Choy, E., and Ruwanpura, J. Y. (2007). "Predicting construction productivity using situation-based simulation models." *Canadian Journal of Civil Engineering*, 33(12), 1585-600.

Christensen, L. C., Christiansen, T. R., Jin, Y., Kunz, J., and Levitt, R. E. (1999). "Modeling and simulating coordination in projects." *Journal of Organizational Computing and Electronic Commerce*, 9(1), 33-55.

Christensen, L. C., Christiansen, T. R., Jin, Y., Kunz, J., and Levitt, R. E. (1997). "Object-oriented enterprise modeling and simulation of AEC projects." *Microcomput.Civil Eng.*, 12(3), 157-170.

Chung-Yee Lee¹, and Lei, L. (2001). "Multiple-project scheduling with controllable project duration and hard resource constraint: some solvable cases." *Annals of Operations Research*, 102 287-307.

Cios, K. J. (2007). *Data mining : a knowledge discovery approach*. Springer, New York.

Codd, E. F., Codd, S. B., and Salley, C. T. (1993). *Providing OLAP (On-line Analytical Processing) to user-analysts: an IT mandate*. Codd and Date, San Jose, CA, USA.

Cody, W. F., Kreulen, J. T., Krishna, V., and Spangler, W. S. (2002). "The integration of business intelligence and knowledge management." *IBM Syst J*, 41(4), 697-713.

Collopy, F. 1., and Armstrong, J. S. (1992). "Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations." *Management Science*, 38(10), 1394-414.

Construction Industry Institute. Cost/Schedule, and Controls Task Force. (1988). *Work packaging for project control*. The Institute, Austin, Tex.

Creswell, J. W., Qualitative inquiry and research, and Design. (2007). *Qualitative inquiry & research design : choosing among five approaches*. Sage Publications, Thousand Oaks.

Crowder, R., and Sim, Y. (2004). "An approach to extracting knowledge from legacy documents." *2004 ASME Design Engineering Technical Conferences and Computers and Information in Engineering Conference, September 28, 2004 - October 2*, American Society of Mechanical Engineers, Salt Lake City, UT, United states, 253-259.

Cuadrado-Gallego, J. J. 1., Fernandez-Sanz, L., and Sicilia, M. -. (2006). "Enhancing input value selection in parametric software cost estimation models through second level cost drivers." *Software Quality Journal*, 14(4), 339-57.

Davenport, T. H., Prusak, L., and Agl_com. (2000). *Working knowledge : how organizations manage what they know*. Harvard Business School Press, c1998, Boston, MA.

Dean, B. V. 1., Denzler, D. R. 1., and Watkins, J. J. (1992). "Multiproject staff scheduling with variable resource constraints." *IEEE Trans.Eng.Manage.*, 39(1), 59-72.

Deckro, R. F. 1., Winkofsky, E. P., Hebert, J. E., and Gagnon, R. (1991). "A decomposition approach to multi-project scheduling." *Eur.J.Oper.Res.*, 51(1), 110-18.

Demian, P., and Fruchter, R. (2006). "An ethnographic study of design knowledge reuse in the architecture, engineering, and construction industry." *Research in Engineering Design*, 16(4), 184-95.

Desouza, K. C. (2003). "Barriers to effective use of knowledge management systems in software engineering." *Commun ACM*, 46(1), 99-101.

Diekmann, J. E., and Al-Tabtabai, H. (1992). "Knowledge-based approach to construction project control." *Int.J.Project Manage.*, 10(1), 23-30.

Disterer, G. (2002). "Management of project knowledge and experiences." *Journal of Knowledge Management*, 6(5), 512.

Doloi, H. (2007). "Twinning motivation, productivity and management strategy in construction projects." *EMJ - Engineering Management Journal*, 19(3), 30-40.

Dong, T., Tong, R., Zhang, L., and Dong, J. (2007). "A knowledge-based approach to assembly sequence planning." *Int J Adv Manuf Technol*, 32(11), 1232-1244.

Duc Thanh Luu, Ng, S. T., and Swee Eng Chen. (2003). "A case-based procurement advisory system for construction." *Adv.Eng.Software*, 34(7), 429-38.

Duen-Ren Liu, and Hsu, C. (2004). "Project-based knowledge maps: combining project mining and XML-enabled topic maps." *Internet Res.: Electron.Networking Appl.Policy*, 14(3), 254-66.

Dumond, J. (1992). "In a multi-resource environment, how much is enough?" *Int J Prod Res*, 30(2), 395-410.

Dumond, J., and Mabert, V. A. (1988). "EVALUATING PROJECT SCHEDULING AND DUE DATE ASSIGNMENT PROCEDURES: AN EXPERIMENTAL ANALYSIS." *Management Science*, 34(1), 101-118.

Duverlie, P. 1., and Castelain, J. M. 1. (1999). "Cost estimation during design step: parametric method versus case based reasoning method." *Int J Adv Manuf Technol*, 15(12), 895-906.

Egbu, C. O. (2004). "Managing knowledge and intellectual capital for improved organizational innovations in the construction industry: an examination of critical success factors." *Engineering Construction and Architectural Management*, 11(5), 301-15.

Ei-Diraby, T. A., Lima, C., and Feis, B. (2005). "Domain taxonomy for construction concepts: Toward a formal ontology for construction knowledge." *J.Comput.Civ.Eng.*, 19(4), 394-406.

Eldin, N. N. (2005). "The effect of early freezing of scope on project schedule." *Cost Engineering (Morgantown, West Virginia)*, 47(2), 12-18.

El-Diraby, T. E., and Kashif, K. F. (2005). "Distributed ontology architecture for knowledge management in highway construction." *J.Constr.Eng.Manage.*, 131(5), 591-603.

El-Rayes, K., and Moselhi, O. (2001). "Optimizing resource utilization for repetitive construction projects." *J.Constr.Eng.Manage.*, 127(1), 18-27.

Elwany, M. H., Korish, I. E., Aly Barakat, M., and Hafez, S. M. (1998). "Resource smoothening in repetitive projects." *Computers & Industrial Engineering*, 35(3-4), 415-418.

Engineering News Record (ENR) (2007) "Top international design firms",

Website Address:

http://enr.construction.com/people/topLists/topIntlDesign/topIntlDesign_1-50.asp

Engwall, M., and Jerbrant, A. (2003). "The resource allocation syndrome: the prime challenge of multi-project management?" *International Journal of Project Management*, 21(6), 403-409.

Ester, M., Kriegel, H. -, Sander, J., and Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, AAAI Press, Menlo Park, CA, USA, 226-31.

Fabi, B., and Pettersen, N. (1992). "Human resource management practices in project management." *International Journal of Project Management*, 10(2), 81-88.

Fan, H. (2007). "Leveraging operational data for intelligent decision support in construction equipment management." PhD thesis, University of Alberta (Canada), Canada.

Fan, H., AbouRizk, S., Kim, H., and Zaiane, O. (2008). "Assessing residual value of heavy construction equipment using predictive data mining model." *J.Comput.Civ.Eng.*, 22(3), 181-191.

Fan, H., Kim, H., AbouRizk, S., and Seung Heon Han. (2008). "Decision support in construction equipment management using a nonparametric outlier mining algorithm." *Expert Syst.Appl.*, 34(3), 1974-82.

- Faraj, I., and Alshawi, M. (1999). "A modularised integrated computer environment for the construction industry: SPACE." *Electron.J.Inf.Technol.Constr.*, 4.
- Fatemi Ghomi, S. M. T., and Ashjari, B. (2002). "A simulation model for multi-project resource allocation." *Int.J.Project Manage.*, 20(2), 127-130.
- Fayek, A. R., Revay, S. O., Rowan, D., and Mousseau, D. (2006). "Assessing performance trends on industrial construction mega projects." *Cost Engineering (Morgantown, West Virginia)*, 48(10), 16-21.
- Fayyad, U. M. (1996). "Advances in knowledge discovery and data mining." 611.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From data mining to knowledge discovery in databases." *AI Magazine*, 17(3), 37.
- Fendley, L. G. (1968). "Toward development of complete multiproject scheduling system." *Industrial Engineering*, 19(10), 505-515.
- Feniosky Pena-Mora, Sang Hyun Lee, and Park, M. (2006). "Dynamic planning and control methodology for strategic and operational construction project management." *Autom.Constr.*, 15(1), 84-97.
- Fiori, C., and Kovaka, M. (2005). "Defining megaprojects: Learning from construction at the edge of experience." *Construction Research Congress 2005: Broadening Perspectives - Proceedings of the Congress*, American Society of Civil Engineers, Reston, VA 20191-4400, United States, San Diego, CA, United States, 715-724.
- Fitsilis, P., Gerogiannis, V., and Kameas, A. (2006). "Extracting and maintaining project knowledge using ontologies." *Proceedings of the 1st International Workshop on Technologies for Collaborative Business Process Management-TCoB 2006*, INSTICC Press, Paphos, Cyprus, 85-96.

Forde, B. W. R., Russell, A. D., and Stiemer, S. F. (1989). "Object-oriented knowledge frameworks." *Eng Comput (New York)*, 5(2), 79-89.

Froese, T. (2006). "Emerging information and communication technologies and the discipline of project information management." *Revised Selected Papers*, Springer-Verlag, Ascona, Switzerland, 230-40.

Froese, T. (2003). "Future directions for IFC-based interoperability." *Electron.J.Inf.Technol.Constr.*, 8.

Froese, T., Fischer, M., Grobler, F., Ritzenthaler, J., Yu, K., Sutherland, S., Staub, S., Akinci, B., Akbas, R., Koo, B., Barron, A., and Kunz, J. (1999). "Industry foundation classes for project management-a trial implementation." *Electron.J.Inf.Technol.Constr.*, 4.

Froese, T. M., and Paulson, B. C., Jr. (1994). "OPIS: an object model-based project information system." *Microcomput.Civil Eng.*, 9(1), 13-28.

Froese, T. (1995). "Models of construction process information." *Part 1 (of 2)*, ASCE, New York, NY, USA, Atlanta, GA, USA, 5-12.

Fruchter, R., and Demian, P. (2002). "CoMem: Designing an interaction experience for reuse of rich contextual knowledge from a corporate memory." *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM*, 16(3), 127-147.

Fruchter, R., Demian, P., Yin, Z., and Luth, G. (2003). "Turning A/E/C Knowledge into Working Knowledge." *Towards a Vision for Information Technology in Civil Engineering: Proceedings of the Fourth Joint International Symposium on Information Technology in Civil Engineering, Nov 15-16 2003*, American Society of Civil Engineers, Reston, VA 20191-4400, United States, Hashville, TN, United States, 143-155.

Fuller, M. A., Valacich, J. S., and George, J. F. (2008). *Information systems project management : a process and team approach*. Pearson Prentice Hall, Upper Saddle River, NJ.

Gallego, J. J. C., Rodriguez, D., Sicilia, M. A., Rubio, M. G., and Crespo, A. G. (2007). "Software project effort estimation based on multiple parametric models generated through data clustering." *Journal of Computer Science and Technology*, 22(3), 371-378.

Gamst, G., Meyers, L. S., and Guarino, A. J. (2008). *Analysis of variance designs : a conceptual and computational approach with SPSS and SAS*. Cambridge University Press, Cambridge; New York.

Georgy, M. E. (2008). "Evolutionary resource scheduler for linear projects." *Automation in Construction*, 17(5), 573-583.

Georgy, M. E., Chang, L., and Zhang, L. (2005). "Prediction of engineering performance: A neurofuzzy approach." *J.Constr.Eng.Manage.*, 131(5), 548-557.

Georgy, M. E., Chang, L., and Zhang, L. (2005). "Utility-function model for engineering performance assessment." *J.Constr.Eng.Manage.*, 131(5), 558-568.

Giovinazzo, W. A. (2000). *Object-oriented data warehouse design : building a star schema*. Prentice Hall, Upper Saddle River, NJ.

Gonçalves, J. F., Mendes, J. J. M., and Resende, M. G. C. "A genetic algorithm for the resource constrained multi-project scheduling problem." *European Journal of Operational Research*, In Press, Corrected Proof.

Gunasekaran, A. (2007). *Modeling and analysis of enterprise information systems*. IGI Pub., Hershey, PA.

Gunduz, M., and Hanna, A. S. (2005). "Benchmarking change order impacts on productivity for electrical and mechanical projects." *Build. Environ.*, 40(8), 1068-1075.

Gupta, V. K., Chen, J. G., and Murtaza, M. B. (1997). "A learning vector quantization neural network model for the classification of industrial construction projects." *Omega*, 25(6), 715-27.

Hai Chen Tan, Carrillo, P. M., Anumba, C., Bouchlaghem, N., Kamara, J. M., and Udeaja, C. E. (2007). "Development of a methodology for live capture and reuse of project knowledge in construction." *J.Manage.Eng.*, 23(1), 10-26.

Hajjar, D., Abourizk, S., and Hunka, D. (1999). "Improved project control through advanced data acquisition technologies." .

Hajjar, D., and AbouRizk, S. (1999). "Symphony: an environment for building special purpose construction simulation tools." *1999 Winter Simulation Conference Proceedings (WSC)*, IEEE, Piscataway, NJ, USA, Phoenix, AZ, USA, 998-1006.

Hajjar, D., and AbouRizk, S. M. (2002). "Unified modeling methodology for construction simulation." *J.Constr.Eng.Manage.*, 128(2), 174-185.

Halkidi, M., and Vazirgiannis, M. (2001). "Clustering validity assessment: finding the optimal partitioning of a data set." *Proceedings 2001 IEEE International Conference on Data Mining*, IEEE Comput. Soc, San Jose, CA, USA, 187-94.

Han, J., Fu, Y., Wang, W., Chiang, J., Zaiane, O. R., and Koperski, K. (1996). "DBMiner: Interactive mining of multiple-level knowledge in relational databases." *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, Montreal, Can, 550.

Han, J., and Kamber, M. (2006). *Data mining : concepts and techniques*. Morgan Kaufmann; Elsevier Science distributor, San Francisco, Calif; Oxford.

- Hanlon, E. J., and Sanvido, V. E. (1995). "Constructability information classification scheme." *J.Constr.Eng.Manage.*, 121(4), 337-345.
- Hari, S., Egbu, C., and Kumar, B. (2005). "A knowledge capture awareness tool: An empirical study on small and medium enterprises in the construction industry." *Engineering Construction and Architectural Management*, 12(6), 533-67.
- Haugen, K. K. (1996). "Stochastic dynamic programming model for scheduling of offshore petroleum fields with resource uncertainty." *Eur.J.Oper.Res.*, 88(1), 88-100.
- Hegazy, T., and Ersahin, T. (2001). "Simplified spreadsheet solutions. II: Overall schedule optimization." *J.Constr.Eng.Manage.*, 127(6), 469-475.
- Hegazy, T., and Ersahin, T. (2001). "Simplified spreadsheet solutions: I: Subcontractor information system." *J.Constr.Eng.Manage.*, 127(6), 461-468.
- Hegazy, T., and Wassef, N. (2001). "Cost optimization in projects with repetitive nonserial activities." *J.Constr.Eng.Manage.*, 127(3), 183-191.
- Hegazy, T. (1999). "Optimization of resource allocation and leveling using genetic algorithms." *J.Constr.Eng.Manage.*, 125(3), 167-175.
- Hegazy, T., and Kassab, M. (2003). "Resource optimization using combined simulation and genetic algorithms." *J.Constr.Eng.Manage.*, 129(6), 698-705.
- Hegazy, T., and Petzold, K. (2003). "Genetic optimization for dynamic project control." *J.Constr.Eng.Manage.*, 129(4), 396-404.
- Hegazy, T., Shabeeb, A. K., Elbeltagi, E., and Cheema, T. (2000). "Algorithm for scheduling with multiskilled constrained resources." *J.Constr.Eng.Manage.*, 126(6), 414-421.

Henninger, S. (1995). "Developing domain knowledge through the reuse of project experiences." *Proceedings of the ACM SIGSOFT Symposium on Software Reusability, Apr 28-30 1995*, ACM, New York, NY, USA, Seattle, WA, USA, 186-195.

Hewage, K.N., Jergeas G.F. and Ruwanpura, J.Y (2008). "IT usage in Alberta's building construction projects: current status and challenges" *Automation in Construction*, v 17, n 8, p 940-7.

Hiremath, H. R., and Skibniewski, M. J. (2004). "Object-oriented modeling of construction processes by unified modeling language." *Autom. Constr.*, 13(4), 447-68.

Ho, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. Chapman & Hall/CRC, Boca Raton.

Howson, C. (2008). *Successful business intelligence : secrets to making BI a killer app*. McGraw-Hill, New York.

Huang, J. C., and Newell, S. (2003). "Knowledge integration processes and dynamics within the context of cross-functional projects." *Int.J.Project Manage.*, 21(3), 167-176.

Huemann, M., Keegan, A., and Turner, J. R. (2007). "Human resource management in the project-oriented company: A review." *International Journal of Project Management*, 25(3), 315-323.

Hyari, K. 1., and El-Rayes, K. (2006). "Optimal planning and scheduling for repetitive construction projects." *J.Manage.Eng.*, 22(1), 11-19.

Industrial Reports Inc., (2008). Website Address: <http://industrialreports.com/>

Inmon, W. H. (2005). *Building the data warehouse*. Wiley, Indianapolis, IN.

Inmon, W. H., and Nesavich, A. (2008). *Tapping into unstructured data : integrating unstructured data and textual analytics into business intelligence*. Prentice Hall, Upper Saddle River, NJ.

Inmon, W. H., O'Neil, B. K., and Fryman, L. (2008). *Business metadata : capturing enterprise knowledge*. Elsevier/Morgan Kaufmann, Amsterdam; Boston.

Ishida, K., and Kitagawa, H. (2008). "Detecting current outliers: Continuous outlier detection over time-series data streams." *19th International Conference on Database and Expert Systems Applications, DEXA 2008*, Springer Verlag, Heidelberg, D-69121, Germany, Turin, Italy, 255-268.

Isidore, L. J., and Back, W. E. (2002). "Multiple simulation analysis for probabilistic cost and schedule integration." *J.Constr.Eng.Manage.*, 128(3), 211-219.

Jennex, M. E. (2007). *Knowledge management in modern organizations*. Idea Group Pub., Hershey, PA.

Jergeas, G. (2008). "Analysis of the front-end loading of Alberta mega oil sands projects" *Project Management Journal*. Vol. 39, no. 4, p. 95-104.

Jergeas, G., and Der Put, J. V. (2001). "Benefits of constructability on construction projects." *J.Constr.Eng.Manage.*, 127(4), 281-290.

Jiang, G., and Shi, J. (2005). "Exact algorithm for solving project scheduling problems under multiple resource constraints." *J.Constr.Eng.Manage.*, 131(9), 986-992.

Jiang, M. F., Tseng, S. S., and Su, C. M. (2001). "Two-phase clustering process for outliers detection." *Pattern Recog.Lett.*, 22(6-7), 691-700.

Jin, Y., Levitt, R. E., Kunz, J. C., and Christiansen, T. R. (1995). "Virtual Design Team: A computer simulation framework for studying organizational aspects of concurrent design." *Simulation*, 64(3), 160-174.

Jung, Y., and Kang, S. (2007). "Knowledge-based standard progress measurement for integrated cost and schedule performance control." *J.Constr.Eng.Manage.*, 133(1), 10-21.

Jung, Y., and Woo, S. (2004). "Flexible work breakdown structure for integrated cost and schedule control." *J.Constr.Eng.Manage.*, 130(5), 616-625.

Jyh-Bin Yang. (2007). "Developing a knowledge map for construction scheduling using a novel approach." *Autom.Constr.*, 16(6), 806-15.

Kamara, J. M., Anumba, C. J., and Carrillo, P. M. (2002). "A CLEVER approach to selecting a knowledge management strategy." *Int.J.Project Manage.*, 20(3), 205-211.

Kamara, J. M., Anumba, C. J., Carrillo, P. M., and Bouchlaghem, N. (2003). "Conceptual framework for live capture of project knowledge." *Proc., CIB W078 Int. Conf. on Information Technology for Construction—Construction IT: Bridging the Distance, CIB, Waiheke Island, New Zealand, 178–185.*, .

Kamara, J. M., Augenbroe, G., Anumba, C. J., and Carrillo, P. M. (2002). "Knowledge management in the architecture, engineering and construction industry." *Construction Innovation*, 2(1), 53.

Kandil, A., and El-Rayes, K. (2006). "Parallel genetic algorithms for optimizing resource utilization in large-scale construction projects." *J.Constr.Eng.Manage.*, 132(5), 491-498.

Kara, S. 1., Kayis, B. 1., and Kaebnick, H. 1. (2001). "Concurrent resource allocation (CRA): a heuristic for multi-project scheduling with resource constraints in concurrent engineering." *Concurrent Eng.: Res.Appl.*, 9(1), 64-73.

Karaa, F. A., and Nasr, A. Y. (1986). "RESOURCE MANAGEMENT IN CONSTRUCTION." *J.Constr.Eng.Manage.*, 112(3), 346-357.

Kasvi, J. J. J., Vartiainen, M., and Hailikari, M. (2003). "Managing knowledge and knowledge competences in projects and project organisations." *Int.J.Project Manage.*, 21(8), 571-82.

Kerzner, H. (1979). *Project management : a systems approach to planning, scheduling, and controlling*. Van Nostrand Reinhold, New York.

Kim, K., and Garza, D. L. (2005). "Evaluation of the resource-constrained critical path method algorithms." *J.Constr.Eng.Manage.*, 131(5), 522-532.

Kim, K., and Garza, D. L. (2003). "Phantom float." *J.Constr.Eng.Manage.*, 129(5), 507-517.

Kimball, R., and Merz, R. (2000). *The data Webhouse toolkit : building the Web-enabled data warehouse*. John Wiley & Sons, New York.

Knight, K., and Fayek, A. R. (2002). "Use of fuzzy logic for predicting design cost overruns on building projects." *J.Constr.Eng.Manage.*, 128(6), 503-512.

Knight, K., and Fayek, A. R. (2000). "A preliminary study of the factors affecting the cost escalation of construction projects." *Canadian Journal of Civil Engineering*, 27(1), 73-83.

Kodama, M. (2007). *Project-based organization in the knowledge-based society*. Imperial College Press; distributed by World Scientific Pub. Co. Pte, London; Singapore.

Koh, J. L. Y., Lee, M. L., Hsu, W., and Kai Tak Lam. (2007). "Correlation-based detection of attribute outliers." *Proceedings*, Springer-Verlag, Bangkok, Thailand, 164-75.

KPMG (2003). "Insights from KPMG's European Knowledge Management Survey 2002/2003"

Kum-Khiong Yang¹, and Chee-Chuong Sum¹. (1997). "An evaluation of due date, resource allocation, project release, and activity scheduling rules in a multiproject environment." *Eur.J.Oper.Res.*, 103(1), 139-54.

Kum-Khiong Yang¹, and Chee-Chuong Sum¹. (1993). "A comparison of resource allocation and activity scheduling rules in a dynamic multi-project environment." *J.Oper.Manage.*, 11(2), 207-18.

Kunz, J. C., Christiansen, T. R., Cohen, G. P., Jin, Y., and Levitt, R. E. (1998). "The Virtual Design Team." *Commun ACM*, 41(11), 84-91.

Kurtulus, I., and Davis, E. W. (1982). "MULTI-PROJECT SCHEDULING: CATEGORIZATION OF HEURISTIC RULES PERFORMANCE." *Management Science*, 28(2), 161-172.

Kwak, Y. H., and Watson, R. J. (2005). "Conceptual estimating tool for technology-driven projects: Exploring parametric estimating technique." *Technovation*, 25(12), 1430-1436.

La Bella, A., Canzano, D., and Grimaldi, M. (2004). "Critical capabilities and performance in the aerospace industry: A knowledge management approach." *2004 IEEE Aerospace Conference Proceedings, Mar 6-13 2004*, Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States, Big Sky, MT, United States, 3962-3969.

Laslo, Z., and Goldberg, A. I. "Resource allocation under uncertainty in a multi-project matrix environment: Is organizational conflict inevitable?" *International Journal of Project Management*, In Press, Corrected Proof.

Leake, D. B., and Wilson, D. C. (2001). "A case-based framework for interactive capture and reuse of design knowledge." *Appl.Intell.*, 14(1), 77-94.

- Lee, D. (2005). "Probability of project completion using stochastic project scheduling simulation." *J.Constr.Eng.Manage.*, 131(3), 310-318.
- Lee, K., Chin, S., and Kim, J. (2003). "A core system for design information management using Industry Foundation Classes." *Computer-Aided Civil and Infrastructure Engineering*, 18(4), 286-98.
- Leseure, M. J., and Brookes, N. J. (2004). "Knowledge management benchmarks for project management." *Journal of Knowledge Management*, 8(1), 103-16.
- Lessard, C. S., and Lessard, J. (2007). *Project management for engineering design*. Morgan & Claypool Publishers, San Rafael, Calif.
- Leu, S., Yang, C., and Huang, J. (2000). "Resource leveling in construction by genetic algorithm-based optimization and its decision support system application." *Automation in Construction*, 10(1), 27-41.
- Leu, S., and Hung, T. (2002). "An optimal construction resource leveling scheduling simulation model." *Canadian Journal of Civil Engineering*, 29(2), 267-275.
- Levitt, R. E., Thomsen, J., Christiansen, T. R., Kunz, J. C., Jin, Y., and Nass, C. (1999). "Simulating project work processes and organizations: Toward a micro-contingency theory of organizational design." *Management Science*, 45(11), 1479-1495.
- Lewis, J. P. (2008). *Mastering project management : applying advanced concepts to systems thinking, control & evaluation, resource allocation*. McGraw-Hill, New York.
- Liberda, M., Ruwanpura, J., and Jergeas, G. (2003). "Construction Productivity Improvement: A Study of Human, Management and External Issues." *Construction Research Congress, Winds of Change: Integration and Innovation in Construction, Proceedings of the Congress*, American Society of Civil Engineers, 33-40.

Liebowitz, J., and Megbolugbe, I. (2003). "A set of frameworks to aid the project manager in conceptualizing and implementing knowledge management initiatives." *Int.J.Project Manage.*, 21(3), 189-98.

Liebowitz, J. (2006). *Strategic intelligence : business intelligence, competitive intelligence, and knowledge management*. Auerbach Publications, Boca Raton, FL.

Lova, A., Maroto, C., and Tormos, P. (2000). "A multicriteria heuristic method to improve resource allocation in multiproject scheduling." *European Journal of Operational Research*, 127(2), 408-424.

Love, P. E. D., Fong, P. S., and Irani, Z. (2005). "Management of knowledge in project environments." 242.

Lozon, J. P., and Jergeas, G. F. (2008). "Evaluating best practices for oil sands projects." *52nd Annual Meeting of AACE International and the 6th World Congress of ICEC on Cost Engineering, Project Management, and Quantity Surveying, June 29, 2008 - July 2*, Association for the Advancement of Cost Engineering, Toronto, ON, Canada, .

Lu, H., and Issa, R. R. A. (2005). "Extended production integration for construction: A loosely coupled project model for building construction." *J.Comput.Civ.Eng.*, 19(1), 58-68.

Lu, M. (2000). "Productivity studies using advanced ANN models." PhD thesis, University of Alberta (Canada), Canada.

Lu, M., and AbouRizk, S. M. (2000). "Simplified CPM/PERT simulation model." *J.Constr.Eng.Manage.*, 126(3), 219-226.

Lu, M., and Lam, H. (2008). "Critical path scheduling under resource calendar constraints." *J.Constr.Eng.Manage.*, 134(1), 25-31.

- Lu, M., and Li, H. (2003). "Resource-activity critical-path method for construction planning." *J.Constr.Eng.Manage.*, 129(4), 412-420.
- Luiten, G. T., Tolman, F. P., and Fischer, M. A. (1998). "Project-modelling in AEC to integrate design and construction." *Comput.Ind.*, 35(1), 13-29.
- Ma, Z. (2006). *Database modeling for industrial data management : emerging technologies and applications*. Idea Group Pub., Hershey, PA.
- MacMahon, C., Lowe, A., and Culley, S. (2004). "Knowledge management in engineering design: personalization and codification." *J.Eng.Des.*, 15(4), 307-25.
- Malhotra, Y. (2005). "Integrating knowledge management technologies in organizational business processes: getting real time enterprises to deliver real business performance." *Journal of Knowledge Management*, 9(1), 7-28.
- Mandel, R. (2007). "Building blocks: Constructing refineries, one module at a time." *Welding Design and Fabrication*, v 80, n 5, may 15, 2007, 80(5),.
- McCahill, D. F., and Bernold, L. (1993). "Resource-oriented modeling and simulation in construction." *J.Constr.Eng.Manage.*, 119(3), 590-606.
- Menches, C. L., and Hanna, A. S. (2006). "Quantitative measurement of successful performance from the project manager's perspective." *J.Constr.Eng.Manage.*, 132(12), 1284-1293.
- Merkle, D. 1., Middendorf, M., and Schmeck, H. (2002). "Ant colony optimization for resource-constrained project scheduling." *IEEE Transactions on Evolutionary Computation*, 6(4), 333-46.
- Messner, J. I. (2003). "An Architecture for Knowledge Management in the AEC Industry." *Construction Research Congress, Winds of Change: Integration and Innovation in Construction, Proceedings of the Congress, Mar 19-21 2003*, American Society of Civil Engineers, Honolulu, HI., United States, 889-896.

Mezher, T., Abdul-Malak, M. A., Ghosn, I., and Ajam, M. (2005). "Knowledge management in mechanical and industrial engineering consulting: A case study." *J.Manage.Eng.*, 21(3), 138-47.

Milton, N. J. (2005). *Knowledge management for teams and projects*. Chandos Pub., Oxford.

Mohamed, Y. (2002). "A framework for systematic improvement of construction systems." PhD thesis, University of Alberta (Canada), Canada.

Mohamed, Y., and AbouRizk, S. M. (2005). "Framework for building intelligent simulation models of construction operations." *J.Comput.Civ.Eng.*, 19(3), 277-291.

Mohamed, Y., and AbouRizk, S. M. (2006). "A hybrid approach for developing special purpose simulation tools." *Canadian Journal of Civil Engineering*, 33(12), 1505-1515.

Mohamed, Y., Borrego, D., Francisco, L., Al-Hussein, M., Abourizk, S., and Hermann, U. (2007). "Simulation-based scheduling of module assembly yards: Case study." *Engineering, Construction and Architectural Management*, 14(3), 293-311.

Mohanty, R. P., and Siddiq, M. K. (1989). "Multiple projects — Multiple resources constrained scheduling: A multiobjective analysis." *Engineering Costs and Production Economics*, 18(1), 83-92.

Moon, S. W., Kim, J. S., and Kwon, K. N. (2007). "Effectiveness of OLAP-based cost data management in construction cost estimate." *Autom.Constr.*, 16(3), 336-44.

Morris, P. W. G., and Pinto, J. K. (2007). *The Wiley guide to project organization & project management competencies*. John Wiley & Sons, Hoboken, N.J.

Morris, P. W. G., and Pinto, J. K. (2007). *The Wiley guide to project technology, supply chain & procurement management*. John Wiley & Sons, Hoboken, N.J.

Morris, P. W. G., and Pinto, J. K. (2007). *The Wiley guide to project, program & portfolio management*. J. Wiley & Sons, Hoboken, N.J.

Nasira, S. and Abd.Majid, M. (2006). "PROJECT COST PERFORMANCE FORECASTING SYSTEM", Proceedings of the 6th Asia-Pacific Structural Engineering and Construction Conference, Kuala Lumpur, Malaysia

Nassar, N. K. (2005). "An integrated framework for evaluation, forecasting and optimization of performance of construction projects." PhD thesis, University of Alberta (Canada), Canada.

Navon, R., and Sacks, R. (2007). "Assessing research issues in Automated Project Performance Control (APPC)." *Autom.Constr.*, 16(4), 474-84.

Nissen, M. E., and Levitt, R. E. (2004). "Agent-based modeling of knowledge dynamics." *Knowledge Management Research & Practice*, 2(3), 169-83.

Nonaka, I., and Takeuchi, H. (1997). "The knowledge-creating company: how Japanese companies create the dynamics of innovation." *Technological Forecasting and Social Change*, 55(1), 99.

O'Leary, D. E. (1998). "Enterprise knowledge management." *Computer*, 31(3), 54-61.

Olson, H. (1995). "Quantitative "versus" qualitative research: the wrong question." *Canadian Association for Information Science Meeting on Connectedness: Information Systems, People, Organizations (CAIS/ACSI '95)*, Univ. Alberta, Edmonton, Alta., Canada, 40-9.

O'Neill, J. J. (1989). *Management of industrial construction projects*. Nichols Pub., New York.

OLAP Council (1997). "OLAP Council White Paper"

Oni, A. I. (2008). "Linking project controls implementation planning and project Success: The northern Canadian mining projects lessons." *52nd Annual Meeting of AACE International and the 6th World Congress of ICEC on Cost Engineering, Project Management, and Quantity Surveying*, Association for the Advancement of Cost Engineering, Morgantown, WV 26501, United States, 8.

Orange, G., Burke, A., and Beam, J. (2000). "Organisational learning in the UK construction industry: a knowledge management approach." *Proceedings of ECIS 2000: 8th European Conference on Information Systems*, Vienna Univ. Econ. & Bus. Adm, Vienna, Austria, 599-606.

Owolabi, A., Anumba, C. J., El-Hamalawi, A., and Harper, C. (2006). "Development of an industry foundation classes assembly viewer." *J.Comput.Civ.Eng.*, 20(2), 121-131.

Padman, R. 1., Smith-Daniels, D. E. 1., and Smith-Daniels, V. L. 1. (1997). "Heuristic scheduling of resource-constrained projects with cash flows." *Naval Research Logistics*, 44(4), 365-81.

Park, M. (2005). "Model-based dynamic resource management for construction projects." *Autom.Constr.*, 14(5), 585-98.

Park, M., and Pena-Mora, F. (2003). "Dynamic change management for construction: introducing the change cycle into model-based project management." *System Dynamics Review*, 19(3), 213-42.

Payne, J. H. (1995). "Management of multiple simultaneous projects: a state-of-the-art review." *International Journal of Project Management*, 13(3), 163-168.

Petkova, O., and Petkov, D. (2003). "Improved understanding of software development productivity factors to aid in the management of an outsourced

project." *Journal of Information Technology Cases and Applications (JITCA)*, 5(1), 5-22.

Prietula, M. J., Carley, K. M., and Gasser, L. G. (1998). "Simulating organizations : computational models of institutions and groups." 248.

Project Management Institute (PMI) (2008). "Project Management – Body of Knowledge (PM-BoK)".

Raghu, T. S., and Vinze, A. (2007). "A business process context for Knowledge Management." *Decis.Support Syst.*, 43(3), 1062-1079.

Rajpathak, D. G., Motta, E., Zdrahal, Z., and Roy, R. (2006). "A generic library of problem solving methods for scheduling applications." *IEEE Trans.Knowled.Data Eng.*, 18(6), 815-828.

Rankin, L. K., Lozon, J. P., and Jergeas, G. F. (2005). "Detailed execution planning model for large oil and gas construction projects." *33rd CSCE Annual Conference 2005*, Canadian Society for Civil Engineering, Montreal, H3H 2R9, Canada, Toronto, ON, Canada, 130-1.

Reda, R. M. (1990). "RPM. Repetitive project modeling." *J.Constr.Eng.Manage.*, 116(2), 316-330.

Rezgui, Y. (2001). "Review of information and the state of the art of knowledge management practices in the construction industry." *Knowl.Eng.Rev.*, 16(3), 241-54.

Riggs, J. L., Brown, S. B., and Trueblood, R. P. (1994). "Integration of technical, cost, and schedule risks in project management." *Comput.Oper.Res.*, 21(5), 521-33.

- Robinson, H. S., Carrillo, P. M., Anumba, C. J., and Al-Ghassani, A. M. (2005). "Knowledge management practices in large construction organisations." *Engineering Construction and Architectural Management*, 12(5), 431-45.
- Rozenes, S., Vitner, G., and Spraggett, S. (2006). "Project Control: Literature Review." *Project Management Journal*, 37(4), 5.
- Rozenes, S., Vitner, G., and Spraggett, S. (2004). "MPCS: Multidimensional Project Control System." *International Journal of Project Management*, 22(2), 109-118.
- Rujirayanyong, T., and Shi, J. J. (2006). "A project-oriented data warehouse for construction." *Autom. Constr.*, 15(6), 800-807.
- Russell, A., and Froese, T. (1997). "Challenges and a vision for computer-integrated management systems for medium-sized contractors." *Canadian Journal of Civil Engineering*, 24(2), 180-190.
- Salem, O. M. (1999). "Infrastructure construction and rehabilitation: Risk-based life cycle cost analysis." PhD thesis, University of Alberta (Canada), Canada.
- Savin, D., Alkass, S., and Fazio, P. (1996). "Construction resource leveling using neural networks." *Canadian Journal of Civil Engineering*, 23(4), 917-925.
- Sawhney, A. 1., Bashford, H. 1., Walsh, K. 1., and Mund, A. 1. (2001). "Simulation of production homebuilding using Symphony." *Proceedings of the 2001 Winter Simulation Conference*, IEEE, Arlington, VA, USA, 1521-7.
- Schindler, M., and Eppler, M. J. (2003). "Harvesting project knowledge: A review of project learning methods and success factors." *Int.J.Project Manage.*, 21(3), 219-228.
- Seidman, C. (2001). *Data mining with Microsoft SQL server 2000 technical reference [electronic resource]*. Microsoft Press, Redmond, Wash.

Senouci, A. B., and Adeli, H. (2001). "Resource scheduling using neural dynamics model of Adeli and Park." *J.Constr.Eng.Manage.*, 127(1), 28-34.

Seong Yong Ohm¹, and Chu Shik Jhon¹. (1992). "A branch-and-bound method for the optimal scheduling." *92CH3078-3*, IEEE, Boston, MA, USA, 8-6.

Settas, D., Bibi, S., Sfetsos, P., Stamelos, I., and Gerogiannis, V. (2006). "Using Bayesian belief networks to model software project management antipatterns." *4th International Conference on Software Engineering Research, Management and Applications, SERA 2006*, Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States, Seattle, WA, United States, 117-124.

Shaheen, A. A. I. (2005). "A framework for integrating fuzzy set theory and discrete event simulation in construction engineering." PhD thesis, University of Alberta (Canada), Canada.

Shahin, A. A. (2007). "A framework for cold weather construction simulation." PhD thesis, University of Alberta (Canada), Canada.

Shelbourn, M. A., Bouchlaghem, D. M., Anumba, C. J., Carillo, P. M., Khalfan, M. M. K., and Glass, J. (2006). "Managing knowledge in the context of sustainable construction." *Electron.J.Inf.Technol.Constr.*, 11 57-71.

Shen, Q., Guo, J., Zhang, J., and Liu, G. (2008). "Using Data Mining Techniques to Support Value Management Workshops in Construction." *Tsinghua Science and Technology*, 13(2), 191-201.

Shi, J. J., and Rujirayanyong, T. (2006). "A project-oriented data warehouse for construction." *Autom.Constr.*, 15(6), 800-7.

Shuai, D., Lu, C., and Zhang, B. (2006). "Entanglement partitioning of quantum particles for data clustering." *30th Annual International Computer Software and Applications Conference, COMPSAC 2006*, Institute of Electrical and Electronics

Engineers Computer Society, Piscataway, NJ 08855-1331, United States, Chicago, IL, United States, 285-290.

Soibelman, L., and Kim, H. (2002). "Data preparation process for construction knowledge generation through knowledge discovery in databases." *J.Comput.Civ.Eng.*, 16(1), 39-48.

Solomon, P. J., and Young, R. R. (2007). *Performance-based earned value*. Wiley, Hoboken, N.J.

Song, L. (2004). "Productivity modeling for steel fabrication projects." PhD thesis, University of Alberta (Canada), Canada.

Song, L., and AbouRizk, S. M. (2006). "Modeling labor productivity in steel fabrication." *2006 AACE International Transactions - 50th Annual Meeting*, Association for the Advancement of Cost Engineering, Morgantown, United States, Las Vegas, NV, United States, 25-1.

Song, L., and AbouRizk, S. M. (2006). "Virtual shop model for experimental planning of steel fabrication projects." *J.Comput.Civ.Eng.*, 20(5), 308-316.

Song, L., Al-Battaineh, H. T., and AbouRizk, S. M. (2005). "Modeling uncertainty with an integrated simulation system." *Canadian Journal of Civil Engineering*, 32(3), 533-42.

Song, L., Wang, P., and AbouRizk, S. (2006). "A virtual shop modeling system for industrial fabrication shops." *Simulation Modelling Practice and Theory*, 14(5), 649-662.

Song, P. X. (2007). *Correlated data analysis : modeling, analytics, and applications*. Springer Verlag, New York.

Statistics Canada Catalogue, (2008). Website Address:

<http://www.statcan.gc.ca/start-debut-eng.html>

Stefatos, G., and Ben Hamza, A. (2007). "Cluster PCA For outliers detection in high-dimensional data." *2007 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2007*, Institute of Electrical and Electronics Engineers Inc., New York, NY 10016-5997, United States, Montreal, QC, Canada, 3961-3966.

Stumpf, A. L., Ganeshan, R., Chin, S., and Liu, L. Y. (1996). "Object-oriented model for integrating construction product and process information." *J.Comput.Civ.Eng.*, 10(3), 204-212.

Stuurstraat, N., and Tolman, F. (1999). "A product modeling approach to building knowledge integration." *Autom.Constr.*, 8(3), 269-75.

Sullivan, M. (2007). *Statistics : informed decisions using data*. Pearson Prentice Hall, Upper Saddle River, N.J.

Sumathi, S., and Sivanandam, S. N. *Introduction to Data Mining and its Applications. Studies in Computational Intelligence, Volume 29*. Springer, .

Sure, Y., Staab, S., and Studer, R. (2002). "Methodology for development and employment of ontology based knowledge management applications." *SIGMOD Record*, 31(4), 18-23.

Tan, H. C., Carrillo, P. M., Anumba, C., Bouchlaghem, N., Kamara, J. M., and Udejaja, C. E. (2007). "Development of a methodology for live capture and reuse of project knowledge in construction." *J.Manage.Eng.*, 23(1), 10-26.

Taniar, D. (2008). *Data mining and knowledge discovery technologies*. IGI Pub., Hershey.

Teicholz, P. (1993). "Forecasting final cost and budget of construction projects." *J.Comput.Civ.Eng.*, 7(4), 511-529.

The Data Warehousing Institute (TDWI) (2008). Website Address:

<http://www.tdwi.org/>

Thompson, B. (2004). *Exploratory and confirmatory factor analysis : understanding concepts and applications*. American Psychological Association, Washington, DC.

Toklu, Y. C. (2002). "Application of genetic algorithms to construction scheduling with or without resource constraints." *Canadian Journal of Civil Engineering*, 29(3), 421-429.

Tolman, F. P. (1999). "Product modeling standards for the building and construction industry: Past, present and future." *Autom. Constr.*, 8(3), 227-235.

Tormos, P. 1., and Lova, A. 1. (2003). "An efficient multi-pass heuristic for project scheduling with constrained resources." *Int J Prod Res*, 41(5), 1071-86.

Tormos, P. 1., and Lova, A. 1. (2001). "A competitive heuristic solution technique for resource-constrained project scheduling." *Annals of Operations Research*, 102 65-81.

Tserng, H. P., and Lin, Y. (2004). "Developing an activity-based knowledge management system for contractors." *Autom. Constr.*, 13(6), 781-802.

Turban, E., Aronson, J. E., and Liang, T. (2005). *Decision support systems and intelligent systems*. Pearson/Prentice Hall, Upper Saddle River, NJ.

Turk, Z. (2006). "Construction informatics: Definition and ontology." *Advanced Engineering Informatics*, 20(2), 187-199.

Udaipurwala, A., and Russell, A. D. (2002). "Computer-assisted construction methods knowledge management and selection." *Canadian Journal of Civil Engineering*, 29(3), 499-516.

Van der Velde, Robert R., and van Donk, Dirk Pieter. (2002). "Understanding bi-project management: Engineering complex industrial construction projects." *Int.J.Project Manage.*, 20(7), 525-533.

- van Donk, D. P., and Riezebos, J. (2005). "Exploring the knowledge inventory in project-based organisations: A case study." *Int.J.Project Manage.*, 23(1), 75-83.
- Vaziri, K., Carr, P. G., and Nozick, L. K. (2007). "Project planning for construction under uncertainty with limited resources." *J.Constr.Eng.Manage.*, 133(4), 268-276.
- Wales, R. J., and AbouRizk, S. M. (1996). "An integrated simulation model for construction." *Simul Pract Theory*, 3(6), 401-20.
- Walter, T. J. 1. (1997). "A framework for integrating design automation with computer aided parametric estimating (CAPE)." *NAECON 1997*, IEEE, Dayton, OH, USA, 417-22.
- Wang, E. T. G., Chia-Lin Lin, Jiang, J. J., and Klein, G. (2007). "Improving enterprise resource planning (ERP) fit to organizational process through knowledge transfer." *Int.J.Inf.Manage.*, 27(3), 200-212.
- Wang, P. (2006). "Production-based large scale construction simulation modeling." PhD thesis, University of Alberta (Canada), Canada.
- Wang, Y., and Chen, M. (2004). "A collaborative knowledge production model for knowledge management in complex engineering domains." *2004 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, The Hague, Netherlands, 5050-5.
- Wickramasinghe, N., and Von Lubitz, D. K. J. E. (2007). *Knowledge-based enterprise : theories and fundamentals*. Idea Group Pub., Hershey PA.
- Wiley, V. D. 1., Deckro, R. F., and Jackson, J. A., Jr. (1998). "Optimization analysis for design and planning of multi-project programs." *Project Management and Scheduling: Fifth International Workshop*, Elsevier, Poznan, Poland, 492-506.
- Williams, S., and Williams, N. (2007). *The profit impact of business intelligence*. Elsevier/Morgan Kaufmann, Amsterdam; Boston.

Witten, I. H., and Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Morgan Kaufman, Amsterdam; Boston, MA.

Woo, J., Clayton, M. J., Johnson, R. E., Flores, B. E., and Ellis, C. (2004). "Dynamic Knowledge Map: Reusing experts' tacit knowledge in the AEC industry." *Autom. Constr.*, 13(2), 203-207.

Wrembel, R., and Koncilia, C. (2007). *Data warehouses and OLAP: concepts, architectures, and solutions*. IRM Press, Hershey PAU; London.

Xiao, W., and Wei, Q. (2007). "Development of an integrated project management information system for aerial enterprises and its key technologies." *2006 IEEE International Conference on Systems, Man and Cybernetics*, Institute of Electrical and Electronics Engineers Inc., New York, NY 10016-5997, United States, Taipei, Taiwan, 4073-4077.

Yeo, K. T., and Ning, J. H. (2006). "Managing uncertainty in major equipment procurement in engineering projects." *Eur.J.Oper.Res.*, 171(1), 123-34.

Yin, P., and Wang, J. (2006). "Ant colony optimization for the nonlinear resource allocation problem." *Applied Mathematics and Computation (New York)*, 174(2), 1438-1453.

Yu-Cheng Lin, Lung-Chuang Wang, and Tserng, H. P. (2006). "Enhancing knowledge exchange through web map-based knowledge management system in construction: Lessons learned in Taiwan." *Autom. Constr.*, 15(6), 693-705.

Zaiane, O. R. (2008) "Introduction to Data Mining" Class Notes

Zaiane, O. R., Foss, A., Chi-Hoon Lee, and Wang, W. (2002). "On data clustering analysis: Scalability, constraints, and validation." *Proceedings*, Springer-Verlag, Berlin, Germany, 28-39.

Zhen Chen, Heng Li, Kong, S. C. W., and Qian Xu. (2005). "A knowledge-driven management approach to environmental-conscious construction." *Construction Innovation*, 5(1), 27-39.

Zhiliang, M., Wong, K. D., Heng, L., and Jun, Y. (2005). "Utilizing exchanged documents in construction projects for decision support based on data warehousing technique." *Automation in Construction*, 14(3), 405-412.

Zhu, X., and Davidson, I. (2007). *Knowledge discovery and data mining: challenges and realities*. Information Science Reference, Hershey.

Zocher, M., and Thompson, G. (1992). "Cost and schedule baseline development." *Proceedings of the 36th Annual Transactions of the American Association of Cost Engineers - AACE*, Publ by AACE, Morgantown, WV, USA, 3-1.