Asset Management Data Warehouse Data Modelling

Avin Mathew

BE, BCom

School of Engineering Systems

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

August 2008

Abstract

Data are the lifeblood of an organisation, being employed by virtually all business functions within a firm. Data management, therefore, is a critical process in prolonging the life of a company and determining the success of each of an organisation's business functions. The last decade and a half has seen data warehousing rising in priority within corporate data management as it provides an effective supporting platform for decision support tools. A cross-sectional survey conducted by this research showed that data warehousing is starting to be used within organisations for their engineering asset management, however the industry uptake is slow and has much room for development and improvement. This conclusion is also evidenced by the lack of systematic scholarly research within asset management data warehousing as compared to data warehousing for other business areas. This research is motivated by the lack of dedicated research into asset management data warehousing and attempts to provide original contributions to the area, focussing on data modelling.

Integration is a fundamental characteristic of a data warehouse and facilitates the analysis of data from multiple sources. While several integration models exist for asset management, these only cover select areas of asset management. **This research presents a novel conceptual data warehousing data model that integrates the numerous asset management data areas.** The comprehensive ethnographic modelling methodology involved a diverse set of inputs (including data model patterns, standards, information system data models, and business process models) that described asset management data. Used as an integrated data source, the conceptual data model was verified by more than 20 experts in asset management and validated against four case studies.

A large section of asset management data are stored in a relational format due to the maturity and pervasiveness of relational database management systems. Data warehousing offers the alternative approach of structuring data in a dimensional format, which suggests increased data retrieval speeds in addition to reducing analysis complexity for end users. To investigate the benefits of moving asset management data

from a relational to multidimensional format, **this research presents an innovative relational vs. multidimensional model evaluation procedure**. To undertake an equitable comparison, the compared multidimensional are derived from an asset management relational model and as such, **this research presents an original multidimensional modelling derivation methodology for asset management relational models**. Multidimensional models were derived from the relational models in the asset management data exchange standard, MIMOSA OSA-EAI. The multidimensional and relational models were compared through a series of queries. It was discovered that multidimensional schemas reduced the data size and subsequently data insertion time, decreased the complexity of query conceptualisation, and improved the query execution performance across a range of query types.

To facilitate the quicker uptake of these data warehouse multidimensional models within organisations, an alternate modelling methodology was investigated. **This research presents an innovative approach of using a case-based reasoning methodology for data warehouse schema design.** Using unique case representation and indexing techniques, the system also uses a business vocabulary repository to augment case searching and adaptation. The system was validated through a case-study where multidimensional schema design speed and accuracy was measured. It was found that the case-based reasoning system provided a marginal benefit, with a greater benefits gained when confronted with more difficult scenarios.

Keywords

Asset management, data warehousing, conceptual data model, multidimensional model, star schemas, case-based reasoning.

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Avin Mathew
August 2008

Acknowledgements

I would like to firstly acknowledge the guidance and vision provided by my principal supervisor, Associate Professor Lin Ma, for my PhD degree, career, and life. You managed to draw out qualities in myself upon which this work depended. I would also like to thank my associate supervisor, Professor Doug Hargreaves, for his insight and advice. To the team at QUT – Dr. Sheng Zhang, Dr. Yong Sun, and Liqun Zhang – I appreciate your assistance and patience to the numerous questions I asked.

I would like to thank the sponsors of this research – the Australian Government, QUT, CIEAM, SunWater, and MPT Solutions. Without your contributions, I believe I would have overlooked such a marvellous opportunity.

My life has been changed thanks to the Australian-American Fulbright Association, and in particular, Mark Darby and Lyndell Wilson. I thank you for taking a chance on me, and providing such a unique opportunity. Much appreciation must go to Professor Jay Lee for being extremely accommodating during my time in the US, and to the crew at IMS – Patrick, Masoud, Mark, Shijin, Haixia, Carsten, Edzel, and Chen – it was one of my most enjoyable years ever.

Most importantly, I would thank God Almighty for providing me the opportunity to undertake this degree, as well remaining faithful during the times I needed Him the most. I would also like to thank my parents, brother and sister for enduring with my eccentric behaviour and providing an outlet of relief outside of work. Serra, I thank you for all your support and wisdom, and am eagerly looking forward to spending my life with you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

3NF	Third Normal Form
BI	Business Intelligence
CBR	Case-Based Reasoning
CIEAM	CRC for Integrated Engineering Asset Management
CMMS	Computerised Maintenance Management System
CRC	Cooperative Research Centre
CRIS	Common Relational Information Schema
DBMS	Database Management System
DSS	Decision Support System
EAM	Enterprise Asset Management
EAV	Entity Attribute Value
ER	Entity Relationship
ERP	Enterprise Resource Planning
ETL	Extraction, Transformation, and Loading
FMECA	Failure Mode, Effects, and Criticality Analysis
GIS	Geographic Information System
GUI	Graphical User Interface
IMS	Intelligent Maintenance Systems
IT	Information Technology
KPI	Key Performance Indicator
MES	Manufacturing Execution System
MIMOSA	Machinery Information Management Open Systems Alliance
O&M	Operations and Maintenance
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
OSA-EAI	Open Systems Architecture for Enterprise Application Integration
RBD	Reliability Block Diagram
SCADA	Supervisory Control And Data Acquisition
SQL	Structured Query Language
STEP	Standard for the Exchange of Product model data
UML	Unified Modelling Language
XML	Extensible Markup Language

1 Introduction

The last 30 years have led to an explosion in the amount of data available within organisations. Technology enhancements in information and data acquisition systems have allowed for huge volumes of data to be collected and retained within an organisation, while innovations such as the Internet have produced new sources of data such as e-mail systems. Data management has become a critical function for information technology (IT) departments as they attempt to gather, maintain, and analyse organisational data.

Many organisations have turned to data warehousing to bridge the gap of turning data into knowledge, and the last decade and a half has seen corporate decision support systems (DSS) coalescing around data warehousing and online analytical processing (OLAP). Serving as an information management solution that integrates information across domains, organisations, and applications, a data warehouse provides a conduit of accurate and timely information. A data warehouse forms the backbone of the information supply chain to a decision support system, and consequently gives companies a way of turning knowledge into tangible results.

Data warehousing has been successfully applied to a variety of business areas, including asset management. However, the sheer number of asset management systems presents challenges for the integration of data from disparate sources. Asset management data are governed by a variety of systems including ERP (Enterprise Resource Planning) systems, EAM (Enterprise Asset Management) systems, CMMS (Computerised Maintenance Management Systems), MES (Manufacturing Execution Systems), and SCADA (Supervisory Control And Data Acquisition) systems, to name a few. Several standardisation efforts have attempted to harmonise the different asset management data areas, however these efforts have had a less than desirable uptake due to a number of inherent problems in their approach.

There are several additional challenges unique to asset management data warehousing. These include:

- A mixture of related data from high sample rates (e.g. condition monitoring sensor data) and low sample rates (e.g. asset disposal)
- Specialist systems for specific asset management functions (e.g. process control, reliability analysis) that cannot be rolled up into an ERP system
- Complex IT integration with plant and equipment leading to a significant number of legacy systems due to the disparity between asset life (long life) and IT system life (short life)
- Specialist systems use complex algorithms for analysis (e.g. frequency spectrum analysis) as statistical functions provided by OLAP tools are often too simple
- A range of users originating from different areas in the organisation (e.g. engineering, information systems, and management), each with different qualifications (e.g. certificates, diplomas, degrees), and each requiring a different set of KPIs

Despite the evident shortcomings in asset management data warehousing, the academic community has not capitalised on the opportunity to investigate solutions, and little systematic research exists in this area. To address several of the challenges posed by asset management data warehousing, this research undertakes an original and comprehensive investigation into the area of data warehousing, with a focus on data modelling. Data modelling is one of the foundational processes in the design of a data warehouse, and much of a data warehouse's functionality is dictated by its data model. This research is a seminal work in the area of asset management data warehousing data modelling, and attempts to provide a valuable resource for organisations seeking to develop an integrated data warehouse platform for their asset management.

1.1 Asset Management

Despite the term "asset management" being well entrenched within the finance industry (and having at least six other definitions originating from different disciplines) [1], the term was adopted by the engineering profession during the last two decades for the management of physical assets. Examples of physical engineering assets include motors, pumps, pipes, and buildings. The definition of asset management used in this research is "the process of organising, planning and controlling the acquisition, use,

care, refurbishment, and/or disposal of physical assets to optimise their service delivery potential and to minimise the related risks and costs over their entire life through the development and application of intangible assets such as knowledge-based decision-making applications and business processes" [2].

Numerous asset management frameworks have been proposed by governmental bodies, standards committees, and organisations [2]. These frameworks extend the definition detailed above to provide an understanding into the motivations, functions, systems, and information required in the management of assets. The framework on which this research is based is the 12-module CIEAM Asset Management Framework, illustrated in Appendix A. CIEAM (Cooperative Research Centre for Integration Engineering Asset Management) is an Australian research organisation that addresses the field of integrated engineering asset management. The theoretical framework was developed from a comprehensive review of literature and existing frameworks, and provides a clear understanding of the functions within asset management.

Both the definition and framework form the scope for asset management in this research. Both items demarcate the boundaries of asset management, and outline the activities and functions undertaken for the management of engineering assets.

1.2 Data Warehousing

The term "data warehousing" was coined by Inmon [3] in 1990 when he conceived the now-classic definition of a data warehouse as a "subject-oriented, integrated, time-variant and non-volatile [data] collection in support of management decision making processes". The process of data warehousing involves extracting data from source online transactional processing (OLTP) systems, transforming the data, and loading the data into a centralised data warehouse for the purpose of analysis. Over OLTP systems, data warehouses provide numerous benefits [4]: a single view of organisational data, quicker and easier access to data, increased performance of operational systems, integrated data analysis, and the use of OLAP tools.

The benefits of data warehousing were quickly acknowledged by many organisations, and by the mid-1990s, 95 percent of the Fortune 1000 companies in the US had developed or planned to develop a data warehouse [5]. Due to an inexperienced industry and lack of matured technology, it was estimated that one-half to two-thirds of these early data warehousing projects would fail [6]. This trend continues today with

Gartner estimating that 50 percent of data warehouse projects in 2007 would have limited acceptance or would be failures [7]. A data warehousing project can be a risky proposition, but it is not without its commensurate rewards. A data warehouse was directly attributed to lifting one organisation's $60 million loss to profits in excess of $211 million eight years later [8].

Due to the ballooning amounts of data captured within organisations, it is not surprising to see that data management projects were the third top priority IT projects for 2007, also yielding the quickest payback period [9]. Data warehousing, along with its source information systems, are an important aspect of an organisation's information management platform and their importance will continue to grow as organisations clamour to centralise and analyse their data.

1.3 Context

This research has been conducted as part of a larger Australian Research Council (ARC) initiative which aimed to develop an intelligent maintenance decision support system (DSS) for the water utility industry. The goal of the DSS was to employ novel and intelligent algorithms that combine asset operation, condition, and reliability data, to make accurate predictions about asset maintenance. The data required by the DSS could be located anywhere in an organisation, and a data warehouse provided the best mechanism to provide a consistent interface to the data and to guarantee data quality.

The work conducted in this research has progressed well beyond the original scope of the original proposal by attempting to address not just maintenance, but asset management as a whole. This research has also increased its scope by broadened itself from the water utility industry to encompassing the majority of industries that conduct asset management operations. The latter was made possible due to the similarities in information systems and data structures between these different industries and asset types. Despite the industry sponsorship of this research by SunWater and MPT Solutions, and the genesis of this work from within the water utility industry, this research attempts to remain industry-neutral.

1.4 Research Questions

The main aim of this research was in enhancing the body of knowledge of data warehousing data modelling, particularly for the field of asset management. To understand the opportunities, challenges, and experiences by previous researchers and practitioners, the first research question was asked:

"What is the state of asset management data warehousing in scholarly literature and industrial use?"

An understanding of the domain gave a clearer indication on the existing work that had been conducted by researchers, as well as the emphasis placed on the different areas in asset management data warehousing. The distinct lack of data warehousing research in asset management discovered by this research suggested an immaturity within existing data models. As conceptual analysis is typically recognised as the first step within data modelling [10], the second research question was asked:

"What is the constitution of an integrated asset management data warehouse conceptual data model?"

The conceptual data model along with all existing asset management data models present a relational view of asset management data. The field of data warehousing offers a dimensional view of data, and many data warehousing tools and platforms are built around multidimensional data structures. To estimate the potential value of integrated asset management data warehousing using dimensional data structures, the third research question was asked:

"What are the benefits received by moving asset management data from a relational format to a multidimensional format?"

As any eventual data warehouse data models would face challenges with awareness and acceptance, the ability to quickly implement these models within an organisation would increase their desirability. This beckoned the fourth research question:

"How can data warehouse data models be quickly implemented for an organisation?"

In answering these four research questions, five original research activities were undertaken and the outcomes of these activities are briefly described in the section below.

1.5 Outcomes, Innovation, and Significance

This research presents five original contributions to the field of asset management data warehousing to answer the research questions above. The first research question is answered through the first two research outcomes, while the subsequent questions correspond directly with the third to fifth outcomes.

Research Outcome 1:
Review of asset management data warehousing literature

This research presents **the first thorough review and analysis of both academic and non-academic forms of literature in asset management data warehousing**. Relying upon an interpretive epistemology, the review analyses past research into the area from sectoral, geographical, and chronological perspectives, and identifies the innovations in the papers that undertake pure research into asset management data warehousing (as opposed to applying data warehousing to assist research in a tangential area). The review reinforces the originality of all five research outcomes in this work.

The review has implications for both the theory and practice of asset management data warehousing, clearly delineating the limitations of past research and highlighting the numerous avenues available to future research.

Research Outcome 2:
Asset management data management survey

This research presents **a unique cross-sectional survey of asset management data management in industry**. The survey was undertaken to understand issues in data warehousing, information systems, integration, and data retention policies for

organisations that conduct asset management. Employing a positivist epistemology, the survey utilised a questionnaire as the primary means of data collection, and along with triangulative interviews, used descriptive and inferential statistics to highlight certain industry practices and concerns. The survey reinforces the significance of all five research outcomes in this work.

The survey serves as a benchmarking tool for organisations who can compare their data management maturity with their own or other industries. The survey also provides both developers of data management platforms as well as researchers with solid data on where efforts should be focused in order to have a significant impact.

Research Outcome 3:
Asset management conceptual data model

This research presents **a novel asset management conceptual data model that integrates the numerous data areas within asset management**. The thorough ethnographic methodology used in modelling investigated a unique set of data sources (including data model patterns, standards, information systems, and business process models) that described various aspects of asset management data.

The systematic methodology developed for this work incorporates a **diversity and depth unseen in previous research and sets a standard for future work**. The model is **unique to asset management in that it utilises data model patterns and takes into consideration the four data warehousing characteristics** (as stated in Section 1.2).

The model increases the coverage of integrated asset management areas, and serves as a reference for asset management system development, for deriving organisation-specific corporate data models, and for harmonising understanding between the IT and engineering domains.

Research Outcome 4:
Asset management multidimensional model evaluation

This research presents **an original multidimensional modelling derivation methodology from asset management relational models** in addition to **an**

innovative relational vs. multidimensional model evaluation procedure. A derivation is required to make an equitable comparison between the two model types, and the MIMOSA OSA-EAI CRIS is used as a case study. Quantitative quasi-experimental tests show the beneficial effects of multidimensional models upon data size, query conceptualisation complexity, and query execution performance.

The work presented in this thesis is the **first application of multidimensional modelling to integrated asset management and presents a unique derivation methodology**. The use of the **data integration standard is also unseen in previous work**. The evaluation methodology presented is **distinctive due to the metrics and segmentation by query types**.

This work provides quantifiable and factual evidence of the benefits of multidimensional modelling within asset management and gives confidence towards its application. It also provides a simplified data warehousing path for organisations that have adopted the MIMOSA OSA-EAI either for system or data integration.

Research Outcome 5:
Case-based reasoning system for data warehouse schema design

This research presents **a pioneering approach to data warehouse schema design using case-based reasoning**. A thorough system design is proposed, and the design is implemented and tested against the user-driven data warehouse schema design methodology through a case study.

This work presents a **new methodology to data warehouse schema design**. The approach adopts a **unique strategy for case representation identification** and a **novel industry-based thesaurus** for associating case items.

Organisations will be the primary beneficiaries of the system, as provides a mechanism to prototype data warehouse schemas more rapidly and with fewer errors. These two characteristics have significant implications towards a reduction in data warehouse development costs.

Figure 1.1 graphically summarises the five research outcomes and their relationships. The first two outcomes are designated as research preparation, as they indicate given

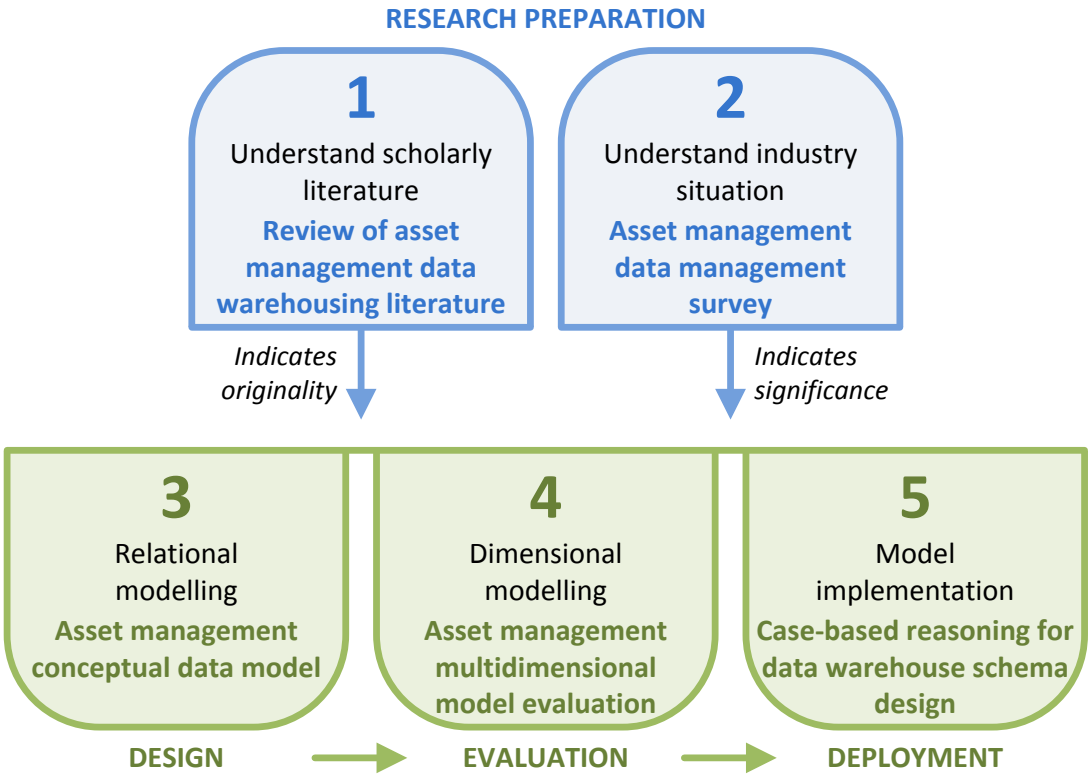


Figure 1.1 – Research outcomes and relationships

an insight into the originality and significance of the three subsequent outcomes. Research Outcomes 3 and 4 respectively fall in the area of relational modelling and dimensional modelling, but have different premises: design and testing, respectively. With Research Outcome 5 covering the deployment of data warehouse data models in organisations, the three latter outcomes neatly align with the linear systems development life cycle [11].

1.6 Thesis Outline

There are seven chapters in this thesis.

Chapter 1 provides a background and context to the research and presents the research questions that are addressed. Five original and innovative outcomes are presented and for each, the chapter includes a justification and the wider implications.

Chapter 2 reviews fundamental literature on data warehousing, and critically analyses asset management data warehousing research. From a comprehensive review of the literature, the originality of the material presented in this thesis is justified.

Chapter 3 describes a cross-sectional survey of asset management information systems and data warehousing conducted in various industries. The survey provides an understanding into the management and use of information systems and data warehousing, and underscores the significance of this research.

Chapter 4 begins the start of the asset management data warehouse data modelling process by presenting a comprehensive conceptual data model. The chapter discusses the rationale behind the modelling methodology, presents a conceptual model for asset management data, and validates the results through four case studies.

Chapter 5 discusses the use of multidimensional schemas for asset management data. An empirical-based entity relationship to multidimensional schema methodology is presented, and the MIMOSA OSA-EAI is used as a case study. The process is validated by taking a section of the model through a rigorous testing procedure involving five query types against eight different data sets.

Chapter 6 presents a case-based reasoning system that allows for the rapid development of data warehouse multidimensional schemas. A detailed system design is presented, and the implemented software is validated by a case study.

Chapter 7 concludes the thesis by revisiting the research process, the contributions to the body of knowledge, and their implications for theory and practice. It also looks at the restrictions imposed in the current work and suggests areas where future research can build upon this foundational asset management data warehousing research.

2 Review of Literature

The classical research methodology dictates that a literature review be undertaken towards the outset of research before any research questions are engaged. In doing so, a researcher is benefited through the comprehension of existing methodologies that may address the research questions, the traps and pitfalls encountered by other researchers, and potential areas for future research.

The primary focus of this literature review in this work was to 1) gain an understanding of the issues and techniques in asset management data warehousing, and 2) identify relevant areas that this research could contribute to the body of knowledge.

2.1 Review Methodology

The methodology adopted for this literature review was primarily a bottom-up methodology (by searching for research papers, as opposed to the top-down approach of starting with prominent academics and practitioners). While a top-down approach was initially trialled, the bottom-up approach uncovered more research. The following sections detail the methodology used in the review.

2.1.1 Selecting Information Sources

The advent of the Internet has led to a universal information explosion, as it facilitates an easier dissemination of information. As such, there are many new tools at a researcher's disposal, which were not available a decade ago. While older forms of information retrieval are still accessible, these new developments mean that these traditional forms have been superseded. For example, manually crawling through the table of contents of journals, conference proceedings, and books has been made easier through abstract and full-text searching. The usefulness of these developments is reflected in the use of online resources in this research (seen in Table 2.1). As the topic of this research is within data management, it seems naturally apt to maximise the use of electronic information sources.

The primary literature search source was that of publisher databases. Publisher databases are a collection of indexing and search services for journals and conference proceedings and contain a mixture of abstract and full-text searching. As numerous journals and conference proceedings are hosted by these sites, a large number of resources can be searched through in a short duration of time. The particular databases

Media type	Source type	Information source
Conference and journal papers	Publisher databases	ACM Portal[1] ASME Digital Library[2] IEEE Xplore[3] ProQuest[4] ScienceDirect[5] SpringerLink[6] ISI Web of Knowledge[7] Wiley InterScience[8]
	Aggregate search	Citeseer[9] EBSCOhost[10] Google Scholar[11]
	Citation indexes	ScienceDirect Google Scholar
Books	Libraries	Queensland University of Technology University of Queensland Griffith University University of Cincinnati
	Online providers	Google Books[12]
Patents	Aggregate search	Patent Lens[13]
Web sites	Web search	Google

Table 2.1 – Media types and their information sources

[1] http://portal.acm.org
[2] http://www.asmedl.org
[3] http://ieeexplore.ieee.org
[4] http://www.umi.com/proquest
[5] http://www.sciencedirect.com
[6] http://www.springerlink.com
[7] http://portal.isiknowledge.com
[8] http://www.interscience.wiley.com
[9] http://citeseer.ist.psu.edu
[10] http://search.ebscohost.com
[11] http://scholar.google.com
[12] http://books.google.com
[13] http://www.patentlens.net

shown in Table 2.1 were selected due to their large collection of publications spanning the information systems and asset management areas. As all are commercial services, the list was constrained by university access to the service.

To broaden the number of searched resources, services such as EBSCOhost and Google Scholar (both are discipline neutral) allow aggregate searches of multiple publisher databases. As these services typically do not have the copyright to distribute the actual article, they do provide a starting point for further searching. Google Scholar also acts similar to Citeseer in that it is an online crawling service (although Citeseer is primarily geared towards computer science and engineering domains), and consequently, can reference self-archived works.

Another tool used for the literature review was that of citation indexes. Citation indexes can be used to establish which later documents cite which earlier documents. This can be used to uncover newer research in the area, or conversely used to find eminent works for those articles with a high citation count.

Libraries are the traditional method of information searching and retrieval, and still form the core source for reference knowledge. Much of library information is stored as books on physical media, although there is an emergence of digital books – some of which were used in this research. Due to the technical nature of the topic, only university libraries were used, and these libraries were selected based on convenience of locality.

While academic institutions remain as the dominant force in research, many research divisions exist in commercial organisations. While the goal of academic research is to improve societal technologies, the goal of most commercial research is to improve the bottom line. Patents are one of the mechanisms used in achieving this goal as they provide an exclusive right to use the discovered research for a limited duration. As most commercial research organisations will patent their research rather than publish them as academic works, patent searching is an important, albeit unorthodox part of the literature review. Patent Lens is a free service that allows searches through US, European, Australian, and WIPO (World Intellectual Property Organization)/PCT (Patent Cooperation Treaty) patent databases.

Another source of relevant literature being used more frequently in research methodologies is Internet articles. While academics typically scorn such articles in preference of traditional literary formats, the popularity of the Internet has resulted in the establishment of web sites devoted to nearly every subject matter imaginable. This approach to gathering literature was explored to a great extent in this research to uncover material that the other tools could not. The results of these searches led to various data warehousing news sites that provided daily information about advancements in data warehousing, online discussion forums that contained technical questions and answers; and case studies presented by vendors.

2.1.2 Searching Information Sources

The search strategy for each information source started with identifying relevant key words and terms around the topic. The search terms “data warehouse” and “data warehousing” produced a plethora of articles on the basic data warehouse theory, while “asset management” produced articles on financial assets. The latter required the addition of related terms such as “physical” or “maintenance” to produce relevant results on engineering asset management.

Combining search phrases (e.g. “asset management data warehouse” and “asset management system integration”) narrowed search results, while the omission of non-core terms (e.g. “asset data warehouse” and “asset management integration”) broadened the search results.

Figure 2.1 shows the results of an example of the process. The chart indicates the total number of results returned from a basic search of different information sources using the terms “asset management data warehouse”. In most cases, the term was split into two phrases “asset management” “data warehouse” to focus the search. In the case of Patent Lens, the total number of patents was limited to the number of unique patents (hence multiple filings of the same patent were grouped). In the case of Google Scholar, 949 results were returned, and the y-axis is truncated for better presentation.

The number of results from each information source is matched against the number of relevant results. This involved perusing each article returned from the search to determine its relevance to asset management data warehousing (loosely related materials are also flagged as relevant). In the case of Google Scholar which returned a significant number of results, the narrowing modifiers described in the above

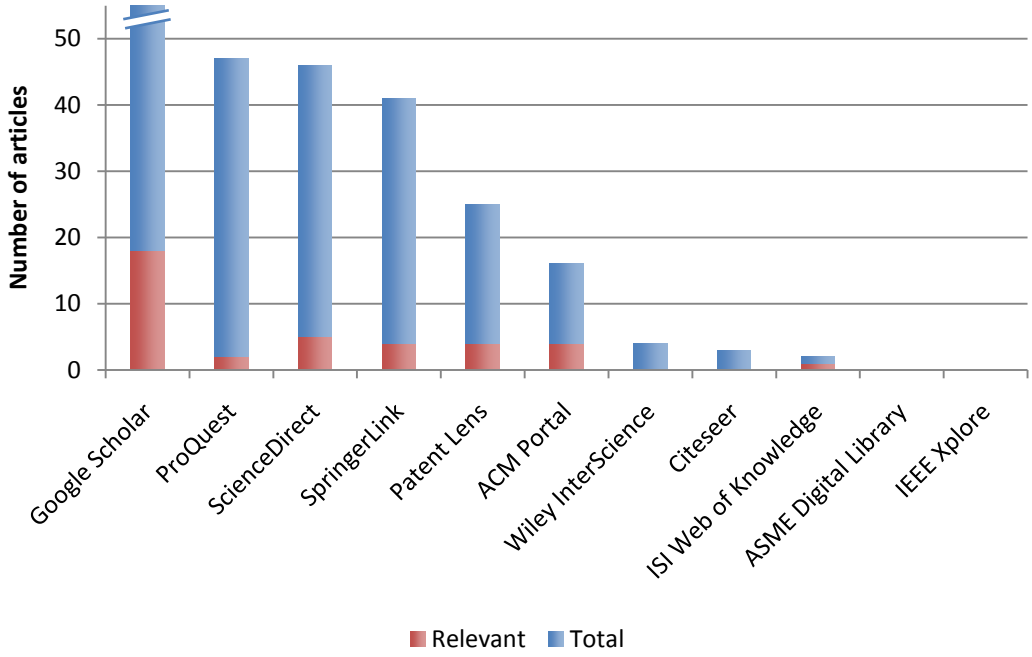


Figure 2.1 – Articles on asset management data warehousing

paragraph were used to refine the results before the results were perused. The initial conclusion drawn from this process was that there was little research in this area as the total number of relevant results encountered was less than 40 (and from these, less than 10 made more than a passing reference to asset management or data warehousing). Hence there would be a reasonably good chance of finding an open area of research.

2.1.3 Managing Information

References were managed with Endnote X (a commercial reference management software) in order to reduce duplication of reading and summarisation. All relevant references from the searches were stored within the software, and each reference was allocated a hierarchical topic label (e.g. data warehousing>>data modelling>>power utilities) for later use. Full bibliographic records including abstracts were stored in the Endnote database to facilitate future searches. This was conducted for all types of media, including web sites.

In all, more than 1000 papers[1] were scrutinised to compile this thesis and more than 200 were used as direct references for this thesis.

2.2 Review of Background Theory

2.2.1 Data Warehousing

A data warehouse can be viewed as a repository of data, created to store historical data of an organisation. Inmon [12] describes a data warehouse as a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

The data warehouse architecture provides logical centralisation by providing a single view of the data [13]. Having the necessary data in one location can reduce the response time for queries, leading to productivity and/or efficiency gains. Data are transferred from operational information systems to the data warehouse such that queries do not affect the function and performance of the operational systems. As a data warehouse is created from requirements provided by the organisation and individuals, the complexity of data warehouses can be large. Companies often find data warehouses to be time consuming and costly.

Although a relatively new field in information systems, literature on data warehousing is vast. Data warehouses are now becoming common as information sources for decision support systems. Winter [14] illustrates that data warehouses can act as a middle layer between operational databases (by performing extraction, transformation, integration, correction) and decision support systems (by performing selection, aggregation, and supplementation).

There is often contention about the differences between data warehouses and operational databases. Hammergren [15] indicates that the major difference between data warehouses and operational database systems is that operational systems are highly structured around the events they manage, whereas data warehouse systems are organised around the trends or patterns in those events. Other differences relate to the size, performance, content focus and tools associated with each system. These facts

[1] The 1000 minimum figure is based on the number of saved journal and conference papers. This does not include books and other physical media, nor does it include web-based articles.

have implications for the current research, as a large portion of asset management data comes from operational databases. To satisfy the data warehouse characteristics, the data warehouse design must focus on supporting queries rather than transactions. Data warehousing has also increasingly added denormalisation, as it improves the non-technical user's ability to navigate through the data [15].

2.2.2 Data Warehouse Lifecycle

Several approaches have been developed that attempt to describe the data warehouse lifecycle [16, 17]. Kimball et al. [18] described a seven stage approach: project planning, business requirement definition, dimensional modelling, physical design, data staging design and development, deployment, and maintenance and growth. Golfarelli and Rizzi [19] suggested a six step design approach: analysis of the information systems, requirement specification, conceptual design, dimensional schema validation, logical design, and physical design. All approaches to the data warehouse lifecycle follow the typical software development lifecycle where the requirements are firstly specified, followed by various design levels, implementation, testing, and then maintenance.

2.2.3 Data Warehouse Design Methodologies

Most data warehouse design is undertaken in an informal way, with fact metrics based on management strategies and goals. Influencing factors of these fact metrics become the corresponding dimensions in the data warehousing system. To formalise the data warehouse design methodology, several development approaches have been developed [20-22]: user oriented, operational oriented, and business process oriented data warehouse development.

User Oriented Design

Poe et al. [23] provided a more concrete method of data warehouse design by using interviews to obtain informational requirements. By interviewing different user groups, a complete understanding of the business can be gained. Ewen et al. [24] also preceded data warehouse design with a set of questions aimed at several key business managers. The planned data warehouse was for a healthcare agency and the questions spanned the areas of membership, benefits, demographics, care management, billing, and rates.

Operational Oriented Design

A more formalised data warehouse design approach is to examine the data models of the underlying operational or application systems to determine the applicable transactions. Golfarelli et al. [25] developed a methodology to derive a data warehouse conceptual model from existing entity relationship (ER) schemata which describe an operational information system. Böhnlein and vom Ende [26] also developed a similar methodology with operational information systems being represented by a Structured Entity Relationship Model (an extension of the ER model).

Business Process Oriented Design

Kimball et al. [18] highlighted the need for a business process oriented data warehouse design strategy. The advantage of a business process oriented development approach is that business process models formally describe the informational requirements of users. Böhnlein and vom Ende [20] proposed a methodology of deriving data warehouse structures from business process models. Using the Semantic Object Model approach to business process modelling, a four step process was developed: (1) identify the goals of the system, (2) analyse the business processes, (3) derive a conceptual object schema, and (4) identify initial data warehouse structures. While this methodology developed is sound in theory, it lacks formal rules and much of the derivation process reverts back to the experience and intuition of the user. Consequently, automation of the technique becomes difficult.

Commentary

The different approaches to data warehouse design relate to differing levels in structuring user requirements. User oriented design loosely structures user requirements, and can consequently lead to problems with the correctness and completeness of the requirements. Users cannot describe their requirements exactly, and with their unfamiliarity with data warehousing and OLAP systems, often do not have a complete picture of their demands [27]. Data models of operational information systems provide a more formalised view of the business, but the limitations they impose stem from their purpose. Data models are targeted for the development of application systems, and do not represent the business completely [20]. They also may not represent the full requirements of users, and in the case of external information, probably do not exist. Business process oriented design uses models that attempt to further formalise the business from different perspectives. The obvious shortcoming of

business process oriented design is the immense requirement of resources (time, human, and financial) in developing the process models.

2.2.4 Data Warehouse Modelling

There is no widely accepted standard for the modelling of data warehouses. The different methods that have been proposed include: Application Design for Analytical Processing Technologies [28], Stars [29], Dimension Modelling [25], Object Oriented Multidimensional Modelling [30], and starER [31]. There are two main schools of thought for modelling, each of which is supported by leading data warehouse advocates: Kimball [18] supports the multidimensional approach (including star and snowflake schemas), whilst Inmon supports the ER approach [32].

2.2.5 Data Warehouse Design Issues

Yao [33] provided a comprehensive summary of the design issues faced in a data warehousing project. These include the:

- Granularity (the level of detail or summarisation held in the units of data)
- Partitioning (the break up of data into separate physical units)
- Data (the three types of data: raw data, aggregated data, and metadata)
- Data sources (the operational systems from which the data warehouse is populated)
- Extraction, Transformation, and Loading (ETL) process (the data management process between the source systems and data warehouse)
- User access (the interface and analytical tools made available for the user)

Other important issues in a data warehouse project that should be addressed are:

- Data warehouse architecture [34] (the arrangement of the data sources, ETL, intermediate stores, warehouses, data marts, user access)
- Data marts [35] (localised warehouses with specific content)
- Conceptual, logical, and physical design [36] (three stages representing respective increases in detail)
- Data warehouse modelling and schema

The points above represent the different factors faced by data warehouse designers. There is research scope in examining these factors from a general asset management viewpoint and the topics that could be covered include defining appropriate

granularities for the different types of asset management data, determining an effective partitioning of data for the different areas of asset management, and investigating commonly used aggregate values. Investigation of some design issues is limited from the perspective of a general domain: items such as data sources, ETL, and user access are largely enterprise-specific and may require a case-by-case examination.

2.3 Review of Asset Management Data Warehousing

2.3.1 Asset Management System Integration

Although not asset management data warehousing specific, there are numerous published works that look at the integration of asset management systems in various industries, including manufacturing [37], mining [38], power utilities [39], transportation [40], and water utilities [41]. As data warehousing has integration at its core, there are several concepts in integration that are relevant for data warehousing, such as data model integration and data transformation.

The relevant works for asset management data warehousing are in the form of asset management system integration standards. There are several integration standards within asset management, including ISO 15926, MIMOSA OSA-EAI, and ISA-95. The most applicable is the MIMOSA (Machinery Information Management Open Systems Alliance) OSA-EAI (Open Systems Architecture for Enterprise Application Integration).

With the numerous asset management systems offered by different vendors, the process of integration can be problematic as many systems have their own unique data exchange interfaces. This leaves businesses facing a dilemma, as different integration techniques bring their own advantages and disadvantages. Purchasing systems from a single vendor leads towards system compatibility, however suppliers may not provide a total asset management solution, and the reliance on one vendor can prove risky. Businesses may purchase a generic bridge that integrates different systems, which may prove more cost effective than internally developing a bridge, but provides less customisation ability and requires updates for new system versions. Another option is to use an industry-standard bridge, which allows businesses to mix different systems with reduced integration costs. However, there may be performance loss compared to a custom solution, and in addition, vendors must be willing to support the standard.

The absence of a standard for asset management data exchange was a driving factor in the formation of MIMOSA and the subsequent development of the OSA-EAI. The MIMOSA OSA-EAI provides a standardised interface for data exchange of operation and maintenance (O&M) data. Consequently, it provides an ER data model named the Common Relational Information Schema (CRIS) that dictates how data should be stored in a database. The CRIS covers the operation and maintenance set of asset management data: asset registry, work/action management, condition monitoring, and reliability information. Current research is being undertaken on storing and communicating algorithm management, intelligent agent management, geospatial tracking, and capability forecasting data.

The MIMOSA OSA-EAI does provide a valuable information model for asset management, but it is also a model designed for an operational rather than an analytical system. The thoroughness of the standard is commendable, as it covers a large portion of the different areas in asset management and the continual revision bodes well for asset management systems. As the standard is relatively new, there has been no research into using it in a data warehousing design. However, it can only provide a starting point for a data warehouse system.

One key advantage of using the OSA-EAI is through software reuse. Software reuse is a technique that aims to develop components that can be implemented in multiple systems with slight or no modification. Reusable components increase the likelihood that prior testing has been undertaken, and that common bugs have been detected. Another significant advantage in reuse is the potential reduction in implementation time for not only the development stage, but also those of design and testing. As MIMOSA have undertaken a comprehensive analysis of the data elements involved in engineering systems, harnessing their effort avoids duplicating existing research.

Another key advantage of using the OSA-EAI is that of data interoperability. Many condition monitoring and operational systems store data in proprietary file formats, making it difficult to use the data with external programs such as analysis tools. Unless the software supports an export facility, organisations are effectively locked into particular tools. Open specifications for data management assist in interoperability between engineering asset management systems.

2.3.2 Data Warehousing in Asset Management

There are many references to asset management data warehousing projects conducted within organisations [42-44], however there is little ongoing scientific research in this area. Those that exist are tangential to data warehousing, and only use it as a means for research into other areas such as data mining [45]. An extensive literature search for research within asset management data warehousing was conducted and is presented below [1]. The analysis is segmented by industry and is largely chronological to show the trends within each industry.

Power Utilities

The majority of scientific asset management research springs from the utilities area and in particular, the power industry. As the power industry has been using SCADA (Supervisory Control And Data Acquisition) systems, EMSs (Energy Management Systems), and DMSs (Distribution Management Systems) in their substations for an inordinate number of years, it is a natural progression to attempt to integrate these systems into one centralised system, akin to data warehousing. Thus it is not as surprising that the majority of asset management data warehousing research originates from this one industry.

In terms of raw data size, the majority of asset management data in the power industry is produced by SCADA systems/EMSs/DMSs. Shi et al. [46] provided several potential data warehousing applications for these types of operational data and some of the challenges that the power industry faced in developing systems. Dahlfors and Trogen [47] continued with this idea by providing two power utility case studies from Europe that integrated SCADA systems and EMSs from a number of substations. The eventual data warehouse system was used for statistical analysis and reporting.

To perform a system-wide analysis of equipment from multiple power stations, Dolezilek [48] designed a health monitoring system that would combine equipment health data. Using DDE (Dynamic Data Exchange) and OPC (Object linking and embedding for Process Control) for communications, different end users from different departments could access the data through tailored interfaces.

[1] It must be noted that while the search is comprehensive, it cannot be logically asserted that all relevant papers pertinent to asset management data warehousing have been reviewed.

Werner and Hermansson [49] discussed the characteristics of a utility data warehouse, which represented a merger between a traditional data warehouse and an online SCADA system. As utility companies have huge amounts of SCADA data, the paper highlighted a need for companies to mine that data. The data warehouse acted both as an online historian (handling SCADA data), and as a general data warehouse for decision support (handling customer and asset management information). The fundamental design considerations included compression due huge amounts of stored SCADA data, and also increasing I/O performance through buffering and indexing.

An end-to-end process was presented by Lavalle [50] whereby a data warehouse system was created for a Mexican power utility. Generation, transmission, and distribution data was taken through an ETL process and loaded into the warehouse. A dashboard system was used for the visualisation of data by management and finding patterns and trends via data mining was the future step in the decision making chain.

As the data update cycle of traditional data warehousing systems is typically at a daily or greater granularity, intra-day reporting is not possible. He et al. [51] proposed an "active" data warehouse that acts as an almost real-time data source that is updated every second. SCADA/EMS data are stored in the system, and the data warehouse also contains rules that can alert operators to potential faults. While the ideal is admirable, the idea lacks any semblance of data warehousing, and is in fact, a normal transactional database under a data warehousing moniker.

While a lot of utility research focuses on putting operational data into the corporate data warehouse, McDonald [52] highlighted the importance of data warehousing non-operational data. While the term non-operational data is not specifically defined, it is assumed to be gathered from systems other than SCADA systems. The same 'non-operational' data classification terminology was used by Kezunovic and Latisko [53] who developed a similar data warehouse but at the substation level. Their main observations were that the SCADA system would determine the features and performance of the data warehouse, and that the level of data security had a significant role when designing the substation warehouse.

Thomas et al. [54] further expanded on the idea of data warehousing of non-operational data in power utilities for the purposes of health assessments. Non-operational data consisted of "records or logs of multiple events such as series of faults,

power fluctuations, disturbances and lightning strikes" while the most valuable non-operational data were "the digitized waveforms that reveal the occurrences during a voltage disturbance or a fault" [54]. The work identified which non-operational data would be warehoused and the user groups that would analyse the data, and conducted a case study using OPC for communications and Oracle for the data warehouse.

The main work in integrating a full spectrum of operational and non-operational asset management data into a data warehouse was produced by Draber et al. [55]. The data warehouse model identified 11 categories of functions: replacement, maintenance and diagnosis, performance, customer information, contract/tender, lifecycle cost, availability, network analysis, supervision, risk management, and power quality. Despite the comprehensive view of asset management functions, the presented analysis model did not harness the advantages of an integrated source of data, instead, processing data as if it were from its native transactional system. The presented case study also did not support the data warehouse model and only transferred operational data from a SCADA source.

A data warehouse was developed by the Mihajlo Pupin Institute for a transmission network company EMS in Serbia [56]. Both a SCADA system and a local technical database containing equipment data fed into the data warehouse via an ETL process. A difference with other systems was the handling of analysis – both the data warehouse and the original SCADA database were used in the decision support analysis, rather than solely the data warehouse. This is important as it highlighted that data warehousing is not a panacea.

Pathak et al. [57] advocated a different approach to the development of a power system asset management system. As a service-oriented architecture was used, each service obtained data directly from the data sources in a federated approach. They illustrated two advantages of this approach: (1) information was always up-to-date at the time of query, and (2) there was no single point of failure as in the case of a centralised data warehouse.

The majority of asset management data warehousing work in the power utilities industry involves the warehousing of operational data, and in particular, SCADA data. The actual mechanisms of the ETL stages are not discussed at length in any of the papers, and it is unclear how the authors deal with aspects such as resolving differing

time granularities and aggregation of high sample rate data. While there are ideas to introduce non-operational data into the data warehouse, the areas of this class of asset management data are quite limited.

Building Infrastructure

While the majority of literature in data warehousing deals with the asset management of dynamic systems, the management of static assets is also a major concern for engineering. As "significant parts of Australia's infrastructure are ageing and nearing the end of their economically useful lives" [58], research into building asset management is vital.

The Department of Housing and Works of Western Australia developed an in-house property and facilities management system in 2002 [59]. Using a data warehouse, it used an innovative approach to manage a list of asset management service providers in the market, as well as providing a tendering system for routine maintenance, general restoration and minor works.

The New York State Department of Transportation developed an "asset management information system" that despite its name, served as a data warehouse [60]. Updated on a daily basis, the warehouse allowed for ad hoc reporting of their highway and bridge management projects. Future plans were to integrate GIS (Geographic Information System), financial, and condition information into the system.

The Chicago Regional Transport Authority also developed an asset management information system for their transit infrastructure [61, 62]. It contained asset locations, conditions, demand, usage and capital improvement information in a web-enabled, multimedia data warehouse [63]. Using an off-the-shelf relational DBMS (database management system), it integrated with a GIS for spatial presentations of data. As opposed to other organisations keeping a tight rein on their corporate data, the data warehouse system did allow for public access in accordance to a data sensitivity policy.

Also following a web-based approach because of distributed users, the Department of Accelerated Rural Development in Thailand developed an asset management data warehouse [64]. The data was taken from several databases across two divisions and contained asset condition and maintenance planning, scheduling, prioritisation, and budgeting information.

A system for managing the operation of transportation assets was patented by Sroub and Mackraz [65] where routes could be dynamically recomputed based on the current state of a transportation network. Having implications within rail networks, the real time traffic data would be transferred to an offline data warehouse for analytical processing.

Maintenance is the common factor between the asset management data warehouses in the building infrastructure industry. As infrastructure assets are static, they do not have an operational nature like other industrial assets. As such, condition data as well as scheduling are the primary data sources for the building infrastructure industry. Spatial data also seems to play a vital role in building infrastructure data warehouse as it supports useful geographic visualisation.

Resources

The need for the integration of data within the resources sector has consequently given birth to many data integration initiatives, the most famous being ISO 15926. Data warehousing is another technique that is being explored, particularly within the oil and gas industry.

The need for open systems standards was expressed by Knights and Daneshmend [66] for the transfer of data between systems in the mining industry. Showing a data warehouse model that incorporated data from mine planning, blast design, maintenance, production control, and finance systems, the paper discussed the relation of the MIMOSA and POSC (Petroleum Industry Open Systems Corporation) standards to data warehousing.

The Kuwait Oil Company developed a system called FINDER which provided an integrated oil exploration and production database based on Oracle [67]. Serving as a corporate data warehouse, two key applications were developed based on the data contained within: production scheduling and production management.

In developing an integrated pipeline information management system, a European gas supplier built a real-time Oracle-based data warehouse [68]. Displaying Key Performance Indicators (KPIs) on a dashboard, it would show gas requirements, sales, statuses of contract fulfilment, and allow for planning based on past data as well as extrapolations of future trends.

As much of resource exploration revolves around location-based data, a spatial data warehouse was created for a multinational oil and gas company [69]. The data warehouse contained exploration data for wells, land and lease outlines, seismic lines, assets, and topographic imagery.

Conducting research into asset management data warehousing issues, Rudra and Nimmagadda [70] investigated the roles of multidimensionality and granularity in Australian oil and gas exploration data. Using a combined approach of ER and multidimensional modelling, they arrived at the conclusion that the data size was dependent on the technique used, and that the technique used was dependent on the content of the data.

In follow up research, object models were developed for oil and gas exploration data sets [71, 72] and their relations to multidimensional models were highlighted. The concluding thoughts of the authors envisioned a system that would incorporate other oil and gas data from areas such as drilling, production, and marketing.

Operation of assets in exploration and production appears as the common data theme within resources asset management data warehousing. More traditional data warehouse issues such as granularity and schema representation are addressed by this research, and there is a higher noted use of multidimensional modelling compared to other industries.

Manufacturing

Asset management for manufacturing companies must deal with the assets that are used in the manufacturing process, as well as the products they create and assemble. The manufacturing industry has specifically designed manufacturing execution systems that often combine functionality from equipment, inventory, operation, and condition systems.

In a proposal for a data mining system for fault diagnosis, process control, and quality control in manufacturing organisations, Büchner et al. [73] indicated that a data warehouse was vital in supporting the system. It would provide a mechanism to integrate the data from various database systems including product and process design assembly, materials planning and control, order entry and scheduling, maintenance, and recycling.

An information systems structure was developed for the manufacturing industry by Park and Favrel [74], to extend data sharing between manufacturing companies. Termed a virtual enterprise, the effort would allow for the manufacturing of products at a lowered cost, higher quality, with less risk, and shorter lead times. The key data source in this infrastructure was a data warehouse that would integrate data from the involved firms.

Organisations with dispersed manufacturing networks have similar properties to multiple organisations in a virtual enterprise. To increase the research knowledge in this area, Lau et al. [75] designed a multidimensional database for the purposes of data mining in a dispersed manufacturing network. Using OLAP and rule-based reasoning, the system integrated product and price, contact address details, design capability, quality control, and delivery time data. Star schemas were used as the multidimensional modelling technique and the system was prototyped on Microsoft SQL Server.

Although not specifically using the term data warehousing, Dabbas and Chen [76] developed a integrated database system for one of Motorola's semiconductor manufacturing laboratories. Passing through an ETL phase, equipment, manufacturing, product, and process data were integrated into a central database from which factory performance reports were generated. The model used for the database was a relational schema based in Microsoft Access.

As the ETL stage requires the majority of the resources in a data warehousing project, a quality based integration mechanism is needed. To fill the gap, Hinrichs and Aden [77] developed a data integration system for data warehouses that had ISO 9001 compliance. The system was paralleled against a manufacturing scenario with materials, production, and product data being integrated.

In a study of why some organisations receive exceptional benefits from data warehousing, Watson et al. [78] used a case study of an unnamed large manufacturing company. The data warehouse was supplied information from over 100 internal and six external data sources and included areas of logistics, manufacturing, and quality. As the project was initiated by the information systems department for ease of data management, the uptake by individual business units was not as strong as its potential.

An operator support system based on a data warehouse was proposed by Abonyi et al. [79]. Using an extensive amount of historical process data from process and quality control data sources, multivariate statistical analysis was conducted to determine relationships at a polyethylene plant in Hungry. Microsoft SQL Server was used as the back-end while Matlab with a database toolbox was used as the front-end.

Materials management is an important aspect of manufacturing as it is one of the first stages for the physical creation of a product. A data warehouse was used as the support infrastructure in a materials selection system proposed by Li [80] as it provided "(a) data integration; (b) data completeness and (c) decision-making support". The warehouse focused on the identification of physical properties of materials, their units, and their capabilities.

The integration of event-based production data with financial and/or purchasing information in manufacturing was patented by Pokorny et al. [81]. The production data consisted of manufacturing delay times and waste produced, while the financial and/or purchasing system could be in the form of a data warehouse system.

It is not surprising that the data warehousing research within the manufacturing industry largely focuses on the integration of process, product, and quality control data. As these form the largest data areas within a manufacturing organisation, it is natural to conclude that they would also receive the most attention. Data mining of this data takes a forefront in asset management data warehousing research, with data warehousing being used as a means to develop algorithms that will improve performance and decrease costs.

Telecommunications

Data warehousing has been used in telecommunications for numerous uses including billing [82], marketing [83], predicting customers churn behaviour [84], and traffic analysis [85]. The main assets of telecommunication systems are its electrical communications stations and network, and network management systems are the general category of systems that deal with asset management.

A large number of data warehousing and decision support system issues faced by the telecommunications industry was addressed by Conine [86]. Looking at different user group requirements, and the composition and integration of existing information

systems, the paper also addressed what areas of asset management should be included in a data warehouse system. These areas included requirements forecasting, inventory management, project management, asset utilisation, and capacity analysis.

Calvanese et al. [87] discussed a data warehouse design that was implemented by Telecom Italia. The data warehouse dealt with customer details, and only nominally touched on the area of asset management by having a link between a Customer entity and a Maintenance Contract entity.

Dealing with the integration of network management and inventory management systems in telecommunication organisations, Liang and Lanmann [88] published a patent that could be used for data warehousing systems. The patent covers the synchronisation between an online asset management database and offline inventory database for the purposes of network utilisation planning.

Despite the numerous applications of data warehousing outside of asset management, the telecommunications industry does not have a significant asset management data warehousing research base. However, forecasting utilisation and capacity are the important issues within telecommunications asset management data warehousing.

Physics Research and Nuclear

The Comprehensive Nuclear Test Ban treaty in 1996 led to the creation of a distributed radionuclide monitoring system that monitors the atmosphere for radioactive particles [89]. The 80 monitoring stations would send their daily results to a central data warehouse located in Vienna, Austria that would subsequently analyse the data to determine if any nuclear explosions had occurred.

As different companies are involved in the design, manufacturing, construction, and operation of nuclear services in Korea, there are various issues with the data sharing across the industry. In order to facilitate the sharing of design and O&M product data, Mun et al. [90] developed a neutral data warehouse that would serve as an information bridge. Acknowledging various product data standards including ISO 15926, the system uses the Nuclear Product Model, a product data standard for the nuclear industry.

While many data warehouse systems have been created by researchers for organisations, data warehouses have also been created to aid the purposes of research.

Product and workflow data were combined into an engineering data warehouse at the CERN particle physics laboratory in Geneva [91]. The information captured within included the mechanical design, calibration and maintenance of particle detectors as well as the tasks performed on particle components.

Conducting measurements is an important area within asset management as they provide a form of evidence for the operation or condition of assets. This is also true in the area of particle and nuclear physics, where the composition of materials are ascertained. These testing procedures generate a large number of numerical results, and issues within aggregation and data compression are important areas.

Defence

The three traditional defence organisations – army, air force, and navy – have a multitude of expensive assets, many of which are mobile. This presents a unique situation in their management, particularly as their successful operation can be the determinant between life and death.

In a proposed integrated vehicle health management system for aircraft, a data warehouse was designed to support offline post-processing of historical condition, usage, maintenance, diagnostic, and prognostic data [92]. The system was validated on a vehicle fuel system case study which was based on the OSA-CBM (Open System Architecture for Condition Based Maintenance).

Considerations were given to the development of a logistics data warehouse by the Australian Submarine Corporation [93]. The main reasoning was to move away from a transactional database to a purpose built data platform for their knowledge discovery techniques. The same intent of moving to a data warehouse platform was expressed by Draper [94] for a health and usage management system for UK military helicopters. However, the goal was to create a central data warehouse to increase data visibility for fleetwide analysis.

Operating history, maintenance history, and software and hardware configuration data were passed to an enterprise data warehouse used by the US military [95]. The system contained embedded sensors that could provide diagnostic and prognostic data to personnel to improve support, increase readiness, and reduce the lifecycle cost of assets.

A data warehouse was developed by the US Navy which supported a diagnostic data mining program for one of its aircraft [96]. Data was collected from three sources: a maintenance actions database, flight data repository database, and a spreadsheet on the correlation between the maintenance actions and flight data.

Asset management data warehousing in the area of defence has largely focused on diagnostics and prognostics to serve as an indicator for maintenance. Much of the data are collected from condition monitoring systems that can produce a huge amount of data at high sample rates. This has implications for the identification of useful data to store within the warehouse (i.e. feature selection), as well as developing aggregation and compression techniques to bring about a more manageable warehouse size.

Commentary

Asset management data warehousing is at the nexus of two domains – information systems and engineering – and the background of authors is typically from one of these two disciplines. From the literature, it appears that the approach to asset management data warehousing is dictated by one's background, and that those from an engineering discipline are less inclined to discuss technical data warehousing issues. However, the engineering-based research has a greater disposition to field-based case studies, while those from information systems-based research will adopt a more theoretical approach. Consequently, it appears that applying the field of information systems to the field of engineering has more worth to the body of knowledge, rather than the converse situation.

An interesting observation is that the papers describing asset management data warehousing are in clearly delineated industries. As seen in Figure 2.2, the power utility industry contained the most number of papers followed by the manufacturing and building infrastructure industry. One speculation is that the industries with a greater distribution are technologically mature to other asset management industries in regards to their information systems implementation. Hence the needs for integration have arisen earlier than other firms. As there are many common asset management functions between industries (e.g. installation, operation, and maintenance are universal activities), then it logically follows that most research by any specific industry will collectively benefit others.

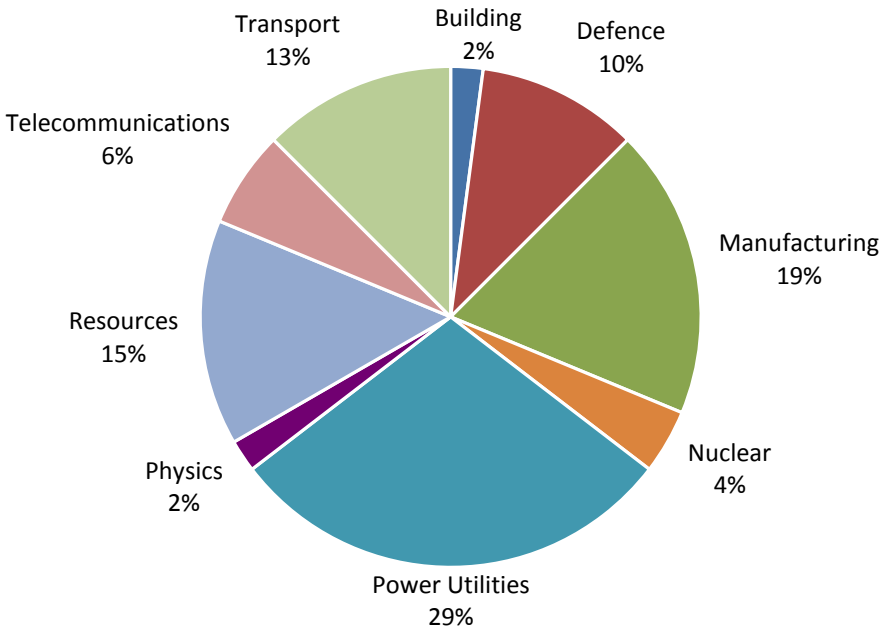


Figure 2.2 – Industry sectors of research

A characteristic observed is the focus of data warehousing of a particular type of data collected by that industry sector. The research in the power utilities industry directs their concentration on data warehousing SCADA and operational data; manufacturing upon process, product, and quality control data; resources and transport infrastructure on GIS and spatial data; and defence on condition monitoring data. A few efforts have tried to combine different areas of asset management data, however, the majority of the research focuses on data warehousing of select data areas.

Similar to the building infrastructure industry's use of spatial data warehousing, another way to visualise the number of papers is through geographic location. Figure 2.3 shows a breakdown of the papers that mention asset management data warehousing by locality of the first named author. The United States by far generates the majority of research articles with 21 papers, Australia second with five papers, and the United Kingdom and China third with three papers. These figures give an indication of each country's need for asset management integration and data warehousing (which are most likely to be constrained by research budgets). It is interesting to note the interest in the field shown by countries with a lower gross domestic product (GDP) per capita [97], including Chile, Serbia, Thailand, Mexico, and Hungary. This indicates that there is a universal need for asset management data warehousing research and not only by organisations in wealthy nations that have mature information systems.

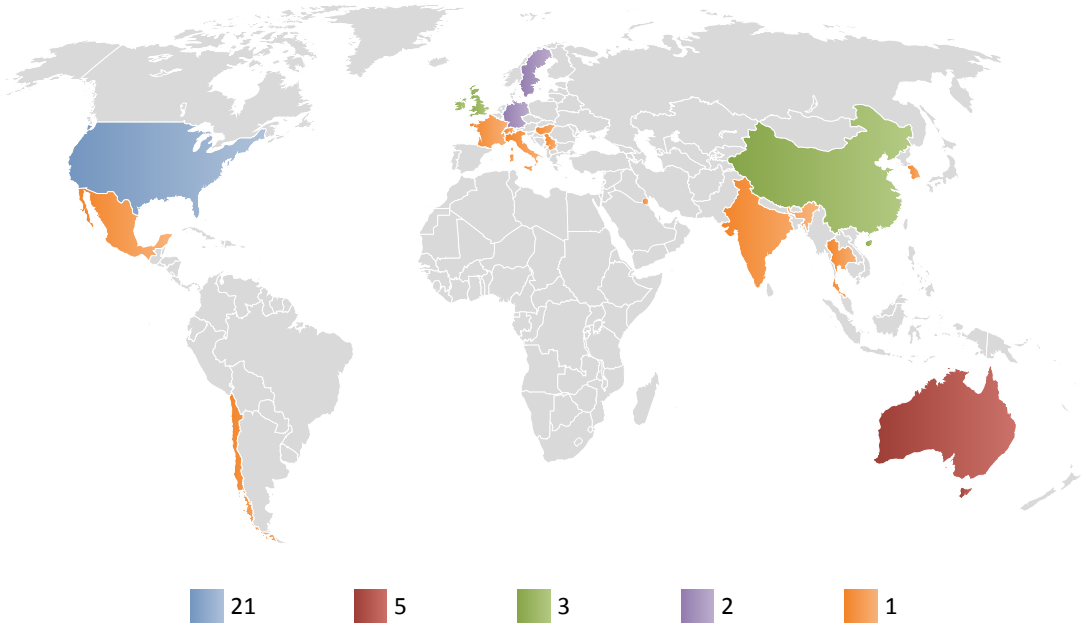


Figure 2.3 – Geographic distribution of research

Further segmentation of the geographical data shows that the United States had a mixture of industries: unsurprisingly consisting of the majority of the defence industry research, as well as participation by power utilities, manufacturing, and building infrastructure. Australia had a focus on resources (although this is due to three papers originating from the same institution), while the papers from the United Kingdom and China were from varied industries. When looking at the research papers into asset management data warehousing theory, the United States and Australia were tied at three papers each, followed by China, India, Sweden, and Switzerland – countries either with a high GDP or a high GDP per capita.

Another interesting observation is in the timing of research. Figure 2.4 shows a graph of the number of papers published that mention asset management data warehousing between 1997 and 2007. The field of data warehousing began in the mid-1980s and research ramped up in the early 1990s. The potential of warehousing asset management data started in the late 1990s, with the apex of research in the early 2000s, and has since dwindled.

Figure 2.4 also illustrates the number of research papers into asset management data warehousing theory as opposed to those applying asset management data warehouses as a means for other research. These twelve papers deal with specific issues in asset management data warehousing: of what does the field consist; what ETL problems can

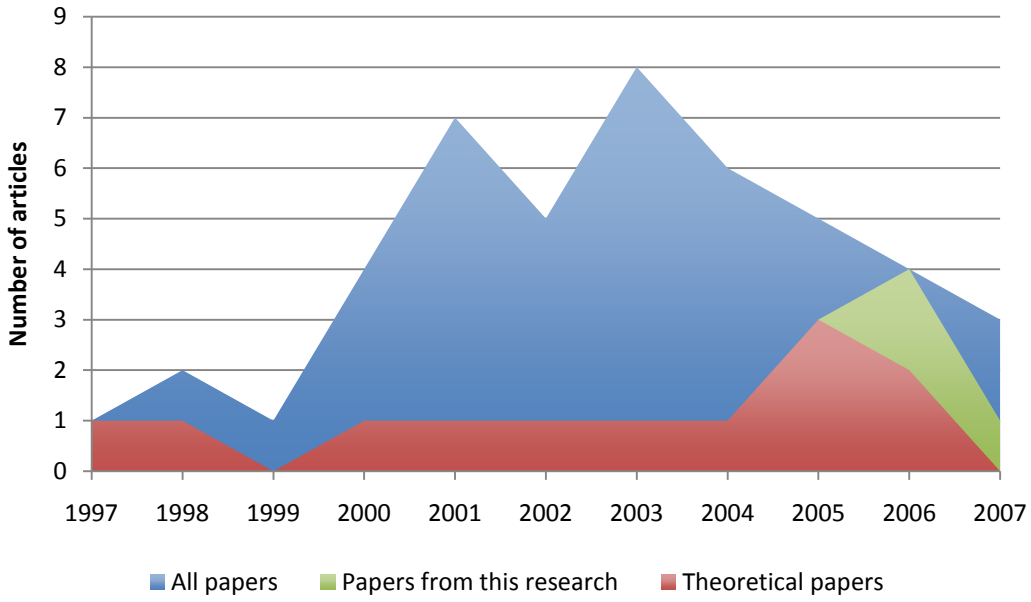


Figure 2.4 – Timeline of asset management data warehousing research

arise and how should these be handled; and how data from different sources can be integrated. By far, the majority of these are from the power utility industry (with the resources sector in second place) and primarily discuss issues with warehousing operational SCADA/EMS data.

Also illustrated by Figure 2.4 is the papers contributed through this PhD research. Three theoretical papers in asset management data warehousing were contributed after 2004, raising the tail end of the distribution. These papers are distinguished as to highlight the impact of this research upon the body of knowledge of asset management data warehousing.

As seen in the above research papers, while there are many indications of asset management data warehousing in use, there is very little research into asset management data warehousing itself. The two common conclusions made on the scarcity of literature are either (1) the area is new and there truly is little research, or (2) others have found that the topic is not worthwhile investigating and have consequently not published. The second premise is discounted as there is research in the area that points to the potential benefits of data warehousing. The first premise is reinforced by the fact that both data warehousing and asset management are relatively new fields in their respective areas and it is understandable why there is little research literature in the combined area of asset management data warehousing.

From the twelve papers that constitute as significant research into asset management data warehousing, three papers survey the area of data warehousing and attempt to relate it to their target industries [66, 80, 86]. These papers broadly review data warehousing specific issues such as data marts, ETL, data formats, potential data and systems to be warehoused, and decision support systems. These papers lack technical details about their industry-specific issues, and much of the material presented is repeated between them, with small injections of industry-specific issues.

A large section of the research within asset management data warehousing focused on incorporating operational data [46, 49, 51, 52, 55]. Most operational data are numeric time-series data, which can have varying degrees of sampling rates. Systems with high sampling rates can acquire massive amounts of data, and particularly when integrating data from many of these systems, issues such as data compression and performance are pertinent. While both Shi et al. [46] and He et al. [51] were aware of data compression issues, only Werner and Hermansson [49] addressed the topic in detail by looking at two different techniques. As a form of lossy data compression, feature extraction is one area that has not been fully researched in operational data warehousing. This is where only meaningful sections of the data set are stored, while the rest is discarded. While an attempt was made to identify non-operational data for integration with operational SCADA data [54], the resulting implementation still only contained data from the SCADA system.

The primary use of data warehousing is for knowledge discovery and data warehousing has brought OLAP tools to the fore. This is another research area that has not been investigated and multidimensional analysis of operational data remains an open area. As multidimensional analysis first requires multidimensional data models, this research investigates the latter in Section 5.

The most relevant technical works in asset management data warehousing all originate from Curtin University of Technology in Western Australia for the resources industry. Fundamental data warehousing issues of multidimensionality, granularity, star schemas, attribute hierarchies, and OLAP cubes are investigated [70]. A comparison was also drawn up between ER, multidimensional, and object models [72] as well as star and snowflake schemas [71]. The data set consisted of spatial data of oil and gas wells used in surveying operations, rather than the engineering assets themselves.

Thus there is still scope within the resources sectors to look at asset management data warehousing issues.

An examination of the innovations presented by these twelve works can be found in Appendix B. Apart from the two earliest papers ([66, 86]) that identify the applicability of asset management data warehousing, only six papers have any real innovation in the field. Much of the innovation comes from identification or proposals of IT topics to asset management (e.g. inter-disciplinary data integration [55], data compression and I/O performance for large data sets [49], standards for integration [52], and multidimensional models [72]). Only two works present detailed implementation and testing methodologies: (1) comparing ER and multidimensional models [70], and (2) a SCADA data warehouse implementation [70]. However these two works do not present sufficient detailed methodologies in replicating their work. When compared with the available data warehouse theory, the previous asset management-based research is fairly rudimentary and any innovations seem quite insipid.

2.4 Implications

From a comprehensive review of the available literature, it can be seen that asset management data warehousing is an area that is ripe for research. While there are numerous indications of data warehouses being applied in engineering contexts, the inclusion of asset management data is fewer, and the amount of pure research into asset management data warehousing is relatively very small. Current data warehousing research in asset management has a myopic view of data and integration, as it only focuses on a few or even just one data area at a time.

By virtue of the sparse amount of research into asset management data warehousing, work into nearly any data warehousing area (e.g. architectures, data modelling, ETL, metadata, OLAP, and visualisation) could be considered original. The challenge is to supplement such originality with innovativeness and significance, and only a few of the past theoretical contributions have these two attributes. The research questions 2-4 presented in Section 1.4 provide suitable scope in providing an original and innovative contribution, and the work in this thesis would be considered as the first real substantial and significant work into integrated data warehousing for asset management.

3 Asset Management Data Management Survey

The last 30 years have seen information systems becoming more pervasive in asset management operations within many organisations. Information systems have progressed to cover a wide variety of areas including asset registration, financial management, process scheduling and control, materials management, maintenance management, condition monitoring, risk management, reliability management, and safety management.

Due to the plethora of systems and unique combinations of information systems within organisations, past investigations into information systems have targeted specific industries or specific systems. Research emphasis within asset management itself has been placed on a few select areas such as control systems, maintenance, condition monitoring, and reliability. This has led to a disparity in the level of research into information systems across the whole of asset management.

The last two decades have also seen an influx of the use of computer-based technology into asset management as breakthroughs significantly increase their functionality and subsequent adoption. Advances in computational power have paved the way for harnessing complex algorithms for the analysis of operation and condition data. Research into database technology has allowed huge volumes of data to be collected and processed as well as spurring on the advent of the data warehouse. The Internet has brought the benefits of information sharing and accessibility to the fore, and corporate system integration and workflow management is now being addressed in current research.

Understanding the adoption and use of the aforementioned technologies allows the IT industry to strategise on where to focus future research and development effort for asset management systems. However, with the availability of competing technologies compounded with a mixture of organisation technology adoption strategies, it can be difficult to identify the current state of technological usage in asset management.

This chapter surveys the current state of information systems and data warehousing in corporate asset management by examining the use of information systems, the justifications behind their use, their integration, data warehousing, and data retention. A questionnaire forms the primary tool of investigation, and the survey methodology and the results analysis in this chapter are exclusively devoted to the questionnaire. Interviews with several organisations are used to confirm discoveries in the questionnaire results analysis. The results provide a rationalisation for the research topics in this thesis, and show the data management needs of the asset management industry.

3.1 Research Design

3.1.1 Research Questions

To understand how data warehousing in asset management was used in the corporate arena, the research questions that directed the survey design were:

- What is the composition of information systems in asset management operations?
- Why are some systems used while others are not, and what improvements can be made to current asset management information systems?
- How is the success of an asset management information system measured?
- What is the level of integration between these information systems?
- What data are regularly discarded and why?
- What is the level of asset management data warehousing activities in organisations?

The research questions cover a broad spectrum of information system areas and provide sufficient scope to be able to delve into a respondent's reasoning. The questions provide a framework to understand the effects of information systems and data within asset management operations in organisations.

3.1.2 Research Methods

This survey used a two-step triangulation methodology. Firstly, the research questions were addressed through the use of a structured questionnaire to form a general impression of the area. Secondly, interviews were conducted in order to extract organisation and industry specific knowledge in addition to providing a validation to the results of the questionnaire.

3.1.3 Survey Type

The research questions indicated the need for current information within asset management and this requirement was the basis for the selection of an exploratory cross-sectional study rather than a longitudinal one. While concerns have been raised about the lack of longitudinal studies in information systems research [98], the timeframe of business process and information system change is typically longer than the period that is open for postgraduate research. Thus the survey employed a cross-sectional design to examine the research areas at a single point in time. However, longitudinal comparisons are made through the survey analysis (Section 3.4) where data are available from past, non-related surveys.

3.1.4 Unit of Analysis and Respondents

An individual organisation was selected as the unit of analysis. Asset management operations are typically facilitated by different departments within an organisation. As many organisations have centrally accessible information systems where an IT department administrates these systems, the smallest common granularity is the organisation. In the case of subsidiary organisations that have ownership, management, and use of independent information systems, these subsidiaries are considered as distinct "organisations".

The targeted respondents were those individuals who could successfully represent their organisation. Personnel with comprehensive knowledge about the information systems in their firm were targeted; particularly those personnel who had a background in asset management and/or information technology. Two survey questions were included to elicit the organisational roles of the respondent to understand their approach to the questionnaire.

3.2 Sampling Procedures

3.2.1 Sampling Type

Both purposive sampling and snowball sampling were used. Purposive sampling involves choosing a sample frame based on predefined characteristics. Organisations that undertake asset management are exclusively investigated, while personnel with knowledge of asset management and/or information systems are targeted. Snowball sampling involves asking the original sample frame to forward the survey to any

suitable participants or recommend potential participants, thus producing a 'snowball' effect. As the fields of asset management and information systems are diverse, allowing respondents to entrust the survey questions to a more suitable colleague would increase the response rate.

Both purposive and snowball sampling are non-probability sampling methods as they involve a non-randomly selected sample frame. While probability sampling methods mitigate sampling error (the sample does not accurately represent the population), such methods require detailed statistics about the population [99]. As complete statistics for the population of asset management organisations around the world are incredibly difficult to acquire, non-probability methods remained the only alternative. Despite a view that non-probability methods cannot be used towards inferential statistics, others express an opposite but conditional view that non-probability methods can make generalisations about a population if they are made with additional expert knowledge [100]. The interviews used in triangulation served as part of the required expert knowledge.

3.2.2 Sample Frames

The ideal distribution of a survey is to include every member in a population. However, this can cause difficulties in the administration of a survey if the population is very large. Hence a sample frame is selected to accurately represent the population. In the case of organisations that conduct engineering asset management, the total worldwide population is extremely large. The Australian Bureau of Statistics suggests that there are more than 529,000 organisations in Australia alone that undertake asset management activities [101]. The majority of organisations in the world have engineering assets, albeit at different scales, management levels, and support by information systems. Thus a more manageable sample frame needed to be chosen.

Two sample frames were selected: the first was member companies sponsoring a US-based research organisation – Center for Intelligent Maintenance Systems, and the second was companies affiliated with the Australian research organisation, CIEAM. These sample frames were selected on the basis of vested interest into this research. Affiliation was determined by past, current, and future membership, in addition to any past official (e.g. meetings and seminars) and non-official interaction (e.g. emails). It was expected that the selection of the sample frame would introduce sample biases towards organisations interested in research.

The Center for Intelligent Maintenance Systems (IMS Center) is a three campus National Science Foundation Industry/University Cooperative Research Center located at the Universities of Cincinnati, Michigan, and Missouri-Rolla. At present, the IMS Center is supported by more than 30 member companies in its research into advanced diagnostics and prognostics of machinery. The industries of these organisations include automobile, aerospace, heavy machinery, electronics, semiconductors, mining, and professional services. Thus a diverse range of companies were included in the sample frame. The interest expressed by the IMS Center in this research was in understanding the information systems at their member companies, and their involvement with information standards. The information would be utilised in interoperability and standardisation projects for their Watchdog® platform.

As mentioned in Section 1.1, the Cooperative Research Centre for Integrated Engineering Asset Management is also a multi-university-based research centre with a focus on industry-directed research into an integrated approach to life-cycle engineering asset management. Although it is supported by less industry participants than the IMS Center, it has a broader research focus and as such, has a greater number of associations with various industry groups.

3.2.3 Representativeness of Samples

Purposive sampling was used on the first sample frame by selecting companies upon the basis of their attendance at an IMS Center Industrial Advisory Board meeting. The questionnaire was distributed to each meeting participant with 120 individuals receiving the form. There were 58 registered participants who were attending on behalf of member companies (with multiple participants per member company), while the rest were visitors from academia and industry. Three sponsor organisations did not attend the meeting.

The second frame also used convenience sampling to select potential candidates by looking at business contacts of personnel affiliated with CIEAM, as well as an email distribution list. As the distribution list was owned by CIEAM, there were concerns regarding the Australian Privacy Act which binds the types of information that organisations can provide to third-parties. To avoid breaching the Act, the researchers could only access the people on the distribution list indirectly.

In order to direct the survey to suitable companies, each potential company was labelled as either “asset owner”, “asset supplier”, “software supplier”, “consulting organisation”, “news and research organisation”, or “academic organisation”. While these labels do not form a complete domain of values, they were sufficient to filter the contact database into the final sample, where the contacts in the first two categories (“asset owners” and “asset suppliers”) were selected.

The accuracy of generalised results from the samples to the sample frames was expected to be reasonable due to the relative size of the samples. As records of membership were well managed and most meetings involved the exchange of contact details, the samples well reflected the frames.

Snowball sampling was used on the second sample by asking recipients to pass on the survey questions to other people in their organisation if they themselves were not applicable. While traditional snowball sampling does not mandate intra-organisation forwarding, this restriction was used to make the survey appear more exclusive in an attempt to raise the response rate.

3.3 Data Collection

3.3.1 Development of Questions

The questions asked of participants were developed directly from the research questions. For inspiration on question design and answer, numerous questionnaires including those within and outside of asset management/information systems were analysed.

The eventual 22 questions shown in Appendix C consisted of closed ended questions as preliminary interviews and research had predicted the most common responses. The questions were grouped into topics based on their subject matter, and each topic was given a heading. The ‘Your Organisation’ topic gave clarifications on how to answer the questionnaire if the organisation was a division of a larger organisation, while the ‘Data Warehousing’ topic gave a short description on data warehousing. The initial questions were designed to be easier to answer such that the respondent would feel more comfortable with the questionnaire. An ‘Other’ response was placed for questions where the set of answers did not comprise of the whole domain of answers, and an ‘Unknown’ response was included for instances where the answer was not known. In

retrospect, an 'Unable to answer' response may have been useful for instances where the respondent could not answer the question due to legal, business, or privacy issues.

The questions, order, and format were continually reviewed over a period of six months. The reviewers were selected from both academia and industry, and their backgrounds included engineering, information technology, communications, business management, and psychology. The primary modifications were to the question design and wording, while ancillary items such as layout and style were largely untouched from the original design. A pilot study with 26 people from ten companies was conducted, giving feedback on the phrasing of several questions and answers.

3.3.2 Implementation

As two different samples were selected, the implementation of the questionnaire was divided across two media. The first implementation was in the form of a paper questionnaire that could be physically delivered to respondents. Microsoft Word was the tool of choice after deliberating on a variety of software packages. The tool provided sufficient functionality in question layout and drawing of form elements. As the questionnaire was to be delivered on US Letter-sized paper, several questions involving tables were physically constrained due to the paper and vertical margin width. While this constraint did not limit the question design, it did affect the aesthetics and, possibly, the text readability. The final paper questionnaire was five double sided black and white pages in length.

The second implementation was that of online forms that would be viewed via the Internet. While there were several vendors who provide online questionnaire design and hosting services (e.g. SurveyMonkey, QuestionPro, and Zoomerang), the questionnaire used custom developed HTML forms that would be hosted on a private server. This provided flexibility that was not available with commercial services.

The HTML forms were designed in Macromedia Dreamweaver and manually optimised by hand. The optimisations, while not necessary for satisfactory rendering by a browser, cleaned the code such that it could be validated as HTML 4.01 Transitional compliant by the W3C Markup Validator. This goal was important for consistent presentation across different browsers. Presentation of the questionnaire was tested against three browsers – Microsoft Internet Explorer 6 & 7, and Mozilla Firefox 2. During the period of the survey, browser statistics showed that these three platforms

covered about 91% to 92% of the browser market [102]. Screen resolution was tested at 800×600, 1024×768, and 1280×1024 as 68% of browsers are set to 1024×768 or lower [102].

Cascading Style Sheets were used to separate presentation and content thereby simplifying the development process, as well as reducing file sizes (the largest page being 45KB). A difference to the paper version, colour was introduced by using a blue hue on most elements. Blue was chosen as it is the most popular consumer colour and provides a more relaxing and comforting environment [103]. Each question block used a gradient fill to improve the aesthetics, and a larger than normal 12 point sans-serif Arial type face was used for text to improve readability for the older target demographic.

The HTML forms were divided over five pages so as to not overwhelm the respondent [104], to minimise any potential data loss [105], and to be able to track the progress of each participant. Welcome and thank you pages were used to introduce the questionnaire and thank the respondent for their participation. Each question page was made to contain a page completion indicator, as it has been shown that users who can view their progress are less likely to drop out from a web questionnaire [106]. “Previous Page” and “Next Page” buttons also allowed the respondent to traverse between pages, with the state of the form controls being saved per session. As HTML forms do not inherently support this capability, a PHP module was used. PHP was selected as a scripting language due to familiarity with the system.

Tectite FormMail was trialled as an out-of-the-box PHP form processor, which would securely email the multi-page form results. It was selected based upon its reputation for functionality and customisability. However, due a host of incompatibilities with the available web servers, FormMail was dropped in preference of writing a form processor from scratch. The PHP script would clean the answers on each page (to prevent spam attacks), and email the results. The developed script also allowed for more flexibility in customising potential responses based on past responses (e.g. hiding data areas in Question 20 based on Question 6).

One advantage of online questionnaires over traditional paper questionnaires is the ability to provide immediate validation of answers, to create branching questions, and transform answers of questions. Validation was performed by correctly using radio and

check box controls – radio controls restricted input to select answers while check box controls allowed multiple answers. Javascript was also used to ensure that the choice 'Other' required selection before a respondent could type their answer (e.g. Question 18). Branching questions and transforming answers were implemented using Javascript to show or hide the whole question or elements of the question. For example, Question 11 asks if the respondent's organisation operates a data warehouse. If answered with a positive, Questions 12 and 13 were shown, otherwise if answered with a negative, Questions 14 or 15 were shown (while Questions 12 and 13 were hidden). This allowed for less confusion and clutter on the page while shortening the perceived length of the questionnaire. Another example is Question 6 which asked if an organisation uses information systems in a particular area. For those areas marked with 'Yes', Question 7 would display those areas, but Question 8 would not. For those areas marked with 'No' or 'Unknown', Question 7 would not display those areas, while Question 8 would. Thus mutable questions could be asked to avoid illogical responses. Due to differences in server-side and client-side scripting, transforming answers for questions on the same questionnaire page used Javascript (e.g. the above example), whilst transforming answers over different pages used PHP (e.g. Question 20 based on Question 6).

Although 94% of browsers have Javascript enabled [102], the questionnaire was also designed to safely fail for those without Javascript. As client-side scripting was used to direct the question flow of the questionnaire, those without Javascript would see instructions that directed users to the correct question based on their responses (while these instructions were hidden for those with scripting enabled). For example, Question 11 instructs users to the appropriate question based on their response, while Question 16 instructs users to proceed to Question 19 if 'None' or Unknown' is selected.

While both Javascript and PHP could be used to enforce mandatory responses (either per page or per questionnaire), this capability was not employed. Per page mandatory responses were avoided to allow participants to browse the entire questionnaire, while per questionnaire mandatory responses were avoided to allow participants to abstain from responding to particular questions. Combined with the back and forward navigation buttons, and allowing comments to be added on each page, this functionality allowed the online version to retain the advantageous properties of a paper version.

3.3.3 Administration

The paper questionnaire was inserted into the back of a ring binder which was handed to IMS industry participants at the IAB. To provide an incentive in filling out the questionnaire, each respondent was offered a two week trial to a patent search service sponsored by MappGlobal IP Strategy Services. Several times during the meeting, the participants were reminded and encouraged to complete the questionnaire, and a collection box was placed at the back of the room for completed questionnaires.

The online questionnaire was setup at http://www.cieam.com/is_survey. The questionnaire was hosted on a CIEAM domain and had endorsing logos from CIEAM, QUT, and IMS for authenticity. The URL (Uniform Resource Locator) was kept simple to assist readability and memorisation. A link from the index page of the CIEAM site to the questionnaire allowed access for unsolicited visitors. One issue with the web server was uptime. As the CIEAM website had a tendency to occasionally go offline (due to a variety of uncontrollable factors), there may have been instances of lost responses.

A list of potential respondents was sourced from business cards that had been collected over a number of years. As these businesses were of different types including asset owners, manufacturers, consulting, and research, each of the 400+ contacts were filtered based on their suitability.

All emails originated from a QUT domain as opposed to a free email account for added authenticity. Using Infacta GroupMail 5, an email template was developed (see Appendix C) that personalised the name, organisation, and a clickable link to the questionnaire (e.g. http://www.cieam.com/is_survey/?i=1 with increments to the value of i). Each email was addressed individually as personalisation of correspondence has been shown to improve survey response rates [107]. No cut off date was specified in the email as most people respond soon or not at all and a date dissuades late responses [108]. The email was sent on a Wednesday at 11am as this day of the week and time of the day has the highest read and click through rate from all business hours [109]. As potential respondents were located in different time zones, emails were scheduled to stagger delivery.

To increase the response rate, multiple people within the same organisation were offered to complete the questionnaire. As completing the organisation field was optional, the personalised link to the questionnaire was used to match the response to

the participant database. For the paper questionnaire, only two responses were from the same organisation. These two were identical – presumably the two participants were sitting next to each other, and provided exactly the same responses. For the online questionnaire, there were three cases where multiple people from the same organisation completed the questionnaire. A merging strategy was followed to combine records that contained different responses. For responses with 'Unknown' values, any known values took precedence. For responses that asked if systems existed, positive responses overrode any negative responses. For the subjective questions, the respondents were contacted in order to achieve a consensual view.

Another method to increase the response rate was sending a follow up email. In order to maintain the appearance of personalised correspondence, the follow up email was styled to look like a Microsoft Outlook reply to the original email. The style matched the same font families, font styles, and colours. As replying to an email places the email header fields (from, sent date, to, and subject) in the body of the message, the email address of the recipient was placed in the "to" field, while the "sent date" field was removed as there were no time records of the original email sent.

3.3.4 Collection

All results were placed into a Microsoft Excel spreadsheet. Each response was allocated a column, with the total number of columns equalling 96. The difference between the 22 questions and 96 columns is that each check box question was allocated a column for each response, increasing the total. If the check box was ticked, then a value would be placed in the column, otherwise the column would remain blank. Each paper response was manually typed into the results spreadsheet. After entering the data, a review was conducted to verify that the responses were entered correctly.

Each electronic response was emailed to the researchers. Each page visited on the questionnaire caused an email to be returned with the current temporary results, while the final questionnaire page caused an email to be sent that contained Comma Separated Value (CSV) data. This could be copied and pasted directly into the spreadsheet to eliminate manual transfer errors. As text field responses could contain commas entered by the respondent, there was a potential for confusion with the text-to-data function within Excel. This function takes a CSV string and converts it to Excel's internal column format. Thus double quotation marks were placed around all text field data to avoid potential errors.

Questionnaire metadata was also collected using PHP environment variables and stored with each response. These included the time of submission, the browser being used, and the IP address. These data were used to correlate each page submission with the whole questionnaire, as multiple participants could cause questionnaire emails to be received intertwined. The identifier in the personalised link was also stored for comparison against the contact database, and the status of each response was inserted as 'Complete' for those who completed the entire questionnaire; or 'Viewed until 2nd page', 'Viewed until 3rd page', 'Viewed until 4th page', or 'Viewed whole survey' depending on the extent of perusal by a participant.

3.4 Survey Analysis

3.4.1 Questionnaire Response Rate

The questionnaire delivered to the IMS Industrial Advisory Board meeting resulted in ten responses. When comparing the number of responses to the total number of questionnaires delivered, the response rate was 8.3% (8/96). When comparing the number of responses to the total number of member companies attending, the response rate was 21.6% (8/37). In either case, the response was fairly poor.

When examining the reasons behind the poor response rate, four factors were identified:

1. As the questionnaire was distributed on five double sided Letter pages, the length may have been dissuasive to recipients. Combined with the short time span in which to complete the questionnaire and the lack of opportunities to consult colleagues for information, recipients may have felt overwhelmed and decided not to participate.
2. There might have been a possible mismatch of questions to the target respondents. The recipients of the questionnaire were largely in management roles and from engineering backgrounds. Their involvement with the information technology operations in their organisation may be limited, and hence the recipients were unable to complete the questionnaire.
3. Not all companies attending the meeting were asset owners, with some being software development companies, others reselling equipment, and others providing professional services. As there was no way to distinguish in advance who would receive each binder, the questionnaire was administered to everyone. Thus

the 'real' response rate would be higher than the above values when removing the non-applicable firms.

4. There may be an unwillingness to share such information lest it detract from an organisational competitive advantage as responses could be used to infer the strengths and weaknesses of a firm. Although providing contact details and the organisation name was optional, respondents may have felt that the risk of completing the questionnaire was too great.

The response rate for the web version of the questionnaire was considerably greater. Table 3.1 shows a breakdown of the targeted recipients into categories of applicability, responses, and views. A total of 182 targeted emails were sent to 116 companies. A screen of the potential participants was conducted beforehand to eliminate non-asset owning organisations as well as individuals who might have left an organisation. Despite these efforts, 55 individuals were not applicable either because a reply to the original email request indicated that the organisation was not in the designated population, or the individual had left the organisation and the email address was invalid.

The individual response rate was 37.0% with 47 responses, while the organisational response rate was 49.4% with 43 responses. While there are no universally agreed upon minimum acceptable rates for surveys, in a analysis of survey response rates in academic literature, Baruch [110] listed a range of response rates between 10% to 96% with a mean of 55.6%. Over the period of analysis (1975 to 1995), the mean response

	Individuals	Relative %	Companies	Relative %
Total	182		116	
Applicable	127	69.8[a]	87	75.0[a]
Responded	47	37.0[b]	43	49.4[b]
Viewed	19	14.2[b]	17	18.4[b]
Responded or viewed	66	52.0[b]	60	69.0[b]

Table 3.1 – Online response/viewing rates

[a] Relative to total
[b] Relative to applicable

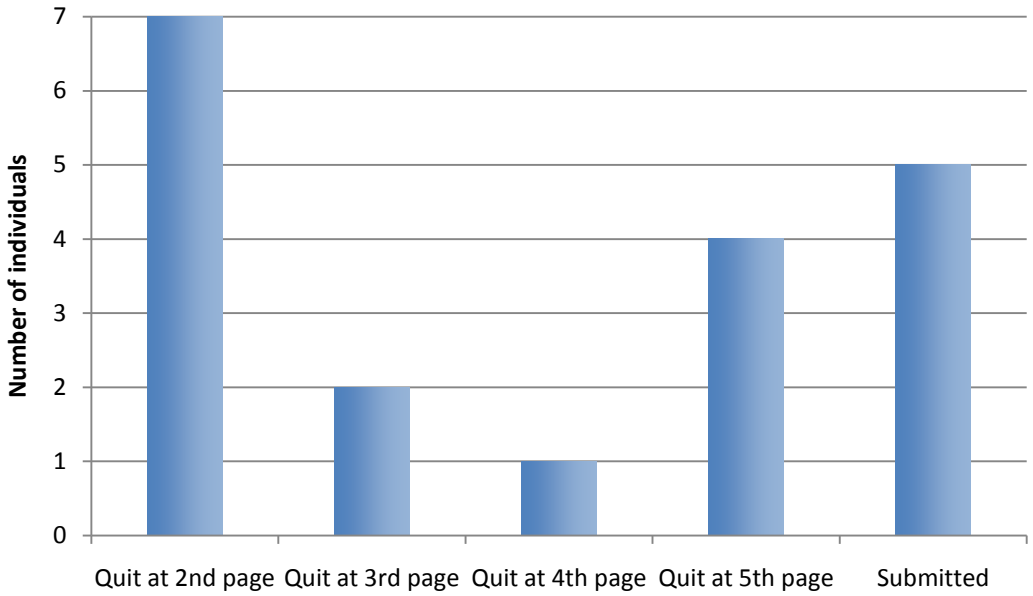


Figure 3.1 – Questionnaire viewing habits

had declined from 64.4% to 48.4% [110], while another study showed that the mean response rate for online surveys was estimated to be 39.6% [111]. The response rate in this study, while just above average, is less important as the questionnaire is being used as a tool to gain insight [112] rather than serve as a descriptive instrument.

Because of the electronic nature of the questionnaire, the number of views could be recorded by monitoring the progression of a participant through the questionnaire. A view was constituted as a perusal of at least one page the questionnaire that may or may not have been submitted. It contained either blank or clearly erroneous data with the latter used to trigger the JavaScript functionality (see Section 3.3.2) to view the branching questions. In total, 19 individuals viewed the questionnaire and Figure 3.1 shows the majority ended the questionnaire at the second page. As the second page provided an adequate representation of the topical content, it provided a suitable decision point for an individual to proceed with viewing the rest of the questionnaire. Thus individuals either quit early (if the survey was not relevant or they were unable to answer) or quit at the end (if they were interested in seeing the type of questions asked).

The combined number of responses and views signifies the total number of individuals that responded to the sent email. Over half the targeted sample clicked on the link

provided in the email and viewed any part of the questionnaire. The rate is surprisingly low and there are several possibilities of why this may be the case: the survey may not have been relevant to the recipient; the individual may not have had time or was unable to undertake the survey; the email may have reflexively been categorised as spam and consequently deleted; or the email may have been categorised as spam and discarded by automatic rules and filters.

3.4.2 Metadata Analysis

While the actual question responses constitute the primary focus of the survey, there are interesting data that can be analysed from the web-based questionnaire. The data were obtained from the web server environment variables and can be considered as questionnaire metadata. Among a host of different variables, the metadata include the operating system and environment, IP address, browser, and time of page request.

Operating System and Browser

All respondents were using a Windows-based operating system, with the majority using Windows XP. There were eight respondents who used Windows 2000, and one who used Windows NT. Considering Windows NT was released in 1996, it is a strange anomaly to see an old unsupported operating system being used on what would supposedly be a business desktop computer.

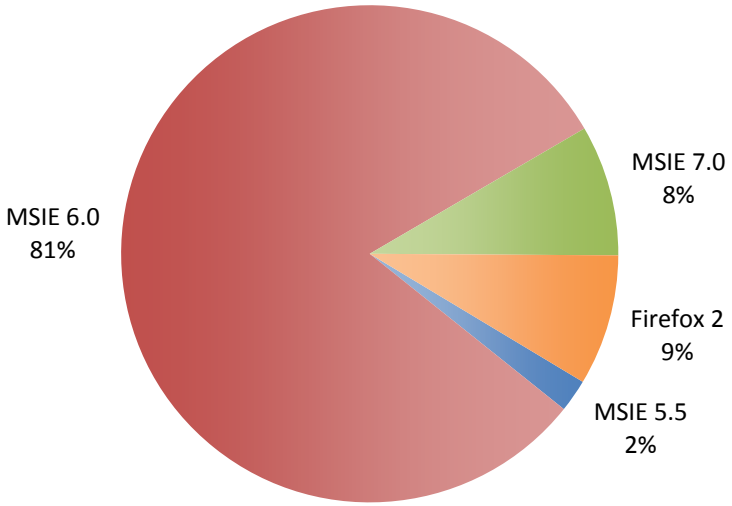


Figure 3.2 – Browser statistics

As seen in Figure 3.2, the makeup of browsers was dominated by Microsoft Internet Explorer 6 (MSIE 6.0) and followed by an equal split between Internet Explorer 7 (MSIE 7.0) and Firefox 2. The individual using Windows NT was also using Internet Explorer 5.5. The significance of browser statistics can dictate what technologies are used in the development of web sites and online applications. Not all browsers are equal, and each support different technologies and standards. Despite MSIE 7.0 being a superior version to MSIE 6.0, the business world is often conservative in its approach to information technology spending [113]. This also appears true within asset management organisations.

Time of Response

Adjusted for the differences in time zones, the majority of responses were returned within two days of the initial email. Figure 3.3 shows that 57% of respondents completed the questionnaire within two days, 77% completed it within one week, and 90% had completed it within two weeks. These results indicate that future online surveys can receive the majority of results within two weeks – a considerable speed improvement over the traditional paper form.

The seemingly spurious outliers towards the tail end of the distribution show that some responses were returned after three weeks. The questionnaire was distributed in

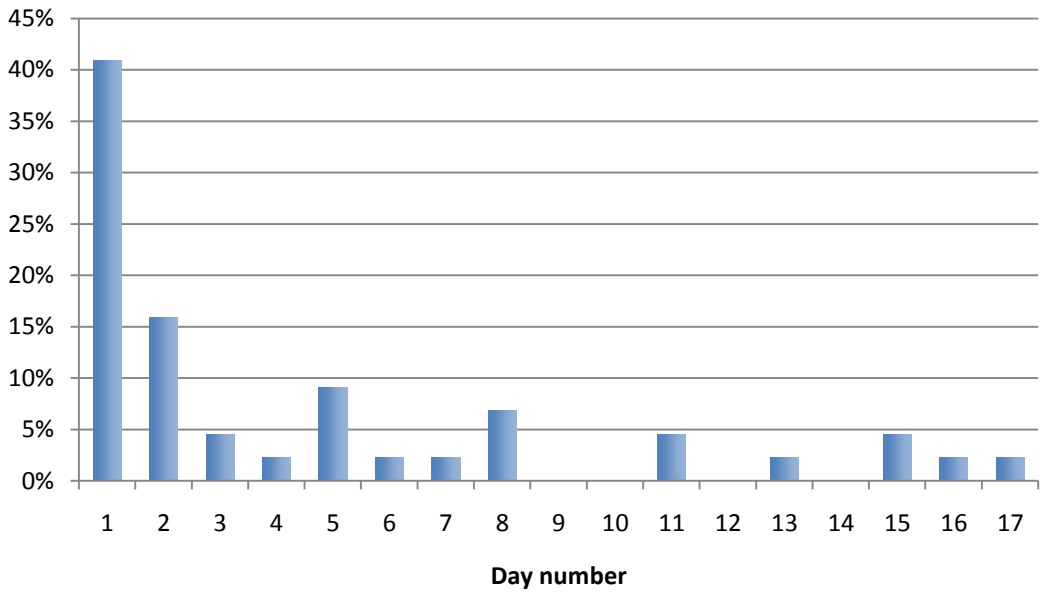


Figure 3.3 – Timeframe of responses

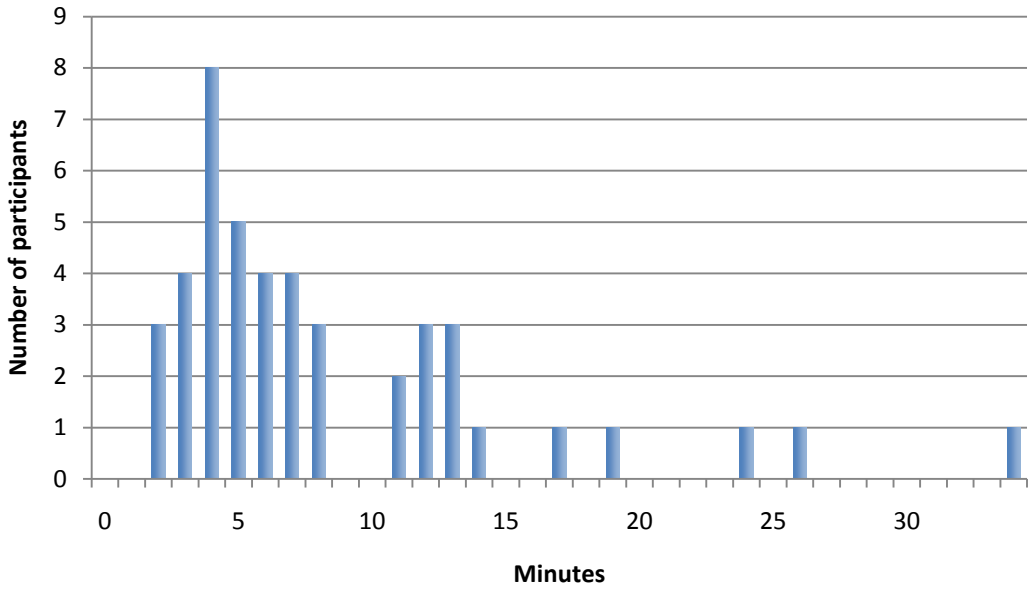


Figure 3.4 – Duration of questionnaire

November and there were no holiday events during this period. As the questionnaire completion time was suggested to take less than ten minutes, it would be expected that the majority of people would have undertaken the survey as soon as they read the email request. There were two instances where individuals replied to the initial email and indicated that they would complete the questionnaire at a later period of time as they were on business travel at the time of receipt.

Duration of Questionnaire

The questionnaire completion time suggestion of less than ten minutes was based on the pilot study. Excluding the two outliers of 3.7 hours and 2 hours, it took participants 8.49 minutes to complete the questionnaire on average (σ = 6.78 minutes). As response rates can drop if the actual duration starts to exceed the expected duration [104], the stated response was reasonably accurate with 68.8% of respondents completing the questionnaire in less than ten minutes (Figure 3.4). The four respondents beyond the 20 minute mark can be assumed to have spent a non-consecutive amount of time, with these individuals attending to other matters in between undertaking the questionnaire.

Miscellaneous Statistics

- 31.8% of respondents qualified their responses with comments, and only 6.8% responded with more than one comment. No questions or clarifications were asked about the survey.
- 24.4% of respondents used the navigation facilities on the questionnaire to modify their answers. This shows that the initial comprehension of certain questions may have been mistaken.

3.4.3 Respondent Analysis

While the metadata analysis was conducted at a granularity of individual respondents, the data analysis in this and subsequent sections are conducted at a granularity of an organisation.

Geographic Distribution of Respondents

The origin of this research can be clearly seen in Figure 3.5 with the majority of organisations based in Australia. There are several participants from Europe, North America, and Asia, while none from Africa or the Middle East. In measuring the willingness to participate from these regions, a comparison between invited respondents and actual respondents sees a decreased response from North America, Europe, and Asia (total composition of invited respondents is Australia 65%, North America 17%, Europe 8%, Asia 9%, and South America 1%).

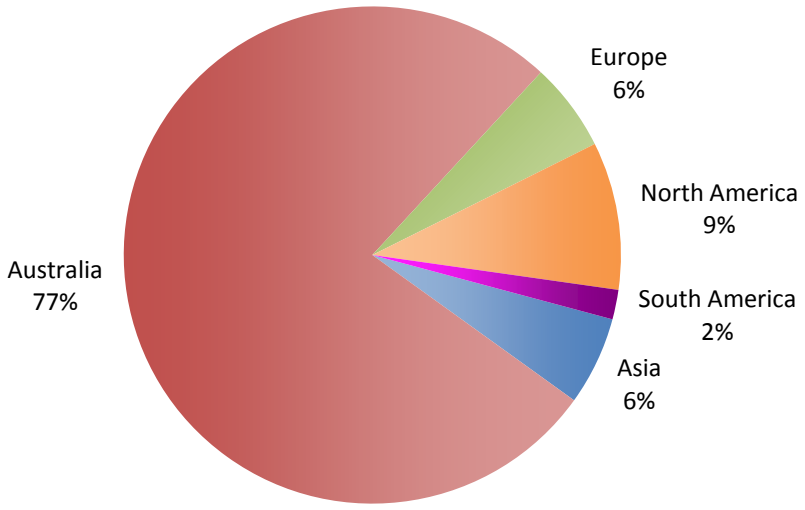


Figure 3.5 – Location of respondents

The implications of these results limit the generalisation of the research findings to asset management data management within Australian organisations, while using other organisations as a consistency check. From the analysis of literature in Section 2.3.2, it would be expected that North American and European organisations would have similar data management attributes as Australian organisations.

Distribution of Respondents by Industry

The distribution of industries participating in the survey is shown in Figure 3.6. Notable representations are from the water utility, power utility, transportation and infrastructure, mining, and manufacturing industries.

There is an inordinate number of utility providers to transportation or mining organisations compared to their relative market sizes (both are roughly 68 times larger than utilities in terms of number in Australia [101]). Two reasons can account for the disparate industry distribution: (1) discrepancies in the sample distribution and (2) the respondent's willingness to participate in the questionnaire. As described in Section 3.2.2, the sample frame distribution was dictated by the association of the participant with the two research centres. The distribution of these organisations clearly does not represent the market distribution, and is instead based on factors such as the organisation's ability to capitalise on potential research benefits, their marketing

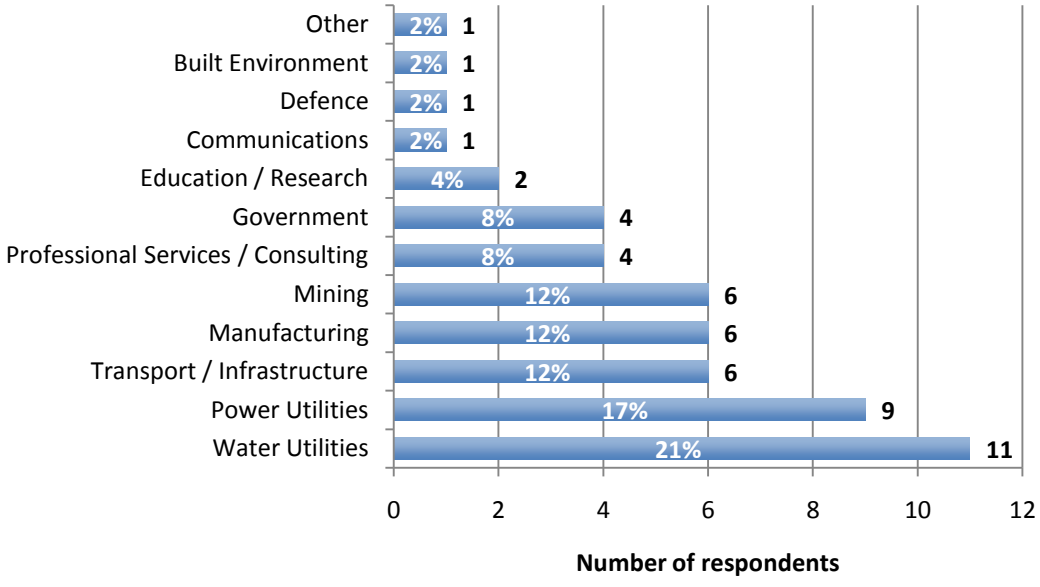


Figure 3.6 – Distribution of respondents by industry

strategies, charitable disposition, and the research centre's ability to attract interest and funding.

The second reason, willingness to participate, is indicated in Figure 3.7 where the percentage of respondents to the total number of survey requests is displayed per industry sector. The communications sector appeared to produce an anomalous result of 100% participation (as there was only one request sent) and has been taken out of the graphical results as not to skew the visual analysis. The other is the manufacturing industry, where there are an extremely low percentage of participants. Both sample frames contained approximately the same number of manufacturing organisations, so the potential issues of a reduced response rate due to the paper media was not applicable. One difference between the manufacturing industry and others is that it produces a physical end product, while the others offer service-based products. While follow up research showed that there were similar elements between the two classes of industries, the difference combined with the background of this research originating from a service-based industry, may have led to less interest from the manufacturing sector.

While the distribution does not mimic the real market distribution, this does not detract from the significance of results. Insight can still be gained into the topic areas of

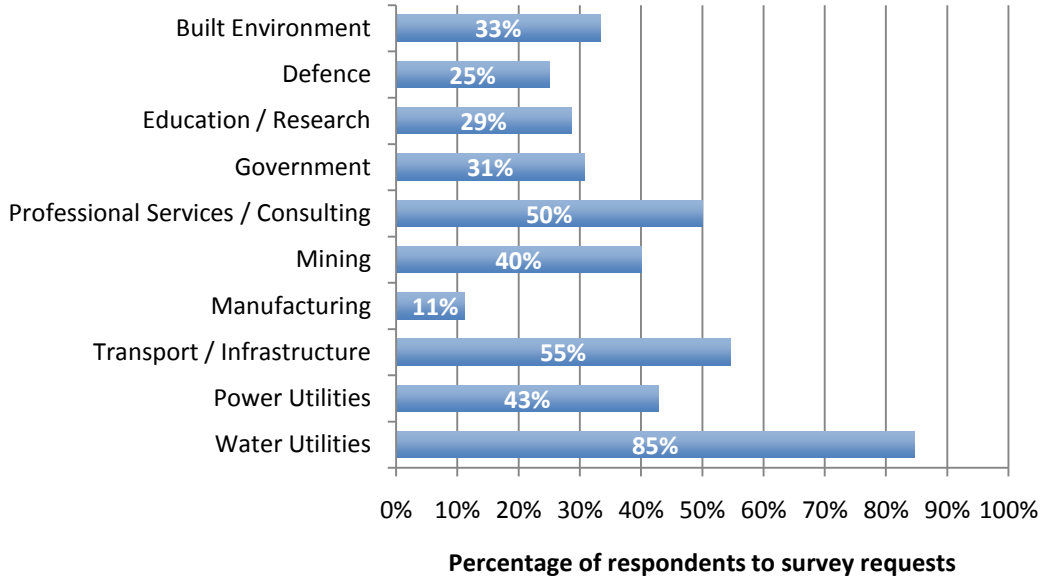


Figure 3.7 – Willingness to participate by industry

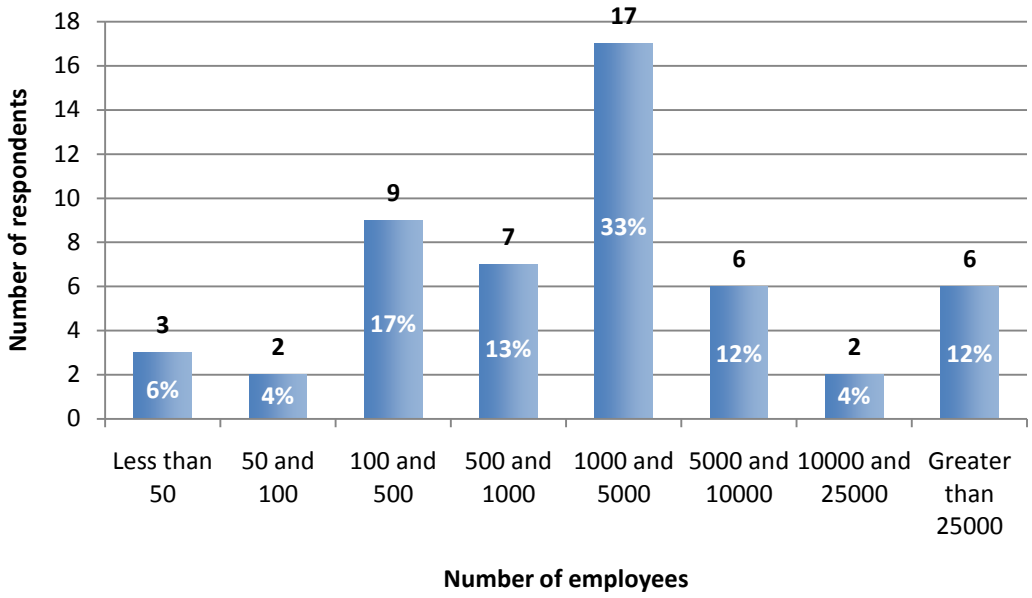


Figure 3.8 – Size of respondents by employment

the survey, although further follow-up work would be required to be able to make definitive declarations.

Size of Respondents

The size of responding organisations was measured in two ways: by employment and by revenue. The size of an organisation was important to gauge the inclusion of both small and large organisations, which could have a possible effect on information system maturity.

There is an adequate representation of organisation sizes distributed over the various categories in Figure 3.8. There are two significant points to note:

- Thirty-three percent of organisations fell within the 1000 to 5000 employee category while a very small number of organisations fell within the 10000 to 25000 category. The categories were based on a logarithmic style distribution, with the lower end dictated by the governmental descriptions of small and medium sized enterprises [114].
- The 100 to 500 category contained 45% of the water utility respondents, and these largely consisted of regional located utilities (as opposed to metropolitan).

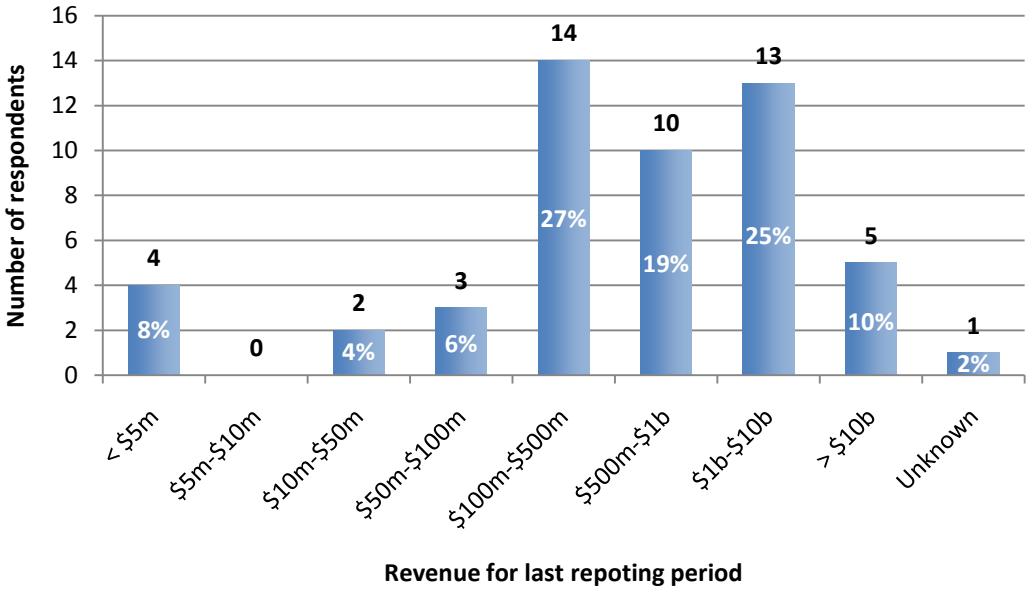


Figure 3.9 – Size of organisations by revenue for last reporting period

As seen with the survey respondents consisting of organisations on both the small and large extremes of employment size, there was also a large spread of stated revenues. The revenue figures in Figure 3.9 were the amount earned over the last annual reporting period adjusted in terms of Australian dollars. As different regions use different fiscal calendars (e.g. Australia use 1st July to 30th June, the US use 1st October to 30th September, and the UK use 6th April to 5th April), the last annual reporting period was used as a common baseline.

As discussed with the number of employees, there were a greater number of larger sized firms in the distribution, and this is also evidenced by a greater number of firms with larger revenues. Eighty-three percent of firms had revenues over $100 million, with a few mega-corporations in the "greater than $10 billion" category having revenues that were in excess of $50 billion and even $100 billion. As the net worth of "property, plant, and equipment" in asset management organisations typically runs into the billions, a left skewed distribution was expected for organisation turnover.

One respondent indicated that they could not divulge the annual revenue of their organisation as it was considered "corporate classified information". However, this information was posted in an annual report and was publicly available on their corporate web site.

3.4.4 Information System Analysis

As information systems form the source systems to a data warehouse, it is important to understand their composition within an organisation. However, due to the large number of system permutations (e.g. a company may have more than one system of a system type) between different organisations, it becomes very difficult in framing appropriate questions for an unattended questionnaire. Hence instead of querying the composition of systems themselves, the systems that managed specific asset management data areas were examined. The slightly different focus involved investigating the major data areas within asset management, whereby seven categories were identified (shown in Figure 3.10). Thus some systems such as GIS do not appear to be explicitly represented in the question, but are implicitly included in the equipment management data area. While there are several other data categories that could have been included, these seven constituted the common major areas.

The results show that most organisations have equipment and work management systems, while the adoption of risk and reliability management systems lags behind. The ordering of the two leaders is slightly strange as most work management systems also include an equipment management component or module. This is also true with materials management, where it is typically a component within the work management

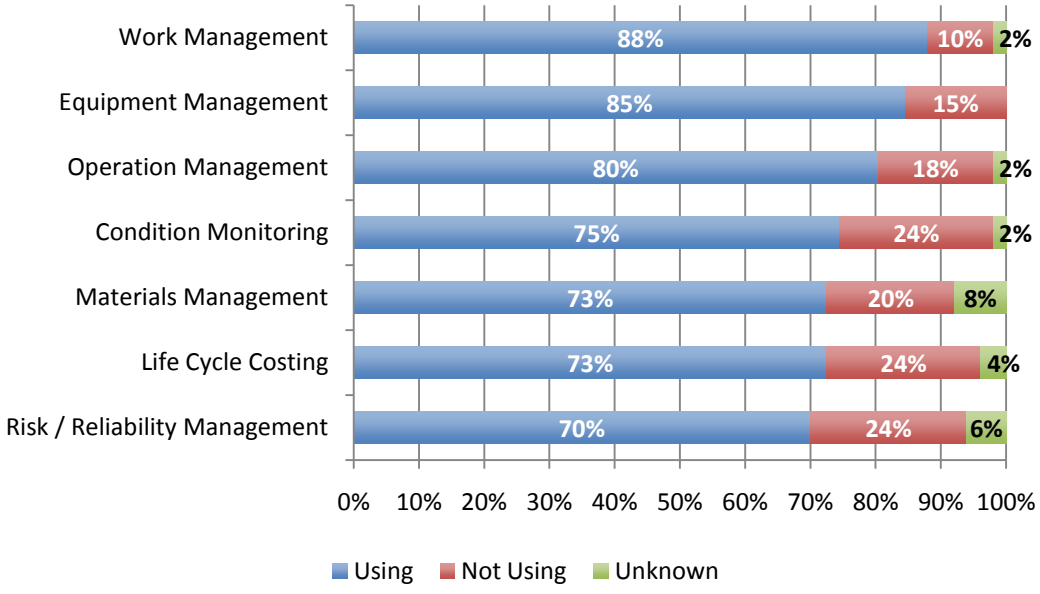


Figure 3.10 – Composition of asset management information systems

system. One possibility could be that organisations are not purchasing these modules in their system packages, or that these modules are not being used to their full capacity and are hence discounted by the respondent. Interviews with several organisations suggest the latter.

The uptake of risk and reliability management procedures into organisational asset management has been slow. A survey in 2001 indicated that only 57% of respondents were including risk management into their asset management to increase safety and reduce maintenance outages [115]. Thus there is a greater than 13% increase, albeit over a six year period.

Unsurprisingly, the top five responding industry sectors (water utilities, power utilities, transportation and infrastructure, manufacturing, and mining) all employed work management systems. Life cycle costing systems within the power utility industry had a usage rate below 40%, while risk management systems within the power utility and mining industries only had a 50% uptake.

Justification of System Selection

To understand why the above information systems were used or not used in organisations, the benefits for using the systems as well as the reasons for not using the system were analysed. Matrix questions were used in these two questions and only one category could be selected to indicate the primary benefit or reason for each area. The selection restriction was enforced to elicit a greater distinction in the results (a multiple selection response would smooth out fluctuations in the responses).

The heat map in Figure 3.11 shows that the “improving business procedures” category was the decisive benefit of asset management systems, nominated with the highest vote for five out of seven areas and tying for first position for two areas. Information systems are often built around a “best-practice” workflow specified by the developer of the system. Due to the lack of workflow customisability in many information systems, an organisation will often need to adopt the system workflow model, rather than adapting the system to the current organisational workflow. While this forced adaptation produces both beneficial and detrimental effects, in the case of asset management, the benefits seem to be greater.

	Improving data entry & quality	Data processing & reporting	Automating repetitive tasks	Improving business procedures	Satisfying regulatory compliance	Other
Work Management	5%	30%	7%	41%	14%	5%
Equipment Management	23%	25%	5%	20%	23%	5%
Operation Management	5%	10%	18%	53%	10%	5%
Condition Monitoring	8%	37%	5%	37%	8%	5%
Materials Management	14%	25%	17%	33%	11%	0%
Life Cycle Costing	3%	32%	0%	54%	5%	5%
Risk / Reliability Management	6%	29%	3%	29%	26%	6%

Figure 3.11 – Benefits of using information systems

The “data processing and reporting” category was the next most selected category, winning one area, jointly winning two areas, and coming runner up in three areas. The interviews in this research found that many organisations generate thousands of reports per day through their information systems, and this supports the high selection of this benefit category. Data processing and reporting is incidentally the primary focus of data warehousing by providing a platform for integrated data analysis.

The reasons for not using asset management information systems are displayed in Figure 3.12. Due to the majority of organisations using systems in each asset management data area, the number of responses to this question was low. Nevertheless, some trends can be observed.

The main reasoning for not using information systems on average was classified in the “Other” category. However, due to the limited space available on the questionnaire,

	No benefit or business case	Unable to find project sponsor	No suitable vendor or product	Unable to successfully use system	Function provided by another method	Other
Work Management	20%	0%	20%	0%	0%	60%
Equipment Management	25%	13%	25%	0%	13%	25%
Operation Management	22%	0%	11%	0%	33%	33%
Condition Monitoring	8%	8%	0%	8%	33%	42%
Materials Management	30%	0%	20%	0%	10%	40%
Life Cycle Costing	25%	17%	8%	8%	17%	25%
Risk / Reliability Management	25%	25%	8%	0%	17%	25%

Figure 3.12 – Reasons for not using information systems

there was no follow up question on what “Other” constituted. The next highest category, “no benefit or business case”, had a mean response rate of 22% (σ = 7%). As the benefits of such systems are well documented, there must be particular circumstances in which these benefits do not apply. As this category is an umbrella category for several different underlying causes, further investigation would be required to investigate this phenomenon. This is also true with the category, "unable to find project sponsor".

Thirty-three percent of respondents to the question indicated that another means was used for managing condition monitoring and operation data. Condition monitoring is often outsourced to dedicated monitoring organisations that specialise in particular monitoring techniques (e.g. vibration, thermography, and oil analysis). The data are stored by the monitoring organisation whilst a report is sent to the asset owner.

Desired System Improvements

Software is released in iterations in order to provide a constant revenue stream for the developer, and in return serving the needs of the purchasing organisation. These needs are constantly shifting due to changes in factors such as technology, competition, regulations, and culture. Eight categories of improvements to asset management information systems were investigated, and are listed in Figure 3.13. As the question looked at general improvements to systems, asset management system specific questions were not included. A ranking style answer mechanism was trialled and considered too difficult for respondents, and consequently, the question asked for the top three prioritised improvements.

The most significant desired improvement was "easier integration with other systems" with 24% of the response. This is definitely a trend seen within the asset management literature reviewed in Section 2.3.2 due to a growing information system maturity across the industry. Integration allows for greater automation of workflows, faster and more complex report generation using data from multiple systems, and an elimination of redundant and inconsistent data. As integration produces numerous additional benefits for systems (as opposed to the other improvements in the list – for example, more responsive systems and easier to customise – that just rehash existing functionality), it produces an attractive option.

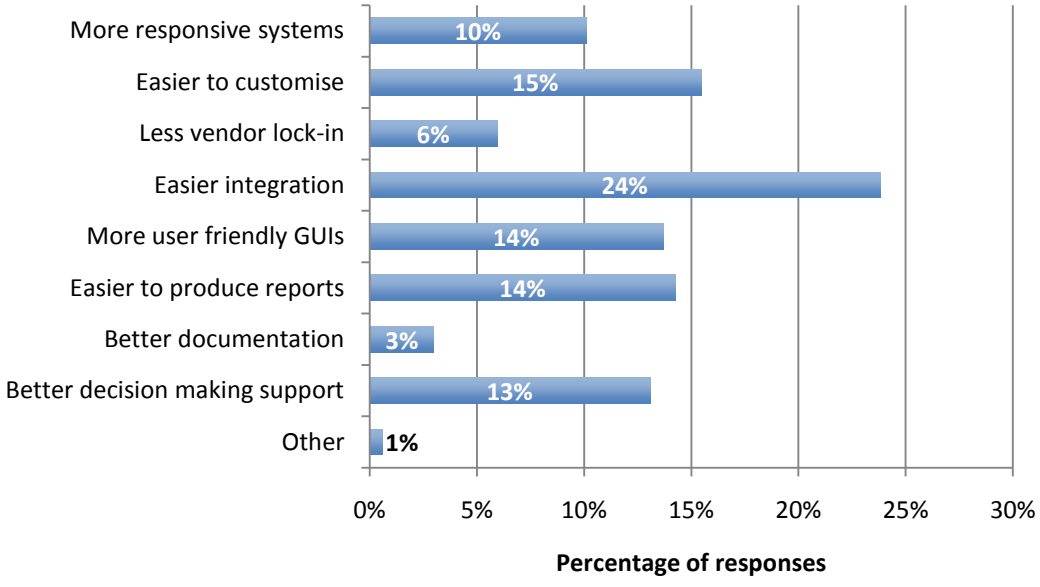


Figure 3.13 – Desired improvements for asset management information systems

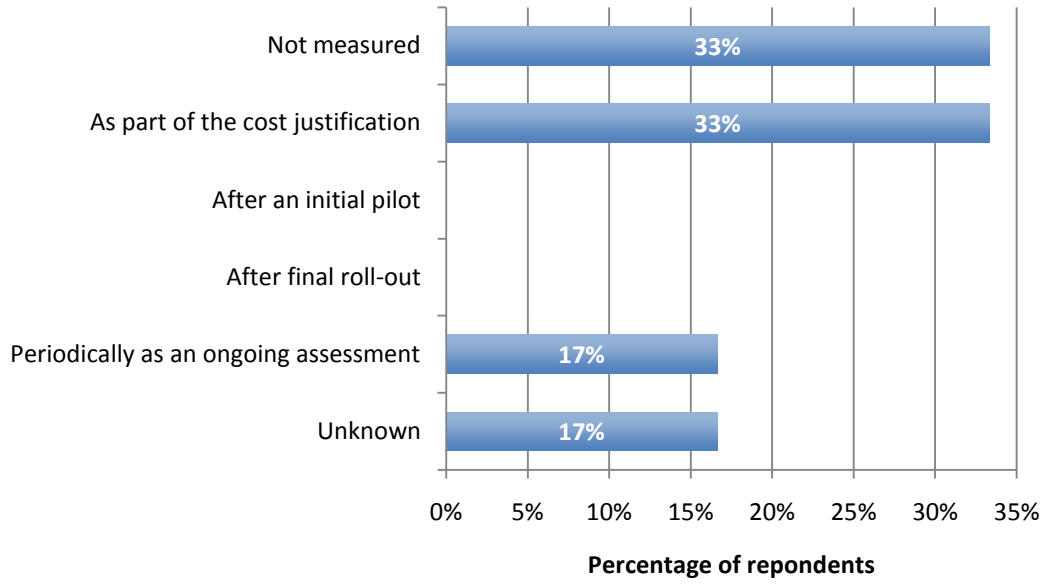


Figure 3.14 – Measurement of return on investment

Equal preference was spread across the areas of customisability, graphical user interfaces (GUIs), reporting, and decision making support. Little preference was given to transitioning between vendors, which would imply that organisations can already easily move their data between different vendors, that organisations do not object to being locked into a particular vendor's solution (benefits outweigh the costs or the organisation has not fully comprehended the situation), and/or that it was not a high priority immediate item that needed to be addressed. As integration would also result in alleviating some issues in moving to different vendors, this correlation may have been recognised by the respondents.

Measuring Return on Investment

Participants were asked when the return on investment of asset management information systems was measured. The return on investment (ROI) is the ratio of money gained or lost relative to the amount invested. When calculated before the start of a project, it can be used to compare various investments to see their relative return for investment selection. When calculated after a project, it can be used to determine the success of an investment. The categories were based the ROI analysis categories presented by Palmer [116].

While at least 50% of responding firms measured the ROI of asset management information system projects, as seen in Figure 3.14, one-third of the companies surveyed did not. In an ROI study conducted by CIO Insight [117], 71% of organisations rated financial justification as important for information technology projects, while only 62% attempted to measure the value of information technology projects through metrics such as ROI. This lower latter figure seems to be also applicable to organisations involved in asset management operations. While it could be a possibility that another measure of profitability or success is used by the firms that do not conduct ROI analysis, it is less likely that it is a classical financial measure (e.g. return on assets, return on capital).

Most organisations that calculated the ROI, estimated it before the outset of their information system projects, while just 17% conducted periodic assessments of the ROI. This is below the market average, where 52% of companies that do measure ROI periodically for information systems [117] is compared to 35% for asset management projects.

3.4.5 Data Warehousing Analysis

With the emphasis on desired improvements in information systems focusing on integration and reporting, it is unsurprising to see in Figure 3.15 that 42% of

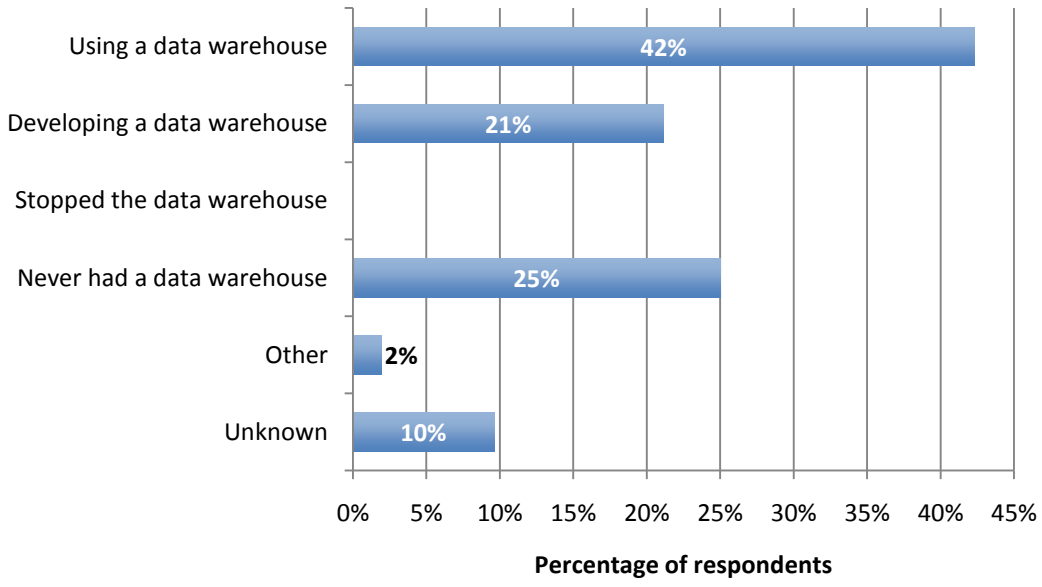


Figure 3.15 – Level of asset management data warehousing

respondents have a data warehouse that is used in their asset management. A further 21% of organisations are developing data warehouses, while 25% are not using a data warehouse. While no organisation had stopped using a data warehouse, the interviews in this research came across instances of failed data warehouse projects where the systems are no longer used as they became too unwieldy to manage.

A study in 2002 indicated that 24% of utilities in North America had implemented a data warehouse with a 13% adoption rate for international utilities [49]. When limiting the scope to utilities, the results from this questionnaire show that 30% of utilities now have implemented data warehousing. As 95% of utilities were based in Australia, there has been just over a doubling in the asset management data warehousing adoption rate over the last five years within the utilities sector.

An interesting observation was seen in a comment on the question - "we use an EAM across functions that means we do not need [a data warehouse] for asset information". The comment shows that many organisations still do not fully understand that data warehouses and transactional systems are not competitors, but are in fact, synergists. While a transactional system can be made to masquerade as a data warehouse, it will be inherently limited by transactional system characteristics.

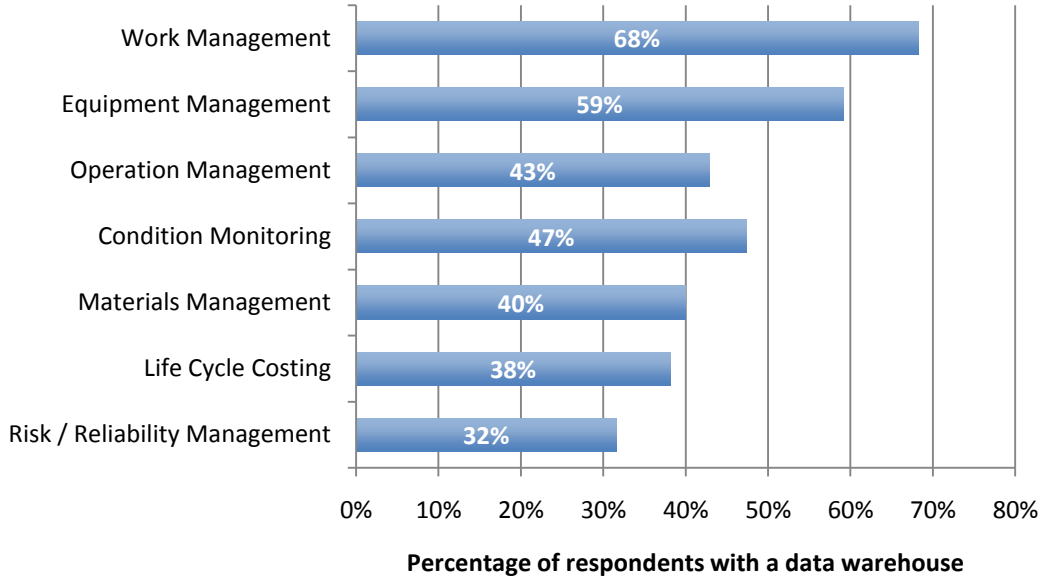


Figure 3.16 – The types of data loaded into the data warehouse

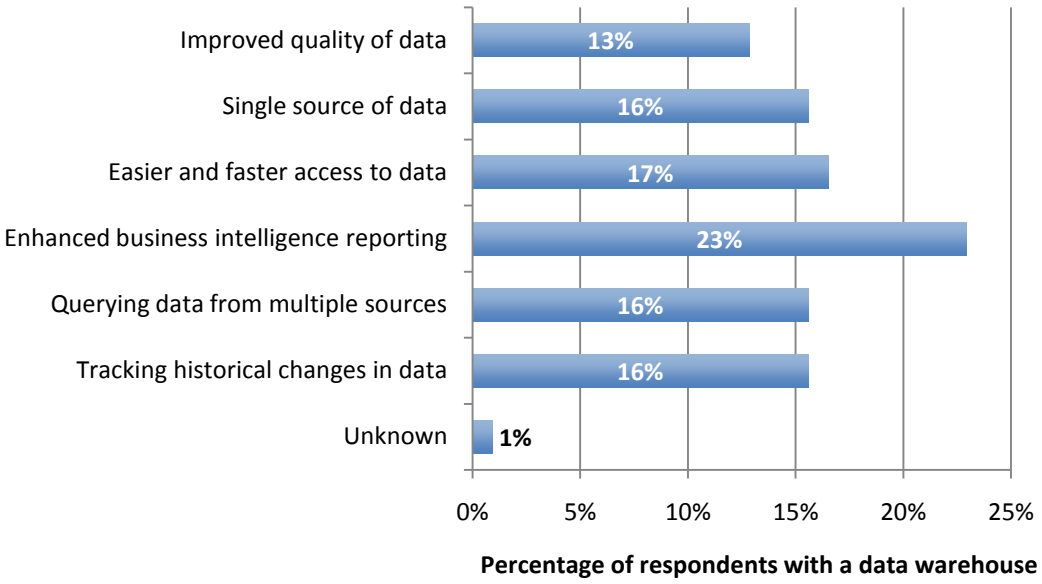


Figure 3.17 – Justifications for data warehousing

As a data warehouse is a generic tool that can be used for any data set, the respondents that were using or developing a data warehouse were asked about the types of asset management data that were integrated in their data warehouse. Figure 3.16 shows a distribution of answers similar to the use of information systems for the seven data areas. Not all asset management data are included within the data warehouse, although work management leads while risk and reliability management trails. There is also a greater spread ($\sigma_{\text{data warehouse}} = 12.5\%$ while $\sigma_{\text{information systems}} = 7.1\%$) which indicates an inability or rejection of a justification in including particular areas into the data warehouse.

Figure 3.17 shows that the primary justification for most organisations in developing a data warehouse was "enhanced business intelligence reporting". As indicated in the desired improvements, reporting was a key area of interest to respondents. From these two premises, it can be concluded that reporting is a critical function within organisations and it will continue to be in the future.

Amongst the other categories, there was little difference in the number of responses despite the inclusion of integration – ("single source of data") and integrated reporting ("querying data from multiple sources") categories. As it would be expected that the latter category would be the primary justification based on the responses to previous

questions, the more technical description of this category may have been less understood compared to the more generic and glamorous sounding "business intelligence reporting" category.

The reasons for not having a data warehouse were asked to those organisations without one. As seen in Figure 3.18, almost one-third of respondents cited a lack of supporting infrastructure, with all of these respondents originating from the utilities industry. Hardware infrastructure is typically less of a problem for data warehousing, as hardware can easily be purchased and made to integrate with existing systems more readily. In the case of utilities, bandwidth can pose a problem for the ETL stage when there are many data sources located in regional areas as there is typically less telecommunications investment in these areas [118]. Software infrastructure is also often a problem, as there are issues with data formats (e.g. proprietary binary data formats), schema formats (e.g. normalisation issues with free text fields), and granularity (e.g. differing measurement granularities between regional sites).

The next notable reason against a data warehouse was the lack of benefits perceived. In support of this, a comment was made that "although not a true data warehouse we do have links which allow reporting across systems" and that "data warehousing is only a hack to cover poorly integrated data". Data warehousing is a situational solution, and not all organisation are able to realise its benefits. Integration is but one area that data

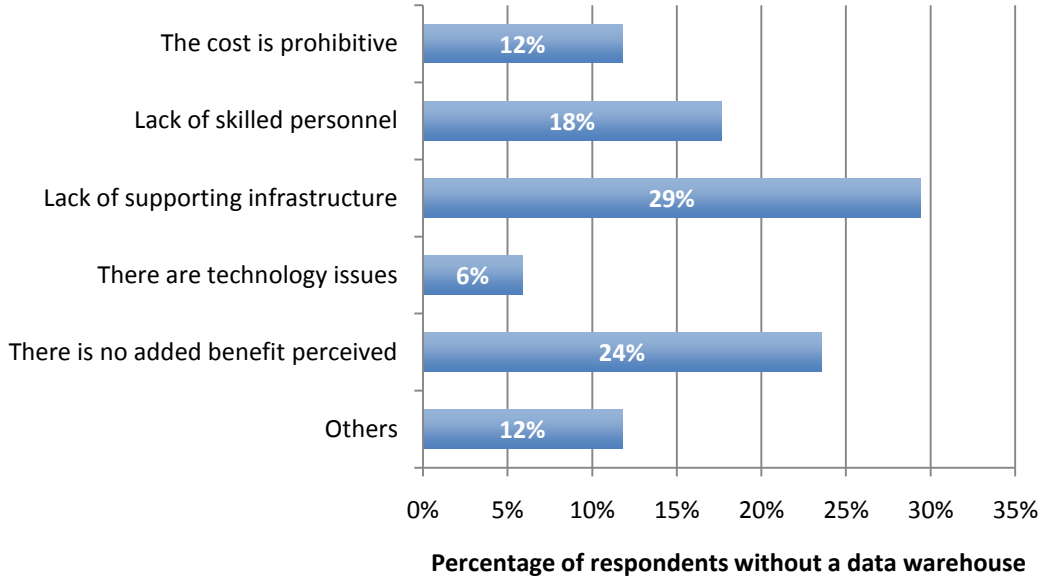


Figure 3.18 – Reasons against a data warehouse

warehousing offers a solution, and there are numerous other advantages as seen in the list of justifications in Figure 3.17.

3.4.6 System Integration Analysis

An important area for data warehousing is that of system integration. Integration involves making an information system accessible to another via an automated process. While data warehousing is an enabler of integration, it does not necessitate system integration, as it provides an indirect rather than a direct method. In addition, a direct method of integration does provide an easier upgrade path to data warehousing than those organisations without it, as it provides an existing infrastructure platform.

The majority of respondents had some type of information system integration within their organisation as seen in Figure 3.19. As the "some systems are integrated" category includes various levels of integration (e.g. automated file copies of data dumps, and sophisticated extraction and insertion through SQL), the category was a catch all for many sub-categories.

A more interesting observation was the "no systems are integrated" category. The respondents were generally smaller organisations that had a smaller number of areas covered by information systems. However, there was one anomalous case of a multi-

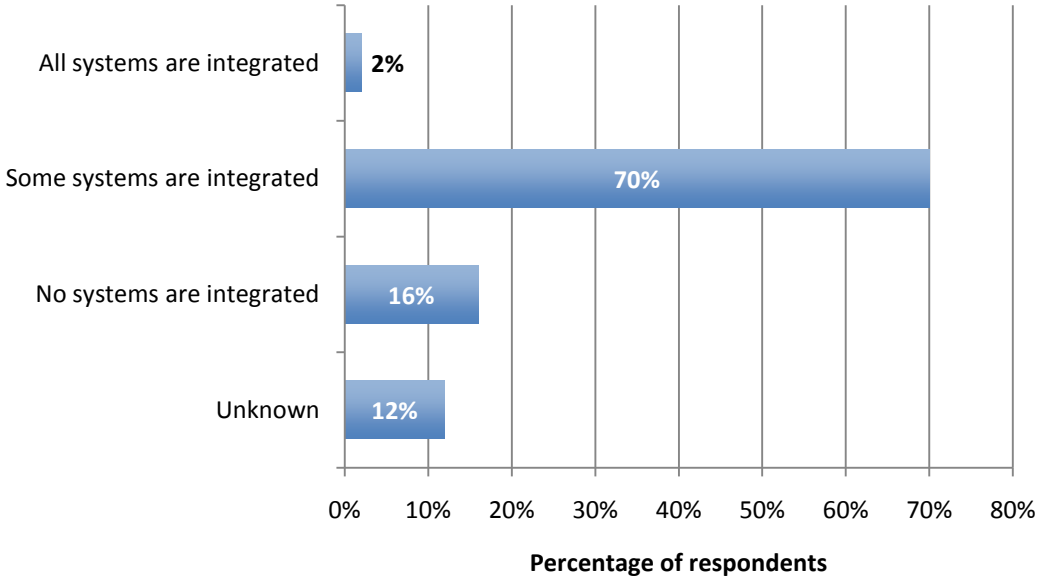


Figure 3.19 – Asset management information system integration

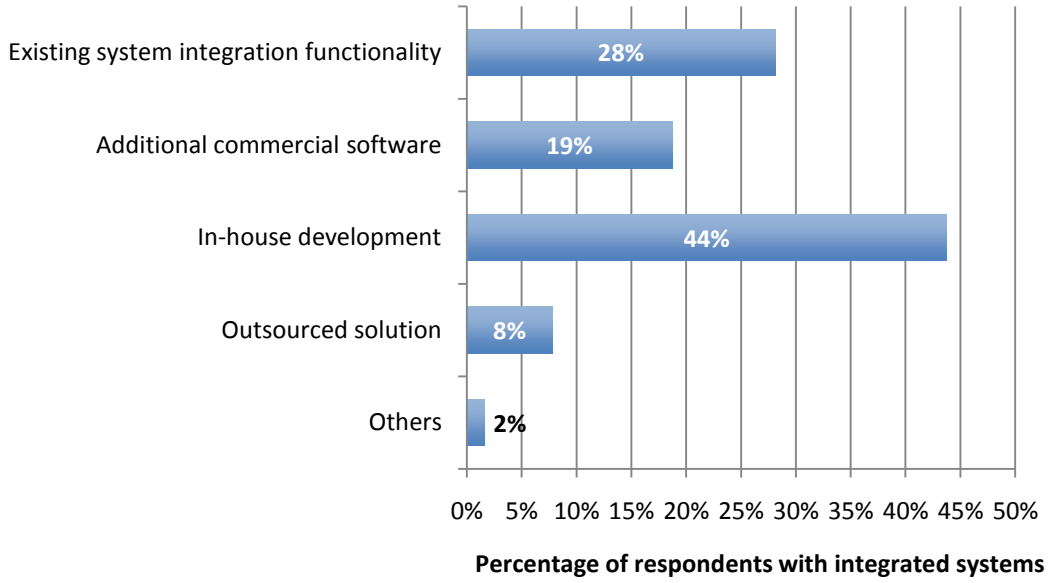


Figure 3.20 – Method of integration

billion dollar organisation that spent very little on asset management information systems, and had a data warehouse but not integrated systems. Presumably, the data warehouse formed the basis for integrated data analysis rather than an individual information system (i.e. ERP or EAM), and automation of other systems was not required at this stage.

Integration can be developed through different mechanisms and Figure 3.20 shows the four most common methods. The in-house development of an integration bridge that spans different systems was adopted by 44% of the respondents with integrated systems. A custom solution is typically the most flexible methodology, but requires a commensurate amount of effort to achieve. Integration functionality already built into existing systems was the next highest method with a 28% response. Consolidation within the asset management information systems market has led to a few key vendors, and these vendors often make data interfaces available whereby other systems can push or pull data through their system. Smaller vendors or those in horizontal markets will develop out-of-the-box integration modules that enable their own software to "talk" to the software of the larger vendors. There are also several providers that focus on developing the integration aspect of systems, and 19% of respondents used this kind of software. These techniques are not mutually exclusive, and 48.6% of the total

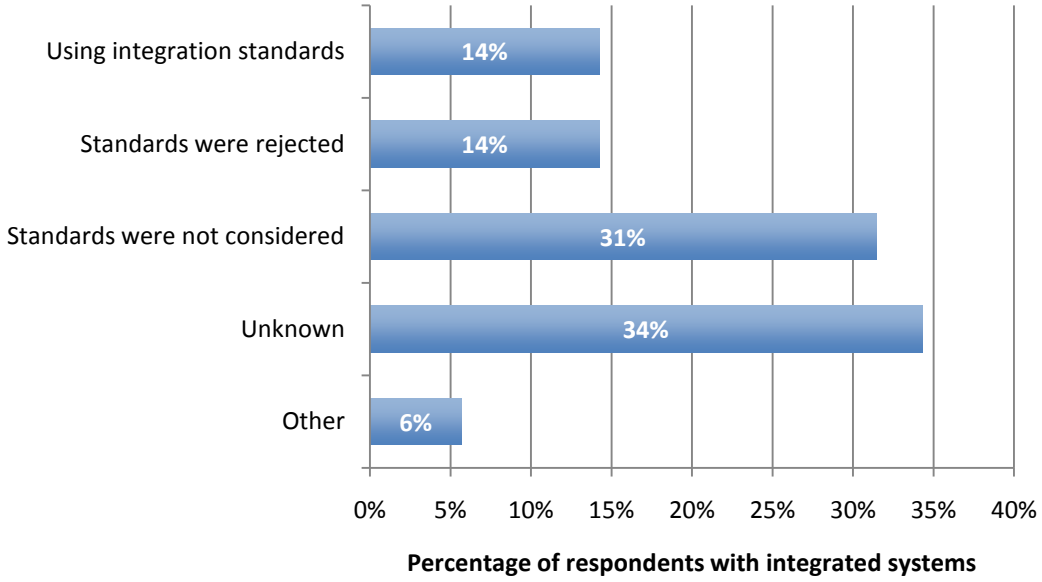


Figure 3.21 – Use of integration standards

respondents with integrated systems used more than one technique to achieve their integration.

While some vendors do make public interfaces to their systems, these interfaces often are designed specifically for the originating system. As discussed in Section 2.3.1, standards have arisen that attempt to formalise these interfaces with common terminology and definitions. These standards can be used by vendors of the systems, their purchasers, or third parties that provide integration services. Standards also have implications within asset management data warehousing as it can provide a platform for quicker ETL development.

From the results in Figure 3.21, 34% of respondents were either not aware of integration standards being used in their organisation while 31% of firms had not considered standards important to their integration efforts. This is a large percentage of organisations that are most likely unaware of the benefits of standardised interfaces (see Section 2.3.1). For the organisations that had investigated standards, there was an equal split with 14% eventually adopting some type of standards and rejecting their use.

One organisation indicated that they used organisational-based standards. While internally developed standards will provide short term savings in IT expenditure, the

case may be different for the long term if new systems are purchased, or if systems need to be connected to sources external to an organisation.

Another used SAP application interfaces as its standard as SAP formed their core information system. With SAP being a dominant player in the ERP space, they can almost dictate which trends, technologies, formats, and workflows certain industries must adopt. Due to the influence of larger EAM vendors, many smaller vendors will ultimately provide integration support for their software using such application interfaces.

One comment on the question indicated that "vendors deliberately use non-standard structures to lock clients". This was a common sentiment expressed in the industry interviews and is particularly the case for condition and operation systems, which often store data in proprietary binary formats. However, the majority of ERP/EAM/CMMS platforms have now moved to relational DBMSs which has liberated data from their governing system. This subsequently allows organisation to integrate *data* in standard formats, although this is one step down from complete standardised *system* integration.

3.4.7 Data Retention Analysis

In a definition of a data warehouse, Marco [119] specifies four preconditions, the last of which states that "a data warehouse holds historical views of data". Transactional systems can store copious amounts of data; however, large amounts of data can lead to system performance degradation. The first solution to improving performance for most organisations is purging the system of superfluous data. This can pose a problem for data warehousing and data mining initiatives, as they can require a comprehensive set of historical data. As observed through the projects associated with this research, a good set of quality historical organisational data spanning years or decades is often difficult or impossible to obtain.

For each of the seven asset management data areas, participants were asked about their data retention policies. Figure 3.22 shows that 27% of respondents discarded their operation data while 24% discarded their condition monitoring data. The primary data type in these two areas is time-series measurement data. Depending on the monitoring policy, the data acquisition process can be data-intensive, producing huge volumes of data for asset operation or condition measurements. The huge amounts of

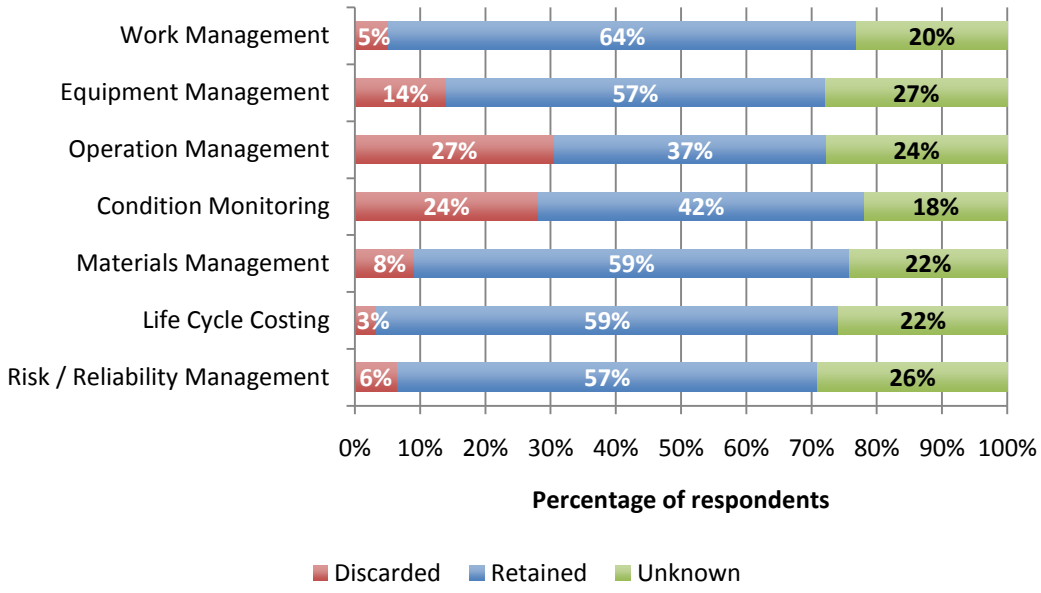


Figure 3.22 – Data retention policy per data area

data from these areas have led to the creation of data historian systems that provide storage, compression, and basic analysis functions for process data.

It is also significant to note that a large percentage of the respondents were uncertain about their organisational data retention policies. One-quarter of the total respondents answered with an "Unknown" response for all of their selected data areas. There is a lack of data lifecycle awareness within organisations with many individuals using corporate systems without knowing the procedures in data collection, maintenance, and purging. There are many assumptions introduced into data from these processes and it becomes important to understand these assumptions when conducting analysis.

The organisations that discarded any type of data were asked for their justification for the policy. Figure 3.23 shows that 31% of respondents found that the usefulness of the data had declined while 12% were uncertain of the usefulness of the data. The first category deals with the *immediate* usefulness or purpose of the data, while the second category covers the *future* usefulness of the data. The benefits of data to organisations may originate years or decades in the future, when the data may be identified to have some use. However, a trade-off must be made by data managers between the current spending for data maintenance and any future benefits in revenue or cost savings.

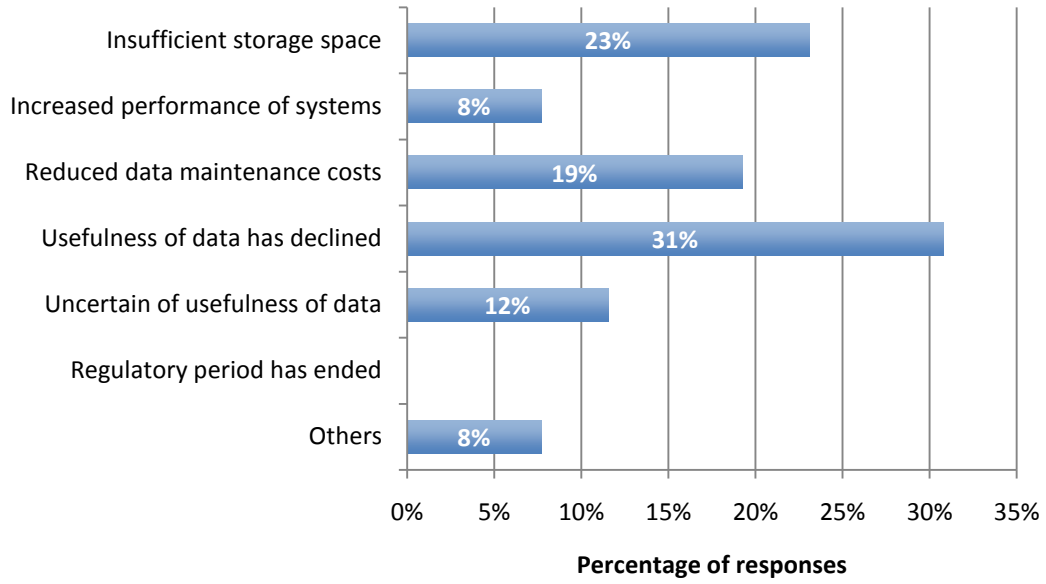


Figure 3.23 – Reasons for discarding data

Despite the increases in storage technology, 23% of respondents cited an insufficient storage space for data. Usefulness, rather than technology or costs, is typically the underlying factor for the selection of this response. Organisations will not spend funds on increasing storage space if the benefit or usefulness of the data is less than the cost to store and analyse it.

The "Other" category provided a comment was that data was discarded due to "plant retirement". If any remaining assets from the plant were not distributed to other plants, or if no historical analysis involving the retired plant would be conducted, then the retirement event would be considered as suitable justification for discarding data.

One respondent made an amusing comment that their systems had "stupid data structures that overwrite history". This research has come across several older process systems that overwrite data on a cyclic FIFO (first in, first out) basis, but will keep aggregated values of the overwritten periods. Designed without the foresight of archiving data, this functionality was included to reduce the memory footprint of the application to maintain performance. While modern systems usually have more options for retaining data, legacy systems can present this problem.

3.4.8 Data Management Analysis

Each participant was asked to give a rating of their organisational data management practices. A six-point Likert scale was used, with values ranging from "Very Poor" to "Very Good". Due to the concerns that an odd-numbered Likert scale can produce a mid-point bias [120], an even number of categories was used. As the categories were not objectively defined, the social definitions of the terms need to be taken into account in the data interpretation.

Figure 3.24 shows that 54% of respondents rated their data management practices as average and 14% as good, below average, or poor. When looking at the top five largest sector groups, t-tests showed that the transportation and infrastructure industry ranked themselves significantly higher than other industries ($p = 0.028$) while the power industry ranked themselves significantly lower than other industries ($p = 0.53$).

3.4.9 Internal Consistency Analysis

A calculation of Cronbach's alpha was attempted to measure the internal consistency for the whole of data set as well as the sub-themes in information systems, data warehousing, integration, and data retention. Due to the branching nature of the questionnaire and the reliance upon combo boxes, only the questions on information system composition, ROI, data warehousing use, integration use, and the overall rating

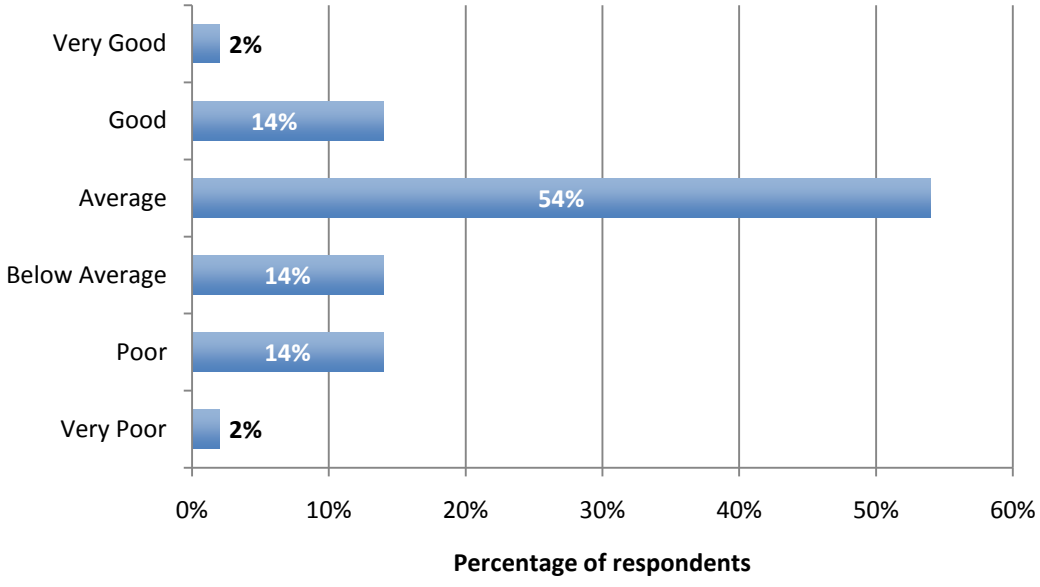


Figure 3.24 – Data management self-rating

could be used. For the overall set of results, the alpha was 0.561 (N=36). For the sub-theme of information systems (composition and ROI), the alpha was 0.743 (N=38). As 0.7 is typically used as the benchmark in social science research [121], the overall figure is a little low. This is due to the smaller number of respondents as compared to their diverse answers, as well as the lower number of applicable results sets due to question design.

3.5 Implications

This exploratory survey was conducted to provide a preliminary examination of the status of information systems and data warehousing within asset management organisations. The responding organisations were from an assorted selection of industries, and were of varied sizes. The survey research questions addressed the composition and use of information systems, as well as issues in data warehousing, integration, and data retention.

The survey found that the majority of organisations use work and equipment management systems while life cycle costing and risk and reliability management systems were lacking in their use. However, these previously esoteric systems are becoming more a part of an organisation's information system architecture as adoption increases for the business processes they support.

One of the more striking revelations of this survey was that the primary justifications for using asset management information systems was to both improve business procedures and data reporting. Streamlining business procedures through workflow automation decreases the overall time and resources required by each procedure, and thus, reducing costs. Data analysis and reporting also provides a method to detect inefficiencies within processes, and provides a platform for continuous improvement. Current information systems are largely based on relational data models, and the organisations using their information systems for data processing and reporting would be interested in Chapter 5 that investigates the benefits of moving to multidimensional structures.

While most organisations have already integrated some of their information systems through in-house development, easier integration was cited as the most desired aspect for next generation systems. However, the lack of knowledge on data integration

standards was evident within these organisations. Standardising on data transmission protocols allows a degree of forward compatibility and decreases the risk of adoption in long-term usage scenarios at the expense of an increased initial cost. With current corporate attitudes focussing on the here and now of profit making, many organisations would take a short-sighted view of a standards-based approach. Chapter 4 discusses asset management data integration at length, and also investigates the significant data model standards in this area.

Asset management data warehousing appears to be firmly established in organisations, with more than half the respondents either developing or operating a warehouse. Naturally, adoption has increased over the past half-decade and the primary reason for adoption is, again, the need for enhanced data analysis and reporting. With the significant uptake of data warehousing for asset management, it appears that this methodology is a step in the right direction, although issues still remain on how to integrate data across different asset management areas. Chapter 6 is pertinent to the two latter groups as it discusses research into case-based reasoning will allow these organisations to quickly develop a data warehousing platform based on other industry cases.

Overall, the major conclusion of this survey is that the use of data management software for asset management in general is yielding favourable results for most users. Technology is being used to automate processes leading to greater efficiencies, and complex data analysis is now becoming mainstream after decades of simple data capture and reporting. There is no clear cut industry or size of organisation leading the charge, nevertheless, all asset management organisations are residually benefiting from the ones that are. These survey results are not meant to provide a definitive description of the field, but are instead a mechanism to provide an exploration into data management in asset management. In the context of this research, the results highlight the significance of the third to fifth research outcomes as justified above.

4 Asset Management Conceptual Data Modelling

One key to a superior system is ensuring it is developed on a solid foundational model. There are a multitude of models that underlie a system, ranging from models that provide the system functionality (data models and activity models) to those that govern its development (cost models and software development models). Data models are not the sole governor of the functionality of a system, but they do have a critical impact, particularly on data driven systems.

The importance of data models with data warehousing systems is even more apparent, as the whole purpose of a data warehouse is to supply information. Data quality issues aside, the data must be stored in the correct format, with the correct relationships, and at the correct granularities. Data warehouses extract data from a variety of systems in order to provide an integrated view of the data through its own model. There are a large number of information system products available for asset management that have been independently developed, and this inevitably leads to disjointed and incompatible models. As there are often multiple paths to achieve the same outcome, the difference in each vendor's knowledge, skills, and resources leads to different products and underlying data models. Despite increasing collaboration between vendors in making products more compatible, much of this cooperation occurs at the application communications level rather than lower data model level. While the former may present a consistent data view from a business or application perspective, it does not solve the issue of internal data model fragmentation. Thus, the challenge in data warehousing is provide an integrated asset management model from several fragmented and potentially incompatible models.

This chapter attempts to provide a comprehensive conceptual data model that presents an integrated view of asset management elements. While a complete model is desirable, a comprehensive model is the best that can be realistically achieved due to the contextual nature of asset management, limited resources, and the evolving nature of asset management and technology. The conceptual model describes the meaning of asset management data and the concepts they represent rather than the purpose of the

data or the physical representation within a database. The conceptual model does not intend to address the semantic issues (reference data, ontologies, etc.) of data integration, but rather, it addresses the syntactic issues (data modelling). This is because the syntactical issues must first be resolved before any semantic issues can be addressed.

By examining a variety of different sources of information on asset management, a holistic approach was utilised in modelling. Ranging from existing data models to analysis methods to business process models, a variety of inputs were used to produce the conceptual data model.

Departing from a pure relational model, object oriented concepts were engaged to provide richer semantics and condensed notation. As the notation language formed the basis for communication, existing notational standards such as entity relationship modelling and the Unified Modelling Language were used.

The overall approach was both verified and validated to ensure the quality of the work. Reviews were conducted by asset management experts and an investigation into the conceptual data model's compatibility with existing models was undertaken for verification. Validation consisted of three case studies on software projects, as well as a review against the CIEAM Asset Management Framework.

4.1 Related Literature

While there are no models that purport to be a conceptual data model for asset management data warehousing, broadening the scope by excluding data warehousing reveals areas of relevant work. These conceptual data models for asset management are found within standards presented by industry bodies and standards organisations, as well as in the data models from which information systems are built.

4.1.1 Standards

There are three relevant standards – ISO 15926, MIMOSA OSA-EAI, and ISA-95 – that are intended for the exchange of asset management data between systems. All originate from different backgrounds and needs, and their influences are apparent in the design of the models. While the primary objective of each standard is in communicating data, the relevance to data modelling stems from the fact that one step in integrating data is identifying the content of the data (which can be expressed as a model).

ISO 15926 is entitled, “Industrial automation systems and integration—Integration of life-cycle data for process plants including oil and gas production facilities”. Its roots are planted in ISO 10303, more commonly known as STEP (Standard for the Exchange of Product model data) whose focus was on providing a taxonomy of equipment related data. Although initially beginning with the process industry, the coverage of ISO 15926 has increased as it has become more generic and less specific to a particular industry. It is now being proposed as an upper level ontology by its primary maintainers [122], and has subsequently received criticisms levelled against its ontological applicability [123]. Ontologies and data models both deal with conceptualisation, however, ontologies are typically generic and task-independent while data models are task-specific and implementation-oriented [124]. ISO 15926 is more the former, with its entities more focused on modelling the universe rather than specifically focusing on asset management. Abstract and generic models require longer periods for implementation, as a greater number of design choices are required. Coupled with the huge nature of the model and lack of supporting technology [125] no complete implementations of ISO 15926 exist at present.

MIMOSA OSA-EAI is entitled, “Open Systems Architecture for Enterprise Application Integration”. It is poised as a standard interface for operations and maintenance data, but its vibration data exchange heritage is apparent with its comprehensive support of XML (Extensible Markup Language) Schema for the transmission of raw measurement data. Apart from the area of measurements, the OSA-EAI also supports the areas of equipment, agents (personnel and organisations), work management, events, equipment health and diagnosis, alarms, and reliability.

ISA-95, which is entitled, “Enterprise-Control System Integration”, was born out of the need for a standard interface between enterprise and process control systems. Thus the areas it covers involve equipment, personnel, materials, capabilities, schedules, and performance. While it is clear that the standard does not try to define the entire gamut of asset management unlike the previous two, it provides unique concepts not touched upon by the others with its capability and performance models.

The relationship between the OSA-EAI, ISA-95, and OPC (a process control standard) is shown in Figure 4.1. It shows elements within asset management where each standard is applicable. From the diagram, the only overlap between the OSA-EAI and ISA-95

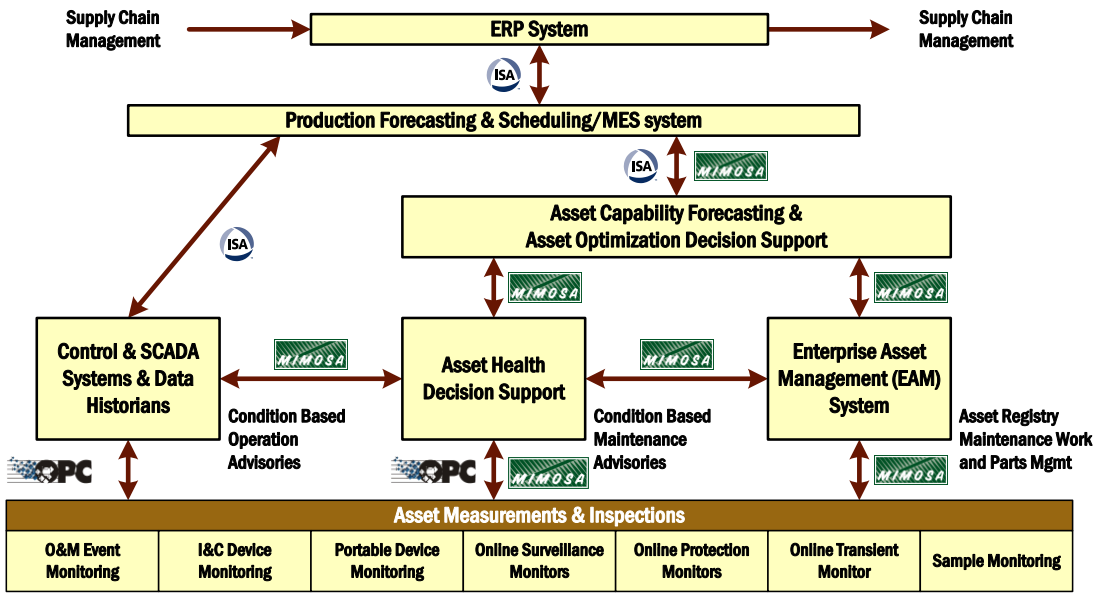


Figure 4.1 – Enterprise systems information network [126]

appears to be within production scheduling, however in reality, the overlap is broader with reiterated definitions of equipment, locations, and personnel.

4.1.2 Information Systems

There are also a multitude of information systems within the EAM (Enterprise Asset Management) category (as well as Enterprise Resource Planning and Computerised Maintenance Management System categories) that are based on asset management data models. As public publishing of these data models is limited in order to maintain a competitive advantage, it is difficult to ascertain the scope of the data models on which the information systems are based. Regardless of their unavailability, their scope is questioned as there are no information systems that support the complete functionality of an asset management conceptual data model. While it is a false inference to assert that this is because no asset management conceptual data models exist, it does strengthen the probability. An organisation would create a complete asset management data model either because (1) it plans to expand their current solutions into other areas, or (2) it plans to integrate their solution with other products. With (1), even if a company develops a data model, it would suffer from publishing restrictions as previously stated. With (2), as the lack of integration between different classes of information systems is the primary motivation behind the above integration standards, the information systems-based data models must be lacking in their scope.

4.2 Conceptual Data Modelling Methodology

The main considerations of a data modelling methodology revolve around the process as well as the process inputs, both of which are discussed below. The outputs of the methodology as well as their validation are discussed in subsequent sections.

4.2.1 Process

The foundation of most data model theory is derived from the ANSI/SPARC three schema architecture proposed by Tsichritzis and Klug [127]. The architecture separates the conceptual, logical (external or application view), and physical (internal view) levels. The purpose of a conceptual data model is to explore high level domain concepts; the purpose of a logical data model is to define the entities, attributes, and relationships for an enterprise project; and the purpose of a physical data model is to design the schema of a database [128]. Moving from the conceptual model to physical model entails an increase of structured information at each level. Thus a conceptual model is required before either a logical or physical model can be developed.

The two important groups of conceptual data modelling methodologies are the ER model and Object Role Modelling (ORM) [129].

Chen [130] proposed the ER model by presenting a technique that described entities and relationships in a graphical format, using a set of shapes and lines. Using a top-down approach, entities are first identified, followed by the relationships between entities and the attributes of the entity. Despite the ER model giving birth to a number of variations, criticisms were levelled at the model's lack of clarity in definitions. Codd [131] said that "the major problem with the entity-relationship approach is that one person's entity is another person's relationship".

The three most popular variations on the ER model are the Information Engineering model [132], IDEF1X [133], and the Oracle Method (formerly CASE*Method) [134]. The Information Engineering model was an attempt to refine the ER model by discarding the notion of a complex relationship, and terming the relationship itself as an entity. Thus every relationship is binary, as only two entities are involved (or one entity for a reflexive relationship). Coming from the family of ICAM Definition Languages, IDEF1X had widespread usage as it was mandated by the US government for all government projects.

Advocates of Object Role Modelling argue that their methodologies capture a broader range of structural features and constraints compared to the ER-based models [10]. ORM does not distinguish graphically between an entity and an attribute on the premise that the relationship between an attribute and entity is conceptually the same as a relationship between two entities. This allows the representation of subtleties not available in ER modelling. ORM also has a rigorous means of dealing with higher arity relationships when the number of objects amongst a relation is greater than two.

The methodology in this chapter has its roots based in the ER model due to the simplicity and pervasiveness of the technique. Despite the application of object-oriented techniques to data modelling "show[ing] exactly the same concepts as ER models" [135], UML object techniques are used to enhance the data modelling process (see Section 4.3.1). Literate modelling [136] is also used to enhance the ER and UML methodologies by providing a narrative of each model.

It is important to note that object modelling itself was not conducted. The intent of object modelling is to model a software system and how it operates, rather than how data are stored. Instead, only the notation elements of object modelling are used, rather than its philosophical purpose.

4.2.2 Modelling Inputs

A comprehensive data model requires a comprehensive modelling process and a thorough knowledge of the domain. In order to understand the field of asset management and the data requirements, seven points of investigation formed the inputs to the modelling process. These were the examination of data model patterns literature, standards, information systems, business process models, interviews, analysis procedures, and business documents.

Data Model Patterns

There are several definitions attributed to the word 'pattern' [137]. A pattern is a template from which something can be derived. A pattern is also reoccurring characteristics of multiple objects. Patterns literature within computer science revolves around providing exemplars of good practice. As one of the aims of this work is to provide a reference model, the characteristics of patterns are significant to note.

Area/Sub-area	Hay	Fowler	Silverston
People and organisations	✓	✓	✓
Assets and objects	✓	✓	✓
Documents	✓		
Contracts	✓	✓	
Ordering and invoicing			✓
Procedures and activities	✓	✓	
Shipment			✓
Work	✓	✓	✓
Accounting	✓	✓	✓
Measurement	✓	✓	
Units	✓	✓	
Ranges		✓	

Table 4.1 – Data model pattern comparison

There are numerous branches of patterns literature ranging from object models, integration models, data models, to metadata models. While the area is broad, the focus for this work is on data model patterns in which there are three seminal works. These are books by Hay [138], Fowler [139], and Silverston [140]. Both Hay [138] and Fowler [139] provide more conceptual models where the focus is on entities, their relationships, and cardinalities. Silverston [140] concentrates on providing logical models that include attributes, and are one step removed from physical models.

The difference in the approaches to the patterns contributes to the differences in the covered business areas as shown in Table 4.1. The table lists the patterns relevant to asset management and the authors that describe the patterns. Silverston [140] tends to be more implementation driven, and goes into domain specific patterns such as ordering and invoicing, and shipment of goods, while the other two authors abstract the detail with the broader categories of contracts and activities, respectively.

There are several works in enterprise object model patterns [136, 141, 142] that are useful to conceptual data modelling. The qualifier ‘enterprise’ is used to differentiate from creational, structural, and behavioural object patterns such as those by the Gang of Four [143]. Despite the impedance mismatch between object models and data models [144], a data model can be derived as object modelling forms a superset of relational modelling. As the aim is to provide a conceptual data model, the object model

patterns need to be abstracted through techniques such as class categorisation to discover the underlying theory upon which they are based.

In a similar vein to patterns, albeit from a commercial perspective, ADRM (Applied Data Resource Management) provide enterprise, business area, data warehouse, and data mart models for various industries [145]. Starting from a common enterprise model, contextual elements are added for various organisations. Thus portfolios are added for financial service companies, rate plans are added for telecommunications companies, policies are added for insurance companies, and metering is added for utilities. There are some patterns that can be used for asset management, but such support is varied.

Standards

Standards are an agreed upon set of rules that are established by an authority [137]. However, the proliferation of standards-issuing bodies signifies the lack of a universal authority. This is true within asset management, where there are a multitude of

Area	Standards
Specifications	PAS-55
Assets / Activities	MIMOSA OSA-EAI ISO 15926 / ISO 10303 – STEP ISA-95 / B2MML ISO 14224 OASIS PPS KKS
Personnel / Organisations	MIMOSA OSA-EAI ISO 15926 ISA-95 / B2MML
Documentation	OMG ManTIS
Life cycle costing	AS 4536
Measurements / Diagnosis / Prognosis	MIMOSA OSA-CBM MIMOSA OSA-EAI ISO 13374 IEEE 1232 – AI-ESTATE
Units	IEEE SI 10-1997
Reliability	ISO 14224 AIAG FMEA-3 SAE J1739 MIL-STD-882
Risk	AS 4360

Table 4.2 – Asset management related standards

standards from different organisations that span the same areas [146]. As Grace Hopper said, “the wonderful thing about standards is that there are so many of them to choose from” [147].

Despite the plethora of standards, their importance on conceptual data modelling lies within the endorsement by authorities. These are experts that have significant domain knowledge, and who may have conflicting views but have come to a compromise to ratify a standard. Thus the established concepts, models, and nomenclature in standards can be applied in the modelling process.

For the patterns they describe, only Arlow and Neustadt [136] reference and comply with standards, primarily those by the International Standards Organization (ISO). Thus money is phrased in terms of ISO 4217, countries in terms of ISO 3166, and books in terms of ISBNs (ISO 2108). The standards referenced only deal with categorical data, and thus their influence on the patterns is only reflected in the model attribute types. For example, as ISO 4217 describes currencies as a three letter code (e.g. “USD”), the Currency class has an alphabeticCode attribute of String type.

There are hundreds of standards that are related to asset management, and it would be an impossible task to examine each. There are several summaries of existing and future standards and how they fit into particular areas of asset management [146, 148-150], but these summaries are not exhaustive. Even when examining data related asset management standards, not all are applicable in conceptual data modelling. For example, IEEE 1451 provides a standard transducer interface, while and IEC 60870 provides a standard controller interface. Despite their importance to asset management, the mechanism used by systems to transmit data is not important (although the data contained within are important).

Table 4.2 shows a list of asset management standards that were used for conceptual data modelling. It is not a complete list of all applicable standards, but it covers a large majority of asset management data areas that are in the conceptual data model. Standards such as MIMOSA OSA-EAI, ISO 15926 and ISA-95 cover multiple areas and provide either object or data models. General standards that are domain independent are used for less traditional asset management activities such as risk management or life cycle costing.

Category	Information System
Enterprise Resource Planning (ERP)	SAP
Enterprise Asset Management (EAM)	Maximo, Mainpac, Hansen
Documentation	Comos
Geographic Information Systems (GIS)	ArcView
Reliability	Relex
Process and control	CitectSCADA
Condition monitoring	Emonitor Odyssey, Watchdog

Table 4.3 – Asset management information systems

Information Systems

As one of the goals of conceptual data modelling is to provide an integrated model from which a data warehouse can be based, it is important to look at information systems. An information system is "the system of persons, data records and activities that process the data and information in a given organization, including manual processes or automated processes" [151]. The term is often used synonymously with computer-based information systems, although information system purists reject the association. For the purposes of convenience, the term 'information system' will primarily indicate computer-based information systems.

Information systems are based upon a data model, which form constituent areas of a total asset management model. Information systems also supply the data that are extracted, transformed, and loaded into a data warehouse.

As with standards, there is an overwhelming number of existing information systems with the key differentiator being their functionality. The areas covered by each information system are similar for their class. Thus most EAM systems will cover asset registries, financial management, materials management, maintenance work management, etc. while reliability systems will cover FMECA (Failure Mode, Effects, and Criticality Analysis), reliability block diagrams, root cause analysis, etc. The similarity between systems in the same class benefits the modelling process by narrowing the systems to be examined to a sample representing the population.

Examination of information systems can be undertaken in two ways. The first is a direct physical inspection of the front end functionality and/or back end database. The database does not directly need to be analysed as the data model can be derived by

looking at the functions the system supports. Particularly with large and older systems that have a huge amount of tables with unintelligible names, examining the front end is more productive, as asset management areas can be studied in manageable portions. The second means to examine information systems is indirectly via a list of supported functions. This can be either gathered from documentation, advertising materials (such as vendor websites), or word of mouth.

Convenience sampling was used in selecting the information systems to be analysed. As many of the systems require intricate deployment setups that can cost thousands and sometimes millions of dollars, the systems of CIEAM and IMS associated organisations were investigated. The list of information systems examined is presented in Table 4.3. The systems listed encompass a wide range of asset management data, particularly as ERP and EAM systems are expansive in their coverage. SAP, Maximo, ArcView, and Emonitor Odyssey form part of the world market leaders in their respective categories [152-154], while the Citect is the Oceania market leader [155] for process control systems. Although popularity does not indicate technical superiority, it does indicate the significance and relevance of functionality required by organisations.

Business Process Models

In order to understand, redesign, and optimise existing business processes, companies are undertaking business process modelling to capture their business processes. These models show the relationship between activities, data, entities, resources, and goals. The connection between activities forms a flow, and branches can be defined using Boolean logic to indicate where decisions are required. As the main goal of data warehousing is to provide a foundation for decision support, business process modelling has an impact on conceptual data modelling by dictating the types of data required to be supported in the model.

Business process models of asset management processes at SunWater were analysed. However, the models did not completely cover all of the asset management business functions within the organisation. Using the ARIS (Architecture of Integrated Information Systems) methodology of business modelling, the models covered the business value added chain, a high level asset management process, and the processes involved in information acquisition and analysis, risk analysis, strategic and operational planning, scheduling, and maintenance work. The models were described through

Category	Organisation
Industry organisations	Assetricity Hunter Water IFS MPT Solutions Pall Corporation Queensland Rail SunWater
Universities	Queensland University of Technology University of Cincinnati University of South Australia
Groups/Committees	Center for Intelligent Maintenance System CRC for Integrated Engineering Asset Management Machinery Information Management Open System Alliance

Table 4.4 – Interviewed organisations

process, data, organisation, and function views, with supplemented descriptions (i.e. literate modelling).

Analysis Methods and Functions

While a large percentage of data warehouse decision support systems solely use OLAP in their analysis, there are domain specific analysis techniques that are more intricate in design and data requirements. These can be in the form of methodologies, or at a lower level, algorithms.

Reliability and condition monitoring are two areas of asset management that use complex analysis methods. The reliability area presents simple reliability measures (e.g. mean time before failure), FMECA and root cause analysis, to reliability prediction using historical failures [156]. There are generally two ultimate goals of condition monitoring: diagnosis and prognosis. Diagnostic methodologies focus on using intelligent classifiers to compare healthy and non-healthy equipment states, while prognostic methodologies focus on using regression to predict the health of an asset. These two areas are also covered by the conceptual data modelling process.

Interviews

One of the metrics by which a decision support system is measured is its ability to satisfy end users. Thus eliciting requirements and knowledge from people within the field who may use an asset management data warehouse is important. While standards

do present the consensual views of people and groups, there may be elements missing from standards. This could be due to issues where it becomes too difficult to reach a consensus, issues that do not have enough resources for standardisation, or issues that do not require standardisation. Thus interviewing experts can fill gaps not discussed within standards. Interviews also have the advantage of exploratory analysis, as topics can evolve.

People from three types of organisations were interviewed. The organisations, which consisted of industry organisations, universities, and committees are shown in Table 4.4. The interviews included structured and unstructured questions, presentations given by the interviewer or the interviewee, and field demonstrations. The interviews also were conducted individually as well as in groups for both the interviewer/s and interviewee/s.

Business Documents

Although data model patterns, standards, information system data models and functionality, business process models, and analysis methods can be embodied as documents, there is other information that this medium provides. Documents include business vision and strategies, policies, manuals, log books, presentations, diagrams, and charts. While the information content in the aforementioned documents could be stored within business process models or information systems, their existence indicates a need for the document format (e.g. technological limitations within business process models/information systems, or a greater utility with a document format).

With the Internet increasing the ease of information dissemination, many business documents, such as annual reports, can be found online. While confidential documents are usually locked away on the corporate Intranet, some documents are made publicly available – particularly those by government organisations. The documents used in this research were a combination of business-supplied as well as Internet-harvested documents. While the majority of their content areas were uncovered via the other six modelling inputs, the documents provided a validation for these areas, as well as insight into business operations around the globe.

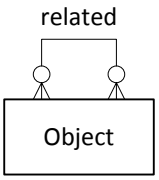


Figure 4.2 – Possible generic asset management data model

4.2.3 Issues in Conceptual Data Modelling

Scope of Integrated Areas

The conceptual data model aims to provide an understanding of how the different areas involved in asset management are integrated. For the areas outside of asset management, it only provides a guide on the actual content. For example, the financial account model is integrated into the asset management model through the account and transaction objects. While the non-asset management models presented in this research are comprehensive and cover all the functionality found in associated patterns literature, it is more than conceivable that these models may be lacking in certain areas for a sole implementation of this non-asset management area.

Generality and Implementation

As postulated and attempted by Hay [138] with his Universal Data Model, there exists a generic data model that is applicable to record every type of data in every type of circumstance. While the answer lies in judiciously using the common Entity Attribute Value (EAV) pattern, the answer is not Hay's Universal Data Model as it cannot represent relationship attributes, value types, and characteristic durations.

Despite the limitless expressiveness of a generic data model, the extreme example shown in Figure 4.2 of calling two related objects, 'a complete asset management data model', seems a little asinine. To be able to derive a useful implementation model, the conceptual data model needs to be more descriptive. However, there is no clear definition on the level of detail required, and hence design choices are made throughout the model.

External Data

An important characteristic to data warehouses is their ability to integrate data from sources external to the organisation. This includes currency rates, stock and future prices, economic metrics (e.g. currency exchange rates), weather, asset and inventory

prices, design data, manufacturer specifications, etc. Some of these data are already integrated into information systems either manually (e.g. entering asset specifications from documents) or automatically (e.g. obtaining data via a weather web service). It could be argued that once data are transferred to an organisation's information systems, they are no longer classified as external data. In either case, the distinction is not important as it is only a superficial categorisation and despite the conceptual data model including support for some external data, no explicit distinction is made.

Time

The concept of time plays an integral role in data warehousing. Time is a fundamental dimension that is found in nearly every data warehousing schema, and forms the basic analysis unit for reports. Many approaches have been devised to provide support for the treatment of time in data models, but most interpret the temporal relation as a sequence of states indexed by points in time [157]. Some temporal models also include the transaction time of creation and modification of these time points to fully capture all temporal information. Most entities in the conceptual data model are designed with a temporal consideration, while the transaction times are handled by metadata, as explained below.

Metadata

Metadata, commonly described as data about data, is an integral part of a physical data model. ISO 15489 and ISO 23081 discuss records management and metadata, and give guidelines on the management of both digital and non-digital data. Traditional metadata include statuses of records, dates of creation, modification, and deletion, as well as users who undertook these events. More advanced metadata can include access control records, retention periods, business context information, and data quality indicators. Metadata plays a notable role within data warehousing, and there are many resources on data warehouse metadata management [158] including modelling metadata [159] and mining metadata [160]. Metadata is not included in the conceptual data modelling, as the model focuses on the content. However, the aforementioned metadata types can be used in conjunction with the model.

Primitive Data Types

Object attributes of primitive data types (e.g. characters, integers, floating-point numbers, Booleans, strings, and time) were identified for the purposes of developing the data model, but were not explicitly described. Thus it is assumed that there is a

'name' attribute of string type for an AGENT, and a floating-point 'amount' attribute for a TRANSACTION. The justification behind the omission is that it is beyond the scope of a conceptual model to provide all the object attributes and the activity would be conducted within a logical data model process. The only exception is with the time data type, as (1) time is vital to data warehousing, and consequently (2) there are some unique uses of time that are not typically found in current information systems and require explanation.

4.3 Modelling Conventions

Hay [138] divides data modelling conventions into three areas: syntactic, positional, and semantic. Syntactic conventions are those that dictate the symbols that are used. Positional conventions involve the arrangement of symbols in relation to other elements. Semantic conventions deal with the grouping of elements based upon their meaning. This section describes the modelling convention used throughout the conceptual data model.

4.3.1 Syntactic Conventions

The symbolic notation used in the following sections is derived from a combination of UML and Crow's Foot notation. UML is a standardised specification language produced by the Object Management Group for object modelling. It uses graphical notation to create an abstract model of a system. Crow's Foot notation was proposed by Everest [161] in order to enhance the clarity of entity relationship notation.

Figure 4.3 presents an example of the modelling notation used in the conceptual data model. The foundation of all models is the entity. An entity is a distinct item of information. An entity type is the definition of a set of entities. For example, an entity of type *Person* can have a name of "Joe Bloggs" and gender of "Male". As the distinction between entity and entity type has been blurred over time in data modelling, references to the terms entity and entity type are synonymous. Entities can either be abstract (e.g. VEHICLE) or possible (e.g. BUS). Entities only can be defined by possible entity types - abstract types entities cannot be declared and only exist to assist with generalisation.

Generalisation is the UML term for what is commonly known as inheritance, and the symbol used is a line with an open triangle on the parent entity. Thus a child entity will

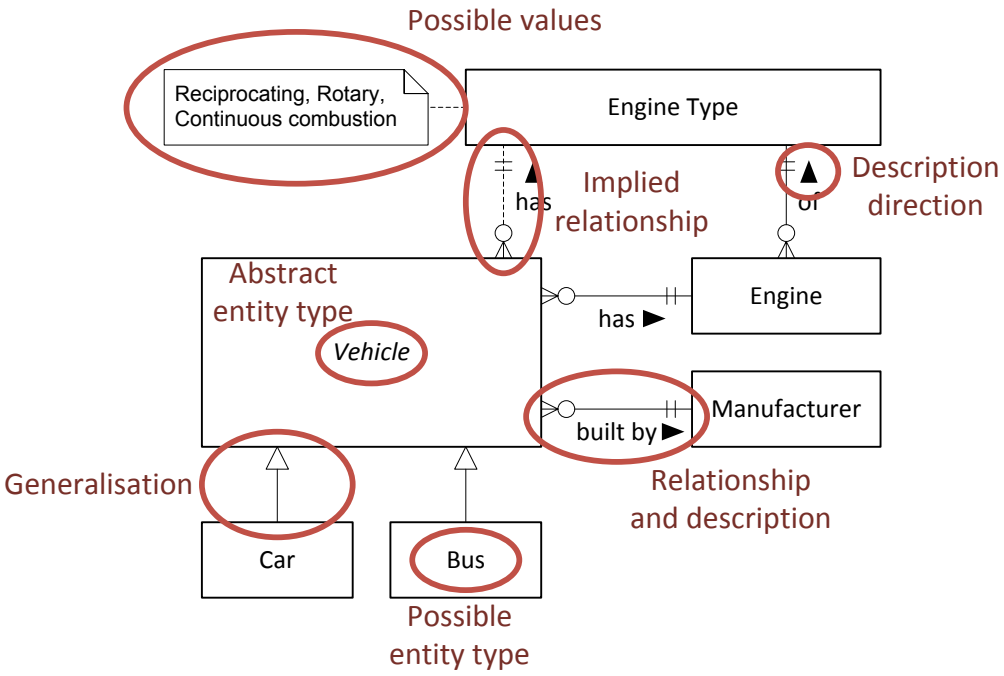


Figure 4.3 – Symbolic notation

inherit the attributes and relationships that exist on the parent entity. For example, CARS and BUSES would have the same attributes and relationships as VEHICLES.

Entity types are always written in singular, as the objective of modelling is to express the relationship of a single entity to zero, one, or many of another entity. However, to provide grammatical flow within the accompanying descriptions, plural forms may be used. The names of entities in the descriptions are shown in small capital letters, as shown in the previous paragraph.

The figure shows relationships between the entities as combinations of solid and dashed lines. A solid line is a normal relationship between two entities, while a dashed line is an implied relationship (i.e. an indirect relationship). Implied relationships are only used to emphasise a particular point in the descriptions.

The ends of relationships use the Crow's Foot notation to define cardinality. The circle represents zero, the dash represents one, and the crow's foot represents many. Thus the circle and dash represent zero or one, the dash and dash represent exactly one, the circle and crow's foot represent zero to many, and the dash and crow's foot represents

one to many. Crow's Foot notation was selected over UML's method as it is a more compact and symbolic method.

Relationships also contain a UML arrow that indicates the direction in which the relationship is meant to be read. For example, a VEHICLE is built by a MANUFACTURER and an ENGINE is of an ENGINE TYPE.

The UML comment symbol is used to indicate potential values that an entity can take. Such comments are usually only used with "type" entities whose values are used to define attributes of other entities. The list of values presented is not intended to be exhaustive, and is only indicative of potential values to aid understanding.

4.3.2 Positional Conventions

The simplest method to arrange a model is through random placement of entities on a page. As this method can be confusing, such a practice has evolved by still randomly placing entities, but with a minimal crossing of lines. Further refinements have seen 90° bends removed, and horizontal/vertical alignment of shapes on a page [138].

The CASE*Method technique orients entities such that the Crows' Foot on a relationship faces to the left and top of the diagram [134]. This has the effect of placing entities representing tangible objects in the lower right area of the diagram, while entities representing less tangible roles, interactions, and transactions move to the upper left.

Fowler [139] prescribes another philosophy that splits a model into a knowledge level and an operational level. The knowledge level forms the rules and types that govern the activities of the operational level. The operational level instantiates the knowledge level and contains more frequently changing information. This is the convention adopted in the conceptual data model although exceptions are made for clarity.

4.3.3 Semantic Conventions

Semantic conventions can be defined at two levels. The first deals with modelling similarities in business situations from different organisations. It is postulated by Hay [138] that despite the intrinsic differences between organisations, the models constructed come from a common set of contentions of thought, i.e. model patterns. Thus, identifying the basic elements of a specific business will lead to identification of

businesses in general. The second is the configuration of elements within the similar business situations. There are reoccurring structures that exist within models, such as roles or activities undertaken by people or organisations, classifications of objects into types, and processing of monetary transactions. At these two levels, a vocabulary of common business situations and structures can be developed that form the semantic conventions when developing data models. This vocabulary is taken from patterns literature, as well as studies into various companies.

4.4 Asset Management Conceptual Data Model

4.4.1 Assets

Akin to how customer relationship management systems are centred around customers, the core foundation of an asset management system is the asset. MIMOSA defines an asset as "an instantiated entity which can be physically tagged with an asset identifier and/or depreciated by an accounting system". ISO 15926 does not explicitly define an asset, however, engineering assets directly fall under the classification of a *Functional Physical Object* which is described as "a physical object that has functional continuity as its basis for identity".

Figure 4.4 shows a simple description of an ASSET. As with the OSA-EAI, an ASSET is a physical instance of a model. At its very least, information about an ASSET include a manufacturer serial number and a company specific identifier, while information about a MODEL at least includes a model number. A MODEL may be a product, equipment, or material or other MODEL TYPE. An ASSET can be comprised of other ASSETS, shown by the

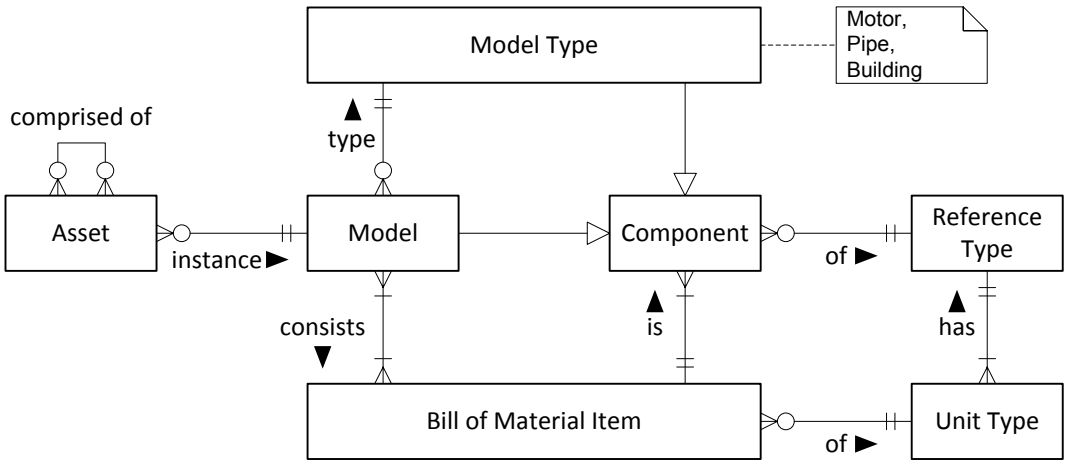


Figure 4.4 – Assets and models

reflexive reference of an ASSET to itself. This allows a pump system ASSET with serial number P001 being comprised of a pump with serial number PU001, motor M001, shaft S001, etc.

What is classically known as an asset type is linked to the MODEL as a MODEL TYPE. MODEL TYPE contains values such as 'pump', 'motor', and 'shaft'. Most asset management data models link model numbers and asset types directly to the asset object, sometimes at the exclusion of the model object, or creating relationships from the asset and model objects to the asset type object. The latter is often done to eliminate the extra join between the ASSET and MODEL TYPE. However, this is converse to database normalisation rules, and the model shown above is more semantically correct.

Each MODEL has a bill of material which is a list of parts (BILL OF MATERIAL ITEMS) required to create the associated ASSET. No bill of material object is created, as there is a one-to-one mapping between a MODEL and its bill of material. Each BILL OF MATERIAL ITEM contains the constituent COMPONENT required as well as the amount. The amount is expressed with the units of UNIT TYPE, and COMPONENTS have a REFERENCE TYPE to indicate which UNIT TYPES can be used to describe the amount. For example, a copper wire COMPONENT may have a reference type of a length in order to use a unit type of centimetres. Units of measurement are more comprehensively discussed in Section 4.4.10.

COMPONENTS can either be MODELS or MODEL TYPES. While MODELS are a necessary inclusion, MODEL TYPES also form COMPONENTS to provide a way of listing generic parts rather than a specific model part. Thus the bill of material for an audio speaker can include two 8Ω resistors, rather than specify a particular model of resistor (e.g. resistor R8O-1A from supplier A).

Attributes

An asset and model often have different properties or attributes attached to indicate the specifications or capabilities of the item. Asset specifications are data provided by the manufacturer on the ratings, operating parameters, functions, price etc. of the model of asset. While they can provide a reflection of how an asset might perform, the specification may not always be indicative of how the asset actually performs. Hence capabilities of each individual asset are measured and recorded by an organisation through assessments (e.g. an asset commissioning process). Thus capabilities record

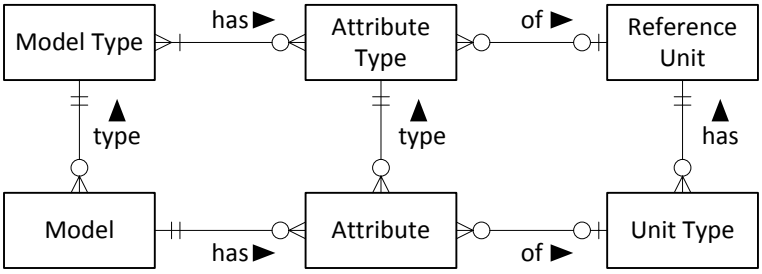


Figure 4.5 – Model specifications

the measured specifications of the instantiated asset, while specifications are those of the asset model.

Figure 4.5 shows that each MODEL has specifications (ATTRIBUTES) of a particular type (ATTRIBUTE TYPE) that are linked to the MODEL TYPE. MODEL TYPES are related to ATTRIBUTE TYPES as common attributes can exist amongst MODEL TYPES. For example, a cylindrical shaft MODEL TYPE has specification types of a length, diameter, and material type. Thus any cylindrical shaft MODEL must have a specified length, diameter, and material type. The specification type also has a REFERENCE UNIT associated while the specification is dictated by a UNIT TYPE.

Most model specifications are static, and revisions of the model to modify specifications will usually result in the creation of a new model with new specifications. Hence, model specifications are not a function of time.

Fowler [139] uses the idea of a *scope* to modify the interpretation of the attribute. This modifier is enacted as part of the Activity/Measurement pattern and its relationship to model specifications is detailed in Section 4.4.11.

Capabilities of ASSETS are represented in Figure 4.6. The construct is similar to that of MODEL specifications. As ASSET and MODEL TYPE are not directly linked, a derived relationship is shown between the two objects to show the link between the knowledge and operational level.

While specifications are static, capabilities of assets have a temporal aspect and are ever changing. An asset may not run as efficiently as it used to, due to entropy of the system. To account for changing capabilities, ATTRIBUTES have applicability over a TIME PERIOD. A TIME PERIOD is an entity that has a beginning and ending point in time. As

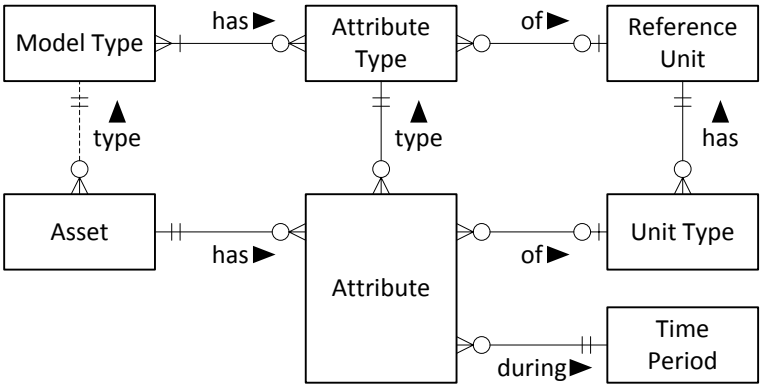


Figure 4.6 – Asset capabilities

ending points in time are not always known, these values are commonly left as *null* (with specific procedures dictated by the data management policies of an organisation).

The use of the common entity, ATTRIBUTE, hints the similarities between specifications and capabilities. While two individual entities could have been used to describe these types of data, a single entity was used to provide a simpler model.

Asset Structure

It was shown earlier that an ASSET was related to another ASSET through a many-to-many relationship. While this is sufficient for most needs, the structure in Figure 4.7 shows the relation of ASSETS through a more flexible structure.

As shown in the previous section, an ASSET is indirectly related to a MODEL TYPE. Each ASSET has a relationship with another ASSET at the operational level, while MODEL TYPES have relationships with ASSOCIATION TYPES at the knowledge level. For example, a 'SCADA system' MODEL TYPE may have an ASSOCIATION TYPE of a 'controller' for the 'pump

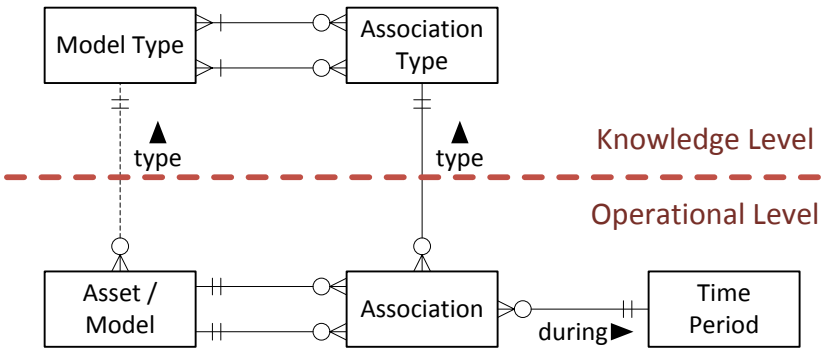


Figure 4.7 – Asset/model association structure

system' MODEL TYPE. By using ASSOCIATION TYPES, the model eschews the limited parent-child relationship found in other models and allows for a broad range of semantic relationships. Temporal aspects of relations are captured through the TIME PERIOD object within the relation. Thus modifications to ASSET relationships (e.g. upgrades) can be captured within the model.

4.4.2 Structural Patterns

The ASSET and MODEL models described above exhibit two generic structural patterns that can be applied to subsequent models to improve their consistency and coherency. These structural patterns build on top of existing generic patterns work, by combining advantages found in various patterns.

Attributable Object Pattern

The ASSET and MODEL ATTRIBUTE structures can be generalised to describe ATTRIBUTES for any object. An abstract ATTRIBUTABLE OBJECT of ATTRIBUTABLE OBJECT TYPE is shown in Figure 4.8 and using the EAV approach, ATTRIBUTES are attached to the ATTRIBUTABLE OBJECT. The knowledge level links the ATTRIBUTABLE OBJECT TYPE to the ATTRIBUTE TYPE via an OBJECT TYPE ATTRIBUTE which specifies the REQUIREMENT TYPE, such that mandatory or optional relationships can be enforced at the operational level. Using a whitelist, only the specified ATTRIBUTE TYPES are allowed to be attributed to an object. Using a blacklist

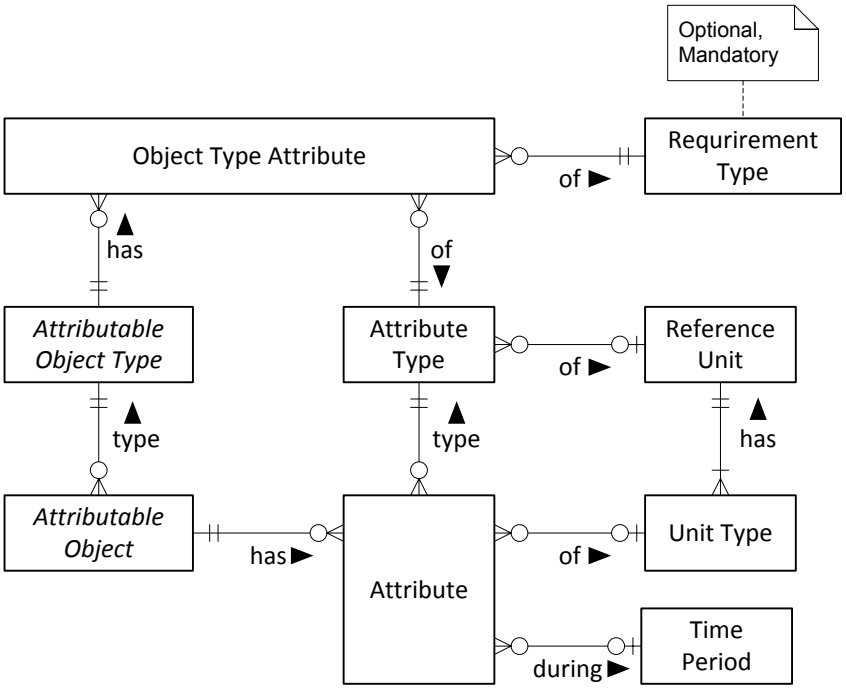


Figure 4.8 – Object attribute pattern

would lead to a scenario where the number of ATTRIBUTE TYPE rules would be proportional to the product of the number of ATTRIBUTE OBJECT TYPES and ATTRIBUTE TYPES. The temporal existence of the attribute is captured by the TIME PERIOD, however the cardinality is zero-to-one to accommodate cases where time is not relevant (e.g. MODEL specifications).

It should be noted that this object attribute pattern is used when there are a significant amount of mutable attributes for an object. Due to the EAV approach, the type of an entity is mutable and non-common attributes cannot be hardcoded into the entity. For example, a 'car' ASSET would have different attributes (gear ratios and fuel tank capacity) to a 'dam' ASSET (water capacity and physical dimensions).

Associable Object Pattern

The ASSET and MODEL ASSOCIATION structures can also be generalised to relate any object to another. Figure 4.9 shows a generic object association pattern that links two different or two equivalent ASSOCIABLE OBJECTS via an ASSOCIATION. At the knowledge level, ASSOCIABLE OBJECT TYPES relate to ASSOCIATION TYPES for the enforcement of rules.

ASSOCIATIONS are used when there are multiple ways of relating ASSOCIABLE OBJECTS. For example, an AGENT can be a *creator* of an ASSET, an AGENT can be a *seller* of an ASSET, and an AGENT can be an *owner* of an ASSET. If only a single relationship can exist between two objects, the ASSOCIABLE OBJECT pattern should not be used, and instead, a hardcoded relationship should be formed.

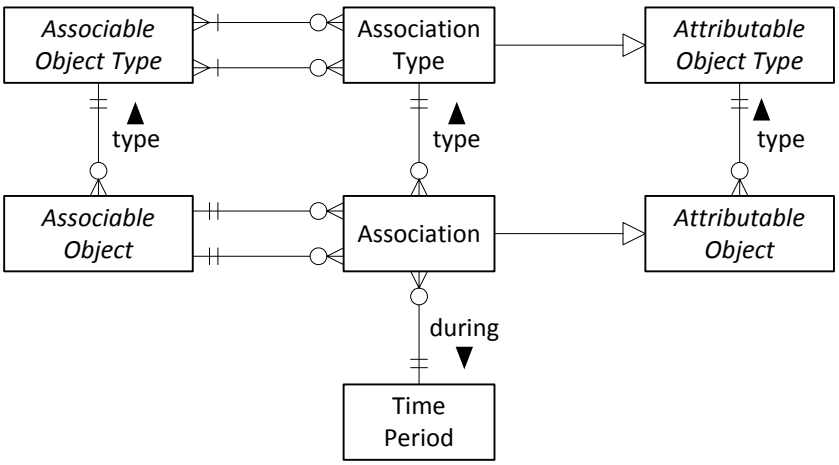


Figure 4.9 – Object association pattern

ASSOCIATIONS can also be used to create groups and/or hierarchies, and giving the ASSOCIATION entity a *name* attribute allows these relationships to be named. For example, three AGENTS may be related to each other through a *colleague* ASSOCIATION TYPE, with the ASSOCIATION *name* being 'Condition Monitoring Department'.

ASSOCIATIONS are also a type of ATTRIBUTABLE OBJECT, and can have their own attributes. For example, a supplier ASSOCIATION between an AGENT and MODEL may have a cost ATTRIBUTE. A TIME PERIOD expresses a validity period where the ASSOCIATION is applicable. For example, a person is an employee of an organisation during a certain period, or a person is the manager of a manufacturing plant during a certain period.

Use of Patterns

The use of the two above patterns is pervasive in the conceptual data model as they provide a method to simplify notation for attributes of and associations between objects. Figure 4.10 provides an example of how Figure 4.6 and Figure 4.7 can be simplified using the patterns and generalisation/inheritance. An ASSET is both an ATTRIBUTABLE OBJECT as well as an ASSOCIABLE OBJECT. As an asset effectively "becomes" an ATTRIBUTABLE OBJECT, it can be substituted into the ATTRIBUTABLE OBJECT pattern structure, and consequently gains the relations to ATTRIBUTES and corresponding REQUIREMENT TYPES, UNIT TYPES, and TIME PERIODS. Similarly, the ASSET entity is substituted into the ASSOCIABLE OBJECT pattern and has relations to ASSOCIATION and the corresponding TIME PERIOD and ASSOCIATION ATTRIBUTES.

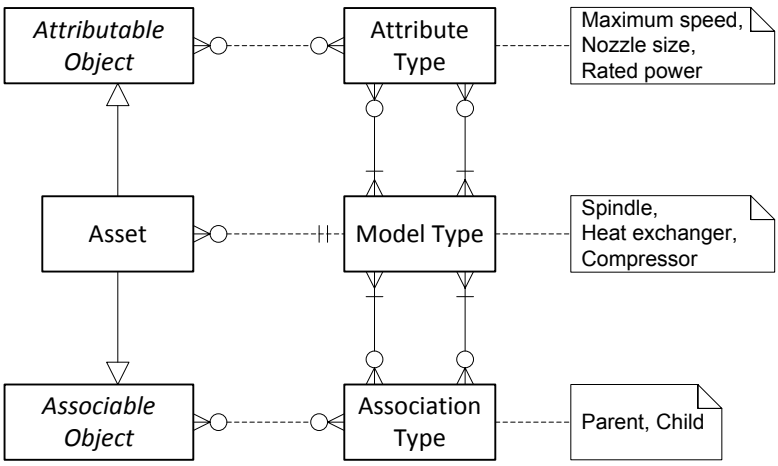


Figure 4.10 – Asset attributes and associations using patterns

It must be noted that ASSOCIABLE OBJECTS are not always of equivalent types, and that two different objects (e.g. an ASSET and an AGENT) can be associated through this technique. Two notational styles are used in the conceptual data model, with both generalisation/inheritance of an ASSOCIABLE OBJECT (usually to denote reflexive associations) as in Figure 4.10, or direct use of an ASSOCIATION object as in Figure 4.15 (for associations between two different entities).

Conclusion

These two EAV structural patterns build on the Universal Data Model [138] by adding functionality at the knowledge level, and include TIME PERIODS and UNIT TYPES. While the structure can be simplified by collapsing these last two items back into the structure (i.e. TIME PERIODS are turned into ATTRIBUTES while UNIT TYPES are represented as ASSOCIABLE/ATTRIBUTABLE OBJECTS), such a modification would increase the complexity in analysing and understanding asset management concepts. Hence they are explicitly represented, and the two object patterns are used exhaustively in the following sections.

4.4.3 Segments

Before assets can be utilised, they are placed in a certain location such that they can fulfil their intended purpose. This 'location' is termed as a *functional location* by SAP, while MIMOSA uses the term *segment*. While a segment is a container on which an asset can be placed, it is not equivalent to a geographical location (although a segment and geographical location can be related). Both segments and geographical locations are arbitrarily divided spatial concepts, and a segment can be positioned on many geographic locations while a geographic location can span many segments.

Figure 4.11 shows an instance where the OBJECT ASSOCIATION pattern is abandoned in preference of typed (or fixed) associations between objects. ASSET, SEGMENT and GEOGRAPHIC LOCATION form the objects, while PLACEMENT and POSITION are the typed associations. While this structure does not have the flexibility compared to the object association pattern's EAV approach, flexibility is not required in this case. The only relationship that ever exists between a SEGMENT and GEOGRAPHIC LOCATION is an arbitrary division of space (i.e. POSITION). The only relationship between a SEGMENT and an ASSET is to provide a container and as such, PLACEMENT is the only required association. This holds true for the entire lifecycle of an ASSET. At creation, the ASSET will be first placed at the SEGMENT on which it is manufactured, while at disposal, the ASSET will be placed on

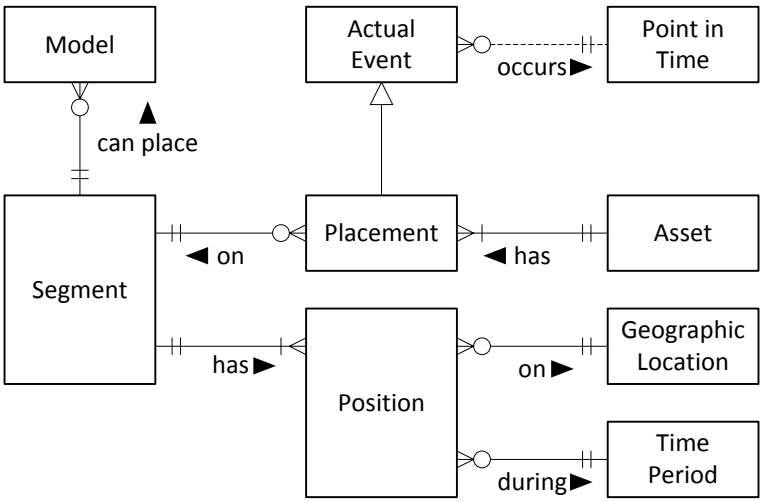


Figure 4.11 – Asset locations

the SEGMENT by which it is disposed (which could be a location external to the organisation).

In terms of temporal applicability, PLACEMENT is a type of ACTUAL EVENT (EVENTS are discussed in Section 4.4.6), and ASSETS can be moved to different SEGMENTS during different POINTS IN TIME. This is the same concept with the *asset on segment* table within the OSA-EAI. As PLACEMENT is an EVENT, the PLACEMENT process (e.g. installation) can be captured through a causal ACTIVITY (see Section 4.4.5). While POSITIONING a SEGMENT on a GEOGRAPHIC LOCATION could also be considered as an EVENT (from a pedantic perspective), it is not considered as such in this model for simplicity.

SEGMENTS can be accorded multiple GEOGRAPHIC LOCATIONS through the POSITION entity. Mobile segments such as vehicles can be positioned at different GEOGRAPHIC LOCATIONS at different TIME PERIODS. For systems that only record static SEGMENTS, the POSITION entity is not required, as the mapping between a SEGMENT and GEOGRAPHIC LOCATION would be 1:1. The SEGMENT would then store the geographic details as entity attributes.

One unique characteristic is the relationship between SEGMENT and MODEL. This relationship provides a list of ASSET MODELS that can be placed on the SEGMENT. This restriction may be due to a legal or business constraint – for example, only refrigeration MODELS with an energy rating of four stars or higher can be used.

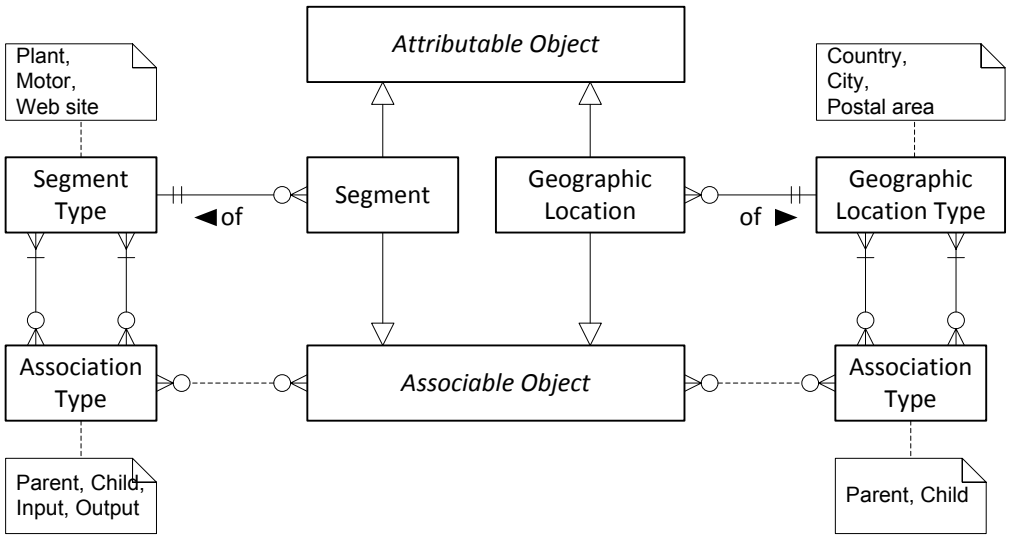


Figure 4.12 – Location attributes and associations

Like ASSETS, SEGMENTS and GEOGRAPHIC LOCATIONS are both ATTRIBUTABLE OBJECTS and ASSOCIABLE OBJECTS as shown in Figure 4.12. SEGMENT ATTRIBUTES can specify capability requirements for which corresponding ASSETS are required to fulfil (e.g. a particular SEGMENT requires a 100 hp motor). These capability requirements can be enacted as forecasts if the TIME PERIOD is set to a future date. Descriptive ATTRIBUTES can be given to GEOGRAPHIC LOCATIONS, such as dimensions, longitude and latitude coordinates, or soil type.

ASSOCIATIONS allow SEGMENTS to be composed of other SEGMENTS (e.g. buildings containing rooms) and GEOGRAPHIC LOCATIONS to be composed of other GEOGRAPHIC LOCATIONS (e.g. countries containing states which contain cities). The term *site* is used within the OSA-EAI to denote an aggregation of segments, but is excluded in this model. A site only serves to provide a hierarchical distinction rather than a functional one, and is effectively a specialisation of a segment. Thus to implement the same functionality in this model, a SEGMENT is designated as a site via its SEGMENT TYPE.

SEGMENT ASSOCIATIONS also allow processes to be modelled for process or manufacturing organisations. Akin to a process diagram, SEGMENTS form the nodes while ASSOCIATIONS form the connections between the nodes, with the type of connection specified by the ASSOCIATION TYPE. Process ACTIVITIES or MEASUREMENTS (see Section 4.4.11) can then occur on or between these SEGMENTS. Such a SEGMENT process model can also serve as a descriptive mechanism for a reliability block diagram (RBD) in reliability analysis.

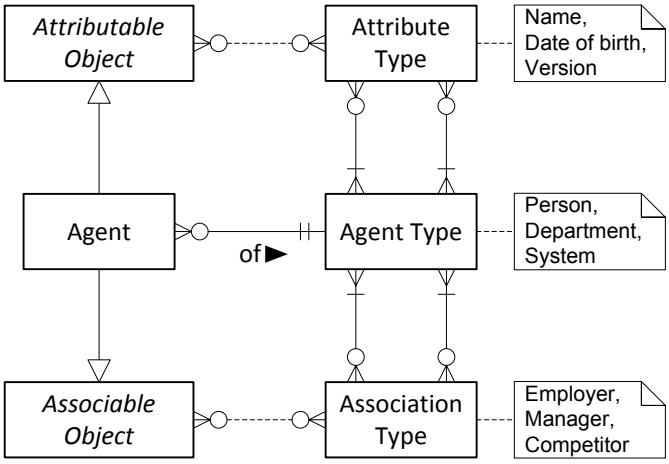


Figure 4.13 – Agent attributes and associations

RBDs are used in calculating the system reliability from subsidiary component reliabilities. Thus the reliability of a process system can be calculated using an ASSET'S reliability ATTRIBUTES (i.e. the component reliability) and the SEGMENT ASSOCIATION structure.

Consideration was given to GEOGRAPHIC LOCATION being a type of ASSET as it shares similar properties in that it can be physically tagged and can depreciate in value. The conceptual data model follows the ISO 15926 convention of making the distinction between a *spatial location* (equivalent to a GEOGRAPHIC LOCATION) and *materialized physical object* (equivalent to an ASSET) with both being a *physical object* (equivalent to a RESOURCE – see Section 4.4.4).

4.4.4 Agents

An agent is defined as "an animate object that makes various types of assessments" by MIMOSA. The given description is vague and limiting however, the types of an agent both broaden and qualify the description. An agent can consist of a person, group, organisation, or software.

The AGENT model uses both the object attribute and object association patterns as shown in Figure 4.13. ATTRIBUTES of an AGENT include the name, address, salary of a person, skills, certifications etc. As ATTRIBUTES can be temporal (e.g. an agent's address may change over time), attributes are associated with a TIME PERIOD in the attribute pattern. AGENTS also have a many-to-many relationship with other AGENTS and multiple roles can be specified. For example, an organisation may employ a person, that person

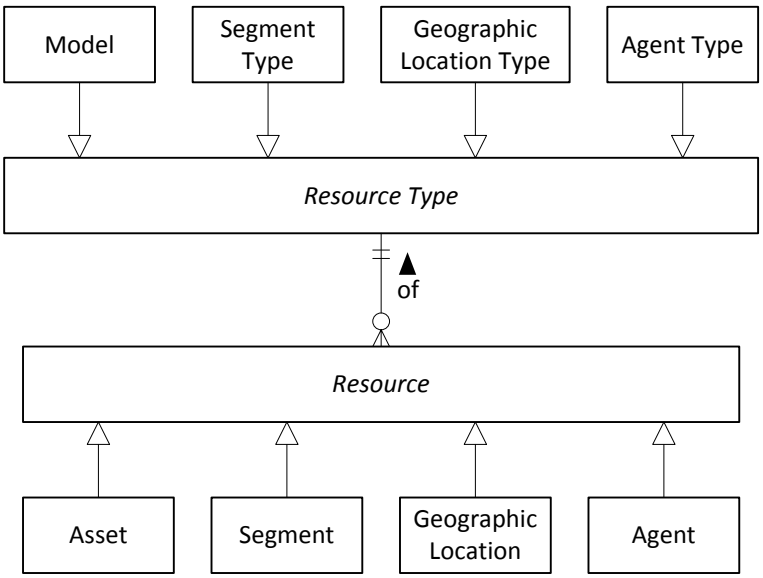


Figure 4.14 – Resources

can be part of a group, that group may manage a computer system. While this structure also allows for intricate customer relationship management functionality, the primary intent is for asset management.

There is an inconsistent overlap between the *enterprise* and *agent* objects within the OSA-EAI, and the above model does away with the enterprise object, since it is semantically equivalent to an AGENT with an enterprise AGENT TYPE. Similar to the *site* object discussed in the previous section, the enterprise object only serves a nominal purpose at the implementation level.

As indicated in Section 4.4.1, ASSETS and GEOGRAPHIC LOCATIONS fall under the *physical object* classification in ISO 15926. This is also true of the AGENTS and SEGMENTS. Figure 4.14 shows these entities collective grouped as RESOURCES. While ISO 15926 uses the nomenclature of "class of <item>" to describe the type of an object, the conceptual data model uses a simpler method by appending the word 'type' to the entity in question.

Having a RESOURCE parent object simplifies the notation in creating relationships between AGENTS and ASSETS, SEGMENTS, and GEOGRAPHIC LOCATIONS as seen in Figure 4.15. AGENTS can be 'owners' or 'managers' of ASSETS, SEGMENTS, and GEOGRAPHIC LOCATIONS. Mobile SEGMENTS (discussed in Section 4.4.3) may be positioned on GEOGRAPHIC LOCATIONS that may be 'owned' by a different organisation. As an AGENT is also listed as a

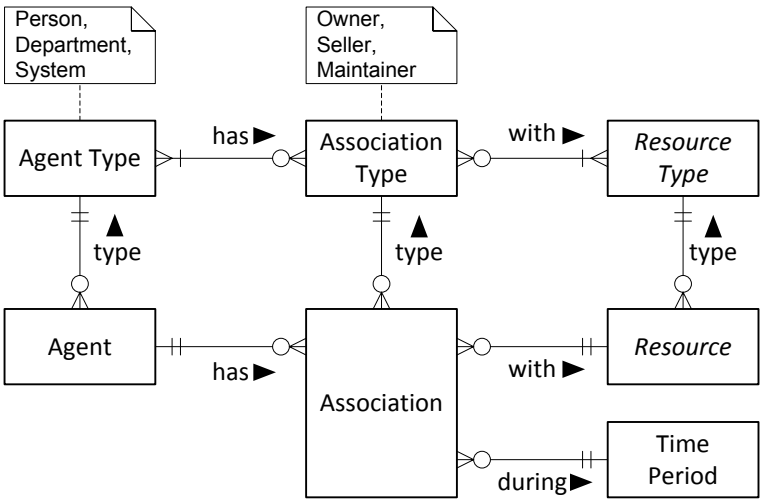


Figure 4.15 – Agent roles with resources

RESOURCE, the ASSOCIATION between an AGENT and RESOURCE replicates that shown in Figure 4.13.

A temptation may arise to further abstract the associations by making RESOURCES itself an ASSOCIABLE OBJECT. However, not all RESOURCES are directly related to each other. For example, ASSET and GEOGRAPHIC LOCATION are indirectly related through a SEGMENT, and a RESOURCE reflexive association would circumvent the function of SEGMENTS.

AGENTS have a special relationship with MODELS in that there are AGENTS who design, manufacture, supply, and provide professional services for a MODEL. As multiple AGENTS can supply a MODEL, the price, taxes, currency type, and other relevant information regarding the cost of an MODEL offered by a supplier are recorded as ATTRIBUTES of the ASSOCIATION, rather than ATTRIBUTES of the MODEL. The TIME PERIOD attached to the attribute pattern also allows variations in these values over time.

4.4.5 Activities

Activities are actions that bring about a change in the environment in conjunction with RESOURCES. The conceptual data model strays from OSA-EAI terminology which terms activities as *work orders*. The term *work order* conveys a limited set of actions (such as installation and maintenance) while the term *activity* suggests a broader set of actions. The more generic definition is required (in particular, when activities are used to define large projects), and hence the deviation from the OSA-EAI Terminology Dictionary.

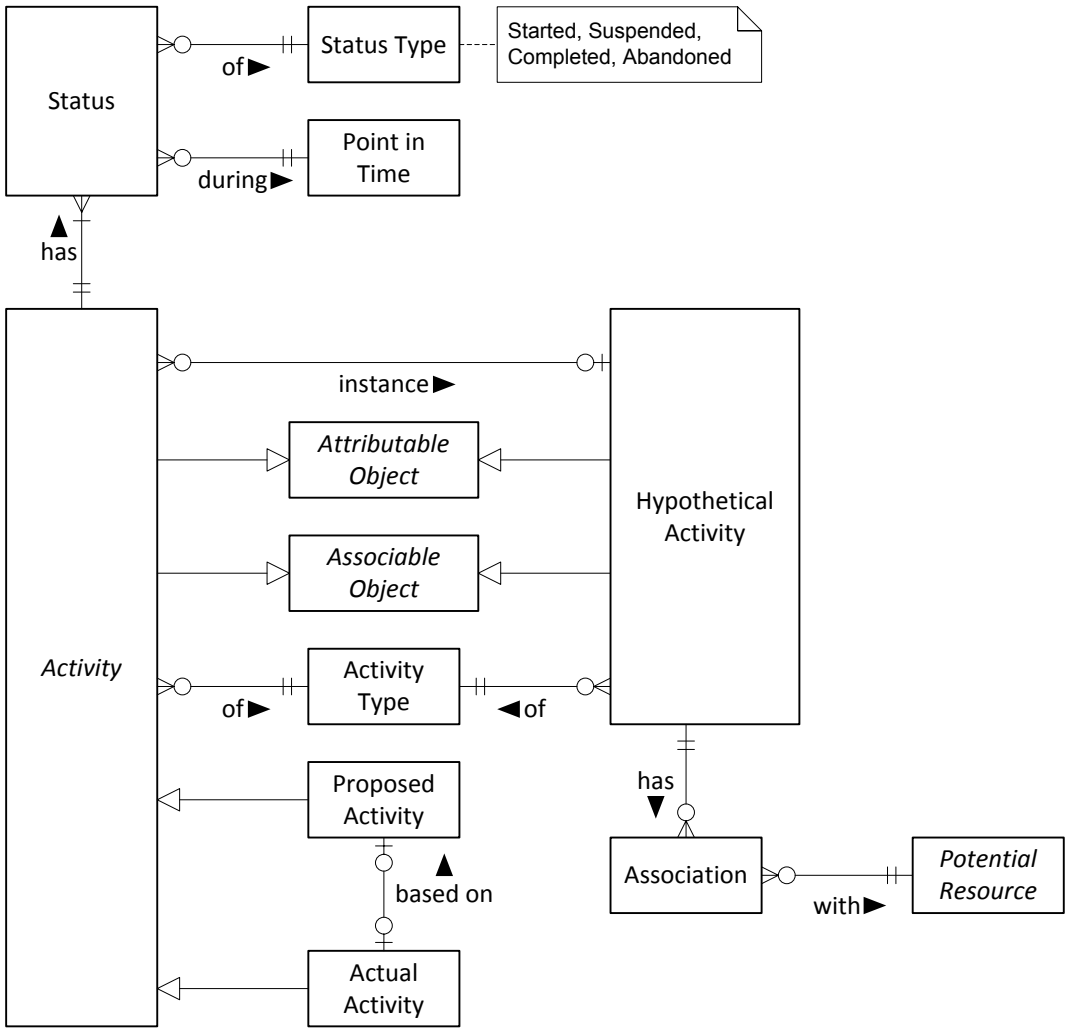


Figure 4.16 – Activity knowledge level

Both ISO 15926 and the OSA-EAI make distinctions between an activity and an event: events are marked as occurring at a single point in time and have no duration, while activities occur between two points in time. For example, the operation of a tool changer machine would be considered as an activity while the start of a tool change process would be considered as an event. Events are discussed in Section 4.4.6.

Figure 4.16 shows the knowledge level of an ACTIVITY. ACTIVITY itself is an abstract object, being comprised of either a PROPOSED ACTIVITY or an ACTUAL ACTIVITY. An ACTUAL ACTIVITY is one that is occurring/has occurred while a PROPOSED ACTIVITY is a plan of an ACTUAL ACTIVITY that will occur. Fowler's [139] planning pattern separates the two, while other patterns authors do not. ACTUAL AND PROPOSED ACTIVITIES are separate entities but are linked to indicate that ACTUAL ACTIVITIES can be based on PROPOSED ACTIVITIES (proposals do not always lead to actions, and conversely, actions may not

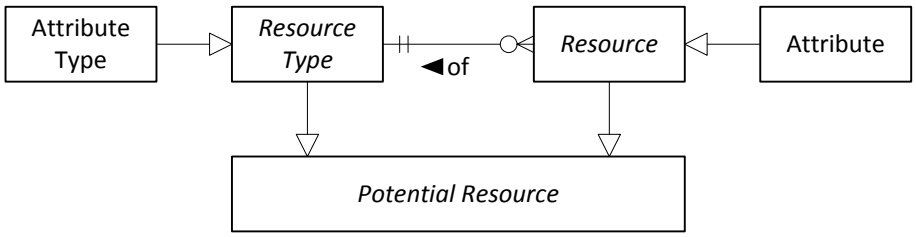


Figure 4.17 – Additional resources

always be based on a plan). ACTIVITIES are an ASSOCIABLE OBJECT and use the association pattern to form groups and hierarchies. Thus large projects can be composed of smaller, linked ACTIVITIES. ASSOCIATIONS also allow for detailing the auditing of ACTIVITIES, as audits themselves are ACTIVITIES.

Both ACTUAL AND PROPOSED ACTIVITIES have a STATUS that can range between the start of an ACTIVITY, through to its suspension, completion, or abandonment. For an ACTUAL ACTIVITY, the STATUS relates to the actions described in the ACTIVITY, while for a PROPOSED ACTIVITY, the STATUS relates to the action of developing the proposal. As the STATUS of an ACTIVITY can change over time, a TIME PERIOD indicates its temporal applicability.

ACTIVITIES are based on a HYPOTHESISED ACTIVITY, which serves as a reference activity or template, and has a possibility of occurring (similar to class instantiation for an object in object oriented programming) as a PROPOSED OR ACTUAL ACTIVITY. HYPOTHETICAL ACTIVITIES form a superset of PROPOSED AND ACTUAL ACTIVITIES and hence within an implementation, if an unforseen ACTIVITY occurs, it is added to the list of HYPOTHETICAL ACTIVITIES.

A list of POTENTIAL RESOURCES is linked to the HYPOTHETICAL ACTIVITY and is named as such because they have the potential to be used in an ACTIVITY. Thus ACTIVITIES may require the use of MODEL TYPES (e.g. rotor), SEGMENT TYPES (e.g. workshop), MODELS (e.g. MD-301), AGENT TYPES (e.g. computer system), ATTRIBUTES (e.g. skills/certifications), ASSETS, SEGMENTS, or AGENTS, as seen by the definition of RESOURCES in Figure 4.14 and its extension in Figure 4.17. This structure allows a large knowledge base to be constructed that contains different potential activities along with the resources that the activities require. The use of the term “hypothetical resource” was avoided as it conveys a meaning that would exclude tangible RESOURCES.

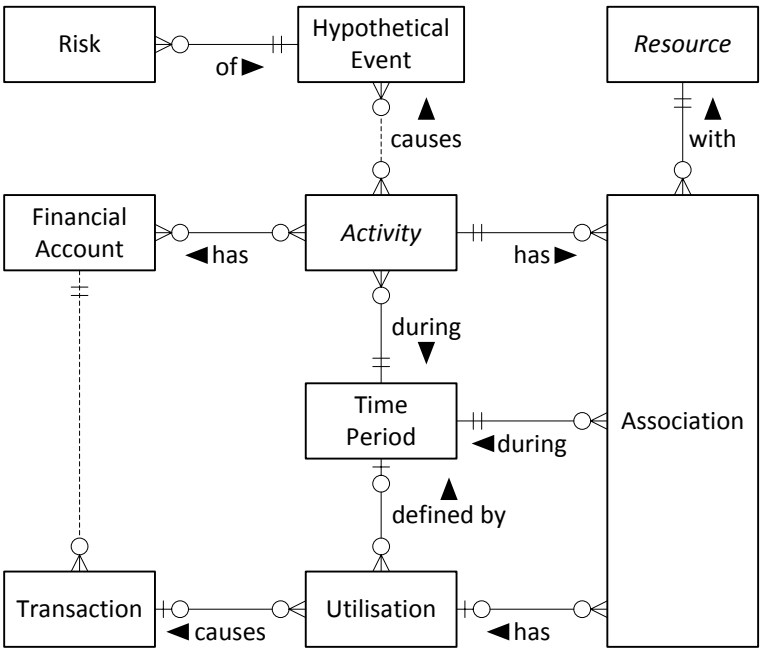


Figure 4.18 – Activity operating level

At the activity operating level (shown in Figure 4.18), the associated RESOURCES are restricted to physical instances, as PROPOSED AND ACTUAL ACTIVITIES move beyond the ambiguity of types and into real instances. While it can be debated that PROPOSED ACTIVITIES should include RESOURCE TYPES (as opposed to solely using actual instances), the solution is dependent on rules set by an organisation.

There is a dual relationship between the ASSOCIATION of RESOURCES and TIME PERIOD. The first is a direct relationship, and is derived from the object association pattern to record the duration of the ASSOCIATION (i.e. time-base allocation of the RESOURCE to the ACTIVITY). The second relationship is indirectly through the UTILISATION object which records when the RESOURCE was actually in use. As UTILISATION does not always involve time (e.g. assets that use counters or the amount of inventory required) the cardinality of UTILISATION to TIME PERIOD has a minimum of zero (with the counter/amount stored as an attribute of the UTILISATION entity).

The activity operating level also shows two characteristics that are missing from existing patterns literature and asset management data models. The first is associating the ACTIVITY with the various RISKS that can befall it. RISKS (discussed in Section 4.4.7) are an assessment of the HYPOTHETICAL EVENTS that are caused by an ACTIVITY.

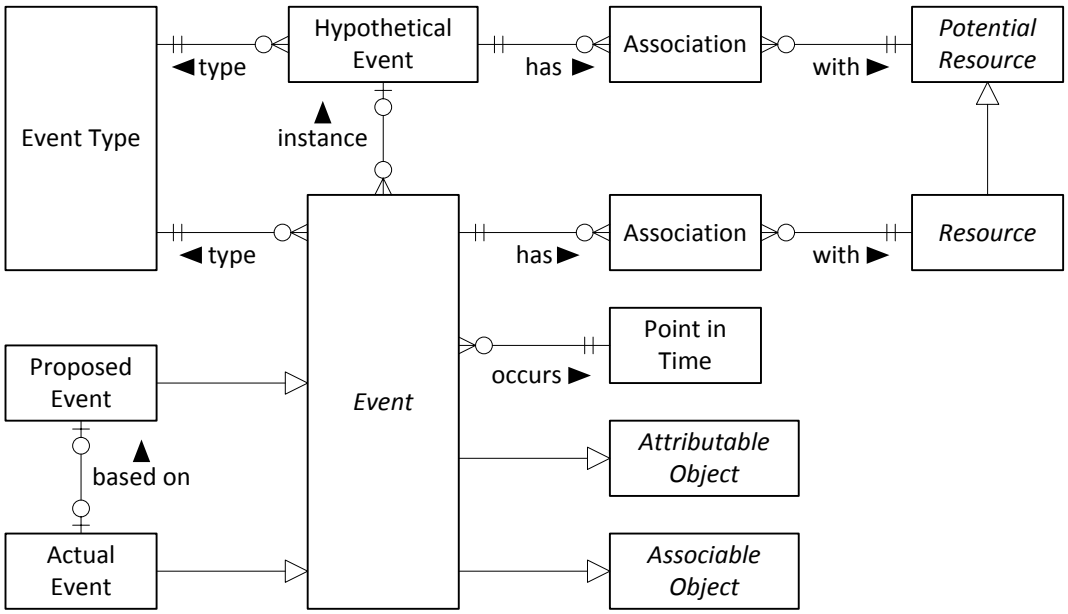


Figure 4.19 – Events

The second often omitted characteristic is FINANCIAL ACCOUNT data. FINANCIAL ACCOUNTS are attached to ACTIVITIES, and record the amount spent or earned in relation to the ACTIVITY. In order to track lifecycle financial data of RESOURCES, any UTILISATION of RESOURCES can cause a TRANSACTION to be posted to a FINANCIAL ACCOUNT. Hence TRANSACTIONS could record the wages of an employee for a maintenance order, the rental price of equipment or locations, or usage-based depreciation of equipment. FINANCIAL ACCOUNTS are discussed in more detail in Section 4.4.8.

4.4.6 Events

As opposed to activities that occur over a duration, events mark a change at a single point in time. Events could mark stages within an asset or system operating procedure, such as the starting of motors, opening and closing of valves, etc. An event could also indicate the failure time of equipment, or occurrence of an accident.

The event model in Figure 4.19 shares similarities with the activity model. PROPOSED AND ACTUAL EVENTS inherit relationships and attributes from EVENT, and ACTUAL EVENTS can be based on PROPOSED EVENTS for planning purposes (e.g. scheduling within a MES). EVENTS can be an instance of a HYPOTHETICAL EVENT, which contains a set of all EVENTS that could possibly occur. While this list can potentially grow to an enormous size, items can be culled based on their likelihood of occurrence or aggregated into more generic events. Thus, the event of an employee slipping due to a banana peel can be

eliminated due to its low likelihood, or aggregated into the event of an employee accident.

As with ISO 15926, one difference between the event and activity models is that EVENTS occur at a POINT IN TIME, rather than over a TIME PERIOD. While a TIME PERIOD is usually implemented as two date and time fields, a POINT IN TIME is implemented as a singular date and time field.

ASSOCIATIONS to RESOURCES for EVENTS and POTENTIAL RESOURCES for HYPOTHETICAL EVENTS are similar to the ASSOCIATIONS to the same objects in the activity model. These allow an EVENT to be linked to the ASSETS, SEGMENTS, AGENTS, etc. that are involved with the EVENT (e.g. the motor ASSET failed). The reflexive associations of an ASSOCIABLE OBJECT are applied to EVENTS to enable grouping (e.g. a machine start up event which has sub-events) and hierarchies of EVENTS (e.g. a breakdown of an employee accident into more detailed types). While this structure can be used to describe a causal chain of events (e.g. an ASSET fault tree), a more flexible structure is described below.

Cause and Effects

Both EVENTS and ACTIVITIES have a reason for why they occurred or they were undertaken, as well as their resulting effects. ISO 15926 provides support for such reasoning through the *cause_of_event* entity, which attributes the occurrence of an event to the occurrence of an activity. Thus the activity of refuelling a truck could cause a "fuel tank full" event. The causal model below follows this behaviour, but expands to allow greater flexibility.

A CAUSE is defined in Figure 4.20, as an occurrence of an ACTIVITY or EVENT that enables or instigates the undertaking of another ACTIVITY or EVENT. An EFFECT is similar – the occurrence of an ACTIVITY or EVENT due to the undertaking of another ACTIVITY or EVENT. As ACTIVITIES and EVENTS are both CAUSES and EFFECTS, there exist reflexive relationships as well as relationships between each. The specified cardinality between a CAUSE and EFFECT is zero-to-many, which is incorrect in a logical worldview as a cause *always* has an effect and an effect *always* has a cause. The reasoning behind the discrepancy is in avoiding an infinite regression/progression. As all CAUSES are essentially EFFECTS, if either side of the relationship was mandatory, the model would be forced to document every single event from before the creation to after the destruction of the universe.

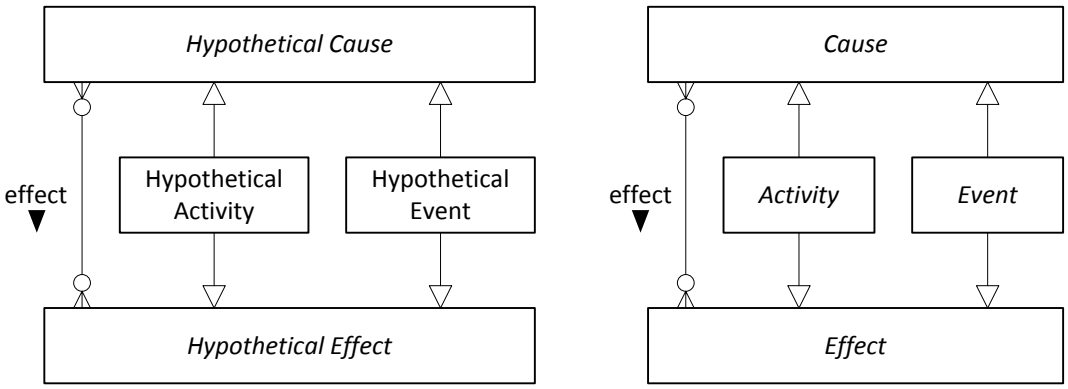


Figure 4.20 – Causes and effects

As is the case in ISO 15926, it can be argued that ACTIVITIES should not be included as a CAUSE, as ACTIVITIES always produce EVENTS and these EVENTS are the true cause. While this is a valid restriction, it does limit the data described by the model. At an implementation level, it can produce redundant data if the EVENT caused by an ACTIVITY is only nominal and there to satisfy the restriction (e.g. the completion of a review ACTIVITY causes a redundant review complete EVENT which triggers an auditing ACTIVITY). Thus to provide flexibility, both ACTIVITIES and EVENTS are considered CAUSES and EFFECTS.

A similar causal structure exists for HYPOTHETICAL CAUSES AND EFFECTS and HYPOTHETICAL ACTIVITIES AND EVENTS.

The cause and effect associations between EVENTS and HYPOTHETICAL EVENTS are particularly important for reliability analysis techniques such as fault tree analysis and root cause analysis that require associations for tree traversal. As the associated objects are both CAUSES (parents) and EFFECTS (children), the resulting implemented structure is a doubly linked tree. Thus traversal of the tree can be done in either direction, as opposed to the OSA-EAI, which only links to the children of an event. While double links can lead to increased performance at the expense of storage resources, the conceptual semantic value is increased for the event model.

4.4.7 Motivation

While the previous sections showed that ACTIVITIES are undertaken by an organisation, and that EVENTS or other ACTIVITIES can be the cause of a particular ACTIVITY, no underlying business reason was given for an ACTIVITY'S purpose. All organisations

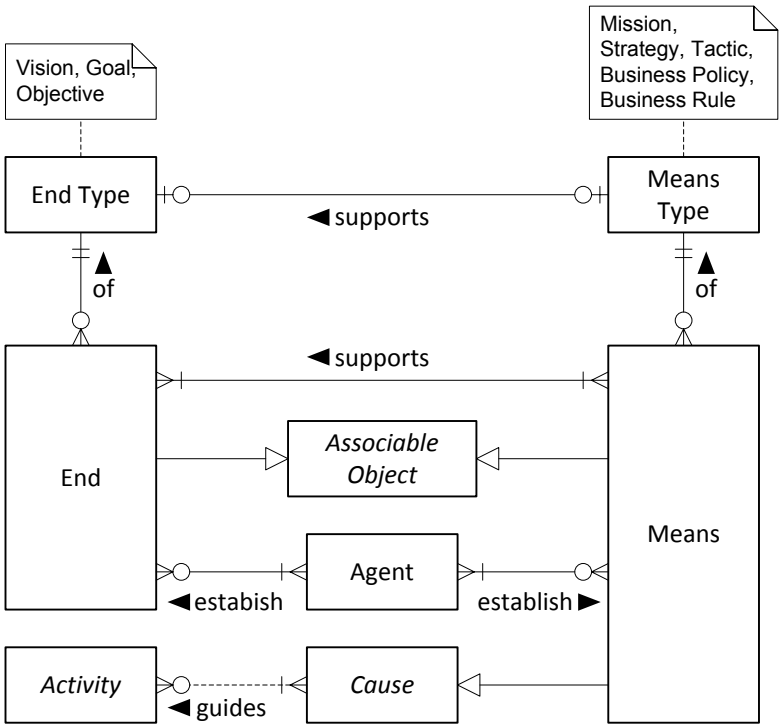


Figure 4.21 – Ends and means

(should) have a vision which is enacted through a mission. This mission is supported by setting various goals and objectives, through which activities are planned and carried out.

The Business Rules Group created the Business Motivation Model (BMM) in order to provide a structure to describe this type of business data, identifying the ends and means of business plans and the influencers that shape the elements of those plans [162]. The model also described assessments of impacts of those influencers, as well as directives that initiate and constrain the operation of the plans. The motivation model expressed below is largely derived from the BMM, and is integrated with the other objects in the conceptual data model.

Ends and Means

The main elements of a business plan are its ENDS and MEANS. An END is something the business sets out to accomplish, such as a vision, goal, or objective. A MEANS is a capability that can be called on, activated, or enforced to achieve ENDS, such as strategies, tactics, policies, and rules. As seen in Figure 4.21, both ENDS and MEANS are ASSOCIABLE OBJECTS, and hence multiple ENDS can be related and similarly, multiple MEANS can be related. As not all MEANS TYPES support END TYPES (missions only support

visions, but not objectives), the relationship between the two entities creates a knowledge level for such rules. The AGENTS who established the ENDS and MEANS define an audit trail, and are made mandatory in the conceptual data model (as opposed to being optional in the BMM).

MEANS are also the first CAUSE to ACTIVITIES and govern how an ACTIVITY is undertaken. As all ACTIVITIES are related to MEANS and all MEANS are related to an END, all organisational ACTIVITIES are consequently enacted to achieve a desired result.

Rule Enforcement

MEANS of a 'directive' MEANS TYPE can be given different ENFORCEMENT RULES (see Figure 4.22) to indicate the extent to which the directive is to be enforced. The Business Rules Group list different ENFORCEMENT LEVELS to indicate how each ENFORCEMENT RULE is upheld. These include strictly enforced, pre and post authorised overrides, overrides with explanations, and guidelines. The ENFORCEMENT RULE also includes an action to be undertaken (HYPOTHETICAL EVENT) if the rule is to be enforced.

If a directive is breached, then the ENFORCEMENT of a rule causes the occurrence of an ACTUAL EVENT. For example, a breach of the directive "company drivers must not receive a moving traffic violation" could cause an "enrolment for safe driving counselling" EVENT.

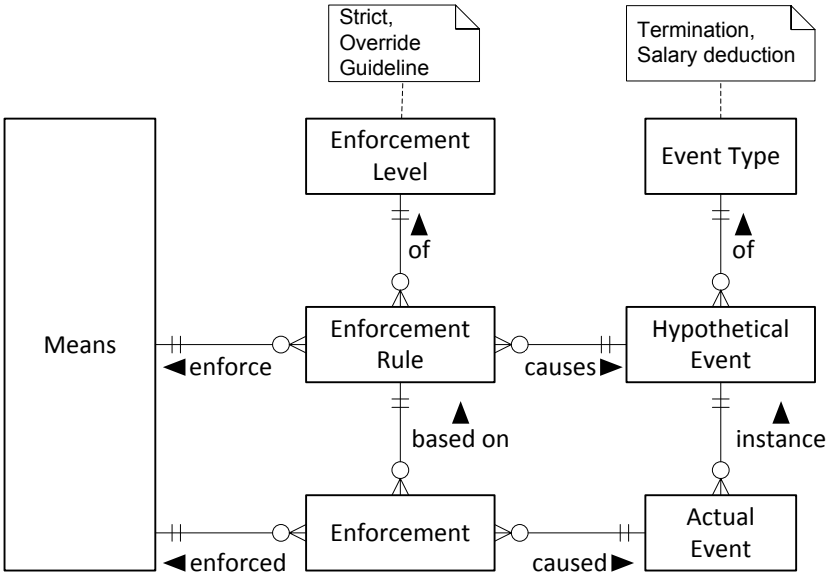


Figure 4.22 – Rule enforcement

Assessment of Influencers

There are various factors that can affect business plans, and these are recognised as INFLUENCERS. INFLUENCERS have the capability to "produce an effect without apparent exertion of tangible force or direct exercise of command, and often without deliberate effort or intent" [162]. INFLUENCERS are categorised as either internal (assumptions, corporate values, habits, infrastructure, issues, management prerogatives, and resources) or external (competitors, customers, environment, partners, regulations, suppliers, and technology). AGENTS can be the source of both internal and external INFLUENCERS.

As seen in Figure 4.23, an ASSESSMENT is made to judge the POTENTIAL IMPACTS of an INFLUENCER upon the ENDS and MEANS of an organisation. An ASSESSMENT is a MEASUREMENT (covered in Section 4.4.11) which is itself an ACTIVITY. Thus an ASSESSMENT has relationships to the assessing AGENTS and can be composed of sub-assessments. While the BMM prescribes a SWOT (strengths, weaknesses, opportunities, and threats) analysis for ASSESSMENT TYPES, any business analysis methodology is supported.

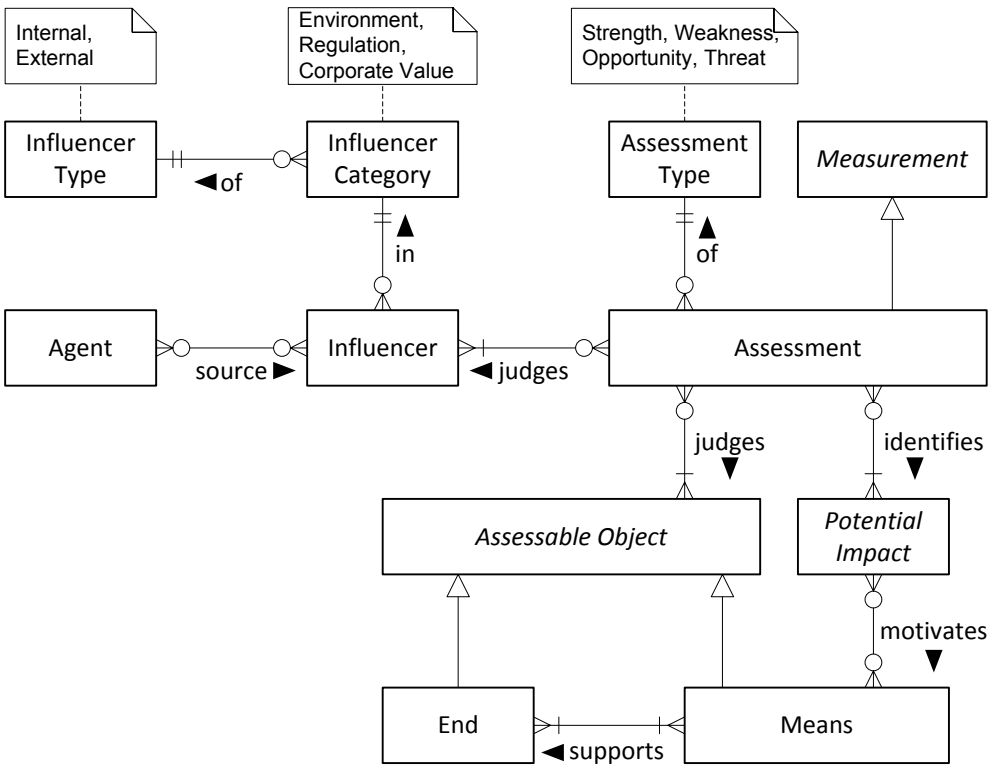


Figure 4.23 – Assessment of influencers

Potential Impacts

An ASSESSMENT can identify significant POTENTIAL IMPACTS upon the MEANS and ENDS, and these impacts are classified as either RISKS or REWARDS. Significant POTENTIAL IMPACTS can provide impetus for directives that govern MEANS or support the achievement of ENDS.

The model in Figure 4.24 expands on the BMM by providing a more comprehensive understanding of RISKS and REWARDS. Risk management has become more prevalent in asset management over the last few decades, however the uptake in integrating risk into information systems has been slow. Compared to the literature on "reward" management, the literature on risk management is vast, and hence the approach to describing POTENTIAL IMPACTS was based on describing RISKS. There are numerous models on defining risk, and the risk and reward model developed below is based on several standards including: AS 4360, AIAG FMEA-3, SAE J1739, and MIL-STD-882.

A risk is "the chance of something happening that will have a [negative] impact on objects" [163]. Incorporated into the conceptual data model, a RISK is the probability of occurrence of a HYPOTHETICAL EVENT. The cardinality of RISK to HYPOTHETICAL EVENT is a one-to-one relationship, and different HYPOTHETICAL EVENTS are resultant of different RISKS. The chance or likelihood of a HYPOTHETICAL EVENT occurring is an attribute of the

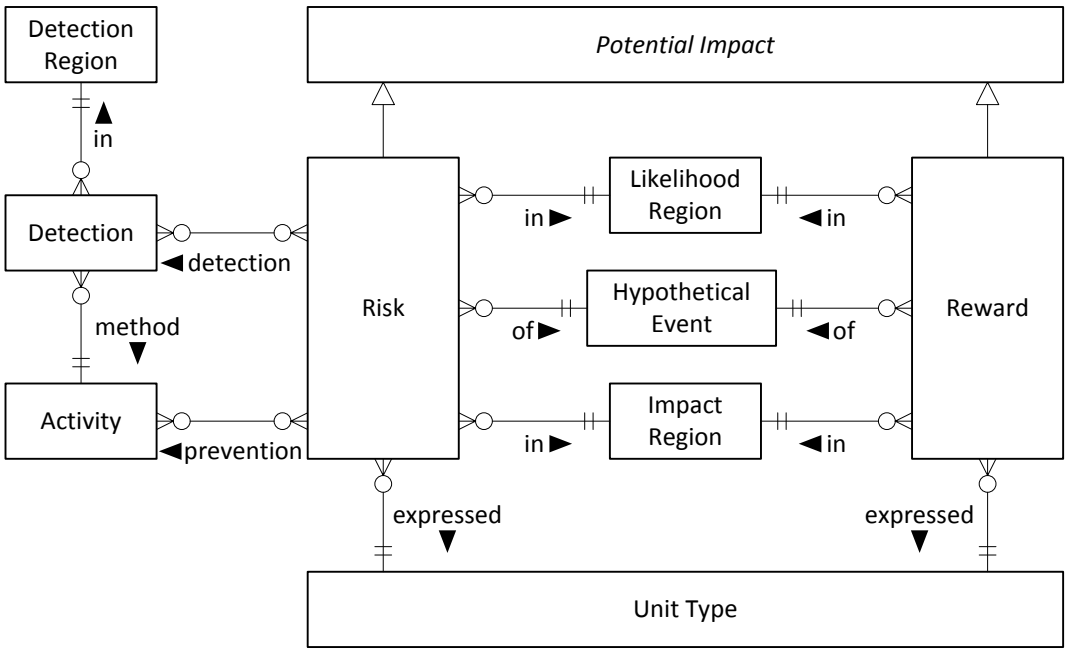


Figure 4.24 – Risks and rewards

RISK entity. The values that it could take are chosen from a LIKELIHOOD REGION, which could contain a categorical scale (e.g. low, medium, and high chance of occurrence), or a numeric scale (e.g. percentage) depending on the business context.

For engineering applications, risk is often quantified through the expression:

$$risk = probability\ of\ an\ incident\ \times severity\ of\ the\ incident$$

in order to compare different risks against each other to prioritise activities [164]. For example, assets with a higher risk of failure may demand a higher level of maintenance attention. Severity is an abstract notion and is defined by the organisation in a similar way to probability, using a categorical or numeric scale for the IMPACT REGION. The term "impact" is used rather than "severity" as the latter conveys the occurrence of a negative event, which is not the desired case with REWARDS.

Failure Mode and Effects Analysis (FMEA) is a risk management technique widely used in reliability engineering. It prioritises failures based on three parameters (probability of occurrence, severity, and probability of detection) to calculate a Risk Priority Number. The first two parameters are the same as the ones described above, while the latter is the likelihood that the failure mode or cause will be detected before it occurs. Hence a RISK is related to zero-to-many detection ACTIVITIES which have a likelihood of detection selected from a DETECTION REGION.

To reduce the likelihood or severity of the RISK, organisations can undertake ACTIVITIES that mitigate or prevent RISKS.

A REWARD is similar to a RISK, as it also measures the likelihood and impact of a **positive** HYPOTHETICAL EVENT. While REWARDS could possibly require detection and encouragement (antithesis of prevention) ACTIVITIES, there is no significant support from neither literature nor industry practice to warrant its inclusion in the model.

4.4.8 Finances

Asset management activities are diverse in that they cover areas generally considered outside the domain of engineering. These include financial, legal, IT, education, and even psychology. Thus consideration must be given in incorporating these elements within the asset management conceptual data model.

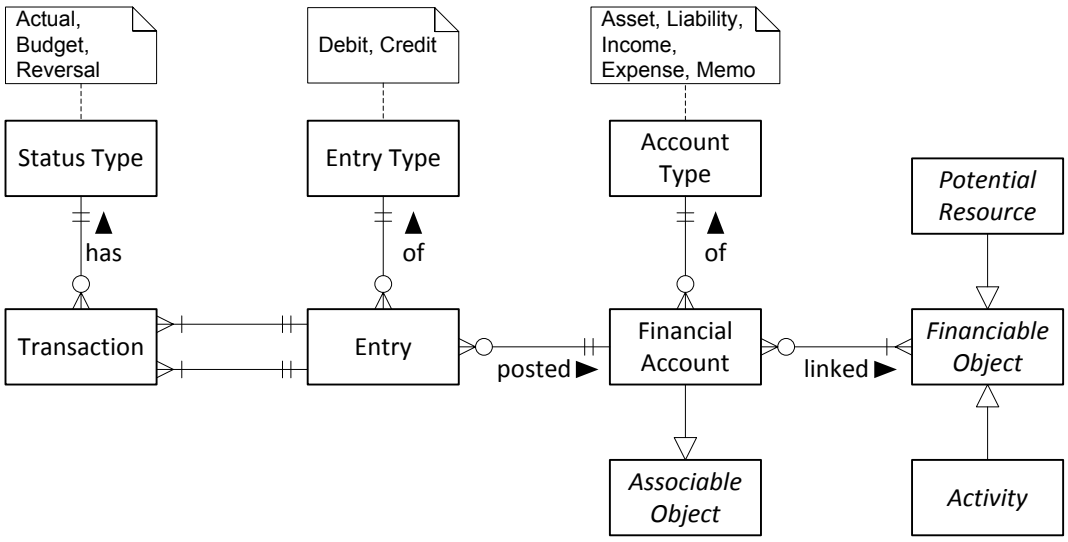


Figure 4.25 – Financial accounts

The financial account model in Figure 4.25 is based on the accounting patterns in literature. Both MIMOSA and ISO 15926 cover cost tracking through attributes of entities. This non-explicit method of modelling is satisfactory for functionality if an implemented system is standalone. However, in an integrated context with other domains, such functionality and usability is lessened for users outside of the engineering domain. That is, these models may not provide sufficient functionality for accounting systems, as the model may be too simplistic.

Each FINANCIAL ACCOUNT is linked to a FINANCIABLE OBJECT, which is either a POTENTIAL RESOURCE or an ACTIVITY (including projects). This relationship provides an indication for the use of the account and at its very least, each FINANCIAL ACCOUNT is connected to an AGENT of type 'enterprise'.

Each FINANCIAL ACCOUNT records TRANSACTIONS (via ENTRIES) related to the FINANCIABLE OBJECT. A FINANCIAL ACCOUNT can be one of several basic kinds - asset, liability, income, expense and memo. The first four are common accounting types, while a memo type is taken from Fowler's [139] financial pattern. It is intended for reporting and indicates an allocation of funds, despite money not actually moving (e.g. allocating income tax to be paid in future). A FINANCIAL ACCOUNT is also an ASSOCIABLE OBJECT such that it can be linked to other accounts.

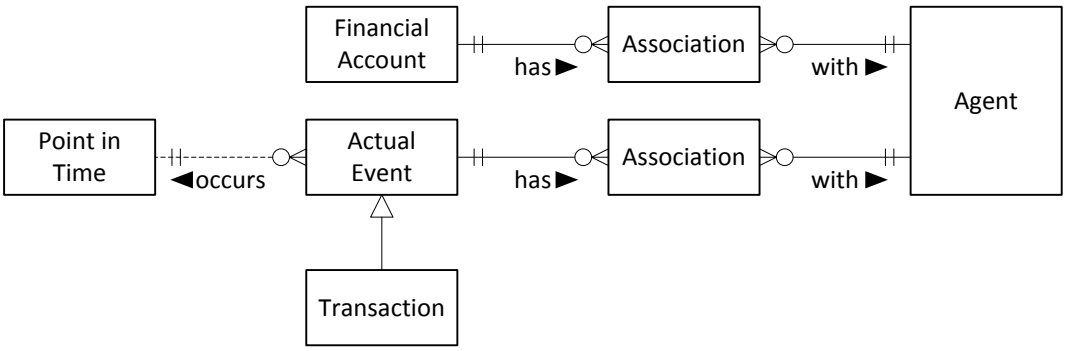


Figure 4.26 – Financial transactions and agents

ENTRIES can either be debited or credited to a FINANCIAL ACCOUNT. Two ENTRIES of each type are required to form a TRANSACTION as per double entry accounting. For example, money spent on purchasing an ASSET would cause a debit in an ASSET account and a credit a cash or liability account depending on the payment option. TRANSACTIONS can be actual amounts, budgeted amounts, or reversal amounts. Reversals are used when previous entries are incorrect, and modifications need to be explicitly shown.

ISO 15489 covers records management and advises that all data have a record of accountability. For example, changes in data should also record the person or system that changed the data values. Thus as shown in Figure 4.26, a FINANCIAL ACCOUNT has an ASSOCIATION with an AGENT to indicate the past/present owners, auditors, and/or supervisors.

A TRANSACTION is a type of an ACTUAL EVENT, and thus shares all the properties of EVENTS. While it could be argued that a TRANSACTION is an ACTIVITY, most financial information systems treat transactions similarly to an EVENT, with transaction times specified at a POINT IN TIME, rather than over a duration as with an ACTIVITY. Thus a TRANSACTION occurs at a POINT IN TIME and has ASSOCIATIONS with AGENTS (e.g. the person who was responsible for the transaction). TRANSACTIONS can consequently have CAUSES (e.g. CONTRACT ITEM) and EFFECTS (e.g. delivery of equipment EVENT).

4.4.9 Contracts

The business of most organisations revolves around agreements or contracts. A contract is a binding set of promises between two or more parties. The most basic and well known contracts are those involved in the purchase and sale of goods. Other types of contracts are rental or lease agreements, permits, insurance policies, and warranties.

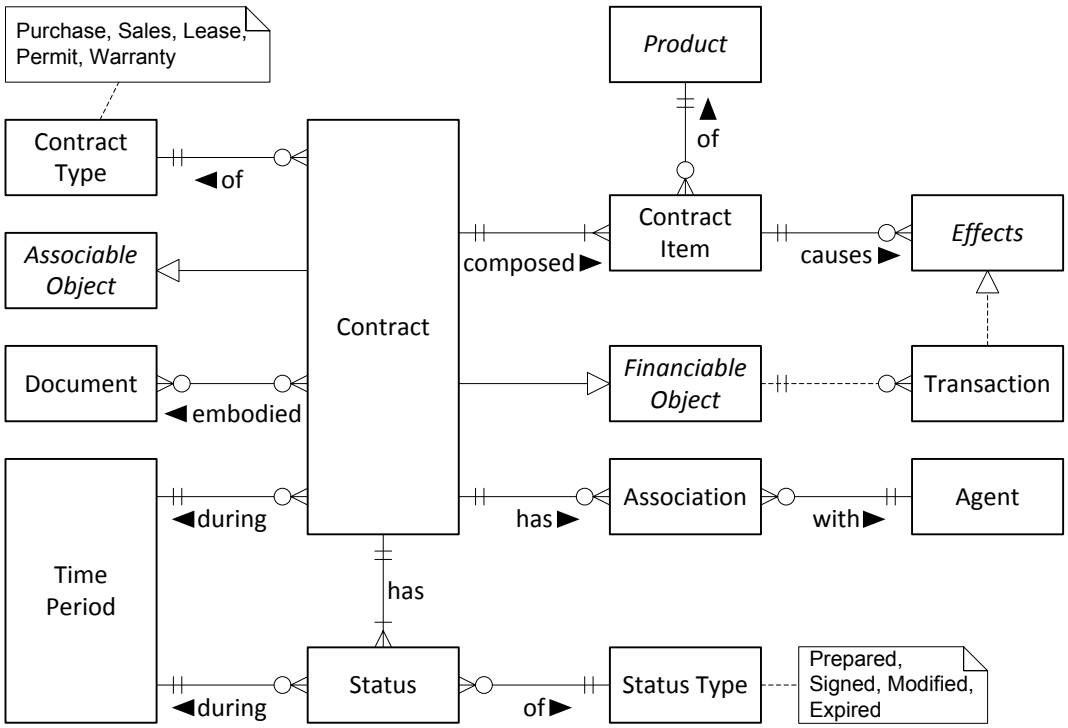


Figure 4.27 – Contracts

While contracts are generally considered to be a procurement or legal area rather than engineering, the implications of contracts upon asset management are undeniable: assets are purchased and sold through contracts; external maintenance or condition monitoring contractors require contracts; and property is leased through contracts.

The model in Figure 4.27 is a generic contract model that can take on multiple types. As an ASSOCIABLE OBJECT, reflexive associations permit sub-contracts or references to other CONTRACTS (e.g. new contracts based upon old). The association pattern is used between CONTRACTS and AGENTS such that different CONTRACT TYPES have bearing on the ASSOCIATION TYPE. Thus with a purchase contract, only buyer and seller associations would be allowed, while with a property rental contract, only landlord and tenant associations would be allowed. A CONTRACT can also be embodied in zero-to-many DOCUMENTS.

As with activities, a CONTRACT also has a lifespan in which it can be accorded different STATUSES. These STATUSES are both predictive in that they can indicate the terms of the CONTRACT, and descriptive in that they can indicate modifications to the CONTRACT. Thus periods of validity can be set up before the CONTRACT is effective, and the CONTRACT can be labelled as void if a breach has been identified.

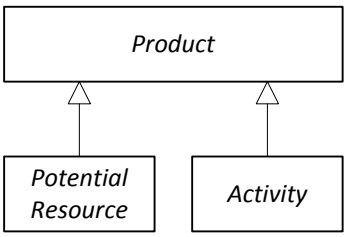


Figure 4.28 – Products

A CONTRACT is composed of one or more CONTRACT ITEMS, which form the basic terms of the CONTRACT. Each CONTRACT ITEM is of a PRODUCT, which are either goods (POTENTIAL RESOURCES) or services (ACTIVITIES) as shown in Figure 4.28. As PRODUCTS encompass two abstract items, they can take on many forms from MODEL TYPES, to SEGMENTS, to MEASUREMENT activities. A CONTRACT ITEM can be the instigator of EFFECTS. For example, an ASSET purchased through a CONTRACT will cause delivery EVENTS to be registered or scheduling of ACTIVITIES for receipt of the item.

CONTRACTS are tied into the FINANCIAL ACCOUNT pattern similar to ACTIVITIES: each CONTRACT can be associated with multiple FINANCIAL ACCOUNTS, and each CONTRACT ITEM causes TRANSACTIONS to be posted to a FINANCIAL ACCOUNT. Thus the price of an ASSET purchased would cause one TRANSACTION, while the shipping charges of the ACTIVITY would cause another TRANSACTION. Separating financial transaction data by CONTRACT ITEMS gives a finer granularity than using the aggregated cost where such detail is lost.

Insurance and Warranties

Insurance and warranties are specialised contracts that are commonly used with ASSETS and SEGMENTS. Insurance allows a party to transfer the risk of a loss to another party in exchange for a premium. Warranties are an assurance by a party to another party that contract terms will be met, and if not, compensation will be applicable.

Both INSURANCE and WARRANTIES in Figure 4.29 are a type of CLAIMABLE CONTRACT, which is a type of CONTRACT. Therefore, they inherit reflexive ASSOCIATIONS, ASSOCIATIONS with AGENTS, STATUSES, FINANCIAL ACCOUNTS, and CONTRACT ITEMS. WARRANTIES are of a particular WARRANTY TYPE, and could be a supplier provided warranty, or a customer purchased warranty. Both WARRANTIES and INSURANCE are taken out to guard against RISKS.

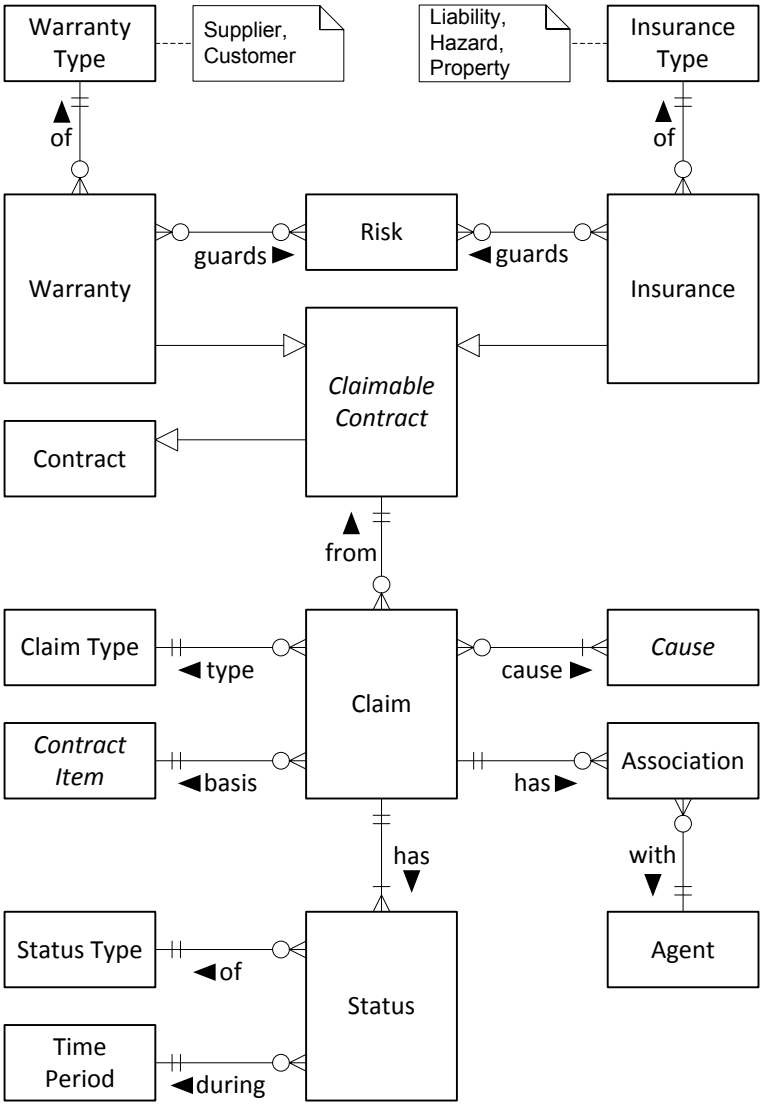


Figure 4.29 – Insurance and warranties

CLAIMS made on claimable contracts are based on one CONTRACT ITEM. Thus multiple CLAIMS are required if many CONTRACT ITEMS are involved. Each CLAIM has one or more CAUSES which form the reason for the CLAIM. AGENTS are ASSOCIATED with each CLAIM in order to indicate the parties on both sides of the CONTRACT involved in processing the CLAIM. CLAIMS also have STATUSES that indicate when they are started, are in progress, or closed.

4.4.10 Units of Measurement

Units of measurement are crucial in providing standardised measures of quantities. The model shown in Figure 4.30 uses the OSA-EAI method of converting between UNIT TYPES via a REFERENCE UNIT. For example, to convert between the UNIT TYPE of Degrees

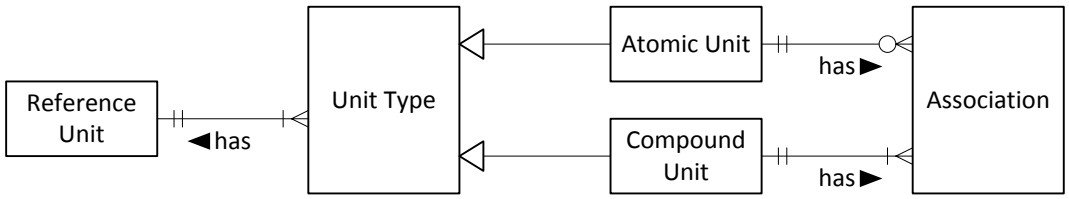


Figure 4.30 – Units

Fahrenheit to Degrees Celsius, the temperature is first converted to Kelvin as the REFERENCE UNIT, then to the target UNIT TYPE. As a REFERENCE UNIT is used, the number of conversion equations between analogous UNIT TYPES is significantly decreased.

A limitation with the OSA-EAI model is that compound units, such as the unit type for speed (distance per time), have no relation to their atomic components, thus limiting the types of conversion that can take place. Converting from meters per second to meters per hour is not possible without adding a new unit type, despite the existence of constituent elements (meters, seconds, and hours). The OSA-EAI model can be improved upon by using Fowler's [139] compound unit pattern, by partitioning each UNIT TYPE into either an ATOMIC UNIT or a COMPOUND UNIT. Through the unit ASSOCIATION, a COMPOUND UNIT is comprised of multiple ATOMIC UNITS. The ASSOCIATION also specifies the mathematical relation of the ATOMIC UNIT to the COMPOUND UNIT. For example, the ATOMIC UNIT meters is the base, the ATOMIC UNIT seconds is divided, and another ATOMIC UNIT seconds is divided, with the resultant COMPOUND UNIT becoming meters per second^2.

4.4.11 Measurements

A measurement is a quantifying observation of objects or activities. In asset management, a large portion of activities involves measurements of assets or locations. These can comprise of both manual and non-manual (computerised) measurements. Measurements include monitoring of asset operations and condition, segment environmental conditions, health assessments by personnel, and stocktaking of inventory. While measurements are often associated with lower level asset management operations, their use in higher level asset management is often through the use of KPIs.

Measurements

A MEASUREMENT is a type of ACTIVITY, and hence it inherits all the attributes of an ACTIVITY. Figure 4.31 shows a MEASUREMENT inheriting from both PROPOSED AND ACTUAL

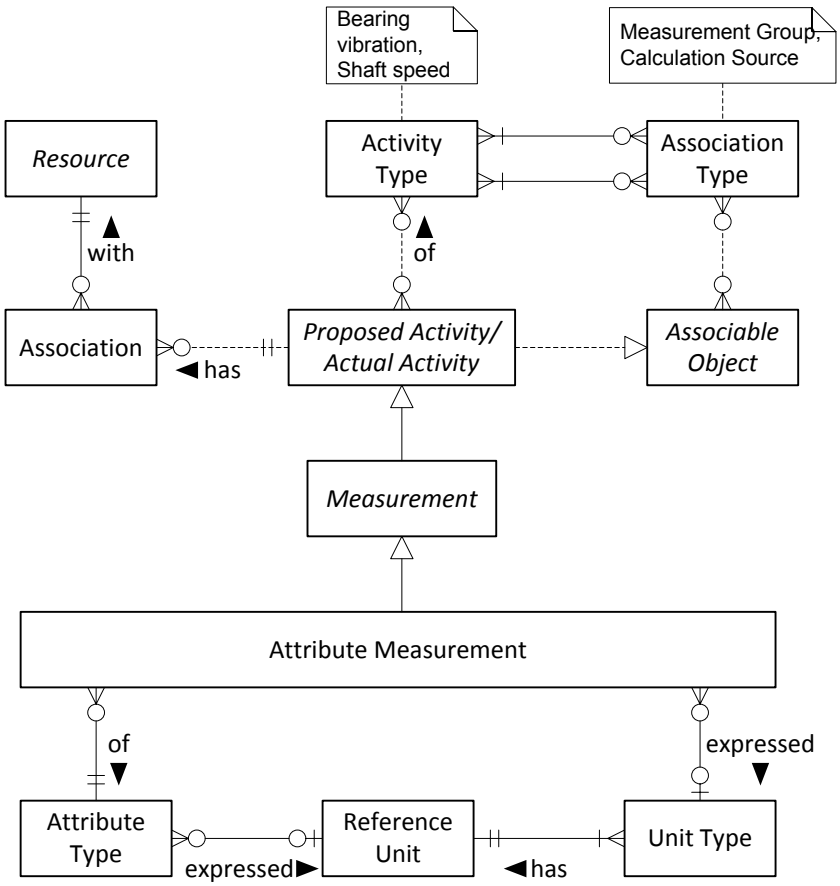


Figure 4.31 – Measurements

ACTIVITIES, as MEASUREMENTS can be of either type. Because of the inheritance, a MEASUREMENT has STATUSES, FINANCIAL ACCOUNTS, RISKS, and ACTUAL RESOURCES attached. It can also be the CAUSE or EFFECT of other ACTIVITIES or EVENTS, and can be based on HYPOTHETICAL ACTIVITIES.

RESOURCES are a vital element within MEASUREMENTS and the type of ASSOCIATION indicates the role of each RESOURCE. For an ASSET that is being measured, there will be additional ASSETS used to conduct the MEASUREMENT, AGENTS who perform the MEASUREMENT, and the SEGMENTS on which the MEASUREMENT occurs.

The (indirect) reflexive ASSOCIATION between MEASUREMENTS allows grouping of measurements (e.g. simultaneous readings from multiple sensors on an asset) as well as indicating which arguments are used in a calculation MEASUREMENT. Both the measurement patterns by Hay [138] and Fowler [139] consider calculations to be a type of MEASUREMENT and this view is shared by the conceptual data model. Calculations

process source arguments (i.e. other MEASUREMENTS) through a function (i.e. HYPOTHETICAL ACTIVITY) to produce a calculated MEASUREMENT. An ordering number field in the MEASUREMENT ASSOCIATION can be used to specify the order of parameters in the calculation.

Hay [138] uses the term *variable type*, Fowler [139] uses *phenomenon type*, and MIMOSA uses *measurement location type* for what is termed ATTRIBUTE TYPE. This term is used to remain consistent with ASSET and SEGMENT ATTRIBUTES introduced in Sections 4.4.1 and 4.4.3. Each MEASUREMENT in the model is of one ATTRIBUTE TYPE whose values include speed, temperature, and weight. At the knowledge level, each ATTRIBUTE TYPE is expressed in a REFERENCE UNIT (e.g. metres per second, Kelvin, or kilogram) while at the operational level, each MEASUREMENT is accorded with a UNIT TYPE of the corresponding REFERENCE TYPE (e.g. kilometres per hour, Celsius, or pound). As some ATTRIBUTE TYPES may not have REFERENCE TYPES or UNIT TYPES (e.g. colour), the lowest cardinality is zero.

The measurement model covers measurement events, sample tests, and health assessments within the OSA-EAI as each of these are considered types of measurements. The OSA-EAI includes data quality types and confidence percentages within the measurement location tables, and change pattern types and likelihood probability for the health assessment tables. Data quality type, confidence percentage, and likelihood probability are aspects of the data itself, rather than topical asset management data, and are discussed within the metadata section in Section 4.2.3. Storing a change pattern type appears to violate normalisation rules in that it can be computed from two MEASUREMENTS (although its inclusion would be justified for the first recorded health assessment of an object).

Regions

Without reference information, a measurement is meaningless. For example, a measured operating temperature of 60°C carries no weight until it is compared against the knowledge of a normal operating temperature of 40°C. To record reference knowledge, a range or region of values is used to partition the attribute space. Fowler [139] uses a similar concept of a range of values for a phenomenon (i.e. an ATTRIBUTE), while the OSA-EAI only uses a range to define alarm regions for signals.

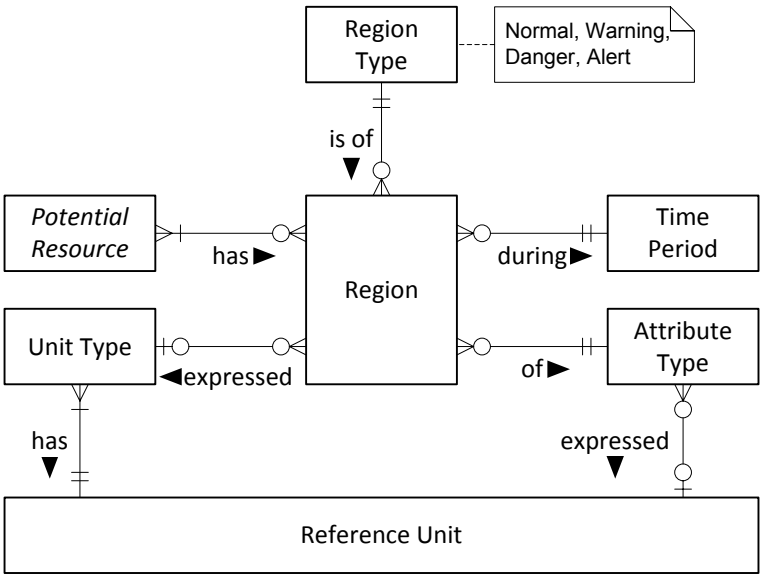


Figure 4.32 – Resource regions

As shown in Figure 4.32, a REGION categorises a range of values of an ATTRIBUTE TYPE for a POTENTIAL RESOURCE into a particular REGION TYPE. For instance, the speed ATTRIBUTE TYPE for a motor RESOURCE can have 0-1000 RPM categorised as a normal operating REGION TYPE, 1000-1250 RPM categorised as warning, and greater than 1250 RPM categorised as an alert. This also applies for inventory ASSETS - bearing lubricant stored on a SEGMENT can have greater than 3 litres as normal operating level, 1 to 2 litres as warning level, and less than 1 litre as an alert which could trigger a procurement ACTIVITY. Some REGIONS can change over time, such as asset wear causing tolerances to increase/decrease, and hence a TIME PERIOD is linked to the REGION.

Measured Attributes

All recorded attributes of an object are observed through measurements. For example, the dimensions and weight of an asset must first be measured before it can be recorded as an attribute of the asset. While many attributes require resources and time to measure, some attributes are trivial to measure and only require human judgement (e.g. colour, shape, liquid/solid/gas states). Despite its triviality, a measurement has still occurred in the strictest definition. Thus to promote consistency and simplicity, all ATTRIBUTES of ASSETS, SEGMENTS, AGENTS, and ACTIVITIES are the result of MEASUREMENTS.

The model shown in Figure 4.33 enhances the ATTRIBUTABLE OBJECT pattern from Section 4.4.2 by including the corresponding MEASUREMENT for an ATTRIBUTE. The advantage of this method is that a MEASUREMENT provides the context to how the

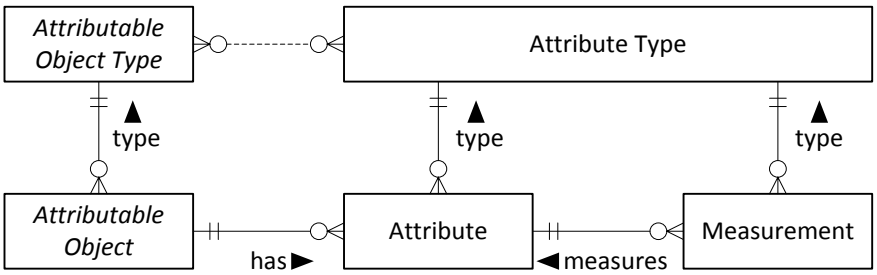


Figure 4.33 – Object measured attributes

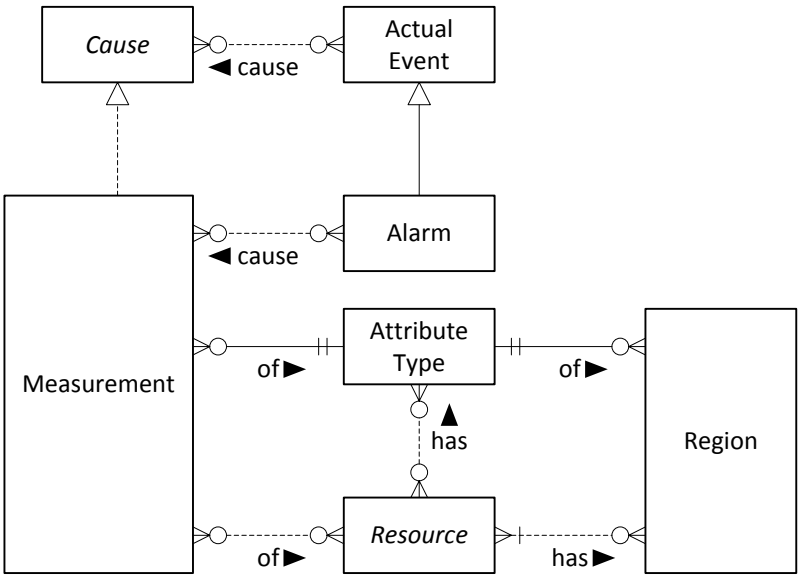


Figure 4.34 – Alarms

ATTRIBUTE was obtained (similar to the scope pattern by Fowler [139]). Thus the measurement of the maximum load on a motor can be associated with contextual temperature and power measurements.

Alarms

During the operation of a process, an incongruity may lead to the failure of the desired process outcome. This could be due to a variety of reasons including asset failure, poor software exception handling, human error, etc. Process systems and condition monitoring systems usually alert users when various thresholds are crossed such that action can be taken.

As an alert or alarm occurs at a particular point in time (i.e. the point when a threshold is crossed), Figure 4.34 shows that an ALARM is a type of ACTUAL EVENT. While "proposed

alarms" do exist in real life (e.g. testing alarm equipment), these are not ALARMS based on an aberrance in operation, but are in fact, MEASUREMENTS of alarm equipment. Through inheritance, ALARMS have different EVENT TYPES, CAUSE and EFFECTS, ASSOCIATIONS with RESOURCES, and HYPOTHETICAL EVENTS.

As an ALARM is an EVENT, and an EVENT has a CAUSE, in the case of process and condition monitoring systems, that CAUSE is a MEASUREMENT. A MEASUREMENT is conducted on a RESOURCE, and if the MEASUREMENT is within an applicable REGION, an ALARM is generated. Thus the relationship to the corresponding MEASUREMENT provides evidentiary substantiation. ALARMS can also be manually triggered (e.g. a person pressing an emergency button), or triggered by another EVENT (e.g. a person opening a monitored fire escape) and hence the relationship to CAUSE rather than directly to MEASUREMENT.

4.4.12 Documents

The universal communication mechanism of organisations, documents serve as a means to disseminate information. Technology has taken documents from stone tablets to papyrus to paper to electronic formats. Modern organisations are usually burdened by the latter two forms which can be used to describe a plethora of different document types – lists, logs, drawings, manuals, certifications, reports, articles, books, forms, mail, etc. Hay [138] notes that the definition of documents should not appear within data models as documents are simply another manifestation of the data contained within

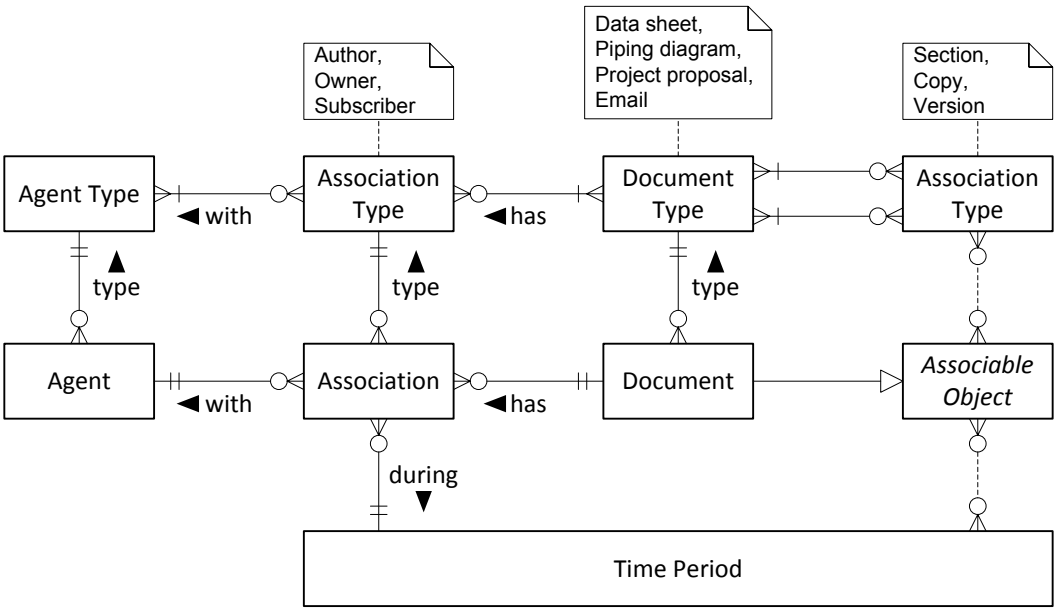


Figure 4.35 – Document association structure and agent associations

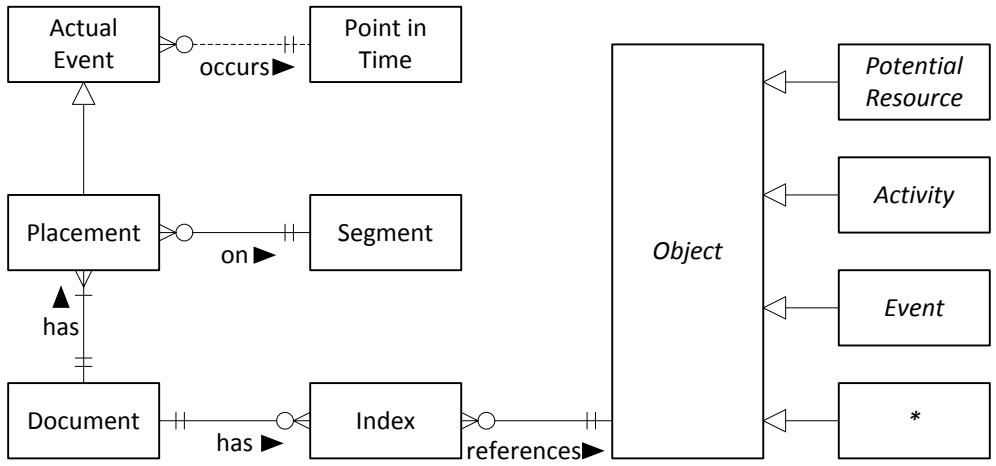


Figure 4.36 – Document location and indexes

the said data models. However, it is important to illustrate the importance of relationships between documents and other data model entities – particularly as documents are a vital ingredient in organisations.

From its data model structure, a DOCUMENT is similar to the ASSET/SEGMENT/AGENT core entities. As shown in Figure 4.35, DOCUMENTS have reflexive ASSOCIATIONS to allow for sub-documents, grouping of documents, copies, and versioning. These different ASSOCIATION TYPES are linked to DOCUMENT TYPES at the knowledge level. The association pattern is used to relate DOCUMENTS and AGENTS to indicate authorship, ownership, or subscriptions. Both ASSOCIATIONS have a TIME PERIOD applicability.

DOCUMENTS can also have a PLACEMENT on a SEGMENT, similar to an ASSET as shown in Section 4.4.3. PLACEMENT is an ACTUAL EVENT which links to a SEGMENT. For intangible electronic DOCUMENT TYPES, the SEGMENT would describe the computer system in which the DOCUMENT is located.

Hay [138] indicates that a document can make references to items within and outside an organisation. Examples include a pay slip that refers to a person and their account, a data sheet that lists attributes of an asset, or a report that outlines the daily activities of a plant technician. There are no limitations on the topics a document can cover, and hence a DOCUMENT can have an INDEX to any OBJECT (shown in Figure 4.36). As every entity is a type of OBJECT (with the asterisk entity being an attempt to indicate as such), a DOCUMENT can be related to a POTENTIAL RESOURCE (e.g. ASSET, AGENT, and SEGMENT), ACTIVITIES, or EVENTS as well as FINANCIAL ACCOUNTS, CONTRACTS, REGIONS, ALARMS, etc.

While the association pattern could be used to relate multiple DOCUMENTS to multiple OBJECTS, a more specific and meaningful association (i.e. INDEX) is used. No applicability time period is placed on the INDEX (as with ASSOCIATIONS) as changes to the referenced object typically results in a new version of the document. For example, the introduction of a new type of ASSET to a plant would (or should) trigger the review and revision of maintenance procedure DOCUMENTS.

The reflexive DOCUMENT ASSOCIATION with ASSOCIATION TYPES, and INDEXES to relevant OBJECTS allows for a flexible document structure. While the text of a document is not likely to be stored in a relational database but instead as a file (e.g. Microsoft Word or PDF), the structure shows how asset management objects can be referenced by documents.

4.5 Experimental Testing

Verification and validation are two stages within software testing to ensure quality. Verification ensures that a system meets the specified requirements while validation is to demonstrate that a system fulfils its intended use when placed in its intended environment [165]. As summarised by Balci [166], verification is about building the model right, while validation is about building the right model.

4.5.1 Verification

Verification was conducted in the form of expert reviews and compatibility with existing work.

Expert Reviews

During the development of the data model, the model would be periodically reviewed by experts within the field of asset management. As asset management is an expansive discipline, it is virtually impossible to find an expert that is knowledgeable in all areas, and thus a variety of experts were engaged. The people who provided input into the model via interviews (see Section 4.2.2) were also asked to analyse both the modelling process and the model itself, enabling a cyclical feedback loop. More than 20 people from these organisations reviewed the model for accuracy and coverage.

These reviews were conducted by presenting firstly a justification of asset management conceptual data modelling; secondly a background on data modelling; thirdly the methodology used in this research; and finally the models themselves. While Microsoft

PowerPoint was used to present the first three items, the modelling tool, Microsoft Visio, was used to present the model. The latter was used as (1) it was difficult presenting readable models that fit on a single presentation slide and (2) changes to the model could be made in real-time.

As the conceptual data model was presented at the same time as being reviewed, a concern was that each expert would not have enough time to examine the model. Due to confidentiality concerns at the time of review, the model could not be given out to reviewers, and all reviews needed to be conducted in the presence of researchers. These sessions lasted one to two hours, with the variation resulting from the amount and type of questions asked.

Compatibility with Existing Work

Section 4.1 showed that whilst there were no equivalent existing models, there were three standards as well as certain information systems that were relevant for asset management conceptual data modelling. These items were used in the verification process by comparing the corresponding areas of this research with the other models. The objects within the conceptual data model were examined to see if they were ideologically compatible with the objects within the compared model. The term *ideologically compatible* is used because a simple one-to-one comparison would produce erroneous results (e.g. the site vs. segment argument in Section 4.4.3), and thus the intent of each object needed to be compared.

As these standards and information systems were used as inputs into developing the conceptual data model (Section 4.2.2), it could be asserted that they cannot be used to verify the conceptual data model. An analogous case is using the same training data as testing data within a pattern classification exercise resulting in an artificially-boosted classification accuracy. The practice of reusing inputs for testing is acceptable in a data model scenario because (1) the process of developing the model (i.e. training) is considerably more complex and time consuming, and (2) there are not many models available for a verification comparison.

4.5.2 Validation

Validation was conducted in the form of both practical and theoretical usability experiments. Four case studies were conducted, each at differing levels for software development. Case Study 1 provides an implementation case of a section of the model;

Case Studies 2 and 3 provide comparisons against two detailed asset management software specifications; and Case Study 3 provides a comparison against a high-level asset management framework.

Case Study 1

Due to the overwhelming nature of the number of asset management areas covered by the conceptual data model, it could not be implemented in its entirety for validation. A complete implementation would require an unattainable amount of resources, and hence a different approach needed to be taken. A partial implementation was conducted in conjunction with the CIEAM research project ID201.

The research project entitled "Integrated Decision Support System for Asset Management in the Water Utility Industry", looked at developing an asset management health system called BUDS (Bottom Up Decision Support). The goal of BUDS was using condition monitoring and reliability data in order to aid in asset renewal decision making. Thus sensor data from condition monitoring and SCADA systems were stored in the database from which diagnoses and prognoses were calculated. The system also stored failure modes, RDBs, and fault trees for reliability prediction calculations. Maintenance cost data was also stored and used with the condition monitoring and reliability predictions to translate the figures into a format that managers could understand – costs and dates (as seen in Figure 4.37).

The software was developed in Microsoft Visual Studio .NET 2003, using C# as the programming language. A relational DBMS (Microsoft SQL Server 2000) and flat files were both used as the storage mechanism as it has been shown that each have their own merits [167]. The conceptual data model was translated to a logical and physical data model for the sections stored in the database. For the objects stored in the relational database, a Structured Query Language (SQL) program was created that would create tables and insert data.

As data was gathered from a variety of systems (ERP/CMMS, condition monitoring, SCADA, Microsoft Access database, and Microsoft Excel spreadsheets), a semi-automated ETL process was developed for each source. In addition to extracting the data from the sources for loading into the data warehouse, a data cleansing process was added for selected sources.

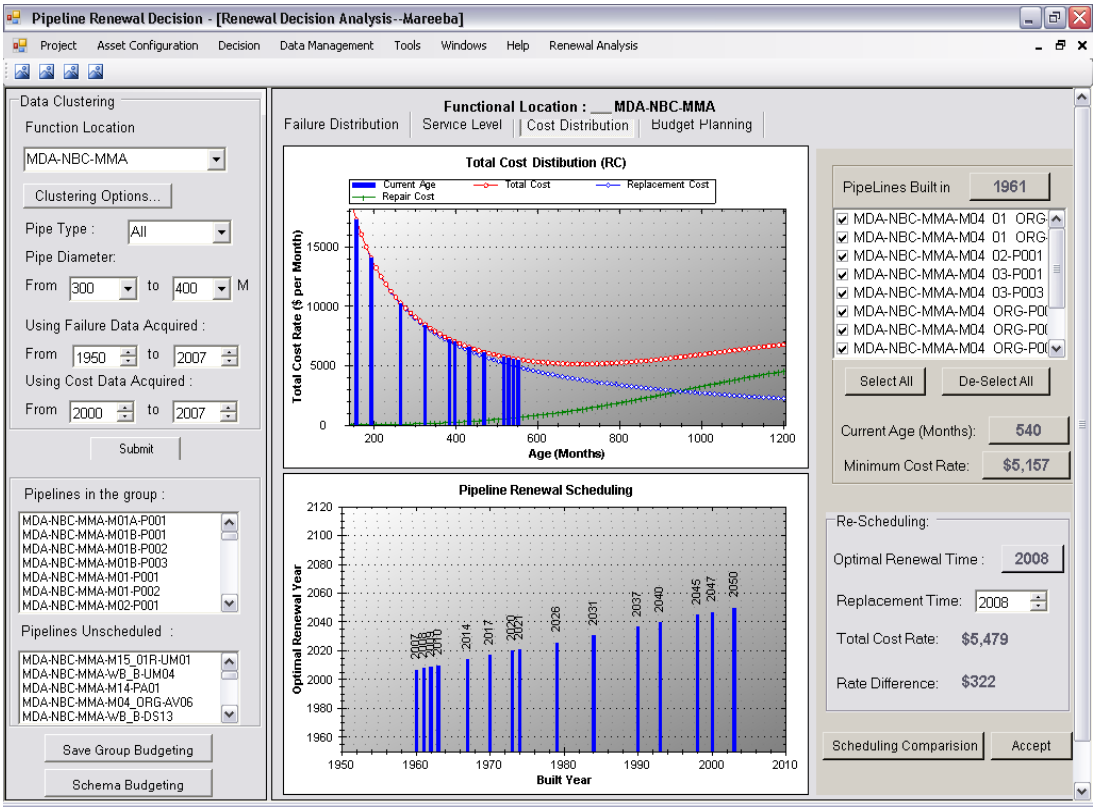


Figure 4.37 – BUDS software screenshot

As the ETL process integrated data from several source systems, analysis could be conducted not only with the aforementioned program, but with any tool that had a database interface (e.g. ODBC). Thus tools such as Microsoft Excel and MATLAB were used for report generation in areas that BUDS did not have the required functionality.

Case Study 2

A subsequent CIEAM research project, ID206, looked at extending the functionality of BUDS to cover additional asset management areas. As the project is currently in progress, only the software requirements were validated against the conceptual data model.

The requirement specifications in ID206 extended those of ID201 to include functionality in the areas of asset operation, inventory, safety procedures, project management, contracts, and asset performance. The specifications itemised specific data requirements and described the required functionality for each area.

A comparison between the specifications and the conceptual data model is shown in Appendix D. For each area in the specifications (indicated by bold text), sub-areas are matched with the equivalent area in the conceptual data model. Only leaf nodes in the specifications are matched, as the parents of leaf nodes only provide a grouping of functions, rather than the functions themselves.

The comparison shows that the conceptual data model adequately covers each area in the specification. As is the nature of research, the specifications are in constant flux and some sections are incomplete, and for these areas, a less detailed comparison can only be made. There are also several instances where multiple data areas in the conceptual data model are referenced. The analysis shows that certain areas are referenced more than others, and this is determined by the research direction and purpose of the software.

Case Study 3

The specifications for an asset management system under development by World In One Technology Sdn Bhd in Malaysia were also compared to the conceptual data model. The system forms the base module for a future system implementation of change management, maintenance management, and fleet management. It currently covers areas in asset registry, configuration, failure, movement tracking, value and depreciation, warranty tracking, measurements and reporting, and performance and reliability measures.

A comparison between the specification and the conceptual data model is shown in Appendix D. As with Case Study 2, only the leaf nodes in the specification are addressed.

The comparison for this case study shows that the conceptual data model covers each area in the specification. While metadata was not modelled, it was addressed by this research, and this is used with the specification area of user permissions. The areas cover similar areas to Case Study 2, with less emphasis on condition monitoring measurements, and more emphasis on metering. Warranties are addressed through the contract model, and currency conversion is addressed through the units of measurement model (using the same method as the OSA-EAI).

Framework area	Relation to conceptual data model
Strategic Planning	Motivation
Asset Ownership	Documents, Motivation
Risk Management	Motivation, Risk
Budgeting & Costing	Finances
Data Management	Assets, Segments
Condition Monitoring	Measurements
Tactical Planning	Activities
Human Resources	Agents
Assets Usage Life Cycle	Activities
Performance Measure	Measurement
Information Systems	N/A
Financial Management	Finances

Table 4.5 – Comparison to the CIEAM Asset Management Framework

Case Study 4

The comparison process was also applied to the CIEAM Asset Management Framework (shown in Table 4.5). While not a software specification, the framework provides guidance on the broader functions within asset management, and through an identification of the data needs of these functions and sub-functions, a comparison could be drawn.

The functions described by the framework are at a higher level than the two previous case studies, and hence only the top level areas in the conceptual data model are compared. In addition, as the framework areas are broad categories, only the primary conceptual data model areas are listed. For example, Documents and Motivation are the primary data areas for Asset Ownership, but in reality, data in the Assets, Segments, Agents, and Activities models would also be used for functions in this framework module.

The Information Systems framework module is listed with a N/A as the sub-functions relate to the creation of information systems, rather than the use of information systems (as is the case with the other eleven areas).

4.6 Innovation

1. A comprehensive and holistic modelling methodology

The methodology presented in this chapter for data modelling examines a comprehensive amount of data sources to derive the eventual conceptual data model. Starting with a holistic view of asset management, rigorous research was conducted before modelling each area. The unique systematic approach distinguishes this research from a rudimentary application of data modelling.

Existing asset management models are born out of sub-disciplines, such as asset codification, operation, or condition monitoring, and as such, provide meticulously detailed models for these sub-disciplines and less detail for other areas. While the same can be said of the origins of this research, the effect is lessened due to the aggregation of data sources.

Most existing models employ a bottom-up approach to model development, concentrating on a section of asset management, and then adding to the existing model as new functionality is required. As these models will attempt to maintain backward-compatibility, the new additions are constrained by certain characteristics of the existing model (e.g. naming schemes, types, and pattern structures). This research employs a top-down approach to modelling, and considers each element within the framework of asset management. Such consideration avoids the constraints enforced by the bottom-up approach and provides for a better structured model.

2. Based on data model patterns literature

The existing asset management data models are typically experiential and are built from scratch. While starting from zero is an advantage as it provides maximum flexibility, the methodology loses the advantages of building upon others' past experiences. One area of experience that has not been captured in previous asset management data models is the use of patterns literature. Patterns literature looks at data modelling from a discipline-neutral perspective. Thus the commonalities in business functions (e.g. asset management, financial management, human resources management, and customer relationship management) of an organisation are all handled similarly. This naturally leads to a more generic model to express a variety of scenarios, and also leads to a more integrated organisational model. The integrative

and generic characteristics from patterns are captured within the conceptual data model.

3. Consideration of data warehousing characteristics

As explained in Section 1.2, a data warehouse is commonly defined by four characteristics – subject-oriented, integrated, time-variant, and non-volatile [3]. All previous data models have been developed with OLTP as the end goal, while the end goal of this research is for data warehousing. As such, the first three[1] characteristics have been built into the model: subject-oriented by focussing on the natural partitions in asset management data; integrated by considering intra- and inter-relations between and to asset management data; and time-variant by judiciously employing temporal elements where possible.

4.7 Significance

1. A benchmark for the diversity and depth of modelling inputs

The innovative methodology sets a benchmark for data modelling not just within the field of asset management, but for other domains. While many data modelling initiatives use external sources as references, this research has not come across any work that thoroughly researches each area under investigation with the depth presented in this work. The presented methodology and subsequent models set a target for future efforts in data modelling.

2. Increased coverage of asset management areas

By virtue of integrating numerous smaller models, the conceptual data model provides a greater coverage of asset management data. Due to the generic approach in conceptual data modelling, and due to the pervasive use of EAV structures and generalisation/inheritance, an expansive area could be covered. The generic approach considerably increases the descriptive flexibility of the model.

3. A model for asset management system development

The foremost purpose of the conceptual asset management data model was to provide a basis upon which an integrated asset management software platform can be derived.

[1] The fourth data warehouse characteristic is not influenced by the underlying data model, but instead, the prescribed maintenance policies.

While the intended software platform involved data warehousing, the model is not only restricted to this domain, and can be used for OLTP database design, software design, and enterprise integration. As data analysis and data management are synergists, the integrated data platform presented by this research will support the development of advanced techniques in addition to unique data mining strategies.

4. A generic reference model for deriving enterprise-specific models

The conceptual data model is based on common enterprise data model patterns and can be used as a reference model. Reference models are usually construed as "good practice" models that contain a wealth of knowledge and experience. Organisations who are seeking to integrate existing asset management systems or develop new asset management systems (either end users or providers) can use the model to derive an organisation-specific or corporate data model. The conceptual data model then serves as a set of principles on the composition and relationship between asset management data areas.

5. A model for understanding asset management data areas and relationships

The conceptual data model serves as a mechanism to understand the varieties of asset management data. From an IT perspective, it enables users to understand the classes of engineering data. From an engineering perspective, it aids users in understanding IT concepts of object and relational forms. The model effectively attempts to bridge these two different fields to harmonise the understanding of asset management data.

4.8 Conclusion

The importance of data modelling is evident in system development, as it has far reaching consequences on the system design. Within a data warehousing context, data models influence low level factors such as storage space and performance speed, to higher level factors such as the types of analysis that can be conducted. With an increasing number of systems and data areas, asset management organisations are seeking to integrate these areas for advanced data warehouse-styled reporting. An understanding of asset management data needs to be in place before integration can occur, and this chapter has attempted to provide one stage in developing this understanding.

The conceptual data model developed by this research provides an integrated view of asset management data. By examining data model patterns, standards, information

systems, business process models, analysis methods, and conducting interviews, a comprehensive model was developed. Relational, object, and literate modelling were combined to harness the benefits of each modelling language, provide a systematic approach, and to present a more intuitive understanding of the data model.

The types of data available within asset management systems are diverse, and the primary areas identified by this research include: Asset, Segment, Agent, Activities, Events, Motivation, Finances, Contracts, Units, Measurements, and Documents. Each of these areas have unique structural elements, and both associative and attributable patterns are clearly evident in their structure. The conceptual data model presents the integration between each area and clearly highlights the interrelationships within asset management data. The models were verified through expert reviews and validated against four case studies that ranged from a low-level asset management software implementation to a comparison against a high-level asset management framework.

5 Asset Management Multidimensional Model Evaluation

For many years, organisations have deployed OLTP systems to automate and record their business activities. The challenges presented by these systems for business analysis led to the development of data warehousing, which focused on providing an environment suitable for business intelligence (BI) and DSS. The advantages of data warehousing over OLTP systems accrue from both its methodology as well as its differences in technology. One of the technological facilitators of data warehousing is found in the approach to data modelling. Multidimensional models present a new method to structure data, and promise benefits of speed as well simplified access for ad-hoc reporting. As these characteristics meld well with the ideology of data warehousing, many data warehouse implementations use denormalised multidimensional models compared to equivalent relational models. There is also an abundance of OLAP tools devoted to BI that operate on multidimensional models.

Past comparisons between relational and multidimensional models have investigated query speed, storage space, and comprehensibility [168, 169]. However, these embed analysis logic within the data retrieval code (i.e. mixing the view and controller in a Model-View-Controller architecture). While this is fine for most business analysis, asset management is different in that data analysis logic often relies on complex algorithms that cannot be represented through SQL and other declarative languages, and often require separate applications or platforms for execution.

The current generation of asset management systems are built upon both relational and flat structures. To test if the same benefits are applicable in asset management, this chapter compares the suitability of multidimensional models over relational models for asset management data warehousing. To do so, physical or at least logical models are required for the evaluation. As the conceptual data model in the previous chapter has not been translated to a logical model, the multidimensional schemas are based on the ER models found within the MIMOSA OSA-EAI version 3.0f. The primary reason of using the MIMOSA OSA-EAI is because it is a standardised model that covers a broad area of asset management data.

5.1 Background Theory

5.1.1 Data Warehouse Schema Modelling

A schema is a description of the structure and rules of an object. In the data management sense of the word, it is the model that defines the data objects, their attributes, their relationships, and rules. There are several methodologies of arranging schema objects, with the most prevalent being the third normal form. Third normal form (3NF) modelling is the classical approach to relational database design whereby data redundancy is minimised through normalisation. Because of normalisation, 3NF schemas typically have a larger number of tables compared to multidimensional schemas.

There are two primary schema methodologies of representing multidimensional data, either as a star or snowflake schema. In a star schema, the data are stored in a central fact table, and surrounded by one or more denormalised dimension tables. It is named as such because of this central structure with radiating points. A snowflake schema is similar to a star schema but allows for normalisation in dimensional tables to remove redundancy, and hence dimension tables can be associated with other dimension tables. Compared to a 3NF schema, multidimensional schemas are highly denormalised. Because of the decrease in complexity due to denormalisation, multidimensional schemas can be more intuitive to non-technical end users who are more familiar with logical entities rather than entities and relationships. They can also provide optimised performance for star queries, and there are a large number of BI tools based around multidimensional schemas.

5.1.2 MIMOSA OSA-EAI

The OSA-EAI provides open data exchange standards in several key asset management areas: asset registry management; work management; diagnostic and prognostic assessment; vibration and sound data; oil, fluid and gas data; thermographic data; and reliability information. These seven areas are defined by the ER model, named CRIS. The CRIS defines asset management entities, their attributes and associated types, and also relationships between entities.

As seen in Figure 5.1, a reference data library sits on top of the CRIS. The library contains reference data compiled by MIMOSA which can be stored by the CRIS and are intended to facilitate communication between MIMOSA-compliant systems. The

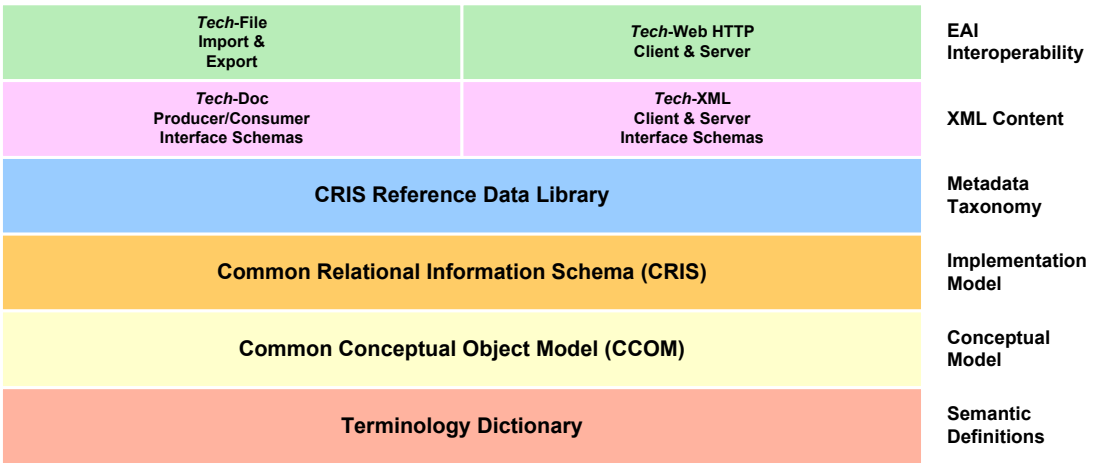


Figure 5.1 – MIMOSA OSA-EAI 3.0f layers

reference data primarily consist of 'type' information such as asset, segment, and event types. However, the largest component of the reference library is manufacturer details.

The OSA-EAI package contains SQL scripts for creating a database based on the CRIS and inserting data from the reference library. A program does not need to implement the CRIS as a database to be MIMOSA-compliant – only the XML schema must be implemented. However, a MIMOSA database implementation makes future development significantly easier in order to comply with the MIMOSA standards.

5.2 Related Literature

5.2.1 Relational Models to Multidimensional Models

There have been many methodologies proposed by researchers to design multidimensional models from ER models. While the use of the term "ER models" is technically a misnomer as multidimensional models can also be represented as ER models, for the purpose of this discussion, the term ER models will be used interchangeably with 3NF models.

While none of the techniques can provide a perfect solution of deriving a multidimensional model due to the lack of structured information contained within an ER model, it is useful to understand their reasoning. Golfarelli et al. [25] based their model around the data warehouse fact. For each fact, an attribute tree was generated which was subsequently optimised, then dimensions, fact attributes, and hierarchies were defined. Apart from the attribute tree optimisation, all other items were identified

through manual means. Cabibbo and Torlone [170] used the methodology of identifying facts and dimensions by restructuring the ER schema to better represent the facts and dimensions, deriving a dimensional graph to succinctly represent the facts and dimensions, and consequently deriving the multidimensional schema dimensions from the graph. Böhnlein and vom Ende [26] used the Structured Entity Relationship Model (SERM) to visualise existency dependencies when identifying data warehouse structures. ER models were converted to SERM, then a standard fact and dimension identification process was undertaken. Moody and Kortink [171] proposed a methodology which classified entities and identifies hierarchies to produce dimensional models. Bonifati et al. [35] investigated the design of a data mart through three steps: a top-down requirement analysis to elicit and consolidate user-requirements, a bottom-up extraction of data marts from a conceptual schema, and a comparison between the two to derive functional data marts. Marotta et al. [172] used a set of transformation rules to trace the mapping between a source logical schema and a data warehouse logical schema. Palopoli et al. [173] designed the system DIKE to extract information and properties from a database schema. The resultant data repository was then used to form the basic structure of a data warehouse schema. Chen and Hsu [174] tried a more automated approach by initially grouping entities within an ER model, assigning weights, then producing a snowflake schema by pruning the graph such that each entity was connected to the fact entity through the least expensive path.

None of the methodologies specified above are fully automatic, but require significant user interaction. In many cases, the methodologies focus on assisting the derivation of dimensions rather than facts. As with all automated or semi-automated data warehouse schema design techniques, these can be used as a starting platform for a designer who may be unfamiliar with the domain or underlying information systems.

5.2.2 Relational and Multidimensional Model Comparisons

Early on within the field of data warehousing, Colliat [168] showed the advantages of a multidimensional database over a relational one for OLAP. The former was magnitudes faster than the latter when used as the data source for a calculation of variance between budgeted and actual sales in a hypothetical beverage company. The research also showed that less disk space was required with the multidimensional approach, and proposed that more programming effort was required to maintain the multidimensional model, although the latter was not formally validated. The work only

involved one type of SQL query – that of calculations – and also did not consider data loading times.

As noted in Section 2.3.2, Rudra and Nimmagadda [70] compared multidimensional and relational structures for oil and gas exploration data and found that data size was reduced when using a combined modelling approach. The work did not highlight any other differences between the types of models.

Schuff et al. [169] investigated qualitative metrics regarding the comprehension of relational and multidimensional models. Participants were given both types of models and were tasked with modifying the schemas. The results were favourable towards multidimensional models, particularly for inexperienced users.

5.2.3 OLAP Benchmarks

There are two primary standards in OLAP benchmarking: the APB-1 OLAP Benchmark from the OLAP Council [175], and the TPC-C, TPC-E, and TPC-H benchmarks from the Transaction Processing Performance Council [176]. These types of benchmark standards simulate different realistic OLAP business situations in order to measure the performance of specific tasks. The performance can be compared cross-platform in order to assess the efficiency and viability of different combinations of software and hardware platforms. The measured operations include data loading techniques, aggregations and calculations, and an assortment of queries.

As these standards are intended for multidimensional models, they do not specify equivalent queries for relational models. However, the principles used in their methodology can still be used and harnessed.

5.3 Multidimensional Modelling Methodology

The methodology in this section was applied to the OSA-EAI CRIS and the results are discussed in Section 5.4. The approach is one that retains much of the structure of the original schema. Thus many entities are similar to their ones in the CRIS, but are collapsed into fewer tables due to denormalisation.

A multidimensional schema is not intended to be a complete replacement for a 3NF schema, but contains a subset of the data. There is a lot of data that could be contained within the OSA-EAI CRIS that is unsuitable and unnecessary for data warehousing. Thus

not all of the CRIS is changed to a multidimensional form, instead, only those parts of the schema intended to be analysed become suitable candidates. Many of the semi-automated techniques that derive multidimensional schema from ER schema ignore this fact and attempt to include every entity in the derived schema, as the techniques cannot process business requirements. Hence entities such as Enterprise and Ordered List are not included in a multidimensional model as they contain little numerical factual information and are ineffective as a dimension.

As discussed in Section 5.2.1, there are no mature automated methodologies for deriving multidimensional schemas from an underlying information source. The methodology used in deriving the multidimensional schemas from the OSA-EAI CRIS is as follows:

Step 1. Identify fact tables by examining the primary data tables
There are several primary data tables that store frequently changing information. Reference tables such as `asset_type` and `segment` do not change often, while data tables such as `meas_event` and `work_order` have rows constantly added. These frequently changing tables are more likely to be fact tables, while reference tables are almost exclusively dimension tables.

Step 2. Identify potential dimensions by examining the foreign keys of primary data tables
Potential dimensions can be located through tables joined to primary data tables via a foreign key. These tables typically have attributes that can be used in dimensions for the primary data tables. In nearly all instances, these foreign key attributes have a finite domain of values, and this is a good indicator for its suitability as a dimension attribute.

Thus, when entity e_1 contains the attribute a, and a is a primary key of entity e_2, then a is designated as a foreign key. If $|V| \in \mathbb{N}^*$ where V is the domain of values for a, then e_1 is likely to be a dimension.

Step 3. Identify conformed dimensions by examining the dimensions prevalent to multiple fact tables
Dimensions may consistently reoccur with different attributes depending on the fact table. Thus in identifying the conformed dimensions, the attribute set of the conformed dimension is the combined set of attributes from the similar dimensions.

When dimension d_1 joined to fact table f_1 and d_2 joined to fact table f_1 are conceptually equal, the common dimension formed is $d_c = d_1 \cup d_2$.

Step 4. Group dimension attributes outside of the conformed dimensions into new dimension

For those dimension attributes that cannot conceptually fit within a common dimension, these are inserted into a dimension specific to the fact table. While not quite a junk dimension, the grouping concept is similar.

Step 5. Identify fact attributes by those attributes with an infinite domain

Several steps are used in identifying fact attributes. In step 2, attributes with a finite domain of values became dimension attributes; alternatively, numeric attributes that have an infinite domain of values are likely to be fact attributes. Pre-calculated attributes, such as differences between start and end times, and differences between scheduled and actual times are also identified. As a lot of data are stored in linked binary/character/numeric tables, commonalities can be identified as fact attributes.

5.4 Asset Management Multidimensional Modelling

5.4.1 Terminology

The schema diagrams presented below append entity names with a (F) or (D). (F) signifies that the entity is a fact, while (D) signifies that the entity is a dimension. The terminology used for entity and attribute names is largely derived from the OSA-EAI Terminology Dictionary and CRIS, however, some terminology used in the conceptual data model is modified to enhance and clarify its meaning.

5.4.2 Conformed Dimensions

Conformed dimensions are the dimensions that have been standardised across multiple business units. While certain dimensions are more amenable to standardisation, others can be more subjective and vary according to business requirements. The advantage of standardising on dimensions is that data from different sources (either data warehouses/marts, or other facts) can be easily combined. This integration of data across business units can lead to unique analysis approaches.

It was found that there are certain dimensions that are universal to almost all areas of asset management data. As it forms a founding dimension for almost all data

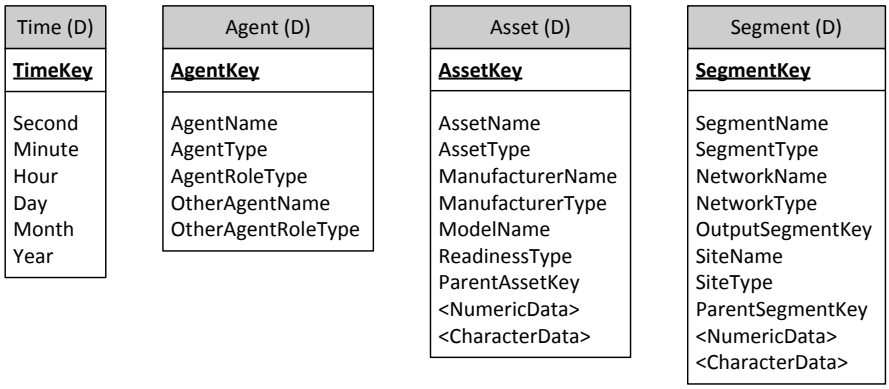


Figure 5.2 – Common conformed dimensions

warehouses, the TIME dimension is an obvious observation (presented in Figure 5.2). However, the dimensions of ASSET, SEGMENT, and AGENT were in most cases, just as ubiquitous.

Time Dimension – The smallest granularity for the TIME dimension is that of a second. While time-based records formatted to the OSA-EAI specifications (which use ISO 8601) can represent fractions of a second, the fractional component is optional. Hence, the lowest required grain for a time record is a second. Other aggregations and derivatives attributes can be included in the dimension definition, such as Quarter or Week Number in Year attributes if deemed useful for decision support.

Agent Dimension – The definition of an agent is "an animate object (person, group, organization, or intelligent agent software) that makes various types of assessments" [177]. Agents are one of the simpler constructs within the OSA-EAI, only consisting of a type, a collection of roles, and roles with other agents.

Asset Dimension – The ASSET dimension combines asset, model, and manufacturer information attributes through denormalisation. It shows the first instance of how the pervasive EAV structures can be represented through a star schema. The OSA-EAI implements three common EAV constructs: those for numeric data, those for alpha-numeric data, and those for binary data. An asset is a weakly typed entity whose real world attributes are mutable through the EAV structures. As a star schema moves towards a denormalised structure, the associated attributes in the EAV structure must be embedded within the asset dimension itself [178]. The schema designer is faced with the problem of determining common numeric and character data attributes to

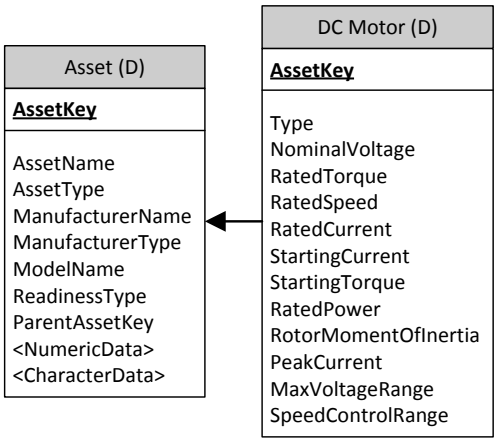


Figure 5.3 – Subclassing the asset dimension

include in the asset dimension, hence why the attributes are marked with < > markers. As not all assets share the same attributes (e.g. a motor has a voltage rating but a pipeline does not), a compatible alternate solution is to subclass the asset entity into strongly typed entities, as shown in Figure 5.3. This procedure to increase flexibility snowflakes the schema, and as a result, the potential usability decreases.

Segment Dimension – The SEGMENT dimension contains the requisite segment name, type, and numeric and character data attributes in a similar vein to the ASSET dimension. Associated networks and sites are included as well as segment hierarchies.

5.4.3 Attribute Hierarchies

Attribute hierarchies are important when using OLAP tools, as they provide users with a convenient approach of wading through data. The TIME dimension can use the classic hierarchy of Year, Month, Day, Hour, etc. while incorporating the aggregate and derivative attributes mentioned in the section above. The ASSET and SEGMENT dimensions include the parent-child hierarchy, allowing drilling down through an asset assembly or through segment children. The SEGMENT dimension contains another parent-child hierarchy through `OutputSegmentKey`, however most OLAP tools cannot process multiple parent-child hierarchies. The ASSET dimension also has an additional model hierarchy which contains the `ManufacturerName`, `AssetType`, and `Model` attributes.

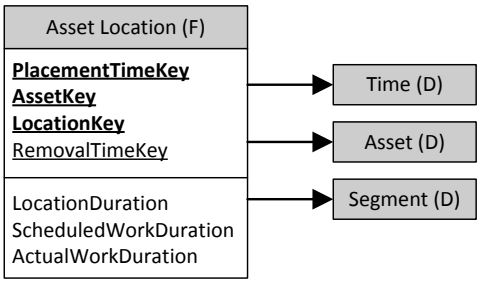


Figure 5.4 – Asset installation star schema

5.4.4 Surrogate Keys

Surrogate keys are used extensively as primary keys for entities in the OSA-EAI. As these keys have no real-world relation to the record they identify, they are omitted from the schema. However, surrogate keys are used extensively in star schemas as they provide advantages of supporting slowly-changing dimensions, referring to records with incomplete data, minimising storage space, and potentially improving join performance. While the surrogate keys used for each dimension are of integer type, they do not necessarily semantically match the surrogate keys in the OSA-EAI.

5.4.5 Configuration Data

Upon acquisition of an asset, details on the information of the asset are typically recorded in a register. These can include the purchase date and price, serial or model numbers, and technical characteristics of the asset. The relationship between assets and assemblies are also recorded. As in dimensional modelling, the characteristic data on assets are recorded in the dimension table as explained in Section 5.4.2.

An asset's location is recorded in MIMOSA through the *asset_on_segment* table. It provides vital information on which segment each asset is located in an organisation. As with the original 3NF schema, the multidimensional version, ASSET LOCATION, reports historical information on previous installations such that asset movement within a firm can be tracked over time. As with all time-related facts with designated start and end times, a pre-calculated duration fact is included to increase query speed. The schema, shown in Figure 5.4, also provides a method of analysing the scheduled and actual installation procedure duration. These two facts are not part of the original schema, but are often stored within work management records and are pertinent to assets that require non-trivial placement (i.e. installation). Subsequently, these fields would require an ETL process for population.

5.4.6 Measurement

Measurements record the condition of an asset, and can consequently trigger a health assessment or register an alarm (see Section 5.4.7). Measurement events record the time a measurement was made at a specified measurement location, along with the associated transducer and data source assets, data records, and confidence levels. Data recorded through transducers are stored in separate tables aligned to their type which is defined by the corresponding metadata. Hence time waveform data are stored distinctly to a single valued amplitude data (see Figure 5.6).

The main fact, MEASUREMENT EVENT, is a denormalised multidimensional representation of the Measurement Event entity in the OSA-EAI schema. As with the ASSET and SEGMENT dimensions in Section 5.4.2, the issue with numeric and character data is revisited. Figure 5.5 also shows that the number of primary keys for the fact table significantly outweighs the number of fact attributes, with this distinction being dictated by the original OSA-EAI model.

Measurement events can have associations with other measurement events, for example, collecting both vibration and RPM readings from a motor. This information is captured through the Measurement Event Association table in the OSA-EAI and through a non-dimension table, RELATED MEASUREMENT EVENT, for the multidimensional schema.

Contention exists over the seemingly duplicate references to asset and segment in the Measurement Event construct in the OSA-EAI [179], and Section 4.4.11 showed that the most appropriate resolution was to eliminate the asset reference and retain the segment reference. However, due to the denormalisation of multidimensional schemas,

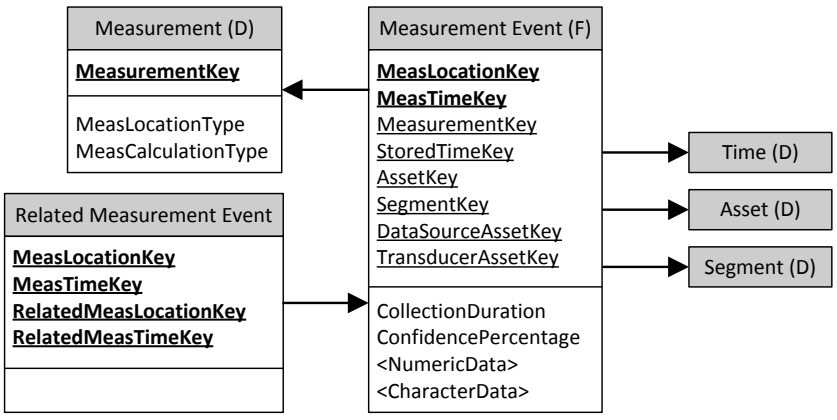


Figure 5.5 – Measurement event star schema

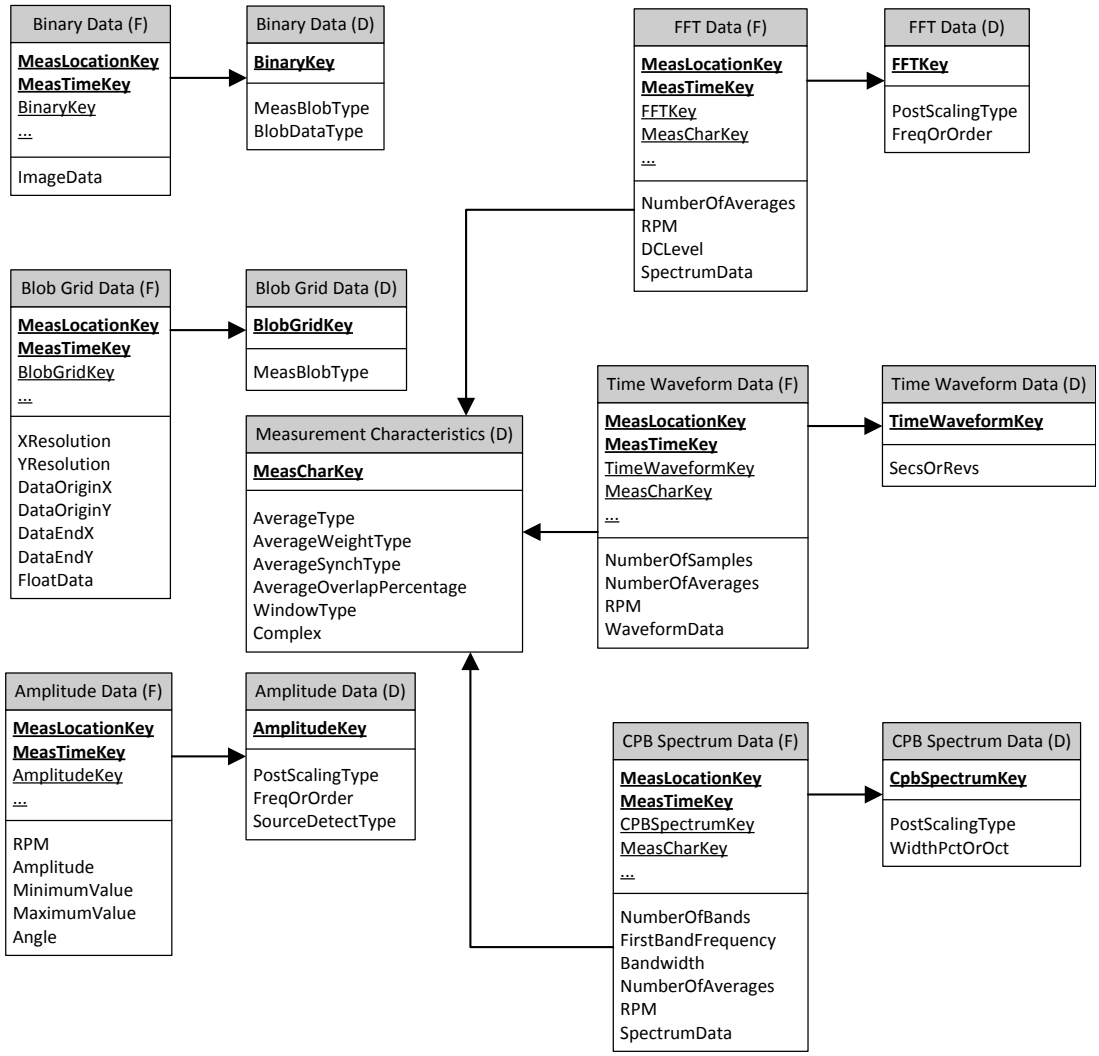


Figure 5.6 – Data table star schemas

both are equally valid, and this allows users to use either the ASSET or SEGMENT dimensions directly without having to drill-across from an ASSET INSTALLATION fact table.

Raw sensor data are stored in fact tables separate from the measurement event (as with the OSA-EAI). The primary keys remain the same as the MEASUREMENT EVENT fact to allow users to drill across to the corresponding sensor data fact by using the common dimensions of SEGMENT and TIME. The other dimensions from the MEASUREMENT EVENT schema are also included, but are simply denoted as '…' to reduce the complexity of the figure. Each fact schema is associated with new dimensions that contain metadata for the sensor readings. Additionally, as the FFT, time waveform, and CPB spectrum data types contain common parameters, these parameters are moved into the conformed

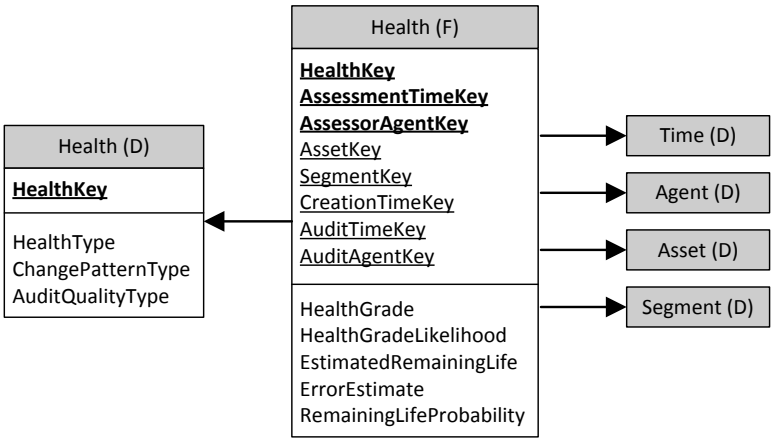


Figure 5.7 – Health data star schema

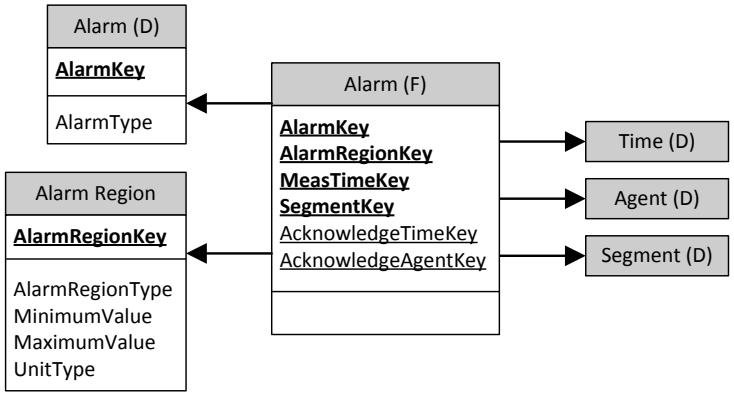


Figure 5.8 – Alarm data star schema

dimension, MEASUREMENT CHARACTERISTICS. This allows users to view facts from different tables through a common dimension.

5.4.7 Health and Alarms

Health and alarm data result from processing measurement data collected from an asset or a segment. Health data in the OSA-EAI provide an indication on the condition of an entity through codified health grades and change patterns, remaining life, and events that substantiate a health assessment. Alarm data record the measurement region that would trigger an alarm, and also any alarms registered.

The HEALTH fact table (shown in Figure 5.7) and corresponding HEALTH dimension contain the above health metrics within the OSA-EAI. An important point to note is that health data are hierarchical, and can be aggregated via assets or segments to provide a

compound figure. For example, a pump system health grade can be derived using a weighted average of the individual health values of its components (pump, motor, and shaft).

The ALARM fact in Figure 5.8 is what is known as a factless fact table. Factless fact tables do not contain any numeric facts, but are simply record an association of different dimensions. While most queries on factless fact tables result in counts of alarms through different dimensions, queries on the ALARM fact can calculate the time between alarm registration and alarm acknowledgement to determine the efficiency of business processes.

5.4.8 Event

An event is a phenomenon that occurs at a single point in time and location. It can include anything that occurs in the physical world, including asset, human, and environmental activities. While measurement events and work order events also constitute as events, they are identified as distinct from events in the OSA-EAI due to their regular occurrence in asset management operations.

Event tables in the OSA-EAI are divided into three discrete categories: those that could happen (hypothesised events), those that are scheduled to occur (proposed events), and those that have happened (actual events). As shown in Table 5.1, these event types

Event Type	Occurrence Mechanism	Linked Characteristics
Hypothetical	Segment Type Asset Type Model Segment Asset	Cause/effect link Numeric data Affected service functions link
Proposed	Segment Asset	Cause/effect link Numeric data Affected service functions link Measurement evidence link Hypothetical events link Actual events link
Actual	Segment Asset	Cause/effect link Numeric data Affected service functions link Measurement evidence link

Table 5.1 – Events

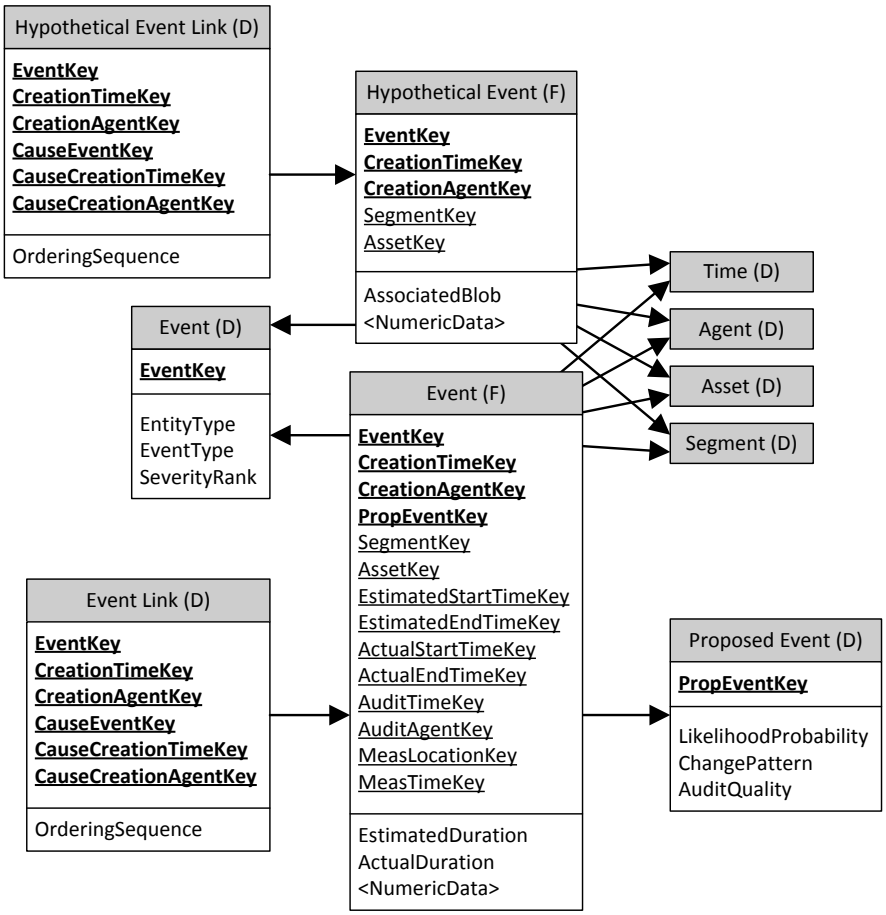


Figure 5.9 – Event star schema

are manifested through different mechanisms (segment type, asset type, model, segment, and asset) and are attributed different combinations of characteristics linked through different tables.

The EVENT star schema, shown in Figure 5.9, summarises all of the 45 event-related tables in the CRIS. The significant reduction in the number of entities comes through judicial selection of entities and attributes. By examining similarities in the structure of the type of events, Proposed and Actual events are combined into one fact table (with the `ActualStartTimeKey` being the indicator). Their combination additionally allows for pre-calculated duration facts. The `EntityType` attribute in the EVENT dimension defines the occurrence location as described above. The EVENT LINK dimensions define the causal link characteristics for EVENT and HYPOTHETICAL EVENT facts.

The numeric data fields indicated in the fact tables are taken from the *ev_num_data_type* table in the OSA-EAI, which stores 17 characteristics including MTBF

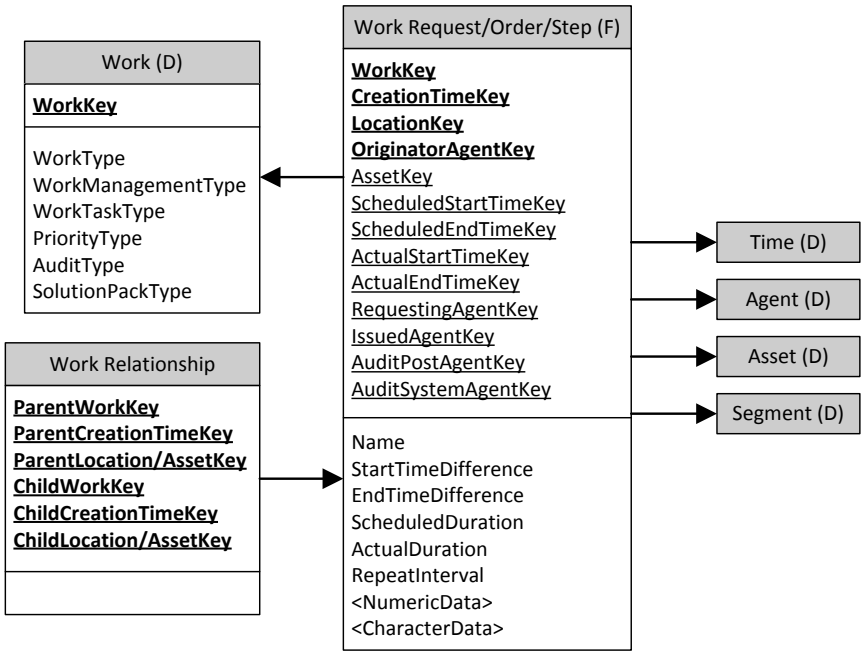


Figure 5.10 – Work request/order/step star schema

and MTTR, safety and environmental impact ratings, and costs. As there are a small number of these numeric data types, all utilised ones could be inserted as fact attributes.

5.4.9 Work Management

Work management is a core area of asset management as it forms the foundation of activities within an organisation. Nearly all firms that conduct asset management will implement work management systems at the very least, as immediate gains in productivity can be harnessed by streamlining work management processes. The work management package in the OSA-EAI can be divided into three functions: work requests, work orders, and work order steps. Requests indicate a need for work, while orders detail the actual work performed. The work management schemas are fundamentally similar to the event schemas as described in the previous section, as work can be considered either as a series of work events, or a singular encompassing work event. Both work and events have time characteristics of scheduled and actual times, audit information, and asset and segment relations. In the OSA-EAI, work and events are distinguished by their distinct intrinsic characteristics (which are stored in the dimension table of the same name).

Work requests, work orders, and work order steps consist of similar attributes and are combined into one fact table as shown in Figure 5.10. One fact table is used as the fact attributes are very similar with the fundamental difference between requests and orders being the data contained in the 'actual' time attributes – as requests are issued before work is conducted, an 'actual' time value would be null. To distinguish requests, orders and order steps, the attribute `WorkType` is included in the WORK dimension. Requests, orders and order steps can be related through the WORK RELATIONSHIP table, which allows work order steps to be aggregated into one order, or work orders to be aggregated into work requests.

Pre-calculated time durations and differences are included in the fact table to simplify queries, and numeric and character data are included (see Section 5.4.2). While the attribute `RepeatInterval` appears to be a likely candidate for inclusion in the WORK dimension, it is also a candidate for data analysis in looking at patterns of repeat intervals over various fact records.

5.5 Multidimensional Model Quality

In order to evaluate the quality of the multidimensional schema produced, metrics must be defined that address characteristics of the schema. There are several papers that address data warehouse quality, but few that address multidimensional model quality.

Calero et al. [180] approached measuring data warehouse quality through a set of numeric metrics based upon schema characteristics. However, the metrics that had a positive correlation with complexity were indicators that were relative – hence a set of multidimensional models in a similar area would be required for comparison. Jarke et al. [181] briefly addressed data warehouse schema quality through five criteria: correctness, completeness, minimality, traceability, and interpretability. Moody [182] expanded the list, albeit for ER models, to eight factors: completeness, integrity, flexibility, understandability, correctness, simplicity, integration, and implementability. Apart from correctness and simplicity which can be evaluated by CASE tools, and implementability through physical implementation, all other factors are ascertained through subjective peer reviews.

As the peer review characteristics largely depend on the deployment of a data warehouse into a business setting, the qualitative component of the quality framework

Loosely constrained joins	What was the total time/cost spent on corrective maintenance during 2005?
Tightly constrained joins	What was the total time/cost spent on corrective maintenance of our pumps on all segments during a 2005?
Calculation	What is the average difference between maintenance actual and scheduled start/end times?
Aggregate	What is the total maintenance account cost for Segment 1 for the last five years ordered by largest to smallest?
Large sort	Which segments (or assets) have been issued the most work requests?

Table 5.2 – Query types and associated questions

was not undertaken. Using CASE tools to develop the model provides inherent syntactical correctness within the model. Implementability was tested indirectly through the experimental query testing in Section 5.6.

5.6 Experimental Testing

The suitability of multidimensional models for asset management was tested through two approaches. The first approach measured the query conceptualisation complexity. Decision support systems employ queries to extract relevant data. These queries may be defined by the developer of the system at design time, or the system may allow for ad-hoc queries at runtime. In either case, the developer or the user is required to formalise the query according to the query language syntax. Less complex queries lead to a shortened design time in addition to requiring a reduced technical capability required by the designer.

The second approach was conducted by examining query execution performance. Many larger organisations run several thousand queries a day to insert data collected from the field, provide information for dashboards, generate reports for managers, and transfer information between systems. Quicker execution of queries provides organisations with more flexibility in executing a greater amount of queries, executing more complex queries, or allowing more timely decisions to be made.

Five different query types were tested against the multidimensional and relational models. The queries were written in SQL and included loosely and tightly constrained joins, calculations, aggregations, and sorting. Loosely constrained joins involved joining a small number of tables, while tightly constrained joins involved joining a large

number of tables. Calculations involve mathematics to calculate facts per record. Aggregations involve combining a large number of records into one using aggregate functions such as `AVERAGE` or `SUM`. Sorting involves setting an order for the resulting data from the query.

For each of the five query types tested, a question was attached to give the query context as shown in Table 5.2. These questions were based on the work management section of the model. This area was selected as a case study due to the pervasive nature of work management data in organisations, as seen by the data management survey (Section 3.4.4).

For each question and schema type, a SQL query was devised to answer the question. The results of the queries from the multidimensional and ER versions were evaluated against each other to ensure correctness. As Microsoft SQL Server was used, SQL functions were based on Transact-SQL. The SQL queries are shown in Appendix F.

5.6.1 Query Conceptualisation Complexity

There are several popular methods for measuring procedural code complexity. These are the lines of code, Halstead complexity measure [183], McCabe cyclomatic complexity [184], and Maintainability Index [185]. The Halstead complexity measure emphasises computational complexity by quantifying the number of operators and operands within the code. Once the rules for identifying operators and operands have been determined, five different Halstead measures can be derived. McCabe cyclomatic complexity measures the number of linearly independent paths through a program. The Maintainability Index provides a prediction of the maintainability of software code over time. The Maintainability Index is calculated using the metrics described above and is defined by:

$$171 - 5.2 \ln aveV - 0.23 aveV(g') - 16.2 \ln aveLOC + 50 \sin \sqrt{2.4 perCM}$$

where $aveV$ is the average Halstead Volume V per module
$aveV(g')$ is the average extended cyclomatic complexity per module
$aveLOC$ is the average count of lines of code per module
$perCM$ is the average percent of lines of comments per module

Query type	Format	Lines of code	Number of joins	Maintainability Index
Loosely constrained joins	ER	32	3	165
	MD	6	2	192
Tightly constrained joins	ER	47	7	158
	MD	10	4	183
Calculations	ER	7	1	189
	MD	2	0	210
Aggregations	ER	20	3	172
	MD	8	2	187
Sorting	ER	6	1	192
	MD	5	1	195

Table 5.3 – Query type characteristics

Due to SQL being a declarative language rather than a procedural language, the Halstead and McCabe measures could not be used, and hence the lines of code and Maintainability Index were the selected complexity measures. In addition, as the primary advantage of multidimensional models is a supposed reduction in the number of joins, this was another metric used in the complexity measurement.

To aid in computing the complexity, Conquest Software Solutions ClearSQL 4.2 was used. It provided several complexity metrics including the ones described above. The lines of code metric (and subsequently the Maintainability Index) was contingent on the program using its own SQL format style, rather than the one presented in Appendix E. Despite this limitation, the results highlight the same conclusion.

As can be seen in Table 5.3, multidimensional models produced the same or improved results for all query types. The lines of code were always fewer due to the reduced number of joins between tables and the use of aggregated facts. The Maintainability Index for multidimensional models was always greater (a greater number is preferable) than the ER model. The number of joins was also reduced due to the smaller number of tables within the multidimensional model. The first two query types testing differently constrained joins, required sub-queries for the relational case but were unneeded with multidimensional models.

5.6.2 Query Execution Performance

Microsoft SQL Server 2005 was selected as the DBMS while testing was performed on two machines. The first was a Dell Inspiron 6000 with an Intel Celeron M 1.40GHz, 1GB RAM, and a 5400 RPM hard disk. The second was a Dell PowerEdge 600SC with an Intel Pentium 4 2.40GHz, 512MB RAM, and a 7200 RPM hard disk. Two computers were used to validate the consistency of results. Creation SQL scripts for the multidimensional models were executed on the target machine along with the MIMOSA CRIS SQL Server database scripts. The SQL scripts contained primary and foreign key constraints. Data from each case study was preformatted before being transferred to the database.

Query execution performance was measured through two categories – metrics relating to time and metrics relating to input/output. The time metrics comprised of the CPU time and elapsed time. CPU time measures the amount of CPU resources used by a query, while elapsed time measures the time from start to completion of query execution. CPU time is the preferred measure as it is independent of the load on the computer. These were activated in SQL Server by running the command `SET STATISTICS TIME ON`.

The input/output metrics consisted of the scan count, number of logical reads, number of physical reads, and the number of read-ahead reads. Scan count is the number of times the tables referenced in a query are accessed. The measure is particularly important when the query contains joins, as a lower number of joins will result in a lower scan count and generally greater performance. Logical reads are the number of pages read from the data cache to produce the query results. Physical reads are the number of data pages read from a disk that are placed into the data cache before the query executes. Read-ahead reads are similar to physical reads in that they measure the number of data pages transferred from a disk to the data cache, but are an attempt by SQL Server to optimise performance by guessing in advance which pages are required. These were activated in SQL Server by running the command `SET STATISTICS IO ON`.

To ensure the state of the computer was equivalent for each query execution, the data and execution caches were cleared after each query completion. The command `DBCC DROPCLEANBUFFERS` was used to clear all data from the cache, while the command `DBCC FREEPROCCACHE` was used to clear the stored procedure cache. Before beginning the next query, a five second delay was introduced to let any disk activity complete.

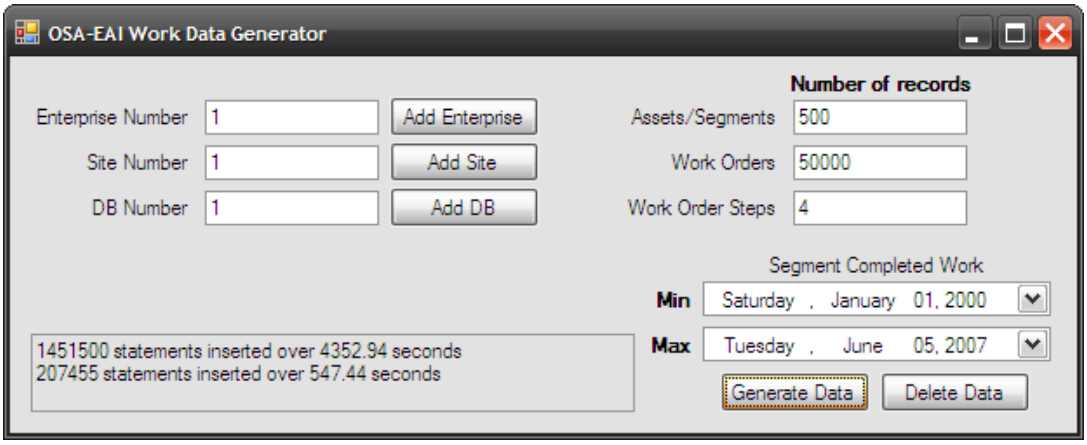


Figure 5.11 – Data set generator

A program was developed in C#.NET which would populate the `asset`, `asset_on_segment`, `segment`, `work_order`, `work_order_step`, `wo_step_num_data`, and `sg_completed_work` tables of the CRIS and the WORK, ASSET, SEGMENT, and TIME tables for the multidimensional model. The program, shown in Figure 5.11, randomly generated values for each column based on the column type. The parameters that could be input by a user were the number of work orders, assets/segments, and work order steps, along with start and end dates in which the work orders took place.

In most academic research, it is preferable to use real-world data sets over simulated data sets. In the situation of testing query execution performance of multidimensional models, the differentiation between real-world and simulated sets is negligible. The content of the data is not important, but rather the characteristics of the data. As the characteristics of the data are largely determined by table field types and these field types are the same for real-world and simulated data sets, there is little between the two types of data sets. Real-world data sets may contain blank fields (which can be simulated) and illegal characters (which would be cleaned by the ETL process), however, simulated data are easier to manipulate – particularly when comparing results from different sized data sets.

The program was used to generate eight data sets whose specifications are shown in Table 5.4 (full details in Appendix F). The combinations of sizes chosen for the work order and segments were based on companies surveyed within industry. This is with exception of data sets 1 and 8 which were included to test the robustness of both

	Data set							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Work Orders	100	500	1000	5000	10000	50000	100000	500000
Segments	50	100	100	500	500	500	1000	5000
ER Size (MB)	0.5	3	5	26	51	255	510	2557
MD Size (MB)	0.2	0.7	1	4	8	36	72	357

Table 5.4 – Tested data set specifications

models by using extremely small and large values for the number of work orders and segments.

When comparing the size of the data in each model, the multidimensional model fares extremely well compared to the ER model. This might at first seem like a misnomer as multidimensional models are denormalised and should generally be larger than their ER counterparts [186]. When comparing the number of SQL insert statements, the results are understandable. For the smallest data set, the number of insert statements is 1.9 times greater for the ER model, while for the largest data set, the number is 7.2 times greater. The disparity in statements and size is due to the EAV structure used within the OSA-EAI. For example, inserting four financial costs (labour, parts, consumables, and miscellaneous) for a work order step requires one statement in the `work_order_step` table and four in the `wo_step_num_data` table. Each of these statements requires a reference to the `work_order_step` table, a reference to a numeric data type (the cost), a value, and a unit type. Compared with one insert statement in the WORK fact table due to the denormalised costs and unit type, there is a marked difference.

Figure 5.12 graphically shows the ratio of ER to multidimensional models for the number of statements, execution time, and data size for the data insertion processes. Corresponding with the example above of the smallest data set, the number of insert statements has a ratio of 1.9, the time of insertion has a ratio of 1.7, while the data size has a ratio of 3.3. As all data sets have ratios above 1.0, it can be definitively concluded that the multidimensional models perform better for the bulk import of data with better comparative performance for larger data sets.

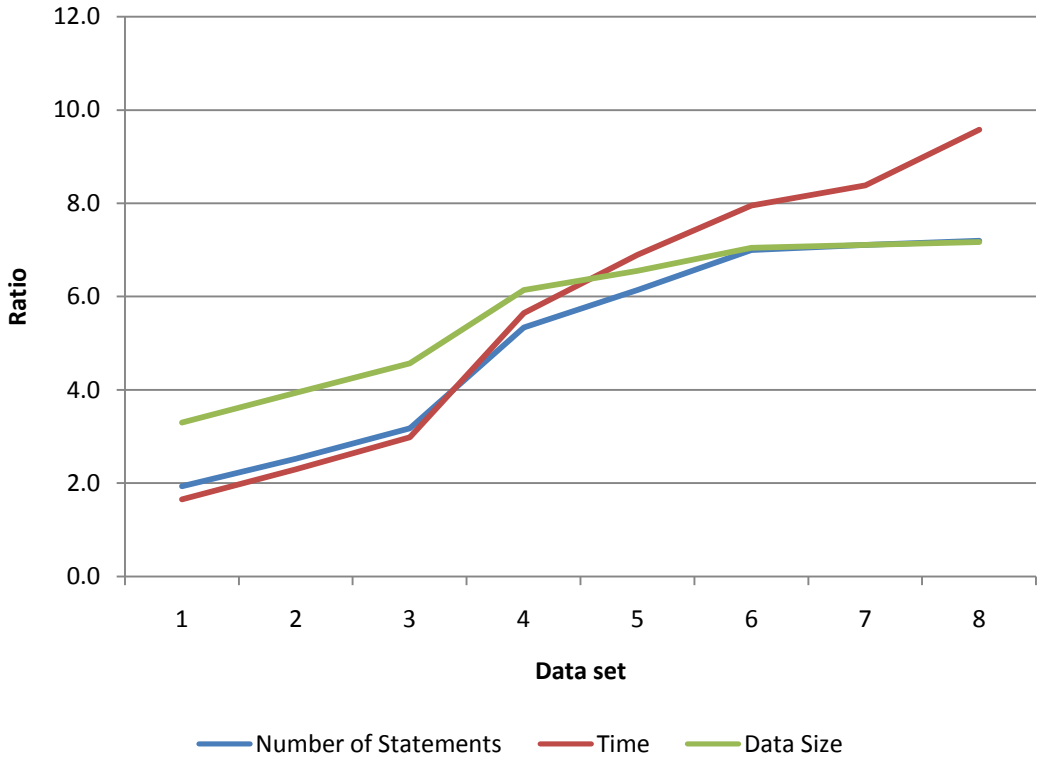


Figure 5.12 – Ratios of ER to multidimensional models for insertion

As variability can occur with the time and input/output measures (due to a non-deterministic allocation of resources by the operating system), each query was executed five times and the results averaged. Figure 5.13 shows the total query time for the first four data sets, with roughly the same histogram produced for the last four data sets, albeit with a larger y-axis (not shown). The figure shows a break down of the total CPU time and elapsed time for each data set per model. In every case, the multidimensional model required less execution time than the ER model, regardless of the data set size. Generally as the data set size increases, the difference in query time becomes more apparent. For the smallest data set, the elapsed time is 1.9 times greater for the ER model, while for the largest data set, the elapsed time is 12 times greater.

The ratio of relational and multidimensional models for execution time for each data set is plotted in Figure 5.14. Highlighting the above results, it shows a seemingly unbounded elapsed time ratio, while the CPU time ratio is more limited. The peak of the CPU time ratio at data set 4 is due to the characteristics of data set size selection – as both the number of work orders and asset/segments increases by a factor of five from data set 3 to 4. In all other cases, the number of work orders or the number of

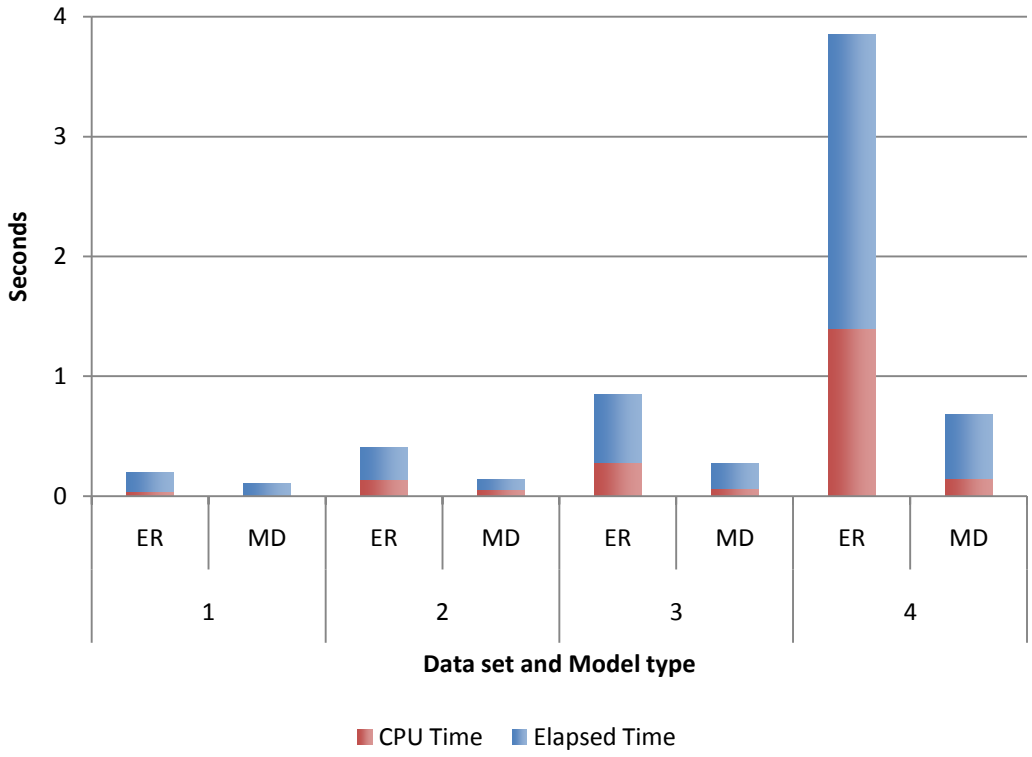


Figure 5.13 – Total query time for small data sets

asset/segments increase by non-proportional amounts, except in the move from data set 7 to 8, where a similar but less pronounced increase is seen again. The latter does not have the same effect as the move from data set 3 to 4 due to the overbearing amount of physical computational resources required to process the queries.

These two figures do not show a complete picture of the tests, due to details being obscured through the aggregation of query types. Figure 5.15 shows the CPU time increase of using multidimensional schemas per query type. The first three query types, show speed increases ranging between 1.4 and 23.4 times faster than using the ER model. However, the aggregation and sorting query types present an anomaly. The figure shows that the aggregation query type led to a speed decrease with an order of 2.4 to 35.9 times. The sorting query type did not present the same substantial gains achieved by the first three query types, achieving an increase of 1.4 times on average.

With the input/output metrics, the results favoured the multidimensional schemas in all but one case: the physical reads measure with the smallest data set. However, as the

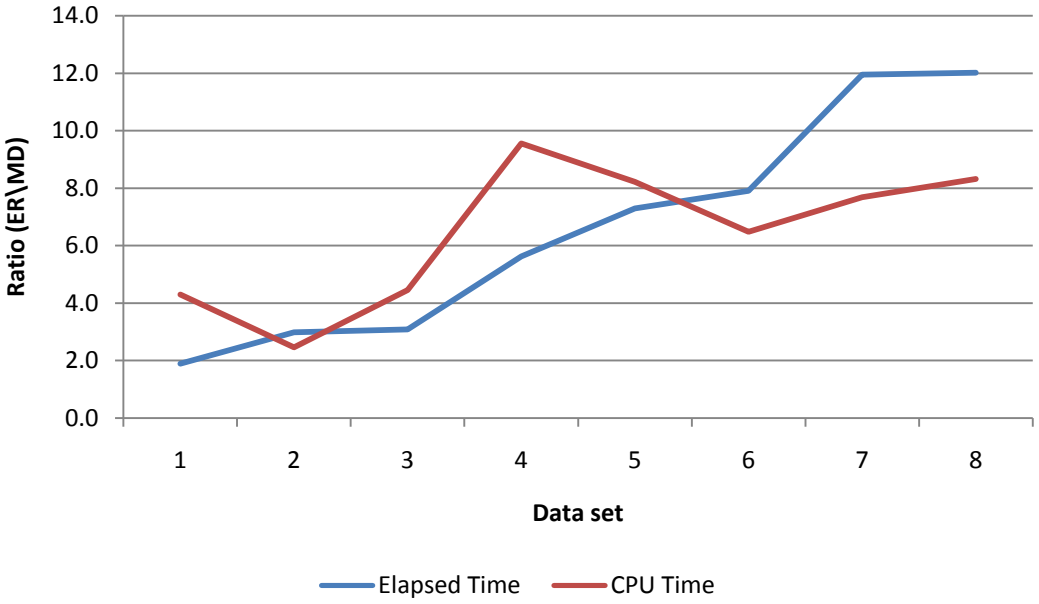


Figure 5.14 – Ratio of relational to multidimensional for execution time

difference is minor (a difference of 2 physical reads), the anomalous result was not considered significant.

The same tests were performed on the second machine to validate the consistency of results. The results were very similar although the time metrics were higher and can be attributed to the smaller amount of RAM available to the machine. The smaller amount of memory can also be attributed to the increased time ratios when comparing ER to multidimensional models. The ratio factors were increased for the second machine.

5.7 Innovation

1. *The application of multidimensional modelling to multiple asset management areas*

Prior research into multidimensional modelling of asset management data has concentrated on a select number of data areas, and in particular, those from GIS systems. This research looks at the multidimensional modelling of a number of asset management areas, including asset configuration, measurement, health and alarm, event, and work management.

2. *A formalised methodology for deriving multidimensional models from ER models*

While numerous methodologies exist for transforming an ER model into a multidimensional equivalent, this research presents an additional technique that was

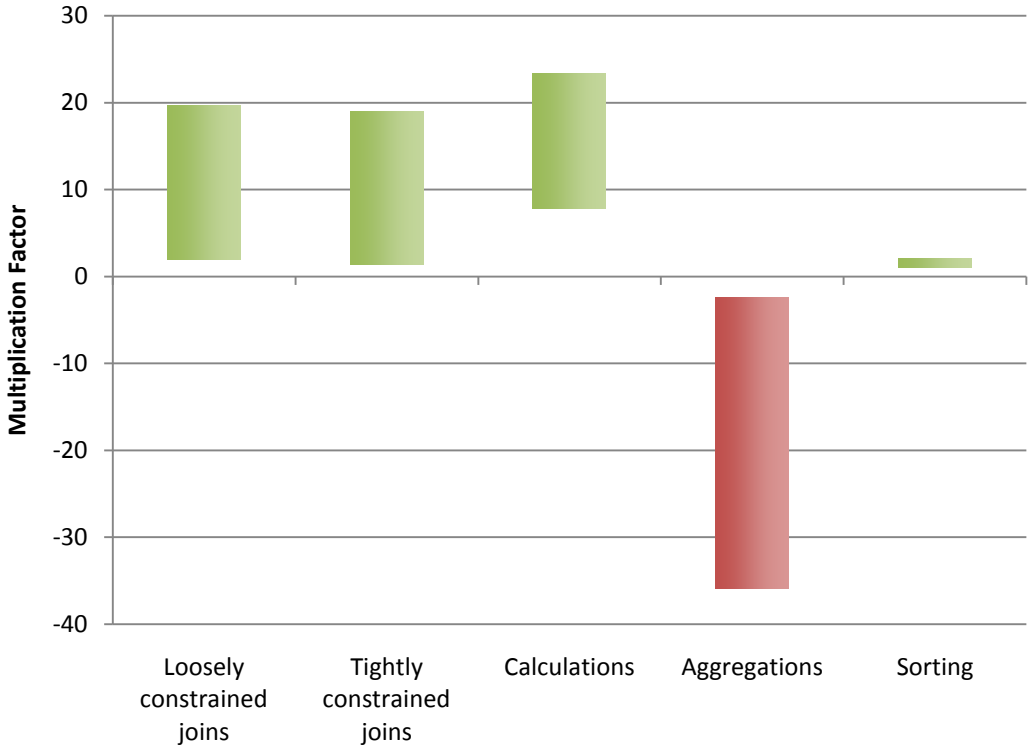


Figure 5.15 – Speed increase range per query type for MD schemas

most suitable for the OSA-EAI. The methodology adopted was derived after reviewing existing methodologies and the OSA-EAI CRIS, with the resulting methodology the most applicable. As the CRIS is weakly typed due to the use of EAV structures, the methodology would be applicable to other weakly typed relational data models.

3. The use of a data integration standard (MIMOSA OSA-EAI) as a case study

Due to the number of organisational ER models compared to standards-based ER models, nearly all multidimensional modelling research uses the former type for case studies. While there is little structural difference between the two, the generality of the latter more readily indicates the potential applicability of the research across organisations.

4. The evaluation methodology of asset management multidimensional models

This research tests both the usability and performance of the schemas per query type. In reviewed asset management multidimensional model research, only the storage efficiency was tested and compared against ER models. This research contrasts ER and

multidimensional models through the storage efficiency, the query conceptualisation complexity, and query execution performance over different query types.

5.8 Significance

1. *Evidence-based confidence for the use of multidimensional models in asset management data warehousing*

The two most notable advantages of multidimensional models are its performance efficiency as well as its human usability [187]. The rearrangement of data into a form suited for the output of data rather that the input of data can increase the execution performance of queries from hours to seconds. Users are presented with more timely information, and the greater availability of resources increases the potential number of queries that can be run. The inherent simplicity of multidimensional models allows "users to understand databases, and allows software to navigate databases efficiently" [188]. These same advantages can be harnessed for asset management processes, in making data and data analysis more accessible for users. By simplifying complex queries, data analysts are more easily and quickly able to answer questions for organisations. The investigation of these factors upon asset management data warehousing should lead to a more confident selection in choosing the most appropriate modelling methodology.

2. *Simplified data warehousing path for MIMOSA-compliant organisations*

The use of the MIMOSA OSA-EAI as a case study also has benefits as the standard provides a generic data model for a section of asset management. The resulting multidimensional model is also just as generic, and there exists a straightforward translation path for data between the ER and multidimensional equivalent. Thus organisations that use the OSA-EAI as a database or as a data integration architecture can readily harness the use of the multidimensional models in this research.

3. *A data warehouse data exchange model*

The CRIS serves as the foundational model for the upper level XML specifications in the OSA-EAI. Similarly, the model presented in this chapter can serve as the foundational element for the exchange of asset management multidimensional data. As the OSA-EAI is in constant flux, the methodology presented in this work can be employed for each update to the standard.

5.9 Conclusion

Asset management data warehousing is an open area of research as new technologies in asset management often leads to new combinations of data required, and methods of arranging that data are always in need. Data warehousing is one possible methodology which is slowly starting to be examined by the engineering asset management community.

This chapter has shown a methodology of turning third normal form schemas such as the MIMOSA OSA-EAI CRIS into a multidimensional model for data warehousing using a formal five step methodology to identify facts, dimensions, conformed dimensions, junk dimensions, and fact attributes.

When applied to the CRIS, four conformed dimensions were identified: Time, Agent, Asset, and Segment. These were used in developing star schemas for the configuration, measurement, health and alarm, event, and work management areas within the OSA-EAI.

Two aspects of the multidimensional models were tested: query conceptualisation complexity and query execution performance. Across five query types, the multidimensional models fared better than their 3NF counterparts for conceptualisation complexity. There were similar favourable results for execution performance testing, with insertion time and all query types (except the aggregation type) performing better with multidimensional models. Contrary to schema theory, the multidimensional models required a smaller amount of storage space compared to their equivalent 3NF schemas.

6 Case-Based Reasoning System for Data Warehouse Schema Design

Data warehousing technologies and methodologies have matured considerably over the past decade. The industry is lucrative, with the average estimated expenditure on data warehousing software in 2006 growing to $5.7 billion [189]. Despite the quick uptake by many organisations, there are still a significant amount of organisations that have not yet adopted data warehousing due to hurdles such as unjustifiable costs. As vast budgets are being allocated for data warehousing projects, the ability to streamline processes within a data warehouse implementation can yield significant immediate cost savings for firms.

This chapter proposes and evaluates a system based on case-based reasoning (CBR) theory to assist in the schema design process. The system is able to automate schema prototyping to consequently reduce the manual design time for data warehouse implementations while increasing the quality of the output. These two factors are intended to subsequently alleviate costs involved in data warehouse schema design.

Using CBR for data warehouse schema design is a logical step as much of the data warehouse literature on schema design is based around illustrating concepts through examples. Techniques and methodologies are presented by using potentially real scenarios and the data warehouse community adapts these examples to their situation. Using CBR for data warehouse schema design formalises this inherent process.

6.1 Background Theory

6.1.1 Case-Based Reasoning

Learning through experience is an important approach that humans employ to comprehend new problems. For instance, the diagnosis and prognosis of a patient exhibiting a set of symptoms is often revisited when the same set of symptoms appears on a different patient. While the prescription may require modification to suit the particular patient, a medical professional can significantly reduce the amount of work by reusing elements of the original diagnostic and prognostic assessment. Sustained

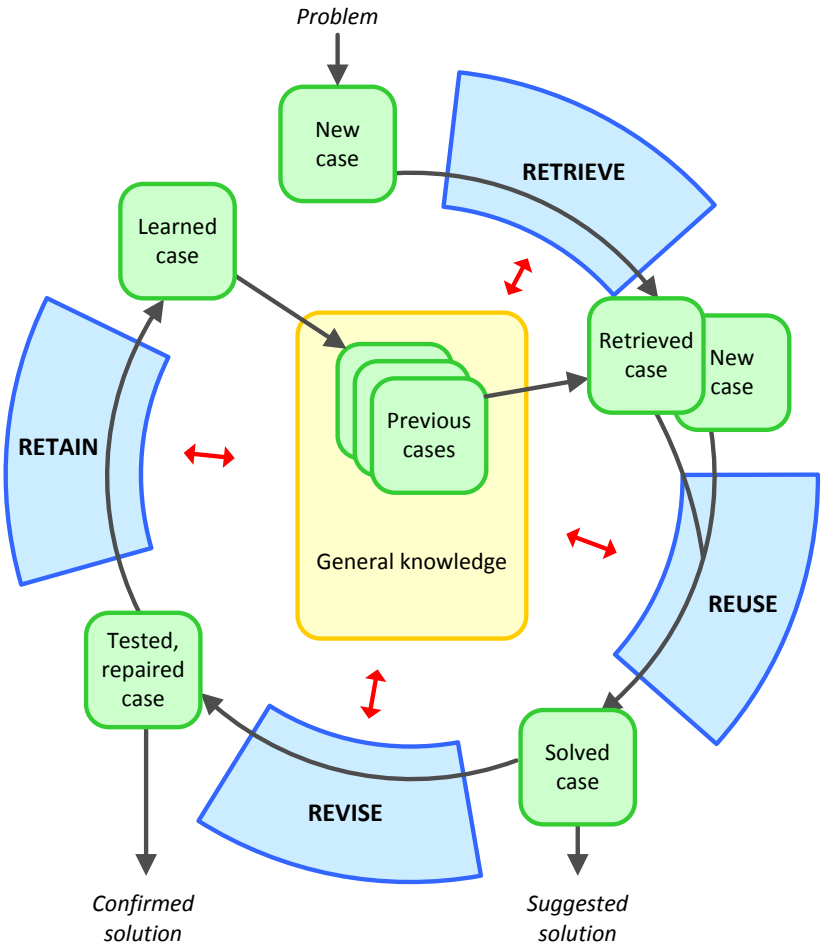


Figure 6.1 – Case-based reasoning lifecycle

learning is a consequence of such reasoning – with each successfully solved problem, the experience is retained to solve future problems; with each unsuccessful solution, the reason for failure is identified and avoided for future problems.

Case-based reasoning is formed on the above methodology. The CBR lifecycle as described by Aamodt and Plaza [190] is shown in Figure 6.1 and involves a reasoning cycle of four processes: (1) retrieving the most similar case/s; (2) reusing the information and knowledge in the case to solve the problem; (3) revising the proposed solution; and (4) retaining the parts of the experience likely to be useful for future problem solving.

The most basic representation of a case is the problem and its corresponding solution. Additional case features can be stored in order to maximise the likelihood of matching the closest case against a given problem during the retrieval stage. A case base or case

library organises and indexes cases by selected features, and case organisation is designed to achieve two goals: (1) to provide efficient searching during the case retrieval stage, and (2) to properly integrate new cases during the case retaining stage.

6.2 Related Literature

Despite the proliferation of both case-based reasoning and data warehouse schema design theory, there have been no attempts in employing case-based reasoning for data warehouse schema design. The closest related area is case-based reasoning for database schema design in which there have been three main efforts.

DES-DS (Design Expert System for Database Schema) – Paek et al. [191] designed the DES-DS with two main components, a Domain Dependent Case Base (DDCB) and a Domain Independent Case Base (DICB). The DICB consisted of nine generalised schema cases that covered different combinations of cardinalities (many-to-many, one-to-many, and one-to-one) and dependencies (partial key, transitive, and full functional). The DDCB contained complete schemas that were hierarchically indexed by a single textual identifier indicating the business area that the schema described. Case representation comprised of the aforementioned business identifier, the schema in the form of a Relation Concept Graph (RCG), and a text description of the schema. Case matching was performed by specifying user requirements in the form of a RCG, and using graph matching techniques. If the CBR system could not match an appropriate case in the DDCB, it would then derive a solution from one of the cases in the DICB.

CSBR (Common Sense Business Reasoning) – Although not based on CBR theory but similar in methodology, Storey et al. [192] introduced a database design system. Knowledge in the system was divided into three components: an Application Case Base (ACB), an Application Domain Base (ADB), and a Naive Business Model (NBM). Each of these components represented a different layer of abstraction, from the more specific ACB which contained actual cases to the more abstract NBM which stored generic business logic. One distinguishing difference compared to the above CBR systems was the use of the NBM as a thesaurus. As user provided terms for entities and attributes may vary compared to the stored cases, the NBM could resolve user terminology to case terminology.

CABSYDD (Case Based System for Database Design) – Choobineh and Lo [193] also designed a case-based reasoning system for database schema design named CABSYDD.

It also comprised of two components, a CBR system and a module that would derive schema from first principles. The case indexing was similar to that used by Paek et al. [191], in that each schema was hierarchically organised by business area. The hierarchy was formalised by categorising cases using a four tiered structure (sector, sub-sector, industry group, and department) based on the North American Industry Classification System (NAICS). Case representation included schemas expressed by Extended Entity Relationship models, textual identifiers for the business area classification, and a textual case description. Matching was performed by calculating the case with the highest matching index score. If no matching cases exist, the system invoked the module that created a new schema from first principles.

While database and data warehouse schema design are similar areas, there are many differences that distinguish the two. The function and purpose, the data stored, the techniques and technologies used for development, the usage, and the priorities are all different and the differences in attributes have been outlined in many classic data warehouse literature [18, 194].

Outside of CBR, there have been a few systems which try and formalise the data warehouse schema design process. There have been efforts to derive data warehouse schemas from underlying operational database schemas [25, 26, 195], from business process models [20, 196], from conceptual graphical models [197], and from XML sources [198]. None of the systems exploit any knowledge about previous data warehouse implementations, leaving an open area of research into investigating CBR-like systems.

6.3 Case-Based Reasoning System

6.3.1 System Architecture

The CBR system architecture, as shown in Figure 6.2, uses the common software three tiered architecture [199].

Data Layer

- The Schema Cases are a collection of the schema design and associated metadata for a particular data warehouse scenario.
- The Industrial Classification Structure provides a method of organising the Schema Cases.

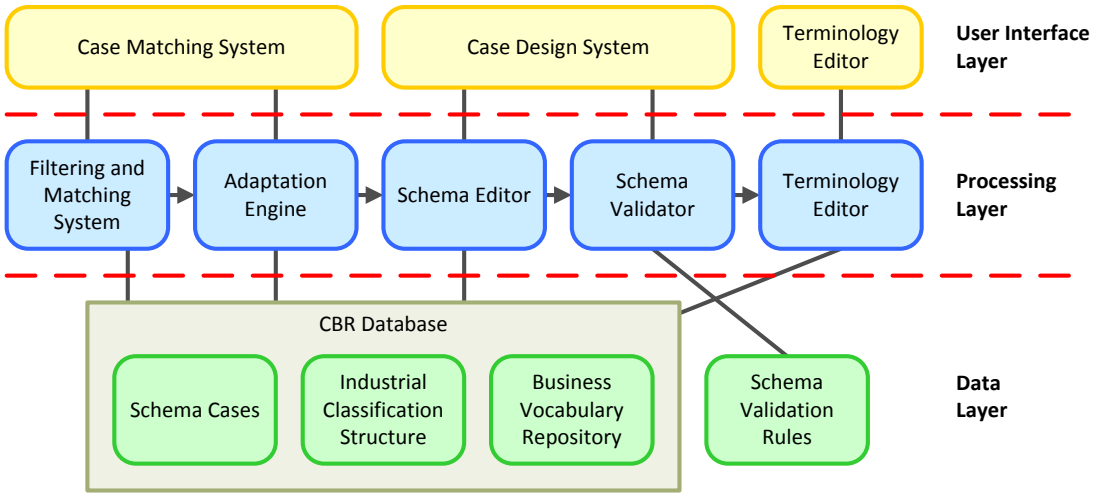


Figure 6.2 – System architecture

- The Business Vocabulary Repository (BVR) relates items between cases via business terminology.
- The Schema Validation Rules are used to enforce schema correctness.

Processing Layer

- The Filtering and Matching System matches the user-specified problem against the Schema Cases, and ranks matched cases according to their relevance.
- The Adaptation Engine automatically modifies matched cases to suit the user-specified problem.
- The Schema Editor allows manual refinement of the case.
- The Schema Validator uses the Schema Validation Rules to ensure schema correctness.
- The Terminology Editor allows for the association of schema items to items in other cases.

User Interface Layer

- The Case Matching System covers the first two CBR processes – retrieval and reuse – and provides an interface to enter the scenario description.
- The Case Design System also consists of two elements that cover the last two CBR processes – revise and retain – and provides an interface to modify schemas.
- The Terminology Editor allows users to visually connect related schema items.

The data and process layers are described in detail in the following sub-sections while the user interface layer is described in Section 6.4 as part of the implementation description.

6.3.2 Case Representation

According to Kolodner [200], a case is "a contextualized piece of knowledge representing an experience" and comprises of two characteristics: the solution itself and the context of the solution. The existing research on CBR for schema design [191-193] focussed on the former characteristic, and largely ignored the second.

By capturing the case context, a greater number of indexes can be formed leading to more efficient case matching. The features selected for each data warehouse schema design case should attempt to encompass all aspects of the design solution. By developing a comprehensive list of features that represents the totality of the experience, more flexible matching scenarios can be implemented.

Feature selection has been approached by structuring the case around a conceptual meta-model for metadata [201]. As seen in Figure 6.3, the model contains the 12 relevant subject areas for metadata: business function, subject area, purpose, steward, location, community and audience, security, data related, time and date, media type, package, and status and version. Two categories in the meta-model (project & process and event) were omitted as they were not relevant for a data warehouse schema case.

The *Artifact* category (a category for the primary data in question) contains an individual data warehouse schema with the schema attributes including the entities (fact and dimensions), relationships, relationship cardinalities, attributes, attribute types, and attribute constraints. Only one fact table is included per schema in order for the attached business process to be meaningful. An important associated category is the *Data related* category, which contains the ETL calculations performed on data sourced from the underlying OLTP system which are necessary to populate each data warehouse fact.

The *Subject area* category contains information on the industrial classification of the organisation that used the data warehouse. The industrial classification is a hierarchical entity, at the top level indicating the sector of business, and at the bottom level indicating the industry of the organisation. This attribute provides support for case

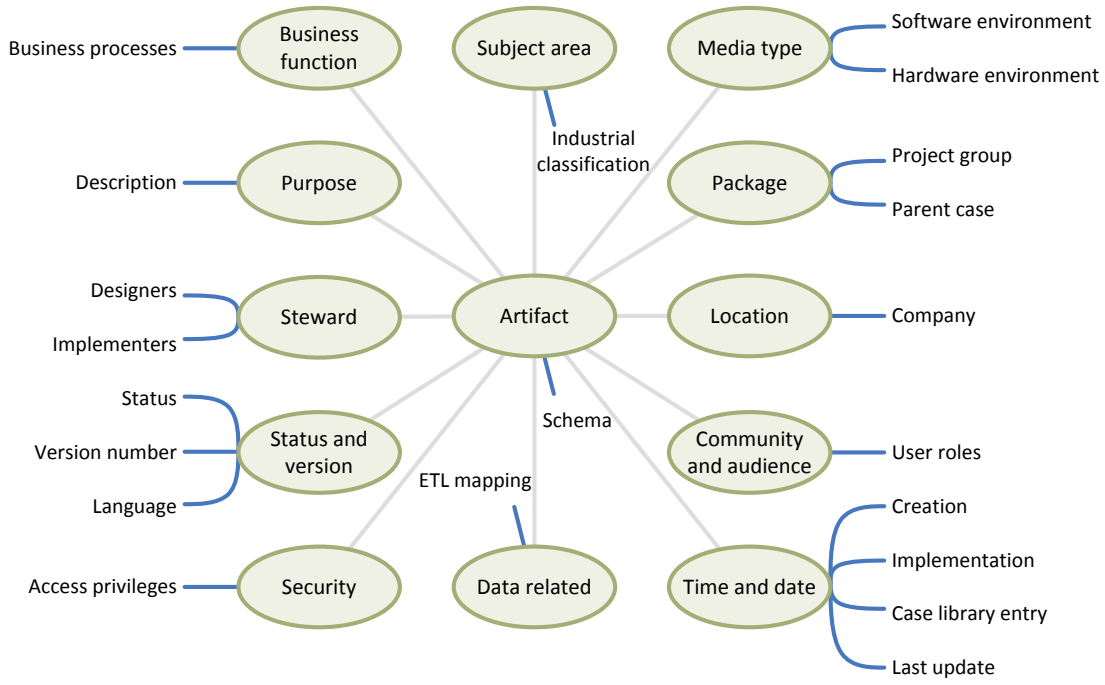


Figure 6.3 – Data warehouse schema design case representation

organisation and is discussed in Section 4.3. The industrial classification is different to the company name which is found in the *Location* category and is only stored for completeness.

Drilling down into the organisation, the related business processes that are supported by the schema are identified through the *Business function* category. The business process description may be a single identifier or multi-valued list of processes that the schema supports depending on the level of granularity required. This is different to the *Purpose* category which contains a textual description of the schema outlining its purpose.

The *Steward, Community and audience,* and *Security* categories list organisational elements, which can consist of individual people, organisational roles, groups of people, or entire organisations. Designer and implementer data are included as they can provide an indication on the quality of the underlying schema. Using these attributes to judge quality through an automatic case matching system is admittedly a difficult task. The user roles attribute indicates the organisation role of the users who utilise the data warehouse in their decision making process. Access privileges are stored to assist user restrictions on cases.

The *Media type* category covers the support technology in terms of the data warehouse and OLTP hardware and software environments, and the OLAP tools used for data exploration.

The *Package* category indicates how the case is related to other cases in the library. The project group identifies a collection of cases from a single data warehouse implementation in the organisation. The parent case is used to point to the case from which the current case is derived or reused.

The *Status and version* category contain the status of the case, such as "in production", "under testing", "under review", or "reviewed", the version, and the language of the case.

The *Time and date* category contains lifecycle dates that can be split into two categories: data warehouse lifecycle dates and CBR lifecycle dates. The design and implementation dates fall under the former category, while the case library entry and last update dates fall under the latter. This category of metadata provides the case matching system with rankings on chronological timing and popularity of schemas.

The above list of attributes does not form the totality of attributes of a data warehouse schema case, but were selected on the basis of being objective and factual attributes. Metrics such as quality and risk are subjective classifications, making it difficult to maintain consistency between cases. Monetary metrics such as costs or cost savings are influenced by many assumptions and factors and typically cannot be allocated to an individual schema. While not all case representation attributes contribute towards automated case matching, they provide information to the user of the CBR system who can use such case information in the manual refinement process.

6.3.3 Case Organisation

Having a set of cases is not sufficient for the retrieval step – the cases must first be organised in a case library. The simplest form of case organisation is having a flat array structure, whereby case matching takes place on a sequential basis. Two other popular techniques for case organisation are hierarchical trees and discrimination networks [200]. Two of the previous case-based reasoning systems for schema design [192, 193] organised the case library through a hierarchical industrial classification system. However no justification was given for the selection of a hierarchical structure.

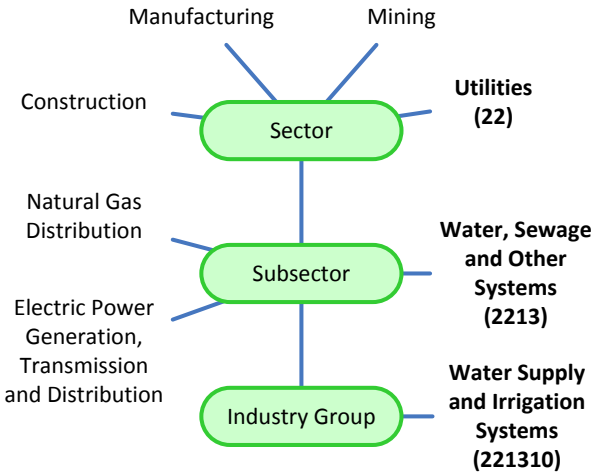


Figure 6.4 – NAICS hierarchy for Water Supply and Irrigation Systems

The advantage of a hierarchical structure is to allow for quicker and efficient matching by restricting the number of cases traversed. This is particularly important when the number of cases is very large. It can also provide a way of clustering cases based on abstraction. Cases higher in the tree will be more general compared to those lower in the tree.

The most effective organisation structure for data warehouse schemas is in using their primary industrial classification. Many industrial classification taxonomies have been published [202] and Figure 6.4 shows an example of the hierarchical structure from the North American Industrial Classification System (NAICS). The number of hierarchical levels between industrial classification taxonomies typically varies between three and four. However, the number of levels only serves as an indicator of searching flexibility. More levels in a system means an increase in the grouping of industries. This can subsequently lead to more flexible case matching scenarios at the expense of servicing additional administrative overheads in case matching and storage due to tree traversal. As industrial classification taxonomies do not include department level information, an additional level can be added to provide a finer grain, as was done in [193].

While case organisation is important when storing data directly to a disk, it becomes less important when using a relational database, which abstracts the physical storage mechanism. While a database does not preclude the use of a hierarchical structure, such a structure is difficult to maintain within a relational database. It also enforces a particular structure upon the database which may not be the most efficient retrieval or

storage methodology. In addition, any advantages in case matching speed will not be apparent due to the limited number of cases and the performance of modern day computing.

6.3.4 Business Vocabulary Repository

As discussed by Gust [203], context independent definitions are infrequently used and for terminology to be meaningful, it must be used within an appropriate context. For example, what is named a "customer" in a restaurant environment is equivalent to a "patient" in a hospital, and a "guest" in a hotel. The Business Vocabulary Repository serves as a thesaurus to associate business terminology from different contexts.

Entities are stored in the BVR using an ontological graph structure which can contain two types of nodes: terminology nodes or schema nodes. Figure 6.5 shows an example of green schema nodes related to blue terminology nodes. Schema nodes represent either facts, fact attributes, dimensions, or dimension attributes from the case library. Terminology nodes act as descriptors or tags for the schema nodes. Schema nodes can only connect to terminology nodes, although multiple terminology nodes can be connected. All associations within the BVR represent equivalence and as the purpose of the BVR is to serve as a conduit to synonyms, associations to other thesaurus term types (broader, narrower, related, converse, or homonyms) [204] are not required. The content used within terminology nodes is not important, but needs to remain consistent as its only purpose is to serve as a link between schema nodes.

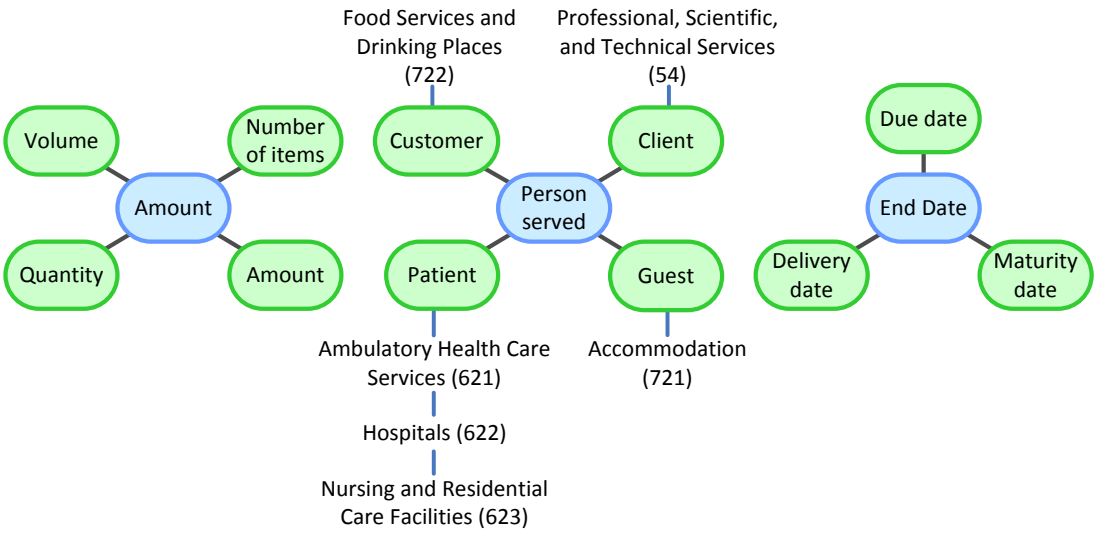


Figure 6.5 – Using industrial classification for business context association

In order to define a meaningful business context, terminology nodes are associated with zero or more industrial classification categories indirectly through schema nodes. Multiple industrial classifications can be attached to a term, as certain terms can be used in multiple industries (e.g. the Patient schema node in Figure 6.5). The attached industrial classifications can then be used during the case matching and automatic adaptation processes to give a distance ranking. By calculating the distance between the current and desired context, terms can be ranked in order of (1) exactly matching the industrial classification, (2) matching a classification in the same hierarchy, or (3) not matching.

The BVR system described above is a revision of a previous design. The previous BVR system design specified that each terminology node was directly connected to each related terminology node. This structure meant that there would be $\sum_{k=1}^{n-1} k$ associations between each node n in each terminology cluster. An implementation revealed the flaw as this method had severe performance implications for situations where the number of nodes in a group was large. The tag-based system reduced the number of associations in a cluster to n.

6.3.5 Case Filtering and Matching

To avoid computing a ranking score for every single case in the case library, cases are first filtered to determine their relevance. As shown in Figure 6.6, filtering is conducted according to four filter measures: industrial classification, business process,

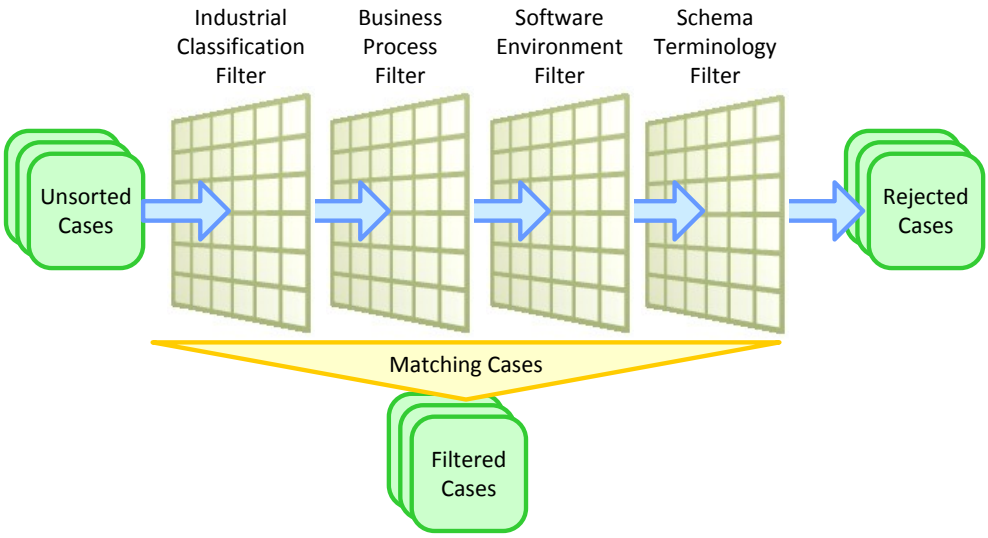


Figure 6.6 – Case matching process

environment, and schema terminology. These four measures are chosen on their relevance and ease for users to define search terms. The cases that match any filter condition are retained for the subsequent ranking stage, while those that do not match any filter are rejected.

Industrial classification filter – The industrial classification attribute is particularly important to a case as it provides an indicator to contextual terminology (as described in the previous section). As the industrial classification is a hierarchical structure, searching for cases based on their industrial classification can be performed at different hierarchical levels. Specific industries do not need to be targeted and case matching may take place on entire sectors or sub-sectors, which can produce more generalised results.

Business process filter – One of the characteristics of a data warehouse is "subject-oriented" and these "subjects" are often determined by the business processes the warehouse supports. Adequately representing a business process in a succinct manner is difficult, and issues with granularity are apparent. The approach used in this research is to use a simple textual name of the business process for the filter. For example, a user may set filter parameters to match schemas that support a certain business process, such as a pump maintenance or pipe manufacturing process. The BVR is then used to match equivalent processes between industries. Simplicity is paramount, as users most likely will not want to input business process information in detail. However, this filter has much scope for expansion, particularly if business process descriptions can be automatically read from a file generated from a business process modelling tool.

Software environment filter – Software environments encompass both the information systems from which data will be extracted, to the type of data warehouse system in which data are loaded. These two elements give an indication on potential types of calculations and aggregations that can be performed, more so when enterprises have information systems with similar database schemas. For example, an enterprise that uses the same asset management information system as one in the case library can use the same calculations and aggregations, since the data sources are the same.

An issue with both the filtering of business process and software environment information is matching entities with the same definition. In the case of the industrial classification, the definition of each industrial classification is rigidly defined by a

national governmental body. However, there are no standard enumerations of business process types or software environments, and their definitions and descriptions can be subjective. Hence issues arise when similarly named business processes in other organisations are different in meaning. For example, a pump maintenance process may involve a bearing lubrication step in one organisation, but the step may be omitted in another organisation. Increasing the granularity (e.g. including a description of every step in the business process instead of just the name of the process) will provide more accurate matches, but at the expense of increasing the complexity of the system.

Schema Terminology filter – As with the three CBR systems for database design [191-193], filtering is also performed by matching any keywords in the case (including the description). While the BVR could be used by this filter to locate equivalent terms from other industries, this approach could (1) produce a considerable decrease in performance if the number of user-specified filter parameters are large and all match terms in the BVR and/or (2) increase the number of non-relevant cases to rank.

6.3.6 Case Ranking

The case filtering process may identify zero or more cases that are applicable to the problem. In instances where multiple cases are identified, cases need to be ranked according to their degree of match. The ranking procedure is a four step process and is as follows:

Step 1. Select attributes that can be use to determine similarity

The attributes chosen are those used in the filtering process (industrial classification, business process, software environment, and schema terminology), as well as the date attributes of creation/implementation and reuse dates.

Step 2. Normalise each attribute into a scale between 0 and 1

The first measure chosen is the industrial classification. The tree structure can be used to ascertain a cardinal score by using the distance from the selected node to the node of a matched case. Hence, those cases in the same branch hierarchy will rank more favourably than those outside the hierarchy. As relationships between the top level categories in the industrial classification are not described, all categories outside the hierarchical branch (i.e. in a different sector) receive the same score of 0.

The business process, software environment, and schema terminology attributes are all nominal, using qualitative rather than quantitative values. They cannot be translated to an ordinal scale for ranking, because developing such a scale for each is too time-consuming. Hence scaling for the software environment attribute becomes a simple matter of giving a matching attribute a score of 1 and a non-matching attribute a score of 0. For the business process, attributes can either: match, match a BVR result, or not match. Cases that match receive the highest ranking score, next are those that required the BVR to match, and the lowest ranking is given to those that do not match. Likewise, schema terminology uses the BVR to produce a normalised count of the number of matches of the search terms.

Date attributes have ordinal values and can influence the ranking score in two ways: preference is given to those cases that have had a higher usage for case derivation and to those cases that are newer. Cases with a higher reuse count indicate a design that is more easily abstracted, and newer designs will typically indicate a greater relevance for modern systems and processes.

Step 3. Assign a weight to each attribute depending on the influence of the attribute to the case.

The determination of weights is largely empirical and requires adjusting parameters until the best outcome with the least error occurs.

Step 4. Calculate the case ranking.

A ranking score R for case c is calculated from the weighted sum of squared differences:

$$R(c) = 1 - \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \big(1 - sv(a_i)\big)^2$$

where w_i is the weight of the i-th attribute a_i, and $sv(a_i)$ is the scaled value of the i-th attribute of case c. A higher ranking score indicates a closer matching case.

6.3.7 Automatic Case Adaptation

There are typically two adaptation approaches in CBR theory: structural adaptation and derivational adaptation [205]. Structural adaptation involves a substitution technique where incompatible elements of the old solution are swapped with relevant

elements to solve the new problem. Derivational adaptation involves applying the reasoning path for developing the solution to the old problem to the solution to the new problem.

Structural adaptation is used in this system for two reasons. Firstly, the data stored in each case lacks a problem description. Without a problem description, the generic data warehouse schema derivational methodology cannot be utilised. The reason why a problem description is not stored is due to the breadth of problem specifications. Trying to match new problems to old to find the correct derivational path is too time consuming for users, hence why this CBR system limits the problem representation to the four filtering methods described in Section 6.3.5. The second reason why structural adaptation is used is due to the fundamental nature of using case-based reasoning for data warehouse schema design: that there is a core set of schema patterns. There are a number of commonalities between schemas with slight derivations depending on the problem. In this case, a substitution technique is more appropriate.

The highest ranked case is chosen for structural adaptation if it is not within the desired industrial classification and business process. The case library and BVR are consulted to adapt the matched case to a new case and the quantity and quality of data in these two sources will determine the degree of automation possible.

Schemas can be modified by changing the fact attributes, the dimensions themselves, or the dimension attributes. Automatic adaptation provides suggestions on modifications to these items and these suggestions need to be accepted or rejected by the schema designer. Other case data, such as the reuse score for items and dates, are used to break ties in non-definitive circumstances.

Dimension adaptation – For many multidimensional schemas within the same data warehousing project, commonalties arise in the dimensions used. Often there is a subset of dimensions that is particular to an organisation or industry that is used repeatedly with different facts. Thus the first stage in automatic adaptation is the examination of similar dimension groupings in very closely linked industries. The BVR is consulted to determine similarity through the equivalence associations between nodes, suggests potential dimensions that are frequently used with the desired business process, and suggests dimensions that are often used in the desired industry.

Attribute adaptation – Additional adaptation takes place by using the BVR to locate the equivalent terminology for fact and dimension attributes. If an equivalent term exists in the desired business context, then the located term is automatically substituted. If no equivalent terms exist in the desired business context, a list of equivalent terms can be generated ranked by their distance to the current industrial classification and presented to the user in the manual refinement stage. Extraneous or irrelevant attributes may exist in the fact or dimension tables, and as automatic adaptation is not able to determine the relevance of attributes, such attributes require removal during the manual refinement stage.

6.3.8 Manual Refinement Adaptation

Manual refinement serves as a confirmation of automatic adaptation and reinforces the learning of "good" cases. The suggestions made through automatic adaptation are presented to the schema designer along with a likelihood score. This score is calculated based on the commonality for facts and dimensions, and industrial classification distance for terminology. Each confirmation records a positive score for the item for future adaptation.

6.3.9 Case Storage

The process of inserting newly defined cases into the case library is fairly trivial. Case metadata needs to be ascertained – some can be automatically determined (e.g. the parent schema for derived cases, date information, and industrial classification, business process, software environment, and keywords used in the filtering and matching stage) while others must be input manually by the user.

6.4 Implementation

The case-based reasoning system specified above was implemented using Microsoft Visual Studio.NET and SQL Server 2005 for validation testing of the system. A relational database was selected in order to use SQL queries to simplify data manipulation procedures. SQL queries are considered as inferior for CBR systems by some, as they cannot extract cases based on similarity without augmenting the database with explicit knowledge on the relationship between concepts [206]. This issue was avoided by only using SQL queries for data insertion and extraction, while ranking and similarity calculations were performed within the procedural program code.

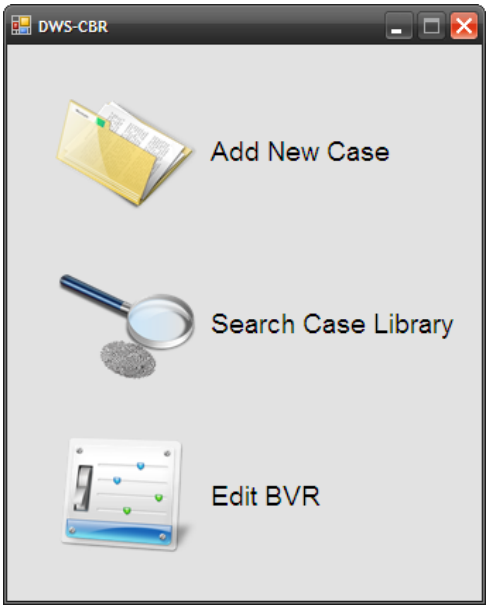


Figure 6.7 – CBR launch pad

Users are first presented with a launch pad dialog (displayed in Figure 6.7) that provides three functions: (1) searching the case library, (2) adding a new case to the system, or (3) editing the BVR.

6.4.1 Case Searching

The first step in searching the case library is to define the problem as required by the parameters in Section 6.3.5. Users are presented with the dialog in Figure 6.8. Under certain classification systems, organisations can have more than one industrial classification, hence three fields are provided. Three fields are similarly provided for business processes. The artificial limitation to three values is done because there are not many cases that require more than three values, and the restriction keeps the interface simple.

The industrial classification, business process, and software environment fields allow for auto-completion. The auto-completion for the industrial classification fields display all possible classifications while the other fields only display those items stored in the case library. Auto-completion for industrial classification is provided to assist the user in finding classifications within a specified classification system. Classifications can also be filtered and selected from the list on the right, or their numbers can be manually input. Auto-completion for business processes is limited to those already in the case library. If the user-specified business process does not exist, then the fact table to

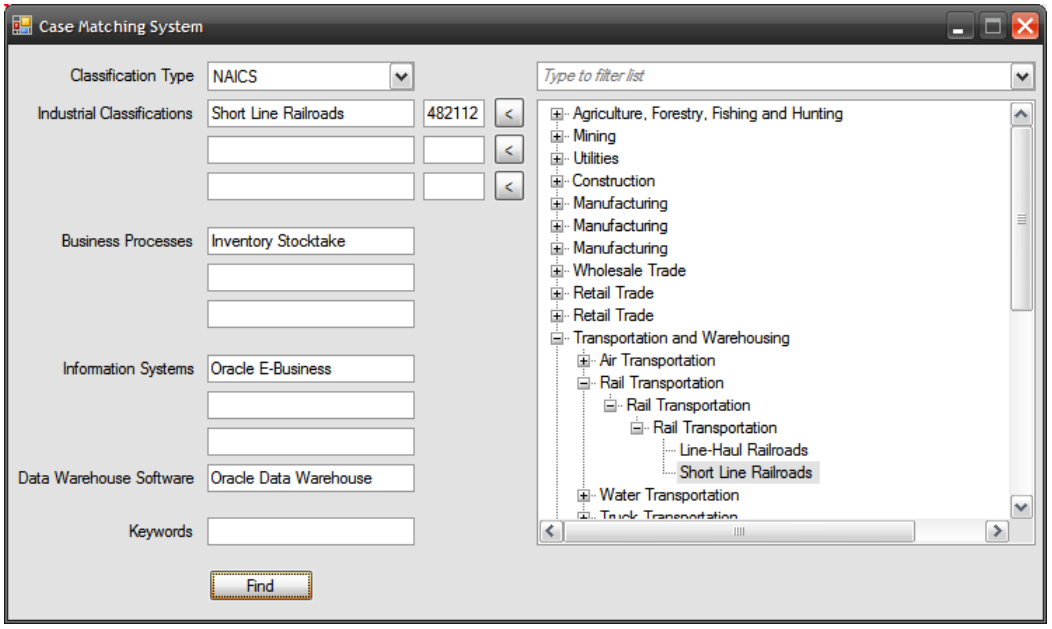


Figure 6.8 – Case searching

which the problem requires will not exist (unless the business process has an alternate name). In such a case, the business process would be given a normalised score of zero (see Section 6.3.6). However, excluding the business process entirely eliminates unnecessary filtering and ranking computation. The limited auto-completion on the software environment fields follows the same reasoning.

6.4.2 Case Adaptation

Case ranking and automatic adaptation is performed transparently to the user, with the user being presented with the results of automatic adaptation in a screen similar to Figure 6.9. Facts are ranked in order of relevance, and the top ranked fact is displayed to the user. Fact and dimension entities are differentiated by colour, with fact tables in red and dimension tables in blue (or green if it is automatically adapted). The user interface provides rudimentary control, allowing users to add, edit, and remove dimensions and attributes. Zoom and scroll functionality is implemented for larger schemas. Due to limitations of the drawing control, the layout of the fact and dimensions is not presented in the traditional star formation, although no functionality is impaired by this constraint.

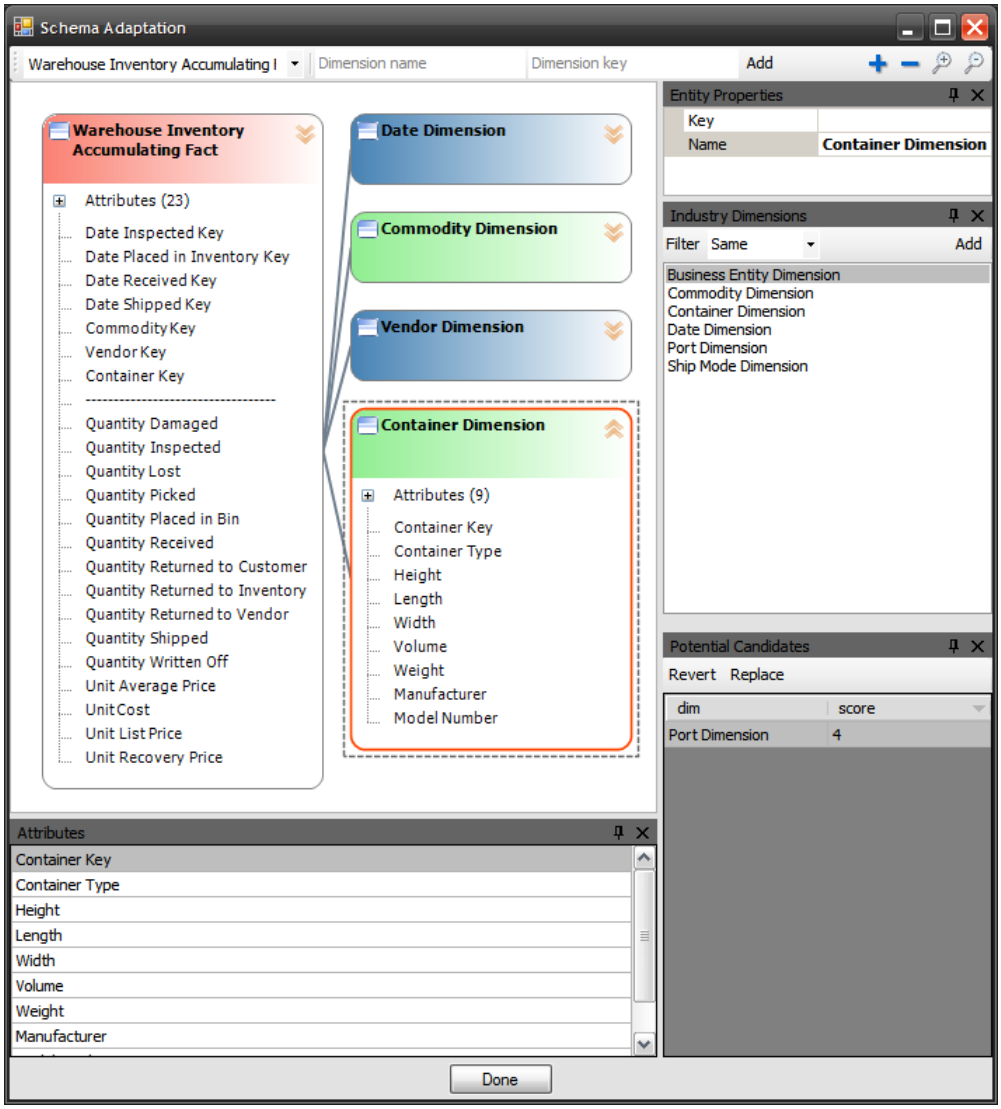


Figure 6.9 – Schema adaptation

The example in Figure 6.9 shows the results of case searching from parameters in Figure 6.8. The case library was developed from reference data warehouse schemas from literature while the BVR was developed manually for these schemas. The rail transportation case was adapted from a food manufacturer case as can be partially seen from the fact table name, which was left unmodified. The example shows two automatically adapted dimensions – `Product Dimension` became `Commodity Dimension` while `Warehouse Dimension` became `Container Dimension`.

A list of common industry dimensions is presented to the user to aid in schema construction. This list is useful in scenarios when similar dimension patterns arise among companies within the same industry. This list is generated based upon the

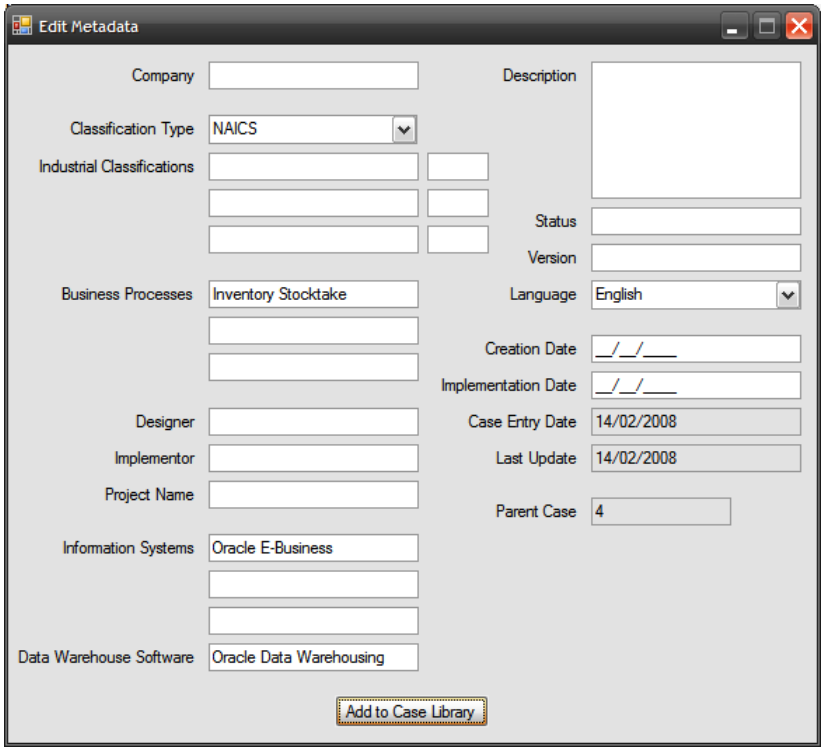


Figure 6.10 – Editing metadata

distance between the selected industrial classifications and other cases in the case library. The user can select the preferred distance such that cases that are in the Same, Close, Far, or Distant industries are visible.

For each fact attribute, dimension, and dimension attribute in the selected case, a list of potential candidates is generated. Candidates are ranked by their industrial classification distance and reuse count, and the industrial classification must be greater than a user-defined minimum distance. As candidates are items that are possibly a better match than the current item, if the selected fact has the same industrial classification as a desired classification, then candidates are ranked solely on their reuse count. As per automatic adaptation, the top ranked candidate is swapped in place for each schema item. Automatic swaps can be approved by the user to improve the adaptation algorithm, while poor adaptation results can be undone or manually swapped with other candidates.

It is possible to derive illogical schemas if the entered business process does not logically fit into the industrial classification. For example, a car manufacturing-based

fabrication process does not make sense in a financial services organisation. In such a case, little to no adaptation will occur.

6.4.3 Case Storage

Once the manual schema adaptation is complete, the next stage is to attach case metadata to the schema through the form in Figure 6.10. The fields entered during the case searching process are automatically copied to the appropriate fields. Case Entry Date, Last Update, and Parent Case are automatically determined by the system and cannot be edited.

Before the case is added via SQL to the database, validation is conducted to ensure that the case is properly formed (e.g. required fields are checked for a non-empty values). Some validation is inbuilt into the form, such as the use of combo boxes, masked fields, or auto-complete fields. Schema validation is largely built into the design process, with restrictions on the addition, editing, and deletion of schema items.

The case is stored in the database via a series of SQL statements. Ordering is important, particularly with the number of foreign keys per table for describing the schema.

6.4.4 Case Insertion

There are two methods to initially populate the case library: (1) using SQL scripts, or (2) using the Add Case GUI. Due to the system using a database, the case library can be accessed independently of the CBR. Consequently, cases represented as SQL can be inserted into the database, allowing bulk insertion of cases.

The Add Case GUI is similar to the schema adaptation form in Figure 6.9, however it does not provide a list of related dimensions or candidates. After completion of schema creation, the edit metadata form in Figure 6.10 is displayed before the case is added to the system.

6.4.5 BVR Editing

Editing the Business Vocabulary Repository is made straightforward through the user interface. Cases are sorted by facts, and changing the selected fact will determine which dimensions and attributes are displayed. Once a schema item is selected, a list of current tags is displayed. Tags can be added to the schema item from a list of available tags, and if the relevant tag does not exist, a new one can be created. For the currently

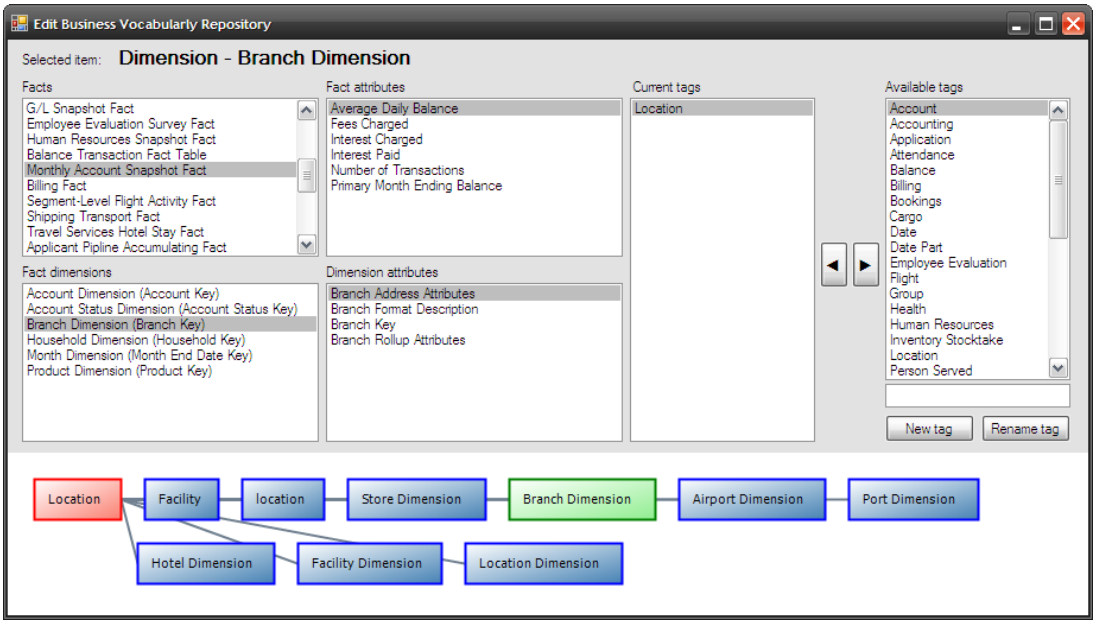


Figure 6.11 – Business Vocabulary Repository editor

selected tag, a graph is display which displays all the schema items of the same type which are also associated with the tag. The graph displays the current tag, the selected schema item, and related schema items as colour-coded nodes (red being the tag, blue being the item name, and green being the currently selected item). The graph provides a visual display to aid the user in ensuring that the semantic meaning of schema items is consistent.

An example can be seen in Figure 6.11. The `Monthly Account Snapshot Fact` is first selected, and the Branch Dimension is subsequently selected to display the tags. The only tag that is displayed is `location`. Nine dimensions which are related to the `location` tag are shown in the graph. These range from dimensions from businesses in hospitality to airport to shipping industries.

6.5 Experimental Testing

As pointed out by Tichy [207], "experimentation is central to the scientific process [as] only experiments test theories". Thus if possible, a system should be subjected to empirical evaluation to prove its scientific merit. For knowledge-based systems, evaluation includes validation of the system outcomes as well as verification of system utility and effectiveness [208].

6.5.1 Methodology

Two hypotheses were tested: the ability of the CBR system to improve the speed of schema development as well its ability to aid the translation accuracy from requirements to a functional star schema. As improvement is a relative term, the *difference* in speed and accuracy needed to be measured. Thus a comparison of schema design with and without the CBR system was undertaken. To facilitate this goal, participants were used to test both methods. Participants were given schema design tasks which outlined the requirements and the queries that the schema would need to support. Two tasks were given, and the complexity of the second task was intended to be slightly more difficult.

The tasks were set within an imaginary scenario where a company developing a decision support system required the creation of star schemas for their data warehouse backend (see Appendix C for the scenario and task descriptions). The scenario was based upon several test cases that were inserted into the data warehouse. As described in Section 6.4.2, these cases were taken from well known literature on data warehousing schema design and spanned a multitude of industries and business processes. While it might initially seem that the CBR system had an unfair advantage as the scenario was based upon the cases within, a comparison could not be made if the presented scenario had no relation to any of the cases. This is due to the design of the CBR adaptation model – if no matching industrial classification or business process is found, then no adaptation can take place (i.e. in the situation where there is no relevant case data, the CBR system would only provide suggestions on common dimensions).

As there was two tasks as well as two methods (with and without the CBR system), two ordering biases were introduced. In order to eliminate biases in the sampling technique, the group was split into four as shown in Figure 6.12. Technique bias occurs because there are two techniques – without CBR and with CBR. To minimise the bias, the "without CBR" technique was performed first in Groups 1 and 3, while the "with CBR" technique was performed first in Groups 2 and 4. Task bias occurs because there

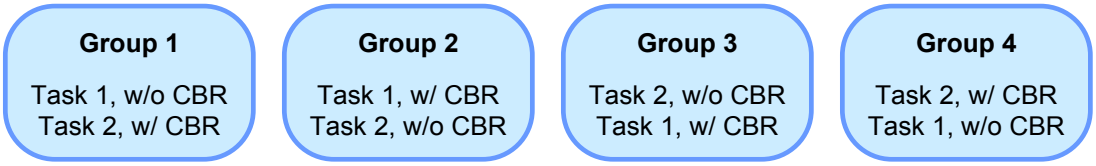


Figure 6.12 – CBR evaluation groups

are two tasks – Task 1 and Task 2. To minimise this bias, Groups 1 and 2 performed Task 1 first while Groups 3 and 4 performed Task 2 first. As interface usability was not being measured, schemas to both tasks were completed on paper as participants had more experience in writing their solutions on paper as opposed to using the CBR system.

To investigate the first hypothesis – that the CBR system improves the speed of schema development – the duration of each task was quantitatively measured from beginning to conclusion. This included the perusal of the task description, questions posed to the supervisor to aid understanding, schema design time, and review. As the scenario description preceded both tasks, the time taken for this section was not included. Both the CBR and non-CBR tools were set up beforehand to exclude preparation times, although any impact would have been negligible. A stopwatch was used to measure the duration as a high precision instrument was not required.

The second hypothesis – that the CBR system improves the translation accuracy – used a mixture of quantitative and qualitative measurements. For each fact attribute, dimension, and dimension attribute, the number of items matching and excluded from the task solution were assessed against a reference solution. The matching items increased the accuracy while the excluded ones decreased the accuracy. Items that were not in the specifications were judged based on their probability of being justified within the specifications. Justified items increased the accuracy while non-justified items decreased the accuracy.

Because there were four groups, it was preferred that the total number of people be a multiple of four in order to equalise the number per group. Twelve people were used in the testing procedure: nine of which were PhD candidates, one who had completed a Masters degree, and two who had completed a Bachelors degree. These participants had varying levels of data modelling and data warehousing experience.

6.5.2 Results Analysis

The full evaluation results for each participant are shown in Appendix G. The duration for each measurement was rounded off to the nearest minute as some individuals did not record the seconds duration, and thus the smallest common granularity was a minute. A flexible marking scheme was used for the schema accuracy characteristics, where the correct number of items was compared against the potential correct number.

Characteristic	Method	M	SD	t	df	p
Duration	Manual	21.99	7.13	.22	10	0.825
	CBR	21.17	5.19			
Fact attributes	Manual	77.8%	23.4%	.08	10	0.941
	CBR	76.4%	38.9%			
Dimensions	Manual	76.7%	22.5%	.07[a]	6.76[a]	0.304
	CBR	87.8%	9.6%			
Dimension attributes	Manual	64.8%	32.1%	.96	10	0.462
	CBR	80.7%	39.8%			

Table 6.1 – Comparison of manual and CBR methods for Test 1

Characteristic	Method	M	SD	t	df	p
Duration	Manual	23.30	6.40	.97	9	0.357
	CBR	19.36	6.93			
Fact attributes	Manual	36.0%	25.10%	-1.93	9	0.086
	CBR	68.9%	30.32%			
Dimensions	Manual	60.7%	22.91%	-2.41[a]	6.46[a]	0.058
	CBR	87.8%	14.24%			
Dimension attributes	Manual	74.3%	18.32%	-2.53[a]	4.92[a]	0.067
	CBR	94.5%	6.81%			

Table 6.2 – Comparison of manual and CBR methods for Test 2

This marking scheme incorporated the style of a participant's response as well as incorporating metrics for effort.

Task 1

As seen in Table 6.1, the six participants using the CBR system had a slightly lower mean for duration (21.17 mins) than the six participants using the manual method (21.99 mins). The CBR system also fared significantly better for dimension and dimension attribute correctness (87.8% and 80.7% vs. 76.7% and 64.8%) while the manual method fared marginally better for fact attributes (77.8% vs. 76.4%). However, as seen by the p-values, participants using either the manual or CBR methods showed no significant difference between these measured characteristics.

[a] The t and df were adjusted because variances were not equal

Task 2

One of the participants did not complete Task 2 using the manual method and hence there were only five samples using the manual method for Task 2, while there were six samples for the CBR method.

As seen in Table 6.2, the average duration for using the CBR system (19.36 mins) was notably lower than the manual method (23.30 mins) for Task 2. Each of the schema correctness measures fared significantly better for the CBR method than their manual method counterparts: fact attributes, dimensions, and dimension attributes showed more than a 20% increase. No significant difference was found for duration ($p = 0.357$), however the schema correctness measures exhibited more significant differences than in Task 1. Fact attributes scored almost double (68.9% vs. 36.0%) and had a p-value of 0.086 with the effect size, d, of 0.32 – a small to medium effect according to Cohen [209]. Dimension accuracy of the CBR method (87.8%) differed significantly to that of the manual method (60.7%) with p-value of 0.058. The effect size, d, was 1.46 indicating a very large effect. There was also a significant difference for means of each method for dimension attributes (94.5% vs. 74.3%) with a p-value of 0.067 and again producing a very large effect size with $d = 1.61$.

Discussion

Task 1 showed no significant difference between any measured characteristic while Task 2 showed a significant difference for fact attributes, dimensions, and dimension attributes for $\alpha = 0.10$. The small number of samples may have impeded the results, and a larger number of participants would provide more definitive results. As participants had no previous experience in using the CBR system, it would also be expected that the gap would increase with increased familiarity.

Figure 6.13 displays the self-ranking for each participant on their experience level in data modelling. Following a normal distribution, there were various levels of experiences for data model and star schema design. However, as participants were divided into the four testing groups, the data modelling distribution between each group did not retain the normal distribution characteristics due to the small sample size.

The differing difficulty of tasks may have also been an influencing factor. While problem difficulty can only be estimated, the general consensus was that Task 1 was

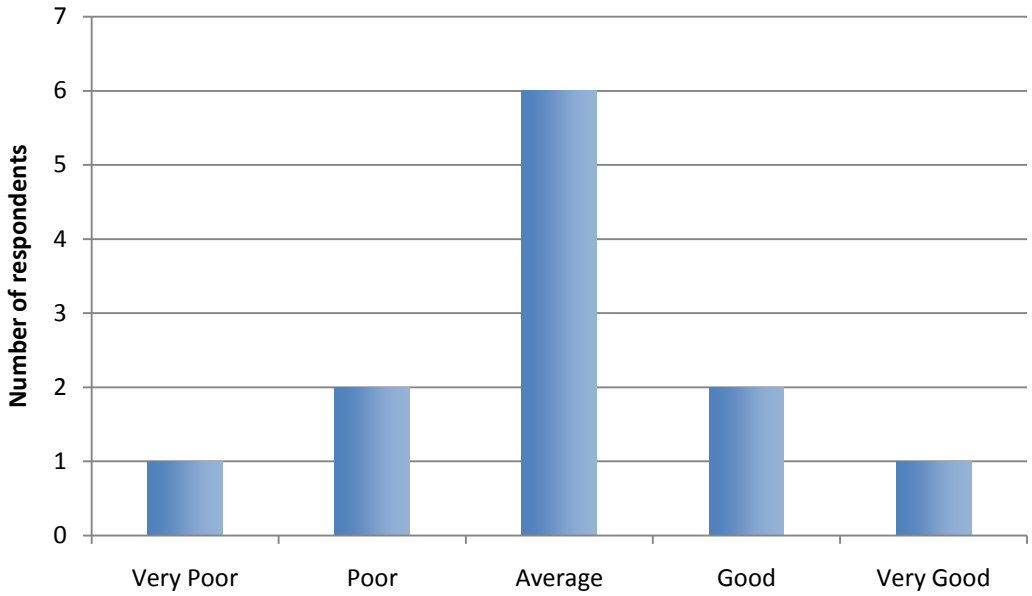


Figure 6.13 – Data modelling experience

easier than Task 2. A task that is easier would negate the difference between the two methods, with the manual method may even becoming quicker as the step of searching is eliminated.

Participants were asked to measure how the CBR system aided them in their schema design. Figure 6.14 shows that more than one-quarter of the respondents indicated that the CBR system provided innovative ideas. Thus for scenarios with vague specifications that provide only a rough guideline, the CBR system can be used to enhance schemas. However, the enhancements would be dependent on the content within the schema library.

More than 20% of respondents also believed that the CBR system improved the quality of their designs, and promoted a quicker design process. The statistics on the measured characteristics seem at ends with these two results, and it can be asserted that participants were more confident about the speed and accuracy of their CBR schemas as compared with their manually-designed equivalents.

While the statistics of the measured characteristics indicated a less than absolute confirmation of the two hypotheses, it can be concluded that the CBR system does provide a marginal benefit to the speed of schema development and translation

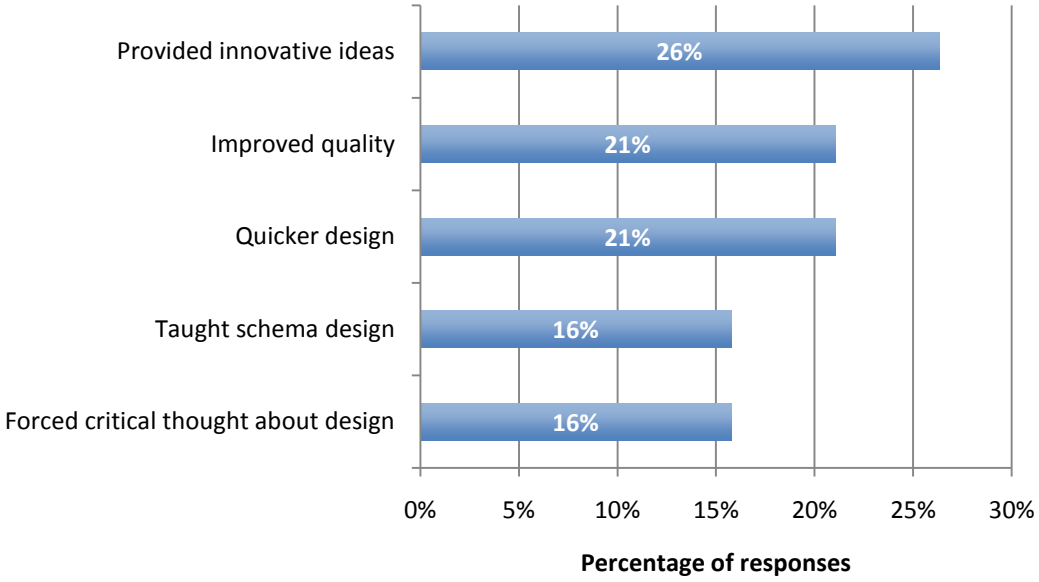


Figure 6.14 – The benefits of CBR design

accuracy (more so with difficult problems), and in the cases where that margin is zero, the user is still more confident of their schema designs.

6.6 Innovation

1. *A new methodology for data warehouse schema design*

As described in Section 2.2.3, the existing methodologies proposed for data warehouse schema design can be categorised into three groups: user oriented, operational oriented, and business process oriented. A case-based reasoning approach to data warehouse schema design can be categorised as a fourth method – case oriented – as it presents a new approach not seen in the other three.

2. *A unique case representation identification methodology*

The approach to case representation is unique as case metadata is identified via a meta-model for metadata. This enables a more structured approach in identifying the constituency of a data warehouse schema case.

3. *The use of an industry-based thesaurus for relating case items*

The use of an industry-based thesaurus to approximate the semantic equivalence of items between cases is unique to case-based reasoning for schema design. Most of these existing case-based reasoning systems associate case items directly, or though a

common reference point (e.g. tagging). The inclusion of the meaning of terminology by different industries provides enhanced functionality for case searching, ranking, and adaptation.

6.7 Significance

1. Rapid prototyping of data warehouse schemas

The goals of case-based reasoning are to provide a quicker and more accurate solution than a non-CBR method. The CBR system for data warehouse schema design exhibits these two benefits and this system will allow for an improved data warehouse development process. This will enable organisations to develop systems with a reduction in cost, and lower the barriers to entry for data warehousing.

2. Design of schemas in fields of non-expertise

Along with rapid schema prototyping, the additional benefits from a CBR system can be found in harnessing the experience of previously successful designs for clients in the same or similar industries. Because designs can build upon an existing knowledge base, designers can also propose solutions in domains where the designer is not an expert. While designs will need to be validated for correctness, the approach can provide an initial model which can be further refined.

6.8 Conclusion

There are still numerous asset management and non-asset management organisations that have not adopted data warehousing as part of their data management strategy. The research into case-based reasoning for data warehouse schema design provides an incentive to such companies by providing a lower cost path to a data warehousing system.

This research into case-based reasoning showed several academic novelties. The unique case representation and indexing methodology allowed users to employ a diverse number of searching methods. Cases were matched based on the industrial classification, business process, software environment of the problem, and the Business Vocabulary Repository could be used to enhance the system by expanding the breath of case searching as well as suggesting contextualised items for schema design.

When testing against manual schema design techniques, the CBR system produced marginally quicker and more correct schemas. The CBR system provided greater assistance for more difficult schema problems, and also enhanced a user's confidence in their designs.

Professional service organisations specialising in data warehouse design will be the primary beneficiaries of this research. Much of the system is geared towards having a quality set of data – in this case, data warehouse schema cases. Individual organisations that develop their own data warehousing solutions will not typically have access to schemas used by other industry organisations for reasons of privacy and competition. Organisations that focus on offering professional services in data warehousing have greater opportunities for developing a considerable knowledge base from their past projects.

7 Conclusion

Data warehousing plays a vital role within an organisation's IT infrastructure, forming the basis for many business intelligence and decision support systems. Data warehousing research has typically focused on domains outside of asset management, and consequently, there are numerous issues still to be addressed in the warehousing of asset management data. This thesis forms part of the body of knowledge into asset management data warehousing and signals the beginning of research into integrated data warehouse data modelling for asset management.

7.1 Research Overview

Chapter 1 introduced the background and context of the research, and indicated the direction and method of investigation as dictated by four research questions. These questions enquired about the state of asset management data warehousing within academic research and industry use; the constitution of an integrated asset management data warehouse conceptual data model; the benefits of proceeding to multidimensional models; and the rapid implementation of such models within an organisation.

Chapter 2 addressed the first component of the first research question in developing an understanding of the research into asset management data warehousing. A comprehensive systematic approach was undertaken to extensively search for all literature in asset management data warehousing. Numerous papers of data warehousing research into tangential areas were uncovered, however only a few delved into pure research for asset management data warehousing. The lack of research in this area provided significant opportunities to add value to the body of knowledge.

Chapter 3 addressed the second component of the first research question in developing an understanding into asset management data warehousing in industry. A triangulated survey was conducted through the use of a questionnaire that was validated through

organisation-specific interviews. The questionnaire was implemented both as a paper and web-based instrument, and deployed to two different samples. The questionnaire probed data warehousing topics in source information systems, integration, and data retention, as well as the reasoning behind their justification. The results showed a significant use of data warehousing within asset management organisations, although many organisations lacked the integration between systems and data.

Chapter 4 presented a conceptual data model for asset management data warehousing. The model provides a mechanism for integrating asset management data acquired from source systems. Data model patterns literature, standards, information systems, business process models, analysis methods, interviews, and business documents were analysed to gain an understanding of the asset management domain. The resulting conceptual data models used both object and relational techniques with terminology consistent with the MIMOSA OSA-EAI. Eleven data areas were deemed important for asset management: assets, segments, agents, activities, events, motivation, finances, contracts, units of measurement, measurement, and documents. Due to the reliance on EAV structures, it was posited that all types of asset management data would fit within these eleven areas. This assertion was verified through expert reviews and a comparison against MIMOSA OSA-EAI, ISO 15926, and ISA-95. Sections of the model were also validated through the implementation of an asset management software system, a comparison against the specifications of two software systems, and a comparison against the CIEAM Asset Management Framework.

Chapter 5 investigated the use of multidimensional models in asset management. The models were derived from the MIMOSA OSA-EAI CRIS as it provides a standard model that can be used within a multitude of industries. The five-step methodology for developing the schemas from the ER model in the CRIS was based on existing methodologies as well as empirical observations from dealing with the CRIS. The entire CRIS was turned into star schemas and their suitability was tested through two approaches: measuring query conceptualisation complexity and query execution performance. Across five different query types, the multidimensional model fared better than the ER counterpart for query conceptualisation complexity. Over eight data sets of varying size, multidimensional models only fared better than their equivalent ER models for four of the five query types when testing query execution performance. One unexpected observation was the disk space efficiency of the multidimensional models.

Contrary to theory, the multidimensional models required magnitudes less disk space compared to the ER models due to the OSA-EAI's heavy reliance on EAV structures.

Chapter 6 illustrated a case-based reasoning system for data warehouse schema design. Three approaches to data warehouse schema design are generally accepted, and this research proposed a novel fourth approach. Case representation was based on a meta-model for metadata, case organisation was based on indexing via an industrial classification system, and a business vocabulary repository was used as a knowledge base in forming semantic links between schema items in different schemas. Case searching involved a four-stage filter and the resulting cases were ranked through the weighted score of squared differences of normalised attributes. Automatic adaptation used the BVR to locate potential schema item substitutions. The system was implemented and tested by 12 individuals against manual schema design techniques. Testing showed a slight improvement of schema accuracy and correctness when using the CBR system, and more so for more difficult problems.

7.2 Research Contributions

This research presents several novel ideas in the area of data warehousing for asset management and has implications for both academic knowledge and industry practice. The work covered in this thesis is considered exploratory in nature as there is no previous evidence of research into data integration and multidimensional modelling for asset management data warehousing, and case-based reasoning for data warehouse schema design. However, outside of the specific areas of asset management and data warehousing respectively, there has been much discussion in data integration, multidimensional modelling, and case-based reasoning. However, the unique amalgamation of these areas is the foundation for the majority of this thesis' contributions.

7.2.1 Review of Literature

The extensive review of literature showed the potential areas of research that could be undertaken in asset management data warehousing. Due to the unique characteristics of each field, only a few researchers have presented ideas over the past decade and there are numerous avenues available for future work.

7.2.2 Asset Management Data Management Survey

The survey of data management in industry shows various characteristics within information system and data warehousing management across industries and the results can be used in a benchmark of organisational data management performance. As was done in this research, the results also show researchers where to focus their efforts to provide significant research contributions in data management. Additionally, developers of information systems and data warehouses can understand the justifications in using particular systems, and accommodate the desire of functionality in future systems.

7.2.3 Asset Management Conceptual Data Modelling

The conceptual data model for asset management data warehousing investigates the integration of asset management data. Organisations can use the model to structure their corporate information and data warehousing systems; information system developers can use the model to design systems as part of an integrated asset management platform; and researchers can use the model as a basis for data analysis algorithms for data sourced from multiple systems.

7.2.4 Asset Management Multidimensional Model Evaluation

The usability and performance benefits of multidimensional modelling of asset management data are clear, giving confidence towards their use in data warehousing both in academic and industrial use. The use of MIMOSA OSA-EAI as a case study also provides a convenient path for organisations using or planning to use the standard to quickly develop associated schemas. This research also highlights the issues for MIMOSA and other standards bodies for designing data warehouse-friendly data integration standards.

7.2.5 Case-Based Reasoning for Data Warehouse Schema Design

The unique case representation methodology, and use of an industry-driven thesaurus enhances the techniques found in previous CBR systems to enable a more effective search and adaptation process. Organisations involved with data warehouse schema design will particularly benefit from this CBR system as it provides increases in the speed and accuracy of schema prototyping.

7.3 Implications for Future Research

While this research has attempted to be thorough and comprehensive in its approach, the physical constraints of time and resources have required artificial limitations on its scope. These factors are discussed below, and are addressed in terms of where future research can build upon the work presented in this thesis.

7.3.1 Asset Management Data Management Survey

The largest limitation with the interpretation and generalisation of results is the constraining sample size. While a larger sample frame and sample could be used to confirm the results, a greater benefit would be reached in conducting another cross-sectional survey in future and comparing the results in this research to form a longitudinal study. This would show the changing trends within asset management information systems and data warehousing, and these trends could be compared to other industries.

7.3.2 Asset Management Data Modelling

The goal of a data model is to provide a supporting data platform for the development of asset management information systems, and in particular, asset management data warehousing. The conceptual stage is only the first component of the data modelling process, and future research efforts will lie in investigating how the conceptual data model can translate to logical and subsequent physical models. This would involve the identification of entities, attributes, and types, as well as determining suitable places for EAV structures or strongly typed (hard-coded) tables (e.g. in the same vein as measurements derive from activities, and insurance and warranties derive from contracts). This will enable a quicker asset management system development process, as the foundational elements in system analysis will have been already conducted.

The areas covered by the model do not cover all areas of asset management, but cover the ones generally pertinent to organisations. The inputs to the modelling process are existing models and systems, and the conceptual data model creates a new model from these existing sources. As there are currently no existing models that describe the legal aspects specific to asset management, this area has not been specifically modelled (although contracts and documents would fall into this area). There are other business specific areas that have not been covered, and future research can lie in identifying the more esoteric areas associated with asset management.

The conceptual data model competes with the current integration standards albeit only at a conceptual level. As it approaches the field of asset management from a holistic perspective, the model is uniquely different to the existing standards. However, standards already have critical mass in terms of their branding, potential reach, and reference data – something that this research does not have. To enhance this research's effectiveness, further work can be done in building an associative model that indicates the distinct and overlapping areas between various asset management standards and the conceptual data model. This work would herald a standards-based asset management data platform.

The conceptual data model is also a starting point for a capability maturity model for data analysis. Associations between the model and different analysis techniques will need to be developed – simpler techniques will generally have fewer associations, while more complex, integrated techniques would require a greater number of associations. The associative model can then be used to measure the "maturity" of an asset management organisation's data analysis capability.

Another research direction is mapping the model against a generic asset management business process model. While research into asset management process modelling is currently in its infancy [210], one aspect of the modelling process is the identification of input and output data specifications for process functions. The asset management data model can serve as the data structure for these asset management processes. The addition of a process model gives insights into the workflow of an organisation and can give rise to systems based on a model-driven architecture (MDA). The MDA approach structures systems based on an underlying model. Thus the asset management data model would indicate how data are stored within an organisation while the asset management process model would designate the workflow of the data.

7.3.3 Asset Management Multidimensional Modelling

The ultimate goal for research into asset management data warehouse data modelling is to develop a set of multidimensional schemas for the whole of asset management. These schemas would be accompanied by methods of analysis, and would be easily tailored to a specific organisation's situation. The approach in converting an ER asset management data model to a multidimensional model would be required for the logical model described in the previous section. The eventual model would contain the

advantages of both – an integrated data model for asset management with exceptional performance and usability for analysis.

Multidimensional models are not a panacea, and often relational or flat structures are preferred methods to store data. Different areas of asset management data may be more suited to the different techniques, and future research can investigate which areas are more tractable to the three storage mechanisms.

MDX is increasingly becoming a more popular substitute for SQL when querying multidimensional structures. As a native querying language for multidimensional models, MDX claims to provide increased usability by using dimensional concepts and increased performance when used with pre-calculated cubes. A comparison of MDX and SQL could be conducted for both query conceptualisation complexity and execution performance. For the former, a method of comparing the effectiveness of two different languages would first need to be formulated, particularly if a quantitative technique is preferred.

As with large organisations having a diverse asset base, there are smaller organisations that have a more specific set of assets. As discussed in Section 5.4.2, the non-Time common dimensions are loosely typed and employ an attribute that determines the dimension's type. The loosely typed asset dimension disallows organisations from viewing data from specific asset type attributes (e.g. a power rating of a motor). A possible extension to the above research could involve an automated selection process of strongly typed dimensions based on usage patterns within an organisation, or an automated identification of subclasses and common binary/character/numeric attributes for use in a snowflake schema.

7.3.4 Case-Based Reasoning for Data Warehouse Schema Design

Only a subset of the metadata was used in the eventual system, and there are many more parameters that can be used for case searching and adaptation (including subjective types). For example, the use of case designer or design time metadata can be used to ascertain the quality of schemas in the matching process. A method of quantifying and allocating costs or changes in income due to the data warehouse would provide another metric of case suitability. While a greater number of parameters does not always lead to a better model, when correctly used, they have the potential of creating a more robust and effective model. The parameters also do not need to be

simple text or number fields, and if representation issues are addressed, these parameters can taken on more complex types such as source ER data models or business process models.

Case-based reasoning typically adapts solutions from a singular matched case. A different approach to adaptation is in deriving a solution from a collection of cases. This would require the CBR system to not only store the schema, but also a justification for each individual element in the schema (facts, dimensions, attributes, types, relationships, etc.). This approach would allow a finer grained justification of elements for a problem scenario. To fully harness such additional information, a more comprehensive system for problem representation would also be required (more than just industrial classification, business process, and software environment). As mentioned above, ER and business process models could serve as the basis for problem representation, and these two are selected as they are two proven structured data warehousing modelling methodologies.

The idea of using templates in case-based reasoning [191] could also be applied to data warehouse schemas. Generic templates for business processes can be inserted in the library, and combined with justifications for each schema item (as described above). Schemas could then be generated from the top-down for particular required processes.

Bibliography

[1] J. Woodhouse. (2006, August). *PAS-55 - Asset Management: concepts & practices* [Online]. Available: http://www.reliabilityweb.com/art06/pas_55_01.htm

[2] R. F. Stapelberg, "Preliminary literature review and survey analysis report," CIEAM, Brisbane, Australia, 2006.

[3] W. H. Inmon, *Building the data warehouse*, 1st ed. Wellesley, MA: QED, 1990.

[4] G. Furlow, "The case for building a data warehouse," *IT Pro,* vol. 3, no. 4, pp. 31-34, July 2001.

[5] META Group, "Industry overview: New insights in data warehousing solutions," in *Information Week*, 1996, pp. 1-27.

[6] B. H. Wixom and H. J. Watson, "An empirical investigation of the factors affecting data warehousing success," *MIS Quarterly,* vol. 25, no. 1, pp. 17-41, March 2001.

[7] Gartner. (2005, February 24). *Gartner says more than 50 percent of data warehouse projects will have limited acceptance or will be failures through 2007* [Online]. Available: http://www.gartner.com/press_releases/asset_121817_11.html

[8] B. L. Cooper, H. J. Watson, B. H. Wixom, and D. L. Goodhue, "Data warehousing supports corporate strategy at First American Corporation," *MIS Quarterly,* vol. 24, no. 4, pp. 547-567, December 2000.

[9] Computerworld. (2007, January 1). *2007 vital signs* [Online]. Available: http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9006866

[10] T. Halpin, *Conceptual schema and relational database design*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[11] Justice Management Division, "Systems Development Life Cycle Guidance Document," United States Department of Justice, January 2003.

[12] W. H. Inmon, "What is a data warehouse?," *Prism Tech Topic,* vol. 1, no. 1, 1995.

[13] E. A. Rundensteiner, A. Koeller, and X. Zhang, "Maintaining data warehouses over changing information sources," *Communications of the ACM,* vol. 43, no. 6, pp. 57-62, 2000.

[14] R. Winter, "The current and future role of data warehousing in corporate application architecture," presented at the Annual Hawaii International Conference on System Sciences, Maui, Hawaii, USA, 2001.

[15] T. Hammergren, *Data warehousing: Building the corporate knowledge base*. London: International Thomson Computer Press, 1996.

[16] C. H. Lee. (2004). *Data warehousing processes* [Online]. Available: http://www.1keydata.com/datawarehousing/processes.html

[17] J. Ostling and R. Cintron-Allen, "Steps to successful data warehousing for telehealth/telemedicine," in *Symposium on Applications and the Internet Workshops*, San Diego, California, USA, 2001, pp. 115-119.

[18] R. Kimball, L. Reeves, M. Ross, and W. Thornthwaite, *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. New York City, New York, USA: John Wiley and Sons, 1998.

[19] M. Golfarelli and S. Rizzi, "A methodological framework for data warehouse design," in *ACM International Workshop on Data Warehousing and OLAP*, Washington, D.C., USA, 1998, pp. 3-9.

[20] M. Böhnlein and A. U. vom Ende, "Developing data warehouse structures from business process models," *Bamberger Beiträge zur Wirtschaftsinformatik,* no. 57, 2000.

[21] B. List, R. M. Bruckner, K. Machaczek, and J. Schiefer, "A comparison of data warehouse development methodologies - Case study of the process warehouse," in *Database and Expert Systems Applications*, Aix-en-Provence, France, 2002, pp. 203-216.

[22] R. Mattison, *Data warehousing: strategies, technologies, and techniques*. New York City, New York, USA: McGraw-Hill, 1996.

[23] V. Poe, P. Klauer, and S. Brobst, *Building a data warehouse for decision support*, 2nd ed. Upper Saddle River, New Jersey, USA: Prentice Hall PTR, 1997.

[24] E. F. Ewen, C. E. Medsker, and L. E. Dusterhoft, "Data warehousing in an integrated health system: building the business case," in *ACM International Workshop on Data Warehousing and OLAP*, Washington, D.C., USA, 1998, pp. 47-53.

[25] M. Golfarelli, D. Maio, and S. Rizzi, "Conceptual design of data warehouses from E/R schemes," in *Hawaii International Conference On System Sciences*, Kohala Coast, Hawaii, USA, 1998, pp. 334-343.

[26] M. Böhnlein and A. U. vom Ende, "Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems," in *ACM International Workshop on Data Warehousing and OLAP*, Kansas City, Missouri, USA, 1999, pp. 15-21.

[27] R. G. Little and M. L. Gibson, "Identification of factors affecting the implementation of data warehousing," presented at the Annual Hawaii International Conference on System Sciences, Maui, Hawaii, USA, 1999.

[28] D. Dulos, "A new dimension," *Database Magazine,* vol. 19, no. 3, pp. 32-37, 1996.

[29] S. Peterson. (1996). *Stars: A pattern language for query optimized schema* [Online]. Available: http://c2.com/ppr/stars.html

[30] J. Trujillo and M. Palomar, "An object oriented approach to multidimensional database conceptual modeling," presented at the ACM International Workshop on Data Warehousing and OLAP, Washington, DC, 1998.

[31] N. Tryfona, F. Busborg, and J. G. B. Christiansen, "starER: A conceptual model for data warehouse design," presented at the ACM International Workshop on Data Warehousing and OLAP, Kansas City, MO, 1999.

[32] J. M. Firestone, "Dimensional modeling and ER modeling in the data warehouse," Executive Information Systems Inc., Wilmington, DE, White Paper 8, June 1998 1998.

[33] K. Yao, "Design issues in data warehousing: A case study," MCompSc, Concordia University (Canada), 2003.

[34] SAS Institute, "Data warehouse administration," Cary, NC, White Paper, June 2000.

[35] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, and S. Paraboschi, "Designing data marts for data warehouses," *ACM Transactions on Software Engineering Methodology,* vol. 10, no. 4, pp. 452-483, 2001.

[36] E. Sperley, *The enterprise data warehouse*. Upper Saddle River, New Jersey, USA: Prentice Hall PTR, 1999.

[37] X. W. Xu and Q. He, "Striving for a total integration of CAD, CAPP, CAM and CNC," *Robotics and Computer-Integrated Manufacturing,* vol. 20, no. 2, pp. 101-109, April 2004.

[38] J. A. Meech, "Integration of Mine and Mill Systems," in *IEEE Advanced Process Control Symposium*, Vancouver, Canada, 2000, pp. 45-53.
[39] T. Werner and C. Vetter, "Data integrity in electric utility IT systems: A case study," in *IEEE Conference on Emerging Technologies and Factory Automation*, Catania, Italy, 2005, pp. 781-784.
[40] K. J. Dueker and J. A. Butler, "A geographic information system framework for transportation data sharing," *Transportation Research Part C: Emerging Technologies,* vol. 8, no. 1-6, pp. 13-36, 2000.
[41] G. Flowers and J. Bayles, "Computerized work management one city, one system," *Computers & Industrial Engineering,* vol. 11, no. 1-4, pp. 280-284, 1986.
[42] Hyperion Solutions. (2001). *Thames Water* [Online]. Available: http://www.hyperion.com/downloads/thames.pdf
[43] Hyperion Solutions. (2003). *Yorkshire Water* [Online]. Available: http://www.hyperion.com/downloads/uk/yorkshirewater.pdf
[44] Oracle Corporation. (2003). *Yarra Valley Water* [Online]. Available: http://www.oracle.com/pls/cis/Profiles.print_html?p_profile_id=100705
[45] V. Babovic, J.-P. Drécourt, M. Keijzer, and P. F. Hansen, "A data mining approach to modelling of water supply assets," *Urban Water,* vol. 4, no. 4, pp. 401-414, December 2002.
[46] D. Shi, Y. Lee, X. Duan, and Q. H. Wu, "Power system data warehouses," *IEEE Computer Applications in Power,* vol. 14, no. 3, pp. 49-55, July 2001.
[47] F. Dahlfors and L. Trogen, "Integrated network management improves the services of utilities," ABB Automation Systems, Västerås, Sweden, 2001.
[48] D. J. Dolezilek, "Understanding, predicting and enhancing the power system through equipment monitoring and analysis," presented at the Western Power Delivery Automation Conference, Spokane, WA, April, 2000.
[49] M. Werner and U. Hermansson, "Integrated utility data warehousing-a prerequisite to keep up with competition on electricity markets," in *International Conference on Power System Management and Control*, London, UK, 2002, pp. 130-135.
[50] M. M. Lavalle, R. Molina, N. Jacome, L. Argotte, I. Galvan, A. Martinez, and G. Arroyo, "SICORP: An enterprise data warehouse for a Mexican utility company," presented at the PowerCON, New York, NY, 2003.
[51] X. He, G. Wang, and J. Zhao, "Research on the SCADA /EMS system data warehouse technology," in *IEEE/PES Transmission and Distribution Conference and Exhibition: Asia and Pacific*, Dalian, China, 2005, pp. 1-6.
[52] J. D. McDonald, "Substation automation - IED integration and availability of information," *IEEE Power and Energy Magazine,* vol. 1, no. 2, pp. 22-31, 2003.
[53] M. Kezunović and G. Latisko, "Requirements specification for and evaluation of an automated substation monitoring system," presented at the CIGRE, Calgary, Canada, September, 2005.
[54] M. S. Thomas, D. Nanda, and I. Ali, "Development of a data warehouse for non-operational data in power utilities," presented at the IEEE Power India Conference, New Delhi, India, April, 2006.
[55] S. Draber, E. Gelle, T. Kostic, O. Preiss, and U. Schluchter, "How operation data helps manage lifecycle costs," presented at the International Conference on Large High-Voltage Electric Systems, Paris, France, August, 2000.
[56] S. Krstonijević, N. Čukalevski, G. Jakupović, N. Damjanović, and S. Cvetićanin, "Real-time transformer dynamic loading application - Implementation and practical use," in *Deregulated Electricity Market Issues in South-Eastern Europe*, Istanbul, Turkey, 2007, pp. 62-68.

[57] J. Pathak, Y. Li, V. Honavar, and J. D. McCalley, "A service-oriented architecture for electric power transmission system asset management," in *International Conference on Service-Oriented Computing*, Chicago, IL, 2006, pp. 26-37.

[58] Engineers Australia, "Australian infrastructure report card," Barton, ACT, Australia, August 2005.

[59] A. Shah, T.-H. Tan, and A. Kumar, "Building infrastructure asset management: Australian practices," presented at the CIB World Building Congress, Toronto, Canada, May, 2004.

[60] T. Clash and J. Delaney, "New York State's approach to asset management: A case study," *Transportation Research Record,* vol. 1729, pp. 35-41, 2000.

[61] J. Hall, R. Robinson, and M. Paulis, "Enterprisewide spatial data integration of legacy systems for asset management: The case of the Illinois Department of Transportation," *Transportation Research Record,* vol. 1917, pp. 11-17, 2005.

[62] E. Wittwer, J. Bittner, and A. Switzer, "The fourth national transportation asset management workshop," *International Journal of Transport Management,* vol. 1, no. 2, pp. 87-99, October 2002.

[63] S. L. Yoder and J. Delaurentiis, "The framework for a regional transit asset management system," *Institute of Transportation Engineers Journal,* vol. 73, no. 9, pp. 42-47, September 2003.

[64] P. Herabat, D. Satirasetthavee, and A. Amekudzi, "Web-based rural road asset-management system," *Transportation Research Record,* vol. 1855, pp. 105-111, 2003.

[65] B. Sroub and J. Mackraz, "System and method for managing transportation assets," U.S. Patent 20030135304, July 17, 2003.

[66] P. F. Knights and L. K. Daneshmend, "Open systems standards for the mining industry," in *Information Technologies in the Minerals Industry*, Athens, Greece, 1997, pp. 3-10.

[67] G. H. Harrison and F. Safar, "Modern E&P data management in Kuwait Oil Company," *Journal of Petroleum Science and Engineering,* vol. 42, no. 2-4, pp. 79-93, April 2004.

[68] R. Hensel and R. Oelhaf, "Optimization of pipeline operation using integrated information management," *OIL GAS European Magazine,* vol. 30, no. 1, pp. 14-18, 2004.

[69] B. Kaufman, "GIS inside an independent oil and gas company," presented at the ESRI User Conference, San Diego, CA, August, 2004.

[70] A. Rudra and S. Nimmagadda, "Roles of multidimensionality and granularity in warehousing Australian resources data," in *Hawaii International Conference on Systems Sciences*, Big Island, HI, 2005, pp. 216-221.

[71] S. Nimmagadda and H. Dreher, "Mapping and modeling of oil and gas relational data objects for warehouse development and efficient data mining," in *International IEEE Conference on Industrial Informatics*, Singapore, 2006, pp. 1201-1206.

[72] S. L. Nimmagadda, H. Dreher, and A. Rudra, "Data warehouse structuring methodologies for efficient mining of Western Australian petroleum data sources," presented at the IEEE International Conference on Industrial Informatics, Perth, Australia, August, 2005.

[73] A. G. Büchner, S. S. Anand, and J. G. Hughes, "Data mining in manufacturing environments: Goals, techniques and applications," *Studies in Informatics and Control,* vol. 6, no. 4, pp. 319-328, 1997.

[74] H. K. Park and J. Favrel, "Virtual enterprise - Information system and networking solution," *Computers & Industrial Engineering,* vol. 37, no. 1, pp. 441-444, October 1999.

[75] H. C. W. Lau, B. Jiang, W. B. Lee, and K. H. Lau, "Development of an intelligent data-mining system for a dispersed manufacturing network," *Expert Systems,* vol. 18, no. 4, pp. 175-185, September 2001.

[76] R. M. Dabbas and H.-N. Chen, "Mining semiconductor manufacturing data for productivity improvement -- an integrated relational database approach," *Computers in Industry,* vol. 45, no. 1, pp. 29-44, May 2001.

[77] H. Hinrichs and T. Aden, "An ISO 9001:2000 compliant quality management system for data integration in data warehouse systems," presented at the International Workshop on Design and Management of Data Warehouses, Interlaken, Switzerland, June, 2001.

[78] H. J. Watson, D. L. Goodhue, and B. H. Wixom, "The benefits of data warehousing: why some organizations realize exceptional payoffs," *Information & Management,* vol. 39, no. 6, pp. 491-502, May 2002.

[79] J. Abonyi, P. Arva, S. Nemeth, C. Vincze, B. Bodolai, Z. D. Horváth, G. Nagy, and M. Németh, "Operator support system for multi product processes-application to polyethylene production," in *European Symposium on Computer Aided Process Engineering*, Lappeenranta, Finland, 2003, pp. 347-352.

[80] Y. Li, "Building the data warehouse for materials selection in mechanical design," *Advanced Engineering Materials,* vol. 6, no. 1-2, pp. 92-95, 2004.

[81] M. R. Pokorny, D. G. B. Barber, P. A. Bush, J. H. Hise, W. S. M. Shun Hoo, C. E. Markham, J. R. Matheus, J. S. Mork, K. S. Nygaard, G. D. Shaffer, and J. A. Stambuk, "Integrating event-based production information with financial and purchasing systems in product manufacturing " U.S. Patent 20030154144, August 14, 2003.

[82] S. Daskalaki, I. Kopanas, M. Goudara, and N. Avouris, "Data mining for decision support on customer insolvency in telecommunications business," *European Journal of Operational Research,* vol. 145, no. 2, pp. 239-255, 2003.

[83] J. Faltys. (2000, March). *Rules-based software for telecommunications targeted marketing* [Online]. Available: http://www.dmreview.com/dmdirect/20000331/2110-1.html

[84] Y. Lian, R. H. Wolniewicz, and R. Dodier, "Predicting customer behavior in telecommunications," *Intelligent Systems,* vol. 19, no. 2, pp. 50-58, 2004.

[85] Q. Chen, M. Hsu, and U. Dayal, "A data-warehouse/OLAP framework for scalable telecommunication tandem traffic analysis," in *International Conference on Data Engineering*, San Diego, CA, 2000, pp. 201-210.

[86] R. Conine, "The data warehouse in the telecommunications industry," in *Network Operations and Management Symposium*, New Orleans, LA, 1998, pp. 205-209.

[87] D. Calvanese, L. Dragone, D. Nardi, R. Rosati, and S. M. Trisolini, "Enterprise modeling and data warehousing in Telecom Italia," *Information Systems,* vol. In Press, Corrected Proof, 2005.

[88] Y. Liang and M. Lanmann, "Integration of network management and network equipment inventory management," European Patent EP 1 251 656 B1, October 18, 2006.

[89] M. Reilly, "Design of computerized maintenance management system for radionuclide monitoring," Bachelor's thesis, University of Virginia, Charlottesville, VA, 2001.

[90] D. Mun, J. Hwang, S. Han, H. Seki, and J. Yang, "Sharing product data of nuclear power plants across their lifecycles by utilizing a neutral model," *Annals of Nuclear Energy,* vol. In Press, Corrected Proof.

[91] F. Estrella, Z. Kovacs, J. Le Goff, R. McClatchey, and I. Willers, "The design of an engineering data warehouse based on meta-object structures," in *Advances in Database Technologies*, Singapore, 1998, pp. 145-156.

[92] K. Keller, D. Wiegand, K. Swearingen, C. Reisig, S. Black, A. Gillis, and M. Vandernoot, "An architecture to implement integrated vehicle health management systems," in *IEEE AUTOTESTCON*, Valley Forge, PA, 2001, pp. 2-15.
[93] A. Woolley, "Collins class submarine systems analysis through data mining," in *International Conference on Health and Usage Monitoring*, Melbourne, Australia, 2003, pp. 17-26.
[94] A. Draper, "The operational benefits of health and usage monitoring systems in UK military helicopters," in *International Conference on Health and Usage Monitoring*, Melbourne, Australia, 2003, pp. 71-79.
[95] L. M. Keeney and R. Rhoads, "Embedded diagnostics and prognostics synchronization for army transformation," in *IEEE Aerospace Conference*, Big Sky, Montana, USA, 2004, pp. 3733-3741.
[96] R. J. Meseroll, C. J. Kirkos, and R. A. Shannon, "Data mining navy flight and maintenance data to affect repair," in *IEEE AUTOTESTCON*, Baltimore, MD, 2007, pp. 476-481.
[97] Central Intelligence Agency, *The World Factbook 2007*. Washington, DC, 2007.
[98] A. Pinsonneault and K. L. Kraemer, "Survey research methodology in management information systems: An assessment," *Journal of Management Information Systems,* vol. 10, no. 2, pp. 75-106, 1993.
[99] E. K. Foreman, *Survey Sampling Principles*. New York, NY: Marcel Dekker, 1991.
[100] M. Williams, "Generalization in interpretive research," in *Qualitative Research in Action*, T. May, Ed. London, UK: Sage Publications, 2002, pp. 125-143.
[101] Australian Bureau of Statistics, "Counts of Australian businesses, including entries and exists," Canberra, February 26 2007.
[102] Refsnes Data. (2007). *Browser Information* [Online]. Available: http://www.w3schools.com/browsers/default.asp
[103] Roper Starch and Pantone, "Consumer color preference study," Carlstadt, NJ, 1999.
[104] J. Lumsden, "Online-questionnaire design guidelines," in *Electronic surveys and measurements*, R. A. Reynolds, R. Woods, and J. D. Baker, Eds. Hershey, PA: Idea Group Reference, 2007, pp. 44-64.
[105] Q. Ma and M. McCord, "Web survey design," in *Electronic surveys and measurements*, R. A. Reynolds, R. Woods, and J. D. Baker, Eds. Hershey, PA: Idea Group Reference, 2007, pp. 9-18.
[106] D. Andrews, B. Nonnecke, and J. Preece, "Conducting research on the Internet: Online survey design, development and implementation guidelines," *International Journal of Human-Computer Interaction,* vol. 16, no. 2, pp. 185-210, 2003.
[107] D. R. Schaefer and D. A. Dillman, "Development of a standard e-mail methodology: Results of an experiment," *Public Opinion Quarterly,* vol. 62, no. 3, pp. 378-397, 1998.
[108] M. Lang, "Dual-mode electronic survey lessons and experiences," in *Electronic surveys and measurements*, R. A. Reynolds, R. Woods, and J. D. Baker, Eds. Hershey, PA: Idea Group Reference, 2007, pp. 65-75.
[109] eROI, "Email marketing statistics by day and time," Portland, OR, 5 July 2007.
[110] Y. Baruch, "Response rate in academic studies - A comparative analysis," *Human Relations,* vol. 52, no. 4, pp. 421-438, April 1999.
[111] C. Cook, F. Heath, and R. L. Thompson, "A meta-analysis of response rates in web- or Internet-based surveys," *Educational and Psychological Measurement,* vol. 60, no. 6, pp. 821-836, 2000.
[112] Educational Benchmarking Inc. (2005, November 18). *Determining an Acceptable Survey Response Rate* [Online]. Available: http://kb.webebi.com/article.aspx?id=10007&cNode=5K3B4O

[113] N. Mukherji, B. Rajagopalan, and M. Tanniru, "A decision support model for optimal timing of investments in information technology upgrades," *Decision Support Systems,* vol. 42, no. 3, pp. 1684-1696, December 2006.
[114] Department for Business Enterprise & Regulatory Reform. (2008, January 12). *Thresholds for small and medium-sized companies and groups* [Online]. Available: http://www.dti.gov.uk/bbf/financial-reporting/acc-audit-developments/page16361.html
[115] C. Ablitt and I. Partridge, "Experiences in implementing risk-based inspection," in *World Congress on Engineering Asset Management*, Harrogate, UK, 2007, pp. 56-64.
[116] N. Palmer, "A survey of business process initiatives," Business Process Trends, January 2007.
[117] A. E. Alter. (2006, July 19). *July 2006 survey: IT value, productivity metrics still not trustworthy* [Online]. Available: http://www.cioinsight.com/article2/0,1540,1991030,00.asp
[118] W. D. Wilde and P. A. Swatman, "Federal government policy and community objectives in regional telecommunications: A SISP-based study of Ballarat," *Journal of Theoretical and Applied Electronic Commerce Research,* vol. 1, no. 1, pp. 16-31, April 2006.
[119] D. Marco, *Building and managing the meta data repository, a full life-cycle guide*. New York, NY: John Wiley & Sons, 2000.
[120] R. Garland, "The mid-point on a rating scale: Is it desirable?," *Marketing Bulletin,* vol. 2, pp. 66-77, 1991.
[121] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd ed. New York, NY: McGraw-Hill, 1994.
[122] M. West, "ISO 15926 – Integration of lifecycle data," presented at the Upper Ontology Summit, Gaithersberg, MD, March, 2006.
[123] B. Smith, "Against idiosyncrasy in ontology development," in *Formal Ontology and Information Systems*, Baltimore, MD, 2006, pp. 15-26.
[124] P. Spyns, R. Meersman, and M. Jarrar, "Data modelling versus ontology engineering," *SIGMOD Record,* vol. 31, no. 4, pp. 12-17, December 2002.
[125] H. Teijgeler, "The process industries and the ISO 15926 semantic web," OntoConsult, Netherlands, 7 August 2007.
[126] OpenO&M for Manufacturing Joint Working Group, "Condition based operations for manufacturing," in *Maintenance Technology*, 2005.
[127] D. Tsichritzis and A. C. Klug, "The ANSI/X3/SPARC DBMS framework report of the study group on database management systems," *Information Systems,* vol. 3, no. 3, pp. 173-191, 1978.
[128] G. Simsion and G. Witt, *Data modelling essentials*, 3rd ed. San Francisco, CA: Morgan Kaufmann, 2004.
[129] J. R. Venable and J. C. Grundy, "Integrating and supporting entity relationship and object role models," in *Object-Oriented and Entity-Relationship Modelling*, Gold Coast, Australia, 1995.
[130] P. Chen, "The entity relationship model - Towards a unified view of data," *ACM Transactions on Database Systems,* vol. 1, no. 1, pp. 9-36, 1976.
[131] E. F. Codd, *The relational model for database management*. Reading, MA: Addison-Wesley, 1990.
[132] J. Martin, *Information Engineering*. Englewood Cliffs, NJ: Prentice Hall, 1990.
[133] T. A. Bruce, *Designing quality databases with IDEF1X information models*. New York, NY: Dorset House, 1992.
[134] R. Barker, *Case*Method: Entity relationship modelling*. Wokingham, England: Addison-Wesley, 1990.

[135] D. C. Hay, "Making data models readable," *Information Systems Management,* vol. 15, no. 1, pp. 21-33, January 1998.
[136] J. Arlow and I. Neustadt, *Enterprise patterns and MDA: Building better software with archetype patterns and UML*. Boston, MA: Addison-Wesley, 2003.
[137] Merriam-Webster, *Merriam-Webster's Collegiate Dictionary*, 11th ed. Springfield, MA: Merriam-Webster, 2003.
[138] D. C. Hay, *Data model patterns: Conventions of thought*. New York, NY: Dorset House, 1996.
[139] M. Fowler, *Analysis patterns: Reusable object models*. Reading, MA: Addison-Wesley, 1996.
[140] L. Silverston, *The data model resource book: A library of universal data models for all enterprises*, 2nd ed. New York, NY: John Wiley and Sons, 2001.
[141] P. Coad, *Object models: Strategies, patterns, & applications*, 2nd ed. Upper Saddle River, NJ: Prentice Hall PTR, 1997.
[142] J. Nicola, M. Mayfield, and M. Abney, *Streamlined object modeling: Patterns, rules, and implementation*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.
[143] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns: Elements of reusable object-oriented software*. Reading, MA: Addison-Wesley, 1994.
[144] A. Behm, A. Geppert, and K. R. Dittrich, "On the migration of relational schemas and data to object-oriented database systems," in *International Conference on Re-Technologies for Information Systems*, Klagenfurt, Austria, 1997, pp. 13-33.
[145] Applied Data Resource Management. *Data Environments* [Online]. Available: http://www.adrm.com/7_products.htm
[146] A. Koronios, D. Nastasie, V. Chanana, and A. Haider, "Integration through standards: An overview of international standards relevant to the integration of engineering asset management," in *World Congress on Engineering Asset Management*, Harrogate, UK, 2007, pp. 1066-1088.
[147] S. Garfinkel, D. Weise, and S. Strassmann, *The UNIX-HATERS Handbook*. San Mateo, CA: IDG Books, 1994.
[148] A. Johnston, "Enabling open standards-based interoperability for operations & maintenance people, processes and systems," MIMOSA, Tuscaloosa, AL, 2006.
[149] S. Mills, "Condition monitoring and diagnostics of machines: Standards in progress," ISO, 2002.
[150] Wolfson Maintenance, "Guidelines for types and locations of measurements & ISO standards references," Manchester, UK, 1998.
[151] ISO, "Information technology - Security techniques - Information security management systems - Overview and vocabulary," in *ISO 27000*, 2008.
[152] M. E. Cline, "Mapping solutions under $500," in *ONLINE*. vol. 29, 2005, pp. 27-30.
[153] H. Liang, Y. Xue, W. R. Boulton, and T. A. Byrd, "Why Western vendors don't dominate China's ERP market," in *Communications of the ACM*. vol. 47, 2004, pp. 69-72.
[154] T. O'Hanlon, "Computerized maintenance management and enterprise asset management best practices," Reliabilityweb.com, Fort Myers, FL, 2005.
[155] S. Wooldridge. (2007). *MyCitect News Mar 2007 Oceania* [Online]. Available: http://www.citect.com/index.php?option=com_content&task=view&id=422&Itemid=458
[156] Y. Sun, L. Ma, J. Mathew, W. Wang, and S. Zhang, "Mechanical systems hazard estimation using condition monitoring," *Mechanical Systems and Signal Processing,* vol. 20, no. 5, pp. 1189-1201, July 2006.
[157] P. Terenziani, R. T. Snodgrass, A. Bottrighi, M. Torchio, and G. Molino, "Extending temporal databases to deal with Telic/Atelic medical data," *Artificial Intelligence in Medicine,* vol. 3581, pp. 58-66, 2005.

[158] M. Staudt, A. Vaduva, and T. Vetterli, "Metadata management and data warehousing," University of Zurich, Zurich, Switzerland, 1999.
[159] J. Poole, D. Chang, D. Tolbert, and D. Mellor, *The Common Warehouse Metamodel: An introduction to the standard for data warehouse integration*: Wiley, 2001.
[160] S. J. Graves, T. H. Hinke, and S. Kansal, "Metadata: The golden nuggets of mining data," presented at the IEEE Metadata Conference, Silver Spring, MD, April, 1996.
[161] G. Everest, "Basic data structure models explained with a common example," in *Fifth Texas Conference on Computing Systems*, Austin, TX, 1976, pp. 39-45.
[162] Business Rules Group, "The business motivation model - Business governance in a volatile world," September 2007.
[163] Standards Australia, *Risk management - AS/NZS 4360*, 3rd ed. Sydney, Australia: Standards Australia International, 2004.
[164] M. Modarres, *What every engineer should know about reliability and risk analysis* vol. 30. New York, NY: Marcel Dekker, 1992.
[165] Software Engineering Institute, "Capability Maturity Model® Integration v1.2," Pittsburgh, PA, 2007.
[166] O. Balci, "Principles of simulation model validation, verification, and testing," *Transactions of the Society for Computer Simulation International,* vol. 14, no. 1, pp. 3-12, March 1997.
[167] K. Qi and Z. He, "Data management in monitoring and diagnosis system for electromechanical equipment based on LabVIEW," *Computer Measurement & Control,* vol. 13, no. 1, pp. 24-30, 2005.
[168] G. Colliat, "OLAP, relational, and multidimensional database systems," *ACM SIGMOD Record,* vol. 25, no. 3, pp. 64-69, 1996.
[169] D. Schuff, K. Corral, and O. Turetken, "Comparing the effect of alternative data warehouse schemas on end user comprehension level," in *SIGDSS Pre-ICIS Workshop on Decision Support Systems*, Las Vegas, NV, 2005.
[170] L. Cabibbo and R. Torlone, "A logical approach to multidimensional databases," in *International Conference on Extending Database Technology: Advances in Database Technology*, Valencia, Spain, 1998.
[171] D. L. Moody and M. A. R. Kortink, "From enterprise models to dimensional models: A methodology for data warehouse and data mart design," in *Design and Management of Data Warehouses*, Stockholm, Sweden, 2000.
[172] A. Marotta and R. Ruggia, "Data warehouse design: A schema-transformation approach," in *International Conference of the Chilean Computer Science Society*, Copiapo, Chile, 2002, pp. 153-161.
[173] L. Palopoli, G. Terracina, and D. Ursino, "Experiences using DIKE, a system for supporting cooperative information system and data warehouse design," *Information Systems,* vol. 28, no. 7, pp. 835-865, 2003.
[174] Y. T. Chen and P. Y. Hsu, "An efficient and grain preservation mapping algorithm: From ER diagram to multidimensional model," in *International School and Symposium: Advanced Distributed Systems*, Guadalajara, Mexico, 2005, pp. 331-346.
[175] OLAP Council. (1998, November). *APB-1 OLAP Benchmark Release II* [Online]. Available: http://www.olapcouncil.org/research/APB1R2_spec.pdf
[176] Transaction Processing Performance Council. (2007). *TPC Benchmarks* [Online]. Available: http://www.tpc.org/information/benchmarks.asp
[177] MIMOSA, "OSA-EAI Terminology Dictionary V3.0," 2006.
[178] P. M. Nadkarni, L. Marenco, R. Chen, E. Skoufos, G. Shepherd, and P. Miller, "Organization of heterogeneous scientific data using the EAV/CR representation," *Journal of the American Medical Informatics Association,* vol. 6, no. 6, pp. 478-493, 1999.

[179] S. Zhang, A. Mathew, L. Ma, M. Hodkiewicz, and J. Mathew, "Integration of pump/motor SCADA data into engineering asset management," in *World Congress on Engineering Asset Management*, Gold Coast, Queensland, Australia, 2006.

[180] C. Calero, M. Piattini, C. Pascual, and M. A. Serrano, "Towards data warehouse quality metrics," in *International Workshop on Design and Management of Data Warehouses*, Interlaken, Switzerland, 2001, pp. 2:1-10.

[181] M. Jarke, M. A. Jeusfeld, C. Quix, and P. Vassiliadis, "Architecture and quality in data warehouses: An extended repository approach," *Information Systems,* vol. 24, no. 3, pp. 229-253, 1999.

[182] D. L. Moody, "Metrics for evaluating the quality of entity relationship models," in *International Conference on Conceptual Modeling*, Singapore, 1998, pp. 211-225.

[183] M. H. Halstead, *Elements of software science (operating and programming systems series)* vol. 7. New York, NY: Elsevier, 1977.

[184] T. J. McCabe, "A complexity measure," *IEEE Transactions on Software Engineering,* vol. SE-2, no. 4, pp. 308-320, 1976.

[185] K. D. Welker and P. W. Oman, "Software maintainability metrics models in practice," *Crosstalk, Journal of Defense Software Engineering,* vol. 8, no. 11, pp. 19-23, 1995.

[186] J. Han, N. Stefanovic, and K. Koperski, "Selective materialization: An efficient method for spatial data cube construction," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Melbourne, Australia, 1998, pp. 144-158.

[187] T. Martyn, "Reconsidering multi-dimensional schemas," *ACM SIGMOD Record,* vol. 33, no. 1, pp. 83-88, March 2004.

[188] R. Kimball, *The data warehouse toolkit: The complete guide to dimensional modeling*. New York, NY: John Wiley & Sons, 2002.

[189] D. Vesset and B. McDonough, "Worldwide data warehouse platform tools 2006 vendor shares," International Data Corporation, Framingham, MA, July 2007.

[190] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *Artificial Intelligence Communications,* vol. 7, no. 1, pp. 39-59, 1994.

[191] Y. K. Paek, J. Seo, and G. C. Kim, "An expert system with case-based reasoning for database schema design," *Decision Support Systems,* vol. 18, no. 1, pp. 83-95, 1996.

[192] V. C. Storey, R. H. L. Chiang, D. Dey, R. C. Goldstein, and S. Sudaresan, "Database design with common sense business reasoning and learning," *ACM Transactions on Database Systems,* vol. 22, no. 4, pp. 471–512, 1997.

[193] J. Choobineh and A. W. Lo, "CABSYDD: Case-based system for database design," *Journal of Management Information Systems,* vol. 21, no. 3, pp. 281-314, 2004.

[194] W. Inmon, "What is a data warehouse?," *Prism Tech Topic,* vol. 1, no. 1, 1995.

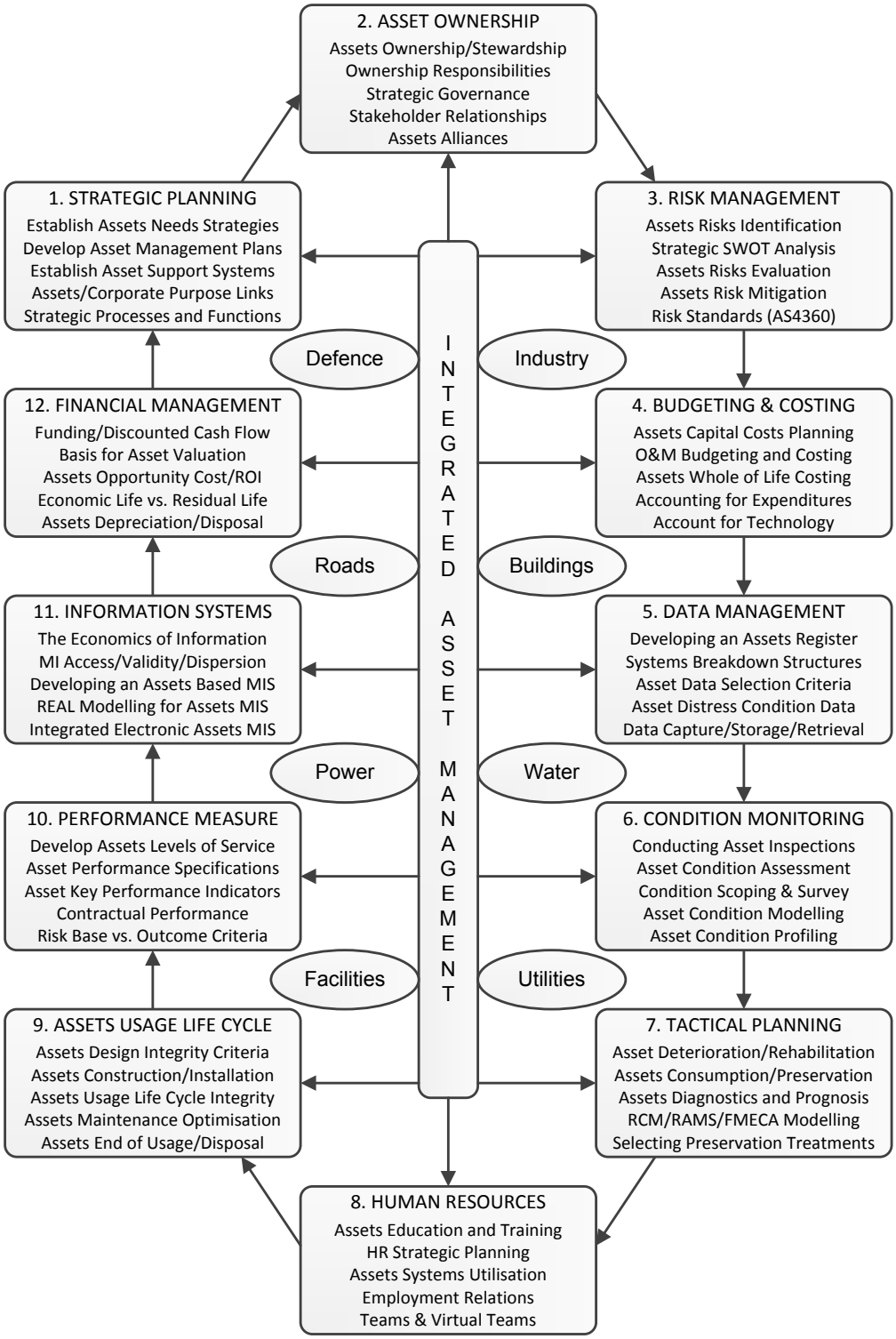
[195] L. Palopoli, G. Terracina, and D. Ursino, "DIKE: A system supporting the semi-automatic construction of cooperative information systems from heterogeneous databases," *Software: Practice and Experience,* vol. 33, no. 9, pp. 847-884, 2003.

[196] C. Kaldeich and J. O. e. Sá, "Data warehouse methodology: A process driven approach," in *Advanced Information Systems Engineering*, Riga, Latvia, 2004, pp. 536-549.

[197] K. Hahn, C. Sapia, and M. Blaschka, "Automatically generating OLAP schemata from conceptual graphical models," in *ACM International Workshop on Data Warehousing and OLAP*, McLean, Virginia, USA, 2000.

[198] M. Golfarelli, S. Rizzi, and B. Vrdoljak, "Data warehouse design from XML sources," in *ACM International Workshop on Data Warehousing and OLAP*, Atlanta, Georgia, USA, 2001.
[199] W. W. Eckerson, "Three tier client/server architecture: Achieving scalability, performance, and efficiency in client server applications," *Open Information Systems 10,* vol. 1, no. 3(20), pp. 1-12, January 1995.
[200] J. L. Kolodner, *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
[201] R. S. Seiner, "A conceptual meta-model for unstructured data," in *The Data Administration Newsletter*, Pittsburgh, Pennsylvania, USA, 2003.
[202] S. Borschiver, P. Wongtschowski, and A. Antunes, "A classificação industrial e sua importância na análise setorial," *Ciência da Informação,* vol. 33, no. 1, pp. 9-21, 2004.
[203] H. Gust, "Representing word meanings," in *Text Understanding in LILOG: Integrating Computational Linguistics and Artificial Intelligence*, 546 ed, O. Herzog and C.-R. Rollinger, Eds.: Springer-Verlag GmbH, 1991, pp. 127-142.
[204] H. M. Townley and R. D. Gee, *Thesaurus-making: Grow your own word-stock*. London: Andre Deutsch, 1980.
[205] C. K. Riesbeck and R. C. Schank, *Inside case-based reasoning*. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates, 1989.
[206] I. Watson, "Case-based reasoning is a methodology not a technology," *Knowledge-Based Systems,* vol. 12, no. 5-6, pp. 303-308, 1999.
[207] W. F. Tichy, "Should computer scientists experiment more?," *Computer,* vol. 31, no. 5, pp. 32-40, May 1998.
[208] D. A. Waterman, *A guide to expert systems*. Boston, MA: Addison-Wesley, 1985.
[209] J. Cohen, *Statistical power and analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
[210] Y. Sun, L. Ma, and J. Mathew, "Asset management processes: Modelling, evaluation and integration," in *World Congress on Engineering Asset Management*, Harrogate, UK, 2007, pp. 1847-1856.

Appendix A CIEAM IAM Framework



Appendix B Literature Analysis

Paper Metadata

Author/s	Year	Topic	Paper type	Country	Applicable industry
Knights & Daneshmend	1997	Open systems standards for the mining industry	Conference	Chile	Resources
Conine	1998	The data warehouse in the telecommunications industry	Conference	USA	Telecommunications
Draber, Gelle, Kostic, Preiss & Schluchter	2000	How operation data helps manage lifecycle costs	Conference	Switzerland	Power utilities
Shi, Lee, Duan & Wu	2001	Power system data warehouses	Journal	China	Power utilities
Werner & Hermansson	2002	Integrated utility data warehousing-a prerequisite to keep up with competition on electricity markets	Conference	Sweden	Power utilities
McDonald	2003	Substation automation-IED integration and availability of information	Journal	USA	Power utilities
Li	2004	Building the data warehouse for materials selection in mechanical design	Journal	USA	Manufacturing
He, Wang & Zhao	2005	Research on the SCADA-EMS system data warehouse technology	Conference	China	Power utilities
Nimmagadda, Dreher & Rudra	2005	Data warehouse structuring methodologies for efficient mining of Western Australian petroleum data sources	Conference	Australia	Resources
Rudra & Nimmagadda	2005	Roles of multidimensionality and granularity in warehousing Australian resources data	Conference	Australia	Resources
Thomas, Nanda & Ali	2006	Development of a data warehouse for non-operational data in power utilities	Conference	India	Power utilities
Nimmagadda & Dreher	2006	Mapping and modeling of oil and gas relational data objects for warehouse development and efficient data mining	Conference	Australia	Resources

Paper Analysis

Author/s	Year	Data warehousing areas	Asset management areas	Validation method	Innovation and significance
Knights & Daneshmend	1997	Conceptual architectural design	Planning, blast design, maintenance, production control	None	Identification of asset management data warehousing applicability
Conine	1998	High level data area requirements analysis	Network requirements forecasting, inventory management and material logistics, transport capacity delivery, project management, network asset utilisation, excess capacity analysis	None	Identification of asset management data warehousing applicability
Draber, Gelle, Kostic, Preiss & Schluchter	2000	High level data area and function requirements analysis User interface Architecture Data sources	Life cycle costs, availability, network analysis, power quality, asset supervision, risk management	Detailed use case	Inter-discipline data integration
Shi, Lee, Duan & Wu	2001	Potential applications	SCADA, EMS, and DMS operational data	None	None
Werner & Hermansson	2002	Detailed conceptual architectural design Data compression and I/O performance Potential applications	SCADA operational data	None	Discussion of data compression and I/O performance for data warehousing
McDonald	2003	Conceptual architectural design Integration	SCADA operational data Standards Communication protocols	None	Relation of standards to integration
Li	2004	Medium level architectural design	Material attributes	None	None

Author/s	Year	Data warehousing areas	Asset management areas	Validation method	Innovation and significance
He, Wang & Zhao	2005	ETL tool design High level architectural design Potential applications	SCADA, EMS, and DMS operational data	Implementation	None
Nimmagadda, Dreher & Rudra	2005	ER and multidimensional schemas Cubes and OLAP	Petroleum surveys, permits	Case study	Proposing multidimensional schemas
Rudra & Nimmagadda	2005	Multidimensional schemas Granularity	Petroleum surveys, permits	Case study	Comparison of ER and multidimensional data sizes
Thomas, Nanda & Ali	2006	Conceptual architectural design Integration	SCADA, and EMS operational data Asset attributes Environmental conditions	Detailed implementation	Detailed implementation
Nimmagadda & Dreher	2006	Multidimensional schemas Object models	Petroleum surveys, permits	None	None

Appendix C Information Systems Survey

Questionnaire

INTRODUCTION

This survey aims to assist organisations in improving their information system infrastructure for their asset management.

We are seeking responses from people who have knowledge on the various information systems at their organisation. The survey should take less than 10 minutes to complete. As some questions are dependent upon others, please answer all questions in order. A summary of the results will be sent to all participants once the data is analysed.

Please be assured that all information you provide is securely stored and will remain strictly confidential. Research findings will be presented in aggregate form and no individuals or organisations will be identifiable.

We greatly appreciate your time and valuable contribution to this research.

For any enquiries regarding this survey, please contact either:

Avin Mathew, Phone: +61 7 3138 9156
or
Associate Prof. Lin Ma

YOUR ORGANISATION

If your organisation is an independent division of a larger organisation, answer questions on behalf of your division.

1. Where is your organisation located?

- ☐ Australia
- ☐ Europe
- ☐ North America
- ☐ South America
- ☐ Asia
- ☐ Africa

2. What is the primary sector for your organisation?

- ☐ Built Environment
- ☐ Communications
- ☐ Defence
- ☐ Government
- ☐ Heavy Manufacturing
- ☐ Light Manufacturing
- ☐ Mining
- ☐ Power Utilities
- ☐ Water Utilities
- ☐ Transportation and Infrastructure
- ☐ Professional Services and Consulting
- ☐ Other: _____

3. How many full time equivalent staff are employed by your organisation?

- ☐ Less than 50 people
- ☐ Between 50 and 100 people
- ☐ Between 100 and 500 people
- ☐ Between 500 and 1000 people
- ☐ Between 1,000 and 5,000 people
- ☐ Between 5,000 and 10,000 people
- ☐ Between 10,000 and 25,000 people
- ☐ Greater than 25,000 people
- ☐ Unknown

4. For the past financial year, what was the turnover of your organisation?

- ☐ Less than $5 million
- ☐ Between $5 million and $10 million
- ☐ Between $10 million and $50 million
- ☐ Between $50 million and $100 million
- ☐ Between $100 million and $500 million
- ☐ Between $500 million and $1 billion
- ☐ Between $1 billion and $10 billion
- ☐ Greater than $10 billion
- ☐ Unknown

5. For the past financial year, what was the total budget for *asset management information systems* projects?

- ☐ Less than $100,000
- ☐ Between $100,000 and $250,000
- ☐ Between $250,000 and $500,000
- ☐ Between $500,000 and $1 million
- ☐ Between $1 million and $5 million
- ☐ Between $5 million and $10 million
- ☐ Between $10 million and $50 million
- ☐ Greater than $50 million
- ☐ Unknown

INFORMATION SYSTEMS

6. Does your organisation use information systems for these areas?

	Yes	No	Unknown
Equipment and Product Data Management *Includes asset registry, specifications, breakdown structure, bill of material, documentation, etc.*	☐	☐	☐
Life Cycle Costing *Includes tracking revenues and costs during the life of assets, projects, etc.*	☐	☐	☐
Operation Scheduling / Process Control *Includes scheduling, utilisation, performance sensor data, control codes, etc.*	☐	☐	☐
Materials Management *Includes standards, acquisition, quality control, flow, etc. of inventory.*	☐	☐	☐
Maintenance and Work Management *Includes inspections, work orders, breakdown, repairs, etc.*	☐	☐	☐
Condition Monitoring *Includes measuring equipment condition through vibration, acoustics, ultrasound, oil analysis, etc.*	☐	☐	☐
Risk and Reliability Management *Includes risk registry, warranties, insurance, FMECA, reliability block diagrams, etc.*	☐	☐	☐

7. For the areas answered with a 'Yes' in Question 6, what is the greatest benefit provided by the systems?

	Improving data entry and quality	Data processing and reporting	Automating repetitive tasks	Improving business procedures	Satisfying regulatory compliance	Other
Equipment and Product Data Management	☐	☐	☐	☐	☐	☐
Life Cycle Costing	☐	☐	☐	☐	☐	☐
Operation Scheduling / Process Control	☐	☐	☐	☐	☐	☐
Materials Management	☐	☐	☐	☐	☐	☐
Maintenance and Work Management	☐	☐	☐	☐	☐	☐
Condition Monitoring	☐	☐	☐	☐	☐	☐
Risk and Reliability Management	☐	☐	☐	☐	☐	☐

8. For the areas answered with a 'No' or 'Unknown' in Question 6, what is the primary reason for not using systems?

	No benefit or business case	Unable to find project sponsor	No suitable vendor or product	Unable to successfully use system	Function provided using another method	Other
Equipment and Product Data Management	☐	☐	☐	☐	☐	☐
Life Cycle Costing	☐	☐	☐	☐	☐	☐
Operation Scheduling / Process Control	☐	☐	☐	☐	☐	☐
Materials Management	☐	☐	☐	☐	☐	☐
Maintenance and Work Management	☐	☐	☐	☐	☐	☐
Condition Monitoring	☐	☐	☐	☐	☐	☐
Risk and Reliability Management	☐	☐	☐	☐	☐	☐

9. When do you measure the Return on Investment (ROI) for your asset management information systems?

- ☐ ROI is not measured
- ☐ ROI is measured in advance as part of the cost justification
- ☐ ROI is measured after an initial pilot
- ☐ ROI is measured upon final roll-out
- ☐ ROI measurement is done periodically as part of an ongoing assessment
- ☐ Unknown

10. In your opinion, what are the *top three* items you would like to see in next generation asset management information systems?

- ☐ Faster and more responsive systems
- ☐ Easier to customize
- ☐ Easier to transition between vendors
- ☐ Easier integration with other systems
- ☐ More user friendly Graphical User Interfaces
- ☐ Easier to produce reports
- ☐ Better documentation
- ☐ Better support for decision making

Others: _____

DATA WAREHOUSING

A data warehouse is a database that integrates cleansed data from different systems for the purpose of analysis/reporting.

11. What is your situation regarding using a data warehouse in your asset management?

- ☐ We are using a data warehouse for asset management - *Proceed to Question 12*
- ☐ We are developing a data warehouse that will be used in our asset management - *Proceed to Question 12*
- ☐ We stopped using our data warehouse - *Proceed to Question 14*
- ☐ We have never had a data warehouse - *Proceed to Question 15*
- ☐ Unknown
- ☐ Other: _____

If you answered 'Unknown' or 'Other', proceed to Question 16

12. At present, information from which asset management areas are loaded into your data warehouse? Tick all that apply.

- ☐ Life Cycle Costing
- ☐ Operation Scheduling / Process Control
- ☐ Materials Management
- ☐ Maintenance and Work Management
- ☐ Condition Monitoring
- ☐ Risk and Reliability Management

13. What were the main justifications for developing the data warehouse? Tick all that apply.

- ☐ Improved quality of data
- ☐ Single source of data
- ☐ Easier and faster access to data
- ☐ Enhanced reporting and business intelligence
- ☐ Simplifying querying data from multiple sources
- ☐ Tracking historical changes in data
- ☐ Unknown

Others: _____

Proceed to Question 16

14. What were the main reasons for stopping the data warehouse? Tick all that apply.

- ☐ The maintenance costs were too high
- ☐ We did not have suitably skilled personnel to operate and manage the warehouse
- ☐ We did not have the right infrastructure to support the warehouse
- ☐ The data warehouse did not provide sufficient benefit
- ☐ Unknown

Others: _____

Proceed to Question 16

15. What are the main reasons for not having a data warehouse? Tick all that apply.

- ☐ The cost is prohibitive
- ☐ We do not have suitably skilled personnel to operate and manage the warehouse
- ☐ We do not have the right infrastructure to support the warehouse
- ☐ There are technology issues
- ☐ There is no added benefit perceived
- ☐ Unknown

Others: _____

INTEGRATION

16. What is your level of information system integration for automated transfer of data between your systems?

- ☐ All our asset management information systems are integrated
- ☐ Some of our asset management information systems are integrated
- ☐ None of our asset management information systems are integrated
- ☐ Unknown

If you answered 'None' or 'Unknown', proceed to Question 19

17. If you have integrated asset management systems, specify the method of integration below. Tick all that apply.

- ☐ Integration is supported directly by our information systems
- ☐ Through the purchase of additional commercial software / service
- ☐ Through in-house development of a data/communications bridge
- ☐ Unknown

Other: _____

18. If you have integrated asset management systems, what is your situation regarding using asset management standards (e.g. OPC, STEP, MIMOSA) for integration?

- ☐ Standards were not considered
- ☐ Standards were considered, but we decided not to pursue it
- ☐ We are currently using standards for integration
- ☐ Unknown

Other: _____

DISCARDING DATA

20. Does your organisation routinely discard any of the following asset management data from your information systems?

	Yes	No	Unknown
Equipment and Product Data Management	☐	☐	☐
Life Cycle Costing	☐	☐	☐
Operation Scheduling / Process Control	☐	☐	☐
Materials Management	☐	☐	☐
Maintenance and Work Management	☐	☐	☐
Condition Monitoring	☐	☐	☐
Risk and Reliability Management	☐	☐	☐

If you answered 'No' or 'Unknown' to all areas, proceed to Question 22

21. What are the reasons for discarding your sensor data? Tick all that apply.

- ☐ Insufficient storage space
- ☐ Increased performance of systems after discarding
- ☐ Reduced data maintenance costs
- ☐ Usefulness of data has declined
- ☐ Uncertain of usefulness of data
- ☐ Unknown

Other: _____

RATING

22. How would you rate your organisation's overall data and information system management?

- ☐ Very Poor
- ☐ Poor
- ☐ Below Average
- ☐ Average
- ☐ Good
- ☐ Very Good

YOUR DETAILS

Providing your details will give us a better indication on your approach to the survey.

What is your primary job area?

For example, 'Accounting', 'Asset management', 'Human resources', 'Information technology', 'Marketing', 'Production'

_____

If you would like to be notified of the results of the survey, please enter your email address and organisation.

Email address: _____

Organisation name: _____

THANK YOU

Thank you for participating in the Information Systems in Asset Management Survey. If you have supplied a contact email, we will be in contact with you shortly.

GroupMail Template

Subject: Survey on Information Systems in Asset Management

Dear *!*FIRST NAME*!*,

I am a CIEAM PhD student from Queensland University of Technology under the supervision of Associate Professor Lin Ma. She advised me that you are knowledgeable in asset management and that your input would be very valuable to our research. We are running a survey that investigates how different industries use information systems in their asset management operations.

The survey can be accessed online at: *<WLINK(!*URL*!)ENDWLINK>*

It should take less than 10 minutes of your time to complete. If you think someone else in *!*ORGANISATION*!* may be better positioned to answer the survey, feel free to forward the survey to them. All responses will be kept confidential and we will be sending all participants a summary report of the findings.

If you have any questions or comments, please do not hesitate to contact me. We thank you very much for your time and valuable contribution to this research.

Cheers,
Avin Mathew

PhD Student
CRC for Integrated Engineering Asset Management
Queensland University of Technology
Email: a.mathew@qut.edu.au
Phone: (61 7) 3138 9156

Appendix D Case Study Comparisons

Case Study 2

Comparison with ID206 specifications:

ID206 Specifications	Relation to conceptual data model
Asset information	
Design and manufacture	Asset specifications
Asset hierarchy	Asset relationships
Characteristic information and asset category	
Operating parameters and measurement location list	Measured attributes, Measurements
Installation	Asset characteristics, Activities
Standard operation	Activities
Standard maintenance	Activities
Standard safety management	Risks
Responsible role	Agent relationships
Operation	
Shift and log management	Documents
Operation management	
Operation simulation	Hypothetical events
Electrical locking control	Hypothetical events
System working layout chart management	Segments
Operation management based on workflow	Activities
Routine work	
Routine work initialization (configuration)	Activities
Alarming	Alarms
Execution	Activities
Analysis and Assessment	Activities
Experiment management	Measurements
Routine inspection	
Definition	Activities
Trigger	Alarms
Feedback data record	Measurements
Execution and confirmation	Activities
Abnormal data analysis	Measurements
Operation guidance	

Real-time operation guidance	Activities
Off-line operation guidance	Activities
Operation analysis	
System performance	Measurements
Personnel performance (link to payroll)	Agents, Activities
Asset maintenance	
Unplanned maintenance (failure-based maintenance)	
Failure report	Documents
Failure analysis and maintenance measurement	Measurements
Failure processing	Measurements
Assessment of failure-fixing result	Measurements
Summary report (link to knowledge-base)	Documents
Planned maintenance (preventive maintenance)	
Configuration for plant asset management	Asset characteristics, Segments
Planned maintenance process	Hypothetical activities
Planned maintenance assessment	Hypothetical activities
Condition-based maintenance (predictive maintenance)	
Configuration	Asset characteristics
Condition based maintenance	Measurements
Maintenance work order management	Activities
Disposal	Activities
Reason for disposal	Ends
Disposal procedure (work flow)	Activities
All information related to this disposal	Documents
Inventory Management	
Configuration for fundamental information (store)	Segments, Documents
Inventory Requirement Management	
Project	Activities
Work order	Activities
Individual requirement planning	Ends
Inventory Purchasing Plan Management	Means
Inventory Arrival management	Means
Inventory transfer into the storeroom	Activities
Inventory distribution	Activities
Un-used inventory return	Activities
Stock checking and summarization report (monthly)	Segments, Documents
Stock level reminding	Segments
Stock analysis	Assets
Supplier management	Agents
Connection with financial system	Financial accounts

Reliability and Safety	
Reliability and safety specification	Documents
Safety	Risks
Safety Monitoring	Risks, Activities
Reliability monitoring	Risks, Activities
Asset reliability assessment	Risks
Safety assessment	Risks
Project	
Standard project definition	Activities, Documents
Project management	Activities
Measures for un-satisfied technical aspects (standard)	Risks, Activities
Contract management	Contacts
Performance management	
Performance indicator definition	Documents
Objective of each performance indicator	Ends
Performance analysis based on performance indicator	
Financial analysis	Financial accounts
Reliability analysis	Risks
Safety analysis	Risks
Asset health condition analysis	Measurements

Case Study 3

Comparison with World In One Technology Asset Management System:

System Specifications	Relation to conceptual data model
Users	
User access rights	Metadata
User login	Metadata
Asset register equipment/component	
Data storage	N/A
General	Model specifications, Asset capabilities
Configuration	Asset relationships
Maintenance	Hypothetical activities
Meters	Measurements
Financial	Financial accounts
Status	Measurement regions
Warranty	Warranty contracts
Statistics	Measurements

User defined fields	Asset capabilities
Business unit	
Data storage	N/A
General	Agents
Configuration	Agent relationships
Meters	Measurements
Business unit financial	Financial accounts
Status	Events
Statistics	Statistics
User defined fields	Agent attributes
Reliability & maintainability statistics	
Availability	Events, Activities, Measurements
Utilisation	
MTBF	
MTTR	
MTBR	
OEE	Events, Activities, Hypothetical activities, Segment relationships, Measurements
Meters	
Calculated readings	Measurements
Automated meter readings	Events, Measurements
Meter configuration form	Measurements
Meter data entry form	Measurements
Meter reading corrections	Measurements
Meter replacement	Activities, Asset capabilities
Administration	
Asset group definition	Asset relationships
Asset status definition	Events Activities
Asset templates	Model specifications
Chart of accounts	Financial accounts
Currencies	Units of measurement
Units of measure	Units of measurement
Departments	Agents
Employees	Agents
Teams	Agents
Manufacturer	Agents
Supplier	Agents

Appendix E SQL Queries

Loosely Constrained Joins

Entity Relationship

```
SELECT SUM(duration), SUM(data_value)
FROM (SELECT (DATEDIFF(dd, gmt_started, gmt_completed) + 1) duration,
SUM(data_value) data_value
FROM work_order, sg_completed_work, work_manage_type, wo_step_num_data,
work_num_data_type
WHERE work_order.work_order_db_site = sg_completed_work.work_order_db_site
AND work_order.work_order_db_id = sg_completed_work.work_order_db_id
AND work_order.work_order_id = sg_completed_work.work_order_id
AND work_order.wm_type_db_site = work_manage_type.wm_type_db_site
AND work_order.wm_type_db_id = work_manage_type.wm_type_db_id
AND work_order.wm_type_code = work_manage_type.wm_type_code
AND sg_completed_work.work_order_db_site =
wo_step_num_data.work_order_db_site
AND sg_completed_work.work_order_db_id = wo_step_num_data.work_order_db_id
AND sg_completed_work.work_order_id = wo_step_num_data.work_order_id
AND sg_completed_work.work_ord_step_seq =
wo_step_num_data.work_ord_step_seq
AND wo_step_num_data.wn_db_site = work_num_data_type.wn_db_site
AND wo_step_num_data.wn_db_id = work_num_data_type.wn_db_id
AND wo_step_num_data.wn_type_code = work_num_data_type.wn_type_code
AND (work_num_data_type.name = 'Cost, Parts, Actual'
OR work_num_data_type.name = 'Cost, Consummables, Actual'
OR work_num_data_type.name = 'Cost, Internal Labor, Actual'
OR work_num_data_type.name = 'Cost, External Labor, Actual'
OR work_num_data_type.name = 'Cost, Extra Expenses, Actual')
AND DATEPART(yy, gmt_completed) = '2005'
AND work_manage_type.name LIKE 'Maintenance%'
GROUP BY gmt_started, gmt_completed) sq
```

Star Schema

```
SELECT SUM(ActualDuration), SUM(TotalCost)
FROM FactWork, DimTime, DimWork
WHERE FactWork.ActualEndTimeKey = DimTime.TimeKey
AND FactWork.WorkKey = DimWork.WorkKey
AND DimTime.Year = '2005'
AND DimWork.WorkManagementType LIKE 'Maintenance%'
```

Tightly Constrained Joins

Entity Relationship

```
SELECT SUM(duration), SUM(data_value)
FROM (SELECT work_order.work_order_id, (DATEDIFF(dd, gmt_started,
gmt_completed) + 1) duration, SUM(data_value) data_value
FROM work_order, sg_completed_work, segment, asset_on_segment, asset,
asset_type, work_manage_type, wo_step_num_data, work_num_data_type
WHERE work_order.work_order_db_site = sg_completed_work.work_order_db_site
AND work_order.work_order_db_id = sg_completed_work.work_order_db_id
AND work_order.work_order_id = sg_completed_work.work_order_id
AND sg_completed_work.segment_site = segment.segment_site
AND sg_completed_work.segment_id = segment.segment_id
AND segment.segment_site = asset_on_segment.segment_site
AND segment.segment_id = asset_on_segment.segment_id
AND asset_on_segment.asset_org_site = asset.asset_org_site
AND asset_on_segment.asset_id = asset.asset_id
AND asset.as_db_site = asset_type.as_db_site
AND asset.as_db_id = asset_type.as_db_id
AND asset.as_type_code = asset_type.as_type_code
AND work_order.wm_type_db_site = work_manage_type.wm_type_db_site
AND work_order.wm_type_db_id = work_manage_type.wm_type_db_id
AND work_order.wm_type_code = work_manage_type.wm_type_code
AND sg_completed_work.work_order_db_site =
wo_step_num_data.work_order_db_site
AND sg_completed_work.work_order_db_id = wo_step_num_data.work_order_db_id
AND sg_completed_work.work_order_id = wo_step_num_data.work_order_id
AND sg_completed_work.work_ord_step_seq =
wo_step_num_data.work_ord_step_seq
AND wo_step_num_data.wn_db_site = work_num_data_type.wn_db_site
AND wo_step_num_data.wn_db_id = work_num_data_type.wn_db_id
AND wo_step_num_data.wn_type_code = work_num_data_type.wn_type_code
AND (work_num_data_type.name = 'Cost, Parts, Actual'
OR work_num_data_type.name = 'Cost, Consummables, Actual'
OR work_num_data_type.name = 'Cost, Internal Labor, Actual'
OR work_num_data_type.name = 'Cost, External Labor, Actual'
OR work_num_data_type.name = 'Cost, Extra Expenses, Actual')
AND asset_type.name LIKE 'Pump%'
AND segment.name LIKE 'Segment%'
AND DATEPART(yy, gmt_completed) = '2005'
AND work_manage_type.name LIKE 'Maintenance%'
GROUP BY work_order.work_order_id, gmt_started, gmt_completed) sq
```

Star Schema

```
USE OSAEAISimulateDW;
SELECT SUM(ActualDuration), SUM(TotalCost)
FROM FactWork, DimAsset, DimSegment, DimTime, DimWork
WHERE FactWork.AssetKey = DimAsset.AssetKey
AND FactWork.LocationSegmentKey = DimSegment.SegmentKey
AND FactWork.ActualEndTimeKey = DimTime.TimeKey
AND FactWork.WorkKey = DimWork.WorkKey
AND DimAsset.Type LIKE 'Pump%'
AND DimSegment.Name LIKE 'Segment%'
AND DimTime.Year = '2005'
AND DimWork.WorkManagementType LIKE 'Maintenance%'
```

Calculation

Entity Relationship

```
SELECT AVG(DATEDIFF(dd, gmt_started, start_after_gmt)), AVG(DATEDIFF(dd,
gmt_completed, end_after_gmt))
FROM work_order_step, sg_completed_work
WHERE work_order_step.work_order_db_site =
sg_completed_work.work_order_db_site
AND work_order_step.work_order_db_id = sg_completed_work.work_order_db_id
AND work_order_step.work_order_id = sg_completed_work.work_order_id
AND work_order_step.work_ord_step_seq =
sg_completed_work.work_ord_step_seq
```

Star Schema

```
SELECT AVG(StartTimeDifference), AVG(EndTimeDifference)
FROM FactWork
```

Aggregate

Entity Relationship

```
SELECT segment.name, SUM(data_value)
FROM segment, sg_completed_work, wo_step_num_data, work_num_data_type
WHERE segment.segment_site = sg_completed_work.segment_site
AND segment.segment_id = sg_completed_work.segment_id
AND sg_completed_work.work_order_db_site =
wo_step_num_data.work_order_db_site
AND sg_completed_work.work_order_db_id = wo_step_num_data.work_order_db_id
AND sg_completed_work.work_order_id = wo_step_num_data.work_order_id
AND sg_completed_work.work_ord_step_seq =
wo_step_num_data.work_ord_step_seq
AND wo_step_num_data.wn_db_site = work_num_data_type.wn_db_site
AND wo_step_num_data.wn_db_id = work_num_data_type.wn_db_id
AND wo_step_num_data.wn_type_code = work_num_data_type.wn_type_code
AND segment.name = 'Segment No. 1'
AND (work_num_data_type.name = 'Cost, Parts, Actual'
OR work_num_data_type.name = 'Cost, Consummables, Actual'
OR work_num_data_type.name = 'Cost, Internal Labor, Actual'
OR work_num_data_type.name = 'Cost, External Labor, Actual'
OR work_num_data_type.name = 'Cost, Extra Expenses, Actual')
AND sg_completed_work.gmt_completed >= DATEADD(yy, -5, GETDATE())
GROUP BY segment.name
ORDER BY SUM(data_value) DESC
```

Star Schema

```
SELECT DimSegment.Name, SUM(TotalCost)
FROM FactWork, DimSegment, DimTime
WHERE FactWork.LocationSegmentKey = DimSegment.SegmentKey
AND FactWork.ActualEndTimeKey = DimTime.TimeKey
AND DimSegment.Name = 'Segment No. 1'
AND DimTime.Date >= DATEADD(yy, -5, GETDATE())
GROUP BY DimSegment.Name
ORDER BY SUM(TotalCost) DESC
```

Large Sort

Entity Relationship

```
SELECT segment.user_tag_ident, COUNT(*)
FROM sg_completed_work, segment
WHERE sg_completed_work.segment_site = segment.segment_site
AND sg_completed_work.segment_id = segment.segment_id
GROUP BY segment.user_tag_ident
ORDER BY COUNT(*) DESC
```

Star Schema

```
SELECT DimSegment.Name, COUNT(*)
FROM FactWork, DimSegment
WHERE FactWork.LocationSegmentKey = DimSegment.SegmentKey
GROUP BY DimSegment.Name
ORDER BY COUNT(*) DESC
```

Appendix F Query Results

Data Set Specifications

	Data set							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Segments	50	100	100	500	500	500	1000	5000
Work Orders	100	500	1000	5000	10000	50000	100000	500000
Work Order Steps	400	2000	4000	20000	40000	200000	400000	2000000

Number of Statements Inserted

	Data set							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
ER	3050	14800	29300	146500	291500	1451500	2903000	14515000
MD	1575	5862	9227	27424	47455	207455	408451	2016456

Insertion Time in Seconds

	Data set							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
ER	3.67	21	45	323	706	4353	8857	44824
MD	2.22	9	15	57	102	547	1056	4682

Space Used in Megabytes

	Data set							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
ER	0.53	2.58	5.19	25.53	51	255	510	2557
MD	0.16	0.66	1.14	4.16	8	36	72	357

CPU Time in Seconds

		Data set							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Loosely constrained joins	**ER**	21.8	59.8	121.8	606	538	3625	8215	42944
	MD	6.2	31	22	47	66	247	484	2185
Tightly constrained joins	**ER**	12.6	40.8	75	531	1065	287	2575	2231
	MD	0	18.6	19	28	59	159	328	1612
Calculations	**ER**	6	31.2	74.8	219	444	2138	4172	32084
	MD	0	3.2	3.2	19	57	210	459	2278
Aggregations	**ER**	0	0	0	6.2	16	35	31	41
	MD	3.2	3.2	12.6	34	38	162	325	1466
Sorting	**ER**	0	6.2	9.2	38	50	209	494	2922
	MD	0	0	6.2	19	38	194	419	2103

Elapsed Time in Seconds

		Data set							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Loosely constrained joins	**ER**	67	134	259	1314	2108	14198	26529	135550
	MD	16	28	54	146	189	1002	1475	6238
Tightly constrained joins	**ER**	51	118	235	1222	2425	6734	12840	22893
	MD	31	40	86	166	248	895	1367	5562
Calculations	**ER**	16	56	167	689	1428	8490	25019	120192
	MD	12	15	32	99	168	855	1419	5981
Aggregations	**ER**	34	57	96	194	361	2009	1813	924
	MD	21	26	59	143	168	858	1429	5345
Sorting	**ER**	31	41	88	428	727	4237	19513	71675
	MD	25	27	42	130	194	902	1481	6097

Scan Count

		Data set							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Loosely constrained joins	**ER**	6	6	6	914	6	6	6	3005
	MD	2	3	3	3	3	3	3	3
Tightly constrained joins	**ER**	8	8	8	917	1799	1548	46	11661
	MD	2	5	5	5	5	5	5	5
Calculations	**ER**	2	2	2	2	2	2	2	2
	MD	1	1	1	1	1	1	1	1
Aggregations	**ER**	9	17	34	34	80	371	365	314
	MD	2	2	2	2	3	3	3	3
Sorting	**ER**	2	2	2	2	2	2	2	2002
	MD	2	2	2	2	2	2	2	2

Logical Reads

		Data set							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Loosely constrained joins	**ER**	167	279	546	44725	5200	26335	52716	279323
	MD	186	86	143	514	958	4514	8958	44514
Tightly constrained joins	**ER**	133	1078	2139	44760	88176	17363	35434	135496
	MD	66	92	149	527	971	4527	8979	44602
Calculations	**ER**	36	156	311	1496	2949	15149	30353	152431
	MD	9	45	89	445	889	4445	8889	44445
Aggregations	**ER**	97	224	350	347	788	3571	3653	3299
	MD	29	78	171	571	923	4479	8927	44518
Sorting	**ER**	25	98	194	910	1769	9218	18490	111659
	MD	12	48	92	452	896	4452	8900	44491

Physical Reads

		Data set							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Loosely constrained joins	**ER**	4	4	7	4	6	7	9	87
	MD	3	1	1	1	1	1	1	1
Tightly constrained joins	**ER**	6	7	8	7	7	24	19	1429
	MD	5	2	2	2	3	3	2	4
Calculations	**ER**	1	1	3	2	3	4	6	70
	MD	4	0	0	0	0	0	0	0
Aggregations	**ER**	6	14	4	4	2	4	48	35
	MD	5	4	5	1	1	1	2	2
Sorting	**ER**	1	1	3	3	3	1	5	77
	MD	4	0	0	1	1	1	1	1

Read-Ahead Reads

		Data set							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Loosely constrained joins	**ER**	63	270	527	2723	5278	26673	53064	283666
	MD	20	96	171	525	972	4524	8964	44504
Tightly constrained joins	**ER**	60	281	543	2757	5292	14651	36212	56388
	MD	8	94	163	541	982	4534	8993	44593
Calculations	**ER**	43	159	309	1582	3024	15206	31325	155884
	MD	0	63	112	456	903	4455	8892	44443
Aggregations	**ER**	3	24	232	378	846	4010	1650	1835
	MD	8	63	112	498	939	4491	8927	44507
Sorting	**ER**	32	107	195	1030	1876	9300	19435	104824
	MD	8	71	120	474	921	4473	8904	44491

Appendix G CBR Evaluation Task

Evaluation Design Tasks

You have been assigned the task of developing a data warehouse for a manufacturer of machinery, including motors, pumps, and gearboxes. The purpose of the data warehouse is to provide staff with sales and shipping information for their newly developed decision support system (DSS). The data warehouse backend requires the structure of data to be in a star schema format. Notation for table and attribute names should be meaningful (since this is a conceptual design), and can contain letters, numbers, and spaces.

Task 1

The data warehouse will need to store the purchase orders from each customer. Orders are defined by the order number, quantity of a product, and dollar amount of a product. Products have information such as the type, model number, dimensions (height, width, length, weight). Customer information will be transferred from the Customer Relationship Management system and includes the customer's name, organisation name, shipping address, billing address. Each customer is also allocated a sales representative that is allocated by their district. Warehouse promotions will often occur in order to move stock, and products purchased during this period will have a discounted price. Promotions have specific start and end dates, and the DSS should be able to calculate the revenue lost in running the promotion.

The following is a list of questions that the schema should be able to provide answers for:

1. What is the total revenue earned from motors during the last quarter?
2. Which are the highest selling products
3. Who has dropped out of our top ten customers over the past two years?
4. For each sales representative, which districts have been their most profitable?

5. Did the increase in sales numbers during promotions offset the reduced product cost and promotion cost during the last year?

Task 2

The data warehouse will also need to store the shipping information of each product to the customer. As the journey of a product to the customer can travel through multiple legs (intermediate stops), dates and locations of departure and arrival should be recorded for both the entire journey as well as per leg. Shipping can done through multiple means, including truck, ships, and aircraft. Fees associated with shipping need to be stored, so they can be charged to the customer's account. The role of people/departments/organisations should be included, such as the people/departments responsible for organising delivery, the people/companies shipping the item, and the recipient person/organisation.

The following is a list of questions that the schema should be able to provide answers for:

1. Which legs account for the largest percentage of time against the whole voyage?
2. How many products are currently being delivered?
3. Which legs consist of the highest fees?
4. Has the amount of shipping time and fees changed compared to five years ago?

Evaluation Survey

1. At what level would you rate your data modelling skills?

☐ Very Good

☐ Good

☐ Average

☐ Poor

☐ Very Poor

2. In which ways did the CBR system aid you? Tick all that apply.

☐ Forced critical thought about schema elements

☐ Improved the quality of the schema

☐ Quicker design of the schema

☐ Identified missing elements in specifications

☐ Other _____

3. What was the ease of use of the CBR system?

☐ Very Easy

☐ Easy

☐ Neutral

☐ Difficult

☐ Very Difficult

Evaluation Results

Fact attribute, dimension, and dimension attribute scores are listed in the form of *<total correct>/<correct + incorrect + missing>*.

Task 1

Participant	Method	Task Order	Duration (mins)	Fact Attributes	Dimensions	Dimension Attributes
1	Manual	1st	21	11/12	4/5	20/22
2	CBR	1st	21	4/4	5/6	16/18
3	CBR	2nd	16	5/6	4/5	0/15
4	Manual	2nd	24	3/4	4/5	11/14
5	Manual	1st	10	4/4	6/6	48/49
6	CBR	1st	27	3/4	5/6	21/22
7	CBR	2nd	20	4/4	4/5	32/32
8	Manual	2nd	30	4/4	3/6	12/17
9	Manual	1st	28	2/4	5/5	5/18
10	CBR	1st	18	0/4	5/5	15/15
11	CBR	2nd	17	4/4	5/5	22/22
12	Manual	2nd	28	2/4	3/6	5/22

Task 2

Participant	Method	Task Order	Duration (mins)	Fact Attributes	Dimensions	Dimension Attributes
1	CBR	2nd	13	4/5	4/6	20/22
2	Manual	2nd	16	1/2	5/6	9/13
3	Manual	1st	26	3/5	4/6	11/13
4	CBR	1st	21	1/3	4/5	25/30
5	CBR	2nd	10	5/5	5/5	26/26
6	Manual	2nd	-	-	-	-
7	Manual	1st	30	1/2	4/5	11/17
8	CBR	1st	29	1/3	4/5	26/28
9	CBR	2nd	23	2/3	5/5	26/26
10	Manual	2nd	17	0/3	2/5	9/9
11	Manual	1st	28	1/5	2/6	9/17
12	CBR	1st	21	5/5	5/5	26/26